Título	*Revista da Associação Portuguesa de Linguística*, n.º 10
Organização	Adelina Castelo Alexandra Fiéis Cristina Flores

Revisão	Ana Cristina Silva
Data	Outubro de 2023
ISSN	2183-9077

Nota prévia

O número 10 da *Revista da Associação Portuguesa de Linguística* é composto por 17 artigos selecionados, previamente apresentados no XXXVIII Encontro Nacional da Associação Portuguesa de Linguística, que teve lugar nos dias 26 e 28 de outubro de 2022 na Faculdade de Letras da Universidade de Lisboa.

Este encontro científico incluiu duas conferências plenárias: "O acento em português e a transferência fonológica em L2", na qual foi orador Guilherme D. Garcia, da Université Laval, e "Discourse Expectations in Causal Discourse", proferida por Oliver Bott, da Bielefeld University.

Participantes ligados a diferentes instituições de ensino superior e centros de investigação nacionais e estrangeiros apresentaram 38 comunicações orais e oito pósteres. Foi também neste Encontro Nacional da Associação Portuguesa de Linguística que foi inaugurada uma exposição em homenagem a Maria Helena Mira Mateus.

Os artigos publicados neste número 10 da Revista da Associação Portuguesa de Linguística mostram bem a diversidade de interesses e perspetivas teórico-metodológicas que podemos contemplar na investigação atualmente realizada em Portugal em Linguística e em áreas de interface, abrangendo áreas como Aquisição (L1 e L2), Fonologia, Linguística Clínica, Linguística Computacional, Linguística de *Corpus*, Linguística Educacional, Semântica, Sintaxe e Sociolinguística. Cada um destes artigos foi submetido a uma avaliação por revisores da respetiva área científica.

Além de agradecer a imprescindível colaboração dos revisores dos artigos agora publicados e o trabalho crucial da comissão científica que avaliou os resumos submetidos ao Encontro, gostaríamos de expressar a nossa gratidão também aos oradores convidados e a todos os participantes e assistentes, pelo enriquecimento resultante destas partilhas e discussões científicas. Finalmente, deixamos o nosso agradecimento à Comissão Organizadora do XXXVIII ENAPL.

As editoras
Adelina Castelo, *Alexandra Fiéis* e *Cristina Flores*

Nova versão do teste de repetição de pseudopalavras LITMUS-QU-NWR-EP

Letícia Almeida[1]*, Christophe dos Santos*[2]*, Maria João Freitas*[1]

[1]Universidade de Lisboa, Faculdade de Letras, Centro de Linguística
[2]UMR 1253, iBrain, Université de Tours, Inserm, Tours, France

Resumo

Neste artigo, apresentamos um teste de repetição de pseudopalavras (LITMUS-QU-NWR) criado para avaliar a fonologia de crianças monolingues e bilingues. Este teste, disponível em várias línguas, caracteriza-se por conter uma parte tendencialmente universal, ou seja, independente da língua-alvo, e uma parte dependente da língua-alvo. Apresentamos os passos de adaptação do teste ao português europeu e os critérios de seleção dos parâmetros fonológicos a serem incorporados na versão portuguesa. Estes foram motivados pelos padrões de aquisição fonológica exibidos por crianças monolingues portuguesas, assim como pelos resultados de um primeiro estudo piloto que levou a uma reformulação do teste.

Palavras-chave: Teste de repetição de pseudopalavras, aquisição fonológica, sílaba, complexidade fonológica, bilinguismo.

Abstract

In this paper, we present a nonword repetition test (LITMUS-QU-NWR) created to assess the phonology of monolingual and bilingual children. This test, available in several languages, has two parts. One part is quasi universal, i.e., independent of the target language, and the other part is language-dependent. We present the steps involved in the adaptation of the test into European Portuguese and the criteria for selecting the phonological parameters to be incorporated into the Portuguese version. Those were motivated by the phonological acquisition patterns exhibited by monolingual Portuguese children, as well as by the results of an initial pilot study that led to a reformulation of the test.

Keywords: Nonword repetition test, phonological development, syllable, phonological complexity, bilingualism.

1. Introdução

O uso de testes de repetição de pseudopalavras (RPP) é cada vez mais frequente na prática clínica, pois estes revelaram serem bastante eficazes para o diagnóstico de perturbação do desenvolvimento da linguagem (PDL) (Bishop et al., 1996). Recentemente, foi mostrado que alguns testes de RPP ainda apresentam a vantagem de neutralizar a variável bilinguismo, o que permite alargar a sua utilização a contextos em que as crianças avaliadas são expostas a mais do que uma língua (Almeida et al., 2019; Schwob et al., 2021). Entre os testes de RPP que tomam em consideração, na sua construção, os contextos bilingues, o teste LITMUS-QU-NWR (*Language Impairment Testing in a Multilingual Society - Quasi Universal - NonWord Repetition*) tem como objetivo principal avaliar a fonologia infantil através da complexidade fonológica. Este teste, que foi adaptado para várias línguas, mostrou ser eficaz em distinguir crianças com desenvolvimento típico de crianças com PDL, monolingues e bilingues, em francês, alemão e árabe libanês (Abi-Aad & Atallah, 2020; Grimm, 2022; Tuller et al., 2018). Em português europeu (PE), poucos são os testes de RPP que avaliam a complexidade

fonológica e, no nosso conhecimento, nenhum foi desenhado com o propósito de poder ser aplicado em contextos multilingues.

Este trabalho tem como objetivo principal apresentar a adaptação ao PE do teste de RPP LITMUS-QU-NWR. Assim, apresentaremos os critérios de seleção dos parâmetros fonológicos a serem incorporados na versão portuguesa e os critérios para a reformulação do teste após um primeiro estudo piloto (Catarino, 2019; Catarino et al., 2021).

1.1. **A pertinência dos testes de repetição de pseudopalavras na avaliação fonológica**

Tradicionalmente, a fonologia infantil é avaliada através de testes estandardizados que solicitam a produção de palavras, tipicamente em tarefas de nomeação de imagens ou de repetição (p.e., para o PE, o teste ALPE, Mendes et al., 2013). No entanto, estas tarefas apresentam duas desvantagens: estão baseadas no conhecimento lexical prévio das crianças e foram estandardizadas numa população de crianças monolingues. Existem, no nosso conhecimento, alguns testes estandardizados para crianças bilingues em alemão e em inglês-espanhol. No entanto, não existem testes de avaliação da fonologia estandardizados para crianças bilingues em PE. O facto de avaliarmos crianças bilingues com tarefas validadas e estandardizadas em crianças monolingues impede um diagnóstico válido das crianças bilingues pela ausência de normas específicas para essas crianças, que possuem conhecimento lexical noutra língua que não a avaliada. Isto poderá conduzir a um sub ou sobre diagnóstico de crianças bilingues (veja-se Armon-Lotem & Grohmann, 2021). Além disso, utilizar um teste de avaliação fonológica baseado em conhecimento lexical prévio das crianças coloca em desvantagem as crianças que, por alguma razão, tenham um léxico mais reduzido, uma vez que o conhecimento lexical previamente adquirido está correlacionado com o desempenho das crianças em tarefas que envolvam produção ou repetição de palavras. Ora, sabemos que as crianças bilingues possuem um léxico reduzido numa língua em comparação com os monolingues dessa mesma língua, uma vez que o seu conhecimento lexical está distribuído pelas duas línguas (Cattani et al., 2014). Além disso, as crianças com PDL também apresentam um défice de vocabulário, pelo que o uso de instrumentos que solicitem, além das suas capacidades fonológicas, as suas representações lexicais, não é o ideal. Assim, alguns autores têm desenvolvido testes de RPP: uma vez que estes testes têm por base unidades com uma estrutura fonológica característica da língua-alvo, estes permitem avaliar a fonologia sem recorrer a unidades lexicais da língua (Chiat, 2015).

No entanto, há que ressalvar que os testes de RPP não são todos semelhantes. Existem múltiplas variáveis manipuladas na sua construção que condicionam o formato final do teste e o que este avalia. Assim, no nosso conhecimento, os testes de RPP podem manipular as seguintes variáveis: a extensão da pseudopalavra, a proximidade lexical, a probabilidade fonotática e a complexidade fonológica. O resultado das variáveis manipuladas na construção do teste e o peso dado a cada variável resulta em testes de RPP diferentes que avaliam aspetos distintos. Por exemplo, um teste de RPP que manipula essencialmente a extensão da pseudopalavra está sobretudo a avaliar a memória fonológica. Um teste de RPP que manipula as variáveis proximidade lexical ou probabilidade fonotática recorre mais ao conhecimento lexical prévio. Finalmente, um teste centrado na complexidade fonológica é mais suscetível de avaliar a fonologia. Existem testes de RPP adaptados ao PE construídos em torno destas quatro variáveis: o teste EP-CNRep (Cruz-Santos, 2009) manipula a extensão de palavra; o Instrumento de Repetição de Pseudopalavras (Ribeiro, 2011) manipula as variáveis extensão de palavra, proximidade lexical e complexidade articulatória. O teste de Repetição de Pseudopalavras Linguística e Morfologicamente Motivadas (Coutinho, 2014) é construído em torno da probabilidade fonológica. Finalmente, o teste LITMUS-QU-NWR-EP (Almeida & dos Santos, 2015) foi construído controlando a complexidade fonológica (veja-se Catarino & Almeida, 2022, para uma descrição detalhada dos vários testes de RPP disponíveis em PE).

Além de serem testes de aplicação fácil e rápida, os testes de RPP apresentam vantagens comparativamente aos testes de vocabulário e de fonologia baseados no léxico: as crianças nunca ouviram as pseudopalavras antes, logo o seu desempenho não deverá ser afetado pela sua experiência linguística nem pelo

conhecimento lexical prévio. Isto torna estas tarefas particularmente interessantes para avaliar crianças que possam ter um inventário lexical reduzido, como é o caso das crianças com PDL e das crianças bilingues. Tendo em conta esta vantagem, o uso de testes de RPP pode servir dois objetivos: por um lado, identificar crianças com PDL e, por outro lado, poder ser aplicado em contextos bilingues.

Os testes de RPP têm sido cada vez mais utilizados na prática clínica, por conseguirem identificar crianças com PDL e, logo, funcionarem como potenciais instrumentos de rastreio. Com efeito, estas crianças costumam ter um desempenho baixo nas tarefas de RPP (Bishop et al., 1996; Conti-Ramsden et al., 2001). Além disso, estes testes também são utilizados em crianças com outras patologias em que pode existir défice fonológico, como na dislexia (Ramus et al., 2013) ou no autismo (Williams et al., 2013). Recentemente, algumas provas de RPP têm sido aplicadas em crianças bilingues e estas parecem ter o potencial de identificar crianças com PDL em situação de bilinguismo (veja-se Schwob et al., 2021, para uma revisão das provas de RPP utilizadas em contextos bilingues). No entanto, as características da prova de RPP utilizada parece influenciar o poder discriminatório da prova nos contextos multilingues (Schwob et al., 2021).

Na próxima secção, iremos descrever as provas de RPP LITMUS, cuja caraterística principal é a de terem sido criadas especificamente para avaliar a fonologia das crianças independentemente de uma eventual situação de bilinguismo.

1.2. **Os testes LITMUS-QU-NWR**

Como vimos na secção anterior, o número de instrumentos destinados a avaliar crianças monolingues e bilingues num só teste é escasso. Para remediar essa escassez, foi criada, na Europa, uma ação COST (*European Cooperation in Science and Technology)* sobre o tema do bilinguismo e das perturbações da linguagem (COST Action IS0804, 2009–2013). No âmbito dessa ação foram concebidos instrumentos de avaliação da linguagem que permitissem discriminar crianças bilingues com desenvolvimento típico e crianças bilingues com PDL nos domínios da sintaxe, do léxico e da fonologia. No caso da fonologia, a escolha recaiu sobre um instrumento de RPP, pois, como vimos acima, estes possuem a vantagem de não recorrerem ao conhecimento lexical previamente adquirido e, logo, permitirem não colocar em desvantagem as crianças bilingues em comparação com as crianças monolingues. Assim, os testes LITMUS-QU-NWR (Ferré & dos Santos, 2015) foram especificamente concebidos para avaliar a fonologia independentemente da língua falada pela criança.

Foi decidido que os testes LITMUS de RPP seriam focados na avaliação da complexidade fonológica através da escolha de estruturas silábicas e de fonemas específicos. Com efeito, estudos prévios tinham apontado para as dificuldades das crianças com PDL no processamento da estrutura silábica (Ferré et al., 2012).

Como referido na secção anterior, a maioria dos testes de RPP manipula a extensão da palavra, o que faz com que as capacidades de memória fonológica interfiram grandemente no sucesso neste tipo de prova. Como o objetivo principal era o de criar um teste que avaliasse as capacidades fonológicas dos participantes, optou-se por tentar reduzir o efeito da memória fonológica no teste, uma vez que essa influência pode esconder, pelo menos em parte, as habilidades fonológicas reais dos participantes. Para tal, optou-se por reduzir a extensão das pseudopalavras. Estudos prévios mostraram que o efeito da extensão da palavra começa a ocorrer assim que o item possui mais de duas sílabas (Poncelet & van der Linden, 2003). Por esta razão, a extensão dos itens a incluir no teste foi limitada a três sílabas, ou seja, a três vogais.

Para limitar o impacto do conhecimento lexical no desempenho da prova, as PP foram criadas a partir de blocos elementares de segmentos e tipos silábicos. Este método permitiu manipular os diferentes blocos a serem combinados, a fim de obter diferentes graus de complexidade fonológica (i.e., fricativa/oclusiva, ou, ausência/presença de encontros consonânticos tautossilábicos e heterossilábicos) sem recorrer a palavras existentes no léxico da língua-alvo (Ferré & dos Santos, 2015).

De forma geral, os testes LITMUS-QU-NWR são constituídos por dois tipos de itens a fim de avaliar, por um lado, o maior número de línguas possível e, por outro lado, a complexidade fonológica da língua avaliada. O primeiro tipo de item é considerado independente da língua (LI). Neste tipo, os itens possuem fonemas e

restrições fonotáticas presentes na maioria das línguas do mundo (Maddieson et al., 2013). No entanto, é de notar que estes itens na realidade são quase universais, e não totalmente universais, uma vez que é impossível abstrairmo-nos completamente da fonologia da língua-alvo. O segundo tipo de itens é considerado dependente da língua (LD). Neste tipo, os itens possuem características fonotáticas específicas de determinadas línguas. A descrição da escolha dos itens LD para o PE será efetuada na Secção 3.1.

Os itens da parte LI do teste, semelhantes em todas as versões, foram selecionados em torno de três variáveis: a sílaba, os segmentos e a sequencialidade. A variável acento não foi tomada em consideração, sendo que as adaptações do teste a várias línguas respeitam o padrão acentual da língua-alvo.

No que concerne à sílaba, num primeiro momento foram selecionados os tipos silábicos a integrarem a parte LI do teste. O tipo silábico considerado como mais simples, tanto do ponto de vista da tipologia linguística como da aquisição, constitui a base da parte LI. Trata-se do tipo silábico C(onsoante)V(ogal), considerado como o tipo silábico universal por excelência, pois faz parte do inventário silábico de todas as línguas do mundo descritas até à data. Foram adicionados dois outros tipos silábicos, CCV e CVC#, mais complexos do que o tipo silábico CV, que fazem parte do inventário silábico da maioria das línguas do mundo. Com efeito, analisado um conjunto de 515 línguas, 88% possuem ou um ataque ramificado (sílaba CCV) ou uma consoante em posição final de palavra (sílaba CVC#) (Maddieson et al., 2013). Assim, três tipos silábicos integram a parte LI do teste LITMUS-QU-NWR: CV, CCV e CVC#.

A nível segmental, a escolha das consoantes a integrarem a parte LI do teste também foi baseada na sua aquisição precoce e na sua presença na maioria das línguas do mundo. A escolha recaiu sobre duas consoantes oclusivas, [p] e [k], porque este modo de articulação é o primeiro a ser adquirido. Além disso, a grande maioria das línguas do mundo possui no seu inventário fonológico uma consoante oclusiva labial e uma consoante oclusiva dorsal, independentemente do vozeamento. O facto de o teste incorporar estas duas oclusivas permite contrastar os pontos de articulação labial e dorsal, sendo o primeiro menos complexo do que o segundo. Foi também introduzido no teste um contraste de modo de articulação com a inserção da consoante fricativa [f], sendo este último modo de articulação mais complexo do que o modo oclusivo. De modo a construir os ataques ramificados, ou seja, a sílaba CCV, a líquida [l] foi selecionada, em detrimento da rótica, por apresentar menos variabilidade fonética. Finalmente, apenas três vogais estão representadas no teste: [a], [i] e [u]. Estas constituem as vogais mais presentes nas línguas do mundo (Maddieson et al., 2013). Assim, em termos segmentais, a parte LI do teste possui 4 consoantes ([p, k, f, l]) e três vogais ([a, i, u]).

Em termos de sequencialidade, tanto a sequencialidade dos segmentos como a dos tipos silábicos pode aumentar a complexidade dos itens criados. No que toca aos segmentos, as sequências que alternam o modo e/ou o ponto de articulação são consideradas mais complexas. Por outro lado, os itens iniciados por uma consoante labial são considerados mais simples. Quanto à sequencialidade dos tipos silábicos, nos itens com três sílabas, a sílaba que ocorre em segunda posição é geralmente menos proeminente acusticamente e mais difícil de processar, principalmente se nela ocorrer um ataque ramificado. A Tabela 1 sintetiza as variáveis tidas em conta aquando da construção da parte LI do teste.

Tabela 1. Pontos de complexidade controlados na criação dos itens do teste original

	Segmentos		Sílabas (σ)	Sequências		
				Segmentos		Posição dos CCV e CVC quando 3 σ
- (menos)	Ocl	Lab	CV	Ponto e modo ident.	Lab-Dor	Em 1.ª e 3.ª posição
Exemplos	[p]	[p]	[ka]	[pupa]	[puka]	[klipafu]
+ (mais)	Fric	Dors	CCV e CVC#	Ponto e/ou modo dif.	Dor-Lab	Em 2.ª posição
Exemplos	[f]	[k]	[kla]	[puka] / [fuka]	[kapi]	vs [kuflapi]

2. **Dados da aquisição da fonologia em PE**

Nesta secção, forneceremos informação sumária sobre a aquisição, por crianças portuguesas monolingues, das estruturas fonológicas representadas no LITMUS-QU-NWR-PE, tanto as integradas na versão inicial da prova como as inseridas na versão que apresentamos no presente artigo (estruturas universais e estruturas específicas da língua alvo). Recorreremos a estudos realizados com base em dados longitudinais naturalistas (Correia, 2009 (5 crianças); Costa, 2010 (5 crianças); Freitas, 1997 (7 crianças); Santos, *em preparação* (3 crianças)) e a estudos realizados a partir de dados transversais experimentais (Guimarães et al., 2014 (cerca de 400 crianças); Amorim, 2014 (80 crianças); Mendes et al., 2013 (cerca de 1000 crianças); Ramalho, 2017 (87 crianças)). Note-se que diferentes faixas etárias para limiar de aquisição nos diferentes estudos citados decorrem de procedimentos distintos tanto nas recolhas de dados como no seu registo e tratamento. A título ilustrativo, diferentes instrumentos são usados para elicitação da produção, sendo que o instrumento em Ramalho (2017) é fonologicamente mais complexo do que os usados em Mendes et al., (2013), em Guimarães et al., (2014) e em Amorim (2014), o que pode ter impacto na identificação da aquisição das várias estruturas por faixa etária.

Como referimos, do ponto de vista da estrutura da palavra prosódica, as pseudopalavras nas provas LITMUS são monossílabos, dissílabos e trissílabos, não podendo exceder mais do que 3 núcleos preenchidos e 7 segmentos. Sabemos que estas três extensões de palavra estão disponíveis nos sistemas das crianças portuguesas antes dos 3;0 de idade e que não existem restrições ao uso de palavra mínima no estádio inicial (Ramalho, 2017; Vigário et al., 2006).

Quanto aos padrões acentuais existentes em PE, pé jâmbico e trocaico na periferia direita da palavra, sabemos, com base em dados longitudinais naturalistas, que ambos são dominados pelas crianças portuguesas antes dos 3;0 (Correia, 2009). Os trabalhos transversais experimentais desenvolvidos para o PE (todos a partir dos 3;0) não relatam problemas de aquisição relativamente ao acento de palavra. Note-se que as palavras terminadas em /l/ ou /ɾ/ e acentuadas foneticamente na última sílaba em PE podem desencadear inserção de vogal em final de sintagma entoacional, [ɨ] no PE padrão, vogal essa que preenche prosodicamente a categoria vazia que constitui o marcador de classe nestas palavras (mar /maɾ+Ø/ [máɾ]/ [máɾɨ]; sol /sɔl+Ø/ [sɔ́ɬ]/ [sɔ́lɨ]). Esta inserção de vogal desencadeia ressilabificação da consoante final (de coda para ataque), ilustrada na alteração do alofone da lateral (em PE, [l] em ataque e [ɬ] em coda) e legitima a análise fonológica destas

palavras como troqueus, estando o marcador de classe vazio associado à posição rítmica fraca do pé trocaico final (Mateus & Andrade, 2000). Estes factos da gramática alvo poderão estar na base da ordem de aquisição coda final >> coda medial para /l/ e /ɾ/, registada na tabela 3, uma vez que as crianças portuguesas começam por produzir /l/ e /ɾ/ finais com paragoge de vogal (Freitas, 1997), em conformidade com o alvo.

Ainda em termos prosódicos, referimos acima que o LITMUS-QU-NWR-PE usa os padrões silábicos CV, CCV e CVC(#). Como para a aquisição de todas as línguas do mundo, sabemos, com base em dados longitudinais naturalistas recolhidos a partir dos 0;10 meses, que o padrão CV está disponível em PE desde o início da aquisição (Freitas, 1997). A ramificação dos constituintes silábicos acontece gradualmente e seguindo a ordem rima ramificada >>ataque ramificado. A ramificação da rima dá-se por disponibilização da coda, que hospeda inicialmente a fricativa palatal /ʃ/ (representada como segmento subespecificado, /S/ ou /s/, conforme os autores (Mateus & Andrade, 2000; Rodrigues, 2003)), sendo a variante alofónica [ʃ] a mais frequente, no sistema alvo e nos dados das crianças. Sabemos, também com base em dados longitudinais naturalistas, que a coda fricativa em final de palavra é adquirida por volta dos 2;0, provavelmente, por efeito da interface fonologia-morfossintaxe na aquisição: a coda fricativa final em PE desempenha, maioritariamente, o papel de marcador de número plural nos não verbos e de marcador de pessoa/número no sistema verbal, sendo, por isso, *locus* para ativação de concordância sintática, o que lhe confere proeminência gramatical no sistema alvo, podendo a confluência de informação fonológica e morfológica numa mesma estrutura desencadear a sua aquisição precoce (Freitas, 1997). Posteriormente, ocorre a sua aquisição em posição medial, em média, já na faixa etária dos 3;0 – 3;6 (Guimarães et al., 2014; Mendes et al., 2013), embora Ramalho (2017) registe a estabilização da coda fricativa medial na faixa etária dos 4;0 – 4;11. Por fim, a coda passará a acomodar, por esta ordem, a vibrante /ɾ/ e a lateral /l/. Ainda no domínio da rima, o núcleo ramificado de tipo VG é de emergência precoce mas de aquisição tardia (Correia, 2004; Freitas, 1997). A fricativa palatal ocorre também, como noutras línguas, nos chamados grupos sC em posição inicial de palavra, como em *escova* [ʃkóvɐ]. Estas estruturas sC, presentes no teste LITMUS, são adquiridas cedo em PE (Freitas & Rodrigues, 2003), normalmente na faixa etária dos 3;0-3;6, contrariamente ao que acontece noutras línguas, nas quais a sua aquisição é tardia (veja-se o caso do neerlandês, em Fikkert, 1994).

Na Tabela 2, demonstramos a aquisição lenta de ditongos no PE, com base em dados longitudinais naturalistas (Freitas, 1997), com intervalos etários de cerca de 1 ano a não revelarem aumento das taxas de sucesso.

Tabela 2. Aquisição de ditongos em PE (desenvolvimento típico)

Crianças	Intervalo etário	% de sucesso
João	1;11 – 2;8	52% – 61%
Inês	1;8 – 1;10	57% – 63%
Marta	1;2 – 2;2	51% – 79%
Luís	1;10 – 2;11	76% – 66%
Raquel	1;10 – 2;10	52% – 63%
Laura	2;3 – 3;3	69% – 69%
Pedro	2;7 – 3;7	61% – 75%

Por fim, e em termos globais, é adquirido o ataque ramificado, com um estudo a registar a faixa etária dos 4;0 – 5;0 (Mendes et al., 2013), mas outros reportando o facto de o processo de aquisição ser lento, terminando após os 6;0 (Amorim, 2014; Guimarães et al., 2014; Ramalho, 2017). Quando consideradas as consoantes associadas a C2, os relatos na literatura não são coincidentes: Mendes et al., (2013) e Amorim (2014) referem que /l/ em C2 favorece a aquisição, quando comparado com /ɾ/ em C2, reportando a ordem C/l/ >> C/ɾ/; contrariamente, Ramalho (2017) relata a ordem C/ɾ/ >> C/l/, com os ataques com lateral a não atingirem o limiar dos 50% na faixa etária dos 5;0 – 6;0, o que contrasta com os ataques com /ɾ/, com 69% na mesma faixa etária (diferenças estatisticamente significativas). Na base desta diferença estarão, provavelmente, questões metodológicas relacionadas com os juízos de valor dos avaliadores relativamente à epêntese de vogal entre C1 e C2, frequente nos enunciados infantis da fase final de aquisição e, segundo Veloso (2006), presente no sistema alvo em estruturas com a lateral em C2. Tal pode ter levado os terapeutas da fala em Mendes et al. (2013) a considerarem formas como [fiɫóɾ] para *flor* como sucessos; tal não aconteceu em Ramalho (2017), o que conduziu à identificação de taxas de sucesso mais baixas para CCV. Também Amorim (2014) excluiu das suas contagens todas as formas produzidas com epêntese de vogal entre C1 e C2. Como referido em Catarino et al. (2021), dado que a epêntese da vogal origina uma configuração fonológica não convergente com a do sistema adulto (C1V.C2V), a sua produção pode ser, no nosso entender, contabilizada como desviante. Tais diferenças de critérios metodológicos na análise dos dados infantis têm, naturalmente, impacto na identificação das ordens de aquisição descritas.

A maior parte dos trabalhos tem demonstrado o impacto da estrutura silábica na aquisição segmental em PE, tal como relatado para outras línguas (Almeida, 2011; dos Santos, 2007; Fikkert, 1994; Polo, 2018; Rose, 2000). Na Tabela 3, sintetizamos informação sobre as idades de aquisição de /ʃ, ɾ, l/ em coda e em ataque ramificado, considerando o efeito da interface segmento – estrutura silábica.

Tabela 3. Idades de aquisição em coda e em ataque ramificado em PE (desenvolvimento típico)

Estruturas	Mendes et al. (2013)	Guimarães et al. (2014)	Amorim (2014)	Ramalho (2017)
Coda fricativa	3;6 – 4;0	3;0 – 3;5	3;0 – 3;5	Até 4;0 – 4;11
Coda vibrante	4;0 – 4;6	4;0 – 4;5	medial: 4;6 – 4;11 final: 4;0 – 4;5	medial: após 6;0 final: 4;0 – 4;11
Coda lateral	5;0 – 5;6	após 5;11	medial: depois de 5;0 final: 4;0 – 4;5	após 6;0
Ataque ramificado com vibrante	4;6 – 4;11	após 5;11	inicial: 4;6 – 4;11 medial: após 5;0	após 6;0
Ataque ramificado com lateral	4;0 – 4;6	*Não testado*	4;0 – 4;5	após 6;0

Três outras consoantes integram o inventário de consoantes nas provas LITMUS: /p, k, f/. No caso do PE, sabemos que cada uma destas consoantes pode ocorrer em ataque simples ou como C1 de ataque ramificado. Em ambos os casos, estão adquiridas antes dos 3;0 anos de idade (Amorim, 2014; Costa, 2010; Guimarães et al., 2014; Mendes et al., 2013; Ramalho, 2017).

Por fim, e com base em dados longitudinais naturalistas (3 crianças portuguesas monolingues avaliadas entre os 0,11 e os 4;10, em Santos, *em preparação*), podemos colocar a hipótese de as vogais /a, i, u/ em posição tónica serem adquiridas antes dos 3;0, na ordem /a/ >> /i/ >> /u/. Já em posição átona, a vogal alta /i/ parece ser adquirida tardiamente (nos dados em foco, aos 4;1 na Inês e não adquirida aos 4;10 na Joana, data do final da recolha). Nos casos de /a/-[ɐ] e de /u/-[u] em posição átona, a aquisição parece depender do contexto morfológico, como observado em Freitas (2004, 2007). Em posição final de palavra, /u/-[u] e /a/-[ɐ] átonos são maioritariamente marcadores de classe e adquiridos antes das mesmas vogais em posição pré-tónica no radical. No caso do /a/-[ɐ] átono, a aquisição de ambas as estruturas dá-se até aos 2;6; quanto ao /u/-[u] átono, existe um intervalo de tempo substancial entre a aquisição de /u/ como marcador de classe (antes dos 2;0) e como parte de radical (na faixa etária dos 3;6 – 4;0).

3. A adaptação do teste LITMUS-QU-NWR ao PE

3.1. LITMUS-QU-NWR-EP – Versão 1.1

3.1.1. Descrição da prova

Como cada língua tem a sua própria fonologia, foi necessário fazer alguns ajustes para ter um instrumento adaptado à língua em que a adaptação estava a ser realizada, ou seja, neste caso, o PE. Como descrito acima, o teste original apresenta as seguintes consoantes [p, t, k, f, l], bem como a fricativa [s], incorporada na parte LD do teste original, em posição de coda interna. No entanto, em PE, a única fricativa que pode ocupar a posição de coda é [ʃ] (Mateus & Andrade, 2000). Por isso, um dos primeiros ajustes realizados foi a substituição da fricativa anterior [s] da versão original pela fricativa posterior [ʃ].

Relativamente à acentuação na versão original (francesa), cujos itens têm um número máximo de três vogais, o acento foi sempre atribuído à última sílaba, uma vez que o acento de palavra em francês recai sempre sobre a última sílaba de grupo entoacional (Di Cristo, 1999). Em PE, nas palavras com mais de uma sílaba, a maioria dos substantivos, adjetivos e advérbios é acentuada na penúltima sílaba, ou seja, existe maioritariamente um padrão acentual trocaico (Mateus & Andrade, 2000). No entanto, existem palavras com mais de uma sílaba cujo acento se encontra na última sílaba ('peru' [pɨ.ˈ**ru**], 'javali' [ʒɐ.vɐ.ˈ**li**]) ou seja, um padrão acentual iâmbico. O acento, quando a palavra tem três ou mais sílabas, também pode ser encontrado na antepenúltima sílaba: 'árido' [ˈ**a**.ri.du]). De acordo com Vigário et al. (2006), o acento de palavra tem a seguinte distribuição, por frequência, em PE: proparoxítonas – 1,99%; paroxítonas – 76,44%; oxítonas – 21,6%. Assim, na primeira adaptação da prova LITMUS ao PE, optou-se por alternar os padrões acentuais troqueu e iâmbico e por não incorporar os casos de acentuação na antepenúltima sílaba, residuais na língua (1,99%). De modo a respeitar as propriedades fonotáticas do PE, apenas os itens terminados em [ɐ] e [u] foram criados com um padrão acentual trocaico. Enfim, a fim de controlar o acento, os pares de itens com o mesmo número de sílabas e os mesmos tipos silábicos (e.g., #CV.CCV.CV#.) foram criados com o mesmo padrão acentual (e.g., [piklɐˈflu] e [kuflɐˈpi] vs [fiˈkuplɐ] e [kuˈpiflɐ]). Ao todo, 40% dos itens polissilábicos receberam o padrão acentual trocaico, abaixo da frequência na língua (76,44%).

A prova original LITMUS, como já foi referido, divide-se em duas partes. Uma parte LI, comum a todas a versões, e uma parte LD. A parte LD contém as características específicas da língua em que o teste é efetuado. O teste original manteve a possibilidade de ter itens com uma fricativa na coda interna e no final da palavra (e.g., [kiʃpa], [kifuʃ]), bem como uma fricativa na posição inicial. Do teste original, foi conservada a possibilidade de ter itens com uma fricativa como coda interna e como coda em final da palavra, bem como uma fricativa em posição inicial de grupo consonântico, no início de palavra (e.g., [ʃkla], [ʃkafu]). Além disso, para esta adaptação, introduzimos a possibilidade de criar itens contendo um núcleo ramificado, uma das especificidades fonológicas do português. Deste modo, nos itens criados podem ser encontrados núcleos ramificados após ataques simples ou complexos (e.g., [klaw] ou [kiw]). Assim, os formatos silábicos característicos da parte LD para o PE são CVC# (em que a última consoante pode ser a fricativa palatal), CVC no meio de palavra (com a lateral e a fricativa em posição de coda medial), CCVC#, ou seja, monossílabos com ocorrência simultânea de ataque ramificado e de coda final, CVG, CVGC (em que a última consoante é a fricativa palatal), sCV e sCCV.

Esta adaptação inicial foi usada num estudo piloto realizado por Catarino (2019), que confirmou a utilidade da tarefa para o PE. Neste estudo foram observadas crianças portuguesas monolingues com desenvolvimento típico e crianças com PDL fonológica. Os dados revelaram que o LITMUS-NWR-EP 1.1 permitiu discriminar perfis de desenvolvimento típico e atípico. Foram analisados em detalhe itens com ataques ramificados de tipo *Obstruinte+Lateral*; a estrutura apresentou-se como um potencial marcador clínico para a identificação de PDL em PE. Na sequência dos resultados de Catarino (2019), alguns ajustes pareceram necessários, como adiante veremos.

3.1.2. **Aplicação da prova**

A prova LITMUS-QU-NWR-PE (1.1) foi testada em Catarino (2019), numa amostra de 21 crianças monolingues, residentes em Lisboa, com desenvolvimento fonológico típico (DT), previamente identificado como tal pelos professores/educadores e pelos encarregados de educação dos participantes, não tendo as crianças sido avaliadas com outras provas antes e/ou após a aplicação do LITMUS-QU-NWR-PE. A amostra encontra-se sumariamente caracterizada na Tabela 4 (adaptada de Catarino et al., 2021; para mais informações, consulte-se Catarino, 2019).

Tabela 4. Constituição da amostra para o DT em Catarino (2019)

	N	Faixa etária	Escolaridade	Alfabetização	Contacto formal com pseudopalavras
G1 - PrEs	9	5;08 – 6;04	Pré-escolar	Não	Não
G2 – 2.º Ano	12	7;05 – 8;03	1.º Ciclo: 2.º Ano	Sim	Sim

O Grupo Pré-Escolar (PrEs), à data da recolha, não tinha sido exposto ao código escrito em situação formal de sala de aula, nem a tarefas de pseudopalavras. O Grupo – 2.º Ano tinha sido exposto, antes da data da recolha, a atividades de leitura, de escrita e de manipulação silábica de pseudopalavras, decorrentes das Metas Curriculares para o 1.º Ciclo do Ensino Básico (Buescu et al., 2012, 2015).

A prova foi ainda aplicada a 4 crianças com desenvolvimento fonológico atípico (DA), diagnosticadas formalmente por terapeuta da fala com Perturbação do Desenvolvimento da Linguagem (PDL fonológica), com idades compreendidas entre os 5;11 e os 10;01 e tempos de intervenção em terapia da fala entre os 6 e os 12 meses, à data da recolha dos dados. A Tabela 5 caracteriza a amostra com DA e foi retirada de Catarino (2019: 109).

Tabela 5. Constituição da amostra para o DA em Catarino (2019)

	S1	S2	S3	S4
Sexo	M	M	F	M
Idade	6;02	5;11	10;01	6;11
Escolaridade	JI	JI	5.º ano	1.º ano

3.1.3. Resultados globais para o LITMUS-QU-NWR-PE (1.1)

Apresentamos, nesta secção, alguns resultados obtidos na aplicação do LITMUS-QU-NWR-PE em Catarino (2019) que nos permitem propor alterações à versão 1.1 da prova.

Para a amostra relativa ao desenvolvimento fonológico típico, foram identificadas, em Catarino (2019), taxas globais de sucesso de 53,4% para o G1 e de 75,1% para o G2. Quanto às quatro crianças com PDL fonológica, as taxas de acerto por sujeito foram as seguintes: S1 = 36,2%; S2 = 37;7%; S3 = 41,4%; S4 = 23,2%. Estes valores são muito baixos, quando comparados com os de estudos anteriores conduzidos noutras línguas: em francês, Ferré et al. (2015) notam que a taxa global de acerto é de 90,6% para o LITMUS-QU-NWR-FR nas crianças monolingues francófonas com cinco e seis anos de idade, com desenvolvimento típico. As crianças monolingues francófonas com PDL também apresentam taxas de acerto superiores às quatro crianças avaliadas em Catarino (2019), atingindo, em média, 53,2% de sucesso na prova. Os resultados das crianças portuguesas monolingues com DT para o LITMUS-QU-NWR-PE também estão abaixo daqueles obtidos com outros instrumentos de RPP em PE. Ribeiro (2011) reporta entre 80% e 90% de acerto em crianças portuguesas monolingues entre os seis e os oito anos de idade, com desenvolvimento típico. Estas taxas de acerto baixas, apesar de discriminarem entre as crianças com DT e as crianças com PDL, levaram-nos a considerar reformular o teste, como referido em Catarino et al. (2021). Como não queríamos aumentar o número de itens do teste, decidimos examinar o comportamento das crianças face às estruturas nele presentes.

No desenvolvimento típico, a coda /l/ revelou-se problemática (taxas de acerto: G1 = 55%; G2 = 67%), tal como o ataque ramificado com /l/ (taxas de acerto: G1 = 61%; G2 = 85%). Por esta razão, decidimos manter ambas as estruturas, uma vez que elas parecem complexas para as crianças.

Os grupos consonânticos de tipo #[ʃ]C (*n* = 9) mostraram um efeito de teto, com 100% de produção conforme ao alvo nas crianças com DT e 80% nas crianças com PDL.

Os monossílabos de tipo CCV (*n* = 3) apresentaram taxas de acerto de 90,4% nas crianças com DT e de 91.6% nas crianças com PDL, não revelando capacidade de discriminação entre os dois grupos de crianças.

Por fim, os ditongos (*n* = 8) também atingiram taxas elevadas nos dois grupos de crianças avaliadas, com 88.1% de acerto nas crianças com DT e 80.6% nas crianças com PDL. Tendo em conta que as taxas obtidas para estas três estruturas não permitiram discriminação entre os dois grupos de crianças, optámos por eliminá-las do teste. Assim, optámos por inserir, na versão 1.2 do instrumento, outras estruturas específicas do PE, apresentadas na secção seguinte.

3.2. **LITMUS-QU-NWR-EP – Versão 1.2**

Com base nos resultados observados nas crianças portuguesas, foi necessária uma reflexão sobre uma reformulação do teste previamente a um processo de redução do mesmo, seguindo o exemplo do LITMUS original, cuja versão final (versão 3) contém apenas 30 itens, o que facilita a sua aplicação em contexto clínico (dos Santos et al., 2020; dos Santos et al., 2021). Em qualquer adaptação de um instrumento de avaliação, o primeiro passo consiste em identificar os itens pouco ou não informativos, ou seja, aqueles que são muito bem-sucedidos (acima de 90% de sucesso) ou muito malsucedidos (abaixo de 10% de sucesso). Em ambos os casos, a variação entre as crianças será nula ou muito reduzida. Por conseguinte, estes itens não serão eficazes para discriminar entre crianças com e sem dificuldades (Dickes et al., 1994). Assim, os itens monossilábicos com o formato CCV e os itens com estruturas sC em início de palavra, que obtiveram ambos taxas de acerto acima de 90%, revelando-se, portanto, não informativos, foram retirados do teste. É também de salientar que as estruturas sC em início de palavra são adquiridas relativamente cedo em PE (Freitas, 1997). Os itens com núcleo ramificado também foram retirados do teste. É de salientar que, embora a taxa de sucesso dos itens que contêm esta estrutura seja ligeiramente inferior a 90%, estes itens não permitem apresentar qualquer diferença significativa entre o grupo DT e o grupo PDL. Assim, os formatos silábicos CVG, CVGC, sC, sCC e #CCV# foram retirados da segunda versão do teste.

Retirar da segunda versão da tarefa as estruturas sC em início de palavra e os núcleos ramificados permite acrescentar pontos de complexidade específicos do PE que não estavam incluídos na primeira versão. Na primeira versão da adaptação, decidimos seguir a versão original, que tinha descartado a vibrante /ɾ/ para a criação de itens, em grande parte devido à sua elevada variabilidade (Walsh Dickey, 1997; Wiese, 2001). Na nova versão da adaptação do PE, decidimos finalmente introduzir esta vibrante na parte LD do teste porque esta consoante parece ter características específicas em PE que podem ser relevantes para a avaliação. Em primeiro lugar, seja em posição de ataque ramificado ou em posição de coda, a vibrante tem uma maior representatividade do que a lateral /l/. Outro aspeto interessante é o facto de haver uma assimetria entre a sua aquisição e a da lateral (as estruturas complexas com vibrante /ɾ/ sendo adquiridas mais cedo do que aquelas com lateral; Amorim, 2014; Mendes et al., 2013; Ramalho & Freitas, 2018), o que pode talvez refletir-se no desempenho de uma tarefa de repetição de pseudopalavras. Enfim, do ponto de vista teórico, existem diferentes propostas sobre a silabificação da sequência *Obstruinte+Líquida* na posição inicial de palavra (#OL). Uma primeira posição teórica considera que todas as sequências #OL são tautossilábicas (Mateus & Andrade, 2000), porque, por um lado, seguem o princípio da sonoridade (Selkirk, 1984) e, por outro, cumprem a condição de dissemelhança da sonoridade (Selkirk, 1984; Vigário & Falé, 1994). Uma segunda posição teórica postula que as sequências #Obstruinte-Vibrante (#OVib) e #Obstruinte-Lateral (#OLat) não silabificam da mesma maneira (Veloso, 2006). As sequências #OVib seriam tautossilábicas e as sequências #OLat seriam heterossilábicas. Veloso (2006) elabora três argumentos para defender a sua hipótese. Em primeiro lugar, em termos diacrónicos, as sequências #OLat do latim desapareceram no galego-português (lat. 'clamar', 'blandum '→ [ʃɐˈmaɾ], [ˈbrɐ̃du]). Apenas as sequências #OVib foram preservadas. Em segundo lugar, após o reaparecimento das sequências #OLat em português, por exemplo como resultado da relatinização na fixação renascentista da

norma escrita, foram observados casos de epêntese vocálica entre a obstruinte e a lateral, por exemplo na poesia ('flor '→' f[ɪ]lor'). Finalmente, com base num estudo sobre a segmentação silábica no início da escolaridade, Veloso (2006) observa 88.6% de divisão heterossilábica para #OLat e somente 32.4% para #OVib. Esta possível diferença de segmentação poderia também intervir no desempenho de uma tarefa de repetição de pseudopalavras. Assim, estes argumentos levaram-nos a introduzir a vibrante /ɾ/ na nova versão do teste. A vibrante foi inserida para criar itens da parte LD: foi utilizada para criar itens com ataque ramificado e com coda final (formatos CCV e CVC#, já presentes na parte LI). Também foi utilizada para criar itens com coda interna (formato CVC) e monossílabos com o formato CCVC. Ou seja, nesta segunda versão da prova, o formato silábico foi reduzido (retiraram-se 5 tipos silábicos) mas a diversidade segmental foi acrescida, com a introdução da vibrante alveolar.

Por fim, como, em PE, o padrão trocaico é o mais frequente (Mateus & Andrade, 2000), para que a nova versão do teste tenha uma distribuição acentual mais próxima da língua-alvo, a proporção de troqueus presentes no teste foi aumentada para 70%, o que vai ao encontro da frequência registada para o PE em Vigário et al. (2006) – 76,44%. Na Tabela 6, sintetizamos as propriedades fonológicas dos itens presentes na versão 1.2 do LITMUS-QU-NWR-EP, dando um exemplo para cada estrutura.

Tabela 6. Propriedades dos itens do LITMUS-QU-NWR-EP (1.2)

	CV	**CCV**	**CVC#**	**CCV e C#**	**#CCVC#**	**CVC$_m$**
Exemplo	[ˈlifɐ]	[ˈflikɐ]	[faɬ]	[fluˈkif]	[fluʃ]	[ˈpuɾfɐ]
n	6	22	22	2	8	12

Com estes ajustes, esperamos ter melhorado o teste, tanto do ponto de vista clínico como teórico. Do ponto de vista clínico, esperamos que esta nova versão aumente a qualidade de discriminação do teste entre crianças com e sem PDL. A par deste objetivo clínico, esperamos que o teste possa contribuir para o debate teórico sobre uma potencial diferença de segmentação entre as sequências #OLat e #OVib.

4. Considerações finais

Neste trabalho, apresentámos um instrumento de RPP, o LITMUS-QU-NWR, assim como os passos da sua adaptação ao PE. Assim, uma primeiro versão do instrumento foi aplicada a crianças com DT e a crianças com PDL (Catarino, 2019), e os resultados desse estudo piloto levaram-nos a reformular o instrumento, tendo em consideração as propriedades fonológicas do PE, os padrões de aquisição exibidos por crianças monolingues portuguesas e diversas propostas teóricas sobre o funcionamento do PE.

Acreditamos que o LITMUS-QU-NWR-EP possa ser um instrumento de RPP pertinente no contexto português, por controlar a complexidade silábica e poder ser aplicado tanto a crianças monolingues como bilingues, com e sem PDL. Tendo em conta os resultados promissores da prova nas outras línguas para que foi traduzida (Abi-Aad & Atallah, 2020; Almeida et al., 2016; Grimm, 2022), esperamos que a versão portuguesa contribua positivamente para a área da linguística clínica e teórica, possibilitando identificar crianças com PDL e testar hipóteses teóricas sobre fonologia. Depois de esta versão ser validada, o próximo passo será criar uma versão reduzida, que possa ser aplicada ainda mais facilmente e rapidamente em contexto clínico.

Referências

Abi-Aad, Karine & Christel Atallah (2020) L'épreuve répétition de non-mots : LITMUS-NWR-LIBAN. In Racha Zebib, Philippe Prévost, Laurie Tuller & Guillemette Henry (orgs.), *Plurilinguisme et troubles spécifiques du langage au Liban*. Presses universitaires de l'Université Saint-Joseph, pp. 79–92.

Almeida, Letícia (2011) *Acquisition de la structure syllabique en contexte de bilinguisme simultané portugais-français.* Dissertação de doutoramento, Universidade de Lisboa.

Almeida, Letícia, Sandrine Ferré, Marie-Anne Barthez & Christophe dos Santos (2019) What do monolingual and bilingual children with and without SLI produce when phonology is too complex? *First language* 39 (2), pp. 158–176.

Almeida, Letícia, Sandrine Ferré, Eléonore Morin, Philippe Prévost, Christophe dos Santos, Laurie Tuller & Racha Zebib (2016). L'identification d'enfants bilingues avec Trouble Spécifique du Langage en France. *SHS Web of Conferences*, 27, 10005. https://doi.org/10.1051/shsconf/20162710005

Almeida, Letícia & Christophe dos Santos (2015) *LITMUS-QU-NWR-EP (European Portuguese)*. Tese de mestrado, Université François-Rabelais de Tours.

Amorim, Clara (2014) *Padrão de aquisição de contrastes do PE: a interação entre traços, segmentos e sílabas*. Dissertação de doutoramento, Universidade do Porto.

Armon-Lotem, Sharon & Kleanthes K. Grohmann (2021) *Language Impairment in Multilingual Settings. LITMUS in action across Europe.* John Benjamins Publishing Company.

Bishop, Dorothy V., Tony North & Chris Donlan (1996) Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology & Psychiatry* 37 (4), pp. 391–403. https://doi.org/10.1111/j.1469-7610.1996.tb01420.x

Buescu, Helena Carvalhão, José Morais, Maria Regina Rocha & Violante F. Magalhães (2012) *Metas curriculares de português – Ensino Básico: 1.º, 2.º e 3.º ciclos propostas pela equipa de português*. Ministério da Educação e da Ciência. Disponível em http://static.publico.pt/docs/educacao/metasportugues.pdf

Buescu, Helena Carvalhão, José Morais, Maria Regina Rocha & Violante F. Magalhães (2015) *Programa e metas curriculares de português do Ensino Básico*. Ministério da Educação e da Ciência. Disponível em https://www.dge.mec.pt/sites/default/files/Basico/Metas/Portugues/pmcpeb_julho_2015.pdf

Cattani, Allegra, Kirsten Abbot-Smith, Rafalla Farag, Andrea Krott, Frédérique Arreckx, Ian Dennis & Caroline Floccia (2014). How much exposure to English is necessary for a bilingual toddler to perform like a monolingual peer in language tests? I*nternational Journal of Language & Communication Disorders* 49 (6), pp. 649–671. https://doi.org/10.1111/1460-6984.12082

Catarino, Inês (2019) *Produção de consoantes em ataque no 1o ciclo do Ensino Básico, em contexto de repetição de pseudopalavras*. Dissertação de mestrado, Universidade de Lisboa.

Catarino, Inês & Letícia Almeida (2022) A repetição de pseudopalavras na avaliação fonológica clínica. In Maria João Freitas, Marisa Lousada & Dina Caetano Alves (orgs.), *Linguística clínica: Modelos, avaliação e intervenção*. Language Science Press, pp. 211–234. https://doi.org/10.5281/zenodo.7233229

Catarino, Inês, Letícia Almeida, Christophe dos Santos & Maria João Freitas (2021) Sobre o impacto da constituência silábica na repetição de pseudopalavras: Dados preliminares do desenvolvimento típico para a validação do LITMUS-QU-NWR-EP. *Revista da Associação Portuguesa de Linguística,* 8, pp. 87–104. https://doi.org/10.26334/2183-9077/rapln8ano2021a7

Chiat, Shula (2015) Nonword repetition. In Sharon Armon-Lotem, Jan de Jong & Natalia Meir (orgs.), *Assessing multilingual children: Disentangling bilingualism from language impairment*. Multilingual Matters, pp. 123–148.

Conti-Ramsden, Gina, Nicola Botting & Brian Faragher (2001) Psycholinguistic markers for Specific Language Impairment (SLI). *Journal of Child Psychology and Psychiatry*, 42 (6), pp. 741–748. https://doi.org/10.1111/1469-7610.00770

Correia, Susana (2009). *The acquisition of primary word stress in European Portuguese*. Dissertação de doutoramento, Universidade de Lisboa.
Costa, Teresa (2010). *The acquisition of the consonantal system in European Portuguese: Focus on place and manner features.* Dissertação de doutoramento, Universidade de Lisboa.
Coutinho, Diana (2014) *Processamento fonológico de pseudopalavras linguisticamente motivadas em crianças com dislexia*. Dissertação de mestrado, Universidade do Algarve.
Di Cristo, Albert (1999) Le cadre accentuel du français contemporain : Essai de modélisation. *Langues*, 3 (2), pp. 184–205.
Cruz-Santos, Anabela (2009) *Cognitive-linguistic processing markers for the identification of European Portuguese speaking school-age children with specific language impairment*. Dissertação de doutoramento, Universidade do Minho.
Dickes, Paul, Jocelyne Tournois, André Flieller, Jean-Luc Kop (1994) *La psychométrie*. Éditions PUF.
dos Santos, Christophe (2007). *Développement phonologique en français langue première : Une étude de cas*. Dissertação de doutoramento, Université Lumière Lyon 2.
dos Santos, Christophe, Sabine Frau, Sandrine Labrevoit & Racha Zebib (2020) L'épreuve de répétition de non-mots LITMUS-NWR-FR évalue-t-elle la phonologie? *SHS Web of Conferences*, 78. https://doi.org/10.1051/shsconf/20207810005
dos Santos, Christophe, Sabine Frau, Sandrine Labrevoit, Prisca Martin, Philippe Prévost & Racha Zebib (2021, 17–18 junho) *Phonological assessment and non-word repetition test reduction* [Apresentação de poster]. International Child Phonology Conference (ICPC), Lethbridge, Canadá.
Ferré, Sandrine, Laurice Tuller, Eva Sizaret & Marie-Anne Barthez (2012) Acquiring and avoiding phonological complexity in SLI vs. typical development of French: The case of consonant clusters. In Philip Hoole, Lasse Bombien, Marianne Pouplier, Christine Mooshammer & Barbara Kühnert (orgs.), *Consonant clusters and structural complexity.* De Gruyter, pp. 285–308.
Ferré, Sandrine & Christophe dos Santos (2015) Un test de répétition de non-mots pour évaluer la phonologie des enfants bilingues. *LiDiL* 51, pp. 11–34. https://doi.org/10.4000/lidil.3678
Fikkert, Paula (1994) *On the acquisition of prosodic structure*. HIL
Freitas, Maria João (1997) *Aquisição da estrutura silábica do Português Europeu*. Dissertação de doutoramento, Universidade de Lisboa.
Freitas, Maria João & Celeste Rodrigues (2003) On the nature of sC-clusters in European Portuguese. *Journal of Portuguese Linguistics,* 2 (2), pp. 55–86. https://doi.org/10.5334/jpl.28
Freitas, Maria João (2004) The vowel [ɨ] in the acquisition of European Portuguese. In *Proceedings of GALA 2003.* LOT, pp. 163–174.
Freitas, Maria João (2007) On the effect of (morpho)phonological complexity in the early acquisition of unstressed vowels in European Portuguese. In Pilar Prieto, Joan Mascaró & Maria-Josep Solé (orgs.), *Segmental and prosodic issues in Romance phonology*. John Benjamins Publishing, pp. 179–198.
Grimm, Angela (2022) The use of the LITMUS quasi-universal nonword repetition task to identify DLD in monolingual and early second language learners aged 8 to 10. *Languages* 7 (3), pp. 218. https://doi.org/10.3390/languages7030218
Guimarães, Isabel, Carina Birrento, Catarina Figueiredo & Cristiana Flores (2014) *Teste de Articulação Verbal - TAV*. Oficina Didáctica.
Maddieson, Ian, Sébastien Flavier, Egidio Marsico & François Pellegrino (2013) LAPSyD: Lyon-Albuquerque Phonological Systems Databases. In *Proceedings of INTERSPEECH 2013*, pp. 3022–3026. https://doi.org/10.21437/Interspeech.2013-660
Mateus, Maria Helena & Ernesto d'Andrade (2000) *The phonology of Portuguese*. Oxford University Press.
Mendes, Ana, Elisabete Afonso, Marisa Lousada & Fátima Andrade (2013). *Teste fonético fonológico – Avaliação da Linguagem Pré-Escolar - ALPE*. Designeed, Lda.

Polo, Nuria (2018) Acquisition of codas in Spanish as a first language: The role of accuracy, markedness and frequency. *First Language* 38 (1), pp. 3–25. https://doi.org/10.1177/0142723717724244
Ramalho, Ana Margarida (2017) *Aquisição fonológica na criança: Tradução e adaptação de um instrumento de avaliação interlinguístico para o português europeu.* Tese de doutoramento, Universidade de Évora.
Ramalho, Ana Margarida & Maria João Freitas (2018) Word-initial rhotic clusters in typically developing children: European Portuguese. *Clinical Linguistics and Phonetics* 32 (5–6), pp. 459–480. https://doi.org/10.1080/02699206.2017.1359857
Ramus, Frank, Chloe R. Marshall, Stuart Rosen & Heather K. J. van der Lely (2013) Phonological deficits in specific language impairment and developmental dyslexia: Towards a multidimensional model. *Brain* 136 (2), pp. 630–645. https://doi.org/10.1093/brain/aws356
Ribeiro, Vânia (2011) *Instrumento de avaliação de repetição de pseudopalavras.* Dissertação de mestrado, Instituto Politécnico de Setúbal e Universidade Nova de Lisboa.
Rodrigues, Celeste (2003) *Lisboa e Braga: Fonologia e variação*. FCG/FCT.
Rose, Yvan (2000) *Headedness and prosodic licensing in the L1 acquisition of phonology*. Dissertação de doutoramento, McGill University.
Schwob, Salomé, Laurane Eddé, Laure Jacquin, Mégane Leboulanger, Margot Picard, Patricia Ramos Oliveira & Katrin Skoruppa (2021) Using nonword repetition to identify developmental language disorder in monolingual and bilingual children: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research* 64 (9), pp. 3578–3593. https://doi.org/10.1044/2021_jslhr-20-00552
Veloso, João (2006) Reavaliando o estatuto silábico das seqüências obstruinte+ lateral em português europeu. *DELTA* 22, pp. 127–158. https://doi.org/10.1590/S0102-44502006000100005
Vigário, Marina, Maria João Freitas & Sónia Frota (2006) Grammar and frequency effects in the acquisition of prosodic words in European Portuguese. *Language and Speech* 49 (2), pp. 175–203. https://doi.org/10.1177/00238309060490020301
Tuller, Laurie, Cornelia Hamann, Solveig Chilla, Sandrine Ferré, Eléonore Morin, Philippe Prevost, Christophe dos Santos, Lina Abed Ibrahim & Racha Zebib (2018) Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders* 53 (4), pp. 888–904. https://doi.org/10.1111/1460-6984.12397
Vigário, Marina, Sónia Frota & Fernando Martins (2006) A ferramenta FreP e a frequência de tipos silábicos e classes de segmentos no Português. In *Textos Seleccionados do XXI Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 675–687.
Walsh Dickey, Laura (1997) *The phonology of liquids*. Dissertação de doutoramento, University of Massachusetts.
Wiese, Richard (2001) The phonology of /r/. In Tracy Alan Hall (org.) *Distinctive feature theory*. De Gruyter, pp. 335–368.
Williams, David, Heather Payne & Chloe Marshall (2013) Non-word repetition impairment in autism and specific language impairment: Evidence for distinct underlying cognitive causes. *Journal of Autism and Developmental Disorders* 43 (2), pp. 404–417. https://doi.org/10.1007/s10803-012-1579-8

Extraction of target structures in learners' corpora: CQL queries for the exploitation of COPLE2

Raquel Amaro[1,2], *Alexandre Carreira*[1,2], *Alice Vieira*[1,2], *Cláudia Castro*[1,2], *Esmeralda Leong*[2]

[1]Centro de Linguística da Universidade Nova de Lisboa
[2]Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa

Abstract

Foreign language (FL) or second language (L2) corpora are sets of productions by non-native speakers, learners of a given language, which contemplate the errors and well-formed structures produced. These serve different research objectives, such as studies on language acquisition (LE and L2), phenomena of linguistic interference or analysis and diagnosis of LE/L2 proficiency levels. In the context of this research, the definition of the learner's proficiency level is often relevant, and this is done, typically, through the analysis of the presence or absence of errors in the learners' productions, based on mappings of typical or expected errors and well-formed structures for a given level of proficiency. However, contrary to the learner's error – which is explicitly marked in the corpus and whose typology and methodology of analysis constitutes a subtopic of investigation on its own –, the well-formed structures, and in particular the target structures (well-formed structures expected in the learners' productions of a given level of proficiency), are not easily identifiable in the corpora. The work presented here aims to fill this gap in COPLE2 – *Corpus of Portuguese Foreign/Second Language* through the use of expressions in CQL – *Corpus Query Language*. Based on pre-identified target structures and on the information made available in COPLE2, such as morphosyntactic tagging and different levels of information and annotation (learner production, teacher correction, normalized form, lemma, etc.), we propose query expressions in CQL that easily allow any user to immediately extract examples of target structures by proficiency level. The construction of the query expressions implies the definition and testing of the best strategies for each case and requires the systematization of linguistic rules and patterns of occurrence of the phenomena in question, but also the definition of ways to circumvent the limitations inherent to the corpus annotation, on the one hand, and the query language, on the other.

Keywords: Corpus Query Language, Learner's Corpus, target structures, PFL, L2.

Resumo

Os *corpora* de língua estrangeira (LE) ou língua segunda (L2) são conjuntos de produções de falantes não nativos, aprendentes de uma dada língua, que naturalmente incluem os erros e os acertos produzidos. Estes *corpora* servem diferentes objetivos de investigação, tais como estudos sobre aquisição de LE e L2, fenómenos de interferência linguística ou análise e diagnóstico de níveis de proficiência de LE/L2. No contexto da investigação destes tópicos, a definição do nível de proficiência do aprendente é muitas vezes relevante e esta é feita, tipicamente, através da análise da presença ou ausência de erros nas produções dos aprendentes, tendo como base mapeamentos entre erros e acertos típicos ou expectáveis para um dado nível de proficiência. No entanto, contrariamente ao erro do aprendente – que é explicitamente marcado no *corpus* e cuja tipologia e metodologia de análise constitui por si um subtópico de investigação –, os acertos, e em particular as estruturas-alvo (estruturas bem formadas expectáveis em produções de aprendentes de um dado nível de proficiência), não são facilmente identificáveis nos *corpora*. O trabalho que aqui se apresenta visa, assim, colmatar essa lacuna no COPLE2 – *Corpus de Português Língua Estrangeira*/*Segunda* através da utilização de expressões de pesquisa em CQL – *Corpus Query Language*. Tendo por base estruturas-alvo pré-identificadas e a informação

disponibilizada no COPLE2, tal como a etiquetagem morfossintática e os diferentes níveis de informação e anotação (produção do aprendente, correção do professor, forma normalizada, lema, etc.), são propostas expressões de pesquisa em CQL que permitem facilmente a qualquer utilizador a extração imediata de exemplos de estruturas-alvo por nível. A construção das expressões de pesquisa implica a definição e a testagem das melhores estratégicas e exige a sistematização de regras linguísticas e de padrões de ocorrência dos fenómenos em causa, mas também a definição de formas de contornar as limitações inerentes à anotação do *corpus*, por um lado, e à linguagem de pesquisa, por outro.

Palavras-chave: *Corpus Query Language*, *Corpus* de aprendentes, estruturas-alvo, PLE, L2.

1. Introduction

Foreign language (FL) or second language (L2) corpora are well-known resources, used for many different research purposes, such as studies in language acquisition, linguistic interference and analysis, diagnosis of FL/L2 proficiency or proficiency profiling (Tracy-Ventura & Paquot, 2020). Depending on the purposes and investment in their compilation and curation, these resources usually contemplate the annotation of the errors produced by the learners. Whatever is not annotated/corrected/identified is a well-formed structure. However, studying FL/L2 phenomena can require as much information on the ill-formed structures as well as on the well-formed structures, at several levels.

For instance, determining the learner's proficiency level requires analysing the presence (or absence) of errors in the learners' productions, which is performed based on the mapping of typical or expected errors, on the one hand, but also of typical or expected well-formed structures for a given level of proficiency (Gramacho et al., 2019; Talhadas, 2016). This means that a B1 level learner is expected not to make common spelling errors, for instance, but also that he/she is expected to produce complex sentences using subordination (and not only simple coordinated ones). However, contrary to the learner's error, which is typically explicitly marked in these corpora and whose analysis and typology constitutes a research topic on its own (Castello et al. 2016), the well-formed structures, i.e., whatever is consistent with the grammar rules of the FL/L2 is not marked. But this does not necessarily mean that all that is not marked, and therefore, well-formed, is of the same nature or relevance. In particular, to perform many of the analysis mentioned earlier, it is necessary to identify target structures, which consist of well-formed structures expected in the learners' productions of a given level of proficiency. However, these productions are not marked, making it quite difficult to extract them from the corpora.

Based on the target structures identified in the POR Nível project (Gramacho et al., 2019) and making use of the information available in COPLE2 – (del Río & Mendes, 2018; Mendes et al., 2016), we present query expressions that allow any user to easily and immediately extract examples of target structures by level. The construction of the query expressions in *Corpus Query Language* (CQL) (Cambridge, 2012; Christ, 1994) implies the definition and testing of strategies and requires the systematization of linguistic rules and patterns of occurrence of the phenomena in question, but also the definition of ways to circumvent the limitations inherent to the corpus annotation, on the one hand, and to the query language, on the other.

In the next sections, we present the process of building the CQL expressions, discussing the methodology used, the list of target structures considered, and the levels and type of information annotated and encoded in COPLE2, as well as the different CQL expressions defined considering the units of analysis or phenomena to be tackled, namely if we were operating at sequences of characters or at word form level, if we were considering multiword expressions or if we were dealing with longer distance phenomena such as subject-verb agreement. We present an analysis and evaluation of the results obtained, focusing on the invalid results extracted, to allow for a more accurate use of the extraction expressions, but also to inform the design of future query expressions. The usefulness and impact of the work developed, as well as possible implementations of the work here

presented, are briefly commented on the final remarks section. All query expressions provided in the paper are ready to be immediately and freely used.

2. Definition of the CQL expressions

The definition of adequate CQL expressions, with good performance, is the main part of the work put forth in this paper. In order to present the reflection and the specific decisions taken in process, we divided this section into several parts: the first consists of a brief presentation of the methodology followed; the second presents the target structures aimed at; the third section describes the information available in COPLE2, which can be used in the extraction queries; the fourth presents and discusses the CQL expressions created; and the fifth and final section evaluates and analyses the results achieved, relating them to the data and resources in use.

2.1. Methodology

To build the CQL expressions for extracting the relevant data (i.e., a specific set of target structures), we devised the following methodology:

1. Identification of the phenomena at stake: selection and description of the relevant target structures, according to existent literature on the subject for European Portuguese.
As further presented in Section 2.2, we collected the target structures from the ones described in the POR Nível project (Gramacho et al., 2019), for the proficiency levels A1 to C1.
However, more than collecting lists of phenomena, it is necessary to investigate and describe how these phenomena reflect in the language data. For instance, target structures related to the use of specific verb tenses and moods can either consist in querying specific part-of-speech tags (e.g., Simple Conditional mood/tense) or specific verb expressions (e.g., Future tense using auxiliary verb *ir* + Infinitive).

2. Mapping of the target structures with linguistic rules and/or occurrence patterns. After establishing the set of target structures to be extracted, we searched for specific examples in the corpus, based on linguistic rules (e.g., the internal structure of the noun phrase in European Portuguese) and or frequent/potentially occurring expressions (e.g., subordinative conjunctions; frequent modal verbs, etc.).

We performed diverse simple queries for each target structure to validate and extend, whenever relevant, the patterns based on linguistic knowledge. This allowed us to identify relevant tag sets and other information included in the retrieved examples (e.g., personal pronouns, common nouns, simple and complex proper nouns), as well as to account for less predictable cases, such as adverbs modifying adjectives in post-adjective position (e.g., "*Pessoas menos fortes* <u>*mentalmente*</u>" ≅ 'People less strong <u>mentally</u>').

3. Building the CQL query expressions. After the mapping phase, building the actual CQL query expression required several other steps to accommodate the limitations of the data available (e.g., COPLE2 does not have syntactic analysis), on the one hand, but also to determine how to use the several layers of annotation available (e.g., normalized form, level of proficiency, tokenization, lemma, part-of-speech tagging).

The main tasks in this phase concerned transforming structural information in linear information (e.g., which elements, in which relative order, and with which level of optionality, compose a Noun Phrase in European Portuguese) and how to use and combine several levels of annotation (document vs. text annotation) in a single query. The queries were built from the simpler to the more complex expressions, or from the more general expressions to the more specific, in an iterative process.

4. Testing and tunning of the CQL expressions. The CQL expressions were tested in the COPLE2 corpus to fine-tune them. This means that queries that over generate (i.e., queries that produced results that were not consistent with the target structures, along with results that were) were complemented with more specific tags or even specific lexical items, to gather gains in precision; queries that under generate (i.e., queries that produced results that were consistent with the target structures but missed many others) were redesigned to be more comprehensive.

5. Evaluation. This final step consists of the overall evaluation of the results achieved with the CQL expressions. This meant using the expressions and analysing, manually, the extracted results. More extensive results were evaluated through the analysis of random samples. The main purpose of this step is to evaluate the productivity of the queries, on the one hand, but also to understand where the extraction errors occurred and why, thus informing the use of the expressions.

This methodology was followed with good results. The following sections present and discuss relevant parts of the steps taken, explaining in detail the options taken.

2.2. **Target structures**

The target structures we considered in the work here presented were the ones considered in the project POR Nível (Gramacho et al., 2019), listed in Amaro et al. (2020). For the purposes of profiling proficiency levels for PLE, the POR Nível project considered the identification of uses divergent and convergent with target structures, having as bases the curricular content of guidance documents, such as PLE Referential Camões (*Referencial Camões PLE*), the syllabuses of the Instituto de Cultura e Língua Portuguesa courses and the syllabuses of the Portuguese Language and Culture Course of the Faculdade de Ciências Sociais e Humanas da NOVA University Lisbon.

The target structures described covered the orthography/spelling domain, morphology and syntax domain and vocabulary. According to the authors, the target structures were identified in specifically compiled corpora for each of the domains. After the identification, the target structures were categorized (Gramacho et al., 2019, p. 174).

Table 1 details the target structures we accounted for, considering the domain (or level of analysis), the phenomenon at stake, the specific target structure and proficiency level it relates to and an example for each structure.

Table 1. Target structures to be extracted, adapted from Amaro et al. (2020, p. 13)

Domain	Phenomenon	Target structure / Proficiency level	Example
Phonology /Spelling	Grapheme-phoneme correspondence	<ç> with [s] value / A2	*peço* ('I ask')
		<x> with [z] value / B2	*e**x**ijo* ('I demand')
	Stress marking	C1	*vigil**â**ncia* ('survaillance'), *respons**á**vel* ('responsible')
	Nasality before consonant	<m> before <p> or <b>; <n> in the rest of the cases / A2	*co**m**plicado* ('complicated'), *e**n**dereço* ('address')
Morphology & Syntax	Verb Tense, Mood and Aspect	Future with Auxiliary *ir* / A2	***vou morar*** ('I **will live**')
		Simple conditional / C1	***gostaria** de exprimir* ('I **would** like to express')
		Subjunctive *Pretérito Perfeito Composto* / C1	***tenha feito** muitos progressos* ('**had made** many progresses')
	Subject-Verb Agreement	B2	***alguém conhece*** ('**someone knows**')
	Conjunctions and Connectors	*mas* / A1	*Sou suíça **mas** moro em Lisboa há muitos anos* ('I'm Swiss **but** I live in Lisbon since many years now')
		por isso / A2	***por isso** tenho de ir* ('**therefore** I have to go')
		para que / B1	***para que** seja* ('**so** it be')
Lexicon	Specific vocabulary	Academic life / A1	***disciplina** favorita* ('favourite **subject**')
		Urban space / B2	***arredores** das cidades* ('city's **outskirts**')

The thirteen target structures were retrieved from the table presented in Amaro et al. (2020, p. 13), except for i) argument-marker prepositions marking argument of predicative nouns, ii) false friends and iii) idiomatic/frozen expressions. The last two cases were not retrieved because these were not listed in the table. However, if defined as pre-established lists of words or expressions, these can also be easily extracted, as we will demonstrate in the next sections.

2.3. **Information annotated in COPLE2**

The COPLE2 – *Learner Corpus of Portuguese L2* (del Río & Mendes, 2018; Mendes et al., 2016), a resource developed by CLUL – Centro de Linguística da Universidade de Lisboa, is composed of written and oral texts produced by foreign students who are learning Portuguese as a foreign language (PFL) or second language (L2). It also contains language proficiency certification exams. COPLE2 is freely available online (in http://teitok.clul.ul.pt/cople2/) and presents detailed information at the metadata level, concerning the informant and the text produced, as well as various types of linguistic annotation (del Río & Mendes, 2018), as schematized in the Figures 1 and 2 below.

Figure 1. Information in COPLE2 at document level

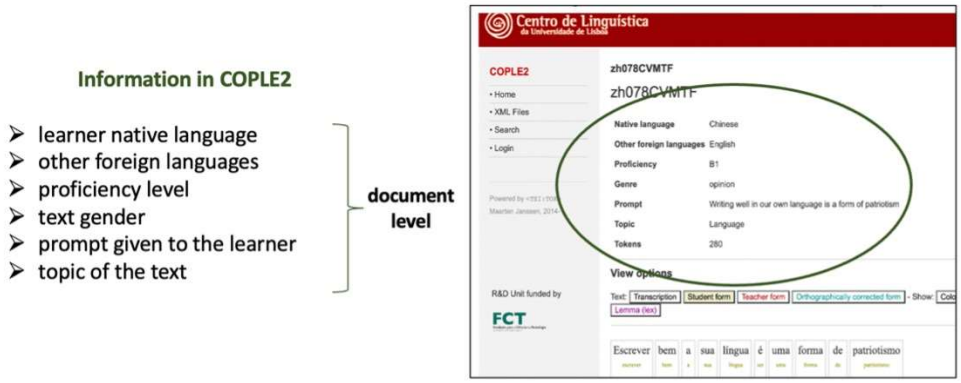
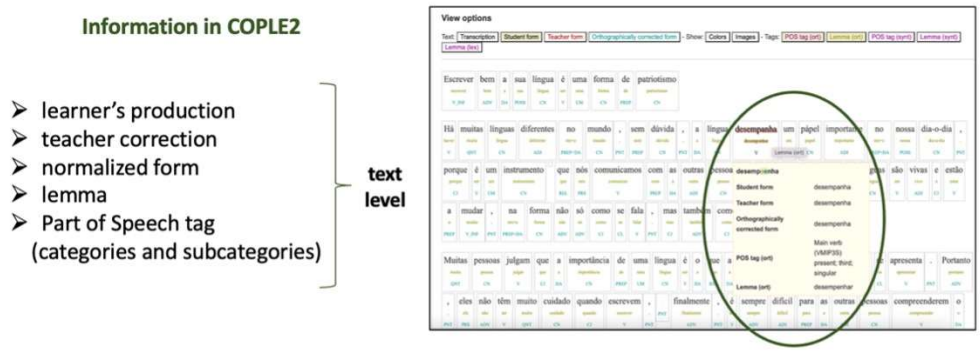


Figure 2. Information in COPLE2 at text level

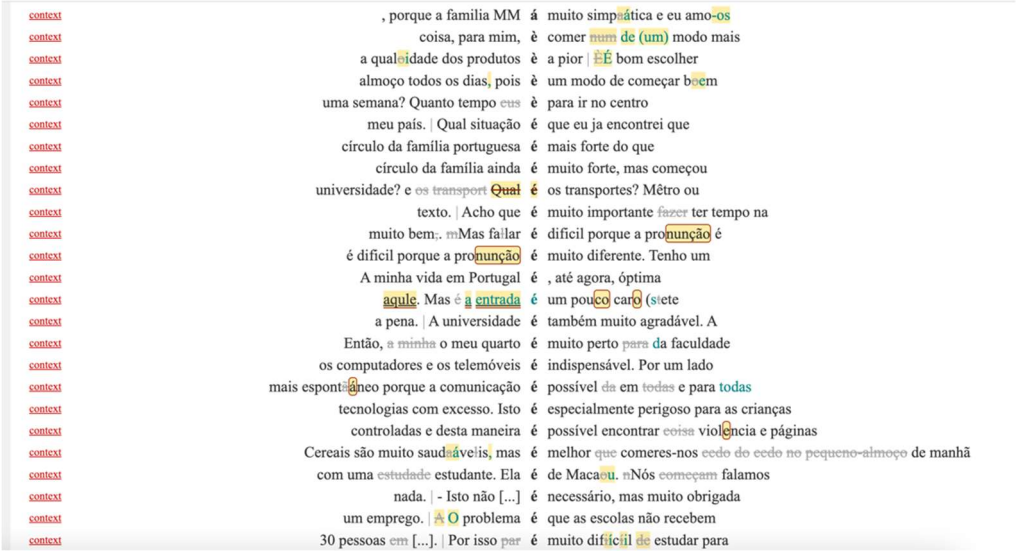


The part-of-speech and lemma information are automatically tagged. The rest of the information is derived from the source texts' metadata (e.g., proficiency level, prompt, topic, learner's native language), retrieved from the source text (teacher's correction), or introduced by the COPLE2 annotators (e.g., normalized form: orthographic normalization, morphosyntactic normalization and lexical normalization).

This rich annotation is associated with a CQL query system, making it possible to combine the different types of features and variables available. The corpus provides data necessary to many research topics, such as such as identifying general errors in PFL/L2 learning or identifying specific errors that may result from transfers of native languages or of other previously acquired foreign languages, enabling the development of applications and didactic materials in the area of PFL/L2 learning and teaching.

The extraction of information requires formulating CQL expressions that can be more or less straightforward depending on the phenomena to be extracted. If these correspond to phenomena specifically annotated (and tagged) in the corpus, it is a matter of knowing and using the specific tags. For instance, a CQL query such as `[form!=nform]` extracts orthographic errors, `[form!=reg]` extracts morphosyntactic errors. Also, the several layers of annotation can be visible in the COPLE2 visualizer (see Figure 3 below).

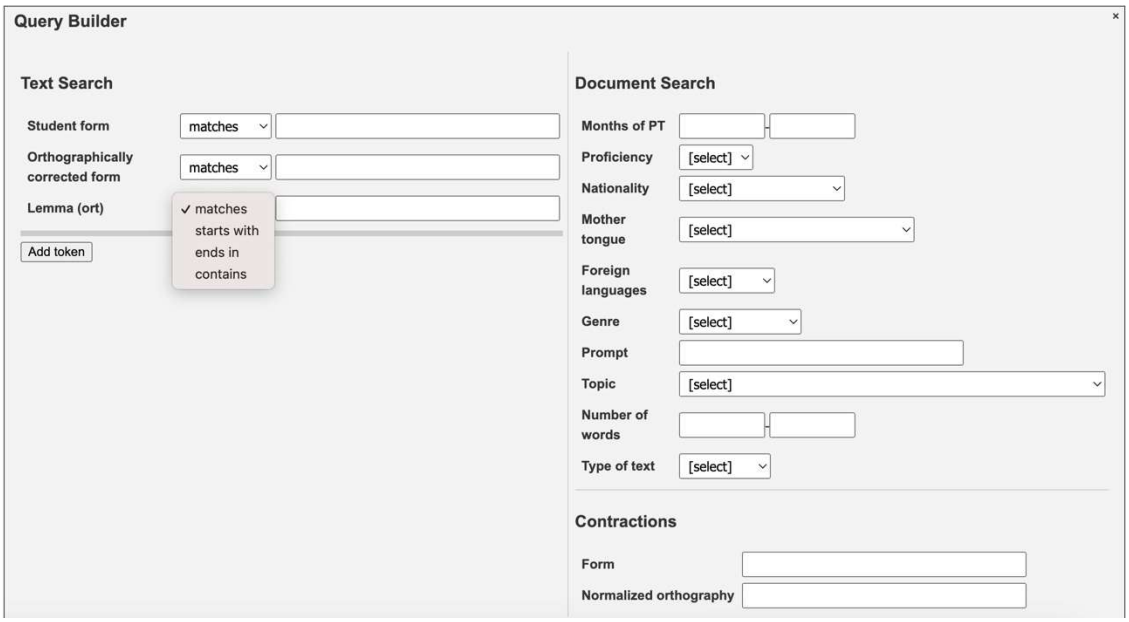
Figure 3. COPLE2 with visualization of error annotation, September 2023



Besides the query command line, the COPLE2 platform provides a Query builder.[1] As shown in the Figure 4 below, this tool helps the user build CQL expressions for one or more token (option 'Add token'), considering three options in the text - 'Student form', 'Orthographically corrected form' and 'lemma' -, and using four string operators - 'matches', 'starts with', 'ends with' and 'contains'. It also allows for combining text and document search parameters (left and right columns of the builder, respectively). This is, in fact, a very useful tool that helped us to know the CQL code for searching parameters at the document level (for instance, 'match.text_mothertongue' is the attribute to search for documents from learners with a specific mother tongue). However, the Query Builder does not replace CQL proficiency, and querying for more complex phenomena requires further knowledge and investment of time.

[1] COPLE2 webpage provides a help page for CQL query builder (http://teitok.clul.ul.pt/cople2/index.php?action=querybuilderhelp) and well as access to further information on CQL in https://cwb.sourceforge.io/documentation.php

Figure 4. COPLE2 Query Builder



The rich annotation system and information used in COPLE2 is a key factor to enable the successful extraction of these structures, as explained in the following sections.

2.4. **CQL expressions**

Considering the target structures to be extracted, and to map the phenomena to specific sets and types of CQL queries, we searched for specific examples in the corpus, based on linguistic rules. This allowed us to identify the relevant tags, possible sets of linear strings of part-of-speech tags, and relevant tags at the document and text levels. From the analysis devised at this stage, it was possible to group the queries by the units these consider - units smaller than words, word forms, multiword expressions, and phrases -, as these require different strategies.

2.4.1. **Sets of characters**

The definition of character sets allowed us to account for the listed phonology/spelling target structures, as these concern the correct use of specific characters in specific sequences, as demonstrated in the Table 2, below.

Table 2. CQL expressions for phonology/spelling target structures

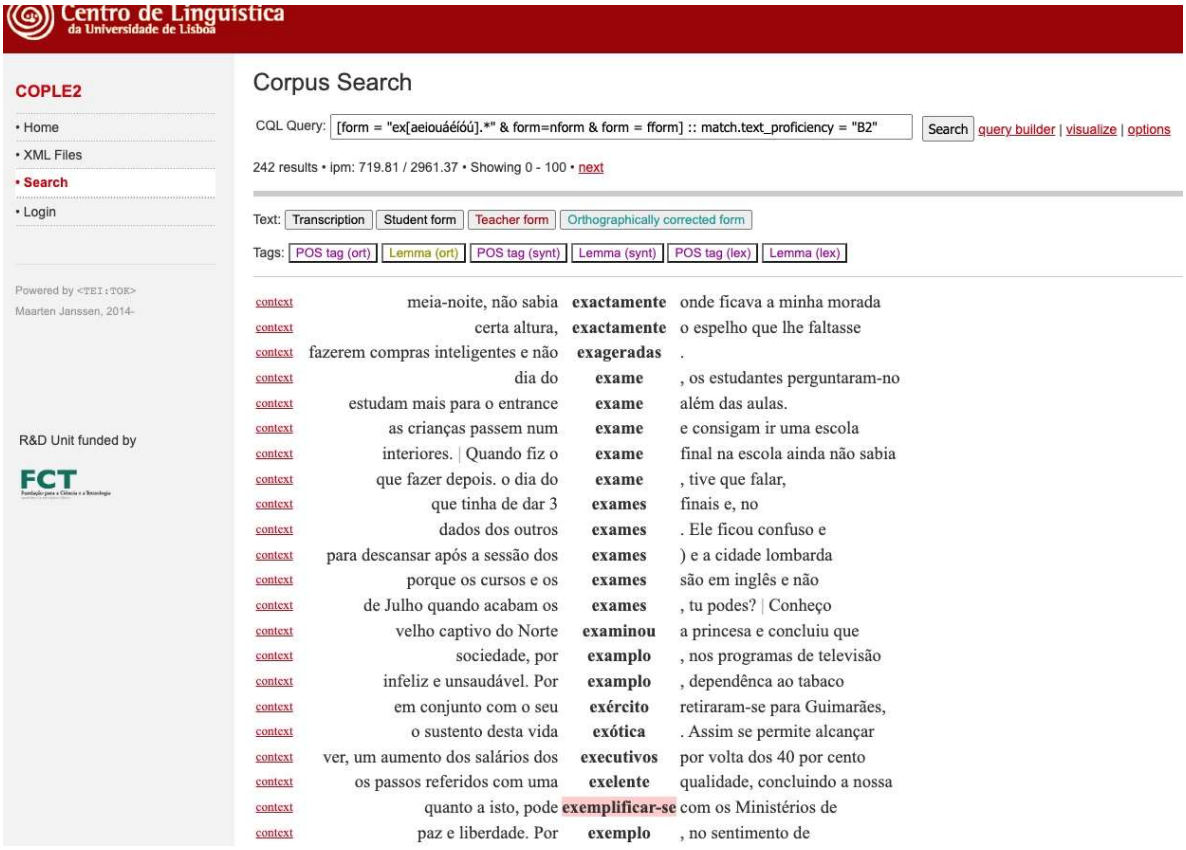
Nr	Target structure / Proficiency level	CQL expression	Results
1	<ç> with [s] value / A2	`[form = ".*ç.*" & form=nform & form = fform] :: match.text_proficiency = "A2"`	594
2	<x> with [z] value / B2	`[form = "ex[aeiouáéíóú].*" & form=nform & form = fform] :: match.text_proficiency = "B2"`	242
3	stress marking / C1	`[form = ".*(À|Á|É|Í|Ó|Ú|Ã|Õ|Â|Ê|Ô|à|á|é|í|ó|ú|â|ê|ô|ã|õ).*" & form = nform & form = fform] :: match.text_proficiency = "C1"`	5313
4	<m> before <p> or <b>; <n> in the rest of the cases / A2	`([form=".*m(b|p).*" & form=nform & nform = fform] | [form=".*n(b|c|d|f|g|h|j|k|l|m|n|p|q|r|s|t|v|x|z|ç).*" & form=nform & form = fform]) :: match.text_proficiency = "A2"`	5763

For instance, to extract the use of the grapheme <ç> (CQL expression 1), we used the 'form' attribute to search for a specific character in a given word (`form = ".*ç.*"`). To assure that the grapheme was correctly used, we checked if the word form used in the learner's production was coincident with the values in the 'nform' and 'fform' attributes (`& form=nform & form = fform`), that is, cases where there was no normalization of the form (`nform`) and/or no correction of the teacher (`forma`). Since this case is expected for A2 proficiency level students, we restricted the query to the document level value "A2" of the attribute 'text_proficiency'. The same strategies were used in the following cases, with the addition of sets of characters that correspond to letters plus diacritics used in European Portuguese (e.g., `áéíóú`) or the sets of letters corresponding the consonants after <n> or <m>.

Also, CQL uses several notations from regular expressions: operators, such as AND (`&`) and OR (`|`), wildcards to represent variables, such as any character (`.`), quantifiers, such as 0 or more times (`*`), or notations such as one of the characters of the set (`[ ]`) and substrings (`( )`). These can operate at different levels, namely character strings (e.g., `.*m(b|p).*`) or attribute level (e.g., `[form = ".*ç.*" & form=nform & form = fform]`). As described above, COPLE2 also allows the combination of annotations at text and document levels (`[form = ".*ç.*" & form=nform & form = fform] :: match.text_proficiency = "A2"`).

The CQL expressions presented in the table above can be directly used (copied and pasted into the query box) in the search system of COPLE2. Figure 5 below shows the results obtained with the CQL expression 2, <x> with [z] value in B2 proficiency level, at current date.

Figure 5. Query results for the CQL expression nr. 2 - <x> with [z] value / B2, September 2023



2.4.2. **Word lists**

The extraction of several of the target structures required CQL expressions considering specific word lists. The definition of these word lists, in turn, could include simple (or atomic) words and multiword expressions, as well as the delineation of the semantic fields to be consider.

As depicted in the Table 3, CQL expressions considering simple words were used to extract target structures related to the domains of morphology and syntax and of lexicon.

Table 3. CQL expressions for target structures with simple word lists, results September 2023

Nr.	Target structure / Proficiency level	CQL expression	Results
5	Simple Conditional / C1	`[pos="VMIC.*" & form=nform & form=fform & form=reg & form=lex] :: match.text_proficiency = "C1"`	55
6	Atomic connectors / A1 and B1	`[form="mas" & form=nform & form=fform & form=lex & pos="C.*"] :: match.text_proficiency = "A1"`	115
7		`[form="mal|embora" & form=nform & form=fform & form=lex & pos="C.*"] [pos!="V.*" & form=nform & form=fform & form=lex]{0,6} [pos="VMS.*" & form=nform & form=fform & form=lex & form=reg]:: match.text_proficiency = "B1"`	6
8	Academic life vocabulary / A1	`[lemma = "estudar|estudante|aluno|escolar|universidade|aula|turma|cantina|professor|tpc|exercício|colega|exame|curso|teste|estojo|caderno|lápis|biblioteca|currículo|faculdade" & form=nform & form=fform & form=lex]:: match.text_proficiency = "A1"`	485
9	Urban space vocabulary / B2	`[lemma = "edifício|habitante|estrada|cidade|passear|trânsito|zona|rio|margem|bairro|loja|redor|centro|aldeia|rua|apartamento|prédio|local|miradouro|supermercado" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	299

The definition of these query expressions was based on the compilation of word lists from different sources: atomic connectors were collected from the ones described in Gramacho et al. (2019); the vocabulary related to academic life for A1 level were collected from the Portuguese Foreign Language handbooks *Passaporte para Português (A1/A2)* and *Português XXI Nível A1*; the vocabulary related to urban space for B2 level were collected from the Portuguese Foreign Language handbook *Aprender Português (Nível B2)*.

The specific lists of words collected can change, whenever relevant. For instance, if, in a given class, the vocabulary taught at level A1 focuses on daily life and holidays instead of academic life, the set of lemmas used in the CQL expression above can be replaced. What is relevant here is that is more practical and simpler to use one query expression for simple words and another for expressions.

The combination of features to be used in the queries can differ, depending on the phenomena we are extracting. For instance, to extract vocabulary (simple words), the relevant attribute for the query is 'lemma' (so we can extract all inflected forms of a given word, for instance, for the lemma 'habitante' (≅ inhabitant): *habitante* (≅ inhabitant)*, habitantes* (≅ inhabitants)*, habitantezinho* (≅ little inhabitant), ...). COPLE2 has different levels of normalization that can be used to further improve the queries and that are relevant for these cases: *nform*: orthographic normalization, *reg*: morphosyntactic normalization and *lex*: lexical normalization (cf. del Río & Mendes, 2018; Mendes et al., 2016). This way, to assure we are extracting correct uses we also checked if the values of the attributes 'form', 'nform', 'fform' and 'lex' corresponded. The 'reg' attribute was used to assure the correct use of inflected forms, typically in the case of verbs.

The feature relevant for extracting specific verb forms, such the Conditional Mode/Tense is part-of-speech category, verb, (`pos="`**`V`**`MIC.*"`) and subcategories, in the Conditional Mode/Tense (`pos="V`**`MIC`**`.*"`).

The relevant features to extract appropriate occurrences of atomic connectors, on the other hand, can involve the combination of a specific word form with a specific part-of-speech tag (as in `form="mas" &`

`pos="C.*"]`) or the combination of more features, such as a specific word form with a specific part-of-speech tag (`form="mal|embora" & pos="C.*"`) occurring close to, or at a distance of, 5 words of a main verb in the Subjunctive Mode (`{0,6} [pos="VMS.*"]`), with words that are not verbs in between (`[pos!="V.*"]{0,6}`). That is, combining features of the queried word forms with features from co-occurring forms, as illustrated in Figure 6, line 7, below. Also in these cases, to assure we extracted correct uses, the values of the attributes 'form', 'nform', 'fform', 'lex' and 'reg' were checked for all occurring elements.

Figure 6. Query results for the CQL expression nr. 7 - Atomic connectors / A1 and B1, September 2023

2.4.3. **Multiword expressions**

The CQL expressions considering multiword expressions (MWE), that is, a specific sequence of words, were used to account for target structures related to complex verb tenses (auxiliary verbs + main verbs), complex connectors and vocabulary related to academic life and urban space. As described for simple word lists, the different phenomena can require the different combination of features, and different levels of complexity, as showed in Table 4.

Table 4. CQL expressions for target structures with MWE (results in September 2023)

Nr.	Target structure / Proficiency level	CQL expression	Results
10	Future with Auxiliary *ir* / A2	`[lemma="ir" & pos="VMIP.*" & form=nform & form=fform & form=lex & form=reg] [pos="R.*" & form=lemma]? [pos="VMN" & form=lemma & form!="ir"] :: match.text_proficiency = "A2"`	191
11	Subjunctive *Pretérito Perfeito Composto* / C1	`[pos="VASP.*" & lemma="ter" & form=nform & form=fform & form=lex & form=reg] [pos="VMP" & form=nform & form=fform & form=lex & form=reg] :: match.text_proficiency = "C1"`	2
12	Complex connectors / B1	`[form="já|dado|visto|para|antes|depois|logo|sempre|até|desde|ainda|tanto"& form=nform & form=fform & form=lex] [form="que"& form=nform & form=fform & form=lex] [pos!="V.*" & form=nform & form=fform & form=lex]{0,6} [pos="V.*" & form=nform & form=fform & form=lex & form=reg] :: match.text_proficiency = "B1"`	33
13		`([form="todas" & form=nform & form=fform & form=lex] [form="as" & form=nform & form=fform & form=lex] [form="vezes" & form=nform & form=fform & form=lex] [form="que" & form=nform & form=fform & form=lex])|([form="apesar" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex]) [pos!="V.*" & form=nform & form=fform & form=lex]{0,6} [pos="V.*" & form=nform & form=fform & form=lex & form=reg] :: match.text_proficiency = "B1"`	5
14	Academic life vocabulary / A1	`[lemma="sala" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex] [lemma="aula" & form=nform & form=fform & form=lex] :: match.text_proficiency = "A1"`	0
15		`[lemma="trabalho" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex] [form="casa" & form=nform & form=fform & form=lex] :: match.text_proficiency = "A1"`	9
16	Urban space vocabulary/B2	`[form="a" & form=nform & form=fform & form=lex] [form="pé" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	3
17		`[lemma="junta" & form=nform & form=fform & form=lex] [form="de" & form=nform & form=fform & form=lex] [form="freguesia" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	0
18		`[lemma="câmara" & form=nform & form=fform & form=lex] [form="municipal" & form=nform & form=fform & form=lex] :: match.text_proficiency = "B2"`	0

The CQL expressions here presented range from simple word sequences (and not lemma) of 2, 3 or more elements such as *a pé* (≅ by foot), *sala de aula* (≅ classroom), *todas as vezes que* (≅ every time that), to sequences of specific part-of-speech category and subcategories and lemma, such as the Subjunctive tense that is composed of the auxiliary verb *ter* (≅ to have) in the Subjunctive Mode, Present Tense (`pos="VASP.*" & lemma="ter"] [pos="VMP"]`) immediately followed by a main verb in Past Participle form (`[pos="VMP"]`).

To account for some flexible MWE included in this set, such as Future with auxiliary verb *ir* (≅ to go), the CQL expression designed also accommodated optional material that can co-occur within the expression such as adverbs (`[lemma="ir" & pos="VMIP.*" & form=nform & form=fform & form=lex & form=reg] **[pos="R.*" & form=lemma]?** [pos="VMN" & form=lemma & form!="ir"] :: match.text_proficiency = "A2"`). The Figure 7, below, show some of the occurrences extracted with more complex examples.

Figure 7. Query results for the expression 13. Complex connectors / B1, September 2023

2.4.4. **Phrases: agreement**

The target structures requiring the levelling of more complex structures concerned subject-verb agreement. The challenge here is that the corpus, although presenting a fine-grained level of part-of-speech annotation, is not syntactically parsed. This way, to assess subject-verb agreement we need to describe, in a linear form, the sequence of word classes that can occur between the core noun of the subject noun phrase (NP) and the verb, and to encode this in CQL. To do so, it is necessary to describe what can follow the core noun within the noun phrase, which lead us to the need to describe the possible configurations of the noun phrase in European Portuguese. For instance, in the sentence (1), below, we have a prepositional phrase and an adverb between the core elements that must agree (in bold).

(1) O **brasão** de Évora ainda **mostra** esse ato heróico de o Geraldo Geraldes (COPLE2, en020CAATI_1)
' The **coat** of arms of Évora still **shows** this heroic act by Geraldo Geraldes '

We considered the following major configurations for the noun phrase in European Portuguese, represented here in simple context-free grammar notation, where NP stands for noun phrase, AP for adjective phrase, AdvP for adverbial phrase, DP for determiner phrase and PP for prepositional phrase. The round brackets indicate optionality:

(2) NP --> Pronoun
NP --> (DP) (AP) Noun (AP) (PP)
DP --> (Quantifier) Determiner (Possessive)
AP --> (AdvP) Adjective (PP)
PP --> Preposition NP
AdvP --> (Adv) Adv
Determiner --> Article
Determiner --> Indefinite
Determiner --> Demonstrative

The configurations in (2) accounts for NPs such as *ele* (≅ he), *Maria* (≅ Mary) or *o saudável incitamento à participação ativa de toda a sociedade civil na resolução dos seus mais urgentes problemas* (≅ the healthy encouragement of the active participation of all civil society in the solution of its most urgent problems).[2]

Besides the elements that are included in the noun phrase, we also accounted for elements that can occur before the verb, as schematized in (3) below:

(3) NP_Subject (Adverb) (Clitic) Verb

The next step is, thus, to transform these schemas in linear expressions in CQL. As demonstrated below, and since there are no structural elements that we can refer to, these can amount to large expressions.

(4) Adverb phrase structure in CQL
```
[pos="R.*" & form=nform & form=fform & form=lex]? [pos="R.*" & form=nform & form=fform & form=lex]
```
which is equivalent to `[pos="R.*" & form=nform & form=fform & form=lex]{1,2}`

However, given that in our case adverbs are always optional, the formulation to be used can be `[pos="R.*" & form=nform & form=fform & form=lex]{0,2}.`

(5) Determiner phrase structure in CQL
```
([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & !contr & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) |
```

[2] Example built based on an excerpt of the editorial note of the newspaper *Expresso* of 24/06/2021. Available at https://expresso.pt/expresso/nota-da-direcao/2021-06-24-Nota-editorial-do-Expresso-a-explicacao-para-um-cabecalho-nao-neutral-c18ec16f

```
([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform
& form=lex])
```

The determiner phrase includes quantifiers (`pos="BQ.*"`) and uses the negation of the attribute 'contr' to exclude determiners and quantifiers within contracted forms, such as *dos* (≅ of the) or *nalguns* (≅ in some). This expression allows for the following combinations:

(6) o/este/aquele *Noun* (≅ the/this/that *Noun*)
todo o/este/aquele *Noun* (≅ all the/this/that *Noun*)
o/este/aquele meu *Noun* (≅ the/this/that my *Noun*)
todo o/este/aquele meu *Noun* (≅ all the/this/that my *Noun*)
algum *Noun* (≅ some *Noun*)

(7) Adjective phrase structure in CQL

```
[pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" &
form=nform & form=fform & form=lex]
```

(8) Prepositional phrase structure within noun phrases in CQL

```
[pos="S.*" & form="d.*" & form=nform & form=fform & form=lex]
(([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform &
form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex]
[pos="BP.*" & form=nform & form=fform & form=lex]?) |
([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform
& form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2}
[pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" &
form=nform & form=fform & form=lex] ([pos="R.*" & form=nform &
form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform &
form=lex])?
```

Prepositional phrases here consider are the ones introduced by the preposition *de* (≅ of), in contracted and non-contracted forms, typically used within the noun phrase, taking as complement a noun phrase. That explains the removal of the condition `!contr` (not contracted) expressed within the determiner phrase. The expression for the prepositional phrase allows extracting cases such as:

(9a) ... é tão diferente **da arquitectura mexicana**... (COPLE2, es010CVMTD)
'... it's so different **from Mexican architecture**... '
(9b) ... embora seja uma **das minhas férias favoritas**, tenho... (COPLE, en022CAMTF_2
'...even though it's one **of my favourite vacations**, I have...'
(9c) ... uma ilha pequena que faz parte **desse grande país**... (COPLE2, de048CVATI)
'... a small island that is part **of that great country**'
(9d) ... só precisas **de algumas camisolas desportivas**, porque a equipa... (COPLE2, it093CSITF_2)
'... you just need **[of] some sports jerseys**, because the team...'

With these expressions we can, then, build the queries for extracting our target structures.

The basic structures of a noun phrase in European Portuguese, described in (2), can, finally, by captured by the following CQL expression, in (10).

(10) Noun phrase structure in CQL

```
([pos="P.*" & form=nform & form=fform & form=lex & form=reg & !contr])|( (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & !contr & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) |([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))?([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex] ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])?)?)
```

These expressions can be quite long, and noun phrases can have as core elements pronouns and nouns, with different features concerning agreement, namely in what concerns Person features. Also, different types of nouns, common and proper nouns, occur in different structures. For these reasons, we decided to divide the queries into four separate expressions accounting for:

i) subject-verb agreement with personal pronoun,

ii) subject-verb agreement with common noun,

iii) subject-verb agreement with proper noun, and

iv) subject-verb agreement with relative pronoun.

Table 5, below, presents the CQL expressions for each of them.

Table 5. CQL expressions for target structures concerning subject-verb agreement (results in September 2023)

Nr.	Target structure / Proficiency level	CQL expression	Results
19	Subject-verb agreement -personal pronoun /B2	([pos="P.*S1" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*1S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*S2" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*2S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*S3|PI.*S" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form =fform & form=lex]{0,2} [pos="V.*3S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*P1" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*1P" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*P2" & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*2P" lemma!="haver" & form=nform & form=fform & form=lex & form=reg]|[pos="P.*P3|PI.*P " & form=nform & form=fform & form=lex & form=reg & !contr] [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*3P" lemma!="haver" & form=nform & form=fform & form=lex & form=reg]) :: match.text_proficiency = "B2"	226
20	Subject-verb agreement common noun /B2	(([pos="N.S*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.S*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform &	1228

		form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex] ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])?)? [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*3S" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg]) | ([pos="N.P*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.P*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])? [pos="E|N.*" & form=nform & form=fform & form=lex] ([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.*" & form=nform & form=fform & form=lex])?)? [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos="V.*3P" & lemma!="haver" & form=nform & form=fform & form=lex & form=reg])):: match.text_proficiency = "B2"	
21	Subject-verb agreement proper noun /B2	([pos="TMS" & form=nform & form=fform & form=lex]? [pos="E" & form=nform & form=fform & form=lex]{1,3} ([form="de|do|da|dos|das" & form=nform & form=fform & form=lex] [pos="TMS" & form=nform & form=fform & form=lex]? [pos="E" & form=nform & form=fform & form=lex]{0,3})? [pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="K.*" & form=nform & form=fform & form=lex & form=reg]? [pos = "V.*3S" & lemma!="haver" & form=fform & form=lex & form=reg]) :: match.text_proficiency = "B2"	206
22	Subject-verb agreement relative pronoun /B2	(([pos="N.S*" & form=nform & form=fform & form=lex]([pos="R.*" & form=nform & form=fform & form=lex]{0,2} [pos="A.S*" & form=nform & form=fform & form=lex])? ([pos="S.*" & form="d.*" & form=nform & form=fform & form=lex] (([form="tod(o|a)s?" & pos="BQ.*" & form=nform & form=fform & form=lex]? [pos="DA.*|BD.*" & form=nform & form=fform & form=lex] [pos="BP.*" & form=nform & form=fform & form=lex]?) | ([form!="tod(o|a)s?" & pos="BQ.*" & !contr & form=nform & form=fform & form=lex]))? ([pos="R.*" & form=nform & form=fform &	255

```
form=lex]{0,2} [pos="A.*" & form=nform &
form=fform & form=lex])? [pos="E|N.*" &
form=nform & form=fform & form=lex] ([pos="R.*" &
form=nform & form=fform & form=lex]{0,2}
[pos="A.*" & form=nform & form=fform &
form=lex])?)? [form="que" & form=nform &
form=fform &form=lex] [pos="R.*" & form=nform &
form=fform & form=lex]{0,2} [pos="K.*" &
form=nform & form=fform & form=lex & form=reg]?
[pos="V.*3S" & lemma!="haver" & form=nform &
form=fform & form=lex & form=reg]))|
(([pos="N.P*" & form=nform & form=fform &
form=lex]([pos="R.*" & form=nform & form=fform &
form=lex]{0,2} [pos="A.P*" & form=nform &
form=fform & form=lex])? ([pos="S.*" & form="d.*"
& form=nform & form=fform & form=lex]
(([form="tod(o|a)s?" & pos="BQ.*" & form=nform &
form=fform & form=lex]? [pos="DA.*|BD.*" &
form=nform & form=fform & form=lex] [pos="BP.*" &
form=nform & form=fform & form=lex]?) |
([form!="tod(o|a)s?" & pos="BQ.*" & !contr &
form=nform & form=fform & form=lex]))?
([pos="R.*" & form=nform & form=fform &
form=lex]{0,2} [pos="A.*" & form=nform &
form=fform & form=lex])? [pos="E|N.*" &
form=nform & form=fform & form=lex] ([pos="R.*" &
form=nform & form=fform & form=lex]{0,2}
[pos="A.*" & form=nform & form=fform &
form=lex])?)? [form="que" & form=nform &
form=fform &form=lex] [pos="R.*" & form=nform &
form=fform & form=lex]{0,2} [pos="K.*" &
form=nform & form=fform & form=lex & form=reg]?
[pos="V.*3P" & lemma!="haver" & form=nform &
form=fform & form=lex & form=reg])):
match.text_proficiency = "B2"
```

The expressions presented here allow us to extract subject-verb agreement cases such as the ones presented (11), ranging from simple sequences of pronoun-verb to long and more complex expressions involving MWE proper nouns with treatment forms, complex noun phrases, and relative pronouns with some distance to the core noun of the noun phrase they refer to.

(11a) ... mas **ele continuou** e finalmente chegou ao destino... (COPLE2, zh059CAATI_1)
'... but **he continued** and finally reached the destination...'

(11b) Quando **alguém parece** que precisa alguma coisa... (COPLE2, ja023CAATF)
'When **someone seems** to need something...'

(11c) ... recomendo que o **Governo de Portugal procure** maneiras adicionais... (COPLE2, en091CVATF)
'... I recommend that the **Government of Portugal looks** for additional ways...'

(11d) ... mas a **superficialidade da nova comunicação rápida não facilita** conversas sobre temas importantes... (COPLE2, de007CVATD)
'... but the **superficiality of the new rapid communication does not facilitate** conversations on important topics...'

(11e) Acredito que muitas **caraterísticas dos portugueses hoje são** relacionadas com a História Portuguesa. (COPLE2, zh065CAATF)
'I believe that many **characteristics of the Portuguese today are** related to Portuguese History.'

(11f) Os **exemplos dessa injustiça encontram-se** por toda a parte, ... (COPLE2, ru013CAATF_3)
'The **examples of this injustice are** everywhere, ...'

(11f) ... são as pequenas **coisas do dia-a-dia que fazem** a diferença e o enriquecimento cultural... (COPLE2, de016CVATI)
'... it's the **little everyday things that make** a difference and cultural enrichment... '

(11g) A solidão, o stresse e a depressão são as **doenças das sociedades modernas que causarão** mais prejuízos, humanos e económicos, no seculo XXI. (COPLE2, zh007CVATF)
'Loneliness, stress and depression are the **diseases of modern societies that will cause** the most human and economic damage in the 21st century.'

(11h) ... temos o exemplo dos **governos de extrema esquerda que se transformaram** em ditaduras nos países do Leste da Europa. (COPLE2, it015CVATF)
'... we have the example of **far-left governments that turned** into dictatorships in Eastern European countries.'

Before discussing, in more detail, the analysis and evaluation of the results achieved, it is necessary to refer some practical decisions taken. The first one concerns the option to not consider the material that can occur before the core noun in the subject noun phrase. We decided not to consider this material since it would make the expressions lengthier and, with that, with more possibility of errors. Additionally, it is the core noun that determines the agreement features. Also, the internal structure of the proper noun phrase is different from the one determined for noun phrases in general. This option is related to the fact that many proper nouns are tagged with the E tag in the corpus, instead of other part-of-speech tags (e.g., *Banco/E Alimentar/E* and not *Banco/N Alimentar/A*). So, the CQL expression considers sequences of adjacent proper nouns, mediated or not by the preposition *de*, instead of searching for regular sequences expressing regular noun phrase structures. Finally, all the expressions, as presented in (3), consider the possibility of adverbs and clitics occurring between the subject and the verb. The examples in (12) illustrate these cases, highlighted in italics.

(12a) Por exemplo, ele *não* sabia que... (COPLE2, en014CAATD)
'For example, he did *not* know that...'

(12b) Eu *pessoalmente* vejo e acredito que... (COPLE2, zh018CVATF)
' I *personally* see and believe that...'

(12c) Contudo, é sempre importante para os consumidores *não se* deixarem... (COPLE2, nl010CVATI)
'However, it is always important for consumers *not to* [*themselves*] let...'

(12d) Se vocês *não me* pagarem até Abril, vou... (COPLE2, ja019CAATD)
'If you do *not* [*me*] pay by April, I'll...'

3. Analysis and evaluation of results

To evaluate the results, we conducted a manual revision of the cases extracted. Table 6 presents the evaluation of the results in terms of number of extracted occurrences, number of bad results, i.e., occurrences that do not correspond to the target structure queried, and the percentage of good results obtained.

Table 6. Evaluation of results (September 2023)

Target structure / Proficiency level		Query Nr.	Number of extracted occurrences	Bad results	Percentage of good results
<ç> with [s] value / A2		1	345	0	100%
<x> with [z] value / B2		2	145	0	100%
Stress marking / C1		3	2 940	0*	100%
Nasality / A2		4	3 596	0*	100%
Simple Conditional/C1		5	55	0	100%
Future with Auxiliary *ir* / A2		10	191	1	99,5%
Subjunctive *Pretérito Perfeito Composto* / C1		11	2	0	100%
Connectors	Atomic / A1	6	115	0	100%
	Atomic / B1	7	7	0	100%
	MWE / B1	12	33	1	97%
	MWE / B1	13	5	1	80%
Academic life vocabulary	Atomic / A1	8	485	0	100%
	MWE /A1	14	0	0	-
	MWE / A1	15	9	9	100%
Urban space vocabulary	Atomic / B2	9	299	12	96%
	MWE /B2	16	3	3	100%
	MWE / B2	17	0	0	-
	MWE / B2	18	0	0	-
Subject-verb agreement/B2	personal pronoun	19	226	0	100%
	common noun	20	1 228	6*	98%
	proper noun	21	206	25	91,9%
	relative pronoun	22	255	4	98,4%

*The evaluation was performed over a sample of 25% of cases randomly selected.

Our overall evaluation is that the CQL expressions built allow the extraction of very reliable results. In 22 CQL expressions, we only performed below 98% in four cases, of these only one below 91,9%. Besides assessing the precision of the results, we also analysed the bad results to understand possible reasons for their occurrence and how to address these, if possible.

Several reasons account for most of the bad results, as illustrated below.

i. Errors in the part-of-speech tag.

(13) ... **visto**_CommonNoun **que é** imenso facil para saber como foi passado... (COPLE2, zh022CVA1TF)
'... **since**_CommonNoun it is really easy to find out how it was spent...
(bad result from CQL expression 22: Subject-verb agreement/B2 relative pronoun)

ii. Missing annotation.

(14a) ... **aquelos**_missing normalized form **que mais tiverom**_missing normalized form suceso. (COPLE2, it101CSATF)
'... **thoso**_ missing normalized form that **more hod**_missing normalized form succes.

(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

(14b) ... **apesar de ninguém pode**_missing normalized form ser o "superman"... (COPLE2, zh044CAMTI)
'... **although no one can**[inflected form]_missing normalized form be "superman"...'
(bad result from CQL expression 13: Complex connectors/B1)

iii. Linear order coincidences related to the annotation of locutions.

(15a) De **maneira**_CommonNounSingular **que**_RelativePronoun **faça**_VerbSingular isto, temos dois passos... (COPLE2, zh013CVATD)
'In such a **manner**_CommonNounSingular **that**_RelativePronoun [it] **does**_VerbSingular this, we have two steps...'
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

iii. Linear order coincidences related to the recursive nature of noun phrase and prepositional phrase structures.

(15b) ... por isso cualquer problema_CoreNoun de uma **pessoa do povo é** considerado... (COPLE2, es036CVATF)
'... for that any problem_CoreNoun of a **person of the people is** considered...'
(bad result from CQL expression 20: Subject-verb agreement common noun/B2)

iv. Linear order coincidences related to sentence subjects.

(16a) ... [ter o tempo para ganhar mais **dinheiro**]Subject **é** mais importante. (COPLE2, zh007CVATF)
'... [having the time to earn more **money**]Subject **is** more important.'
(bad result from CQL expression 20: Subject-verb agreement common noun/B2)

(16b) ... [viver no **estrangeiro**]Subject **é** como uma faca de dois grumes... (COPLE2, zh081CVATI)
'... [living in a foreign **country**]Subject **is** like a double-edged knife...'
(bad result from CQL expression 20: Subject-verb agreement common noun/B2)

iv. Linear order coincidences related to omitted subjects.

(17a) ... e "roubaram" todas as **riquezas que** [Omitted subject 'eles'] **lá tinham**. (COPLE2, es032CVATF)
'... and "stole" all the riches that '[Omitted subject 'they'] **there had**.
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

(17b) ... a **língua que** [Omitted subject 'ele'] **fala**, a tradição que segui,... (COPLE2, zh077CVATF)
'... the **language that** [Omitted subject 'he'] **speaks**, the tradition [I] followed, ...
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

vi. Linear order coincidences related to indefinite subject constructions with *-se*

(18a) ... o mesmo, com a **diferença que hoje se**_clitic **migra** para viver, porque... (COPLE2, es087CVATI1)
'... the same, with the **difference that today one migrates** to live, because...'
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

(18b) ... , com a **esperança que não se**_clitic **volte** atrás e que não sejam... (COPLE2, it101CSATF)
'..., with the **hope that one does not go** back and that [they] do not be...'
(bad result from CQL expression 22: Subject-verb agreement relative pronoun/B2)

vii. Lexical ambiguity

(19a) Nós gostamos de ser o **centro** do mundo, ... (COPLE2, zh017CVATF)
'We like to be the **center** of the world...'
(bad result from CQL expression 9: Urban space vocabulary/B2)

(19b) ... gente que vive na **rua** e que não tem... (COPLE2, de026CVATD)
... people that live in the **street** and that does not have...'
(bad result from CQL expression 9: Urban space vocabulary/B2)

The bad results listed here can be expected considering the fact that i) we are dealing with automatically tagged data, which means that even with high scores, there is always a small margin for error; ii) we are dealing with data annotated only at the part-of-speech level, with no syntactic analysis or parsing; iii) and we are dealing with natural language data that carries structural and lexical ambiguity. For these reasons, and considering the percentage of good results achieved, we feel quite confident that the CQL expressions here devised are efficient, robust, and sufficiently motivated.

4. Final remarks

The CQL expressions proposed in this paper allow for the extraction of the target structures aimed at with very good results. Besides being ready to use, these expressions can also be easily adapted to accommodate different sets of vocabulary (e.g., in CQL expressions 8 and 9, for instance), or to select for different levels of proficiency depending on the curricula in use (e.g., changing the value of the 'match.text_proficiency' attribute).

Although not representing an innovative perspective on the data or on the process - CQL and regular expressions have been widely used since many decades -, it is our belief that having linguistically motivated expressions, with high degree of precision, can be very useful for researchers working in LE and FL acquisition, in phenomena of linguistic interference, and analysis and diagnosis of LE/L2 proficiency levels. For instance, besides extracting data by specific proficiency levels, it is possible to correlate level of proficiency, mother tongue of the learner and the occurrence of target structures and/or errors. Moreover, the work here depicted provides usable query expressions that are not easily built using the COPLE2 Query Builder. As shown in Figure 3, the options available to the user do not cover the range of attributes and of operators used in the CQL expressions we propose here.

The description of the process presented here also contributes to further enhancements, either in what concerns the development of other CQL expressions to extract other phenomena, either in what concerns the development of the COPLE2 query interface. For instance, the queries presented here can be added to the COPLE2 web interface as pre-built queries. This way, the users could use these queries immediately and would not have to build the CQL expressions from scratch.

The possibility of extracting data not explicitly marked in the corpus, such as the target structures considered here, potentiates the use of COPLE2 corpus for studying these types of structures in terms of frequency of occurrence, frequency by proficiency level, relation between target structures and mother tongue (for instance, adding the attribute `:: match.text_mothertongue` to the CQL expressions to further restrict the queries), as well as testing and validating hypothesis in larger data sets. It can also easily and immediately contribute to the building of didactic materials based on real data and real difficulties.

Acknowledgments

We are very grateful to the anonymous reviewers for their comments and suggestions.
We thank Joana Oliveira for her involvement and valuable input in the earlier stages of this work.

Funding

Part of this research is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020 and UIDP/LIN/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

References

Amaro, Raquel, Susana Correia, Carolina Gramacho & Amália Mendes (2020) Automatização no diagnóstico de nível de língua: Anotação e versatilidade dos recursos. *Revista da Associação Portuguesa de Linguística* (7), pp. 1–20. https://doi.org/10.26334/2183-9077/rapln7ano2020a1

Cambridge (2012) Cambridge Sketch Engine - Using Corpus Query Language (CQL) (1.3). Cambridge University Press. Available at https://www.cambridge.org/sketch/help/userguides/CQL%20Help%201.3.pdf

Castello, Erik, Katherine Ackerley & Francesca Coccetta (eds.) (2016) *Studies in learner corpus linguistics*. Peter Lang.

Christ, Oli (1994) A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, pp. 23–32. https://doi.org/10.48550/arXiv.cmp-lg/9408005

del Río, Iria, & Amália Mendes (2018) Error annotation in a Learner Corpus of Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). European Language Resources Association (ELRA), pp. 4116–4119. Available at https://aclanthology.org/L18-1

Gramacho, Carolina, Ana Madeira, Cláudia Martins, Nélia Alexandre, Jorge Pinto & Susana Correia (2019) Por nível: Construção e validação de um teste de colocação para o Português Língua Estrangeira–resultados de um estudo-piloto. *Revista da Associação Portuguesa de Linguística* (5), pp. 172–189. https://doi.org/10.26334/2183-9077/rapln5ano2019a13

Mendes, Amália, Sandra Antunes, Marteen Janssen & Anabela Gonçalves (2016) The COPLE2 Corpus: A learner corpus for Portuguese. In *Proceedings of the 10th Language Resources and Evaluation Conference – LREC'16*, pp. 3207–3214. Available at https://aclanthology.org/L16-1

Talhadas, Rui (2016) Mapping grammatical structures onto proficiency levels. In *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*. Available at http://propor2016.di.fc.ul.pt/wp-content/uploads/2016/07/RuiTalhadasPROPORSRW2016.pdf

Tracy-Ventura, Nicole & Magali Paquot (eds.) (2020) The Routledge handbook of second language acquisition and corpora. Routledge.

Verbo-suporte escondido “com rabo de fora”: Construções complexas com verbo-suporte reduzido

Jorge Baptista [1,2] *e Nuno Mamede* [2,3]

[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
[2] INESC-ID Lisboa, Human Language Technology Lab
[3] Universidade de Lisboa, Instituto Superior Técnico

Resumo

Os verbos-suporte são elementos gramaticais que servem principalmente para ‘conjugar’ os nomes predicativos. A redução (ou elipse) desses verbos é um fenómeno muito comum em textos. No processo de desenvolvimento de um Léxico-Gramática das construções com verbos-suporte e nomes predicativos em português, torna-se crucial estabelecer definições precisas para combinações de verbo-substantivo nas quais o verbo funciona como verbo-suporte. Essa diferenciação é necessária, mesmo quando as construções parecem formalmente idênticas, para representar adequadamente o significado de expressões linguísticas. Este estudo revisita alguns desses cenários e tem como objetivo contribuir para estabelecer critérios formais que facilitem essa distinção.

Palavras-chave: nome predicativo, verbo-suporte, redução, Léxico-Gramática, português.

Abstract

Support verbs are grammatical elements that primarily serve to 'conjugate' predicative nouns. Reduction (zeroing or ellipsis) of these verbs is a highly common phenomenon in texts. In the process of developing a Lexicon–Grammar for support verb constructions with predicative nouns in Portuguese, it becomes crucial to establish precise definitions for verb-noun combinations in which the verb functions as a support verb. This differentiation is necessary, even when the constructions appear formally identical, to adequately represent the meaning of linguistic expressions. This study revisits some of these scenarios and aims to contribute to establishing formal criteria that facilitate this distinction.

Keywords: predicative noun, support-verb, reduction/zeroing, Lexicon-Grammar, Portuguese.

1. Introdução

Analisamos neste artigo um aspeto das *construções com verbo-suporte* (CVS) e nome predicativo, adotando o quadro teórico-metodológico da Gramática Transformacional de Operadores (Harris, 1964, 1982, 1991), sob a perspetiva do Léxico-Gramática (M. Gross, 1981, 1996, 1998). A frase (1) ilustra uma dessas construções:

(1) O Pedro tem (uma grande) coragem

1.1. Aspetos gerais das construções com verbo-suporte

Os nomes predicativos são o núcleo predicativo das CVS e são responsáveis pela seleção dos argumentos e pelas propriedades distribucionais, estruturais e transformacionais dessas construções. Os verbos-suporte (*Vsup*) são elementos (praticamente) vazios de sentido, não têm distribuição característica e são elementos gramaticais, funcionando como elemento auxiliar do nome predicativo, servindo, praticamente, para “conjugar”

os nomes predicativos (M. Gross, 1981). É este nome predicativo que é o núcleo da predicação, já que os *Vsup* servem basicamente para veicular os valores gramaticais de tempo-modo-aspeto e pessoa-número, que o nome predicativo não pode suportar. Para uma descrição mais abrangente das CVS e das propriedades características que permitem identificar um verbo como *Vsup*, numa dada combinação verbo-nome, recomendamos ao leitor a consulta da extensa bibliografia sobre o tópico (para uma síntese recente, Fotopoulou et al., 2021), com especial referência, para o português, a Ranchhod (1990), Baptista (1997a, 2005a), Baptista e Mamede (2020b) e Rassi (2023), entre muitos outros.

Antes de prosseguirmos, faremos uma breve nota sobre os determinantes e modificadores das CVS. Estes podem variar de acordo com as restrições de seleção impostas pelo nome predicativo a essa construção e, mais especificamente, pela escolha do *Vsup* (Baptista, 2005a; Ranchhod, 1990; Rassi, 2023). Alguns desses nomes são geralmente mais naturais com um determinante indefinido, eventualmente acompanhados de um modificador adjetival (por exemplo, *O Pedro tem uma enorme coragem*); ou com um determinante *zero* (E), ou seja, sem determinante nem modificador adjetival (por exemplo, *O Pedro tem coragem*), como sucede no exemplo (1). (Também devemos considerar os casos em que o determinante da CVS é fixo, de que resulta um certo grau de cristalização da expressão, como se vê em *O Pedro está em* o/**um*/E (*verdadeiro*) *auge da sua carreira*; neste exemplo, apenas a forma com o artigo definido é aceitável). Nos exemplos deste artigo, vamos abstrair-nos dessas diferentes distribuições sempre que essas diferenças não sejam relevantes para a argumentação.

1.2. **Conceitos de verbo-suporte e de variante de verbo-suporte**

Este artigo aborda a questão da determinação do estatuto de um verbo como *Vsup* em certas combinações de verbo e nome. Por essa razão, é importante, em primeiro lugar, definir e delimitar os conceitos de *Vsup* elementar e de variante de *Vsup*. O *Vsup elementar* é o suporte semanticamente mais neutro e possui uma distribuição lexical mais ampla para a construção de uma CVS com um determinado nome predicativo. Frequentemente, as CVS apresentam *variantes do Vsup elementar* selecionado pelo nome predicativo (M. Gross, 1981). Essas variantes introduzem diferentes nuances semânticas na CVS, principalmente de natureza *aspetual* (1a) ou *estilística* (1b):

(1a) O Pedro tem $_{\text{neutro}}$ / ganhou $_{\text{incoativo}}$ / perdeu $_{\text{terminativo}}$ coragem
(1b) O Pedro tem $_{\text{neutro}}$ / acalenta $_{\text{var.estilística}}$ / nutre $_{\text{var.estilística}}$ esperança de ganhar o prémio

Em cada CVS, é o nome predicativo que determina não só a seleção do *Vsup* elementar, mas *igualmente* das variantes de *Vsup* com que se constrói.

Por outro lado, um mesmo nome predicativo (isto é, a mesma *palavra*) pode entrar em diferentes CVS. Assim, por exemplo, *coragem* e o seu antónimo *cobardia/covardia* apresentam duas CVS com distribuições muito diferentes de *Vsup* elementares e das respetivas variantes:

(2) O Pedro tem/perdeu/ganhou coragem

(3) O Pedro tem/?perdeu/*ganhou cobardia/covardia

(4) O Pedro ?fez/cometeu/perpetrou/praticou uma grande cobardia/covardia

(5) *O Pedro fez/cometeu/perpetrou/praticou uma grande coragem

Nos exemplos acima, o nome predicativo *cobardia/covardia* apresenta duas construções sintáticas distintas. No exemplo (3), o nome predicativo indica uma característica psicológica ou moral (uma **qualidade** ou **defeito**) de um sujeito humano. (Destacamos em **negrito** as classes semânticas ou a interpretação dos

predicados, bem como os papéis semânticos). Nessa construção, é utilizado o *Vsup* elementar *ter* e a CVS corresponde à nominalização da frase *O Pedro é cobarde/covarde*. Já no exemplo (4), o mesmo nome passa a designar uma **ação**, que é *qualificada* pelo sujeito da enunciação. Nessa construção, é utilizado o *Vsup* elementar *fazer* e, de forma geral, o significado desta frase corresponde à interpretação da frase (6):

(6) O Pedro$_i$ fez/cometeu/perpetrou/praticou uma ação que foi grande cobardia/covardia da sua$_i$ parte

Ora, o nome *coragem* só apresenta a construção (2), que denota uma **qualidade**, tendo por variantes (aspetuais) do *Vsup* elementar *ter* os *Vsup perder* (variante terminativa) e *ganhar* (variante incoativa). Por seu turno, o nome predicativo *cobardia/covardia*, nesta construção (3), não parece admitir tão bem estas variantes (especialmente *ganhar*, para a qual não se encontrou atestação, já que a frase com *perder* pode ser facilmente atestada, embora com uma frequência baixa). Quanto à construção (4), que denota **ação**, apenas o nome *cobardia/covardia* a admite, apresentando, a par do *Vsup* elementar *fazer*, o conjunto de variantes estilísticas *cometer*, *perpetrar* e *praticar*. O nome predicativo *coragem* não admite de todo esta construção (5), qualquer que seja o *Vsup* ou a variante escolhida.

Observa-se que a construção com o *Vsup fazer* está amplamente atestada no português do Brasil (PB), mas não na variedade europeia (PT), como pode ser facilmente verificado no corpus *PtTenTen18* (Kilgarriff et al., 2014; Wagner Filho et al., 2018), acessível por meio da plataforma Sketch Engine[1], a que recorremos para a atestação de algumas das construções aqui estudadas. No entanto, considera-se que *fazer* é o *Vsup* elementar desses nomes que permitem a variante *cometer*; de facto, esse *Vsup* é especificamente selecionado por vários nomes predicativos que denotam **atos ilícitos** (ou criminosos), como por exemplo, *O Pedro fez/cometeu um crime* (cf. Chacoto, 2005). A indicação de *fazer* como o *Vsup* elementar desse subconjunto de nomes predicativos permite regularizar/generalizar a descrição linguística das CVS, considerando *cometer* como uma variante "especializada" do verbo *fazer*, que identifica uma classe particularmente homogénea de predicados semânticos.

Vemos, assim, que: (i) um nome predicativo pode entrar em diferentes CVS, passando a denotar predicados semânticos de natureza diferente em cada uma dessas construções; (ii) a definição das diferentes CVS de um mesmo nome predicativo implica não só a determinação da respetiva frase elementar, com o correspondente *Vsup* elementar, mas também a indicação das (principais, ou mais frequentes) variantes desse *Vsup*, igualmente selecionadas por esse nome predicativo em cada construção.

Naturalmente, podem ser observadas algumas regularidades, como é o caso da variante *cometer* para um subconjunto das CVS com o *Vsup* elementar *fazer*, com nomes que denotam um certo tipo de **ação**. Contudo, esse trabalho de descrição linguística mais detalhada requer uma análise individual, *caso a caso*, da distribuição das variantes de *Vsup* em cada CVS. Isso pode ser percebido ao comparar o par de antónimos *coragem/cobardia*, quando usados no sentido de **qualidade** (ou **defeito**), os quais apresentam variantes aspetuais distintas (cf. Fotoupolou et al., 2021).

Essa descrição mais detalhada é um trabalho em andamento, que faz parte de um projeto mais amplo de construção de um léxico-gramática das construções com *Vsup* (CVS) do português europeu. Esse projeto teve início com Baptista e Mamede (2020a, 2020b) e abrange a problemática abordada neste artigo.

1.3. **Definição do problema**

O problema que pretendemos analisar neste artigo é o seguinte: Certas combinações verbo-nome parecem relevar do mesmo processo de variação de *Vsup* (7a):

[1] https://app.sketchengine.eu [consultado em 31/05/2023].

(7a) O Pedro demonstrou/mostrou/revelou uma certa coragem

mas permitem a reconstituição do *Vsup* subjacente, *ter*, numa oração infinitiva (7b):

(7b) O Pedro$_i$ demonstrou/mostrou/revelou **ter$_i$** uma certa coragem$_i$

Estes verbos (Baptista & Mamede, 2020a; classes **09x**) apresentam geralmente uma construção com sujeito humano e uma oração subordinada completiva-objeto na posição de complemento direto, admitindo na completiva um sujeito correferente do sujeito do verbo, o que é assinalado nos exemplos abaixo pelos índices de correferência 'i'; nessa completiva, reencontramos a CVS subjacente com o *Vsup ter* expresso (8a):

(8a) O Pedro$_i$ demonstrou/mostrou/revelou que tinha$_i$ coragem

Os verbos da classe de construções **09x** apresentam ainda um complemento indireto (dativo) (8b):

(8b) O Pedro$_i$ demonstrou/mostrou/revelou *ao João/-lhe* que tinha$_i$ coragem

A maioria dos verbos pertencentes a estas classes exprime um significado de verbo *de dizer* (*verbum dicendi*), ou seja, um predicado de **comunicação** (sobretudo os verbos da classe **09I**). Este complemento indireto (dativo) corresponde ao **interlocutor**/destinatário da **mensagem**; este último papel semântico é expresso pela oração completiva na função de complemento direto do verbo; o sujeito da construção é interpretado como um **agente-locutor**. Sobre esta classe de verbos, veja-se, entre outros, Baptista (2010) e Reis et al. (2021). Em muitos casos, este complemento dativo encontra-se omitido, portanto, nos exemplos apresentados, não consideramos o complemento dativo.

No entanto, neste caso, não é obrigatória a correferência entre o sujeito do verbo da oração principal e o sujeito da oração subordinada, como se pode verificar na frase (9a):

(9a) O Pedro demonstrou/mostrou/revelou que a Ana tinha coragem

A redução da completiva finita a infinitiva (ou *desfinitização*) pode igualmente ocorrer (9b), acompanhada eventualmente de inversão do sujeito da oração infinitiva:

(9b) ?O Pedro demonstrou/mostrou/revelou a Ana ter coragem/ter a Ana coragem

Pelo contrário, as variantes do *Vsup* elementar não permitem a reconstituição desse verbo, (10a)-(10b):

(10a) O Pedro ganhou/perdeu (*ter) coragem
(10b) O Pedro acalenta/nutre (*ter) esperança de ganhar o prémio

O *Vsup* elementar pode ainda ser reconstituído noutros contextos sintáticos, independentemente da referência de *coragem* com o sujeito, concretamente, numa oração relativa (11), a qual está transformacionalmente associada às formas resultantes da redução dessa relativa (12-13), seja sob a forma de pronome oblíquo (12) num complemento *de N*, seja sob a forma desse complemento a pronome possessivo (13):

(11) O Pedro$_i$ demonstrou/mostrou/revelou a coragem que E$_i$/ele$_{i,j}$/a Ana$_j$ tinha

(12) O Pedro$_i$ demonstrou/mostrou/revelou a coragem dele$_{i,j}$/dela$_j$

(13) O Pedro$_i$ demonstrou/mostrou/revelou a sua$_{i,j}$ coragem

Portanto, é pssível observar na oração relativa com *Vsup* expresso um sujeito de *ter coragem* que não é correferente ao sujeito dos verbos *demonstrar*, *mostrar* e *revelar*. Como esperado, parece haver uma leitura preferencial em (11) no caso de um sujeito reduzido (E) ou sujeito pronominal (*ele*) na oração relativa; nomeadamente, a primeira redução, a zero, leva a uma situação de correferência com o sujeito da oração principal, ao passo que a forma reduzida pronome permitirá ambas as interpretações. Nas formas reduzidas a pronome (12) e (13), essa dupla interpretação também se verifica.

1.4. **Hipótese**

M. Gross (1998, p. 35) sugeriu que as construções em francês equivalentes às frases de (2a), com *demonstrar*, *mostrar* ou *revelar* e um nome predicativo como *coragem*, fossem analisadas como variantes **intensivas** de *Vsup*, que serviriam para a expressar "uma modalidade de exteriorização da qualidade associada ao nome suportado" (tradução nossa).

Pelo contrário, a hipótese de análise que aqui avançamos é a de que estes verbos *não* são *Vsup*, já que parece observar-se, neste contexto, a possibilidade de reconstituição nestas construções do *Vsup* subjacente da CVS de base desses nomes predicativos. Tal permitiria analisar estas frases como frases complexas, resultantes da concatenação de duas frases elementares: (i) uma frase elementar com os verbos (*plenos* ou *distribucionais*) *demonstrar*, *mostrar* ou *revelar* e (ii) uma CVS e nome predicativos encaixada sob aqueles verbos. Essa observação nessas frases permitiria descartar a análise destes verbos como sendo *Vsup* e, pelo contrário, considerá-los como verbos plenos (ou distribucionais).

Em rigor, tratar-se-ia de aplicar o princípio geral harrissiano da *ordem parcial de entrada das palavras na frase*, e que constitui a primeira e mais importante das restrições fundamentais deste quadro teórico (Harris, 1991, p. 60). Este princípio estabelece que se *A* entra na frase imediatamente após *B* (*A>B*) e *se não houver nenhum elemento C tal que A>C>B*, dizemos que *A* é um operador sobre *B* e que *B* é o argumento *imediato* de *A* (tradução nossa). É este princípio geral que permite considerar, por exemplo, que verbos como *continuar* não deveriam ser tratados como *Vsup* e, sim, como verbos auxiliares (aspetuais) (Baptista et al., 2010; Baptista & Crismán, 2021) de um *Vsup* em frases como *O Pedro* continua a estar sob *vigilância médica = O Pedro* continua sob *vigilância médica* (cf. *O Pedro* está sob *vigilância médica*). Já outros verbos parecem funcionar (sobretudo) como verdadeiros *Vsup*: *O Pedro* está/ficou sob *vigilância médica* cp. **O Pedro* ficou a estar sob *vigilância médica*.

Para confirmar essa hipótese, verificamos *sistematicamente* as combinações dos nomes predicativos do léxico-gramática do português (Baptista & Mamede, 2020b) com estes verbos e as variantes do *Vsup* elementar, apoiados nas ocorrências destas combinações em *corpora*. Se se verificar a possibilidade de regularmente reconstituir sob os verbos *demonstrar*, *mostrar* ou *revelar* e um complemento direto preenchido por um nome predicativo como *coragem* o *Vsup* elementar desse nome; ou, inversamente, se se puder formar regularmente, a partir da frase com o *Vsup* expresso, a frase com o nome predicativo como complemento direto daqueles verbos; em ambos os casos, sem que a frase mude de significado ou perca a sua aceitabilidade; será descritivamente mais adequado descrever aqueles verbos como verbos plenos e as frases com *Vsup* reduzido como frases complexas resultantes da concatenação da construção desses verbos plenos com uma CVS e nome predicativo.

A estrutura deste artigo é a seguinte: na Secção 2, descrevemos a metodologia adotada neste estudo, começando pela apresentação dos recursos e ferramentas utilizados (2.1), seguida pelo procedimento de extração dos pares verbo-nome a partir do *corpus* (2.2). Na Secção 3, realizamos uma análise e discussão dos resultados obtidos por meio desse procedimento, levando em consideração as análises concorrentes apresentadas neste trabalho. Na Secção 4, resumimos as conclusões decorrentes dessa análise.

2. Metodologia

Nesta secção, apresentamos sucintamente a metodologia seguida nesta investigação.

2.1. Recursos e ferramentas

Para verificar sistematicamente as combinações verbo-nome relevantes, utilizamos os seguintes recursos linguísticos e ferramentas, que serão descritos de forma sucinta já a seguir:

1. O Léxico-Gramática dos nomes predicativos do português (Baptista & Mamede, 2020b): Trata-se de uma base de dados linguísticos que contém aproximadamente 9.120 entradas de nomes predicativos (Figura 1). Essas entradas são organizadas em formato de tabela, em que as CVS figuram nas linhas e as diferentes propriedades léxico-sintáticas e semânticas são representadas nas colunas. A base de dados fornece informações detalhadas sobre cada CVS de um nome predicativo, incluindo: (i) o número de argumentos, isto é, o sujeito e os eventuais complementos; (ii) as restrições distribucionais quanto ao preenchimento lexical de cada posição argumental, usando principalmente os traços distribucionais de humano/não-humano; (iii) os papéis semânticos correspondentes, a partir da listagem de papéis semânticos elaborada para a descrição das construções verbais e apresentada em Baptista e Mamede (2020a); (iv) as preposições (ou conjunto de preposições) que introduzem os complementos; (v) os *Vsup* selecionados pelos nomes predicativos em cada construção, tanto nas construções *standard*, de orientação 'ativa' (isto é, com sujeito **agente**), e.g. *O Pedro deu/pregou um soco a/em_o João*, como nas construções *conversas* (Baptista, 1997a, 2005a; G. Gross, 1989) de orientação 'passiva' (isto é, com sujeito **objeto** ou **paciente**), e.g. *O João levou/apanhou um soco do Pedro*; (vi) algumas propriedades transformacionais, como, por exemplo, (a) a possibilidade de *pronominalização dativa* de alguns complementos de tipo humano introduzidos por *a*, (b) bem como os complementos dativos resultantes de reestruturação de grupo nominal (Baptista, 1997b); ou (c) a possibilidade de o nome entrar numa construção *simétrica* (Baptista, 2005b); ou ainda, (d) o facto de a CVS admitir o que chamámos de *construção neutra* (Baptista et al., 2022a, 2022b); (vii) a possibilidade de o nome predicativo poder ser encaixado sob um *verbo-operador* (M. Gross, 1981), seja este um verbo-operador causativo, como *dar* em *Esse acidente* deu *ao Pedro medo de andar de avião*; cf. *O Pedro tem medo de andar de avião*) ou um verbo-operador de ligação, como *ter* em *O Pedro tem o João às suas ordens* cf. *O João está às ordens do Pedro*; (viii) construções com *negação obrigatória*, por exemplo, *O Pedro* (**não*) *tem um chavo*; *O Pedro* (**não*) *esteve pelos ajustes*). Além disso, é fornecido um exemplo ilustrativo de cada construção.

2. Uma listagem de cerca de 175 verbos que entram na formação de construções com *Vsup* (Baptista & Mamede, 2020a). Nesta listagem, é fornecido apenas um exemplo da CVS, sem considerar a análise da extensão lexical dos predicados nominais que selecionam esses *Vsup*.

3. O *corpus* CETEMPúblico (Rocha & Santos, 2000), que é constituído por mais de 190 milhões de palavras, integralmente processado pela cadeia de processamento de linguagem natural STRING (Mamede et al., 2012), especificamente desenvolvida para processamento computacional do português. Trata-se de um sistema híbrido, que combina métodos de análise baseados em regras com aprendizagem automática, e apresenta uma ampla cobertura lexical. Além das informações lexicais e morfossintáticas associadas a cada unidade lexical (categoria morfossintática, lema, flexão e outras informações sintático-semânticas), o sistema, usando o analisador sintático (*parser*) XIP (Aik-Moktar et al., 2002) analisa as frases procedendo, primeiro, a uma segmentação em constituintes sintáticos mínimos (*chunks*; e.g. grupos nominais: NP, preposicionais: PP, adjetivais AP, adverbiais: ADVP, etc.); e representando em seguida as relações sintáticas como relações (binárias) de dependência sintática entre palavras, por exemplo, determinante: DETD, entre um determinante definido e o nome núcleo de um NP; ou entre os núcleos dos

chunks previamente constituídos, por exemplo, sujeito: SUBJ, complemento direto: CDIR, modificador: MOD, etc.

2.2. **Extração de combinações verbo-nome**

Na estratégia delineada em Baptista et al. (2015), Rassi et al. (2015), a identificação de uma CVS na cadeia de processamento automática STRING implica o emparelhamento das expressões no texto com os padrões descritos no léxico-gramática dos nomes predicativos. A relação sintática de base entre o *Vsup* e o nome predicativo é capturada através de 3 relações de dependência (14) – (16): CDIR (complemento direto), PREDSUBJ (predicativo) e MOD (complemento preposicional; a dependência MOD captura ainda outras configurações sintáticas, por exemplo, entre um nome e um adjetivo, mas que não são relevantes para este estudo). Neste artigo, além disso, não estudamos as CVS em que o nome predicativo ocupa a posição de sujeito (Baptista, 2022). Assim, quando se verifica a dependência sintática de base adequada entre um determinado nome predicativo e um dos verbos que pode funcionar como seu *Vsup*, o sistema extrai então uma nova dependência, SUPPORT, entre estes dois elementos (note-se que podem ter de verificar-se outras condições, que aqui não consideramos). Os seguintes exemplos ilustram essas relações e as dependências sintáticas de base extraídas pelo sistema, assim como a dependência SUPPORT que identifica a CVS:

(14) O Pedro tem medo do escuro CDIR(tem,medo) SUPPORT_STANDARD(medo,ter)

(15) O Pedro está de greve PREDSUBJ(está,greve) SUPPORT_STANDARD(greve,estar)

(16) O Pedro goza de boa reputação MOD(goza,reputação)
SUPPORT_STANDARD(reputação,gozar)

Procedeu-se, então, à extração automática, a partir do *corpus* CETEMPúblico já processado pela STRING, de todas as combinações verbo-nome em que o nome fosse um nome predicativo presente no respetivo léxico-gramática e que tivesse sido extraída entre estes elementos uma relação de dependência sintática precisa (CDIR, PREDSUBJ ou MOD). Uma das vantagens desta abordagem face a outras estratégias (por exemplo, usando outras ferramentas de pesquisa em *corpora*, igualmente disponíveis), consiste em tirar partido da desambiguação morfossintática prévia, bem como da complexa análise sintática já efetuada pelo sistema STRING; nomeadamente, a identificação das cadeias verbais, constituídas por sequências de verbos auxiliares e um verbo principal (Baptista et al., 2010; Baptista & Crismán, 2021), ou dos casos em que não há uma relação sintática direta entre o verbo e o nome predicativo ou, ainda, das situações em que possam ter sido extraídas dependências entre outros elementos e o nome predicativo em análise.

Ao todo, foram extraídos do *corpus* mais de um milhão (1.113.417) de pares verbo-nome predicativo, distribuídos pelas diferentes relações de dependências, e de que se apresenta na Tabela 1 um pequeno extrato. Esta listagem serve, entre outros objetivos, para identificar combinatórias verbo-nome linguisticamente interessantes, por exemplo, associando à frequência encontrada uma medida de associação apropriada (chi-quadrado, informação mútua, etc.; Manning & Schutze, 1999; Trindade, 2020). Serve igualmente de base ao estudo aqui apresentado e a sua consulta revela aspetos interessantes. (Como os verbos selecionados apenas apresentam uma construção transitiva direta, a coluna da dependência PREDSUBJ está vazia.)

Naturalmente, nem todas as combinações verbo-nome representam uma CVS. Por exemplo, as combinações *ganhar/perder-ação* estão provavelmente relacionadas com o emprego jurídico (*ação judicial*) e não com o uso deste nome em CVS como sucede em *interpor/levantar/mover uma ação*.

Figura 1. Léxico-Gramática dos nomes predicativos do português (extrato com base em Baptista & Mamede, 2020b)

Tabela 1. Combinações verbo-nome predicativo no corpus CETEMPúblico processado pela STRING

nome	Verbo	cdir	predsubj	mod	total
abaixo-assinado	*Demonstrar*	0	0	1	1
abaixo-assinado	*Mostrar*	2	0	1	3
abaixo-assinado	*Revelar*	0	0	2	2
abaixo-assinado	*Ter*	8	0	3	11
abandono	*Demonstrar*	3	0	4	7
abandono	*Ganhar*	0	0	5	5
abandono	*mostrar*	0	0	4	4
abandono	*perder*	1	0	5	6
abandono	*revelar*	1	0	6	7
abandono	*ter*	9	0	22	31
abertura	*demonstrar*	27	0	5	32
abertura	*ganhar*	1	0	21	22
abertura	*mostrar*	96	0	21	117
abertura	*perder*	2	0	24	26
abertura	*revelar*	24	0	17	41
abertura	*ter*	111	0	133	244
abordagem	*demonstrar*	1	0	5	6
abordagem	*ganhar*	0	0	3	3
abordagem	*mostrar*	1	0	24	25
abordagem	*perder*	1	0	2	3
abordagem	*revelar*	3	0	15	18
abordagem	*ter*	67	0	40	107
aborrecimento	*mostrar*	1	0	0	1
aborrecimento	*ter*	6	0	1	7
aborto	*ganhar*	0	0	2	2
aborto	*mostrar*	0	0	2	2
aborto	*perder*	0	0	3	3
aborto	*revelar*	1	0	0	1
aborto	*ter*	6	0	26	32
ação	*demonstrar*	6	0	20	26
ação	*ganhar*	24	0	36	60
ação	*mostrar*	5	0	46	51
ação	*perder*	20	0	47	67
ação	*revelar*	10	0	25	35
ação	*ter*	501	0	433	934

Nota. Extrato de alguns nomes com frequência ≥5 e que ocorrem em relação sintática com os verbos ter, ganhar, perder, demonstrar, mostrar e revelar; a relação sintática é capturada pelas dependências CDIR (complemento direto), PREDSUBJ (predicativo) e MOD (complemento preposicional). Apresenta-se o número de ocorrências por par verbo-nome e por dependência e o total de expressões encontradas

É possível verificar que o mesmo par verbo-nome pode frequentemente apresentar relações de dependência sintática de base diferentes. Por exemplo, o nome *aborrecimento* aparece não só como complemento direto (capturado pela dependência CDIR) do verbo *ter* (concordâncias retiradas do corpus CETEMPúblico; os códigos indicam o respetivo extrato):

> *par=ext9338-soc-95a-2*: O jovem de 23 anos encontrava-se sem emprego, depois de **ter tido aborrecimentos** na loja de materiais de construção onde trabalhou durante algum tempo .

mas também num constituinte analisado pelo *parser* como dele dependente pela relação MOD:

par=ext133888-clt-soc-92b-2: Assim, será fácil ganhar o interesse por uma actividade que ainda hoje **é tida como um aborrecimento** por muitas das crianças .

No caso dos verbos exemplificados na Tabela 1, tratando-se de verbos com uma construção transitiva direta, observa-se regularmente um número elevado de ocorrências com a dependência CDIR. Tal torna estas combinações bons candidatos ao estatuto de CVS. Por outro lado, a ocorrência da dependência MOD resulta de vários fenómenos, geralmente de problemas de ambiguidade, como a análise (incorreta) do artigo *a* como preposição, ou devido a relações extraídas entre elementos a uma maior distância (os dois elementos podem não estar justapostos). Apenas analisando as ocorrências concretas em que essas dependências foram extraídas é possível determinar com rigor se a extração de dependências foi adequada e se se trata ou não de uma CVS.

Finalmente, ao fazer esta análise sistemática, pode verificar-se que certas combinações não se observam no corpus, ainda que pudessem ser perfeitamente naturais. É o caso dos pares *demonstrar/revelar aborrecimento*, que não foram encontrados mas podem ser perfeitamente construídos (*O Pedro demonstrou/revelou aborrecimento por ter de fazer isso*; *Isso demonstrou/revelou o aborrecimento do Pedro em ter de fazer isso*), enquanto que a combinatória *mostrar aborrecimento* se encontra atestada. Por outro lado, a ausência de atestação para as variantes aspetuais *ganhar* e *perder* do *Vsup ter* da CVS de *aborrecimento* parece confirmar a intuição linguística de que este nome nesta CVS não admite tais variantes (*O Pedro teve/*ganhou/*perdeu um grande aborrecimento*).

Apesar destas limitações, esta listagem constitui uma ferramenta muito útil, que permite organizar melhor essa análise, focando nos casos mais frequentes, por exemplo, ou ordenando-os por relevância com base em medidas de associação. Permite ainda contextualizar as dependências-alvo no conjunto de dependências (potencialmente relevantes) que o nome estabelece com outros verbos e, assim, determinar o conjunto de *Vsup* e respetivas variantes de uma dada CVS.

Sendo muito elevado o número de pares verbo-nome predicativo diferentes, decidimos focar-nos, neste artigo, em 6 verbos (*ter*, *ganhar*, *perder*, *demonstrar*, *mostrar* e *revelar*), que relevam do problema de determinar o estatuto destes verbos como *Vsup* ou suas variantes, como foi apresentado acima. Na Tabela 2 apresenta-se o número total de pares verbo-nome predicativo, o número de ocorrências por dependência sintática, o número total de ocorrências.

Tabela 2. Distribuição geral dos pares verbo-nome predicativo, por verbo, no corpus, para os verbos ter, demostrar, mostrar, revelar, ganhar e perder

Vsup	par V-N	Tot-cdir	Tot-predsubj	Tot-mod	Total
Ter	3.656	349.519	0.000	182.030	531.549
mostrar	1.993	12.676	0.000	15.833	28.509
demonstrar	1.399	5.594	0.000	5.590	11.184
revelar	1.924	10.128	0.000	10.407	20.535
ganhar	1.687	20.605	0.000	16.138	36.743
perder	1.843	18.851	0.000	14.821	33.672

Nota. número de pares verbo-nome; número de ocorrências por dependência, número total de ocorrências.

As combinatórias de *ter* com um nome predicativo representam não só o maior número de pares (cerca de 40%) do conjunto de verbos aqui analisados, mas também ocorrem um número de vezes muito superior (4 vezes mais a soma das ocorrências de todos os outros verbos). Os verbos *mostrar* e *revelar* ocorrem num número semelhante de pares (1.993 e 1.924, respetivamente), enquanto *demonstrar* surge em bastante menos pares (1.399). As variantes aspetuais *ganhar* e *perder* surgem num número intermédio de pares (1.687 e 1.843, respetivamente). Considerando o número de ocorrências total, estes verbos variam entre 36,7 mil (*ganhar*) e 11,2 mil ocorrências (*demonstrar*), este último aparentemente bastante menos frequente do que os restantes. Contudo, os verbos *ganhar* e *perder* ocorrem com muito mais frequência (36,7 mil e 33,6 mil ocorrências, respetivamente) do que os verbos *mostrar*, *demonstrar* e *revelar*. Tal pode sinalizar o uso predominante de

ganhar e de *perder* como variantes aspetuais de *ter* (seria necessário consultar esta base de dados para os nomes que se apresentam construídos simultaneamente com *ter* e/ou com *ganhar*/*perder*).

Finalmente, considerando as duas dependências representadas (CDIR/MOD), a distribuição das ocorrências é interessante: *ter* ocorre quase duas vezes mais (1,92) com o nome predicativo como CDIR do que como MOD, enquanto *ganhar* e *perder* apenas 1,27 vezes. A predominância de *ter* explica-se naturalmente como decorrente de se tratar do *Vsup* elementar, logo mais neutro do ponto de vista do significado do que as suas variantes aspetuais *ganhar* e *perder*. Pelo contrário, o verbo *demonstrar* surge praticamente com o mesmo número de ocorrências com o nome na função de CDIR/MOD. Já os verbos *mostrar* e *revelar* surgem em proporções relativamente semelhantes (0,8 e 0,97, respetivamente), com uma ligeira preponderância para MOD. Esta distribuição parece reforçar a distinção entre os dois conjuntos de verbos.

Ora, embora se possam fazer algumas generalizações, cada nome predicativo pode apresentar propriedades específicas, pelo que a análise das variantes de uma dada construção *Vsup* se deve fazer *caso a caso*. Nas linhas que se seguem exemplificamos a situação com o nome predicativo *coragem*.

O nome *coragem* ocorre no *corpus* combinado com um verbo 3.137 vezes. Sem surpresa, *ter* (927) é o verbo mais frequente, seguido de *ser* (217) e de *fazer* (85). Todos estes verbos, noutras combinações, são frequentemente *Vsup*, ou, pelo menos, verbos com valor gramatical e sem distribuição caraterística. Os verbos seguintes, por ordem decrescente de frequência, são todos verbos plenos: *assumir* (68), *enfrentar* (57) e *falar* (38). A Tabela 3 apresenta a distribuição geral dos pares verbo-nome para o nome *coragem* no *corpus* indicando o número de ocorrências observado. Como se pode observar, a maioria das ocorrências de *coragem* corresponde à combinação com *ter* (927), variando os restantes verbos entre 29 e 44 ocorrências, exceto o verbo *perder* que apenas ocorre 9 vezes. Predominantemente, o nome surge como CDIR destes verbos, ainda que se verifique um número interessante de ocorrências de coragem como MOD de *ter*.

Tabela 3. Distribuição dos pares verbo-nome, com o nome coragem, no corpus e os verbos ter, ganhar, perder, demostrar, mostrar e revelar

Nome	verbo	cdir	predsubj	mod	total
Coragem	*demonstrar*	26	0	4	30
Coragem	*ganhar*	39	0	5	44
Coragem	*mostrar*	34	0	2	36
Coragem	*perder*	5	0	4	9
Coragem	*revelar*	28	0	1	29
Coragem	*ter*	794	0	133	927

Nota. Número de ocorrências por dependência; listagem por ordem alfabética do verbo

3. **Resultados**

Procedeu-se, em seguida à análise das combinações verbo-nome específicas. Para tal, extraíram-se as concordâncias destes pares, o que foi feito em duas etapas:

(i) primeiro as sequências verbo-nome admitindo o verbo *ter* entre eles e uma janela de 0 a 5 palavras entre cada elemento; o objetivo foi capturar os casos com o *Vsup* elementar do nome predicativo explícito; e

(ii) de seguida, as sequências verbo-nome que pretendemos analisar, em que o nome predicativo ocorre como complemento direto do verbo.

Para as primeiras, a análise das concordâncias deveria demonstrar a possibilidade de *redução* da CVS na completiva infinitiva, confirmando o estatuto destes verbos como verbos plenos e não como *Vsup* dos nomes predicativos nessas frases. Para as segundas, a análise das concordâncias deveria demonstrar a possibilidade de

reconstituição da CVS encaixada na forma de uma completiva infinitiva com o *Vsup* do nome predicativo explícito. Nas linhas seguintes apresentamos alguns extratos do *corpus* que emparelharam com os dois padrões de procura acima referidos. Os padrões de procura encontram-se expressos por expressões regulares no formalismo CQL (*corpus query language*) e estão aqui simplificados. Apresenta-se o identificador do extrato, para referência, e assinala-se a negrito as palavras que emparelham com os termos da expressão regular. Por limitações de espaço e maior clareza, alguns exemplos encontram-se truncados.

i. Sequências com verbo-suporte explícito

Procura: [lema="demonstrar"] []{0,5} [lema="ter"] []{0,5} [lema="coragem"].

> Par=ext208861-opi-98b-1: Só lhes resta mesmo é **demonstrar** que **têm coragem** para mudar alguma coisa.
> Par=ext531627-soc-94b-1: Estamos a ser úteis e **demonstramos** que em Portugal a AMI **tem** estrutura e **coragem** para ser de facto «a presença humanitária portuguesa no mundo».

Procura: [lema="mostrar"] []{0,5} [lema="ter"] []{0,5} [lema="coragem"].

> Par=ext101718-des-95ª-1: [...], Sorin e Pena [...] foram os heróis de uma equipa aguerrida e que **mostrou** sempre **ter coragem** para ser campeã do mundo.
> Par=ext378313-opi-97ª-2: Tarefa difícil, para a qual o Governo não **mostra ter** a força nem a **coragem** indispensáveis [...]

Procura: [lema="revelar"] []{0,5} [lema="ter"] []{0,5} [lema="coragem"].

> Par=ext1030139-clt-94ª-1: [...] Gibson **revela** não **ter coragem** para sustentar a figura, [...]

Os exemplos acima demonstram a situação em que não há uma relação *direta* entre, por um lado, os verbos *demonstrar*, *mostrar* e *revelar* e, por outro lado, o nome predicativo. Este último ocorre numa construção com *Vsup* explícito, encaixada como uma completiva-objeto daqueles verbos. Para todas as ocorrências encontradas (cerca de uma centena, com o nome *coragem*) é *sempre* possível reduzir a completiva infinitiva, deixando o nome predicativo como complemento direto do verbo principal ((17)-(19); retomamos os exemplos acima):

(17) Só lhes resta mesmo é **demonstrar** (*que têm*) **coragem** para mudar alguma coisa.

(18) [...], Sorin e Pena [...] foram os heróis de uma equipa aguerrida e que **mostrou** sempre (*ter*) **coragem** para ser campeã do mundo

(19) Tarefa difícil, para a qual o Governo não **mostra** (*ter*) a força nem a **coragem** indispensáveis [...]

Note-se, no exemplo (19), a coordenação de duas CVS, v.g. *ter força* e *ter coragem*. Por outro lado, os casos em que a CVS surge sob uma negação, é possível parafrasear a construção fazendo subir o advérbio de negação para a função de modificador do verbo principal e então reduzir o *Vsup*:

(20) [...] Gibson **revela** *não ter* **coragem** para sustentar a figura, [...]

= (21) [...] Gibson *não* **revela** (*ter*) **coragem** para sustentar a figura, [...]

Inversamente, as frases em que o verbo principal surge negado (19) são parafraseáveis pelas construções complexas em que o advérbio de negação tem escopo sobre a CVS (22):

(19) = (22) Tarefa difícil, para a qual o Governo **mostra** *não ter* a força nem a **coragem** indispensáveis [...]

Verificou-se igualmente que os casos em que não há correferência entre o sujeito do verbo e o sujeito do nome predicativo (completivas finitivas) são em número muito menor dos que os casos (completivas infinitivas) em que essa correferência existe ((23); retomamos o exemplo acima):

(23) Estamos a ser úteis e **demonstramos** que em Portugal a AMI **tem** estrutura e **coragem** para ser de facto «a presença humanitária portuguesa no mundo».

ii. Sequências com o verbo-suporte implícito

Trata-se de padrões semelhantes aos anteriores, mas em que apenas se considera o par verbo-nome predicativo analisado:

Procura: [lema="demonstrar"] []{0,5}[lema="coragem"].

Par=ext90546-clt-95ª-1: No entanto, [...] o grupo **demonstrou coragem** e criatividade suficientes para dar a volta à questão.
Par=ext32603-opi-97ª-2: Mais uma vez, **mostrou coragem** e frontalidade.
Par=ext492850-soc-95b-2: Mota **revelou coragem** [...].

Em numerosos casos, é efetivamente possível reconstituir uma CVS subjacente, como sucede nos exemplos (24) – (25), abaixo, que retomamos das concordâncias anteriores:

(24) O grupo **demonstrou** (*ter*) **coragem** e criatividade suficientes para dar a volta à questão

(25) Mais uma vez, **mostrou** (*ter*) **coragem** e frontalidade

(26) Mota **revelou** (*ter*) **coragem**

Note-se, já agora, os dois exemplos de coordenação, em que *coragem* ocorre coordenado com *criatividade* e *frontalidade*, ambos nomes predicativos que se constroem com o *Vsup ter*:

(27) O grupo demonstrou (ter) criatividade

(28) [Ele] mostrou (ter) frontalidade.

Além disso, é possível encontrar, sob os verbos principais, o grupo nominal (27) resultante da *relativização* da CVS:

par=ext1307532-clt-94b-2: O interesse em que ela parecesse verdadeira era do próprio Céline que assim quis **demonstrar** uma **coragem** que não *tinha*.

Cp. (24) Céline quis **demonstrar** a **coragem** que *tinha*.

Naturalmente, o facto de se tratar de verbos plenos (ou distribucionais) permite encontrar diferentes situações em que se observam outros processos de concatenação e redução de CVS. No caso de construções com sujeito não humano (e.g. *decisão*):

par=ext360124-des-93b-2: «Mas esta *decisão* já **revelou** muita **coragem**! "

dadas as restrições distribucionais sobre o sujeito do nome predicativo *coragem* na CVS (um nome humano), não é possível uma correferência entre o sujeito do verbo *revelar* e o sujeito do nome predicativo, pelo se torna necessária uma análise mais complexa. Esta construção corresponderia à redução da CVS sob outra construção do verbo *revelar* (classe **01T**; Baptista & Mamede, 2020a):

(29) Mas esta decisão (*tomada por alguém*) já **revelou** (*que uma certa pessoa tinha*) muita **coragem**!

Na classe **01T**, estes verbos não exprimem um predicado de **comunicação**, como sucede com os verbos da classe **09I**. O sujeito destas construções *não* é um **agente-locutor** e é frequentemente preenchido por nomes não humanos, com conteúdo proposicional, mas sem valor de **causa**. Não raro, encontram-se situações com sujeito humano que, não obstante, apresentam esta segunda interpretação (não agentiva), a par do valor comunicativo típico de **09I**:

[s/ ID] Bispo de Díli **revelou «coragem»**

(30) O facto de o Bispo de Díli ter feito isso/a ação do Bispo de Díli revelou (*que* [ele] *tinha*) coragem

A redução da CVS a grupo nominal (M. Gross, 1981) e, neste, do sujeito do nome predicativo a pronome oblíquo ou possessivo, cf. (7) – (10), é muitas vezes indicativa de situações semelhantes, em que se verifica a ausência de correferência entre o sujeito do verbo principal e o sujeito do nome predicativo:

par=ext962580-nd-91b-2: A consumação dos riscos **demonstra** a **coragem** dos agentes das reformas, [...]

(31a) A consumação dos riscos **demonstra** *que os agentes das reformas têm* **coragem**
(31b) A consumação dos riscos **demonstra** *a* **coragem** *que os agentes das reformas têm*
(31c) A consumação dos riscos **demonstra** *a* **coragem** *deles*
(31d) A consumação dos riscos **demonstra** *a sua* **coragem**

Contudo, é igualmente possível encontrar casos em que a correferência de sujeitos autoriza a redução da CVS; a pronominalização pelo pronome oblíquo (32c) num complemento *de N* parece, nesse caso, ser muito menos aceitável, sendo preferível a redução ao grupo nominal com possessivo (32d):

par=ext424593-clt-95b-1: [...] uma dupla de heróis que são chamados a **demonstrar** *a sua* **coragem**, intrepidez, inteligência e fundo humano [...].

(32a) Esses heróis$_i$ são chamados a **demonstrar** *que* [$^{?/?*}$eles$_i$] *têm/tinham* **coragem**
(32b) Esses heróis$_i$ são chamados a **demonstrar** *a* **coragem** *que* [$^{?/?*}$eles$_i$] *têm/tinham*
(32c) ?*Esses heróis$_i$ são chamados a **demonstrar** *a* **coragem** *deles*$_i$
(32d) Esses heróis$_i$ são chamados a **demonstrar** *a sua*$_i$ **coragem**

Apesar de, nesta secção, nos termos centrado nas combinações dos verbos *mostrar*, *demonstrar* e *revelar* com o nome predicativo *coragem*, verificaram-se sistematicamente para todas as 926 ocorrências verbo-nome predicativo do léxico-gramática do português (Tabela 3) as relações acima descritas entre estes verbos e este nome. Não se observaram casos de possibilidade de reconstituição de *ter* sob as suas variantes aspetuais *ganhar* ou *perder*, o que confirma o estatuto diferente destes verbos, como suporte deste nome predicativo.

4. Conclusões

Com base na análise realizada, chegamos à conclusão de que a construção com *Vsup* (CVS) encaixada e com *Vsup* explícito se encontra amplamente atestada para os pares verbo-nome aqui estudados, nomeadamente com os verbos *demonstrar*, *mostrar* e *revelar* (todos na construção pertencente à classe **09I**, conforme descrito em Baptista & Mamede, 2020a). No corpus analisado, encontramos exemplos desse tipo de construção, nos quais é sempre possível reduzir a CVS, deixando o nome predicativo como complemento direto desses verbos. Essa redução pode deixar o sujeito do nome predicativo expresso como seu complemento, quer como um pronome oblíquo *de N*, quer como um pronome possessivo.

Par=ext208861-opi-98b-1: Só lhes resta mesmo é **demonstrar** que **têm coragem** para mudar alguma coisa.

(29) Só lhes resta mesmo **demonstrar** (*que têm*) **coragem**

par=ext905668-des-98b-1: «Quis **mostrar** que **tenho coragem**, que não desisto, enfim, que sou um lutador.

(30) Quis **mostrar** (*que tenho*) **coragem**

Ressalvam-se os casos com negação da CVS, em que uma forma equivalente pode ser obtida por elevação da negação para modificador do verbo principal:

par=ext1030139-clt-94ª-1: [...] Gibson **revela** não ter **coragem** [...].

(31) Gibson **revela** não ter **coragem** = Gibson não **revela** (ter) **coragem**

Em contrapartida, nos casos em que se encontra no *corpus* a combinação dos verbos-alvo e os nomes predicativos, frequentemente é possível reconstruir o *Vsup* da CVS subjacente desses nomes predicativos:

par=ext508852-des-98b-2: [...] Koehlke [...] **demonstrou coragem** [...].

(32) Koehlke demonstrou (*que tinha/ter*) coragem

par=ext138070-pol-91b-2: Se alguns desesperavam, outros insistiam em **mostrar coragem**.

(33) Outros insistiam em **mostrar** (*que tinham*) **coragem**

Assim, é possível generalizar o fenómeno, descartando a análise destes verbos *demonstrar*, *mostrar* e *revelar* como *Vsup* destes nomes predicativos. Inversamente, as variantes de Vsuplementar nunca admitem a reconstituição deste verbo numa completiva ou numa relativa encaixada, o que confirma o seu diferente estatuto sintático.

Notamos, porém, que a coocorrência do par verbo-nome pode ser fortuita e a correferência de sujeitos ser puramente acidental, já que os verbos-alvo entram também em construções com sujeito não-agentivo (classe **01T**, Baptista & Mamede, 2020a), que não é correferente ao sujeito do nome predicativo. Estes casos, embora menos frequentes, também se encontram abundantemente atestados:

par=ext435581-pol-98ª-2: A experiência falhou, mas **mostrou** a **coragem** e a visão de explorador espacial de Wan Hu, diz a legenda

(34) A experiência **mostrou** a **coragem** de Wan Hu

A relação de significado, indicada por M. Gross (1998 p. 35), entre estes verbos e os nomes predicativos com que se combinam na posição de complemento direto, nomeadamente a "exteriorização da qualidade associada ao nome predicativo", pode perfeitamente ser derivada pela *composição* dos significados desses verbos *demonstrar*, *mostrar* e *revelar* (verbos de **comunicação**, da classe **09I**, Baptista & Mamede, 2020a) com o significado da CVS subjacente (na completiva encaixada).

A falta de correferência entre o sujeito do verbo principal e o nome predicativo, ou a interpretação não agentiva do sujeito (mesmo que preenchido por um nome humano) nas combinações entre esses verbos

demonstrar, *mostrar* e *revelar* e os nomes predicativos, em que estes aparecem como seu complemento direto, relevam de outra construção verbal, mais especificamente, pertencente à classe **01T**, definida de acordo com Baptista & Mamede (2020a). Nesse caso, a redução da construção com *Vsup* (CVS) a um grupo nominal, conforme proposto por M. Gross (1981), e sua colocação na posição de complemento desses verbos *demonstrar*, *mostrar* e *revelar*, embora seja um processo transformacional distinto, permite obter uma interpretação global que pode ser derivada igualmente por *composição* do significado da construção verbal com o significado da CVS encaixada. Em ambas as situações, podemos, pois, descartar a análise destes verbos *demonstrar*, *mostrar* e *revelar* como *Vsup* destes nomes predicativos.

A análise aqui proposta é, certamente, extensível a várias outras situações léxico-sintáticas, cujo levantamento e descrição sistemáticos estão em curso, no âmbito da elaboração de um léxico-gramática das construções com *Vsup* (CVS) e nome predicativo do Português.

Agradecimentos

A investigação para este trabalho foi parcialmente apoiada por fundos nacionais através da Fundação para a Ciência e a Tecnologia (Proj. ref.ª UIDB/50021/2020). Os autores gostariam de agradecer a leitura atenta e minuciosa dos revisores anónimos, que nos permitiu clarificar alguns passos do texto.

Referências

Ait-Mokhtar, Salah, Jean-Pierre Chanod & Claude Roux (2002) Robustness beyond shalowness: Incremental dependency parsing. *Natural Language Engineering* 8 (2/3), pp. 121–144. https://doi.org/10.1017/S1351324902002887

Baptista, Jorge (1997a) Sermão, tareia e facada. Uma classificação das construções conversas dar-levar. *Seminários de Linguística* 1, pp. 5–37.

Baptista, Jorge (1997b) Conversão, nomes parte-do-corpo e restruturação dativa. In *Actas do XII Encontro da Associação Portuguesa de Linguística* (Vol. 1). APL, pp. 51–59.

Baptista, Jorge (2005a) *Sintaxe dos predicados nominais com* ser de. Fundação Calouste Gulbenkian & Fundação para a Ciência e o Ensino Superior.

Baptista, Jorge (2005b) Construções simétricas: Argumentos e complementos. In *Estudos de homenagem a Mário Vilela*. FLUP, pp. 353–367.

Baptista, Jorge (2010) *Verba dicendi*: *A structure looking for verbs*. In Takuya Nakamura, Éric Laporte, Anne Dister & Cédrick Fairon (orgs.), *Les tables. La grammaire du français par filemenu*. *Mélanges en hommage à Christian Leclère*. CENTAL & Presses Universitaires de Louvain, pp. 11–20.

Baptista, Jorge (2022) Support verb constructions with predicate noun in subject position. *BULAG - Bulletin de linguistique appliquée et générale* 40. Presses Universitaires de Franche-Comté, pp. 379–397.

Baptista, Jorge & Rafael Crismán (2021) Auxiliary verb constructions in Portuguese and Spanish. A comparative study. Construcciones verbales auxiliares en portugués y español. Un estudio comparativo. *Revista de Lenguas Modernas* 34, pp. 39–57. https://doi.org/10.15517/rlm.v0i34.41462

Baptista, Jorge & Nuno Mamede (2020a) *Dicionário gramatical de verbos do português*. Universidade do Algarve Editora.

Baptista, Jorge & Nuno Mamede (2020b) Syntactic transformations in rule-based parsing of support verb constructions. Examples from European Portuguese. In Alberto Simões et al. (orgs.), *Proceedings of 9th Symposium on Languages, Applications and Technologies* (SLATE 2020), pp. 11:1–11:4. https://doi.org/10.4230/OASIcs.SLATE.2020.11

Baptista, Jorge, Nuno Mamede & Fernando Gomes (2010) Auxiliary verbs and verbal chains in European Portuguese. In *Computational Processing of the Portuguese Language: Proceedings of the 9th International Conference, PROPOR 2010*. Springer, pp. 110–119.

Baptista, Jorge, Nuno Mamede & Sónia Reis (2022a) Support verb constructions across the Ocean Sea. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022*. LREC, pp. 26–36. Disponível em http://www.lrec-conf.org/proceedings/lrec2022/workshops/MWE/pdf/2022.mwe2022-1.6.pdf
Baptista, Jorge, Sónia Reis & Nuno Mamede (2022b) Nomes predicativos e construções neutras. *Revista da Associação Portuguesa de Linguística* 9, pp. 13–30. https://doi.org/10.26334/2183-9077/rapln9ano2022a2
Baptista, Jorge, Amanda P. Rassi, Cristina Santos-Turati, Cláudia Dias de Barros, Oto A. Vale & Nuno Mamede (2015) Integrated processing of support verb constructions in Portuguese. In *4th General Meeting of PARSEME*, 19–20 março 2015, Valletta, Malta. Disponível em http://typo.uni-konstanz.de/parseme/images/Meeting/2015-03-19-Malta-meeting/WG2-Baptista-et-al-abstract.pdf
Chacoto, Lucília (2005) *O verbo* fazer *em construções nominativas predicativas*. Tese de doutoramento, Universidade do Algarve.
Fotopoulou, Aggeliki, Éric Laporte & Takuya Nakamura (2021) Where do aspectual variants of light verb constructions belong? In *Proceedings of the 17th Workshop on Multiword Expressions* (MWE 2021), pp. 2–12. Disponível em https://aclanthology.org/2021.mwe-1.2.pdf
Gross, Gaston (1989) *Les constructions converses du français*. Droz.
Gross, Maurice (1981) Les bases empiriques de la notion de prédicat sémantique. *Langages* 63, pp. 7–52.
Gross, Maurice (1996) Lexicon-Grammar. In Keith Brown & Jim Miller (orgs.), *Concise Encyclopedia of syntactic theories*. Pergamon.
Gross, Maurice (1998) La fonction sémantique des verbes support. *Travaux de Linguistique* 37, pp. 25–46.
Harris, Zellig S. (1964) The elementary transformations. In Enry Hiz (ed.), *Papers on Syntax*. D. Reidel Pub. Co., pp. 211–235.
Harris, Zellig S. (1978) Operator-Grammar of English. *Lingvisticae Investigationes* 2, pp. 55–92.
Harris, Zellig S. (1981) The elementary transformations. Transformations and discourse analysis papers 54. In Henri Hiz (ed.), *Papers on syntax*. D. Reidel, pp. 211–235.
Harris, Zellig S. (1982) *A grammar of English on mathematical principles*. John Wiley & Sons/Wiley-Interscience Pub.
Harris, Zellig S. (1991) *A Theory of Language and Information. A Mathematical Approach*. Clarendon Press.
Kilgarriff, Adam, Miloš Jakubíček, Jan Pomikalek, Tony Berber Sardinha & Pete Whitelock (2014) PtTenTen: A *corpus* for Portuguese lexicography. In Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (orgs.), *Working with Portuguese corpora*, pp. 111–130. https://doi.org/10.5040/9781472593641.ch-006
Rassi, Amanda (2015) *Descrição, classificação e processamento automático das construções com o verbo DAR em Português Brasileiro*. Tese de doutoramento, Universidade Federal de São Carlos.
Rassi, Amanda (2023). *O verbo* dar *em português brasileiro*. Descrição, classificação e processamento automático. Letraria.
Rassi, Amanda, Jorge Baptista, Nuno Mamede & Oto A. Vale (2015) Integrating support verb constructions into a parser. In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology* (STIL 2015), pp. 57–61. Disponível em https://aclanthology.org/W15-5608
Reis, Sónia, Nuno Mamede & Jorge Baptista (2021) Predicados de comunicação em português europeu: nominalizações e nomes autónomos. *Revista da Associação Portuguesa de Linguística* 8, pp. 237–259. https://doi.org/10.26334/2183-9077/rapln8ano2021a16
Trindade, João (2020) D*eepString - Syntax Deep Explorer: Integrating multi-corpora support into a corpus analysis tool*. Dissertação de Mestrado, Universidade de Lisboa, IST.
Wagner Filho, Jorge, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. (2018) The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC 2018), pp. 4339–4344. Disponível em https://aclanthology.org/L18-1686.pdf

Trustful Test Suites for Natural Language Processing

Mariana Cabeça[1]*, Marianna Buchicchio*[2]*, Helena Moniz*[3]

[1] Faculdade de Letras da Universidade de Lisboa / Unbabel
[2] Unbabel
[3] Faculdade de Letras da Universidade de Lisboa / INESC-ID

Abstract

Machine Translation (MT) research has witnessed continuous growth, accompanied by an increasing demand for automated error detection and correction in textual content. In response, Unbabel has developed a hybrid approach that combines machine translation with human editors in post-edition (PE) to provide high-quality translations. To facilitate the tasks of post-editors, Unbabel has created a proprietary error detection tool named Smartcheck, designed to identify errors and provide correction suggestions. Traditionally, the evaluation of translation errors relies on carefully curated annotated texts, categorized based on error types, which serve as the evaluation standard or Test Suites for assessing the accuracy of machine translation systems. However, it is crucial to consider that the effectiveness of evaluation sets can significantly impact the outcomes of evaluations. In fact, if evaluation sets do not accurately represent the content or possess inherent flaws, the decisions made based on such evaluations may inadvertently yield undesired effects. Hence, it is of utmost importance to employ suitable datasets containing representative data of the structures needed for each system, including Smartcheck. In this paper we present the methodology that has been developed and implemented to create reliable and revised Test Suites specifically designed for the evaluation process of MT systems and error detection tools. By using these meticulously curated Test Suites to evaluate proprietary systems and tools, we can ensure the trustworthiness of the conclusions and decisions derived from the evaluations. This methodology accomplished robust identification of problematic error types, grammar-checking rules, and language- and/or register-specific issues, leading to the adoption of effective production measures. With the integration of Smartcheck's reliable and accurate correction suggestions and the improvements made to the post-edition revision process, the work presented herein led to a noticeable improvement in the translation quality delivered to customers.

Keywords: Grammar Error Detection, performance assessment, Test Suites, NLP systems evaluation

Resumo

À medida que o estudo da Tradução Automática (TA) tem vindo a expandir-se ao longo do tempo, a necessidade de detetar e corrigir erros em textos tem também aumentado. Neste sentido, a Unbabel combina tradução automática com pós-edição feita por tradutores e linguistas, para, assim, obter traduções de boa qualidade. De modo a assistir os editores nas suas tarefas, foi desenvolvida uma ferramenta proprietária de deteção de erros denominada de *Smartcheck*, que identifica erros e sugere correções para os mesmos. O método mais recente de identificação de erros de tradução baseia-se em textos previamente pós-editados e anotados (categorizando cada erro de acordo com as suas características), que são fornecidos aos sistemas de tradução automática como sendo o padrão de avaliação ou o *corpus* de teste para avaliar a precisão dos sistemas de tradução. Contudo, é de extrema importância considerar que a eficácia dos *corpora* de teste pode ter um impacto significativo nos resultados das avaliações. De facto, se

estes *corpora* não representarem de forma precisa e representativa o conteúdo, as decisões tomadas com base nas avaliações podem inadvertidamente produzir efeitos indesejados. Assim, é de extrema importância criar *corpora* de teste adequados, cujos dados sejam representativos das estruturas necessárias para cada sistema, incluindo ferramentas como o *Smartcheck*. Neste sentido, o presente trabalho permitiu criar e implementar uma nova metodologia de criação de *corpus* de teste bem fundamentada, que pode ser aplicada no processo de avaliação de sistemas de tradução automática e de ferramentas de deteção de erros. Recorrendo à aplicação deste *corpus* de avaliação, tornou-se possível confiar nas conclusões e ilações obtidas posteriormente. Esta metodologia possibilitou também que todo o processo de identificação de erros e avaliação de regras gramaticais se tornasse mais robusto, bem como o de deteção de problemas específicos por língua e/ou registo, permitindo, assim, adotar diversas medidas necessárias em produção. Por meio de sugestões de correção de erros válidas do *Smartcheck* e das melhorias aplicadas ao processo de pós-edição, o presente trabalho demonstrou ser possível aferir a qualidade das traduções que são entregues a diferentes clientes de forma mais cuidada e consistente.

Palavras-chave: Sistemas de Deteção Automática de Erros, avaliação de desempenho, *corpus* de teste, avaliação de sistemas de PLN

1. **Introduction**

In recent years, there has been a significant surge in interest in the automation of Machine Translation (MT). While MT offers faster, more efficient, and cost-effective translation, it has not yet achieved the quality standard set by human translations. In light of this limitation, the technologies developed at Unbabel have successfully addressed this issue by effectively combining the advantages of MT with the high-quality assurance provided by post-editing.

Evaluation plays a crucial role in the life cycle of any system. Whether conducted manually or automatically through the use of metrics such as BLUE (Papineni et al., 2002), METEOR (Lavie & Denkowski, 2009) or COMET (Rei et al., 2020), it is essential to have a means of measuring the quality of the output. However, one aspect of evaluation that often goes overlooked is the validity and trustworthiness of the conclusions drawn from these assessments. Hence, it is of utmost importance to establish a robust evaluation standard that encompasses representative, relevant, and accurate data pertaining to the content being translated. This approach enables the identification of limitations and facilitates their resolution.

In an attempt to further enhance the quality of translations, Unbabel created Smartcheck, a proprietary Grammar Error Detection tool that aids post-editors during the post-edition stage in order to improve their performance in terms of efficiency and accuracy. Smartcheck highlights possible errors and provides suggestions, thereby enabling post-editors to complete their tasks quicker. In essence, it assists them in error detection and correction, ensuring the delivery of high-quality translations to clients in a timely manner.

The work described hereafter aims to improve Smartcheck's performance and underscore the significance of fairness and quality in evaluation data. Specifically, it emphasizes the creation of reliable Test Suites based on a robust methodology and representative data of high quality. Although Smartcheck has several distinct modules, this study will focus on the assessment of rules and spell checkers. Consequently, the objectives of this study were threefold: i) implementation of a methodology on creating reliable Test Suites for testing a proprietary tool on error detection and editing suggestions; ii) evaluation of the performance of this tool; and iii) contribution to the improvement of quality based on the edits suggested.

2. State of the art

MT research has witnessed remarkable growth in recent years, despite encountering certain challenges along the way. However, there remains a great amount of work to be done to ensure high-quality outputs and complete system automation. When a new MT system is trained, it is crucial to analyze its output to detect errors. These errors serve as valuable opportunities to improve the system's output, as they allow for a deeper understanding of the most recurrent errors. This analysis enables the identification of specific linguistic structures that require greater attention to further refine the MT system.

Current systems still struggle to ensure fluency and coherence in their translated sentences. Consequently, numerous attempts have been made to evaluate and monitor translation quality. These efforts encompass both manual revision and automated metrics. In terms of manual evaluation, the state-of-the-art approach relies on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014). The MQM metric is devised based on error annotations with different degrees of severity and aims to achieve a high level of granularity in the evaluation of translations. However, the manual process of annotating errors is inherently time-consuming, as every translation error in a given text must be identified and classified according to this typology.

On the other hand, automated metrics provide a general score for a given system's output within a matter of seconds. Nevertheless, these metrics are considered to have a low level of correlation with human judgment. Despite their efficiency in terms of speed, their ability to accurately align with human assessment is frequently questioned.

The demand for a more detailed and time-efficient evaluation process has significantly increased. As such, the MT community began exploring alternative evaluation approaches as early as the 1990s. Approaches such as the utilization of test *corpora* and Test Suites, which are distinct techniques that should not be confused with one another. The main distinction lies in the fact that test *corpora* serve as repositories of extensive and potentially unrefined data, while Test Suites consist of carefully curated sets of tests that accurately represent the linguistic structures under analysis.

Test *corpora* often lack meticulousness, making it challenging to isolate specific linguistic phenomena being tested. Furthermore, the absence of annotations in most *corpora* further complicates the evaluation process. Conversely, Test Suites, the focus of our work, comprise lists of sentences deliberately compiled to create a controlled *corpus* of exemplary data, referred to as gold standard data. These Test Suites facilitate diagnostic evaluation of an MT system, as the input used for testing is pre-checked, allowing for control over the vocabulary and the specific linguistic phenomena being tested. This evaluation method proves particularly valuable when it is necessary to present language phenomena in a comprehensive and systematic manner or generate various combinations of phenomena (Balkan et al., 1994).

To properly evaluate a system using Test Suites, the set of tests must be constructed in a methodological way by following a specific approach that is considered appropriate for the evaluation's primary objective (King & Falkedal, 1990). This is crucial because, as Dale et al. (2012, p. 58) state, "if we cannot entirely trust our gold-standard data, then we cannot place too much trust in the results of evaluations carried out using that data". Therefore, the creation of high-quality Test Suites demands meticulous planning, coherence, and systematic execution.

According to Balkan (1994), there are different approaches when constructing Test Suites, one of which is the bottom-up approach. In the bottom-up approach, the system is tested and its functions are analyzed, treating them as attributes. For instance, consider the case of spell checkers. Their functions consist of detecting misspelled words and providing plausible corrections; therefore, their reportable attributes consist of detected misspelled words and of the corresponding correct form, found among the corrections the system proposes. Subsequently, each attribute

is associated with a value, usually a percentage, which is calculated by comparing it to a standard, such as Test Suites.

However, for Test Suites to serve their purpose in evaluating spell checkers or MT systems in general, it is essential that the included phenomena are specifically relevant to the intended application (Balkan, 1994). In other words, Test Suites must comprise representative examples of the phenomena or content that one aims to evaluate. Within the bottom-up approach, there are multiple options for constructing system-specific Test Suites, depending on the type of evaluation required. Nonetheless, two distinct evaluation scenarios can be identified: the "black box" scenario, wherein the evaluator lacks access to the internal workings of the system but can still test hypotheses about its internal mechanisms, and the "glass box" scenario, wherein the evaluator has access to the system's rules. In the latter scenario, the writer of the Test Suite can tailor it to the system's rules, enabling a diagnostic evaluation to identify the root causes of system errors through a root-cause analysis.

Test Suites play a vital role in conducting meticulous and comprehensive evaluations, and their applicability extends beyond machine translation systems (Avramidis et al., 2018). They can be employed to monitor and evaluate other complex systems, including grammatical error detection (GED) and correction (GEC) tools. These tools aim to improve the overall quality of final translations, necessitating thorough evaluation procedures.

GED systems receive potentially erroneous sentences as input and identify tokens that violate specific linguistic rules. On the other hand, GEC systems are tasked with accurately correcting the identified errors while preserving the original meaning of the sentence, thereby optimizing translation quality. However, the process of error detection is intricate, involving various Natural Language Processing (NLP) tools and complex dependencies between tokens. Unlike humans, whose inherent linguistic competence enables them to intuitively identify errors in sentences without explicit knowledge of grammar and language rules, systems lack this intuitive decision-making ability. Instead, systems rely on explicit universal dependencies, word classifications, and language rules to detect errors. Given the complexity of these systems, their evaluation must be detailed and precise.

To achieve an accurate evaluation, one effective approach involves using trustworthy Test Suites as a complementary component of the evaluation process. Test Suites help ensure the evaluation is conducted with the necessary level of scrutiny and accuracy, thereby improving the reliability of the findings. By relying on Test Suites, a comprehensive framework that accounts for various linguistic phenomena can be established, facilitating a thorough examination of the system's performance.

Test Suites, according to Balkan (1994, p. 1), are carefully put-together collections of sentences. They're often created artificially, with each sentence crafted to specifically check how a system deals with specific linguistic phenomena or sets thereof. These inputs can assume various forms, such as complete sentences, sentence fragments, or even sequences of sentences. When these test inputs are carefully chosen ahead of time, it becomes possible to control both the words used and the specific language aspects being tested. This control lets the evaluator focus entirely on how the system handles the specific language tasks at hand, without being distracted by issues related to vocabulary, as explained by Balkan et al. (1994, p. 53).

This evaluation methodology proves especially advantageous when addressing the following three significant objectives, as delineated by Balkan et al. (1994): i) presenting linguistic phenomena in a comprehensive and methodical manner; ii) generating various potential combinations of linguistic phenomena; and iii) systematically deriving negative data from positive data, achieved by deliberately contravening grammatical constraints associated with the positive data set (Balkan et al., 1994, p. 53). Thus, the use of Test Suites serves as an invaluable means to meticulously evaluate systems. Ultimately, they diver from test sets in the sense that they are not corseted to a certain percentage of the total amount of data in a *corpus*, but rather tailored to provide examples of linguistic structures to be consistently tested.

3. **Methodology**

As previously mentioned, ensuring the quality of MT involves employing post-edition. Therefore, post-editors play a vital role in the quality assurance process and in order to facilitate their work and motivate them to accomplish tasks, it is of utmost importance to provide them assistance so as to minimize errors while considering time constraints. This is where Smartcheck assumes a crucial role in supporting post-editors. The primary objective of Smartcheck is to assist post-editors in two fundamental ways: first, by detecting potential errors, and second, by automatically suggesting corrections aligned with the customer's style guidelines and many other requirements. Smartcheck examines errors such as inconsistencies in register and formality, adherence to specific client rules, and overall text coherence. Meanwhile, the task of spelling verification is delegated to external NLP services, such as a word aligner, a syntax parser, and a spell checker. One key feature of Smartcheck lies in its integration of custom language rules implemented through a proprietary programming language known as SURF.[1] These rules address various types of issues, including style, fluency, grammar-related errors, dependency problems, among others. Consequently, Smartcheck represents the culmination of multiple NLP modules and hardcoded rules tailored to different language pairs and can be regarded as an augmented multilingual version of a spell checker, as it not only analyzes grammar and orthography but also delves into morphology and style-adapted client rules.

It is important to note that Smartcheck does not substitute the erroneous form with the correct one. Hence, it should be classified as a GED tool rather than a GEC tool. The ultimate decision rests with the post-editor, who has the final say in accepting or rejecting the suggestions. Thanks to Smartcheck, post-editors can achieve higher translation quality, enabling these refined texts to be utilized as training data for MT systems and driving their continual improvement. This is due to the fact that MT systems operate on the principles of machine learning, wherein the quality of translations relies heavily on the training data provided. When high-quality translations are used as training data, the machine learns to emulate the desired output. Conversely, if the training data is of low quality, the resulting output is likely to be inadequate. In essence, the quality of the input data directly affects the quality of the output produced by the MT system. Therefore, it is crucial to ensure the use of reliable and accurate data during the training phase to achieve desirable translation quality.

The methodology presented in this section was established to evaluate the performance of Smartcheck. The initial step involved conducting a baseline analysis to establish a benchmark for future comparisons. Upon gathering these results, we formulated a hypothesis that attributed the low metric values to the evaluation dataframe employed for assessing the system's performance, rather than the actual grammar-checking rules. Consequently, the initial dataframe, referred to as the prior Evaluation Dataframe (prior EDF), underwent a meticulous examination, leading to the conclusion that revisions and updates were necessary. To address this issue, new Test Suites were methodically and systematically developed, incorporating representative examples of the typical content handled by the systems. These modifications were guided by meticulous typologies to ensure the highest possible quality. Shortly after the application of this methodology, the annotation typology and evaluation standard were updated, leading to further revisions in the rules. Some rules were too extensive and new rules were added to address client requirements. The most ignored rules were identified, prompting necessary changes and discarding unhelpful rules. The following and final step involved evaluating the revised rules with the updated evaluation standard to ensure accurate error detection.

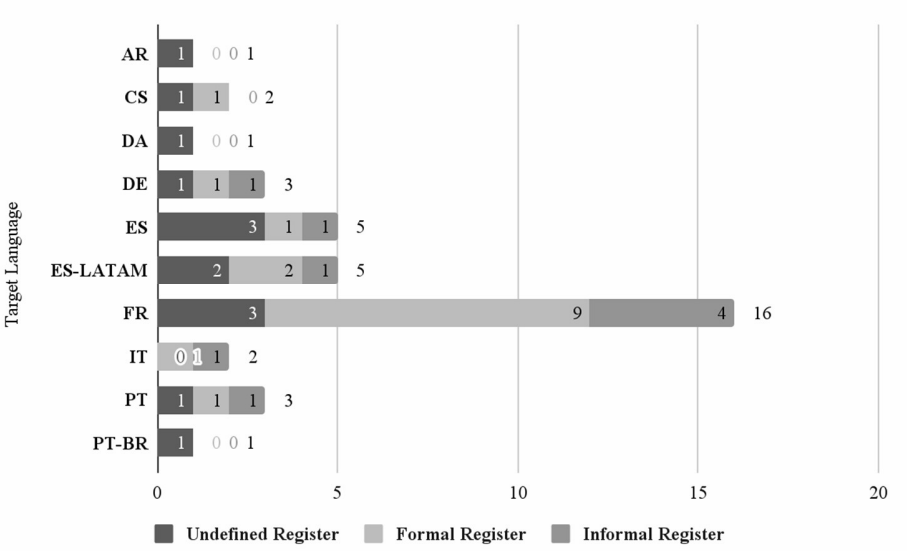
[1] A proprietary programming language that provides a simplified and intuitive interface, enabling linguists to overcome the complexities typically associated with more advanced programming languages.

3.1. **Baseline analysis**

Our research question and motivation stemmed from our initial evaluation, *i.e.* the baseline analysis conducted on custom language rules. This pivotal step involved utilizing quality metrics to assess a total of **39** language-specific rules depicted in Figure 1, all of which were enabled for production by November 5, 2021. This evaluation process prompted us to delve deeper into understanding the effectiveness and impact of these rules, leading us to formulate our research question and pursue further investigation in this area.

Figure 1 provides an overview of the language rules examined during the current baseline analysis. It is important to note that while certain rules were designed for general translations, others were specifically tailored to address formal or informal registers. Furthermore, it is worth highlighting that throughout our analysis, the source language remained consistent as English (EN) across all language pairs examined.

Figure 1. Smartcheck rules: Number of rules per target language



Note. Generic language rules that are not to be applied in any specific register contexts are considered to be in the "Undefined Register" category. Conversely, the "Formal Register" and "Informal Register" categories comprise rules that must only be applied when there is a requirement to use these registers.

The following examples show the difference between each type of rule mentioned above:

(1) **Undefined Register**
Target language: DA; Category: Punctuation
Rule: No punctuation mark must be used after the greeting.

(2) **Formal Register**
Target language: IT; Category: Lexical Register - Formal
Rule: Must use "Cordiali saluti" or "Distinti saluti" in closings.

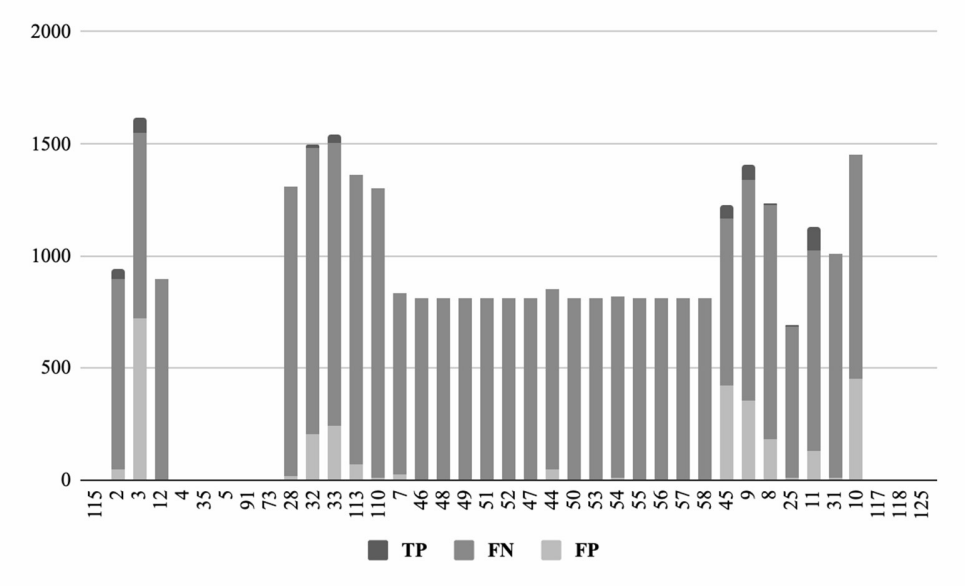
(3) **Informal Register**
Target language: IT; Category: Lexical Register - Informal
Rule: Must use "Ciao", "Saluti" or "Arrivederci" in closings.

It is important to note that our focus in this analysis was on evaluating the individual performance of each rule, rather than considering a set of translated sentences as a whole. Thus, each rule was evaluated with the same set of metrics and the outcome of the evaluation provided us with valuable insights into the frequency with which rules were triggered:

- Cases of True Positives (TPs) – when a rule is correctly triggered due to an existing error in the translation;
- Cases of False Positives (FPs) – when there are no errors in the translation, but a rule is incorrectly triggered nonetheless;
- Cases of False Negatives (FNs) – instances where errors were annotated, but no rule was able to detect and address them.

In order for a rule to achieve high values of accuracy and precision, it is expected that the number of FPs and FNs is minimized in comparison to the number of TPs. However, our analysis revealed a notable disparity, as there were significantly fewer instances of TPs than any other category. This discrepancy is visually depicted in Figure 2, where the y-axis represents the occurrences of rule firings or missed opportunities for firing (corresponding to specific rules identified along the x-axis). This insight highlights the need for further investigation and improvement in order to address the imbalances observed in the performance of the evaluated rules.

Figure 2. Smartcheck rules – TPs, FNs and FPs per rule



The outcomes depicted in Figure 2 provided the necessary data to calculate performance metrics, such as Precision, Recall, and the F1-score, as shown in Figure 3. These metrics are commonly used to assess the effectiveness and overall performance of the rules by offering a holistic assessment of its performance and are calculated in the following manner (Makhoul et al., 1999):

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

On one hand, Precision must be used in order to quantify the accuracy of an automated system such as Smartcheck. It provides insights into how well Smartcheck performs in terms of identifying and flagging actual categories of errors in translations, as opposed to making false identifications. As such, Precision measures the proportion of positive identifications made by Smartcheck that were genuinely correct. A higher Precision score indicates that the system is more accurate in its error detection, as it is capturing a larger proportion of TP errors among the flagged instances. In summary, Precision is concerned with how accurate Smartcheck positive identifications are, emphasizing the avoidance of FP cases.

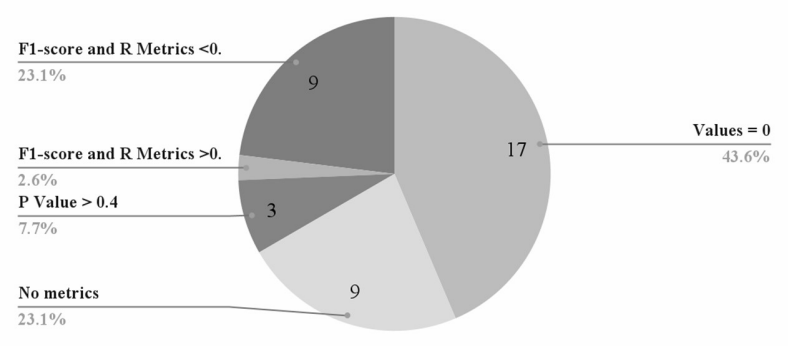
On the other hand, Recall measures the proportion of actual errors within the text that were correctly identified by Smartcheck. It provides insight into how well the system performs in terms of capturing the true errors present in the reference translations. A higher recall score indicates that Smartcheck is more effective at identifying a greater proportion of actual errors within the text, demonstrating its ability to comprehensively detect errors and minimize FP cases. In other words, Recall is concerned with how well Smartcheck captures the actual errors, emphasizing the avoidance of FN cases.

Lastly, the F1-score metric, which is calculated as the harmonic mean between precision (P) and recall (R) as described earlier, represents a balance between these two performance measures.

Regarding Figure 3, it is noteworthy that seventeen rules obtained a score of 0 for all the metrics, including Precision, Recall, and the F1-score. This indicates that these rules did not demonstrate any positive impact or effectiveness in terms of their predictive performance. Additionally, nine other rules experienced a production time out, which prevented the calculation of any metric values. Therefore, the evaluation did not yield any information regarding the performance of these specific rules. Among the evaluated rules, the remaining thirteen rules displayed extremely low values for each metric. None of these rules achieved a score higher than 0.1 for both the F1-score and Recall, except for rule 11, which performed slightly better. Furthermore, only three rules achieved a score higher than 0.4 for Precision.

These results highlight the overall poor performance of the evaluated rules, as evidenced by their consistently low scores across all metrics. The lack of significant positive impact suggests that these rules might not effectively contribute to the desired outcomes or meet the desired standards for overall performance.

Figure 3. Smartcheck rules: Baseline results summary



It is important to note that the observed low results cannot be attributed to the syntax of the rules, as the metrics employed were able to evaluate the rules successfully. The baseline analysis yielded significantly low metric values, indicating a notable performance deficiency. To understand the underlying cause behind Smartcheck rules either failing to flag existing errors in MT outputs or incorrectly flagging correct tokens, a preliminary analysis was conducted to identify potential issues.

The subsequent step involved performing a root-cause analysis to address the results obtained from the previous evaluation. The prior EDF, which serves as the gold standard, should encompass representative examples of the translated content produced by the MT systems and is specifically designed for assessing Error Detection Systems, such as Smartcheck. Smartcheck relies on annotations to provide suggested error corrections to post-editors. As a result, the prior EDF must consist of revised "gold annotations" that require ongoing quality evaluations and frequent updates to ensure their relevance.

To determine the starting point for addressing the issue, instead of individually examining and attempting to enhance the performance of each rule, the focus shifted towards analyzing the data employed for evaluating the rules themselves. Consequently, the evaluation process began with an in-depth analysis of the prior EDF.

3.2. **Root-cause analysis**

The root-cause analysis began with a compilation of data, in which it was required to filter out irrelevant content, and organize the data into distinct files based on specific criteria, namely formal, informal, and generic language-specific data. Due to time constraints only seven language pairs were considered for the analysis.

The data preparation process was conducted in a systematic manner, consisting of two distinct steps. In the first step, the prior EDF was prepared for annotation revision, without any corrections being made at this stage. The second step involved the actual annotation revision process, guided by specific criteria.

The following considerations were taken into account during the data preparation process:

1. Translation step

The focus was exclusively on the MT output, and any post-edited translations were excluded from the analysis. This ensured a clear assessment of the performance of the MT system;

2. Segment duplication
To maintain data integrity and avoid redundant information, duplicated segments were removed from the analysis. This step aimed to streamline the evaluation process and avoid any potential bias resulting from repeated data;

3. Sorting language pair and register
The data was carefully organized based on language pair and register, distinguishing between formal and informal language usage. This allowed for a more targeted evaluation and analysis of the specific linguistic characteristics associated with each language pair and register.

The subsequent step focused on the actual annotation revision process, guided by specific criteria. Annotations were revised according to the MQM-compliant typology for error identification, ensuring a standardized and consistent approach. Additionally, it was crucial to detect incorrect annotations (cases where no errors were present but were erroneously considered as such) as well as missing annotations (instances where errors were present but were not annotated). The severity levels assigned to annotations were also reviewed to determine if they were correctly or incorrectly attributed. Furthermore, the proprietary language guidelines were employed to revise the annotations, aligning them with the desired linguistic standards defined by the language framework.

By following this structured data preparation approach and adhering to the defined criteria, a thorough and reliable assessment of the annotations and their associated linguistic characteristics was achieved and the results are compiled in Table 1. The process ensured data integrity, eliminated duplication, and organized the data based on language pair and register.

Table 1. Prior EDF's analysis: Totals (*Note*: "Ann." stands for "annotations")

Target Language	Total Annotations	Correct Ann.	Incorrect Ann.	Missing Ann.	Correct Severities	Incorrect Severities	Duplicates
DE	1979	1868	111	188	1757	111	312
ES	90	80	10	5	77	3	0
ES-LATAM	1902	1181	721	249	1123	58	189
FR	777	709	68	352	492	217	132
IT	1674	1175	499	244	629	546	458
PT	1303	1183	120	134	553	630	197
PT-BR	930	730	200	326	679	51	534
TOTAL	**8655**	**6926**	**1729**	**1498**	**5310**	**1616**	**1822**

Rather than examining annotations within a dataframe containing duplicated and unverified data, a deliberate choice was made to prioritize the review of more current annotations based on authentic and representative translated content generated by the MT systems. Furthermore, the revision of annotations was deemed imperative, regardless of the selected data. Consequently, it can be deduced that the original dataframe proved inadequate for its intended purpose, highlighting the necessity for an improved evaluation standard.

3.3. **Creation of the New EDF**

A novel and enhanced dataframe was constructed, along with the development of reliable Test Suites. The creation process of the Test Suites followed a similar methodology as presented in Avramidis et al. (2019), Stewart et al. (2022), and Cabeça et al. (2023), encompassing the identification and categorization of errors through

annotations. Contrary to Avramidis et al.'s work, no errors were detected through regular expressions, and there was no requirement for data augmentation. In our work, the errors were meticulously identified through manual annotation, obviating the necessity for regular expressions. The data utilized for this purpose was obtained from our proprietary MT systems, ensuring its inherent representativeness. It is important to emphasize that the new EDF employed in this study does not incorporate gold translations (also known as reference translations), as this aspect falls beyond the scope of our research.

3.3.1. **Data curation**

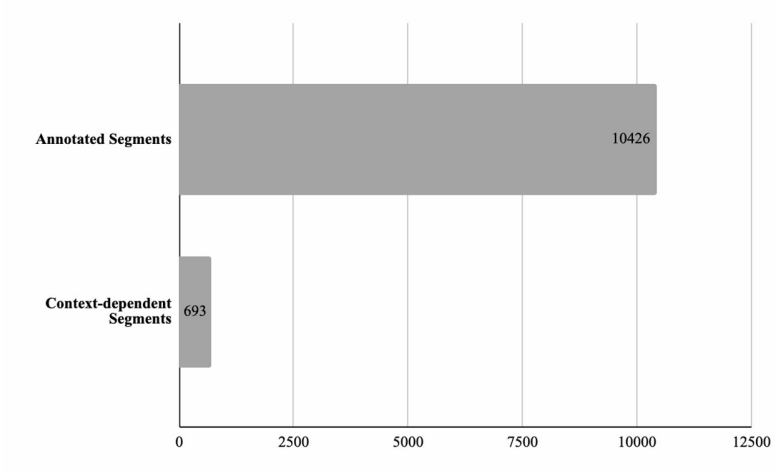
The creation of the new EDF followed a meticulous process, which encompassed the curation of data and the extraction of recently annotated segments. The data utilized for constructing the Test Suites spanned from January 1, 2021, to February 28, 2022, and was meticulously filtered using a methodology akin to the aforementioned filtering step. This data curation step served a practical purpose, as the manual review of annotations necessitates a higher level of effort and data control.

Consistent with all language pairs, duplicates were identified and subsequently removed to ensure the integrity of the dataset. Notably, this data curation step addressed a significant limitation present in the previous dataframe. By scrutinizing the data at a more granular level, it became possible to identify context-dependent annotations that require additional information beyond the individual segment. While the current SURF rules do not explicitly account for context-dependent errors, incorporating this supplementary curation step was imperative. It will enable the new Test Suites to encompass such context-dependent annotations once the SURF rules are updated to accommodate them in future iterations.

As such, during the analysis conducted, three distinct types of context-dependent annotations were identified. The first type involved inconsistencies (translation errors annotated as *Inconsistency*), which are susceptible to nearby segments within the same text. The second type pertained to *Capitalization*, exclusively observed in greetings and closings, dictated by the language specifications for tickets, which are composed in a manner akin to traditional letters. Tickets require distinct guidelines pertaining to capitalization and punctuation, particularly in greetings and closings, highlighting language-specific variations. The third type comprised agreement-related annotations (annotated as *Agreement*), with the referent located outside the segment.

As depicted in Figure 4, among the 10,426 annotations that were thoroughly examined, a mere 6.6% of them, corresponding to a total of 693 segments, were determined to exhibit context-dependent characteristics. This relatively small subset of annotations demonstrated a dependency on surrounding segments or required additional contextual information to accurately evaluate and interpret the linguistic errors present.

Figure 4. New EDF's content: Number of annotated segments and context-dependent segments



3.3.2. **Annotation curation**

The annotation curation process involved a meticulous revision of the previously annotated segments, guided by proprietary language and annotation guidelines. A specific focus was placed on addressing certain types of errors, with priority given to *Orthography*, *Punctuation*, *Register*, and localization challenges such as *Date/time format*, *Currency*, and numeral-related errors. Additionally, particular attention was devoted to selected language pairs, namely DE, ES, ES-LATAM, FR, IT, JA, KO, PT, PT-BR, ZH-CN, and ZH-TW, where linguists and translators carried out more comprehensive reviews. This approach aimed to establish a consensus among the experts involved and ensure consistency throughout the revision process.

The annotation curation step consisted of two distinct assessments, progressively delving into finer details:

- **General Assessment**: This initial phase involved examining the comprehension of the translated text and adhering to language-specific requirements. It entailed scrutinizing aspects related to accuracy, fluency, typography, style, and localization. Errors such as *Additions*, *Omissions*, *Untranslated* content, *MT hallucinations*, *Mistranslations*, *Duplications*, grammar-related issues, and formatting inconsistencies were carefully evaluated within the context of the translated text.
- **Assessment of Customer Support Rules**: This subsequent phase focused specifically on Customer-Service related content. The assessment concentrated on issues pertaining to *Lexical Register*, *Punctuation*, and *Capitalization*. Understanding and implementing appropriate lexical choices, adhering to specific punctuation conventions in greetings and closings, and ensuring correct capitalization based on the sentence structure and part of speech category were key considerations in this evaluation.

3.3.3. **New EDF content**

Through these comprehensive assessments, the new EDF underwent a thorough curation process, resulting in a completed and refined dataset that met the standards set forth for further analysis and evaluation. Thus, the new

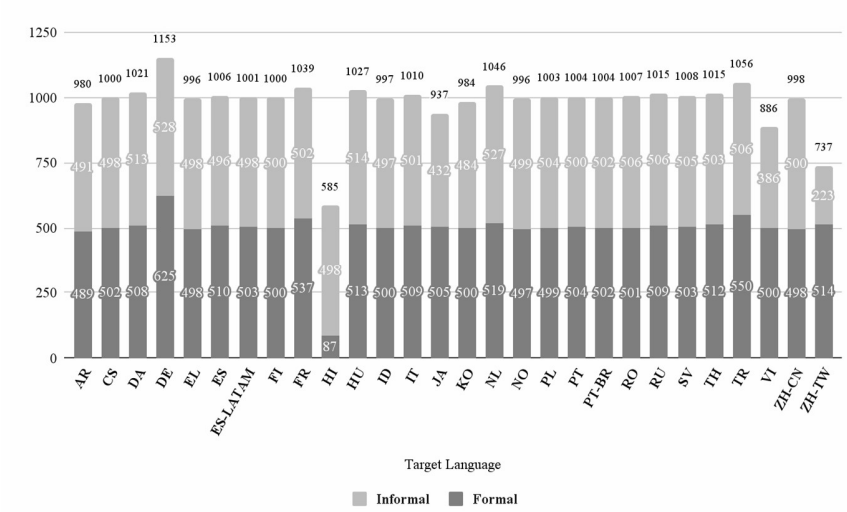
EDF emerged as a comprehensive Test Suite suitable for evaluating Smartcheck rules. The new EDF comprised a total of nearly thirty thousand segments, as depicted in Table 2.

Table 2. New EDF's content: Grand total of formal and informal segments

	Nº of segments
Formal Segments	13894
Informal Segments	13617
TOTAL	27511

These segments encompassed translations for twenty-eight distinct target languages, as illustrated in Figure 5. While it is worth noting that the data for the formal register in HI exhibits a considerably lower number of segments compared to the informal register, the vast majority of languages in the evaluation exhibit a well-balanced distribution of segments. This ensures that the Test Suite consists in a representative sample for most supported languages, enabling a robust evaluation. The focus on maintaining balance in segment counts across languages strengthens the reliability and generalizability of the evaluation results.

Figure 5. EDF's content: Total of formal and informal segments per target language



3.4. **Smartcheck's Performance Optimization**

Each successive stage outlined in this methodology has been dedicated to establishing an appropriate evaluation process to gauge the efficacy of Smartcheck and its error detection rules. Consequently, it is now feasible to conduct a comprehensive assessment of this tool's accuracy and confidently rely on the evaluation results to identify its limitations.

Upon the conclusion of the aforementioned evaluation process, the annotation typology underwent an update to a new version, requiring the adjustment of not only the evaluation standard - the new EDF - but also the rules

themselves. This requirement arises from the fact that Smartcheck relies on annotations to provide suggestions to post-editors. During the revision and updating of the rules, the following was observed: not only were certain rules overly intricate, comprising multiple sub-rules; but additionally, new content-specific rules and rules tailored to specific client requirements needed to be incorporated. In light of this, a thorough analysis was conducted to identify the rules that were most frequently disregarded by the post-editors. Subsequently, the rules that were deemed unhelpful in terms of their limited accuracy in detecting errors were deliberately disabled from Smartcheck.

Consequently, the subsequent step focused on the evaluation of these revised rules, now utilizing the updated EDF, so as to ensure their precise detection of errors.

4. **Results**

The results obtained from the baseline analysis revealed Smartcheck's low performance. The high number of FNs and FPs indicated that Smartcheck was unable to effectively identify the majority of translation issues, and even worse, it flagged problems that did not exist. This led us to formulate a hypothesis that the quality of the data used to evaluate Smartcheck's rules might be a contributing factor to this problem. To address this concern, we made the decision to replace the previous EDF with a new set of gold annotations. The substitution of the EDF with linguist-curated gold annotations represented a notable improvement in the evaluation methodology. The new set of annotations provided a more robust and reliable benchmark for assessing the performance of Smartcheck's rules across all supported language pairs.

4.1. **Rule evaluation comparison between prior EDF and new EDF**

Upon the implementation of the new EDF, it became essential to verify the status of the 39 previously evaluated rules to ensure that they were still enabled in the production environment. This verification was necessary as the baseline analysis had been conducted several months prior to the introduction of the new EDF.

After completing the verification process, Smartcheck was run through numerous MT outputs, and the corresponding metric values were collected and compared. To illustrate this process, let us consider the following example: during the baseline analysis, an issue was identified concerning a punctuation mark in greetings for a certain target language. To automatically detect this language-specific translation problem, Smartcheck referred to Rule A, which specifically addresses punctuation requirements in greetings for that same target language, triggering a warning. It is important to note that Rule A was among the rules evaluated during the baseline analysis.

Subsequently, with the implementation of the new EDF, Smartcheck was once again applied to new MT outputs, and a similar punctuation issue in greetings for the same target language was encountered. Consequently, both sets of MT outputs exhibited the same issues, leading to the activation of the same rule (Rule A) in both instances. If such conditions were met for other existing rules, it would allow for a direct comparison of results between them. In Table 3, we present **7** rules out of the 39 previously evaluated that fulfilled this criterion. Hence, for an unambiguous comparison of rule evaluations, only these seven rules will be considered.

Furthermore, it is important to emphasize that the baseline comparison will focus exclusively on cases of TPs, FNs, and FPs.

Table 3. Rule's comparison between baseline analysis and the new EDF

		Prior EDF			New EDF		
Target Language	**Rule ID**	**TPs**	**FNs**	**FPs**	**TPs**	**FNs**	**FPs**
AR	117		No metrics		4	8	2
CS	125		No metrics		0	40	2
DA	115		No metrics		2	12	1
DE	12	0	0	899	0	21	1
ES-LATAM	28	0	16	1297	3	2	6
	110	3	9	1294	45	8	0
FR	45	59	420	750	123	67	1

Upon comparing the results from both evaluations, two significant inferences were drawn. Firstly, a disparity was observed in terms of FPs between the two evaluations. Secondly, it became possible to evaluate rules that were previously not assessable.

In the baseline analysis, there was a considerable disparity in the number of TPs, FNs, and FPs between the two evaluations. The high number of FP cases in the baseline analysis had a detrimental effect on the system's performance. However, the implementation of the new EDF resulted in a significant reduction in the number of FPs. FN cases also decreased, except for rule 12 in the case of English to German. Additionally, the number of TPs substantially increased for rules 110 and 45. Therefore, it can be concluded that the Smartcheck rules do not possess the low quality that was previously assumed. In other words, solely by modifying the evaluation standard - specifically, the EDF - without altering the rules themselves, the rules used by Smartcheck for error detection exhibit improved performance, albeit not perfect.

On the other hand, for the initial three rules listed in Table 3, no results were available to ascertain their precision, coverage, or overall value for a tool like Smartcheck, as no metrics were obtained. However, due to the introduction of the new EDF, every rule in Smartcheck can now undergo evaluation. This step was crucial in the quality assurance process, as improvements in machine translation systems can only be achieved when existing issues are accurately identified and the necessary adjustments are implemented. Understanding whether a rule fails to trigger when it should or if it incorrectly identifies high-quality translations as problematic is of utmost importance, as it provides guidance for rule revision during the review process.

Nonetheless, drawing conclusions from the evaluation results calls for a certain level of trustworthiness. Since the current EDF comprises non-duplicated data, representative examples that align with the content translated by proprietary MT systems, and annotations reviewed by linguists, it can be inferred that the new evaluation standard and its corresponding results are substantially more reliable and accurate compared to the previous approach.

4.2. **Smartcheck's performance evaluation with new EDF**

The following step in the methodology involved an examination of a correlation pertaining to Smartcheck predictions. Smartcheck's annotations were evaluated in comparison to those of the EDF, while keeping the Smartcheck rules unchanged. Fundamentally, one error detection system would be supplied with the pre-existing data from Smartcheck, while a second error detection system would make use of the reviewed data from the EDF. Subsequently, each system would generate predictions for the same translated segments, identifying tokens deemed erroneous by the system and annotating them accordingly.

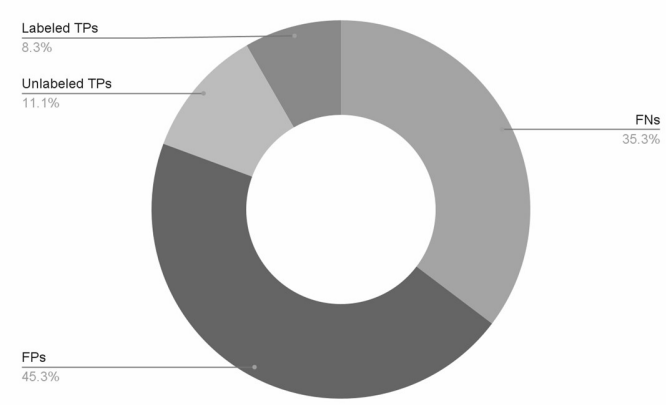
Hence, our objective was to ascertain the disparity between the annotations produced by this grammar-checking tool and a novel, meticulously curated standard. To achieve this, both Smartcheck and the EDF underwent

analysis on various translated segments, and the outcomes were combined and categorized into three distinct sections: performance assessment involving instances of TPs, FPs, and FNs; a comparison of predicted annotations between Smartcheck and the EDF; and a recalculation of metric values associated with Precision, Recall, and the F1-score.

4.3. **Performance assessment**

After running Smartcheck and the EDF on translated segments, the corresponding data was collected. The evaluation focused on cases of FPs, FNs, and TPs. Regarding TPs, they were categorized into two groups: labeled TPs, wherein both the error span and category were correctly identified according to the gold standard, and unlabeled TPs, wherein the error span was correctly identified, yet Smartcheck attributed an error category that did not align with the "gold category." Examining the aggregate count of TPs, FNs, and FPs in Figure 6 allows for the conclusion that Smartcheck considered numerous correct tokens as incorrect, resulting in FP cases. A high prevalence of FPs is detrimental to any error detection system and should be minimized, as it leads to the erroneous perception of statistical evidence that is non-existent. Furthermore, over a third of all annotations were FN cases, indicating that the system overlooked a considerable number of existing translation errors. Nevertheless, FN cases do not have as severe an impact on system performance as FPs, since we have the ability to create new grammar checking rules or review existing ones to reduce the occurrence of FN cases.

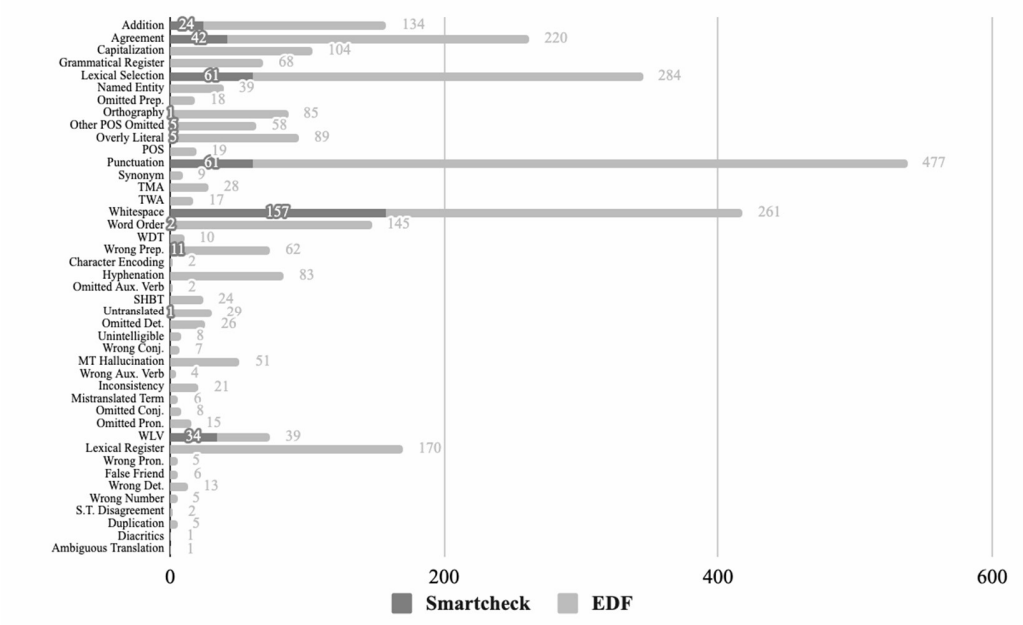
Figure 6. Grand Total of FNs, FPs and TPs when evaluating Smartcheck with the new EDF



4.3.1. **Predicted annotation comparison**

During the comparative analysis, annotations generated by Smartcheck were contrasted with those produced by the EDF using the same dataset from the preceding section. It is important to note that specific target languages and registers were not taken into consideration for the current analysis. Figure 7 presents an overview of the EDF's predicted ideal annotations in comparison to Smartcheck's correctly detected annotations, specifically labeled TP cases. As depicted in Figure 7, Smartcheck failed to achieve the expected objective for all forty-three detected error types in the EDF. The annotations for *Wrong Language Variety*, for instance, came closest to the target. However, for the remaining error types, Smartcheck's annotations significantly deviated from the target count.

Figure 7. Smartcheck labeled TP cases in comparison to gold annotations from the EDF



An analysis of this nature proves highly valuable when evaluating the performance of an error detection system, as it offers a comprehensive understanding of the most problematic error types and allows for the identification of concealed issues that would otherwise go unnoticed.

4.3.2. **Smartcheck evaluation**

The concluding phase of the Smartcheck testing process entailed the computation of Precision, Recall, and F1-score measures, based on the outcomes of the preceding section, specifically the TP, FN, and FP cases. Accordingly, these metrics were individually calculated for each supported target language, with a particular emphasis on discerning the disparity between formal and informal registers.

In order to determine the accuracy of positive identifications made by Smartcheck, the Precision measure was employed. Once computed, the average Precision value for the formal register was slightly higher compared to the average value for the informal register, as evidenced in Table 4.

Table 4. P Average Total per register

Register	Precision Average
Formal	0.3035
Informal	0.3024

To assess the effectiveness of Smartcheck in correctly identifying actual positives based on the new EDF, the Recall measure was employed. The average Recall value for the formal register significantly exceeded that of the informal register, as demonstrated in Table 5.

Table 5. R Average Total per register

Register	Recall Average
Formal	0.2963
Informal	0.2477

The F1-score metric, which represents the harmonic mean of Precision and Recall, was employed to statistically evaluate the performance of the system. As indicated in Table 6, the formal register exhibited a higher average F1-score compared to the informal register.

Table 6. F1-score Average Total per register

Register	F1-score Average
Formal	0.2814
Informal	0.2636

4.4. **Evaluation of spell checkers with new EDF**

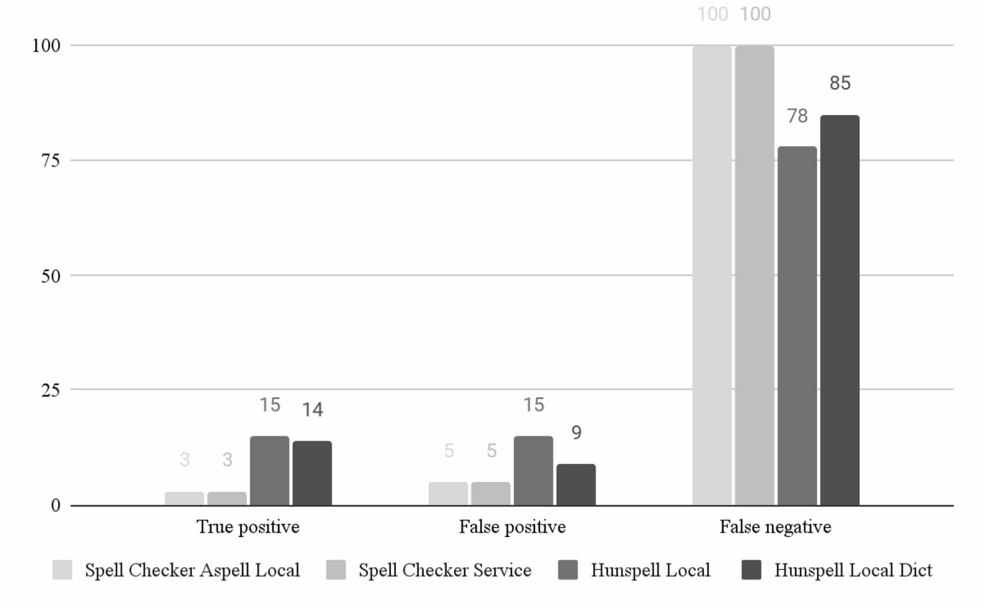
As previously stated, the EDF served as a reliable source of annotated error data in MT outputs, making it the designated gold standard for evaluating error detection systems. In light of this, the EDF's gold data was recently incorporated to assess four different spell checkers used in the production environment with the objective of determining the most suitable candidate for production deployment.

Referring to the data presented in Figure 8, it becomes evident that FN cases exhibited the highest numbers, particularly in relation to the first two spell checkers. This evaluation highlights the significant oversight of existing translation errors (according to the EDF's data) by these spell checkers. Furthermore, the limited number of TP cases indicates a lack of accuracy in error identification. One notable conclusion drawn from the evaluation was that both Hunspell spell checkers outperformed Aspell and the Spell Checker Service. For instance, the Hunspell[2] spell checker exhibited higher TP counts, lower FN totals, and notably higher values for Precision, Recall, and subsequently, F1-score when compared to the others. Based on this assessment, the decision to replace Aspell[3] with

[2] An open source spell checker accessible at http://hunspell.github.io/
[3] An open source spell checker accessible at http://aspell.net/

Hunspell as the spell checker in production was deemed beneficial. Such a decision would not have been possible without the revised and truthful evaluations made possible by the revised EDF.

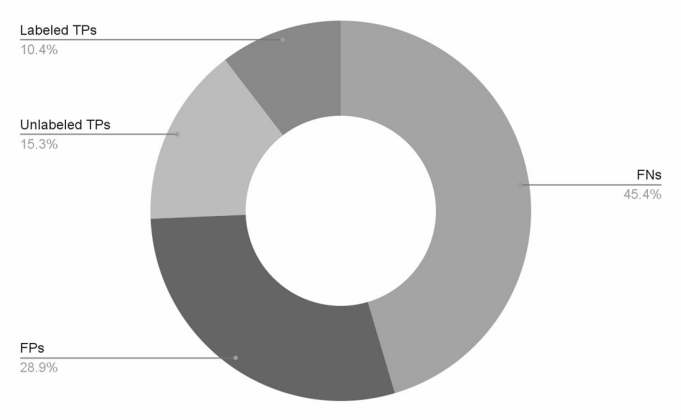
Figure 8. Spell Checkers Evaluation with new EDF as Gold Standard: Cases of TPs, FPs and FNs



4.5. **Smartcheck's Performance Optimization Results**

Upon the requisite modifications implemented in the rules, a subsequent evaluation was conducted to assess their performance, employing the new EDF as the updated evaluation standard once again.

Figure 9. Smartcheck's second evaluation results with the new EDF



Upon careful examination of the final evaluation results, it can be deduced that the outcomes are promising, indicating notable advancements in the functionality of Smartcheck, as depicted in Figure 9. This conclusion is substantiated by the following observations in comparison to the previous evaluation (refer back to Figure 6): Firstly, the total number of TPs has increased from 19.4% to 25.7%, denoting a positive trend in the system's accuracy. Secondly, there has been a significant reduction in the number of FPs, decreasing from 45.3% to 28.9%, representing a notable improvement of 16.4%. Although there has been a rise in the number of FNs from 35.3% to 45.4%, it is important to note that a higher percentage of FNs is preferable over FPs. This preference stems from the fact that it is more desirable for the editor to identify errors independently rather than constantly having to dismiss multiple incorrect suggestions by clicking on the "Ignore" option. Nevertheless, our objective remains to minimize the number of FNs and ultimately eliminate any instances of FPs.

These findings underscore the positive impact of the ongoing efforts to optimize Smartcheck, leading to a more refined and effective tool.

5. **Conclusions**

This paper presents a comprehensive approach to improve the evaluation methodology for error detection systems, with a focus on Smartcheck. We address the limitations of the previous evaluation data set, which failed to capture core errors and led to unreliable results and conclusions. By implementing new Test Suites, we demonstrate the increased trustworthiness of the decision-making process and the ability to evaluate Smartcheck.

Our work contributes to the replicability and visibility of the methodological process involved in creating and curating Test Suites. The main objective was to answer the research question of whether the dataframe used for evaluating Smartcheck rules was suitable for its intended purpose. We propose and implement a robust methodology for creating reliable Test Suites specifically tailored for testing Smartcheck and evaluating custom language rules.

Through this process, we provide valuable insights that contribute to improving translation quality by suggesting edits to post-editors with Smartcheck.

Furthermore, by utilizing the newly developed EDF, we extend the application of the Test Suite beyond its original scope to include the evaluation of spell checkers. This expansion demonstrates the importance of linguistically motivated and scalable Test Suites that can accommodate diverse evaluation objectives.

Additionally, our methodology enables an error-specific evaluation of Smartcheck, striking a good balance between manual and automated metrics. This approach provides a historical perspective on the performance of models, enabling the identification of features that may have been overshadowed by minor improvements when evaluating overall quality. Through these insights, we highlight the significance of small features that can have a substantial impact on the performance and effectiveness of error detection systems.

In conclusion, our research contributes to the advancement of evaluation methodology, ensuring more reliable and valid results for error detection systems. The implementation of our approach demonstrates the importance of carefully curated Test Suites and their role in facilitating informed decision-making processes and improving the overall quality.

References

Avramidis, Eleftherios, Vivien Macketanz, Ursula Strohriegel & Hans Uszkoreit. (2019) Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation* (Vol. 2). Association for Computational Linguistics, pp. 445–454. https://doi.org/10.18653/v1/W19-5351

Avramidis, Eleftherios, Vivien Macketanz, Arle Lommel & Hans Uszkoreit (2018) Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*. Association for Machine Translation in the Americas, pp. 243–248. Available at https://aclanthology.org/W18-2107

Balkan, Lorna (1994). Test Suites: some issues in their use and design. In *Proceedings of the Second International Conference on Machine Translation: Ten years on*. Available at https://aclanthology.org/1994.bcs-1.24

Balkan, Lorna, Doug Arnold & Siety Meije (1994) Test suites for natural language processing. In *Proceedings of Translating and the Computer 16*. Aslib, pp. 51–58. Available at https://aclanthology.org/1994.tc-1.5

Cabeça, Mariana, Marianna Buchicchio, Madalena Gonçalves, Christine Maroti, João Godinho, Pedro Coelho, Helena Moniz & Alon Lavie (2023) Quality fit for purpose: Building business critical errors test suites. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, pp. 451–460. Available at https://aclanthology.org/2023.eamt-1.44

Dale, Robert, Ilya Anisimoff & George Narroway (2012) HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, pp. 54–62. Available at https://aclanthology.org/W12-2006

King, Margaret & Kirsten Falkedal (1990) Using test suites in evaluation of machine translation systems. In *COLING 1990: Papers presented to the 13th International Conference on Computational Linguistics* (Vol. 2). pp. 212–216. Available at https://aclanthology.org/C90-2037

Lavie, Alon, & Michael J. Denkowski (2009) The Meteor metric for automatic evaluation of machine translation. *Machine Translation* 23, pp. 105–115. https://doi.org/10.1007/s10590-009-9059-4

Lommel, Arle, Hans Uszkoreit & Aljoscha Burchardt (2014) Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: Tecnologies de La Traducció* (12), pp. 455–463. https://doi.org/10.5565/rev/tradumatica.77

Makhoul, John, Francis Kubala, Richard Schwartz & Ralph Weischedel (1999) Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*. DARPA, pp. 249–252

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu (2002) BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 311–318. https://doi.org/10.3115/1073083.1073135

Rei, Ricardo, Craig Stewart, Ana C Farinha & Alon Lavie (2020) COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Stewart, Craig, Madalena Gonçalves, Marianna Buchicchio & Alon Lavie (2022) Business critical errors: A framework for adaptive quality feedback. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas* (Vol. 2). Association for Machine Translation in the Americas, pp. 231–256. Available at https://aclanthology.org/2022.amta-upg.17

Algumas notas sobre a quantificação de *Degree Achievements* em Português Europeu

Inês Cantante[1]

[1]Faculdade de Letras da Universidade do Porto/Centro de Linguística da Universidade do Porto

Resumo

A presente investigação aborda a quantificação de *Degree Achievements* (DAs) em Português Europeu (PE). Os DAs têm sido bastante investigados, por se tratarem, na sua maioria, de verbos deadjetivais, que herdam a sua estrutura de adjetivos graduáveis, e que apresentam a particularidade de permitirem uma ambiguidade de leituras relativamente à telicidade do evento que projetam: télicas e atélicas. Por se tratar de verbos que derivam de adjetivos graduáveis, os DAs têm sido considerados, na literatura, predicados graduáveis, o que permite questionar se este tipo de verbo poderá ou não ser quantificado. Dessa forma, o presente trabalho pretende verificar a compatibilidade deste tipo de verbos com os quantificadores *muito* e *pouco* em PE. Tendo em conta a investigação de Quadros Gomes (2011a, 2011b) sobre a modificação de adjetivos graduáveis por *muito* e *bem* em Português do Brasil (PB), pretende-se, neste caso, compreender de que modo a quantificação por *muito* e *pouco* atuará nos DAs, particularmente no que diz respeito à estrutura escalar destes verbos (DAs derivados de escala aberta derivados de escala fechada), bem como à leitura final obtida: télica, atélica ou ambas. Os resultados parecem mostrar que o tipo de escala não influencia a leitura final, já que nenhum dos DAs em análise induz uma leitura télica (de grau máximo, contextualmente definido). De facto, em todos os casos, incluindo os DAs de escala fechada, a leitura obtida é atélica, denotando uma interpretação de processo. Além disso, em PE, uma segunda leitura, relativa à frequência das eventualidades (i.e., a uma repetição do evento denotado pelo verbo), parece estar disponível na maioria dos casos.

Palavras-chave: *Degree achievements*, adjetivos graduáveis, escalas, quantificação verbal, telicidade.

Abstract

The present investigation addresses the quantification of Degree Achievements (DAs) in European Portuguese (EP). DAs have been extensively investigated, as they are mostly deadjectival verbs, which therefore inherit their structure from gradable adjectives. These verbs have the special feature of allowing ambiguous readings regarding the telicity of the event projected by the verb: either telic or atelic. Since they derive from gradable adjectives, DAs have typically been considered, in the literature, gradable predicates, which allows us to question whether or not this type of verb can be quantified. Thus, the present work intends to verify the compatibility of this type of verbs with the EP quantifiers *muito* ('very') and *pouco* ('little'). Taking into consideration the research by Quadros Gomes (2011a, 2011b) on the modification of gradable adjectives by *muito* ('much/very') and *be*m ('well') in Brazilian Portuguese (BP), we aim to understand how these quantifiers will act on DAs particularly in what concerns their basic scalar structure (open-scale and closed-scale DAs) and the final readings available for each case: telic, atelic or both. The results show that the type of scale does not influence the final reading, since none of the analysed DAs induces a telic reading (i.e., the attaining of a maximum degree, contextually defined). In fact, in all the cases, even with closed-scale DAs, the final interpretation is atelic, denoting a process reading. Moreover, in EP, a second reading, concerning the degree of frequency of the eventualities denoted by these verbs (which denotes repetition of the events), seems to be available for most of the analysed examples.

Keywords: Degree achievements, gradable adjectives, scales, verbal quantification, telicity.

1. Introdução

Os Degree Achievements (DAs) são uma classe verbal que tem vindo a ser amplamente estudada na literatura (cf. Abusch, 1986; Dowty, 1979; Hay, 1998; Hay et al., 1999; Kennedy, 2012; Kennedy & Levin, 2008; Kennedy & McNally, 2005; Leal et al., 2015; Rothstein, 2008; e outros). De facto, dos vários estudos sobre esta temática, é possível salientar algumas conclusões comuns, particularmente no que concerne à estrutura dos verbos pertencentes a esta classe, por um lado, e à sua telicidade, por outro. Assim, os vários autores parecem concordar que os DAs deadjetivais derivam de adjetivos graduáveis, dos quais herdam a estrutura escalar. Assim, um DA como *alargar* deverá projetar, à semelhança do adjetivo básico de que deriva (*largo*), uma escala aberta. O mesmo acontece, em sentido oposto, com DAs que derivam de escala fechada, dos quais é exemplo o verbo *encher*. Neste caso, o adjetivo graduável *cheio* só se pode referir a objetos que apresentem a totalidade do seu volume ocupada. No que diz respeito à telicidade destes verbos, uma outra característica que também parece ser consensual é o facto de estes apresentarem a possibilidade de duas interpretações: uma télica e uma atélica. De facto, os DAs aceitam combinar-se tanto com adverbiais temporais do tipo '*em x tempo'* como com adverbiais temporais do tipo '*durante x tempo'*. Note-se que este tipo de adverbiais é frequentemente utilizado para distinguir eventos que apresentam culminação, i.e., télicos, de eventos sem a presença de um fim, i.e., atélicos.

Assim, tendo em conta as considerações tecidas até aqui, e sabendo-se que os DAs têm uma componente graduável, que herdam dos adjetivos de que derivam, a presente investigação tem como objetivo verificar, em primeiro lugar, se este tipo de verbos aceita quantificação, em Português Europeu (PE), e, em segundo, verificar de que forma é que essa quantificação atua, particularmente no que diz respeito à telicidade dos DAs.

Para isso, foram selecionados seis pares de verbos (12 no total), retirados e adaptados da literatura existente, com o objetivo de verificar a sua compatibilidade com os quantificadores *muito* e *pouco*, em PE. O presente artigo organiza-se da seguinte forma: em primeiro lugar, far-se-á uma breve introdução à literatura existente sobre os DAs, apresentando-se, de seguida, alguns conceitos básicos relativos a graduabilidade e escalas, particularmente no que concerne aos adjetivos graduáveis. Segue-se uma reflexão sobre algumas particularidades da quantificação de adjetivos graduáveis, mais propriamente no que diz respeito ao comportamento dos modificadores *muito* e *bem* e às suas especificidades no Português do Brasil (PB) (cf. Quadros Gomes, 2011a, 2011b). Por fim, após a apresentação e discussão dos exemplos em PE, serão apresentadas algumas considerações finais.

2. Degree Achievements (DAs): breve enquadramento teórico

Os DAs são verbos que denotam uma mudança nas entidades sobre as quais predicam. Na verdade, estes verbos, que representam eventos que envolvem uma mudança de estado, projetam um argumento com uma determinada propriedade graduável, cujo grau é obrigatoriamente diferente no início e no fim do evento. Por essa razão, Dowty (1979) analisou-os como verbos de mudança de estado, verificando, no entanto, que esta classe de verbos, em particular, apresentava a particularidade de aceitar adverbiais de duração télicos ('*em x tempo'*) e atélicos ('*durante x tempo'*).

(1) A sopa **arrefeceu** *em dez minutos*.

(2) A sopa **arrefeceu** *durante dez minutos*.

Nos exemplos em apreço, o verbo *arrefecer* projeta uma escala de TEMPERATURA, representando um evento em que a temperatura do argumento, denotado pelo sintagma nominal *A sopa*, varia consoante a progressão do evento. Hay et al. (1999) observam que a diferença entre as duas leituras reside na obtenção ou não de um grau específico da propriedade relevante (interpretado contextualmente relativamente a um standard de comparação) ou de uma leitura em que apenas se atinge um grau diferente (maior ou menor, consoante o

verbo) da mesma propriedade. Desta forma, em (1), a leitura que se obtém é télica, no sentido em que se considera que *a sopa* atingiu um grau, contextualmente determinado, em que se pode afirmar que *a sopa está fria*. No segundo caso, pelo contrário, apenas se diz que o grau de temperatura da sopa é menor, após dez minutos, do que era inicialmente. Tal comportamento levou Abusch (1986) a considerar este tipo de verbos como predicados vagos, que apresentam uma ambiguidade entre duas leituras possíveis, uma do tipo "become adjective" e outra do tipo "become adjective-er" (Abusch, 1986, p. 5). Segundo a autora, esta ambiguidade pode ser resolvida pela existência de um parâmetro de contexto (*context parameter*), que ajuda a compreender se o contexto dado para a comparação é fixo (o que implicaria uma leitura télica) ou não, caso em que a leitura atélica seria possível.[1]

Assim, e apesar de serem tipicamente apelidados de *Degree Achievements* (literalmente, '*culminações de grau'*) (Dowty, 1979), estes verbos apresentam a possibilidade de se comportarem como processos, nas leituras atélicas, e processos culminados, nas leituras télicas. Hay (1998) assume que, de um modo geral, verbos derivados de adjetivos com escala fechada (i.e., com limite máximo ou mínimo) são télicos e verbos derivados de adjetivos com escala aberta (i.e., sem limites) são atélicos. Apesar de se tratar de uma explicação insuficiente para explicar a ambiguidade de leituras dos DAs, esta análise permite considerar os DAs como predicados graduáveis e uma das suas características enquanto tal é que o seu argumento sofre uma mudança numa determinada propriedade (graduável) no decurso do evento.

Ao considerar-se, à semelhança de Hay (1998), Rothstein (2008), Kennedy e McNally (2005) e Kennedy e Levin (2008), que estes verbos herdam a estrutura escalar dos adjetivos de que derivam, torna-se fundamental compreender alguns conceitos relativos à graduabilidade destes adjetivos, o que será feito na secção seguinte.

3. Adjetivos graduáveis: conceitos básicos para a compreensão de predicados graduáveis

Assumindo-se, então, que os DAs herdam as suas propriedades escalares de adjetivos graduáveis, também estes verbos poderão estar associados a dois tipos de escala diferentes: aberta ou fechada.[2] Dessa forma, a determinação da telicidade dos DAs também deverá depender dos adjetivos de grau relevantes, podendo considerar-se que, a escalas fechadas, estarão associadas leituras télicas, estando as leituras atélicas associadas

[1] Os exemplos dados pela autora são os seguintes:

(ia) The Atlantic Ocean is wide and is widening.
(ib) ?The Atlantic Ocean is wide and becoming wide.
(ic) The Atlantic Ocean is wide and becoming wider.
(iia) John is tall and is growing.
(iib) ?John is tall and becoming tall.
(iic) John is tall and becoming taller.

(Abusch, 1986, pp. 5–6)

Nestes casos, a presença de um grau comparativo (exemplos a, e c), que compara, respetivamente, *a largura do Oceano Atlântico* e a *altura do João*, em dois momentos diferentes e não definidos, permite a conceptualização de um contexto que não é tomado como fixo e, por isso, o processo descrito pelo verbo pode continuar a decorrer, não havendo, à partida, um ponto final (*telos*) do evento. No entanto, nos exemplos em b., dados pela autora, as mesmas dimensões de comparação, i.e., *largura do Oceano Atlântico* e *altura do João*, são comparadas relativamente a um standard de comparação que, embora não esteja explicitado, é fixo – note-se que, ao afirmar que *o João é alto*, o falante assume um standard de comparação (que pode corresponder, por exemplo, à altura normal de indivíduos da idade do João), que torna essa frase verdadeira. Consequentemente, a presença deste contexto fixo implica a leitura télica e, por essa razão, impede a continuação do evento (descrita no seguimento da frase).

[2] Um teste utilizado para distinguir adjetivos de escala aberta de adjetivos de escala fechada é a possibilidade (ou não) de modificação pelos *proportional modifiers* (cf. Hay, 1998; Kennedy & McNally, 2005; Leal et al., 2015), i.e., modificadores proporcionais do tipo *completamente*, *parcialmente*, *meio*. De facto, tal modificação está vedada a adjetivos de escala aberta, sendo apenas compatível com adjetivos de escala fechada, já que só este tipo de escala envolve a existência de um conjunto definido e limitado de graus, sobre os quais estes modificadores podem atuar. Veja-se, a título ilustrativo, o exemplo abaixo (retirado de Leal et al., 2015, p. 155):

(i) {completamente/meio} maduro/ vazio – Adjetivos de escala fechada
(ii) ???/*{completamente/meio} saboroso/ mole – Adjetivos de escala aberta

a escalas abertas. No entanto, conforme mostra Kennedy (2012), a introdução de uma medida específica permite 'forçar' leituras télicas em verbos que seriam considerados atélicos:

(3) ?O rio **alargou** *em meia hora*.
(3a) O rio **alargou** 1 metro *em meia hora*.

Em (3), na ausência de um contexto mais abrangente, a frase poderia ser considerada difícil de aceitar, já que o verbo *alargar*, derivado de um adjetivo de escala aberta, parece induzir preferencialmente leituras atélicas, de processo, mais facilmente compatíveis com adverbiais durativos do tipo *durante x tempo*. No entanto, como (3a) permite mostrar, a introdução de uma medida específica, que mede a diferença da largura do rio no início e no fim do evento, marca o ponto em que o evento termina (i.e., o ponto de culminação), e, por isso, a leitura télica passa a ser aceitável.

Hay et al. (1999) notaram, ainda, que também a atelicidade de um evento representado por um verbo tipicamente télico é possível, através da combinação com um adverbial durativo do tipo '*durante x tempo'*, pelo facto de, neste caso, o adverbial funcionar no sentido de cancelar a implicatura de telicidade, dada pelo verbo (Hay et al., 1999, p. 138).[3] Veja-se (4), em que a leitura mais natural é a télica: no fim do evento, o cabelo da Maria estava totalmente seco. Porém, em (4a), a introdução do adverbial temporal induz uma leitura obrigatoriamente atélica e a interpretação é a de que o cabelo da Maria não ficou completamente seco, decorridos dez minutos após o início do evento.

(4) A Maria secou o cabelo.
(4a) A Maria secou o cabelo *durante 10 minutos*.

Ainda assim, nos casos em que a telicidade surge devido à presença de material linguístico que implica uma leitura de medição, a leitura télica não pode ser cancelada pela presença do adverbial, razão pela qual uma frase como (5) não seria aceitável numa leitura atélica.[4]

(5) ?A Maria **alargou** as calças 10cm *durante meia hora*.

Considerando que uma grande parte dos DAs deriva de adjetivos graduáveis (cf. Rothstein, 2008), torna-se fundamental compreender que estes adjetivos representam relações entre objetos e graus, definidos na literatura como "points or intervals partially ordered along some dimension" (Kennedy & McNally, 2005, p. 349). Assim, uma escala configura um conjunto de graus, que representam, cada um, "a different measure of a single gradable property" (Hay, 1998). Uma escala terá sempre uma dimensão de ordenação e uma relação de ordenação, que indica se os valores que configuram os graus estão ordenados de forma crescente ou decrescente

[3] Note-se que este comportamento não se restringe aos DAs, já que, na verdade, também os processos culminados apresentam esta possibilidade. Notem-se os exemplos:
(i) A Maria leu o livro *em meia hora*.
(ii) A Maria leu o livro *durante meia hora*.
Se, no primeiro caso, a leitura obtida é télica, i.e., a Maria acabou de ler o livro, e demorou meia hora a lê-lo do início ao fim, no segundo, a situação é diferente, já que a presença do adverbial *durante meia hora* apenas indica que a Maria esteve envolvida na atividade de ler o livro durante 30 minutos, não havendo indicação de que o livro foi totalmente lido.

[4] Deve notar-se que esta interpretação poderá, eventualmente, ser possível, se for criado um contexto adequado (como, por exemplo, a existência de um concurso de 'alargamento de calças', em que, para um dado intervalo de tempo – do tipo *durante meia hora* –, o concorrente que provoca um maior alargamento nas calças é o vencedor). Note-se, no entanto, que tal contexto levaria à consideração de um intervalo temporalmente delimitado (que corresponderia à duração de uma prova) que obrigaria a uma leitura télica. Além disso, um contexto como este imporia fortes restrições à configuração do Mundo para que tal estado de coisas se pudesse verificar. Assim, apesar de ser teoricamente possível, tal não constitui uma interpretação natural para este tipo de construção, que continua, assim, a ser tida como dificilmente aceitável numa leitura atélica.

(e que reflete, por exemplo, a diferença entre *aquecer* e *arrefecer*). Desta forma, o mesmo objeto pode apresentar, ao longo do evento, diferentes graus de uma determinada propriedade, que, por sua vez, é representada ao longo de uma dimensão específica (temperatura, comprimento, altura, peso, etc.), projetada pelo próprio verbo.

Para Kennedy e Levin (2008), um DA representa uma função de medição que mede, portanto, a diferença no grau de uma dada propriedade graduável, determinada pela base adjetival, que um objeto apresenta no início e no fim do evento. Para estes autores, esta mudança é representada por um valor diferencial ('difference value'), que é um elemento fundamental para a determinação da telicidade do evento: se este valor for exato (i.e., definido) e delimitado, o evento será télico, caso contrário, será atélico (Kennedy & Levin, 2008, p. 8). Os mesmos autores definem o valor diferencial como a "measure of the amount that an object changes as a result of participating in the event described by a DA." (Kennedy & Levin, 2008, p. 163). Em certos casos, a especificação deste valor pode ser dada através de expressões, como sintagmas que representam medidas (*10 cm, 3km*, etc.) ou modificadores de grau (*completamente*).

Estes valores são sempre interpretados de acordo com um **standard de comparação**, que representa "the minimum degree required to 'stand out' in the context relative to the kind of measurement expressed by the adjective" (Kennedy & Levin, 2008, p. 163). Normalmente, o valor do standard de comparação é fixo (no limite máximo ou no limite mínimo de uma escala) e independente do contexto, se o adjetivo tiver escala fechada, ou contextualmente dependente, no caso de o adjetivo projetar uma escala aberta. Estes autores servem-se das características do standard de comparação para explicar a ambiguidade entre as leituras télica e atélica, já que, para que ocorra uma leitura atélica, apenas é necessário que haja um grau mínimo de mudança no objeto que participa no evento, i.e., o valor diferencial pode ser mínimo (cf. (6), em que, para que a frase seja verdadeira, basta que *a rua* tenha alargado 0,0000001 mm). Pelo contrário, para que a leitura seja télica, é necessário que a mudança tenha um valor fixo, definido como um ponto máximo ou mínimo da escala (consoante o verbo projete uma escala que envolva um aumento ou a diminuição do grau), como em (7), em que a frase apenas é verdadeira se a roupa tiver 0% de humidade, ou seja, se estiver completamente seca.[5]

(6) A rua **alargou**.

(7) A roupa **secou**.

Assim, sintetizando, os DAs são tipicamente analisados como predicados graduáveis, que representam funções de medição que avaliam o grau da mudança num objeto com uma determinada propriedade (graduável), como consequência da sua participação no evento. Essenciais para a determinação da telicidade dos DAs são, ainda, a telicidade do valor diferencial e do standard de comparação, projetados pelo próprio verbo.

3.1. **Muito e bem como intensificadores de adjetivos graduáveis: algumas notas**

À semelhança de Kennedy e McNally (2005), que investigaram a distribuição complementar dos modificadores de grau *much*, *very* e *well*, para o Inglês, Quadros Gomes (2011b) também investigou, para o Português do Brasil (PB), a distinção entre os modificadores *muito* e *bem*. Em relação ao Inglês, os autores propõem que os modificadores em análise impõem uma restrição quanto ao tipo de escala que modificam: *very* apenas pode modificar adjetivos que apresentem escala aberta, *much* modifica adjetivos que apresentem escala fechada na ponta inferior e, inversamente, *well* modifica adjetivos cuja escala é fechada na ponta superior. Todavia, Quadros Gomes (2011b) mostra que, em PB, tal restrição não ocorre. Assumindo que estes

[5] Note-se que, embora se trate de um adjetivo de escala fechada, existe a possibilidade de o utilizar em frases como "A camisa secou, mas não completamente: a gola e os punhos ainda estão húmidos". Neste caso, no entanto, o que se refere, na verdade, é que existem partes da camisa que ainda não estão secas. Tal comportamento demonstra que é necessário fazer a distinção entre uma leitura escalar (i.e., de obtenção de um grau máximo numa escala) de uma leitura mereológica, que diz respeito a partes do objeto e, consequentemente, à possibilidade de que algumas dessas partes não tenham (ainda) atingido o ponto máximo da escala, no momento da enunciação.

modificadores se comportam como intensificadores, no sentido em que "'aumentam' o grau da propriedade exibida pelo argumento do AG [adjetivo de grau]" (Quadros Gomes, 2011b, p. 3), a autora mostra que, tanto *bem*, como *muito* podem operar sobre adjetivos de grau de escala aberta ou fechada, não apresentando qualquer restrição de seleção. No entanto, estes modificadores impõem restrições quanto ao resultado da modificação que provocam: quando aplicado a adjetivos graduáveis, *muito* produz sempre um complexo que tem escala aberta e parâmetro relativo, tal significando que o standard de comparação é fornecido pelo contexto. Em (8), a consideração do VOLUME DE OCUPAÇÃO da chávena como elevado pode variar de falante para falante, o mesmo acontecendo em (8a), caso em que a frase pode ser verdadeira para um falante e falsa para outro. Em nenhum dos casos, no entanto, o complexo descrito pelo conjunto [Modificador + Adjetivo de grau] denota a sobreposição com o grau máximo da escala de VOLUME DE OCUPAÇÃO (cf. (8) e TAMANHO (cf. (8a)).

(8) A chávena está **muito** cheia.
(8a) O sapato é **muito** grande.

Pelo contrário, *bem* dá origem a um complexo de escala fechada e parâmetro absoluto (independente do contexto): em (9), a chávena tem de pertencer ao domínio das *chávenas cheias*, para que a frase seja verdadeira; da mesma forma, em (9a), a comparação da altura da Alice pode ser feita, por exemplo, relativamente a raparigas da sua idade, exigindo também que haja uma sobreposição entre a altura da Alice e a parte da escala em que estão situados os *indivíduos altos* da idade da Alice (cf. (9b)).

(9) A chávena está **bem** cheia.
(9a) A Alice é **bem** grande.
(9b) A Alice é bem grande para uma menina da sua idade.

Assim, e conforme afirma a autora, "*muito* não conserva intactas as propriedades do adjetivo que modifica" (Quadros Gomes, 2011b, p. 4), já que, independentemente de modificar um adjetivo de escala aberta ou fechada, este intensificador parece promover sempre uma leitura de escala aberta. Além disso, a autora mostra, também, que o standard de comparação para a avaliação da intensificação por *muito* depende do contexto. Assim, a sua ação enquanto intensificador força uma interpretação de escala aberta do complexo intensificado, independentemente das propriedades escalares do adjetivo básico.

Considerando, à semelhança de Kennedy e McNally (2005), que a propriedade da graduabilidade não é exclusiva de adjetivos, podendo, igualmente, aplicar-se a nomes e verbos, torna-se, então, possível refletir sobre a forma como estes mesmos modificadores atuarão, quando aplicados a verbos como os DAs, que derivam, tipicamente, de adjetivos graduáveis (como, por exemplo, *encher* [*'cheio'*] ou *aquecer* [*'quente'*]). Dessa forma, na próxima secção, serão analisados exemplos de frases com DAs de escala aberta e de escala fechada modificados por *muito* e *pouco*, com o objetivo de compreender de que forma é que estes quantificadores atuam no domínio verbal, particularmente com verbos que têm uma componente escalar (associada ao seu adjetivo básico).

4. **Os dados do PE: quantificação por muito e pouco de DAs deadjetivais**

Para a presente investigação, a seleção dos DAs a analisar foi feita tendo em consideração a relação entre estes verbos e os adjetivos graduáveis a que estão associados, particularmente no que diz respeito às características da escala que partilham: tanto o adjetivo *largo,* como o verbo *alargar* estão associados a uma mesma escala de LARGURA, da mesma forma que o verbo *encher* utiliza a mesma escala que o adjetivo *cheio*, relativa ao VOLUME DE OCUPAÇÃO. Assim, tendo como base a literatura existente, foram selecionados seis pares de verbos, três associados a adjetivos graduáveis com escala aberta e três com escala fechada, num total de doze verbos.

Assim, em análise estão os DAs de escala aberta *alargar*/*estreitar*, *escurecer*/*clarear* e *aquecer*/*arrefecer* e os DAs de escala fechada *abrir*/*fechar*, *encher*/*esvaziar* e *molhar*/*secar*. Os exemplos foram recolhidos e adaptados dos diversos estudos existentes sobre DAs, tanto em Inglês (cf. Kennedy & Levin, 2008; Kennedy & McNally, 2005; McNally, 2017; Rothstein, 2008), como em PE (Leal et al., 2015).

Em relação aos DAs derivados de escala aberta, os verbos *alargar* e *estreitar*, quando modificados por *muito*, parecem induzir uma leitura atélica, i.e., sabe-se que a largura da *rua* no fim do evento representado por *alargar* (cf. (10)) é superior à largura da mesma no início do evento e, no caso de *estreitar* (cf. (11)), a mesma largura (da *estrada*) será menor (pelo facto de o verbo projetar, neste caso, uma escala no sentido decrescente). O mesmo acontece com *pouco*, embora no sentido inverso (cf. (10b) – (11b)). Em nenhum dos casos, porém, a leitura obtida é télica, i.e., não é possível assumir que a *rua*/*estrada* atingiu um grau contextualmente relevante para que possa ser considerada *larga* ou *estreita*.

(10) A rua alargou **muito**.
(10a) A rua alargou **pouco**.

(11) A estrada estreitou **muito**.
(11a) A estrada estreitou **pouco**.

Já em relação ao par *escurecer*/*clarear*, também de escala aberta, os exemplos abaixo mostram que, além das leituras mencionadas para o par anterior, estes verbos apresentam uma possibilidade adicional de interpretação, que se relaciona com a frequência das ocorrências de *escurecer*/*clarear*. Assim, em (13), por exemplo, é possível colocar-se um cenário em que *a sala*, depois de estar clara, tornou a escurecer e voltou a clarear, e assim sucessivamente, num número de ocorrências indefinido, mas que terá, obrigatoriamente, de ser considerado elevado, na presença de *muito*, ou reduzido, na presença de *pouco*. Note-se que a introdução de um adverbial de duração adequado,[6] que privilegie a leitura de frequência, torna os exemplos mais naturais (como o caso dos exemplos em b).[7]

(12) O céu escureceu **muito**.
(12a) O céu escureceu **pouco**.
(12b) *Na última semana*, o céu escureceu **muito**/**pouco**.

(13) A sala clareou **muito**.
(13a) A sala clareou **pouco**.
(13b) *Durante o filme*, a sala clareou **muito**/**pouco**.

[6] Note-se que esta afirmação é válida, também, para o primeiro par de verbos (*alargar*/*estreitar*), se houver um elemento que privilegie a leitura de frequência. Veja-se, a este propósito, um exemplo sugerido por um avaliador anónimo:

(i) Durante as obras dos últimos vinte anos, a avenida *alargou/estreitou* **muito/pouco**.

Se, em (i), a leitura mais natural parece ser a de frequência (em que se assume que *a avenida* sofreu várias alterações ao longo de um período de vinte anos, assumindo-se, simultaneamente, que, em nenhuma dessas alterações, a sua largura se alterou significativamente), tal só acontece devido à presença de um adverbial temporal que indica uma duração longa (*vinte anos*) e que, consequentemente, permite a existência de vários eventos (realização de *obras*) que envolvem a alteração da largura da avenida. De facto, na ausência desta localização temporal, mantém-se a interpretação de mudança de grau verificada anteriormente.

[7] Veja-se, ainda, que esta possibilidade não é restrita à classe dos DAs, podendo aplicar-se a outro tipo de verbos, que não são derivados de adjetivos de grau:

(i) Na semana passada, *choveu* **muito**.
(ii) Hoje, a Maria *caiu* **pouco**.

Tal comportamento demonstra que *muito*/*pouco* podem funcionar como operadores de modificação aspetual, medindo a quantidade de ocorrências dos eventos considerados (*muitas vezes*/*poucas vezes*), sempre de forma imprecisa, registando-a como superior ou inferior ao esperado.

Por fim, a modificação dos verbos de escala aberta *aquecer* e *arrefecer* (cf. (14), (15) – (16)) apresenta, também, uma interpretação relativa à obtenção de um grau de temperatura *superior* (com *muito*) ou *inferior* (com *pouco*) no fim do evento, relativamente ao grau inicial. Além disso, também a leitura de frequência é possível, em que *muito*/*pouco* projetam uma pluralidade eventos de *aquecer*/*arrefecer*, embora esta leitura pareça mais natural em (15), em que o SN *o João*, com função de sujeito, é o responsável pelo evento *aquecer o leite*.[8] Nesse caso, torna-se mais fácil compreender que *o João* pode ter aquecido *o leite* mais do que uma vez, do que, por exemplo, em (14) ou (16), em que a interpretação mais natural parece ser a de mudança de grau, i.e., a de que *a sopa*/*o leite aqueceu*/*arrefeceu* num elevado grau. Dito por outras palavras, a temperatura é diferente (maior ou menor, consoante o verbo e o modificador) no final do evento, comparativamente à sua temperatura no início do mesmo. Mesmo assim, em nenhum dos exemplos é possível uma interpretação télica de que se atingiu uma temperatura contextualmente considerada como *quente* ou *fria*.

(14) O leite *aqueceu* **muito**.
(14a) O leite *aqueceu* **pouco**.

(15) O João *aqueceu* **muito** o leite.
(15a) O João *aqueceu* **pouco** o leite.

(16) A sopa *arrefeceu* **muito**.
(16a) A sopa *arrefeceu* **pouco**.

Passando agora à análise dos DAs derivados de adjetivos graduáveis de escala fechada, é possível observar que as leituras se mantêm. De facto, com *abrir*/*fechar*, existem duas possibilidades de interpretação: (17) pode significar que o grau de abertura *da porta* é maior no final do evento do que no início, e, além disso, que a diferença entre o grau de abertura nos dois momentos é acentuada, pela utilização de *muito* (que marca um grau obrigatoriamente elevado, ainda que impreciso), ou reduzida, no caso de *pouco*. O mesmo acontece, embora em sentido inverso, no caso de (18). No entanto, como segunda possibilidade de interpretação, é possível considerar que *a porta* foi aberta (ou fechada) mais do que uma vez. Neste caso, parece lógico assumir que, entre cada evento de abrir a porta, houve um momento em que a porta voltou a estar fechada. Esta leitura é mais natural em (19), em que o sujeito é *A Maria*, com função agentiva, ou (20), em que *o vento* funciona como causador. Note-se que, mesmo neste caso, a possibilidade de leitura de mudança de grau mantém-se possível (parafraseável por *A Maria fechou a porta num elevado grau*). Esta leitura torna-se mais visível em (19b) e (20b), em que a continuação da frase evidencia que a porta não foi, efetivamente, aberta/fechada por completo.

(17) A porta abriu **muito**.
(17a) A porta abriu **pouco**.

(18) A porta fechou **muito**.
(18a) A porta fechou **pouco**.

(19) A Maria abriu **muito** a porta.
(19a) A Maria abriu **pouco** a porta.
(19b) A Maria abriu **muito** a porta. Na verdade, deixou-a escancarada.

[8] Neste caso em particular, os verbos em análise apresentam um comportamento próximo dos verbos de alternância causativa. Estes verbos têm a particularidade de apresentarem duas 'variantes': uma causativa, que é transitiva, e uma não causativa (inacusativa), em que "o argumento interno direto ocorre como sujeito" (Duarte, 2003, p. 306). Vejam-se os exemplos dados por Duarte (2003, p. 305) para ilustrar esta alternância:
(i) O calor derreteu o gelado.
(ii) O gelado derreteu(-se).

(20) O vento fechou **muito** a porta.
(20a) O vento fechou **pouco** a porta.
(20b) O vento fechou **muito** a porta. Na verdade, deixou-a quase fechada.

Também com *encher*/*esvaziar* (cf. (21) – (22)), as interpretações obtidas são do mesmo tipo, uma que avalia o grau da mudança sofrida pelo argumento verbal, e outra, que avalia a frequência da ocorrência do evento descrito pelo verbo. Apesar de (21a) não parecer tão natural como (21), a explicitação de um contexto apropriado para a interpretação torna a frase aceitável.[9] Em ambos os casos, as leituras são vagas: *muito* e *pouco* medem uma mudança de grau ou uma pluralidade de eventos, que tem de ser acentuada no caso de *muito*, e reduzida no caso de *pouco*, mas que nunca tem um valor ou uma cardinalidade precisamente definida.

(21) A piscina encheu **muito**.
(21a) A piscina encheu **pouco** (apesar de ter chovido muito).

(22) O João esvaziou **muito** a piscina.
(22a) O João esvaziou **pouco** a piscina.

Note-se, ainda, a relevância do complemento no momento da interpretação da frase. Em (22), embora ambas as leituras sejam possíveis, a preferencial parece ser a que envolve uma mudança de grau: *a piscina* tinha, no início do evento, um determinado volume ocupado por uma substância líquida (admita-se que se trata de *água*); ao *esvaziar muito a piscina*, assume-se que *o João* retirou um elevado volume de água da mesma e, consequentemente, *a piscina* encontra-se (muito) menos cheia (embora não se possa assumir que esteja vazia). Em (23), abaixo, também existe a possibilidade de duas interpretações: ou *o João encheu o lava-loiça* até que este ficasse perto da sua capacidade volumétrica total, ou *o João encheu várias vezes o lava-loiça*. No entanto, olhando para (24), a leitura preferencial parece ser a de que *o João esvaziou o lava-loiça várias vezes*; neste caso, o complemento, em interação com o nosso conhecimento do mundo, tem influência na interpretação, já que a capacidade volumétrica de um lava-loiça é bastante inferior à de uma piscina, o que leva a que a ação de o esvaziar seja muito menos demorada, tornando-se mais fácil assumir que possa ocorrer diversas vezes, mesmo num curto intervalo de tempo (por exemplo, uma manhã).

(23) O João encheu **muito** o lava-loiça.

(24) O João esvaziou **muito** o lava-loiça.

Em relação a *molhar*/*secar*, o que está, à partida, em avaliação é uma escala de humidade. Nos exemplos (25) – (26), mantêm-se as duas possibilidades de interpretação: mudança de grau e frequência. No primeiro caso, avalia-se a mudança no grau de humidade do argumento (*a roupa*) entre o início e o fim do evento, por ação da *chuva* e, no segundo, a interpretação é parafraseável por *A chuva molhou a roupa muitas/poucas vezes*. Note-se, no entanto, (26), em que a presença de *ao ar livre*, que define a forma como a roupa foi seca, privilegia a leitura de frequência.

(25) A chuva molhou **muito** a roupa.
(25a) A chuva molhou **pouco** a roupa.

(26) A camisa secou **muito** (ao ar livre).

[9] Note-se que a substituição de *a piscina* por *o estádio* (cf. (i), abaixo, em que se avalia o VOLUME DE OCUPAÇÃO do estádio), tornaria a frase bastante mais natural, o que parece evidenciar que esta leitura é, de facto, possível.

(i) O estádio encheu **pouco**.

(26a) A camisa secou **pouco** (ao ar livre).

Assim, de um modo geral, o que os exemplos em análise parecem mostrar é que, no caso de DAs baseados em adjetivos graduáveis de escala fechada, os quantificadores *muito* e *pouco* parecem atuar no sentido de modificar o grau da propriedade do argumento, de acordo com a dimensão escalar adequada (por exemplo, LARGURA ou TEMPERATURA), como resultado da participação no evento descrito pelo verbo. *Muito* marca uma elevada diferença de grau, o que, de acordo com Kennedy e Levin (2008), corresponderia a um valor diferencial acentuado, ao contrário de *pouco*, que evidenciaria um valor diferencial reduzido entre o início e o fim do evento. No entanto, e apesar de representar um aumento ou diminuição do grau da propriedade relevante, em nenhum dos casos, a leitura obtida é télica, i.e., não pode assumir-se que se atingiu um grau contextualmente relevante da propriedade em causa. Assim, afirmar, sem mais, que *A sopa arrefeceu muito* nunca é equivalente a afirmar que *A sopa ficou fria* (sendo o standard de comparação para *fria* fornecido pelo contexto). Do mesmo modo, afirmar que *A sopa arrefeceu pouco* também não significa que a sopa ficou fria, nem que a sopa continuou quente. Na verdade, tendo em consideração a leitura de grau, *arrefecer pouco* apenas indica que, na representação escalar do evento, o afastamento entre os graus de temperatura no início e no fim do evento é reduzido. Já em relação aos DAs baseados em adjetivos graduáveis de escala aberta, os exemplos analisados mostram que estes se comportam da mesma maneira, i.e., a modificação por *muito*/*pouco* promove leituras atélicas, que envolvem a alteração do grau de uma propriedade, relativa a uma dimensão relevante, projetada pelo verbo, presente no argumento verbal.

Ao investigar a distinção entre o comportamento de *muito* e *bem*, enquanto intensificadores do PB, Quadros Gomes (2011a, 2011b) propõe que *muito*, ao modificar adjetivos graduáveis, forma um complexo que tem sempre escala aberta e parâmetro (i.e., standard de comparação) relativo, dependente do contexto, independentemente de o adjetivo graduável que modifica projetar uma escala aberta ou fechada. O que os dados da presente investigação parecem evidenciar é que os DAs, verbos tipicamente deadjetivais, que herdam a sua estrutura escalar dos adjetivos de que derivam, apresentam um comportamento semelhante. Assim, da mesma forma que o tipo de escala do adjetivo graduável não impõe qualquer restrição à modificação por *muito*, também o tipo de escala associado aos diversos DAs analisados não representa um obstáculo à modificação. Na verdade, o produto da modificação de DAs por *muito* e *pouco* é semelhante ao produto da modificação de adjetivos graduáveis por *muito*: uma leitura atélica, i.e., sem limite (escalar). Em nenhum dos casos, independentemente de se tratar de verbos derivados de adjetivos de escala aberta ou fechada, a leitura obtida após a intensificação é a de que se atingiu um determinado grau máximo da propriedade (contextualmente definido).

Além disso, os exemplos em PE parecem, ainda, pôr em evidência uma segunda possibilidade de interpretação: os quantificadores *muito* e *pouco* não têm apenas a função de intensificação, mencionada acima, podendo também promover uma leitura de frequência, i.e., a iteração de situações, num número indefinido, que terá de ser elevado, para *muito*, e reduzido, para *pouco*. Esta leitura de frequência parece ser possível porque introduz uma culminação, que se repete em todos os casos. Por outras palavras, *encher várias vezes o lava-loiça* implica que, entre cada evento de *encher*, tenha de haver, em primeiro lugar, uma culminação (i.e., *o lava-loiça* tem de *ficar cheio*) e, de igual modo, entre os vários eventos de *encher*, que se repetem, pelo uso de *muito*/*pouco*, tenha de existir um igual número de eventos de *esvaziar*.

5. **Considerações finais**

A presente investigação, ao abordar a possibilidade de quantificação no domínio verbal, focou, apenas, DAs, por serem verbos tradicionalmente associados ao domínio adjetival. Por essa razão, e por ser vasta a literatura que investiga a graduabilidade de adjetivos, esta classe verbal parece permitir uma associação evidente ao domínio da quantificação. Note-se que os DAs são verbos cujos argumentos representam entidades que sofrem uma mudança ao longo do evento, podendo essa mudança ser quantificada.

Recorrendo aos estudos existentes sobre adjetivos graduáveis, a presente investigação põe em evidência duas possibilidades de leitura para os DAs quantificados em PE: uma, em que se avalia o grau de mudança sofrido, ao longo do evento, pela entidade representada pelo argumento verbal, e outra, em que se mede o número de ocorrências do evento descrito pelo verbo. Por quantificarem de forma vaga e imprecisa, os quantificadores em análise não permitem medir a diferença exata no grau da propriedade do argumento no início e fim do evento, nem a cardinalidade precisa do número de ocorrências dos eventos descritos.

A presente investigação pode ser aprofundada, futuramente, de forma a avaliar a compatibilidade destes verbos com outros quantificadores, como o caso de *bastante* e *imenso*, bem como com *proportional modifiers*, do tipo *completamente* ou *parcialmente*. Se, no primeiro caso, a previsão é de que as diferenças comportamentais relativamente a *muito* e *pouco* não sejam significativas (embora tal hipótese tenha de ser verificada), no segundo, a aceitabilidade de frases como *o João aqueceu completamente/parcialmente a sopa* parece ser mais difícil ou, pelo menos, induzir uma leitura diferente das verificadas no presente estudo. Além disso, será também relevante verificar de que forma a quantificação verbal atua com verbos que não sejam inerentemente graduáveis, como o são os DAs.

Agradecimentos

Este trabalho foi financiado pela FCT – Fundação para a Ciência e Tecnologia (Portugal), através da Bolsa de Doutoramento com a referência 2021.04998.BD, e pelo Centro de Linguística da Universidade do Porto (CLUP) (FCT-UIDB /00022/2020).

Deixo, ainda, o meu agradecimento aos avaliadores anónimos, cujo contributo foi fundamental para a melhoria do presente trabalho.

Referências

Abusch, Dorit (1986) *Verbs of change, causation and time*. Center for the Study of Language and Information.

Dowty, David (1979) *Word meaning and Montague grammar. The semantics of verbs and times in generative semantics and in Montague's PTQ*. Reidel.

Duarte, Inês (2003) Relações gramaticais, esquemas relacionais e ordens de palavras. In Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte, Isabel Hub Faria, Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário & Alina Villalva (orgs.), *Gramática da língua portuguesa*. Caminho, pp. 275–321.

Hay, Jennifer (1998) *The non-uniformity of degree achievements* [Apresentação de artigo]. 72th Annual Meeting of the LSA, New York.

Hay, Jennifer, Christopher Kennedy & Beth Levin (1999) Scalar structure underlies telicity in "degree achievements". In *Proceedings of SALT 9*. CLC Publications, pp. 127–144.

Kennedy, Christopher (2012) The composition of incremental change. In Violeta Demonte & Louise McNally (orgs.), *Telicity, change, and state: A cross-categorical view of event-structure* (Part 1). Oxford University Press, pp. 103–121.

Kennedy, Christopher & Louise McNally (2005) Scale structure and the semantic typology of gradable predicates. *Language* 81, pp. 345–381.

Kennedy, Christopher & Beth Levin (2008) Measure of change: The adjectival core of degree achievements. In Louise McNally & Christopher Kennedy (eds.), *Adjectives and adverbs: Syntax, semantics and discourse*. Oxford University Press, pp.156–182.

Leal, António, Luís Filipe Cunha & Idalina Ferreira (2015) Algumas reflexões sobre escalaridade e degree achievements em Português Europeu. In António Leal & Purificação Silvano (orgs.), *Estudos de semântica*. Centro de Linguística da Universidade do Porto/Faculdade de Letras da Universidade do Porto, pp. 153–160.

Quadros Gomes, Ana Paula (2011a) Uma proposta de distinção semântica para os intensificadores 'muito' e 'bem'. *Revista de Estudos Linguísticos* 40 (1), pp. 379–394.
Quadros Gomes, Ana Paula (2011b) A semântica de grau em PB. *Anais do SILEL* 2 (2). EDUFU.
Rothstein, Susan (2008) Two puzzles for a theory of lexical aspect: Semelfactives and degree achievements. In Johannes Dölling, Tatjana Heyde-Zybatow & Martin Schafër (orgs.), *Event structures in linguistic form and interpretation*. Walter de Gruyter, pp. 175–199.

Preprocessing models for speech technologies: The impact of the Normalizer and the Grapheme-to-Phone on hybrid systems

Bruna Carriço[1], *Christopher Shulby*[2], *Helena Moniz* [1,3]

[1]Universidade de Lisboa, Faculdade de Letras, Lisboa, Portugal
[2]Defined.ai, Seattle WA, United States
[3]INESC-ID, Lisboa, Portugal

Abstract

This paper describes the linguistic preprocessing methods on hybrid systems provided by an Artificial Intelligence (AI) international company, Defined.ai. The startup focuses on providing high-quality data, models, and AI tools. The main goal of this work is to enhance and advance the quality of preprocessing models by applying linguistic knowledge. Thus, we focus on two introductory linguistic models in a speech pipeline: Normalizer and Grapheme-to-Phone (G2P). To do so, two initiatives were conducted in collaboration with the Defined.ai Machine Learning team. The first project focuses on expanding and improving a European Portuguese Normalizer model. The second project covers creating G2P models for two different languages – Swedish and Russian. Results show that having a rule-based approach to the Normalizer and G2P increases its accuracy and performance, representing a significant advantage in improving Defined.ai tools and speech pipelines. Also, with the results obtained on the first project, we improved the normalizer in ease of use by increasing each rule with linguistic knowledge. Accordingly, our research demonstrates the added value of linguistic knowledge in preprocessing models.

Keywords: Speech technologies, Normalizer, Grapheme-to-Phone, linguistic knowledge, models.

Resumo

Este artigo descreve os métodos de pré-processamento linguístico em sistemas híbridos fornecidos por uma empresa internacional de Inteligência Artificial (IA), a Defined.ai. A *startup* concentra-se em fornecer dados, modelos e ferramentas de IA de alta qualidade. O objetivo principal deste trabalho é aprimorar e avançar a qualidade dos modelos de pré-processamento aplicando conhecimento linguístico. Assim, focamo-nos em dois modelos linguísticos introdutórios numa arquitetura de fala: o Normalizador e o Grafema-para-fone (G2P). Para isso, foram realizadas duas iniciativas em colaboração com a equipa de *Machine Learning* da Defined.ai. O primeiro projeto concentra-se em expandir e melhorar um modelo de Normalizador para o Português Europeu. O segundo projeto cobre a criação de modelos G2P para duas línguas – Sueco e Russo. Os resultados mostram que ter uma abordagem baseada em regras para o Normalizador e G2P aumenta a sua precisão e o seu desempenho, representando uma vantagem significativa na melhoria das ferramentas e das arquiteturas de fala da empresa. Além disso, com os resultados obtidos no primeiro projeto, melhoramos o normalizador em termos de facilidade de uso, aumentado cada regra com conhecimento linguístico. Dessa forma, a nossa pesquisa demonstra o valor do conhecimento linguístico em modelos de pré-processamento.

Palavras-chave: Tecnologias de fala, Normalizador, Grafema-Fone, conhecimento linguístico, modelos.

1. Introduction

Data preprocessing is a crucial step in building a machine learning model. If data is preprocessed, the results are consistent and of high quality (García et al., 2016). For example, modern speech recognition systems model linguistic entities at multiple levels, sentences, words, phones, and other units, using various statistical approaches (Jurafsky & Martin, 2022). The parameters of these models are usually trained on data, but their accuracy attempts to capture linguistic knowledge.

This text discusses the importance of data preprocessing in building high-quality machine learning models, particularly in speech recognition systems. The Linguistic Processing Module provided by Defined.ai is described, which includes the Normalizer and Grapheme-to-Phone pipelines. Two projects conducted by the Machine Learning Team at Defined.ai have also been discussed: Normalizer expansion for European Portuguese and Grapheme-to-Phone model creation for Swedish and Russian. The study aims to gain insight into how linguistic knowledge impacts speech technologies and how to upgrade and validate Normalizer and G2P models in different languages.

2. Speech technologies: Automatic Speech Recognition

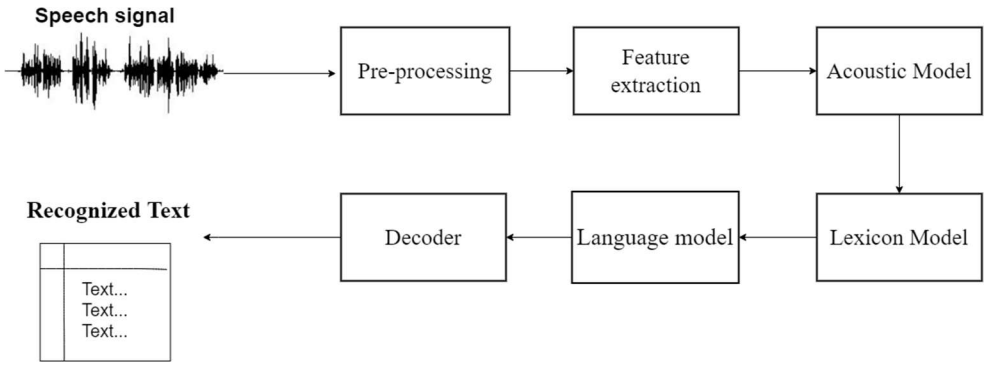
As stated by Alasadi and Deshmukh (2018), speech recognition is an important field that is constantly being developed as an interdisciplinary subfield of computational linguistics. Speech recognition creates technology and methods that empower the acknowledgment and understanding of natural language into a computer-understandable language. This is most generally known as Automatic Speech Recognition (ASR).

Automatic Speech Recognition has its origins back in the 1950s. Early ASR systems had a limited lexicon and were focused on numbers. In 1966, Hidden Markov Models (HMM) became a breakthrough in ASR and have remained state of the art for some time (Hennebert et al., 1994). Over the years, ASR technology has become more reliable and easier to handle due to computer technology and informatics advancements. By the 2000s, speech recognition technology had reached an accuracy rate of approximately 80%, and commercial applications, such as Siri and virtual smartphone assistants, became highly popular.

Classical machine learning models can be classified into two different approaches: Generative[1] and Discriminative[2] approaches. Speech recognition has mostly used the Generative approach in the past decades, and two popular methods based on this approach are HMMs and Gaussian mixture models (GMMs). The basic framework of a speech recognizer system includes three major stages: capture, transducing, and decoding. The acoustic model, lexicon model, and language model are combined during the transducing step. A typical ASR framework includes the components described in Figure 1.

[1] Generative approach: used to learn each language and determine which language the speech belongs to.
[2] Discriminative approach: used to determine the linguistic differences without learning any language.

Figure 1. Basic framework of a speech recognizer system



The field of automatic speech recognition has seen two main approaches over the years: a traditional hybrid approach and an end-to-end (E2E) deep learning approach. The hybrid approach has been dominant for the past decade due to extensive research and training data available. However, the end-to-end approach has gained popularity due to its simplicity, reduced training and decoding time, and comparable accuracy to the hybrid approach (Vielzeuf & Antipov, 2019).

On the other hand, in accordance with (Kurata et al., 2019) an end-to-end Deep Learning approach is a new paradigm in neural network-based speech recognition that has several advantages. Traditional hybrid ASR systems, which consist of an acoustic model, a language model, and a lexicon model, each of which might be sophisticated, require independent training of these components. In contrast, E2E ASR is a unified method with a much simpler training pipeline and models that perform at low audio frame rates. This reduces the amount of time spent on training and decoding. A common end-to-end Deep Learning architecture is the encoder-decoder, which is implemented with RNNs (Recurrent Neural Network).

Errattahi and El Hannani (2017) concluded that the benchmarks have shown that the Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) have proven their efficiency on several Large Vocabulary Continuous Speech Recognition (LVCSR) tasks by outperforming the traditional Hidden Markov Models. LVCSR is characterized by the authors as being the most challenging task in ASR. Recently, Rajadnya (2020) developed an application for continuous speech recognition based on DNN-HMM and Deep Belief Network (DBN) algorithms. This proved that using DNN-HMM with DBN supplies better accuracy than using typical GMM-HMM systems. Deep learning is quickly becoming a standard approach for speech recognition, having effectively replaced all the classical approaches.

In recent years, ASR systems have significantly improved in accuracy, but they still cannot reach human-level accuracy. Factors such as speaker-dependent or speaker-independent models, acoustic models, vocabulary, and language models can influence the performance of ASR systems. System failures can also occur due to issues such as noisy environments and different pronunciations by speakers.

Recently, bias and gender issues in data have also deserved attention from the community. Gender and racial bias are a concept where models and algorithms do not provide optimal services to people of a specific gender or dialect. The disparity of accessible data for both genders and dialects has been proven in recent studies to be its main cause. Brasoveanu and Dotlacil (2020) found bias against women in the performance of speech recognition systems by analyzing the gender representation in different corpora. Bias was also identified against dialect groups; dialect speakers have lower ASR performance than speakers of standard pronunciations (Wassink et al., 2022).

3. Linguistic preprocessing

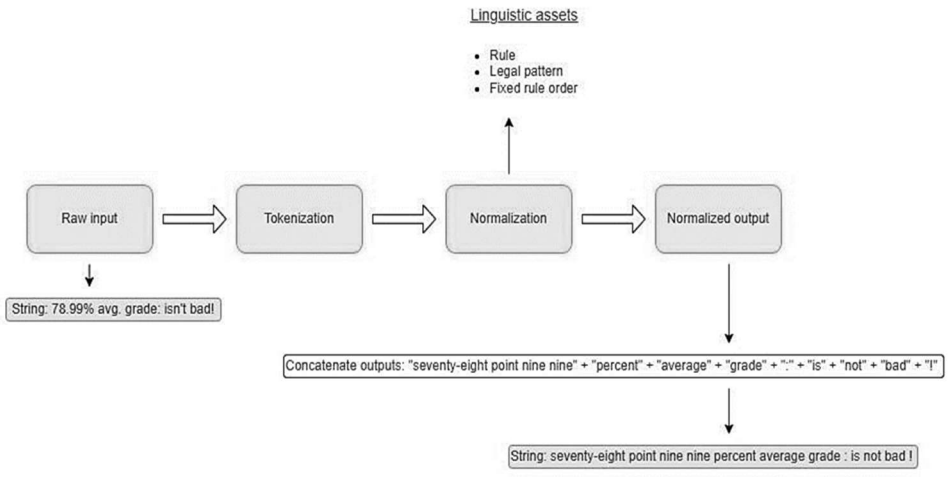
Defined.ai uses a Linguistic Processing Module (LPM) to treat raw text as input and structure it to make it usable for speech technologies that require linguistic knowledge. The LPM is used in many technologies, such as ASR and Text-to-Speech (TTS) systems. Therefore, LPM includes the development of pronunciation lexicons that will provide a mapping between a word's orthographic form and its pronunciation form. Pronunciation lexicons are mainly used to build G2P models to provide pronunciation of OOV words.

In the LPM module are included the Grapheme-to-Phone (G2P) and the Normalizer pipelines. These introductory linguistic models are essential to ensure the quality of data processing and measure its quality. A crucial step since it ensures high-quality data processing throughout the pipelines. By describing these models, we explain the process of converting written text into its spoken form and how we transform graphemes into phonetic transcription, in a stepwise perspective.

Such linguistic models are frequently used as pre-requirements for the preparation of any corpus for ASR systems. At Defined.ai these pre-requirements are one of the company's initiatives that support ASR systems and DefinedData (one of Defined.ai products). The Machine Learning (ML) team ensures the creation and assistance of in-house G2P conversion and Normalization solutions. Consequently, the outcome of this initiative is to create a more robust data collection pipeline and to develop pre-requirements for training data.

The Normalizer pipeline (Figure 2) converts written text into a spoken form that can be used for ASR systems as training data. The normalizer pipeline consists of two main steps - text preprocessing and normalization. Text preprocessing involves cleaning the text by removing unwanted characters and normalizing punctuation, whitespaces, and Unicode characters. Next, tokenization separates the raw input into tokens by considering non-printable characters, spaces, and logical semantic breaks. Normalization involves applying rules to each word of the sentence to produce the spoken form output. The current Normalizer supports 12 natural languages, including English, Portuguese, German, and French, and provides normalized forms for numbers, dates, measurements, time, durations, and symbols. The normalization process is rule-based, making it possible to predict the output of the Normalizer accurately, but new rules can be written to cover new constructions.

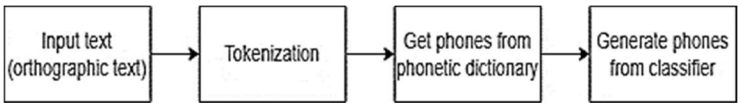
Figure 2. Normalizer pipeline



The G2P pipeline, illustrated in Figure 3, is used to predict the phonetic transcription of a given written word to create and improve ASR systems. The pipeline requires Tokenization and checks vocabulary before

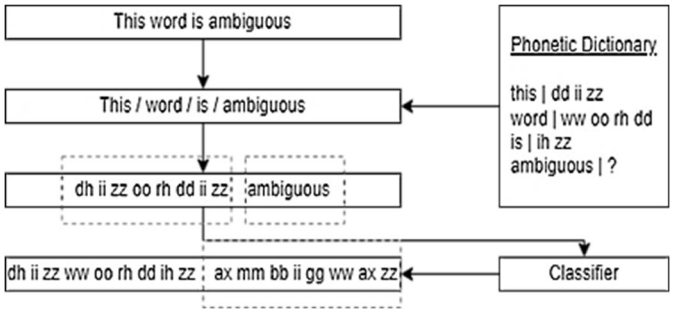
extracting the pronunciation of each word from the lexicon. If a word is not recognized by the lexicon, the Classifier predicts its phonetic transcription.

Figure 3. Grapheme-to-Phone pipeline



The process of pronunciation prediction, shown in Figure 4, starts with the tokenization of orthographic text as input. In the first line, we can see the orthographic text as the input, which is tokenized. After that, we will get the corresponding phonetic transcription of each word that was taken from the pronunciation lexicon. If a word is not recognized by the lexicon (e.g., ambiguous) the Classifier will then predict the phonetic transcription of that given word. The G2P currently supports 18 languages, including English, Portuguese, and Japanese.

Figure 4. Pronunciation prediction process



4. **Phonetic and phonological concepts**

In this section, we introduce the main phonetic and phonological aspects of two distinct languages – Swedish and Russian. We start by introducing the main linguistic concepts used to make a description of each language: phoneme, phone, free variation, and contextual variation.

A phonological segment, also known as a phoneme, is the smallest distinctive unit of sound in a language that can change the meaning of a word. Phonemes are abstract representations of speech sound that exist in the mental grammar of speakers of a particular language. For example, in English, the sounds represented by the letters *p* and *b* are considered different phonemes because they can change the meaning of a word (e.g., *pat* vs. *bat*), even though the actual acoustic properties of the sounds may vary depending on their context or speaker. A phone, on the other hand, is a physical realization of a speech sound. It refers to the actual sound produced by a speaker, including all the variations in pronunciation due to factors like accent, speaking rate, and context.

Free variation refers to a linguistic phenomenon where two or more different sounds can occur in the same linguistic context without affecting the meaning of a word or utterance. In other words, when sounds are in free variation, they can be used interchangeably by speakers without any change in the word's meaning or grammatically. On the other hand, contextual variation is another linguistic phenomenon where the pronunciation of a sound form depends on its position within a word or sentence or its surrounding linguistic context. Contextual variation is typically rule-based and can affect the meaning or grammaticality of words. Resulting from contextual variation of Russian and Swedish, the following processes were deemed important: devoicing at the end of a word, assimilation of voicing between two obstruents.

For the context of this work, we focus mainly on phonetic aspects of such languages since we aim to describe and identify speakers' real pronunciations to implement this knowledge on the G2P model. Thus, this

chapter will mostly cover phonetic transcription and phonological processes that were strictly chosen regarding their relevance to the G2P. In the process of determining the selection criteria, the following factors were considered: (1) Frequency and occurrence – prioritize phonological processes and transcription conventions that are commonly encountered in real-world languages and writing systems; (2) Phonemic variability – consider phonological processes that involve phonemic variability and may pose challenges for G2P conversion. These processes can include allophony, assimilations, and neutralization, which can affect how graphemes are pronounced, and (3) Cross-linguistic applicability – consider phonological processes and transcription conventions that have applicability across multiple languages or writing systems, as they can enhance the versatility of the G2P system.

We begin with a key element of both speech recognition and text-to-speech systems: how words are pronounced in terms of individual speech units called phones. Following (Jurafsky & Martin, 2022) a speech recognition system needs to have a pronunciation for every word it can recognize, and a text-to-speech system needs to have a pronunciation for every word it can say. We model the pronunciation of a word as a string of symbols that stand for phones or segments. A phone is a speech sound and phones are represented with phonetic symbols that bear some resemblance to a letter in an alphabetic language, like English or Portuguese. In the context of this work, we use two different alphabets for describing phones: the International Phonetic Alphabet (IPA),[3] and the DC-Arpabet[4] (phonetic alphabet created by Defined.ai based on the arpabet alphabet). IPA was created in the late 19th century to describe the sounds of all human languages while using a set of transcription principles and its phones. On the other hand, the DC-Arpabet was specifically designed by Defined.ai for US English, while using ASCII symbols. Table 1 shows some examples of an American English phone set using DC-Arpabet symbols for transcribing its consonants, semivowels, vowels, and diphthongs together with their IPA equivalents.

Table 1. Examples of the American English phone set using IPA, and DC-Arpabet symbols

Phonetic Alphabet Symbols		**Classification**
IPA	**DC-Arpabet**	
[b]	[bb]	consonant
[ð]	[dh]	consonant
[s]	[ss]	consonant
[ʃ]	[sh]	consonant
[ʒ]	[zh]	consonant
[j]	[yy]	semivowel
[w]	[ww]	semivowel
[ə]	[ax]	vowel
[ʊə]	[ux]	diphthong

The DC-Arpabet currently contains 239 indexed unique symbols, each composed of at least 2 alphanumerical characters, plus an optional 1 to 3 alphanumerical characters. The symbols are represented by '##' and '_', and whenever an audio includes silence, short pause, unintelligible sounds, vocal noise, or music we represent it using the respective abbreviation shown in Table 2. Building the G2P model for a language also requires a list of characters used to spell that language, that is, its graphemes.

[3] Armstrong and Meier (2005).
[4] Shoup (1980).

Table 2. Symbols and non-phonetic sound representation

Phonetic Alphabet Symbols			Classification
IPA	**DC-Arpabet**	**X-SAMPA**	
##	##	##	sentence break
	_		skipped phone
	Sil		silence
	Sp		short pause
	Zun		unintelligible
	Zvn		vocal noise
	Zmu		

4.1. **Swedish**

This section deals with the phonetics and phonology of Standard Swedish from a synchronistic perspective. We want to focus on describing the authentic pronunciation of speakers. First, we decided to choose one single variant among all existing ones in the Swedish language – Standard Swedish. This variant evolved from the Central Swedish dialects, and most Swedes speak it. Herewith we can predict how most speakers pronounce Swedish. Thus, in this section, we start by briefly describing the Swedish vowel and consonant inventory, where we discuss phenomena crucial to measuring the impact on the G2P. Following, we introduce two frequent features of this language – quantity and retroflexion – which were important in the decisions we had to make for the Swedish G2P model. Also, resulting from the contextual variation of Swedish, the following processes were deemed important: neutralization, vowel reduction, and vowel lengthening.

The Swedish orthographic alphabet consists of nine vowels: <a, e, i, o, u, y, å, ä, ö>. Riad (2014) considers nine distinct vowel phones. Each vowel occurs with long and short variants. The author finds these variants allophones of the same phoneme. The long vowels are [iː, yː, eː, ɛː, øː, ʉː, uː, oː, ɑː] and the short vowels [ɪ, ʏ, ɛ, ø, ɵ, ʊ, ɔ, a], respectively, in Table 3.

Table 3. Standard Swedish orthographic vowels

Orthography	IPA		
	Phoneme	**Long vowel**	**Short vowel**
<i>	/i/	[iː]	[ɪ]
<y>	/y/	[yː]	[ʏ]
<e>	/e/	[eː]	[ɛ]
<ä>	/ɛ/	[ɛː]	[ɛ]
<ö>	/ø/	[øː]	[ø]
<u>	/ʉ/	[ʉː]	[ɵ]
<o>	/u/	[uː]	[ʊ]
<å>	/o/	[oː]	[ɔ]
<a>	/ɑ/	[ɑː]	[a]

In Standard Swedish, several vowel qualities are distinguished in unstressed syllables. The phones /e/ and /ɛ/ neutralize in the short variant, as [ɛ]. The alternations between long and short vowels provide important cues to the phonemic system. Thus, Riad (2014) defends that the lowering of /ø/, and /ɛ/ before a retroflex motivates

the height separation between /a/, which is [low], and /ø/ and /ɛ/ which are [mid]. The two main short allophones of /e/ and /ɛ/ are neutralized as [ɛ]. This results in eight short vowel allophones to match the nine long vowel allophones. However, many dialects have nine long vowels and nine short vowels (Riad, 2014). In some varieties of Standard Swedish, a similar neutralization occurs for short /ø/, where some young speakers have neutralization as [ɵ].

Also, in many cases <e> and <ä> as in *sett* and *sätt* coincide and are both pronounced /e/. This can lead to the mistaken belief that there are only eight short vowels and that [e] and [ɛ] are allophones. Yet, in Standard Swedish, /e/ and /ɛ/ are treated as phones in Standard Swedish. See the vowel phonetic inventory used for Standard Swedish in Table 4.

Table 4. Standard Swedish phonetic vowels and respective examples

Phonetic symbols		**Example (DC-Arpabet)**
IPA	**DC-Arpabet**	
[a]	[aa]	[kk **aa** ll ee nn]
[ɑ:]	[ah]	[**ah** vv ss nn ih tt]
[ɛ]	[eh]	[bb aa ng kk **eh** rr]
[æ:]	[ae]	[ll vv **ae** gg aa rr]
[e:]	[ee]	[bb ll **ee** kk tt]
[ɪ]	[ih]	[ss **ih** vv ii ll aa]
[i:]	[ii]	[dd ee ll tt **ii** dd]
[ɔ]	[oh]	[ee gg **oh** nn]
[o:]	[oo]	[bb **oo** gg eh]
[œ]	[oe]	[ff **oe** ll yy eh tt]
[ø:]	[eu]	[hh **eu** rr ss aa mm mm aa tt]
[ʊ]	[ug]	[yy **ug** nn ss oh nn]
[u:]	[uu]	[kk **uu** nn kk uo rr ss]
[ʏ]	[iu]	[kk ll **iu** ff tt aa]
[y:]	[iy]	[cc **iy** ll rr uo mm]
[ɵ]	[uo]	[ll **uo** dd vv ih gg]
[ʉ]	[uc]	[mm ih nn **uc** tt eh nn]
[j]	[yy]	[oo gg **yy** ih ll tt ih kk tt]

The Swedish orthographic alphabet consists of 14 consonants: <b, d, f, g, h, l, m, n, p, q, r, s, t, v>. Table 5 shows Standard Swedish orthographic consonants and their respective pronunciations. On the other side, the consonant phonetic inventory contains 17 different phones (see Table 8).

Table 5. Standard Swedish orthographic consonants

Orthography	Phonetic symbols	
	IPA	DC-Arpabet
<b>	[b]	[bb]
<d>	[d]	[dd]
<f>	[f]	[ff]
<g>	[g]	[gg]
<h>	[h]	[hh]
<l>	[l]	[ll]
<m>	[m]	[mm]
<n>	[n], [ŋ]	[nn], [ng]
<p>	[p]	[pp]
<k>	[k]	[kk]
<r>	[r], [ʂ]	[rr], [rr ss]
<s>	[s], [ɧ]	[ss], [shx]
<t>	[t]	[tt]
<v>	[v]	[vv]

In contrast to vowels, the opposition length of consonants does not carry any meaning. The consonant system has a double specification of aspiration and voicing in the obstruents. Riad (2014) focuses on the qualitative contrasts of the Swedish consonant system, arranged according to the place of articulation and manner of articulation (see Table 6).

Table 6. Standard Swedish consonants inventory modified and extracted from Riad (2014, p. 49)

	labial,	dental,	alveolar,		
	labiodental	alveolar	palatal	velar	glottal
oral stop	p	t		k	
	b	d		g	
fricative.	f	s	ç		h
fricative/retroflex			ʂ		
fricative/approximant	v				
nasal stop	m	n		ŋ	
Lateral		l			
apical trill		r			

Standard Swedish has palatal and velar voiceless fricatives: /s/, /ʂ/, /ɕ/. The phones [ɧ] and [ɕ] are very similar, although the most prominent phonetic difference lies in its place of articulation. While /ɕ/ has a stable place of articulation, /ʂ/ is subject to contextual conditioned allophony in Standard Swedish, as well as a wide-ranging sociolinguistic variation. In onset position, /ʂ/ can have four different variations – [ɧ], [ɧʷ], [x], [ʂ] –

while in postvocalic position the realization of /ʂ/ is mostly [ʂ] or [ʂː]. Thus, the most common prevocalic realization is [ɧ]. /ʃ/ is also used in some Swedish dictionaries (*e. g.*, *Svensk Ordbok*[5]).

One of the most salient features of North Germanic standard varieties is the quantity system. Stressed syllables are invariably heavy, due to a prosodic condition. This condition is met in either the vowel alone or in a combination of the vowel and the following consonant. In a stressed syllable one segment must be long, either the vowel or the consonant, but not at the same time. Vowels and consonants thus occur in long and short variants, and it is primarily in terms of quantity that these segmental distinctions are made and described.

There are qualitative differences within vowel pairs. Each long vowel has a short counterpart. The long and short consonants in a pair are naturally much more similar in quality. Most consonants have long and short pairs, but there are a few that exhibit a defective quantitative distribution. Two phones never occur directly after a stressed vowel, namely /h/ and /ɕ/, and hence lack long variants altogether. The segments /j/ and / ŋ/, on the other hand, are always long in a postvocalic coda position, provided that the syllable is stressed. The phoneme /ŋ/ never occurs word-initially, but may occur intervocalically and as onset in unstressed positions.

Syllable weight in stressed syllables is phonetically and phonologically clear. Any stressed syllable is bimoraic, where a long vowel is bimoraic, and a short vowel monomoraic. A long consonant is (mono)moraic, and a short consonant is non-moraic. If the vowel is long, then all is fine. If the vowel is short, then the following consonants must be long. In accordance with (Riad, 2014) most of the other Germanic languages lost consonant quantities early on and this has led to rather different quantitative phonologies. In this section, we shall assume that quantity is distinctive in consonants, but some consonants have lexical length, while others become grammatically lengthened or shortened by syllabification. For the vowels, quantity is predictable from prosodic context, when a syllable is stressed or when there is quantitative information (*e. g.*, lexical or positional) in the following consonant.

Another striking feature of Standard Swedish is retroflexion. The retroflexion rule shown in Table 7, creates retroflex sound when two contiguous segments (/s, t, d, n, l/ with a preceding /r/) converge into one element. The output /ʂ, ʈ, ɖ, ɳ, ɭ/ is phonologically distinct from the input segments. Thus, retroflex consonants can appear in most simple words (*e.g.*, *framfar**t**, rampaging),* but can also occur in other articulatory patterns – word boundaries, inflections, compounds, and derivations. In word boundaries, a retroflex consonant emerges if the final letter of a word is an <r> and the initial letter of the following word is <t, d, s, l, n>.

Table 7. Standard Swedish retroflexion rule

Swedish form	**Phonetic transcription**	**English translation**
vår triumf	/voːʈriɵmf/	our victory
hur mår du	/hʉːrmoːɖɵ/	how are you
under sängen	/ɵndeʂɛŋen/	under the bed
eller nej	/ɛleɳɛj/	or not
hur ledsam	/hʉːlesam/	how sad

As for flections, when the genitive <s> is attached to a word ending with <r>, the retroflex /ʂ/ is used (e.g, Peters hus, /peteʂhʉːs/, Peter's house). When a verb ends with a final <r> the retroflex consonants /ɖ/, /ʈ/ occur (e.g., stör-de, /støːɖ /; stör-t, /støːʈ/). Furthermore, this rule also applies to past participles and nouns. Thus, retroflex consonants also occur in compound words (e. g., vårdag, /voːɖɑːg /, spring day) and derived words (varsam, /vɑːʂam/, careful). The phonetic consonants for Standard Swedish are listed in Table 8 along with some examples.

[5] https://svenska.se

Table 8. Standard Swedish phonetic consonants and respective examples

Phonetic symbols		Example (DC-Arpabet)
IPA	**DC-Arpabet**	
[b]	[bb]	[**bb** ee ff uu gg aa dd]
[d]	[dd]	[**dd** ee ll tt uu gg]
[f]	[ff]	[**ff** aa mm ih ll yy]
[g]	[gg]	[**gg** rr oe nn tt]
[h]	[hh]	[**hh** ug tt eh ll]
[l]	[ll]	[**ll** ii nn dd rr ih gg]
[m]	[mm]	[**mm** oh ng aa]
[n]	[nn]	[**nn** ae rr aa]
[ŋ]	[ng]	[rr ii **ng** nn ee rr]
[p]	[pp]	[oe **pp pp** eh nn]
[k]	[kk]	[bb eh **cc** eh nn eh rr]
[r]	[rr]	[**rr** eu kk]
[ʂ]	[rr ss]	[bb aa **rr ss** eh bb eh kk]
[s]	[ss]	[**ss** vv oo rr]
[ɧ]	[shx]	[aa gg aa rr **shx** ih ss mm]
[t]	[tt]	[**tt** eh nn dd eh]
[v]	[vv]	[**vv** ae rr dd]

4.2. **Russian**

This section deals with the phonetics of Standard Russian from a synchronistic perspective. We start by briefly outlining the Russian vowel inventory and consonant inventory. Thus, resulting from the contextual variation of Russian, the following processes were deemed important: neutralization, devoicing at the end of a word, assimilation of voicing between two obstruents, and palatalization.

Finally, we describe and discuss the value of two special signs in Russian phonetic and phonological system. Therefore, to make a plan for training a Russian G2P model, we also need to discuss which phonetic representations are best represented of Russian graphemes, how to treat iotated vowels, and if we should make a contrast between palatalized consonants and non-palatalized consonants.

The modern Russian alphabet consists of 33 letters: 20 consonants (<б, в, г, д, ж, з, к, л, м, н, п, р, с, т, ф, х, ц, ч, ш, щ>), ten vowels (<а, е, ё, и, о, у, ы, э, ю, я), one semivowel (<й>), and two modifier letters or signs (<ь, ъ>) that alter pronunciation of preceding consonants or a following vowel. Table 9 shows the Standard Russian orthographic vowels and their respective pronunciations.

Table 9. Standard Russian orthographic vowels, semivowel, iotated vowels, and signs

Orthography	Phonetic symbols	
	IPA	**DC-Arpabet**
<а>	[a]	[aa]
<е>	[e], [ja], [je], [jo], [ju]	[ee], [yya], [yye], [yyo], [yyu]
<э>	[e]	[ee]
<и>	[i]	[ii]
<о>	[o]	[oo]
<у>	[u]	[uu]
<ы>	[ɨ]	[ie]
<й>	[j]	[yy]
<ё>	[ja], [je], [jo], [ju]	[yya], [yye], [yyo], [yyu]
<ю>	[ja], [je], [jo], [ju]	[yya], [yye], [yyo], [yyu]
<я>	[ja], [je], [jo], [ju]	[yya], [yye], [yyo], [yyu]
<ь>	no phonetic value	
<ъ>	no phonetic value	

In most analyses, the Russian vowel inventory contains five vowel phones: <i, e, a, o, u>. However, studies on Modern Russian (Timberlake, 2014; Yanushevskaya & Bunčić, 2015) claim a sixth vowel <ɨ>. Each one of these vowels is realized as a rich set of allophones ruled by stress and phonological environment. In most cases, these vowels merge into two or four vowels when stressed: /i, u, a/ after hard consonants and /i, u/ after soft consonants.

In orthography, each vowel is represented by two letters. This happens so we can distinguish the not-iotated vowels and the iotated vowels. Vowels in Russian do not have a phonemic distinction of quantity; there are no words distinguished by, for example, a long [aː] as opposed to a short [a]. Thus, Table 10 provides the 11 phonetic vowels in Standard Russian, together with a semivowel.

Table 10. Standard Russian phonetic vowels and respective examples

Phonetic symbols		Example (DC-Arpabet)
IPA	**DC-Arpabet**	
[a]	[aa]	[bb **aa** nn kk uu]
[e]	[ee]	[vv ss tt rr yye zz **ee**]
[i]	[ii]	[vv yye txj yye rr nn **ii** xx]
[o]	[oo]	[mm ii nn **oo** vv aa ttj]
[u]	[uu]	[ss oo oo bb cx tx **uu**]
[ɨ]	[ie]	[tt **ie** ss yya txj yye]
[j]	[yy]	[tt uu tt aa kk oo **vv** aa]
[ja]	[yya]	uu vv aa zj ee nn ii **yya**]
[je]	[yye]	[ii ss tt **yye** ts]
[jo]	[yyo]	[ll **yyo** xx kk aa vv aa]
[ju]	[yyu]	[ll uu txj su uu **yyu**]

As mentioned before, Standard Russian has 20 consonants (see Table 11). However, its phonetic inventory has 32 different consonants, as shown in Table 12. Most of these occur with both a palatalized and a non-palatalized version. The remaining consonants have a single version – the velars are palatalized before front vowels; palatals are either invariably palatalized (e.g., /j/) or invariably not.

Table 11. Standard Russian orthographic vowels

Orthography	Phonetic symbols	
	IPA	DC-Arpabet
<б>	[b], [bʲ]	[bb], [bbj]
<в>	[v], [vʲ]	[vv], [vvj]
<г>	[g]	[gg]
<д>	[d], [dʲ]	[dd], [ddj]
<ж>	[ʒ], [ʒʒ]	[zj], [zx zx]
<з>	[z], [zʲ]	[zz], [zzj]
<к>	[k]	[kk]
<л>	[l], [lʲ]	[ll], [llj]
<м>	[m], [mʲ]	[mm], [mmj]
<н>	[n], [nʲ]	[nn], [nnj]
<п>	[p], [pʲ]	[pp], [ppj]
<р>	[r], [rʲ]	[rr], [rrj]
<с>	[s], [sʲ]	[ss], [ssj]
<т>	[t], [tʲ]	[t], [ttj]
<ф>	[f]	[ff]
<х>	[x]	[xx]
<ц>	[t͡s]	[ts]
<щ>	[t͡ɕ]	[tx]
<ш>	[tʂ]	[txj]
<ч>	[ʂ]	[sj]

Thus, ш and щ share a very similar way of pronouncing their sounds. The letter ш has the phonetic value of [ʃtʃ] (a "palatalized /ʃ/" followed by a sound similar to [tʃ], in which the stop element, represented by t, is weak) or [ʃʃ] (a long "palatalized /ʃ/"). Either of these pronunciations of ш is regarded as correct, but it is common for any speaker to use only one of them. The letter щ has a phonetic value that can switch between [tʃ] (a weak t followed by a "palatalized /ʃ/") and [tɕ] (a weak t followed by a [ɕ]). Either of these is correct but it is pronounced differently depending on the region of Russia (Timberlake, 2014).

One of the most characteristic features of Russian consonantal phonology is that most sounds have both a palatalized and a non-palatalized phonological segment. In accordance with (Bondarko, 2005), palatalized consonants are referred to as soft and non-palatalized consonants as *hard*. Palatalization is an articulation of a consonant in which the blade of the tongue moves toward the hard palate. For example, when a non-palatalized consonant is pronounced, the tip of the tongue is touched near the teeth, while the middle of the tongue lies low in the mouth. In contrast, when the palatalized consonants are pronounced, the tip of the tongue touches behind the upper teeth, and the blade and the middle of the tongue are raised towards the hard palate. Most consonant articulations in Russian have two forms, with or without palatalization. Thus, palatalization in Russian is indicated by adding a diacritic to the phone. According to IPA, we use a palatalized diacritic (/ʲ/) when referring to palatalized consonants (e.g., /bʲ, dʲ, gʲ/). Palatalization is contrastive in word-final and in heterogenic medial coda positions.

Some examples illustrating the palatalization contrast extracted from (Padgett, 2001) are given in (1). Contrasts (1a–b) are prevalent in the language, while (1c) is more limited due to assimilations and neutralizations in that context.

(1a) before back vowels

Mat	foul language	mat'	crumpled
Rat	glad	r'at	row
Vol	ox	v'ol	he led
Nos	nose	n'os	he carried
Suda	court of law	s'uda	here, this way

(1b) word-finally

Mat	foul language	mat'	mother
Krof	shelter	krof'	blood
Ugol	corner	ugol'	(char)coal
v'es	weight	v'es'	entire

(1c) before another consonant

polka	shelf	pol'ka	polka
tanka	tank		
v'etka	branch		
Gorka	hill		

Russian has two modifier signs with no phonetic value. The soft sign *Ь* signals the presence of a soft consonant, and the hard sign *Ъ* signals the presence of a hard consonant. Therefore, they can be used between a consonant or a vowel (*Ь*) and between a consonant and a vowel, between two consonants, or at the end of a word after a consonant (*Ъ*).

In Russian *Ь* has much wider usage than *Ъ*; *Ь* can be used at the end of words or in between two consonants, and it indicates that the preceding consonants are soft. Neither *Ь,* or *Ъ* can be a stand-alone letter or the first letter in a word.

At the same time, there are some letters in Russian that are always hard or always soft and will sound the same way, whether there is a soft sign (Ь), or not. The consonants <ж, ш, ц> are always hard (if these consonants are followed by a soft sign, the sign cannot soften the consonant and serves a purely grammatical purpose. The consonants <ч, щ, й> are always soft. A soft sign following these consonants, once again, serves only a grammatical purpose.

Table 12. Standard Russian phonetic consonants and respective examples

Phonetic symbols		Example (DC-Arpabet)
IPA	**DC-Arpabet**	
[b]	[bb]	[gg **aa** zz bb aa nn kk aa]
[bʲ]	[bbj]	[tt uu rr **bbj** yye rr nn]
[v]	[vv]	[tt uu tt aa kk oo **vv** aa]
[vʲ]	[vvj]	[sj yye **vvj** yye vv]
[g]	[gg]	[sj pp ii **gg** yye llj]
[d]	[dd]	[ee **dd** gg aa rr]
[dʲ]	[ddj]	[bb uu **ddj** tt yye]
[ʑ]	[zj]	[vv rr aa **zj** nn aa yy]
[ʑʑ]	[zx zx]	[pp oo bb yye rr yye **zx zx** yyu]
[z]	[zz]	[bb yye **zz** aa llj]
[zʲ]	[zzj]	[vv oo **zzj** mm yyo ts yya]
[k]	[kk]	[gg rr uu zz oo vv ii **kk** oo vv]
[l]	[ll]	[gg rr yya nn uu **ll**]
[lʲ]	[llj]	[gg uu bb ii tt yye **llj** nn oo]
[m]	[mm]	[dd yye ll aa **mm** ii]
[mʲ]	[mmj]	[vv ii ts yye pp rr yye **mmj** yye rr]
[n]	[nn]	[ii mm yye **nn** nn oo]
[nʲ]	[nnj]	[dd yye **nnj** gg aa mm]
[p]	[pp]	[mm ii tt rr oo **pp** oo ll ii tt]
[pʲ]	[ppj]	[**ppj** rr yye mmj yye rr oo mm]
[r]	[rr]	**[rr** yye zz aa ll]
[rʲ]	[rrj]	[**rrj** aa dd ii tt yye ll yya mm]
[s]	[ss]	**[ss** vv yye tt yya tt]
[sʲ]	[ssj]	[ss vv yya zz aa ll oo **ssj]**
[t]	[t]	[ss mm yye **tt** uu]
[tʲ]	[ttj]	[ss mm oo tt rr yye **ttj]**
[f]	[ff]	[ss pp yye ts ii **ff** ii kk aa]
[x]	[xx]	[ss rr oo kk aa **xx**]
[t͡s]	[ts]	[dd yye mm oo pp pp oo zz ii **ts** ii ii]
[t͡ɕ]	[tx]	[tt yye kk uu **tx** ii xx]
[tʂ]	[txj]	[nn ii nn **txj** yye]
[ʂ]	[sj]	[tt ii **sj** ii nn aa]

5. **Methodology**

To improve version 2 of the European Portuguese normalizer, we first analyze how version 1 of the normalizer performed in comparison to how we want version 2 to work. Secondly, we outline how to build two new G2P models for different languages. Briefly, two transitional questions served as the basis for our research:

(1) How to upgrade a Normalizer into one that covers most of the normalizable tokens? And (2) How to create and validate new G2P models in two different languages?

The Normalizer Linguistic Expansion (NLE) project aimed to expand the rules of the Replacement Maps (RMs) in the Normalizer to cover Real numbers, Symbols, Abbreviations, Ordinals, Measurements, Currency, Dates, and Time, respectively. Table 13 shows an example of the input and its respective normalized output.

Table 13. Normalization process – input and output example

Input	Normalized Output
não mais **cm2** de território inacessível para a polícia eles somaram um buraco de **1577 cm2** na parede	não mais **centímetros quadrados** de território inacessível para a polícia eles somaram um buraco de **mil e quinhentos e quarenta e sete centímetros quadrados** na parede

The goal was to have a more consistent and simpler Normalizer with greater coverage of unambiguous inputs. We added new symbols and abbreviations to the RMs and created Unit Tests (UTs).[6] The Symbols rule was designed to convert non-alphanumeric symbols into their spoken forms, but only when they were not handled by other rules. The Abbreviations rule was created to convert miscellaneous alphabetic or mixed alphabetic-symbolic sequences to their spoken forms (see Table 14). In version 2 of the Normalizer, all abbreviated forms were added to allow for multiple possible expansions due to lexical or inflectional reasons.

Table 14. Normalization process – abbreviations input and output

Input	Normalized output
Wi-Fi	**Uaifai**
r/c	**rés do chão**
O **ap.** fica longe	o **apartamento** fica longe
O **D.r** Duarte trabaha no Hospital Santa Maria	o **doutor** Duarte trabaha no Hospital Santa Maria
Ela é vegetariana, **i. e.**, ela não come carne e peixe	ela é vegetariana, **isto é**, ela não come carne e peixe
O **núm.** de erros é alto	o **número** de erros é alto

The Real Numbers rule converted numeric values of integers and/or decimal values into their spoken forms. The Ordinals rule converted a sequence of a real number and an ordinal marker into their expanded spoken forms, while the Measurement rule converted a sequence of a real number and a measurement unit into their expanded spoken forms. The Currency rule covered currency expressions in their expanded forms. In version 2 of the Normalizer, the Measurement and Currency rules introduced several new abbreviations and symbols that were expanded to their spoken form only when they were unambiguous.

Throughout the NLE project, we expanded several normalizable tokens (symbols, abbreviations, etc.), implemented rules on the various rules, and fixed issues related to varieties confusion between pt-PT and pt-BR. We ensured that pt-PT rules and pt-BR rules were always separate in version 2. The main problem was that several PN rules and assets were shared between different regions of the same language. This had been done in pt-PT and pt-BR for several rules, but mainly due to some orthographic differences between pt-PT and pt-BR.

[6] A Unit Test (UT) is a set of one or more examples to ensure that a unit of code behaves as expected. More specifically, Unit Tests are sets of examples of inputs and expected outputs that test whether the normalizer produces the correct - or expected - output given the input.

To prepare a G2P model for two different languages - Swedish and Russian – an overview of the language had to be performed. Once this task was completed, the next step involved developing a phone set for each language. This implicated identifying the distinct sounds present in the language and creating a set of symbols (phones) to represent them. This phone set serves as the basis for the phonetic transcriptions used in the G2P model. Following this, the phonetic lexicon for each language was mapped and automatically converted to the DC-Arpabet format. DC-Arpabet is a standardized phonetic transcription system used for English, but it can also be applied to other languages. This step involved mapping the phones from the initial phone set to the corresponding DC-Arpabet symbols. The phonetic lexicon was then revised and corrected by native speakers. This step is crucial to ensure the accuracy and quality of the G2P model. During this stage, errors in the phonetic transcriptions were identified and corrected, and missing words or pronunciations were added. Finally, the G2P model was evaluated to assess its performance and accuracy. This involved testing the model on a set of known words and comparing the model's predicted phonetic transcriptions to the correct pronunciations. If necessary, adjustments were made to the model to improve its accuracy.

6. Results and Discussion

6.1. Evaluation metrics

This study discussed the metrics used to evaluate the Normalizer and G2P tools. For the Normalizer, we used accuracy, WER, precision, recall, and F1 score. However, WER may not be an adequate metric since most tokens in a sentence are likely to be irrelevant to the normalizer. We introduced the WERnorm metric to address this issue, which calculates the edit distance over the number of normalized reference tokens. Concerning the G2P, precision, recall, and F1 score were used to evaluate the proportion of correctly predicted grapheme to phone correspondences. It is important selecting appropriate evaluation metrics to accurately assess the performance of these models.

In accordance with Jurafsky and Martin (2022), the word error rate is based on how much the word string returned by the recognizer (the hypothesized word string) differs from a reference transcription. Thus, WER is the proportion of transcription errors that the ASR system makes relative to the number of words that were said. The lower the WER, the more accurate the system.

Equation 1. Word Error Rate formula

$$Word\ Error\ Rate\ = 100\ \times \frac{Insertions\ +\ Substitutions + Deletions}{Total\ Words\ in\ the\ Correct\ Transcript}$$

When evaluating the Normalizer tool, a problem related to using WER as a metric for assessing the normalizer is that in any given sentence – even if we filter only for sentences containing some normalizable expression – most of the tokens are likely to be irrelevant to the normalizer. Thus, WER divides edit distance over all tokens in the reference; however, most tokens we would not expect to be normalized anyway, so applying WER to the Normalizer distorts the results. Instead, using the WERnorm[7] (Word Error Rate over Normalizable tokens) metric, we divide by the number of normalized reference tokens.

Equation 2. Word Error Rate over Normalizable Tokens metric formula

$$WER_{norm}\ = \frac{\sum_{x,y\ \epsilon\ X,Y}\ edit\ distance\ _{x,y}}{\sum_{y\ \epsilon\ Y}\ \max\ (1, normalizable\ tokens_{x,y}\ )}$$

[7] The WERnorm was an evaluation metric proposed and elaborated by the ML team.

Accuracy is also metric for evaluating classification models. It answers the question: "Overall, how often is our model correct?". Thus, Accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally, where 1 represents total accuracy.

Equation 3. Accuracy metric formula

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision is defined as the number of true positives divided by the number of true positives plus false positives. This formula is used to understand the model's accuracy.

Equation 4. Precision metric formula

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall is described as the number of true positives divided by the number of true positives plus false negatives. That is, it calculates the true positives by anything that should have been predicted as positive.

Equation 5. Recall metric formula

$$Recall\ = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 score, also known as F1, is a measure of a model's accuracy on a dataset. F1 score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. A perfect model has an F1 score of 1.

Equation 6. F1 metric formula

$$F1 = \frac{2(Precision * Recall)}{Precision + Recall}$$

6.2. **Normalizer and G2P evaluation**

In the evaluation of version 2 of the normalizer, a set of around 1000 manually normalized and tagged reference sentences were used. The sentences were processed by the normalizer and compared with the reference output. The overview of the evaluation set included the number of sentences, normalizable reference tokens, and total reference tokens used. It can be observed that out of the 23428 reference tokens, only 31.2% were normalizable. All the 1000 sentences had various normalizable tokens, and a corresponding rule was applied to each token. However, the sentences lacked ordinals, resulting in a lower percentage of normalizable ordinals (1.4%) than real numbers (49.9%). This could potentially impact the results.

The normalizer results show that there are statistically significant differences among the rules in both versions of the normalizer. For example, the ordinals rule exhibits the highest percentage in both versions, while abbreviations have the lowest percentage. Real numbers, measurements, currency, time, symbols, and dates show considerable variations in their performance. The analysis of the F1-score results shows a significant performance increase in some of the rules, such as abbreviations and currency. The improvement is higher for abbreviations, currency symbols, and symbols than new real numbers, data formats, or ordinals. Concerning the WER and WERnorm metrics, WERnorm shows that version 2 of the normalizer has, on average, a 4 pp. lower error rate over normalizable tokens compared to version 1. The use of the ordinals and real numbers rule is the largest source of improvement on normalizer WER. Regarding accuracy, version 2 shows an improvement of, on average, 28 pp compared to version 1. The results shown in Table 15 indicate a significant improvement in the normalizer's performance in all the rules.

Table 15. Normalizer's version 1 and version 2 performance regarding WER, WERnorm, and Accuracy metric

Normalizer pt-PT	**WER**	**WERnorm**	**Accuracy**
Normalizer version 1	40,13%	12,47%	46,96%
Normalizer version 2	35,29%	10,58%	74,09%

Regarding the G2P model for Swedish, after doing an analysis of the phonetic lexicon (see Table 16) and the final G2P model, the revision of the lexicon shows that 99% of the orthographic words are correct, with only 1% being incorrect.

Table 16. Swedish phonetic lexicon overview

Swedish Phonetic Lexicon Overview		
	#	**%**
Lexicon entries	25248	100
Corrected words	25067	99
Incorrect words	181	1
Correct transcriptions	22339	88
Incorrect transcriptions	2909	11
WER		11

The most common errors regarding orthographic words are in (1) Double letters (e.g., *u**tt**lardet - u**t**lardet*); (2) Diacritics (e.g., *genéve - genève*). Thus, due to errors in the transcribed audio used to create the initial lexicon, we can observe 1. Segment position alteration - Metathesis (e.g., *vädning - vending*), and (3) Segment suppression in the middle of the word - Syncope (e.g., *lövs**t**edt - löv_tedt*). Consequently, errors in the orthographic words will be followed in the phonetic transcriptions as well.

Regarding phonetic transcriptions, we find 88% of transcriptions are correct and 11% incorrect. The most common errors of incorrect transcription are in the following vowels:

1. Mapping of graphemes <ä, o> - these graphemes have different equivalent phones; however, their occurrence does not always depend on phonological rules (e.g., [**eh** gg oo] – [**ae** gg ug]). /o/ was mostly transcribed as [oo], and /ä/ as [eh] instead of [ae].
2. Vowel reduction in /i/ - weakening of a vowel in an unstressed position (e.g., [hh **ih** ll **ih** ng shx eu] – [hh **ii** ll **ii** ng ss eu]). The generated transcription reduced the vowel /i/ whereas the correct transcription should be [ih] instead of [ii].

Regarding the phonetic lexicon performance, we achieved a WER of 11%. We consider this lexicon to be of good quality and ready to use for the G2P model.

Regarding per phone evaluation, we tested each phone in terms of Precision, recall, and F1-score. We conclude that the best performed phones are consonants [bb, dd, ff, hh, kk, ll, mm, nn, rr, ss, tt, vv]. Phones with a perfect F1 score are the following two: [ff, rr]. Therefore, consonants are performing better than vowels. The phone [ug] has the worst performance with an F1-score of 81%, following [ii] (83%), [uu] (85%), [ah] (86%), and [oo] (87%). The average Precision, Recall, and F1-score present a good result of 96% per phone. Thus, in total, all phones present 97% of accuracy.

Considering the G2P model for Russian, the analysis of the lexicon revision (see Table 17) showed that most of the Russian orthographic words were correct (96%), with only 4% being incorrect.

Table 17. Russian phonetic lexicon overview

Russian Phonetic Lexicon Overview		
	#	**%**
Lexicon entries	27340	100
Corrected words	26286	96
Incorrect words	1054	4
Correct transcriptions	23989	88
Incorrect transcriptions	3351	12
WER		12

We observed the following two errors in orthographic Russian words: (1) Segment suppression in the middle of the word - Syncope (e. g., а_нализа – а**з**нализа), and (2) Diacritics suppression (e. g., взл**е**ты - взл**ё**ты). Consequently, errors in the orthographic words will be followed in the phonetic transcriptions as well.

Regarding phonetic transcriptions, we find 88% of transcriptions are correct and 12% incorrect. The most common errors of incorrect transcriptions are due to:

i. Sonorization - sound change where a voiceless consonant becomes voiced (*e. g.,* [**ss** oo zz dd aa nn ii yye] - [**zz** oo zz dd aa nn ii yye]).

ii. Distinction between vowels and iotated vowels – vowels before <*ж*, *ш,* н> are not iotated (*e. g.,* [aa vv aa nn ss ts **yye** nn uu] - [aa vv aa nn ss ts **ee** nn uu])

iii. Mapping of vowel <и> - its regular equivalent phone is [ii], however it can also be [ie] (*e. g.,* [mm oo ss kk vv **ii**] - [mm oo ss kk vv **ie**]).

We got a reasonable WER of 11% for the phonetic lexicon performance. This lexicon is of good quality and ideal for the G2P model. Regarding per phone evaluation, we tested each phone regarding Precision, Recall, and F1-score. We conclude that the best performed phones are mainly consonants [gg, zx zx, ll, mm, nn, rrj, ts, txj]. Phones with a perfect F1 score are predominantly palatalized consonants [zj, llj, mmj, nnj]. Indeed, consonants are performing better than vowels. The phone [ii] has the worst performance with an F1-score of 83%, following [ie] (85%), iotated vowels [yya, yye, yyo, yyu] (86%), and consonant [zz] (87%). The average Precision, Recall, and F1-score present a good result of 95% per phone. Thus, in total, all phones present 96% of accuracy.

In conclusion, the analysis of the lexicon revision showed that 96% of the words were correct, with 4% of incorrect words mostly due to errors in orthographic Russian words and phonetic transcriptions. The phonetic lexicon performance had a reasonable WER of 12%, and the best-performing phones were predominantly consonants, especially palatalized consonants. The Swedish G2P model showed better performance with a 98% F1-score compared to the Russian model's 96% F1-score, indicating that Swedish orthography is more phonetically based and has a simpler phone set.

7. Conclusion and future work

The present paper describes the work done on the Normalizer and Grapheme-to-Phone models in Speech Technologies, focusing on the importance of linguistic knowledge in preprocessing models. The Normalizer was evaluated, and version 2 showed better performance than version 1. The phonetic lexica for Swedish and Russian were obtained and evaluated, with the Swedish G2P model showing better accuracy than the Russian model. The results obtained led to the expansion and improvement of the Normalizer and G2P models, now

supporting 14 languages. The rule-based approach used in the Normalizer and G2P models increased their accuracy and performance, demonstrating the importance of linguistic knowledge in preprocessing models. The work done on the Normalizer also contributed to discussions on specific vs. language-generic rules. In this sense, we see future possibilities of expanding the coverage of the current normalizer by implementing the real numbers and ordinals rule (best-performed rules) to a module that can be used for various languages. Languages in which real numbers and ordinals are represented similarly (e.g., magnitude/decimal separators, number base). The same could be done with other rules, although this still needs thorough research.

Overall, the study shows the relevance and meaningfulness of having linguistic knowledge in preprocessing models for Speech Technologies.

References

Alasadi, Abdumalik & Ratnadeep Deshmukh (2018) Automatic speech recognition techniques: A review. In *Signal Processing and Computer Vision*, pp. 464–470. Available at https://www.researchgate.net/publication/325296232

Armstrong, Eric & Paul Meier (2005) IPA Chart. Available at https://www.ipachart.com/

Bondarko, Liya (2005) Phonetic and phonological aspects of the opposition of "soft" and "hard" consonants in the modern Russian language. *Speech Communication* 47 (1–2), pp. 7–14. https://doi.org/10.1016/j.specom.2005.03.012

Brasoveanu, Adrian & Dotlacil Jakub (2020) Production-based cognitive models as a test suite for reinforcement learning algorithms. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, pp. 28–37. Available at https://aclanthology.org/2021.cmcl-1.pdf

Errattahi, Rahhal & Asmaa El Hannani (2017) Recent advances in LVCSR: A benchmark comparison of performances. *International Journal of Electrical and Computer Engineering* 7 (6), pp. 3358–3368. http://doi.org/10.11591/ijece.v7i6.pp3358-3368

García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez & Francisco Herrera (2016) Big data preprocessing: Methods and prospects. *Big Data Analytics* 1 (1), pp. 1–22. https://doi.org/10.1186/s41044-016-0014-0

Hennebert, Jean, Martin Hasler & Hervé Dedieu (1994) Neural networks in speech recognition. In *Proceedings of the 6th Microcomputer School of Neural Networks, Theory and Applications*. Swiss Federal Institute of Technology, pp. 23–40. Available at https://www.researchgate.net/publication/2249623

Jurafsky, Daniel & James Martin (2022) *Speech and language processing* (3rd ed.) [Draft]. Stanford University. Available at https://web.stanford.edu/~jurafsky/slp3/

Kurata, Gakuto, Kartik Audhkhasi & Benedict Kingsbury (2019) IBM Research advances in end-to-end speech recognition at INTERSPEECH 2019. *IBM Research Blog*. Available at https://www.ibm.com/blogs/research/2019/10/end-to-end-speech-recognition/ [accessed on 23/12/2022].

Padgett, Jaye (2001) Contrast dispersion and Russian palatalization. In Elizabeth Hume & Keith Johnson (eds.), *The role of speech perception in phonology*. Academic Press, pp. 187–218.

Rajadnya, Kirti (2020) Speech recognition using Deep Neural Network Neural (DNN) and Deep Belief Network (DBN). *International Journal for Research in Applied Science and Engineering Technology*, 8 (5), pp. 1543–1548. https://doi.org/10.22214/ijraset.2020.5359

Riad, Tomas (2014) *The phonology of Swedish*. Oxford University Press.

Shoup, John (1980) Phonological aspects of speech recognition. In Wayne A. Lea (ed.), *Trends in speech recognition*. Prentice-Hall, pp. 125–138.

Timberlake, Alan (2014) *A reference grammar of Russian*. University of California at Berkeley.

Vielzeuf, Valentin & Grigory Antipov (2019) *Are E2E ASR models ready for an industrial usage?*. Cornell University.

Wassink Sophie Groot, Jessica Wingerden van & Poell Rob (2022) Correction: Meaningful work and resilience among teachers: The mediating role of work engagement and job crafting. *PLoS ONE* 17 (5). https://doi.org/10.1371/journal.pone.0269347

Colocação de clíticos em PE L2: Percurso de desenvolvimento e estado final

Alexandra Fiéis[1,2]*, Ana Madeira*[1 2]*, Joana Teixeira*[1,2]

[1]Universidade NOVA de Lisboa, Faculdade de Ciências Sociais e Humanas, Lisboa, Portugal
[2]Universidade NOVA de Lisboa, CLUNL, Lisboa, Portugal

Resumo

Este estudo investiga a aquisição da colocação de clíticos em PE L2, recorrendo a uma tarefa de produção oral induzida e a uma tarefa de juízos de aceitabilidade rápidos. Os participantes são 20 falantes de PE L1 e 30 aprendentes de espanhol L1-PE L2 nos níveis intermédio a quase-nativo. Os resultados mostram que a ênclise estabiliza cedo (pelo menos, na produção) e a próclise se desenvolve sequencialmente, seguindo um percurso semelhante ao observado na aquisição de L1: Negação > Completivas de conjuntivo > Completivas de indicativo > Adverbiais, sujeitos quantificados. Os contextos de desenvolvimento mais tardio são os menos categóricos nas gramáticas nativas, o que pode dar origem a maior variabilidade no input e acrescentar complexidade à tarefa de aquisição. Por isso, os aprendentes necessitam de exposição prolongada a input para descobrir os padrões de colocação de clíticos em PE. A sua aquisição plena parece ser possível, mas apenas no nível quase nativo.

Palavras-chave: colocação de clíticos, português europeu, L2, desenvolvimento, estádio final.

Abstract

This study investigates the acquisition of clitic placement in L2 EP, using an elicited oral production task and a speeded acceptability judgement task. Participants were 20 L1 EP speakers and 30 L1 Spanish-L2 EP adult learners at intermediate to near-native levels. Results show that enclisis stabilizes early (at least in production) and proclisis develops sequentially, following a route similar to that observed in L1 acquisition: Negation > Subjunctive complement clauses > Indicative complement clauses > Adverbial clauses, quantified subjects. The contexts that develop later are the less categorical ones in native grammars, which may give rise to greater input variability and add complexity to the acquisition task. As a result, learners need prolonged input exposure to discover the patterns of clitic placement in EP. Full convergence with the target language seems to be possible, but only at a near-native level.

Keywords: clitic placement, European Portuguese, L2, development, final state.

1. Introdução

Investigação recente tem sugerido que os fenómenos adquiridos tardiamente em língua materna (L1) podem causar dificuldades no desenvolvimento bilingue, principalmente devido a fatores de input (Sorace, 2014; Tsimpli, 2014). A colocação de clíticos em português europeu (PE), que depende de vários fatores (sintáticos, lexicais, semânticos, entre outros), constitui um fenómeno particularmente adequado para investigar esta hipótese, uma vez que os estudos sobre aquisição de L1 constataram que as crianças falantes de PE começam por generalizar a ênclise a contextos de próclise e adquirem conhecimento de alguns desses contextos muito tarde (Costa et al., 2015), o que tem sido relacionado com a variabilidade no input. No domínio de aquisição de língua não materna (L2), estudos preliminares de Gu (2019, 2021, 2022), com falantes nativos de mandarim, sugerem que o percurso de aquisição dos contextos de próclise poderá ser semelhante em PE L1 e

L2. No entanto, não há estudos com falantes quase nativos, não sendo ainda claro se o conhecimento dos contextos de próclise pode ser plenamente adquirido. Assim, é necessária mais investigação para se compreender melhor o percurso de desenvolvimento e o estado final da aquisição deste fenómeno.

O presente estudo investiga a aquisição da colocação de clíticos por falantes nativos de espanhol com nível intermédio, avançado e quase nativo em PE L2, recorrendo a dados de produção induzida e juízos de aceitabilidade.

O artigo está estruturado do seguinte modo: a Secção 2 apresenta uma visão panorâmica dos estudos prévios sobre colocação de clíticos em PE e espanhol e a sua aquisição em PE como L1 e L2; na Secção 3, formulamos as questões de investigação e as predições; a Secção 4 apresenta a metodologia do estudo; os resultados são descritos na Secção 5; e, finalmente, na Secção 6, discutimos os resultados e apresentamos as principais conclusões do estudo.

2. Colocação de clíticos

Nas línguas românicas, os pronomes clíticos, formas fonologicamente fracas, ocorrem sempre em adjacência a um hospedeiro verbal, em ênclise ou em próclise. No entanto, as condições que determinam a colocação pré- ou pós-verbal do clítico relativamente ao verbo variam de língua para língua.

Na próxima Secção, descrevem-se os padrões de colocação de clíticos nas línguas românicas, com especial incidência no PE e no espanhol, as línguas que são alvo de estudo neste trabalho. Nas secções 2.2. e 2.3., descrevem-se alguns estudos sobre a aquisição de colocação de clíticos em PE L1 e L2, respetivamente.

2.1. Na gramática de falantes nativos de PE e de espanhol

Numa língua românica como o francês, os clíticos de objeto são sempre proclíticos (cf. (1)); já em línguas de sujeito nulo, como o espanhol ou o italiano, a colocação dos clíticos depende de um fator morfológico, nomeadamente a finitude: a ênclise ocorre em orações não finitas, enquanto a próclise ocorre em orações finitas (cf. (2), para o italiano, e (3) para o espanhol).[1]

(1) *Pedro m'appelle tous les jours*
Pedro CL-me telefona todos os dias
'O Pedro telefona-me todos os dias'

(2a) *Pedro mi chiama tutti i giorni*
Pedro CL-me telefona todos os dias
'O Pedro telefona-me todos os dias'

(2b) *Pedro ha deciso di chiamarmi*
Pedro AUX decidiu telefonar.CL-me
'O Pedro decidiu telefonar-me'

(3a) *Pedro me llama todos los días*
Pedro CL-me telefona todos os dias
'O Pedro telefona-me todos os dias'

(3b) *Pedro decidió llamarme*
Pedro decidiu telefonar.CL-me
'O Pedro decidiu telefonar-me'

[1] Excetuam-se os contextos de imperativo afirmativo, em que ocorre ênclise nas três línguas.

No PE standard, por seu turno, a colocação proclítica ou enclítica não está estritamente ligada a fatores morfológicos como a finitude (ao contrário do espanhol ou do italiano), podendo observar-se três padrões de colocação dos clíticos em orações finitas: próclise, ênclise e mesóclise.

Assim, observa-se próclise em contextos sintáticos específicos (cf. Duarte & Matos, 2000; entre outros): na presença da negação (4); com sujeitos negativos pré-verbais (5); com alguns advérbios quantificados em posição pré-verbal (*já, também, sempre, só, ainda*…) (6); com alguns sujeitos quantificados em posição pré-verbal (7); em subordinadas finitas introduzidas por complementador (8); em orações com CP lexicalizado em interrogativas e exclamativas-Qu (9); e em orações com constituintes focalizados (10):

(4) O menino não **se** levantou

(5) Ninguém **se** levantou

(6) O menino já **se** levantou (vs. O menino levantou-se já)

(7) Todos os meninos **se** levantaram (vs. Levantaram-se todos os meninos.)

(8a) O menino disse que **se** levantou às 8 horas
(8a') O menino quer que a mãe **se** levante cedo.
(8b') O menino está cansado porque **se** levantou às 8 horas.

(9a) Quem **se** levantou?
(9b) Que cedo **se** levantaram!

(10) Muita coisa **me** contas!

Por outro lado, a ênclise é o padrão que se observa em orações independentes finitas sem proclisadores, como se mostra em (11):

(11a) O menino penteou-**se**.
(11b) O avô chamou a neta e ela abraçou-**o**.

A mesóclise verifica-se em frases sem proclisadores com o verbo no futuro simples e no condicional, como em (12a-b), respetivamente:

(12a) O avô dar-**lhe**-á um presente.
(12b) O avô dar-**lhe**-ia um presente.

Já em orações infinitivas e com complexos verbais a situação é mais complexa, uma vez que, em PE, há subida de clítico, e em algumas situações a cliticização pode ocorrer quer ao verbo finito quer ao verbo não finito. Como estas estruturas não serão alvo de estudo neste trabalho, vamos focar-nos apenas nas orações finitas. Nestas orações, a próclise é determinada não apenas pelo contexto sintático, mas também por outros fatores, nomeadamente lexicais (cf. Martins, 2016).

Alguns trabalhos mostram que em certos contextos de próclise existe variação no PE (cf. Martins, 1994, 2016). Verifica-se que, com a negação, a próclise é categórica, mas que, com sujeitos negativos, há alguma variação entre a posição enclítica ou proclítica. Do mesmo modo, alguns sujeitos quantificados também admitem variação. Em orações completivas, na literatura, há variação entre finitas e não finitas e, dentro das orações finitas, há menos ênclise com conjuntivo do que com indicativo. Finalmente, a própria variação que se observa no estatuto sintático (mais ou menos subordinado) das orações adverbiais explicativas (cf. Lobo, 2003) dá origem a padrões de cliticização menos categóricos.

Já em espanhol, como se mostra acima, a colocação dos clíticos é apenas sensível ao estatuto finito/não-finito dos verbos, situação em que contrasta claramente com o PE.

2.2. **Na aquisição de PE L1**

No que respeita a aquisição de L1, não têm sido observados problemas na colocação de clíticos, na grande maioria das línguas: em italiano (cf. Guasti, 1993; entre outros); em espanhol e catalão (cf. Wexler et al., 2004; entre outros); em francês (cf. Grüter, 2006; Hamann et al., 1996; Pierce, 1992; entre outros); e, finalmente, em grego standard (cf., por exemplo, Marinis, 2000).

Esta situação não será de estranhar, uma vez que, nestas línguas, como vimos, os padrões de cliticização são bastante claros. Contudo, noutras línguas, como o PE (e também o grego cipriota), em que a variação entre ênclise e próclise não depende de finitude, foram identificados problemas na colocação de clíticos (cf. para o PE, Costa et al., 2015; Duarte et al., 1995; e, para o grego cipriota, Neokleous, 2013; Petinou & Terzi, 2002).

Em PE, nos estádios iniciais de aquisição, há uma generalização da ênclise a contextos de próclise (cf. Duarte & Matos, 2000; Duarte et al., 1995), e observa-se uma aquisição mais precoce de próclise em alguns contextos, de acordo com uma escala, situando-se à esquerda os contextos de aquisição mais precoce e à direita os de aquisição mais tardia:

> Negação > Sujeitos Negativos, Completivas (de indicativo) > Advérbios (proclisadores) > Adverbiais (causais com *porque*) > Sujeitos Quantificados (com *todos*).
>
> (Costa et al., 2015)

Segundo Costa et al. (2015), os dados das crianças mostram que os contextos em que existe uma maior dependência de conhecimento lexical (i.e., em que o falante tem de recorrer a conhecimento item a item para determinar se está perante um contexto de próclise ou de ênclise) são adquiridos mais tardiamente. É isto que acontece com os advérbios, os sujeitos quantificados e as orações adverbiais. Note-se que, neste último caso, apenas conhecendo alguns conectores específicos (e.g., *porque* vs. *pois*), é possível ao falante determinar se está perante subordinação adverbial (e, portanto, um contexto de próclise) ou coordenação (e, portanto, um contexto de ênclise) (cf. Lobo, 2003). Crucialmente, como Costa et al. (2015) notam, os contextos em que a posição do clítico está mais dependente de propriedades específicas dos itens lexicais são aqueles em que os próprios adultos exibem alguma variação entre próclise e ênclise. Parece, pois, haver um paralelismo entre o percurso de aquisição e a variação encontrada na gramática do adulto.

O mesmo se observa na aquisição bilingue de PE, cujo percurso de aquisição segue em geral os padrões observados na aquisição monolingue. No entanto, observa-se um desenvolvimento mais lento nos bilingues. Num estudo desenvolvido com crianças falantes de PE como língua de herança com o alemão como L1, Flores e Barbosa (2014), numa tarefa de produção induzida (oral), mostram que as crianças (7–15 anos) falantes de herança apresentam um percurso de desenvolvimento idêntico ao do grupo de controlo (crianças monolingues na mesma faixa etária). Ou seja, em contextos de próclise, as taxas de acerto são mais elevadas com negação do que em subordinadas, e com advérbios e quantificadores, o que parece sugerir que a aquisição dos padrões de colocação dos clíticos é possível, ainda que o seu desenvolvimento possa ser mais tardio do que na aquisição monolingue (possivelmente, devido a input reduzido). Por seu turno, as crianças bilingues de PE/francês estudadas por Casa Nova (2014) e Flores et al. (2016) produzem menos ênclise em contextos de próclise do que os bilingues PE/alemão, mas a diferença não é significativa; e, contrariamente aos outros grupos (controlos monolingues e bilingues PE/alemão), produzem próclise em contextos de ênclise. Este domínio da próclise em contextos neutros, segundo as autoras, poderá resultar de transferência inicial do francês, que só tem próclise, o que terá influência no ritmo de desenvolvimento. Já Tomaz et al. (2019), também com crianças bilingues PE/francês, apresentam resultados no mesmo sentido: embora o percurso de desenvolvimento seja semelhante ao dos monolingues, as crianças bilingues PE/francês produzem taxas mais elevadas de próclise, tanto em

contextos de próclise como de ênclise, o que pode, mais uma vez, resultar de transferência do francês (embora existam diferenças individuais).

Em suma, parece haver indícios de que a colocação de clíticos em português é uma área vulnerável na aquisição quer monolingue quer bilingue.

2.3. **Na aquisição de PE L2**

Mostrámos, na Secção 2.1., que, ao contrário do que acontece em línguas como o espanhol e o italiano, em que a colocação enclítica ou proclítica está relacionada com finitude, não criando obstáculos à aquisição dos padrões alvo na L1, no PE, a posição do clítico depende de uma diversidade de fatores (em particular, fatores sintáticos, lexicais e semânticos), o que contribui para tornar mais complexa a aquisição deste fenómeno. Verifica-se, assim, que o conhecimento dos contextos de próclise constitui uma dificuldade na aquisição de PE L1, como tem sido demonstrado pelos estudos descritos em 2.2. Nos últimos anos, tem sido proposto que os fenómenos adquiridos tardiamente em L1 podem ser de difícil aquisição no desenvolvimento bilingue (Tsimpli, 2014). Dadas as suas características e o que sabemos sobre a sua aquisição em L1, os padrões de colocação de clíticos em PE constituem um fenómeno particularmente adequado para investigar esta hipótese.

Tsimpli (2014) defende que, tanto na aquisição de L1 como na aquisição bilingue por crianças, os fenómenos de aquisição precoce são aqueles que são estritamente sintáticos (como é o caso da colocação de clíticos em espanhol e de alguns contextos de colocação de clíticos em PE, como, por exemplo, a negação), enquanto os fenómenos de aquisição tardia ou muito tardia envolvem interfaces com outras componentes linguísticas como, por exemplo, a semântica ou o léxico (como é o caso de alguns contextos de próclise em PE, tais como as frases com sujeitos pré-verbais quantificados). Estes fenómenos são particularmente vulneráveis a efeitos de input (Sorace, 2014; Tsimpli, 2014). A questão que se coloca, então, é se, na aquisição de PE L2 por adultos, o *timing* de aquisição de diferentes propriedades linguísticas coincide com o que tem sido observado em PE L1 e se depende também da natureza das propriedades em questão.

Estudos anteriores sugerem que, à semelhança do que acontece na aquisição monolingue e bilingue de PE L1 (Costa et al., 2015; Flores & Barbosa, 2014; Flores et al., 2016; Tomaz et al., 2019), a ênclise é adquirida mais cedo do que a próclise na aquisição de PE L2 por adultos; observa-se também uma generalização inicial de ênclise a contextos de próclise, seguida de um desenvolvimento gradual do conhecimento destes contextos, sendo o contexto de negação aquele em que a próclise estabiliza mais cedo (Gu, 2019, 2021, 2022; Madeira et al., 2006; Madeira & Xavier, 2009). Embora este percurso de desenvolvimento pareça não ser significativamente afetado pela L1 dos aprendentes, há indícios de que esta poderá ter um efeito no ritmo de desenvolvimento (Madeira et al., 2006; Madeira & Xavier, 2009). Num estudo exploratório realizado com falantes nativos de mandarim (uma língua sem clíticos pronominais), Gu (2019, 2021) confirma, com base em dados de juízos de aceitabilidade sem pressão de tempo, a existência de variação no desenvolvimento das propriedades que determinam a colocação dos clíticos nos diferentes contextos de próclise e identifica a seguinte escala de aquisição de próclise: negação > frase com advérbio (*também*) / oração adverbial causal (com *porque*) > frase com sujeito quantificado (*todos*). Embora este trabalho não inclua todos os contextos testados nos estudos sobre PE L1, a escala proposta coincide no essencial com a que é observada na aquisição de L1. A principal diferença reside na ordem dos contextos intermédios da escala: no estudo de Gu (2019, 2021), não é clara a ordem por que estabiliza a próclise nos contextos de frase com advérbio e oração adverbial (embora os resultados de Gu (2022), baseados numa tarefa de juízos de aceitabilidade rápidos, pareçam indicar que esta estabilização poderá ocorrer mais cedo nos contextos de frase com advérbio do que nas orações adverbiais).

Também não é claro nos estudos de Gu se a preferência por próclise em contexto de sujeito quantificado é totalmente adquirível.

Tal como acontece na aquisição de PE L1, os atrasos[2] no desenvolvimento do conhecimento da colocação dos clíticos em alguns contextos em PE L2 poderão dever-se à variação que está associada a itens lexicais específicos (Costa et al., 2015), o que explicaria as semelhanças nas sequências de aquisição dos contextos de próclise em L1 e L2 observadas nos estudos realizados até ao momento. Contudo, alguns destes estudos apresentam problemas metodológicos, sobretudo no que se refere à aferição dos níveis de proficiência dos participantes,[3] o que compromete as conclusões que se retiram quanto a sequências de aquisição. Por este motivo, é necessário realizar mais estudos adotando critérios mais rigorosos para a identificação do nível de proficiência dos participantes (cf. Secção 4.1).

Os estudos sobre a aquisição da colocação de clíticos em PE L2 por adultos têm-se debruçado predominantemente sobre o percurso de desenvolvimento deste fenómeno. Embora todos mostrem que alguns contextos são adquiridos mais cedo do que outros, nenhum estudo inclui falantes quase nativos, pelo que não sabemos se é possível desenvolver conhecimento pleno dos contextos que parecem ser de aquisição mais tardia. Tendo em conta que estes contextos envolvem propriedades internas à gramática, i.e. interfaces internas, como léxico-sintaxe (e.g., no caso de sujeitos quantificados), a Hipótese de Interface (HI) (Sorace, 2014; Sorace & Filiaci, 2006) prediz que serão plenamente adquiríveis em L2, mesmo que estejam sujeitos a atrasos de desenvolvimento.

3. Questões de investigação e predições

Neste trabalho, procuraremos responder às seguintes questões de investigação:

QI.1. Os contextos de ênclise são adquiridos cedo em PE L2, tal como em PE L1?

QI.2. O percurso de desenvolvimento dos contextos de próclise em PE L2 assemelha-se ao encontrado na aquisição de PE L1?

QI.3. A convergência com a língua alvo é possível em PE L2 no que diz respeito à colocação dos clíticos em contextos de próclise?

Considerando os resultados dos estudos anteriores sobre o percurso de desenvolvimento dos padrões de colocação dos clíticos em PE descritos na Secção 2, bem como as predições da HI quanto ao estádio final de aquisição em L2, fazemos as seguintes predições em relação às questões de investigação:

P.1. Se, à semelhança do que acontece em PE L1, os contextos de ênclise forem adquiridos cedo em PE L2, os aprendentes terão já conhecimento destes contextos no nível intermédio.

P.2. Se o percurso de desenvolvimento dos contextos de próclise em PE L2 se assemelhar ao que tem tem sido observado na aquisição de PE L1, os aprendentes desenvolverão conhecimento destes contextos de forma gradual: adquirirão primeiro os contextos de aquisição mais precoce e mais tarde os de mais difícil aquisição, de acordo com a escala seguinte (baseada em Costa et al., 2015):

[2] Um avaliador anónimo questiona se se deverá falar de um atraso na aquisição nos falantes L2 em contextos em que há variação, dado que os falantes nativos monolingues também permitem próclise e ênclise nesses contextos. No entanto, é importante notar que, ainda que os falantes nativos permitam ênclise nestes contextos, têm uma preferência significativa por próclise, como mostram os resultados do presente estudo (cf. Secção 5) e do trabalho de Costa et al. (2015). Consideramos, por isso, que existe um atraso no desenvolvimento da L2 apenas quando os aprendentes exibem preferência por ênclise ou não têm preferência clara.

[3] Por exemplo, Madeira et al. (2006) dividem os participantes em função do seu tempo de aprendizagem do PE, estabelecendo uma correlação implícita entre duração da aprendizagem e nível de proficiência, e Madeira e Xavier (2009) assumem o nível de proficiência do curso de português que os participantes estavam a frequentar no momento do estudo.

Negação > Completivas > Adverbiais finitas > Sujeito quantificado.

P.3. Se a convergência com a língua alvo no que diz respeito à colocação dos clíticos em contextos de próclise for possível em PE L2, o padrão de próclise será adquirido em todos os contextos, incluindo os que estão sujeitos a atrasos de desenvolvimento, pelo menos no nível quase nativo.

4. **Metodologia**

Para responder a estas questões, este estudo investiga a aquisição da colocação de clíticos (reflexos) em PE L2 por falantes de espanhol L1 com recurso a duas tarefas experimentais – uma tarefa de produção induzida e uma tarefa de de juízos de aceitabilidade rápidos. Nos pontos seguintes, descrevem-se os participantes do estudo, as tarefas e os procedimentos adotados na análise estatística.

4.1. **Participantes**

Participaram no estudo 30 falantes nativos de espanhol, com os seguintes níveis de proficiência em PE L2: intermédio (n = 10), avançado (n = 10) e quase nativo (n = 10). Todos eles eram filhos de falantes monolingues de espanhol e tinham o espanhol como a sua única L1. Tinham entre 20 anos e 61 anos de idade, e eram aprendentes adultos de PE, tendo na sua maioria começado a estar expostos ao PE a partir dos 16 anos (apenas um participante declarou ter começado a aprender português aos 11 anos). Dos 30 participantes, 21 estavam a viver em Portugal à data do estudo ou já tinham vivido num país de língua portuguesa.

O estudo incluiu ainda um grupo de controlo de falantes nativos de PE, com idades entre os 19 e os 54 anos. Todos eram filhos de falantes monolingues de PE, tinham apenas o PE como sua L1 e tinham residido em Portugal durante toda a sua vida.

As características sociolinguísticas dos participantes (recolhidas através de um questionário de perfil sociolinguístico administrado através do *Google Forms*, que os participantes preenchiam no momento da inscrição no estudo) são apresentadas na Tabela 1.

Tabela 1. Dados sobre os participantes

Grupo	**N**	**Idade**		**Idade de início de exposição ao PE**		**Anos de residência em países de língua portuguesa**	
		M	**Desvio Padrão**	**M**	**Desvio Padrão**	**M**	**Desvio Padrão**
Quase nativo	10	41,4	7,31	20,75	2,94	9,35	7,51
Avançado	10	38,6	10,69	27,6	11,04	2,08	2,39
Intermédio	10	42,2	11,42	35,3	11,34	2,59	3,79
Nativo	20	29,55	10,82	n/a	n/a	n/a	n/a

Na avaliação do nível de proficiência dos falantes não nativos de PE L2 foi utilizada uma versão adaptada do procedimento usado em Sorace e Filiaci (2006) (concebido por White & Genesee, 1996) para identificação de falantes quase nativos. Foram realizadas entrevistas individuais com todos os participantes, com o objetivo de provocar produção oral espontânea a partir de *cartoons* (veja-se o exemplo na Figura 1). Pedia-se ao participante que falasse durante cerca de 3 minutos, respondendo a duas perguntas sobre cada *cartoon*: O que vê no *cartoon*? Qual é a mensagem do *cartoon*?

Figura 1. Exemplo de cartoon[4]

Selecionou-se depois uma amostra da produção de cada participante com cerca de 1,5 minuto, que foi avaliada por três falantes nativos de PE que tinham alguma formação em linguística com base nos seguintes critérios:

morfologia, sintaxe, vocabulário, pronúncia, fluência e impressão geral.[5] Na grelha de avaliação, cada critério estava associado a uma linha contínua de 9 cm, com a designação "não nativo" na extremidade esquerda e "nativo" na extremidade direita, e pedia-se ao avaliador que assinalasse com uma cruz o grau de proximidade da amostra ao nível nativo. Para se garantir que os avaliadores assumiam como ponto de referência as produções de falantes nativos de PE, procedeu-se a uma mistura aleatória de amostras de entrevistas realizadas com falantes nativos de PE com as amostras dos aprendentes de PE L2.

No final, as avaliações foram transformadas em valores discretos, através da sobreposição de um acetato com uma linha de 9 cm dividida numa escala de 18 pontos (1 ponto por cada 0.5 cm), e o nível de proficiência de cada aprendente foi apurado com base na média de pontos atribuídos pelos avaliadores. Os aprendentes de PE L2 a quem os avaliadores atribuíram entre 17 e 18 pontos nos critérios *sintaxe*, *morfologia* e *vocabulário* e 16 ou mais pontos nos restantes critérios, com o máximo de uma exceção (as exceções ocorreram, tipicamente, no critério *pronúncia*), foram considerados quase nativos; os que não obtiveram pontuação para serem classificados como quase nativos e receberam, pelo menos, 15 pontos nos critérios *sintaxe*, *morfologia* e *vocabulário* e 13 ou mais pontos nos outros critérios, com o máximo de uma exceção, foram classificados como avançados; e, finalmente, aqueles que não tinham pontos suficientes para serem classificados como avançados e obtiveram, pelo menos, 13 pontos nos critérios *sintaxe*, *morfologia* e *vocabulário* e 10 ou mais pontos nos restantes critérios, com o máximo de uma exceção, foram considerados intermédios.

4.2. **Tarefas**

Tendo em conta que a colocação dos clíticos é alvo de ensino explícito, foram utilizadas duas tarefas que forçam os participantes a usar primordialmente o seu conhecimento implícito (e.g., Ellis, 2005): uma tarefa de produção oral induzida e uma tarefa de juízos de aceitabilidade rápidos. Foram utilizados apenas clíticos reflexos, a fim de assegurar maior comparabilidade com os estudos de aquisição de PE L1, que usam clíticos reflexos por serem aqueles com que se registam taxas mais baixas de omissão (Costa & Lobo, 2007; Silva, 2009). As tarefas foram administradas por ordem aleatória em momentos diferentes, separadas por um intervalo mínimo de uma semana.

A tarefa de produção induzida consistia numa tarefa de completamento de frases, administrada através da plataforma Gorilla. Para cada item, apresentava-se uma frase incompleta por escrito, acompanhada de uma imagem e algumas palavras. O participante devia ler a frase em voz alta e completá-la de acordo com a imagem.

[4] https://politicalcartoons.com/cartoon/130449

[5] Seguindo a grelha de avaliação usada por Antonella Sorace (c.p.), nas instruções fornecidas aos avaliadores, foram exemplificados os aspetos que deveriam ter em consideração em cada um dos critérios, como se segue: sintaxe (e.g., ordem de palavras); morfologia (e.g., correção da flexão dos verbos, adjetivos e nomes); vocabulário (e.g., adequação/correção das palavras usadas); pronúncia (e.g., sotaque, correção e clareza da pronúncia); fluência (e.g., ritmo de fala, facilidade de elocução); impressão geral (e.g., até que ponto o falante fala bem português, com base nos critérios acima).

Para tal, pedia-se que usasse as palavras fornecidas, podendo usar, adicionalmente, outras palavras que considerasse necessárias. Na Figura 2, mostra-se um exemplo de item de teste.

Figura 2. Exemplo de item do teste de produção oral induzida

A tarefa de produção induzida testou a ênclise e a próclise em 6 condições: frase finita sem proclisador, frase com negação, completiva de indicativo, completiva de conjuntivo, oração adverbial causal (introduzida por *porque*) e frase com sujeito quantificado (com o quantificador *todo/a*). A tarefa incluiu 4 itens por condição, num total de 24 itens experimentais, e 24 distratores. Apresentamos exemplos de itens de teste nas 6 condições na Tabela 2.

Tabela 2. Itens de exemplo da tarefa de produção induzida

Condição	**Exemplo de item**
Frase finita sem proclisador	O avô deu o pente à neta e ela… penteou-se na casa de banho
Frase com negação	O pai deu o pente à filha, mas ela não… se penteou bem
Completiva de indicativo	A menina parecia despenteada, mas o pai viu que ela… se penteou na casa de banho
Completiva de conjuntivo	A menina está despenteada e o pai quer que ela … se penteie com o pente
Oração adverbial	O pai elogiou a filha porque ela … se penteou com o pente
Frase com sujeito quantificado	O assistente trouxe pentes e toda a gente… se penteou no camarim

A tarefa de juízos de aceitabilidade rápidos foi construída com o *software* Psychopy e administrada na plataforma Pavlovia. Em cada item da tarefa, aparecia primeiro um ponto de fixação durante 1500 ms e, em seguida, a frase era apresentada no centro do ecrã palavra por palavra, de forma não cumulativa, a um ritmo de 450 ms por palavra (de acordo com o procedimento habitual nestas tarefas; ver, por exemplo, Bader & Häussler, 2010, e Hopp, 2007). Após a última palavra, pedia-se ao participante que avaliasse o grau de naturalidade da frase numa escala de 1 (nada natural) a 5 (totalmente natural). O participante tinha a opção de não responder, carregando na tecla 'N' para indicar que não sabia a resposta.

A tarefa de juízos tinha um desenho 6 x 2, cruzando as variáveis *tipo de contexto* (os 6 contextos testados na tarefa de produção induzida) e *tipo de colocação do clítico* (ênclise vs. próclise). A tarefa incluía 3 itens por condição, num total de 36 itens experimentais, e 36 distratores. Apresentam-se abaixo dois exemplos de itens de teste.

(13a) Completiva de conjuntivo + ênclise:

A menina está suja e o avô quer que ela limpe-se com a toalha

(13b) Adverbial + próclise:
O treinador repreendeu a jogadora porque ela se sentou no relvado

4.3. **Análise estatística**

A análise estatística dos dados de cada grupo foi realizada com modelos de efeitos mistos, em que foram incluídos como efeitos aleatórios as variáveis *participantes* e *itens* e como efeito fixo a variável *posição do clítico*. Seguindo Cunnings (2012) e Linck e Cunnings (2015), as análises incluíram interceções aleatórias para *participantes* e *itens* e declives aleatórios por participante para a variável *posição do clítico*.

A análise estatística foi conduzida em R, usando o pacote *lme4*. Mais especificamente, foi usada a função *lmer* (modelo misto linear) para a análise dos resultados da tarefa de juízos de aceitabilidade e a função *glmer* (modelo misto linear generalizado), com a especificação "family=binomial", para a análise dos dados da tarefa de produção oral, uma vez que esta produz resultados binários**.** Na tarefa de produção, para cada nível do efeito fixo (ênclise e próclise), as respostas dos participantes foram codificadas como 'produz o clítico na posição de ênclise/próclise' = 1 e 'não produz o clítico na posição de ênclise/próclise' = 0. As respostas sem clítico adjacente ao verbo finito ou em que o contexto sintático em teste foi alterado pelo participante não foram consideradas na análise.

Como a função *glmer* gera output com valores de *p* e a função *lmer* apenas gera valores de *t*, usámos como medidas de significância os valores de *p* e *t*. Como é habitual na literatura (cf. Linck & Cunnings, 2015), um efeito fixo foi considerado estatisticamente significativo (indicado com * nas tabelas) sempre que *p* é inferior ou igual a .05 ou o valor absoluto de *t* é superior ou igual a 2.00.

5. **Resultados**

Nos pontos seguintes, apresentam-se os resultados das tarefas de produção oral induzida e de juízos de aceitabilidade rápidos, com recurso a estatística descritiva e inferencial.

5.1. **Tarefa de produção oral induzida**

Na tarefa de produção oral, a maioria dos participantes produziu respostas com o clítico *se* adjacente a um verbo finito, recorrendo aos tipos de contextos sintáticos que os estímulos pretendiam induzir. No entanto, registaram-se também outros tipos de respostas, nomeadamanente: uso de DP em vez de clítico, como em (14); omissão de clítico, como em (15); alteração do contexto sintático que o estímulo pretendia induzir (e.g., introdução do proclisador *não* em contexto de completiva de indicativo, como em (16)); uso de uma formulação que não requer clítico, como em (17); produção de clítico adjacente a verbo não finito, como em (18); e ausência de resposta. Em cada grupo, a percentagem destas outras respostas é baixa, variando entre 0% e 1.3% por condição experimental no grupo de controlo, entre 3% e 15% no grupo quase nativo, entre 0% e 13% no grupo avançado, e entre 13% e 20% no grupo intermédio. Neste estudo, apenas consideramos as respostas com clítico adjacente ao verbo finito e em que o contexto sintático em teste não foi alterado pelo participante.

(14) A menina está suja e o pai quer que ela limpe o corpo com uma toalha. [AV]

(15) A avó está cansada e o neto quer que ela sente na cadeira. [INT]

(16) A avó trouxe um banco, mas o tio viu que ela não se sentava na cadeira. [QN]

(17) O treinador repreendeu a jogadora porque ela estava sentada no relvado. [QN]

(18) A menina parecia suja, mas o avô viu que ela estava a se limpar com a toalha. [AV]

Como mostra a Figura 3, na tarefa de produção oral induzida, tanto o grupo de falantes nativos de PE como todos os grupos de aprendentes exibem preferência por ênclise em contexto de frase declarativa simples sem proclisador (ênclise vs. próclise: *p*s ≤ .0298; para a análise estatística completa, ver Tabela 3). Nos contextos de negação e nas completivas de conjuntivo, todos os grupos preferem claramente próclise a ênclise (*p*s ≤ .0411). Nos restantes contextos testados, enquanto os falantes nativos revelam uma preferência por próclise (*p*s < .001), o desempenho dos aprendentes de PE L2 varia de acordo com o nível de proficiência. No nível intermédio, exibem opcionalidade entre ênclise e próclise nas completivas de indicativo (*p* = .316) e têm uma clara preferência por ênclise em subordinadas adverbiais e em frases com sujeito quantificado (*p*s ≤ .0277). No nível avançado, a produção de próclise em completivas de indicativo aumenta, ficando a diferença entre próclise e ênclise próxima de significância estatística (*p* = .0798). Com adverbiais e sujeitos quantificados, neste nível, os aprendentes passam a produzir taxas semelhantes de ênclise e próclise (*p*s ≥ .869). Finalmente, no nível quase nativo, os aprendentes de PE L2 convergem com a língua alvo, revelando preferência por próclise em todos os contextos de próclise (*p*s ≤ .0193).

Figura 2. % de produção de ênclise e próclise na tarefa de produção oral induzida

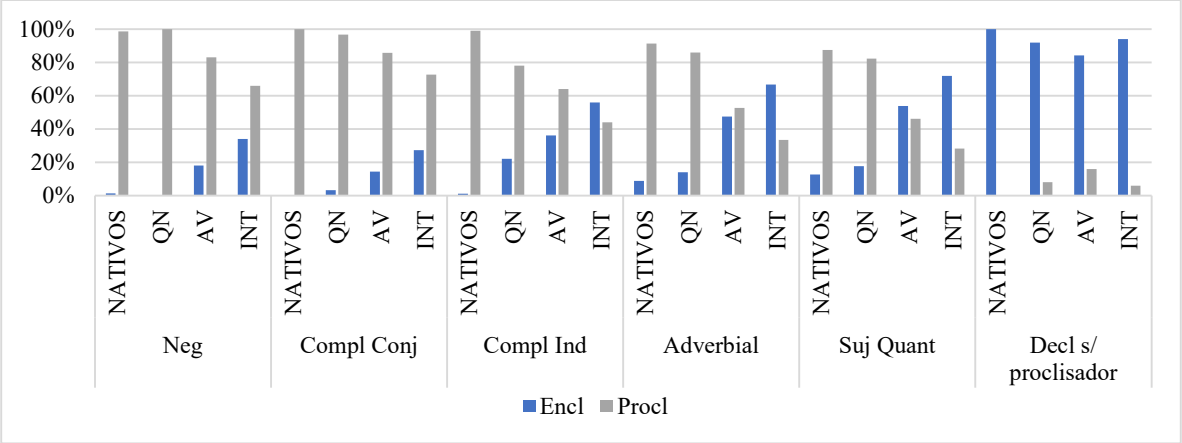


Tabela 3. Diferença entre ênclise e próclise por contexto e grupo na tarefa de produção oral induzida

Contexto	Grupo	Estimativa	Erro padrão	*p*
Frase s/ proclisador	Controlo	-72.09981	15.00119	<.001*
	QN	-6.00570	1.90485	.00162*
	AV	-6.5090	2.2219	<.001*
	INT	-19.363795318	8.910306643	.0298*
Negação	Controlo	64.888900	.002332	<.001*
	QN	4.91020	.85088	<.001*
	AV	22.74266673	8.12633473	.00513*
	INT	19.372507	6.604711	.00336*
Completiva de conjuntivo	Controlo	65.241159	.002785	<.001*
	QN	64.69914140	9.24102190	<.001*
	AV	6.25602	2.72708	.0218*
	INT	5.17673090163	2.53473644180	.0411*
Completiva de indicativo	Controlo	69.06667	.00231	<.001*
	QN	4.24970112387	1.81590816469	.0193*
	AV	4.0714	2.3241	.0798.
	INT	1.65622288146	1.65085362051	.316
Adverbial	Controlo	21.517311172	6.076380768	<.001*
	QN	45.873	2.797	<.001*
	AV	.2801211316	1.6941658433	.869
	INT	-21.69009921	7.05840149	.00212*
Sujeito quantificado	Controlo	40.578148	8.550805	<.001*
	QN	4.68101352043	1.79504916359	.00911*
	AV	.7150556866	4.5967390671	.876
	INT	-6.4045	2.9088	.0277*

Consideremos agora os resultados individuais dos falantes nativos e dos aprendentes de PE nos contextos de próclise. Nas Tabelas 4 e 5, para cada contexto, apresentamos o número de itens com próclise (= P) e ênclise (= E). Os participantes estão ordenados de acordo com o seu desempenho: dos que apresentam taxas mais elevadas de próclise aos que apresentam taxas mais baixas.

Os resultados individuais dos falantes nativos (Tabela 4) mostram que 15 têm próclise categórica, 3 têm casos residuais de ênclise, mantendo uma clara preferência por próclise em todos os contextos, e só 2 não exibem preferência por próclise em todos os contextos. Um destes dois falantes (participante 19) produz percentagens idênticas de ênclise e próclise em orações adverbiais. O outro (participante 20) exibe uma preferência categórica por ênclise com adverbiais e sujeitos quantificados. Em todos os outros contextos, estes falantes preferem próclise. Assim, é apenas com orações adverbiais e sujeitos quantificados que se encontra variação nas preferências dos falantes nativos relativamente à colocação do clítico.

Tabela 4. Número de frases com ênclise e próclise em contextos de próclise por falante nativo de PE

Participante	Negação		Completiva de conjuntivo		Completiva de indicativo		Adverbial		Sujeito quantificado		Total	
	E	P	E	P	E	P	E	P	E	P	E	P
1	0	4	0	4	0	4	0	4	0	4	**0**	**20**
2	0	4	0	4	0	4	0	4	0	4	**0**	**20**
3	0	4	0	4	0	4	0	4	0	4	**0**	**20**
4	0	4	0	4	0	4	0	4	0	4	**0**	**20**
5	0	4	0	4	0	4	0	4	0	4	**0**	**20**
6	0	4	0	4	0	4	0	4	0	4	**0**	**20**
7	0	4	0	4	0	4	0	4	0	4	**0**	**20**
8	0	4	0	4	0	4	0	4	0	4	**0**	**20**
9	0	4	0	4	0	4	0	4	0	4	**0**	**20**
10	0	4	0	4	0	4	0	4	0	4	**0**	**20**
11	0	4	0	4	0	4	0	4	0	4	**0**	**20**
12	0	4	0	4	0	4	0	4	0	4	**0**	**20**
13	0	4	0	4	0	4	0	4	0	4	**0**	**20**
14	0	4	0	4	0	4	0	4	0	4	**0**	**20**
15	0	4	0	4	0	4	0	4	0	4	**0**	**20**
16	0	4	0	4	0	4	1	3	0	4	**1**	**19**
17	0	4	0	4	0	4	0	4	1	3	**1**	**19**
18	0	4	0	3	0	4	0	4	1	3	**1**	**18**
19	1	3	0	4	0	4	2	2	0	4	**3**	**17**
20	0	4	0	4	1	3	4	0	4	0	**9**	**11**

Examinemos agora os resultados individuais dos aprendentes de L2 em contextos de próclise (Tabela 5). No grupo intermédio, só 4 aprendentes apresentam próclise maioritária. Dos restantes aprendentes, 4 ainda têm ênclise dominante, 1 tem taxas quase idênticas de ênclise e de próclise, e 1 aprendente quase não produz clíticos (5 ocorrências em 20 possíveis). Crucialmente, verificamos que a negação e as orações completivas de conjuntivo são os contextos que apresentam mais ocorrências de clítico em posição proclítica no nível intermédio. No nível avançado, o número de aprendentes com próclise dominante aumenta para 6. Embora o uso de próclise aumente em todos os contextos, em orações adverbiais e frases com sujeitos quantificados, a ênclise continua a ser o padrão de colocação do clítico usado por cerca de metade dos aprendentes. No nível quase nativo, a próclise é o padrão dominante para todos os aprendentes, exceto o participante 10, que produz taxas idênticas de ênclise e próclise. Verificamos, assim, que os resultados individuais convergem com os resultados por grupo, mostrando que a próclise é adquirida mais cedo em certos contextos (especialmente o de negação) do que noutros (especialmente adverbiais e sujeitos quantificados).

Tabela 5. Número de frases com ênclise e próclise em contextos de próclise por aprendente de PE L2

Participante		Negação		Completiva de conjuntivo		Completiva de indicativo		Adverbial		Sujeito quantificado		Total	
		E	P	E	P	E	P	E	P	E	P	E	P
INT	**1**	0	4	0	4	0	3	0	3	0	4	**0**	**18**
	2	0	4	0	4	1	3	0	4	3	1	**4**	**16**
	3	0	4	0	4	2	2	2	1	1	3	**5**	**14**
	4	0	4	0	3	1	3	0	3	4	0	**5**	**13**
	5	0	3	0	4	2	2	4	0	3	1	**9**	**10**
	6	0	3	1	1	2	1	4	0	2	0	**9**	**5**
	7	4	0	2	2	3	0	4	0	4	0	**17**	**2**
	8	4	0	3	1	2	0	4	0	3	0	**16**	**1**
	9	4	0	3	1	4	0	3	0	1	0	**15**	**1**
	10	0	1	0	0	1	0	1	0	2	0	**4**	**1**
AV	**1**	0	4	0	4	0	4	0	4	0	4	**0**	**20**
	2	0	4	0	4	0	4	0	4	0	4	**0**	**20**
	3	0	4	0	4	0	4	1	3	0	4	**1**	**19**
	4	0	4	0	4	1	3	1	3	0	4	**2**	**18**
	5	0	4	0	4	0	3	0	3	3	1	**3**	**15**
	6	0	4	0	3	1	3	2	2	4	0	**7**	**12**
	7	0	4	0	3	1	2	3	1	4	0	**8**	**10**
	8	0	4	1	0	4	0	3	0	3	0	**11**	**4**
	9	4	0	1	3	2	0	4	0	4	0	**15**	**3**
	10	3	1	3	1	4	0	4	0	3	1	**17**	**3**
QN	**1**	0	4	0	4	0	4	0	4	0	4	**0**	**20**
	2	0	4	0	4	0	4	0	4	0	4	**0**	**20**
	3	0	4	0	4	0	4	0	3	0	4	**0**	**19**
	4	0	4	0	4	0	4	0	3	0	4	**0**	**19**
	5	0	4	0	4	0	3	0	4	0	4	**0**	**19**
	6	0	4	0	4	0	4	1	3	0	4	**1**	**19**
	7	0	4	0	4	2	1	0	4	0	4	**2**	**17**
	8	0	3	0	4	2	1	0	4	1	3	**3**	**15**
	9	0	3	1	3	1	2	0	4	1	3	**3**	**15**
	10	0	4	1	1	2	1	3	0	3	1	**9**	**7**

5.2. **Tarefa de juízos de aceitabilidade**

Como a Figura 4 mostra, na tarefa de juízos de aceitabilidade, só o grupo de falantes nativos de PE tem preferência por ênclise e rejeita próclise em contexto de frase sem proclisador (ênclise vs. próclise: t = -12.00; para a análise estatística completa, ver Tabela 6). Contrariamente ao que se observou na tarefa de produção induzida, nesta tarefa, os aprendentes de L2 aceitam quer ênclise quer próclise em frases simples sem proclisador. Mesmo o grupo quase nativo, apesar de aceitar significativamente mais ênclise do que próclise neste contexto (t = -2.046), não rejeita a colocação do pronome em posição proclítica (média de aceitação = 4), divergindo, assim, do grupo de controlo.

No contexto de negação, todos os grupos exibem uma clara preferência por próclise (ts ≥ 2.814). Nos restantes contextos de próclise considerados, o grupo de controlo revela uma preferência consistente por próclise em detrimento de ênclise (ts ≥ 6.11). Já as preferências dos grupos de L2 variam segundo o nível de proficiência. No nível intermédio, os aprendentes aceitam próclise ligeiramente mais do que ênclise em completivas de conjuntivo, estando a diferença entre estes dois padrões de colocação próxima de significância estatística (t = 1.906). Em completivas de indicativo, adverbiais e frases com sujeito quantificado, o grupo

intermédio exibe taxas de aceitação de ênclise e de próclise idênticas (*t*s ≤ 1.098), divergindo, deste modo, do grupo de controlo. No nível avançado, os aprendentes passam a ter uma clara preferência por próclise em orações completivas (*t*s ≥ 3.710), independentemente do modo em que estejam. No entanto, continuam a exibir taxas semelhantes de aceitação de ênclise e de próclise em adverbiais e com sujeitos quantificados (*t*s ≤ .632), tal como o grupo intermédio. Por fim, no nível quase nativo, os aprendentes passam a exibir uma clara preferência por próclise em todos os contextos de próclise testados (*t*s ≥ 2.742), convergindo, assim, com o grupo de controlo.

Figura 3. Média de aceitação de ênclise e próclise na tarefa de juízos de aceitabilidade (escala 1 a 5)

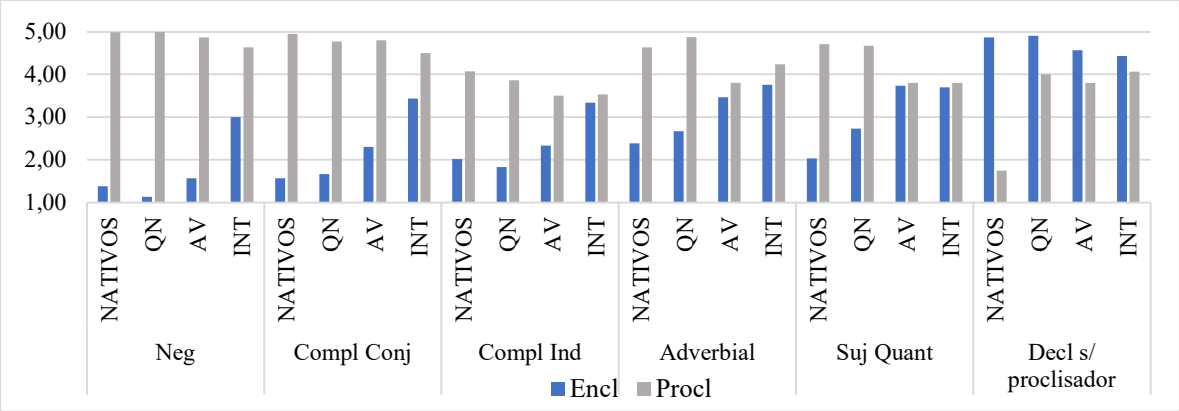


Tabela 6. Diferença entre ênclise e próclise por contexto e grupo na tarefa de juízos de aceitabilidade

Contexto	Grupo	Estimativa	Erro padrão	*t*
Frase s/ proclisador	Controlo	-3.1167	.2597	-12.00*
	QN	-.9000	.4398	-2.046*
	AV	.06667	.64845	.103
	INT	-.3667	.2315	-1.584
Negação	Controlo	3.60000	.13933	25.84*
	QN	3.86667	.14298	27.04*
	AV	3.3000	.3989	8.272*
	INT	1.6333	.5804	2.814*
Completiva de conjuntivo	Controlo	3.38333	.18878	17.92*
	QN	3.1000	.3494	8.871*
	AV	2.5000	.5260	4.753*
	INT	1.0667	.5595	1.906.
Completiva de indicativo	Controlo	2.0507	.2875	7.134*
	QN	2.0309	.5474	3.710*
	AV	1.1667	.5605	2.081*
	INT	.2000	.6591	.303
Adverbial	Controlo	2.2500	.3682	6.11*
	QN	2.3000	.8388	2.742*
	AV	.3333	.5271	.632
	INT	.4847	.4416	1.098
Sujeito quantificado	Controlo	2.6755	.3545	7.547*
	QN	1.9333	.6061	3.19*
	AV	.06667	.64845	.103
	INT	.1000	.3788	.264

Consideremos agora os resultados individuais dos falantes nativos e dos aprendentes de L2 nos contextos de próclise. Nas Tabelas 7 e 8, para cada contexto, apresentamos a mediana de aceitação de próclise (= P) e de ênclise (= E). De modo a facilitar a leitura das tabelas, os participantes estão ordenados de acordo com o seu desempenho, estando no topo os que apresentam medianas mais elevadas de próclise e mais baixas de ênclise.

Os resultados individuais dos falantes nativos (Tabela 7) mostram que 17 aceitam próclise e rejeitam ênclise e 3 aceitam próclise, mas não rejeitam ênclise de forma consistente. De um modo geral, a rejeição de ênclise é mais categórica nas frases negativas e nas orações completivas de conjuntivo do que nos outros contextos de próclise testados.

Tabela 7. Mediana de aceitação de ênclise e de próclise em contextos de próclise por falante nativo de PE

Participante	Negação		Completiva de conjuntivo		Completiva de indicativo		Adverbial		Sujeito quantificado		Total	
	E	P	E	P	E	P	E	P	E	P	E	P
1	1	5	1	5	1	4	1	4	2	5	1	5
2	1	5	1	5	1	5	1	5	1	5	1	5
3	1	5	1	5	1	5	1	5	1	5	1	5
4	1	5	1	5	1	5	3	5	2	5	1	5
5	1	5	1	5	1	5	1	5	1	5	1	5
6	1	5	1	5	1	5	1	5	1	5	1	5
7	1	5	1	5	1	5	2	5	1	5	1	5
8	1	5	1	5	1	5	1	5	1	5	1	5
9	1	5	1	5	1	5	1	5	1	5	1	5
10	1	5	1	5	1	5	1	5	1	5	1	5
11	1	5	1	5	1	5	1	5	1	5	1	5
12	2	5	2	5	2	5	2	5	2	5	2	5
13	1	5	2	5	4	5	5	5	2	5	2	5
14	1	5	2	5	3	5	5	5	2	5	2	5
15	2	5	2	5	2	5	2	5	2	5	2	5
16	1	5	2	5	2	5	2	5	2	5	2	5
17	1	5	1	5	3	3	4	3	4	4	2	4
18	2	5	3	5	4	5	4	4	4	5	3	5
19	3	5	3	5	3	5	3	5	3	5	3	5
20	2	5	3	5	4	5	5	4	5	4	4	5

Os resultados individuais dos aprendentes de L2 em contextos de próclise (Tabela 8) mostram que há uma evolução ao longo dos três níveis de proficiência considerados. No grupo intermédio, encontramos 1 aprendente que parece não ter intuições claras sobre a colocação de clíticos, pois tem uma mediana de 3 para todas as condições. Os restantes 9 aprendentes, de um modo geral, aceitam próclise nos diversos contextos analisados. Contudo, apenas 4 deles apresentam uma clara tendência de rejeição da ênclise. Estes aprendentes são os mesmos que exibem próclise maioritária na tarefa de produção. O contexto em que a rejeição de ênclise é mais acentuada no grupo intermédio é o da negação. No nível avançado, o número de aprendentes que rejeita a ênclise aumenta, sobretudo, em contexto de negação (8 aprendentes) e de orações completivas (6 em completivas de conjuntivo e 7 em completivas de indicativo). Nos outros contextos de próclise analisados, a maioria dos aprendentes continua a aceitar ênclise. Finalmente, no grupo quase nativo, a maioria dos aprendentes aceita próclise e rejeita ênclise em todos os contextos. Há, no entanto, um aprendente que apresenta medianas idênticas de aceitação de próclise e de ênclise.

Tabela 8. Mediana de aceitação de ênclise e de próclise em contextos de próclise por aprendente de PE L2

Participante		Negação		Completiva de conjuntivo		Completiva de indicativo		Adverbial		Sujeito quantificado		Total	
		E	P	E	P	E	P	E	P	E	P	E	P
INT	1	1	5	1	5	2	3	3	4	2	2	1	5
	2	1	5	1	5	1	5	1	5	5	5	1	5
	3	2	5	2	5	4	4	3	5	3	4	2	5
	4	1	5	5	5	2	3	5	5	2	5	2	5
	5	4	4	4	4	4	4	4	5	4	4	4	4
	6	5	5	5	5	5	2	5	5	5	3	5	5
	7	4	5	5	5	4	5	5	4	5	5	5	5
	8	4	5	5	5	5	4	5	5	5	5	5	5
	9	5	4	5	3	4	3	5	4	5	5	5	4
	10	3	3	3	3	3	3	3	3	3	3	3	3
AV	1	1	5	1	5	1	5	1	5	1	5	1	5
	2	1	5	1	5	1	5	1	2	5	5	1	5
	3	1	5	1	5	1	4	4	2	3	5	1	5
	4	1	5	1	5	2	4	2	5	2	5	2	5
	5	1	5	2	5	1	4	4	4	4	2	2	5
	6	1	5	3	5	1	3	5	5	5	5	3	5
	7	1	5	2	5	2	3	3	3	4	3	2	3
	8	1	5	3	5	4	4	5	4	5	5	4	5
	9	5	5	5	5	5	5	5	5	5	3	5	5
	10	2	5	5	5	4	3	5	4	5	3	5	4
QN	1	1	5	1	5	1	5	2	5	2	5	1	5
	2	1	5	1	5	1	4	1	5	1	5	1	5
	3	1	5	1	5	1	5	1	5	2	5	1	5
	4	1	5	1	5	2	5	1	5	1	5	1	5
	5	1	5	1	5	1	4	3	5	3	5	1	5
	6	1	5	1	5	1	5	1	5	1	5	1	5
	7	1	5	2	4	2	5	2	5	3	5	2	5
	8	1	5	1	5	2	5	4	5	4	5	2	5
	9	1	5	3	5	3	5	2	5	3	5	2	5
	10	3	5	3	5	4	5	4	5	4	5	4	5

Em conjunto, os resultados das tarefas de produção e de juízos mostram que os contextos de ênclise não são completamente imunes a problemas em níveis avançados e que a próclise é adquirida mais cedo em alguns contextos (especialmente o de negação e o de completiva de conjuntivo) do que em outros (nomeadamente adverbiais e sujeitos quantificado). A Tabela 9 apresenta uma síntese dos resultados.

Tabela 9. Preferências de colocação do clítico por contexto e grupo nas tarefas de produção e de juízos de aceitabilidade.

Grupo	Tarefa	Frase simples s/ proclisador	Negação	Completiva de conjuntivo	Completiva de indicativo	Adverbial	Sujeito quantificado
Controlos	**Produção**	ENCL	PROCL	PROCL	PROCL	PROCL	PROCL
	Juízos	ENCL	PROCL	PROCL	PROCL	PROCL	PROCL
QN	**Produção**	ENCL	PROCL	PROCL	PROCL	PROCL	PROCL
	Juízos	ENCL / PROCL	PROCL	PROCL	PROCL	PROCL	PROCL
AV	**Produção**	ENCL	PROCL	PROCL	PROCL [a]	ENCL / PROCL	ENCL / PROCL
	Juízos	ENCL / PROCL	PROCL	PROCL	PROCL	ENCL / PROCL	ENCL / PROCL
INT	**Produção**	ENCL	PROCL	PROCL	ENCL / PROCL	ENCL	ENCL
	Juízos	ENCL / PROCL	PROCL	PROCL[1]	ENCL / PROCL	ENCL / PROCL	ENCL / PROCL

[a] Diferença próxima de significância estatística.

6. Discussão e conclusões

Vejamos agora de que forma os resultados descritos na Secção 5 nos permitem responder às três questões de investigação na base deste estudo.

QI.1. Os contextos de ênclise são adquiridos cedo em PE L2, tal como em PE L1?

Para esta primeira questão, predizíamos que, se os contextos de ênclise forem adquiridos cedo em PE L2, tal como em L1, os aprendentes terão já conhecimento destes contextos no nível intermédio. Os resultados da tarefa de produção mostram que todos os grupos de L2 têm um comportamento semelhante ao dos controlos em contextos de ênclise, ou seja, exibem uma preferência clara por ênclise, não produzindo praticamente próclise. Contudo, na tarefa de juízos, ao contrário do que se observa no grupo de controlo, todos os grupos de L2 aceitam quer ênclise quer próclise. Esta assimetria entre os resultados das duas tarefas poderá decorrer da diferente natureza das tarefas. Na tarefa de juízos de aceitabilidade, por estarem sob pressão de tempo, os aprendentes poderão ter maior dificuldade em inibir a L1, que, em frases declarativas simples, apenas permite próclise, divergindo, assim, do PE. Outra hipótese é que os resultados da tarefa de juízos estejam relacionados com a instrução que foi dada aos participantes. Embora todos fossem falantes de PE, ao pedir-se que avaliassem o grau de naturalidade das frases em português, alguns poderão ter considerado que estas seriam naturais noutras variedades do português, como o português do Brasil, por exemplo. Em todo o caso, os resultados não são conclusivos relativamente à aquisição plena da ênclise, pelo que não é possível confirmar com segurança a nossa predição. No futuro, será necessária mais investigação, nomeadamente recorrendo a uma tarefa de juízos de aceitabilidade com áudio em PE e com uma instrução mais clara de que a naturalidade das frases deve ser avaliada tendo em conta a variedade europeia do português.

QI.2. O percurso de desenvolvimento dos contextos de próclise em PE L2 assemelha-se ao encontrado na aquisição de PE L1?

Em relação à segunda questão, predizíamos que, se o percurso de desenvolvimento dos contextos de próclise em PE L2 se assemelhar ao que tem sido observado na aquisição de PE L1, os aprendentes desenvolverão conhecimento destes contextos de forma gradual: adquirirão primeiro os contextos de aquisição mais precoce e mais tarde os de mais difícil aquisição, de acordo com a escala seguinte (baseada em Costa et al., 2015): Negação > Completivas > Adverbiais finitas > Sujeito quantificado. Os resultados de ambas as

tarefas mostram que o conhecimento de próclise em contextos de negação é adquirido precocemente, estando já estabilizado no nível intermédio, o que está em linha com o que se verifica na aquisição de PE L1 e em estudos prévios sobre aquisição de PE L2 (e.g., Gu, 2019, 2021). A preferência por próclise nas completivas desenvolve-se mais tarde, havendo uma diferença entre completivas de indicativo e de conjuntivo. Nas completivas de conjuntivo, que foram as únicas testadas por Costa et al. (2015), os aprendentes começam por manifestar uma preferência por próclise na tarefa de produção no nível intermédio, mas só no nível avançado é que esta preferência é clara nas duas tarefas. Nas completivas de indicativo, a preferência por próclise começa a emergir mais tarde, no nível avançado, o que é visível na tarefa de juízos, e só estabiliza no nível quase nativo, em que a preferência por próclise se manifesta nas duas tarefas. A preferência por próclise nos contextos de adverbiais e sujeitos quantificados é a que se desenvolve mais tarde, apenas no nível quase nativo. Este resultado está de acordo com os dados da aquisição de PE L1 (cf. Secção 2.2) e L2 (cf. Secção 2.3), ainda que, no presente estudo, não se observem diferenças entre os dois contextos. Em suma, os nossos resultados sugerem a seguinte sequência de aquisição de próclise em PE L2: Negação > Completivas de conjuntivo > Completivas de indicativo > Adverbiais causais, Sujeito quantificado. Confirma-se, assim, a predição de que, globalmente, o percurso de desenvolvimento dos contextos de próclise em PE L2 se assemelha ao observado na aquisição de PE L1.

QI.3. A convergência com a língua alvo é possível em PE L2 no que diz respeito à colocação dos clíticos em contextos de próclise?

Por fim, em relação à terceira questão de investigação, a nossa predição apontava para a possibilidade de convergência com a língua alvo no que diz respeito à colocação dos clíticos em contextos de próclise em PE L2. Os nossos resultados confirmam esta predição, uma vez que mostram que a preferência por próclise é totalmente adquirida por falantes quase nativos de PE L2, mesmo nos contextos que estão sujeitos a atrasos de desenvolvimento significativos.

Em conjunto, os resultados deste estudo indicam que os fenómenos que são adquiridos tarde em L1 podem ser igualmente problemáticos na aquisição de L2. O conhecimento de contextos de próclise que são categóricos nas gramáticas nativas adultas, como a negação, estabiliza precocemente em L2, tal como em L1. Já a preferência por próclise em contextos envolvendo algum tipo de variabilidade (seja sintática ou lexical), como orações adverbiais e sujeitos quantificados, é adquirida mais tarde. Estes são os únicos contextos em que a ênclise é também produzida, com alguma expressão, por falantes de PE L1 (por 2 falantes no presente estudo[6] e por 6 falantes em Costa et al., 2015). Na tarefa de juízos, observa-se também que o número de falantes de PE L1 que rejeita a ênclise nestes contextos é menor do que em contextos mais categóricos como a negação.

De igual modo, as diferenças encontradas entre completivas poderão estar relacionadas com o facto de as de conjuntivo serem contextos mais categóricos de próclise nas gramáticas nativas do que as de indicativo, o que é observável nos resultados individuais da tarefa de juízos de aceitabilidade (nas completivas de conjuntivo, a ênclise é rejeitada por 17 dos 20 falantes testados, enquanto, nas de indicativo, só 14 a rejeitam).

Pelo menos no que diz respeito à aquisição de L2 por adultos, os nossos resultados não apoiam a ideia defendida em Tsimpli (2014) de que os fenómenos que se adquirem (muito) tarde são apenas os que envolvem interfaces, uma vez que alguns fenómenos estritamente sintáticos, como a próclise em completivas de indicativo e em orações adverbiais, também se desenvolvem tarde. Tendo em conta que os contextos mais problemáticos na aquisição de PE L2 (e L1) são os menos categóricos nas gramáticas nativas e, consequentemente, menos categóricos no input,[7] podemos concluir que o input desempenha um papel determinante na aquisição da colocação de clíticos, mesmo em contextos que não envolvem interfaces: quanto maior a variabilidade no input, maior a complexidade da tarefa de aquisição de próclise em PE L2. Por isso, nestes casos, os aprendentes

[6] Num grupo de 38 falantes de PE L1 testados no âmbito deste projeto, 6 produzem maioritariamente ênclise nos itens com orações adverbiais e sujeitos quantificados.

[7] Especialmente, as completivas de indicativo, adverbiais causais, e frases com sujeitos quantificados.

necessitam de exposição prolongada a input para descobrir o padrão preferencial de colocação dos clíticos em PE. Embora haja uma sobreprodução e sobreaceitação de ênclise em contextos menos categóricos de próclise até um nível avançado, a convergência total com a língua alvo é possível, mas apenas no nível quase nativo, o que está em linha com as predições da HI.

A existência de um percurso de aquisição semelhante entre L1 e L2 no que diz respeito à colocação de clíticos mostra que, contrariamente ao que alguns autores propõem (e.g., Bley-Vroman, 1989), a aquisição de L1 e de L2 não são dois processos fundamentalmente diferentes, pelo menos, no que se refere a alguns fenómenos morfossintáticos. Os nossos resultados sugerem que ambos os processos podem ser influenciados pelo grau de variabilidade do fenómeno no input.

Em conclusão, os nossos resultados mostram que, em PE L2, a aquisição dos padrões de colocação de clíticos por adultos, de um modo geral, segue o percurso identificado em outras populações, nomeadamente nas crianças monolingues e bilingues de PE L1. Esta é, assim, uma área em que parece haver um percurso de desenvolvimento comum em L1 e L2. Tal como em L1, a aquisição plena do conhecimento da colocação de clíticos parece ser possível em PE L2. Uma questão que fica em aberto é se a aceitação de próclise em contexto de ênclise que observámos na tarefa de juízos de aceitabilidade é um efeito da tarefa ou um efeito da L1 dos aprendentes. Deixamos esta questão para investigação futura.

Referências

Bader, Markus & Jana Häussler (2010) Toward a model of grammaticality judgments. *Journal of Linguistics* 46 (2), pp. 273–330. https://doi.org/10.1017/S0022226709990260

Bley-Vroman, Robert (1989) What is the logical problem of foreign language learning? In Susan Gass & Jacquelyn Schachter (eds.), *Linguistic perspectives on second language acquisition*. Cambridge University Press, pp. 41–68.

Casa Nova, Manuela (2014) *Formas de realização do pronome clítico em português europeu por falantes de herança luso-franceses*. Dissertação de mestrado, Universidade do Minho.

Costa, João & Maria Lobo (2007). Complexidade e omissão de clíticos: O caso dos reflexos. In *Textos seleccionados do XXII Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 303–313.

Costa, João, Alexandra Fiéis & Maria Lobo (2015) Input variability and late acquisition: Clitic misplacement in European Portuguese. *Lingua* 161, pp. 10–26. https://doi.org/10.1016/j.lingua.2014.05.009

Cunnings, Ian (2012) An overview of mixed-effects statistical models for second language researchers. *Second Language Research* 28 (3), pp. 369–382. https://doi.org/10.1177/0267658312443651

Duarte, Inês & Gabriela Matos (2000) Romance clitics and the Minimalist Program. In João Costa (ed.), *Portuguese syntax: New comparative studies*. Oxford University Press, pp. 116–142.

Duarte, Inês, Gabriela Matos & Isabel Faria (1995) Specificity of European Portuguese clitics in Romance. In Isabel Faria & Maria João Freitas (eds.), *Studies on the acquisition of Portuguese*. APL & Colibri, pp. 129–154.

Ellis, Rod (2005) Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition* 27 (2), pp. 141–172. https://doi.org/10.1017/S0272263105050096

Flores, Cristina & Pilar Barbosa (2014) When reduced input leads to delayed acquisition: A study on the acquisition of clitic placement by Portuguese heritage speakers. *International Journal of Bilingualism* 18 (3), pp. 304–325. https://doi.org/10.1177/1367006912448124

Flores, Cristina, Pilar Barbosa & Manuela Casa Nova (2016) A closer look at cross-linguistic influence in the acquisition of Portuguese as a Heritage Language. In Sambor Grucza, Magdalena Olpinska-Szkieko & Piotr Romanowski (eds.), *Bilingual landscape of the contemporary world*. Peter Lang Verlag, pp. 75–94.

Grüter, Therese (2006) *Object clitics and null objects in the acquisition of French*. Tese de doutoramento, McGill, University of Montreal.

Gu, Wenjun (2019) Aquisição de pronomes clíticos de português europeu por falantes de chinês: Dados sobre a colocação. *Revista da Associação Portuguesa de Linguística* (5), pp. 190–206. https://doi.org/10.26334/2183-9077/rapln5ano2019a14
Gu, Wenjun (2021) Aquisição da posição dos pronomes clíticos de português europeu por falantes nativos de chinês. *Textos selecionados do XIII e XIV Fórum de Partilha Linguística.* NOVA FCSH – CLUNL, pp. 59–72.
Gu, Wenjun (2022) Aquisição da posição dos pronomes clíticos em português europeu como L2. *Rotas a Oriente* 2, pp. 205–226. https://doi.org/10.34624/ro.v0i2.27829
Guasti, Maria Teresa (1993) Verb syntax in Italian child grammar: Finite and nonfinite verbs. *Language Acquisition* 3 (1), pp. 1–40. https://doi.org/10.1207/s15327817la0301_1
Hamann, Cornelia, Luigi Rizzi & Ulrich Hans Frauenfelder (1996) On the acquisition of subject and object clitics in French. In Harald Clahsen (ed.), *Generative perspectives on language acquisition*. John Benjamins, pp. 309–334.
Hopp, Holger (2007) *Ultimate attainment at the interfaces in second language acquisition: Grammar and processing*. Tese de doutoramento, University of Groningen.
Linck, Jared & Ian Cunnings (2015) The utility and application of mixed-effects models in second language research. *Language Learning* 65, pp. 185–207. https://doi.org/10.1111/lang.12117
Lobo, Maria (2003) *Aspectos da sintaxe das orações subordinadas adverbiais do português*. Tese de doutoramento, Universidade Nova de Lisboa.
Madeira, Ana & Maria Francisca Xavier (2009) The acquisition of clitic pronouns in L2 European Portuguese. In Acrisio Pires & Jason Rothman (eds.), *Minimalist inquiries into child and adult language acquisition: Case studies across Portuguese*. Mouton de Gruyter, pp. 273–299.
Madeira, Ana, Maria de Lourdes Crispim & Maria Francisca Xavier (2006) Clíticos pronominais em português L2. In *Textos selecionados do XXI Encontro Nacional da Associação Portuguesa de Linguística*. APL & Colibri, pp. 495–510.
Marinis, Theo (2000) The acquisition of clitic objects in Modern Greek: Single clitics, clitic doubling, clitic left dislocation. *ZAS Papers in Linguistics* 15, pp. 259–281. https://doi.org/10.21248/zaspil.15.2000.32
Martins, Ana Maria (1994) Enclisis, VP-deletion and the nature of Sigma. *Probus* 6, pp. 173–205. https://doi.org/10.1515/prbs.1994.6.2-3.173
Martins, Ana Maria (2016) A colocação dos pronomes clíticos em sincronia e diacronia. In Ana Maria Martins & Ernestina Carrilho (eds.), *Manual de linguística portuguesa*. De Gruyter, pp. 401–430.
Neokleous, Theoni (2013) Clitic (mis)placement in early grammars: evidence from Cypriot Greek. In Stravoula Stavrakaki, Marina Lalioti & Polyxeni Konstantinopoulou (eds.), *Advances in language acquisition*. Cambridge Scholars Publishing, pp. 147–155.
Petinou, Kakia & Arhonto Terzi (2002) Clitic misplacement among normally developing children and children with specific language impairment and the status of Infl heads. *Language Acquisition* 10 (1), pp. 1–28. https://doi.org/10.1207/S15327817LA1001_1
Pierce, Amy E. (1992) *Language acquisition and syntactic theory: A comparative analysis of French and English child grammars*. Kluwer.
Silva, Carolina (2009) Assimetrias na aquisição de diferentes tipos de clíticos em português europeu. In *Textos seleccionados do XXIV Encontro Nacional da Associação Portuguesa de Linguística*. APL & Colibri, pp. 527–541.
Sorace, Antonella (2014) Input, timing, and outcomes in a wider model of bilingualism. *Linguistic Approaches to Bilingualism* 4, pp. 377–380. https://doi.org/10.1075/lab.4.3.14sor
Sorace, Antonella & Francesca Filiaci (2006) Anaphora resolution in near-native speakers of Italian. *Second Language Research* 22 (3), pp. 339–368. https://doi.org/10.1191/0267658306sr271oa

Tomaz, Margarida, Maria Lobo, Ana Madeira, Carla Soares-Jesel & Stéphanie Vaz (2019) Omissão e colocação de clíticos por crianças bilingues Português-Francês. *Revista da Associação Portuguesa de Linguística* (5), pp. 85–412. https://doi.org/10.26334/2183-9077/rapln5ano2019a25

Tsimpli, Ianthi (2014) Early, late or very late?: Timing acquisition and bilingualism. *Linguistic Approaches to Bilingualism* 4 (3), pp. 283–313. https://doi.org/10.1075/lab.4.3.01tsi

Wexler, Ken, Anna Gavarró & Vincent Torrens (2004) Feature checking and object clitic omission in child Spanish and Catalan. In Reineke Bok-Bennema, Barte Hollebrandse, Brigitte Kampers-Manhe & Petra Sleeman (eds.), *Romance language and linguistic theories 2002*. John Benjamins, pp. 253–269.

White, Lydia & Fred Genesee (1996) How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research* 12 (3), pp. 233–265. https://doi.org/10.1177/026765839601200301

Tecnologias de fala e a variação de pronúncia do russo no contexto de VoiceInteraction

Anna Havras[1,2]*, Carlos Mendes*[2]*, Gueorgui* Hristovsky[1]*, Sérgio Paulo*[2]*, Helena Moniz*[1,3]

[1]Universidade de Lisboa, Faculdade de Letras, Lisboa, Portugal
[2]VoiceInteraction – Tecnologias de Processamento de Fala, Lisboa, Portugal
[3]INESC-ID, Lisboa, Portugal

Resumo

O presente artigo tem como objetivo descrever o trabalho realizado na *VoiceInteraction*, empresa especializada no desenvolvimento de soluções de processamento de fala, com especial destaque para a transcrição automática, que recorre a um Reconhecedor Automático de Fala (ASR) híbrido. O objetivo principal centrou-se no estudo das características fonéticas da língua russa, tendo em conta quatro tarefas principais: descrição do inventário fonético-fonológico; validação das transcrições de noticiários; validação de um léxico previamente criado; e integração de pausas preenchidas no ASR. O presente trabalho contribuiu para o projeto *Artificial Intelligence and Advanced Data Analysis for Authority Agencies* (AIDA), financiado pela Comissão Europeia no âmbito do programa Horizonte 2020, transcrevendo os dados em língua russa.

Palavras-chave: reconhecimento automático de fala, fonética, pausas preenchidas, língua russa, variedades linguísticas.

Abstract

This article aims to describe the work conducted at *VoiceInteraction*, a company specialized in speech processing solutions, with a particular focus on automatic transcription using a Hybrid Automatic Speech Recognizer (ASR). The primary objective revolved around studying the phonetic characteristics of the Russian language, encompassing four main tasks: describing the phonetic-phonological inventory, validating news transcriptions, validating a previously created lexicon, and integrating filled pauses into the ASR. This work contributed to the Artificial Intelligence and Advanced Data Analysis for Authority Agencies (AIDA) project, funded by the European Commission under the Horizon 2020 program, by transcribing the data in the Russian language.

Keywords: automatic speech recognition, phonetics, filled pauses, Russian language, linguistic varieties.

1. Introdução

O presente trabalho consistiu na validação de um sistema de reconhecimento de fala automático, previamente criado, para a **língua russa**. Esta língua é falada em diferentes regiões dos continentes europeu e asiático. Assim, o objetivo foi o de identificar as variantes fonéticas e variedades linguísticas da língua russa e incorporá-las no reconhecedor, aumentando assim a qualidade do reconhecimento. Foi necessário realizar uma descrição detalhada do sistema fonético-fonológico da língua russa para a validação de um *phoneset* previamente criado e validação do modelo de léxico, que continha as palavras e as suas pronúncias correspondentes. As regiões analisadas foram: a Rússia Europeia, a Bielorrússia e o Cáucaso. Na Rússia Europeia, o russo é a língua oficial; na Bielorrússia, o russo é uma das línguas oficiais do país; e no Cáucaso, o

russo é usado como língua franca, visto que esta era falada na União Soviética e continua até hoje a ser utilizada pelos falantes.

Desta forma, para a validação de um **Reconhecedor Automático de Fala** (ASR) previamente criado para a língua russa, foi necessário recolher dados de fala e de texto de conteúdo noticioso das regiões referidas anteriormente. Os dados de fala foram transcritos automaticamente e editados pelos anotadores humanos, seguindo as orientações de anotação criadas pela empresa, a fim de corrigir os erros presentes nas transcrições automáticas.

Por outro lado, analisámos e validámos um conjunto de **pausas preenchidas**, previamente criado para a língua russa na empresa. Centrámos a nossa atenção na validação das pausas preenchidas, porque se revelaram um tipo de disfluência mais frequentemente utilizado pelos falantes nas transcrições. Observámos que os anotadores humanos apresentaram alguma dificuldade em identificá-las, sendo que o trabalho efetuado corresponde a uma primeira validação do conjunto das pausas preenchidas da língua. A falta de anotação das pausas preenchidas pelos transcritores humanos pode ter um impacto significativo no desempenho do ASR, quando aplicado a tarefas em que a fala é menos preparada e, por isso, estas disfluências são mais frequentes.

Por fim, analisámos o **léxico** da língua russa. O objetivo foi validar as pronúncias geradas automaticamente para as 10 mil palavras mais frequentes de um léxico composto por 400 mil entradas. Esta tarefa permitirá a extração de padrões linguísticos que serão utilizados no módulo *Grapheme-to-Phone* (G2P).

O presente trabalho contribuiu ainda para o projeto *Artificial Intelligence and Advanced Data Analysis for Authority Agencies* (AIDA), financiado pela Comissão Europeia no âmbito do programa Horizonte 2020. Após a validação do sistema de reconhecimento de língua russa, foi possível aplicar o presente trabalho ao AIDA, através da transcrição e da validação dos dados de conteúdo sensível de língua russa.

2. Fonética da língua russa

Para a validação do modelo de léxico e de pronúncia foi necessário recolher regras e regularidades da pronúncia da língua russa. A informação apresentada neste capítulo é uma seleção dos trabalhos de Litnevskaya (2006), Demidov et al. (2013), Kasatkin (2014), Popov (2014), Yanushevskaya e Bunčić (2015), Osipova (2018) e Sokolova (2021).

De acordo com Jurafsky e Martin (2021), a fonética é o estudo dos sons da fala utilizados nas línguas do mundo, como são produzidos no trato vocal humano, como são realizados acusticamente e como podem ser digitalizados e processados. A Tabela 1 apresenta o sistema fonético da língua russa. Na primeira coluna são apresentadas as variantes fonéticas de acordo com o Alfabeto Fonético Internacional (IPA); na segunda coluna, as mesmas variantes são apresentadas de acordo com o X-SAMPA; na terceira coluna, são apresentados os possíveis grafemas correspondentes a cada uma das variantes fonéticas; e finalmente, na última coluna, exemplos com transcrições fonéticas. As vogais são apresentadas no lado esquerdo da tabela, e as consoantes no lado direito.

Tabela 1. Sistema fonético da língua russa

IPA	X-SAMPA	Grafemas	Exemplo	IPA	X-SAMPA	Grafemas	Exemplo
I	i	И	*Ми**л**а* [mʲ**i**ɬə]	p	p	п	***п**ока* [**p**ɐká]
ɪ	I	и, е, э, а, я	*м**е**док* [mʲ**ɪ**dók]	pʲ	p'	п	***п**ятак* [**pʲ**ɪták]
E	e	Е	*м**е**ра* [mʲ**é**ra]	b	b	б	***б**уря* [**b**úrʲə]
ɛ	E	Э	*м**э**р* [m**ɛ́**r]	bʲ	b'	б	***б**юро* [**bʲ**uró]
A	a	а, я	*м**а**л* [m**á**ɬ]	t	t	т	*э**т**аж* [ɪ**t**áʂ]
ɐ	6	а, о	*ж**а**ра* [ʐ**ɐ**rá]	tʲ	t'	т	*жи**т**ь* [ʐɨ**tʲ**]
ə	@	а, о, я	*бур**я*** [búrʲ**ə**]	d	d	д	*ря**д**ы* [rʲɪ**d**ɨ]
Æ	{	я, а	*ч**а**й* [t͡ɕʲ**æ**j]	dʲ	d'	д	***д**ень* [**dʲ**énʲ]
ɨ	1	ы, и, а, е	*м**ы**ло* [m**ɨ**ɬə]	k	k	к	*уро**к*** [ʊró**k**]
ɔ	O	О	*как**о*** [kɐká**ɔ**]	kʲ	k'	к	*ле**к**ий* [lʲóxʲ**kʲ**ɪj]
O	o	О	*м**о**л* [m**ó**ɬ]	g	g	г	*и**г**ра* [ɪ**g**rá]
ɵ	8	ё, о	*вахт**ё**ры* [vɐxtʲ**ó**rɨ]	gʲ	g'	г	*серь**г**и* [sʲɪrʲ**gʲ**ɪ]
U	u	у, ю	*б**у**ря* [b**ú**rʲə]	x	x	х	***х**леб* [**x**lʲép]
ʊ	U	у, ю	*авг**у**ст* [ávg**ʊ**st]	xʲ	x'	х	*арх**и**в* [ɐr**xʲ**íf]
ʉ	}	у, ю	*зл**ю**сь* [zlʲ**ʉ́**sʲ]	f	f	ф	*алло**ф**он* [ɐlɐ**f**ón]
				fʲ	f'	ф	***ф**ильм* [**f**ʲílʲm]
				v	v	в	***в**озди* [**v**ózʲdʲɪ]
				vʲ	v'	в	***в**юга* [**vʲ**úgə]
				s	s	с	*ча**с**ы* [t͡ɕʲɪ**s**ɨ]
				ʐ	z`	ж	***ж**ара* [**ʐ**ɐrá]
				t͡s	ts	ц	***ц**ель* [**t͡s**élʲ]
				t͡ɕ	ts\'	ч	***ч**ай* [**t͡ɕ**áj]
				ɕː	s\:'	щ	***щ**ётка* [**ɕː**ʲɵ́tkə]
				j	j	й	*пусто**й*** [pʊstó**j**]
				ɬ	5	л	*мо**л*** [mó**ɬ**]
				lʲ	l'	л	*ате**л**ье* [ɐtʲ**lʲ**jé][a]
				m	m	м	*съё**м**ка* [sʲjó**m**kə]
				mʲ	m'	м	***м**ера* [**mʲ**éra]
				n	n	н	*пья**н*** [pʲjæ**n**]
				nʲ	n'	н	*де**н**ь* [dʲé**nʲ**]
				r	r	р	*ме**р**а* [mʲé**r**a]
				rʲ	r'	р	*бу**р**я* [bú**rʲ**ə]

[a] Esta palavra também pode ser pronunciada como [ɐtɛlʲjé], com a vogal [ɛ] ao invés de [ɪ] na segunda sílaba.

2.1. **Pronúncia das vogais**

Nesta secção, apresentamos 15 das variantes fonéticas mais frequentes das vogais em russo. Cada uma é apresentada de acordo com os seus possíveis grafemas e exemplos.

As vogais que podem ocorrer em **posição tónica** na língua russa são [a], [æ], [i], [ɨ], [e], [ɛ], [o], [ɔ], [ɵ], [u] e [ʉ]. A Tabela 2 mostra as variantes fonéticas que podem ocorrer em posição tónica e os seus possíveis grafemas correspondentes, seguidos de exemplos, transcrição fonética e o significado equivalente em português.

Tabela 2. Vogais tónicas e os seus grafemas correspondentes (adaptado de Litnevskaya, 2006)

Vogais acentuadas	Grafemas	Palavra	Pronúncia	Significado
[a]	а - [a]	*м**а**л*	[m**á**ɫ]	‘pequeno’
	я - [ʲa]	*м**я**л*	[mʲ**á**ɫ]	‘amassado’
[æ]	а - [a]	*ч**а**й*	[t͡ɕʲ**ǽ**j]	‘chá’
	я - [ʲa]	*п**я**льцы*	[pʲ**ǽ**lʲt͡sɨ]	‘aro
[ɨ]	и - [i]	*М**и**ла*	[mʲ**ɨ́**ɫə]	nome próprio ‘Mila’
[ɨ]	и - [i]	*ж**и**ть*	[ʐ**ɨ́**tʲ]	‘viver’
	ы - [ɨ]	*м**ы**ло*	[m**ɨ́**ɫə]	‘sabão’
[e]	е - [ʲe]	*м**е**ра*	[mʲ**é**ra]	‘medida’
[ɛ]	э - [ɛ]	*м**э**р*	[m**ɛ́**r]	‘Presidente da Câmara’
[o]	о - [o]	*м**о**л*	[m**ó**ɫ]	‘cais’
[ɔ]	о - [o]	***о**кна*	[**ɔ́**knə]	‘janelas’
[ɵ]	о - [o]	*щипач**о**м*	[ɛːʲɪpʁt͡ɕʲ**ɵ́**m]	‘com a pinça’
	ё - [ʲo]	*вахт**ё**ры*	[vɐxtʲ**ɵ́**rɨ]	‘porteiros’
[u]	у - [u]	*б**у**ря*	[b**ú**rʲə]	‘tempestade’
	ю - [ʲu]	*со**ю**з*	[sɐj**ú**z]	‘união’
[ʉ]	у - [u]	*щ**у**рить*	[ɛːʲ**ʉ́**rʲɪtʲ]	‘piscar os olhos’
	ю - [ʲu]	*зл**ю**сь*	[zlʲ**ʉ́**sʲ]	‘Eu zango-me’

Em posição átona podem ocorrer as seguintes variantes fonéticas - [ɐ, ə, ɪ, ɨ, ɔ, ʊ, ʉ]. Os exemplos a seguir representam as regularidades em posições átonas descritas para esta língua.

Após as consoantes fortes (exceto [ʂ], [ʐ] e [t͡ʂ]) podem ser encontrados grafemas como *y* - [u] realizado como [ʊ] (1a); *a* - [a] e *o* - [o] realizados como [ɐ] / [ə] (1b-1e); *ы* - [ɨ] e *e* - [ʲe] realizados como [ɪ] (1f, 1g):

(1a) *y*	[ʊ]	*п**у**стой*	[p**ʊ**stój]	‘vazio’
(1b) *a*	[ɐ]	*с**а**ма*	[s**ɐ**má]	‘sozinha’
(1c) *a*	[ə]	*камер**а***	[kámʲɪr**ə**]	‘câmara’
(1d) *o*	[ɐ]	*п**о**стель*	[p**ɐ**sʲtʲélʲ]	‘cama’
(1e) *o*	[ə]	*г**о**лова*	[g**ə**ɫɐvá]	‘cabeça’
(1f) *ы*	[ɪ]	*м**ы**слитель*	[m**ɪ**slʲítʲɪlʲ]	‘pensador’
(1g) *e*	[ɪ]	*т**е**стировать*	[t**ɪ**stʲírəvətʲ]	‘testar’

Após as variantes [ʂ], [ʐ] e [t͡ʂ], podem ser encontrados grafemas como *y* - [u] realizado como [ʊ] (2a); *a* - [a] e *o* - [o] realizados como [ɐ], [ɨ] ou [ə] (2b-2d); *ы* - [ɨ], *e* - [ʲe] e *u* - [i] realizados como [ɨ] (2e-2h):

(2a) *у*	[ʊ]	***щу**меть*	[ɕʊmʲétʲ]	‘fazer barulho’
(2b) *а*	[ɐ]	*ж**а**ра*	[zɐ̱rá]	‘calor’
(2c) *а*	[ɨ]	*лош**а**дей*	[łəɕɨdʲéj]	‘dos cavalos’
(2d) *а*	[ə]	*мужж**а***	[múzə̱]	‘do marido’
(2e) *о*	[ɨ]	*ш**о**колад*	[ɕɨkɐłát]	‘chocolate’
(2f) *ы*	[ɨ]	*ножниц**ы***	[nózn̪ʲɪt͡sɨ̱]	‘tesoura’
(2g) *и*	[ɨ]	*тиш**и**ною*	[tʲɪɕɨnójʊ]	‘com o silêncio’
(2h) *е*	[ɨ]	*ж**е**лезо*	[zɨ̱lʲézə]	‘ferro’

Após as consoantes suaves, podem ser encontrados grafemas como *у* - [u] e *ю* - [ʲu] realizados como [ʊ]/ [ʉ] (3a-3d); *а* - [a], *и* - [i], *е* - [ʲe] e *я* - [ʲa] realizados como [ɪ] (3e-3h):

(3a) *у*	[ʊ]	*ч**у**десный*	[t͡ɕʲʊ̱dʲésnɨj]	‘adorável’
(3b) *ю*	[ʊ]	*л**ю**бить*	[lʲʊ̱bʲɪtʲ]	‘amar’
(3c) *у*	[ʉ]	*ощ**у**щение*	[ɐɕːʲʉ̱ɕːʲenʲɪje]	‘sentimento’
(3d) *ю*	[ʉ]	*следу**ю**щий*	[sʲlʲédʊjʉ̱ɕːʲɪj]	‘seguinte’
(3e) *а*	[ɪ]	*ч**а**сы*	[t͡ɕʲɪ̱sɨ́]	‘horas’
(3f) *и*	[ɪ]	*м**и**ры*	[mʲɪ̱rɨ́]	‘mundos’
(3g) *е*	[ɪ]	*м**е**нять*	[mʲɪ̱nʲátʲ]	‘mudar’
(3h) *я*	[ɪ]	*п**я**так*	[pʲɪ̱ták]	‘níquel’

Por fim, no início da palavra, podem ser encontrados grafemas como *у* - [u] realizado como [ʊ] (4a); *а* - [a] e *о* - [o] realizados como [ɐ] (4b, 4c); *и* - [i] realizado como [ɪ] (4d); e э - [ɛ] realizado como [ɛ], [ɪ] e [ɨ] (4e):

(4a) *у*	[ʊ]	***у**рок*	[ʊ̱rók] ‘aula’
(4b) *а*	[ɐ]	***а**ллофон*	[ɐ̱łɐfón] ‘alofone’
(4c) *о*	[ɐ]	***о**кно*	[ɐ̱knó] ‘janela’
(4d) *и*	[ɪ]	***и**гра*	[ɪ̱grá] ‘jogo’
(4e) э	[ɪ]	***э**таж*	[ɛ̱táʐ]/ [ɪ̱táʐ]/ [ɨ̱táʐ] ‘andar’

2.2. **Pronúncia das consoantes**

Na língua russa, as consoantes são organizadas em pares de fortes-suaves, com exceção de [ɕ, ʐ, t͡s] - que são apenas fortes, e [t͡ɕʲ, ɕːʲ, j] - que são apenas suaves. Podemos ver todas as consoantes fortes/suaves na Tabela 3.

Tabela 3. Divisão das consoantes em fortes-suaves e pares-ímpares (adaptado de Litnevskaya, 2006)

Consoantes	**Par**	**ímpar**
fortes	[b, v, g, d, z, k, ł, m, n, p, r, s, t, f, x]	[ɕ, ʐ, t͡s]
suaves	[bʲ, vʲ, gʲ, dʲ, zʲ, kʲ, lʲ, mʲ, nʲ, pʲ, rʲ, sʲ, tʲ, fʲ, xʲ]	[t͡ɕʲ, ɕːʲ, j]

As consoantes suaves distinguem-se das consoantes fortes devido a uma articulação adicional - a **palatalização** que se sobrepõe à articulação básica de uma consoante.

Da mesma forma, existe uma divisão das consoantes em vozeadas - pronunciadas com a vibração das cordas vocais e surdas - pronunciadas sem qualquer vibração das cordas vocais. A Tabela 4 apresenta as diferentes consoantes divididas em vozeadas, surdas e pares/ímpares.

Tabela 4. Divisão das consoantes em vozeadas-surdas e pares-ímpares (adaptado de Litnevskaya, 2006)

Consoantes	**par**	**ímpar**
vozeadas	[b, bʲ, v, vʲ, g, gʲ, d, dʲ, z, zʲ, ʐ]	[j, ɬ, lʲ, m, mʲ, n, nʲ, r, rʲ]
surdas	[p, pʲ, f, fʲ, k, kʲ, t, tʲ, s, sʲ, ʂ, x, xʲ]	[t͡s, t͡ɕ, ɕː]

A distinção das **consoantes fortes *versus* suaves** pode ser feita através da grafia. No russo, o [j] em superescrito nas vogais [ʲa], [ʲu], [ʲo] e [ʲe] indica uma propriedade de uma consoante precedente que é suave. Vogais como [i] e [ɪ] também indicam que a consoante precedente é suave:

(5a)	***мал***	[**m**áɬ]	‘pequeno’	*versus*	***мял***	[**m**ʲáɬ]	‘ele amassou’
(5b)	***мол***	[**m**óɬ]	‘cais’		***мёл***	[**m**ʲóɬ]	‘giz’
(5c)	***пэр***	[**p**ér]	‘par’		***перо***	[**p**ʲeró]	‘caneta’
(5d)	***буря***	[**b**úrʲa]	‘tempestade’		***бюро***	[**b**ʲʉró]	‘escritório’
(5e)	***мыло***	[**m**ɨ́ɬə]	‘sabão’		***мило***	[**m**ʲíɬə]	‘agradável’

O sinal gráfico <ь>[1] indica, também, que a consoante precedente par é suave e ocorre em:

— no fim de uma palavra - *конь* [kó**n**ʲ] ‘cavalo’;

— no meio de uma palavra, depois da consoante [l] e seguido por qualquer outra consoante - *полька* [pó**l**ʲkə] ‘dança polonesa’;

— depois de uma consoante vozeada e antes de uma consoante surda - *весьма* [vʲɪ**s**ʲmá] ‘muito’;

— e quando uma consoante suave é seguida por uma das consoantes [gʲ], [kʲ], [bʲ] e [mʲ] - *серьги* [sʲɪ**r**ʲgʲɪ́] ‘brincos’.

Para além disso, quando o sinal gráfico <ь> se encontra antes de [ʲa], [ʲu], [ʲo] e [ʲe], indica que a consoante precedente é suave, mas não suaviza a própria consoante (a consoante e a vogal ocorrem em sílabas separadas, uma vez que o segmento palatalizado [j] se encontra na posição de ataque):

(6a)	***пьян***	[**p**ʲ**j**án] ‘bêbedo’
(6b)	***вьюга***	[**v**ʲ**j**ú**g**ə] ‘nevão’
(6c)	*бульон*	[bʊ**l**ʲ**j**ó**n**] ‘caldo’
(6d)	*ателье*	[ɐtʲɪ**l**ʲ**j**é] ‘Atelier’

Embora a ortografia não nos forneça informações sobre certas consoantes, o facto de elas ocorrerem em certas posições significa que sabemos como pronunciá-las. Isso acontece em contextos como:

[1] O sinal gráfico <ь> também tem a função de marcador de infinitivo quando ocorre em verbos infinitivos e não suaviza a consoante precedente - *учиться* [ʊt͡ɕɪ́**ts**ə] ‘estudar’.

— [n] tem a pronúncia de [nʲ] quando a consoante seguinte é um [t͡ɕʲ] ou [ɕːʲ]:

(7a) *бараба**н**чик* [bərʁbá**n**ʲt͡ɕʲɪk] ‘bateria’
(7b) *бараба**н**щик* [bərʁbá**n**ʲɕːʲɪk] ‘baterista’[2]

— [s] tem a pronúncia de [sʲ] quando a consoante seguinte é [nʲ] ou [tʲ]:

(8a) *ко**с**ть* [kó**s**ʲtʲ] ‘osso’
(8b) *пе**с**ня* [pʲé**s**ʲnʲə] ‘canção’

— [z] tem a pronúncia de [zʲ] quando a consoante seguinte é [nʲ] ou [dʲ]:

(9) *жи**з**нь* [ʐʲí**z**ʲnʲ] ‘vida’

— os fones alveolares dentais fortes, exceto [n] e [ɬ], que precedem os fones labiodentais suaves, podem sofrer um processo de palatalização[3]:

(10a) ***д**верь* [**d**ʲvʲérʲ] / [**d**vʲérʲ] ‘porta’
(10b) *ра**з**ве* [rá**z**ʲvʲe] / [rá**z**vʲe] ‘é’

Em oposição, as **consoantes fortes pares** podem ser identificadas através de:

— ausência do sinal gráfico <ь> que indica que a consoante precedente é forte (não palatalizada):

(11) ба**н**ка [bá**n**kə] ‘jarro’

— quando seguidas de vogais [a, o, ɛ, u, ɨ,][4]:

(12a) ***м**ал* [**m**áɬ] ‘pequeno’
(12b) ***м**ол* [**m**óɬ] ‘cais’
(12c) ***п**эр* [**p**ér] ‘par’
(12d) ***б**уря* [**b**úrʲa] ‘tempestade’
(12e) ***м**ыло* [**m**íɬə] ‘sabão’

— também, o sinal gráfico <ъ> (distinto de <ь>) antes de [ʲa], [ʲu], [ʲo] e [ʲe] indica que a consoante precedente é forte:

(13a) *об**ъя**ть* [ʋ**bjá**tʲ] ‘abraçar’
(13b) *ко**нъю**нкт* [kʋ**njú**nkt] ‘conjunto’

[2] Isto também pode ser explicado através do processo de realização/não-realização dos yers nas línguas eslavas. Os yers alternam com o zero nas sílabas iniciais das palavras, nas sílabas finais das palavras, nas posições prefixais e no meio da palavra. Quando o yer não se reflete na ortografia, pode ainda assim influenciar a pronúncia, afetando a consoante anterior. Para mais pormenores sobre este assunto, consulte Yearley (1995).
[3] De acordo com o *Dicionário Ortográfico* de Borunova, Vorontsova, Eskova e Avanesov (1989), os fones alveolares dentais fortes, exceto [n] e [ɬ], que precedem os sons labiodentais suaves, passam obrigatoriamente por um processo de palatalização. Atualmente, esta realização é rara, e o som forte é preferido.
[4] Em alguns estrangeirismos, uma consoante forte pode ser seguida de [ʲe] - *б**е**кон* [bʲ**e**kón] ‘bacon’.

(13c)	***адъюнкт***	[ɐ**dju**nkt]	‘adjunto’
(13d)	***съест***	[**sjé**st]	‘vai comer’
(13e)	***съёмка***	[**sjó**mkə]	‘sessão’

2.2.1. **Assimilação das consoantes**

Em russo, na grafia existem sequências de duas obstruintes consecutivas. Para simplificar a pronúncia, muitas vezes o primeiro segmento assimila a articulação ou as propriedades acústicas da consoante seguinte, resultando em duas consoantes da mesma qualidade e, consequentemente, na assimilação completa, apagando uma das consoantes.

Podemos observar na Tabela 5, os exemplos de assimilação de consoantes, extraídos de Litnevskaya (2006).

Tabela 5. Assimilação consonântica (adaptado de Litnevskaya, 2006)

Grafemas	**Assimilação das consoantes**				**Significado**
с + ш	[s] + [ʂ]	[ʂː]	***сш****ить*	[**sː**ɨtʲ]	‘costurar’
з + ж	[z] + [ʐ]	[ʐː]	***изж****ить*	[ɪ**ʐː**ítʲ]	‘livrar-se de’
т + ц	[t] + [ts]	[tsː]	***отц****епить*	[ɐ**tsː**ɪpʲítʲ]	‘desengatar’
т + ч	[t] + [tɕ]	[tɕː]	***отч****ет*	[ɐ**tɕː**ót]	‘relatório’
с + щ	[s] + [ɕː]	[ɕː]	*расщ****е****пить*	[rə**ɕː**ɪpʲítʲ]	‘dividir’
с + ч	[s] + [tɕ]	[ʂːʲtɕ]	***с ч****ем-то*	[**sːʲtɕ**émtə]/ [**sː**ʲémtə]	‘com alguma coisa’
т + с	[t] + [s]	[ts]	*мы****тьс****я*	[mɨ**ts**ə]	‘lavar-se’
т + щ	[t] + [ɕː]	[tɕɕː]	***отщ****епить*	[ɐ**tɕɕː**ɪpítʲ]	‘descolar’

Na primeira coluna da tabela, são apresentados os grafemas que serão assimilados, na segunda coluna, a respetiva realização desses grafemas e, na terceira coluna, a realização das consoantes após a assimilação.

Existem também contextos de **assimilação simples**, na qual uma consoante assimila apenas o vozeamento, como em *бу****д****ка* <bu**d**ka> que se transforma em [bú**t**kə] ‘cabine’. A consoante [d] assimila o traço de [-vozeado] da consoante seguinte [k], resultando em [t]. As consoantes surdas pares antes das consoantes vozeadas (exceto [v, vʲ, jʲ, ł, lʲ, m, mʲ, n, nʲ, r, rʲ]) transformam-se em vozeadas, pois assimilam o traço da consoante seguinte, como em *моло****т****ьба* <małat**b**á> → [məłɐ**d**ʲbá] ‘espancamento’ e ***с****дела****т****ь* <**s**délat> → [**z**ʲdʲélətʲ] ‘fazer’.

Além disso, existe um fenómeno de **neutralização da obstruinte** em posição final de palavra:

— a consoante /v/ é realizada como [f], como em *архи****в*** <arxi**v**> → [ɐrxʲí**f**] ‘arquivo’

— a consoante /b/ é realizada como [p], como em *хле****б*** <xlé**b**> → [xlʲé**p**] ‘pão’

— a consoante /z/ é realizada como [s], como em *указ* <ukáz> → [ʊká**s**] ‘decreto’

— a consoante /ʐ/ é realizada como [ʂ], como em *экипа****ж*** <ekipá**ʐ**] → [ɛkʲɪpá**ʂ**] / [ɪkʲɪpá**ʂ**] / [ɨkʲɪpá**ʂ**] ‘tripulação’.

Ao contrário da assimilação, existe na língua russa o fenómeno de **dissimilação consonântica**. Podemos observar este fenómeno fonológico na presença de [g] mais [k], em que [g] dissimila-se de [k] e transforma-se em [x], como em *легко* <le**g**ko> → [lʲí**x**kə] ‘facilmente’ ou *мягкий* <ma**g**kij> → [mʲæ**x**kʲɪj] ‘suave’.

2.2.2. **Apagamento das consoantes**

Esta secção apresentará um conjunto de estruturas em que a combinação de certas consoantes favorece o apagamento de uma delas quando a palavra é realizada. De acordo com Litnevskaya (2006), há casos em que

uma palavra na grafia apresenta três consoantes consecutivas, mas, quando pronunciada, uma delas é apagada (normalmente uma consoante que se encontra no meio). Este processo resulta em sílabas separadas, com uma consoante em posição de coda e a outra em posição de ataque.

Tabela 6. Apagamento das consoantes (adaptado de Litnevskaya, 2006)

Grafemas	**Apagamento das consoantes**				**Significado**
стл	[stɬ]	[sɬ]	*сча**стл**ивый*	[ɕːʲ**sl**ívɨj]	‘feliz’
стн	[stn]	[sn]	*ме**стн**ый*	[mʲé**sn**ɨj]	‘local’
здн	[zdn]	[zn]	*по**здн**ий*	[pózʲ**n**ʲij]	‘atrasado’
здц	[zd͡ts]	[s͡ts]	*под у**здц**ы*	[pʁdʊ**s͡ts**ɨ]	‘sob as rédeas’
ндш	[ndʂ]	[nʂ]	*ла**ндш**афт*	[ɬʁ**n**sáft]	‘paisagem’
нтг	[ntg]	[ng]	*ре**нтг**ен*	[rʲɪ**ng**ʲén]	‘Raio-X’
ндц	[nd͡ts]	[n͡ts]	*голла**ндц**ы*	[gʁɬá**n͡ts**ɨ]	‘holandeses’
рдц	[rd͡ts]	[rt͡s]	*се**рдц**е*	[sʲé**rt͡s**ɨ]	‘coração’
рдч	[rd͡tɕʲ]	[rt͡ɕʲ]	*се**рдч**ишко*	[sʲɪ**rt͡ɕ**íʂkə]	‘coração pequeno
лнц	[ɬn͡ts]	[n͡ts]	*со**лнц**е*	[só**n͡ts**ə]	‘sol’
V1[ɪ]	[V1jɪ]	[V1ɪ]	*мое**го***	[mə**ɪ**vó]	‘do meu’

Na primeira coluna da Tabela 6, podemos observar os grafemas e, na segunda coluna, a sua pronúncia correspondente, na terceira coluna é apresentada a pronúncia após o apagamento de uma das consoantes.

3. **Disfluências**

Como Moniz (2013) mencionou, as disfluências são uma parte crucial da fala humana espontânea (Levelt, 1983; Allwood et al., 1990; Swerts, 1998; Clark e Fox Tree, 2002). Estas são usadas na estruturação do discurso para planear o discurso subsequente e podem introduzir informação nova (Clark e Fox Tree, 2002; Arnold et al., 2003).

Os falantes filtram automaticamente as disfluências, o que parece ser um desafio para os sistemas ASR. Segundo Shriberg (2001, p. 153):

> “There is good reason to study disfluencies, in both theoretical and applied fields. They are frequent - affecting up to ten percent of words and over one third of utterances in natural conversation. For the study of human language, disfluencies provide a window onto underlying processes affecting human speech and language production. On the applied side, disfluencies present a challenge for automatic speech processing, especially since speech recognition models are often trained on read or highly constrained speech (Butzberger et al., 1992).”

Como discutido em Schettino (2022), para melhorar o reconhecimento automático da fala humana espontânea, alguns estudos propuseram inicialmente filtrar as disfluências, detetando-as e cortando-as das transcrições (Bear et al., 1992; Gabrea & O'Shaughnessy, 2000; Heeman & Allen, 1994; Nakatani & Hirschberg, 1994; Oviatt, 1995). Desta forma, os resultados seriam transcrições "limpas" e "perfeitas", mas longe da produção. Consequentemente, a investigação sobre disfluências começou a ganhar interesse, uma vez que se trata de um mecanismo natural da fala espontânea e é uma parte crucial do processamento natural. Na ASR, os investigadores começaram a desenvolver modelos incluindo disfluências, que melhoraram o reconhecimento e o processamento do discurso espontâneo (Heeman & Allen, 2001; Hough & Purver, 2012; Liu et al., 2006). Assim, as disfluências deixaram de ser consideradas um obstáculo, mas um mecanismo linguístico que precisava de ser estudado e incluído na ASR em vez de ser simplesmente eliminado. A sua

inclusão também melhorou a interação homem-máquina, uma vez que é mais natural e assemelha-se à fala humana (Adell et al., 2012; Betz, 2020).

Acreditamos que a investigação sobre as pausas preenchidas na língua russa pode melhorar o sistema de reconhecimento da VoiceInteraction, reduzindo o tempo de processamento e aumentando a qualidade, uma vez que as disfluências, em geral, e as pausas preenchidas, em particular, afetam não só a palavra a ser reconhecida, mas também as sequências adjacentes.

3.1. **Pausas preenchidas**

O presente trabalho focou-se no estudo das pausas preenchidas, visto que estas são o tipo de disfluência mais frequentemente utilizado pelos falantes nos dados de fala analisados, e os anotadores humanos apresentaram algumas dificuldades na identificação das mesmas. Tal como estudado por Barczewska e Igras (2012), a frequência das pausas preenchidas depende em grande medida dos falantes e das tarefas a realizar. No entanto, um falante pode atingir até 10 pausas preenchidas por minuto. Uma vez que os anotadores humanos não transcrevem as pausas preenchidas, a não identificação das pausas causa problemas no ASR.

As pausas preenchidas são usadas pelos falantes para sinalizar reparações e planear as unidades seguintes, e são marcas de trabalho de formulação. São utilizadas em situações de pesquisa lexical, por exemplo, quando o falante ainda está a pensar na palavra correta a produzir, sendo frequente produzir uma pausa preenchida em vez de uma pausa silenciosa. As pausas preenchidas são utilizadas na interação comunicativa para manter ou retomar a palavra, e podem variar em duração, apresentando produção mais longa ou mais curta na formulação do discurso (Moniz, 2013).

Kharlamova (2008) estudou as pausas preenchidas da língua russa através da análise de questionários preenchidos por falantes nativos de russo, análise acústica do discurso na rádio, na televisão e em conferências, e observações de fala em comunicação direta. O estudo revelou as seguintes pausas preenchidas: [e], [ɨ], [m], [n], e [n']. Teng (2015) realizou um estudo de fala espontânea recolhida de sujeitos russos, chineses e americanos, para demonstrar pausas preenchidas universais e específicas de cada língua. Através da medição dos formantes e da observação do espectrograma, as pausas preenchidas utilizadas predominantemente pelos falantes de russo são os sons [a], [am] e [m]. O trabalho de Bogdanova-Belgarian e Baeva (2018) teve como objetivo listar vários elementos não verbais do The Speech Corpus of the Russian Language, entre os quais as pausas preenchidas. Segundo as autoras, além das pausas preenchidas descritas acima, sons [ə] e [əm] também são frequentemente utilizados pelos falantes de russo como pausas preenchidas.

Desta forma, as pausas preenchidas do russo, segundo a literatura, são as apresentadas na Tabela 7. Estas são precedidas de símbolo de percentagem " % ", visto que é desta forma que são sinalizadas pelos anotadores humanos numa transcrição na VoiceInteraction.

Tabela 7. Pausas preenchidas da língua russa de acordo com Kharlamova (2008), Teng (2015), Bogdanova-Belgarian e Baeva (2018) e Teng e Androsova (2022)

Russo	**IPA**	**X-SAMPA**
%а	[a]	[a]
%ам	[am]	[am]
%@	[ə]	[@]
%@м	[əm]	[@m]
%э	[e]	[e]
%ы	[ɨ]	[1]
%м	[m]	[m]
%н(ь)	[n] / [nʲ]	[n] / [n']

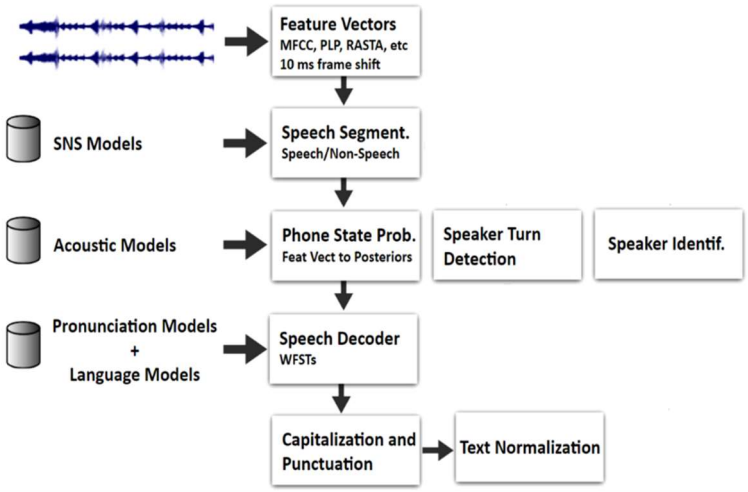
4. Sistema de Reconhecimento Automático de Fala

As tarefas desenvolvidas ao longo do presente trabalho foram aplicadas ao sistema de reconhecimento automático de fala – *Audimus* (Mendes et al., 2019). O *Audimus* é um sistema de reconhecimento de fala, baseado numa arquitetura modular, que recorre a um descodificador de fala baseado em Transdutores de Estado Finito (Weighted Finite-State Transducers - WFSTs), para combinar as probabilidades estimadas pelo modelo acústico, ao nível fonético, com aquelas resultantes do modelo de língua, ao nível da palavra, sendo o modulo de estimação das pronúncias das palavras a ponte entre esses dois módulos. As probabilidades fonéticas são estimadas recorrendo a uma abordagem híbrida, que usa as capacidades de modelação de sequências temporais dos Modelos de Markov Não Observáveis (Hidden Markov Models - HMMs) com as capacidades de identificação e segmentação de padrões das Redes Neuronais (Deep Neural Networks - DNNs). Podemos observar a arquitetura do sistema na Figura 1. Inicialmente, a fala é parametrizada em coeficientes acústicos, calculados a cada 10 ms (milissegundos). Em seguida, aplicando os modelos SNS (*Speech/Non-Speech*), a fala é segmentada. Desta forma, toda a fala humana é preservada e os segmentos que não são fala são filtrados, por exemplo eventos acústicos, ruídos e pausas. Após este processo, são utilizados modelos acústicos baseados em DNN-HMMs, para a realização da probabilidade dos fones das partes de fala do áudio. Nesta fase, o sistema estima as probabilidades de cada um dos fones condicionadas aos valores dos coeficientes acústicos de cada vector extraído anteriormente. Por exemplo, na primeira amostra da fala (t = 1), a onda sonora pode ter as seguintes probabilidades de fones - [e] - 80%, [g] - 10%, [i] - 10%. Utilizando os modelos acústicos, o sistema atribui a probabilidade fonética para descodificar o que falante pode ter dito foneticamente.

Posteriormente, utilizando a probabilidade fonética do passo anterior, o sistema descodifica os fones e as palavras através de modelos de pronúncia e de língua. Neste ponto, o sistema já possui a transcrição da fala humana, por exemplo, “>>[Ana] meu nome é ana”. O sinal >>, o exemplo anterior assinala a mudança de falante e [Ana] é o nome do falante que a deteção de tomada de palavra do falante (*Speaker Turn Detection*) e a identificação do falante (*Speaker Identification*) identificaram.

O resultado do módulo anterior ainda não é o produto final, porque é necessário grafar a maiúscula e pontuar a transcrição. O modelo de grafar as maiúsculas é treinado para grafar as primeiras letras das palavras em casos como a primeira letra do nome próprio ou a primeira letra de uma cidade, por exemplo. Os modelos de pontuação são treinados para pontuar as frases, de acordo com as regras gramaticais de cada língua. Assim, “>>[Ana] o meu nome é ana” é alterado para “>>[Ana] O meu nome é Ana”. O sistema da VoiceInteraction identifica três tipos de pontuação . , e ? , pelo que teria acrescentado o ponto no final do discurso – “>>[Ana] O meu nome é Ana.” O último módulo compreende a normalização do texto. Por exemplo, o que teria acontecido se o sistema de reconhecimento tivesse de compreender dígitos numéricos? Se este fosse o último passo, o sistema teria transcrito o dígito numérico 9 <nove> como “nove”, porque o vocabulário do modelo linguístico não possui numerais. Por isso, o sistema, no final de todos os processos de reconhecimento de fala, tem uma etapa de “Normalização do texto”, em que todos os dígitos são transcritos para dígitos numéricos – “nove” para 9.

Figura 1. Arquitetura do Sistema de Reconhecimento de Fala da *VoiceInteraction* - *Audimus* (apresentação oral de Sérgio Paulo no JEEC (2019))



Por último, o sistema é capaz de detetar e identificar as tomadas de palavra de cada falante. Assim, quando um falante chamado "Ana" está a falar, o sistema é capaz de identificar que é a Ana que está a falar e agrupar todas as suas falas do mesmo falante para ">>[Ana]...." (">>[Ana] O meu nome é Ana.").

5. Metodologia

Esta secção descreverá as etapas realizadas na análise da língua russa, que consistiu em quatro componentes principais:

i. A recolha de dados de fala e texto das diferentes regiões de língua russa;

ii. Revisão das transcrições de noticiários;

iii. Descrição do léxico russo previamente existente na *VoiceInteraction* e revisão das entradas mais frequentes do léxico, juntamente com os ajustes para pronúncias alternativas;

iv. Criação dos procedimentos metodológicos a utilizar na categorização das pausas preenchidas, encontradas ao longo das transcrições e não descritas anteriormente na literatura, tanto quanto nos é dado conhecer.

5.1. Recolha de dados

Foram recolhidos dados de fala e dados de texto de conteúdo noticioso das variedades da língua russa que continham reportagens, entrevistas de rua e debates políticos. O sistema utilizado para este trabalho já havia sido treinado com os dados de russo da Rússia europeia. Assim, foram recolhidos mais dados da variedade da Rússia europeia, incluindo agora também as variedades da Bielorrússia e do Cáucaso. Na Tabela 8, podemos observar os dados de fala recolhidos para cada variedade – 306 horas de fala extraídos de canais de televisão disponíveis em direto, de plataforma de *Youtube* e ainda de estações de rádio. Quanto aos dados de texto, foram recolhidos cerca de 500 milhões de palavras extraídos de jornais, artigos e corpora.

Tabela 8. Dados de fala recolhidos de Rússia europeia, Bielorrússia e Cáucaso

Regiões	**Dados de fala**
Rússia europeia	272h26m24s
Bielorrússia	24h44m40s
Cáucaso	8h48m59s
Total	306h00m03s

O **perfil dos falantes** dos dados recolhidos neste trabalho era constituído por: falantes nativos de língua russa de região de Rússia Europeia; falantes de língua segunda como no caso de falantes de Bielorrússia, visto que o russo faz parte das línguas oficiais do país; e falantes de língua franca provenientes do Cáucaso, já que o russo era falado na União Soviética e continua até hoje a ser utilizado pelos falantes.

5.2. **Transcrição dos dados**

Os dados de fala foram processados na plataforma proprietária, *Calligraphus*, na qual foram transcritos automaticamente e revistos por 4 anotadores humanos, falantes nativos de língua russa e com experiência na área. Para a realização de uma transcrição de qualidade, a *VoiceInteraction* criou um manual de anotação com todas as regras que os anotadores têm de seguir ao transcrever os dados de fala. Este manual já existia para as outras línguas, mas foi necessário verificar se o mesmo podia ser aplicado para a língua russa ou se eram necessárias algumas modificações ao mesmo. A título ilustrativo, algumas das regras do manual de anotação podem ser observadas na Tabela 9.

Tabela 9. Alguns dos símbolos utilizados pelos anotadores humanos no *Calligraphus*

Símbolo	**Uso**	**Exemplo**
[NOTRANS]	usado para indicar conteúdo impercetível ou conteúdo fora do domínio a transcrever (i.e., publicidade, música, entre outros.) que ocorra em grande parte ou na totalidade de um segmento de fala	*"Precisa de um aparelho auditivo? Ligue para o número apresentado no ecrã e aproveite a promoção de hoje."* > ***[NOTRANS]***
[UNK]	usado para indicar fragmentos de fala impercetíveis	*"Nós **[UNK]** escola."*
#	usado para indicar uma palavra soletrada	*"Enviei o ficheiro **#PDF**"*
_	usado como separador em palavras soletradas no plural	*"Ele mostrou os **#_P_D_F's**[a]."*
*	usado para indicar palavras truncadas	*"**Aman*** hoje nós vamos ao parque."*
+	used to indicate syncopated forms (for example, when the speakers pronounce "cause", the human annotator should transcribe it as "because" along with the symbol +)	*"We will go now cause it will rain tomorrow."* > *"We will go now **+because** it will rain tomorrow."*
%	usado para indicar pausas preenchidas	*"Eu **%@m** vou apresentar no ecrã."*
=	usado para indicar palavras prolongadas	*"Estas são **as=** novidades de hoje."*

[a] Se o underscore não fosse adicionado, o "s" seria reconhecido pelo sistema como sendo uma letra soletrada > [pedeɛfɛs] ao invés de [pedeɛfʃ].

Os anotadores transcreveram o conteúdo do áudio de acordo com os símbolos apresentados na Tabela 9. Para garantir que o reconhecedor de fala só é treinado com dados de fala de qualidade e que a transcrição é o

mais próxima possível do que foi proferido, também foi verificado se os anotadores seguiram o manual de anotação. Foram utilizadas 4 horas de áudio (um áudio de cada canal de televisão) como amostra para rever o manual de anotação e observar como os anotadores humanos o utilizam, analisar o desempenho do reconhecedor de fala nas diferentes variedades da língua russa e, finalmente, observar o desempenho dos anotadores humanos.

A verificação das transcrições teve como objetivo verificar os seguintes aspetos:

i. verificar se o manual de anotação era replicável para a língua russa;

ii. determinar quais seriam as áreas mais problemáticas para transcrever automaticamente e o seu impacto nos anotadores humanos;

iii. verificar se os anotadores humanos seguiram o manual de anotação;

iv. analisar as abordagens baseadas nos dados recolhidos para melhorar o léxico com a potencial adição de novas palavras e novas pronúncias para palavras já muito frequentes;

v. analisar o conjunto de pausas preenchidas para a língua russa no *Audimus*, com base na literatura.

5.3. **Análise das transcrições**

Analisámos as transcrições das diferentes regiões falantes de língua russa, utilizando o manual de anotação criado para as outras línguas. Assim, foi possível estabelecer as estruturas problemáticas para um reconhecedor de fala transcrever e, subsequentemente, o seu impacto nos transcritores. Áreas como a sobreposição de fala, o ruído de fundo e as disfluências, entre outros mecanismos de fala e eventos acústicos, demonstraram ser complexas de anotar para o reconhecedor de fala e para os anotadores humanos. Deste modo, prosseguimos com a resolução dessas estruturas.

De forma a resolver alguns dos erros mais frequentemente ocorridos nas transcrições, foi criada uma lista dos erros e a sua respetiva resolução. O objetivo é transmitir esta informação aos transcritores, para melhorar o seu desempenho nas transcrições da língua russa e incluí-la no futuro manual de anotação.

Um dos erros mais comuns foi o de não colocação/sinalização das pausas preenchidas. O sistema ASR não reconheceu as pausas preenchidas e os anotadores humanos também não as identificaram ou transcreveram-nas como discurso imperceptível - [UNK] / [NOTRANS]. Podemos observar um exemplo na Figura 2, antes da correção dos erros, e na Figura 3, após a correção.

Figura 2. Erros antes da correção

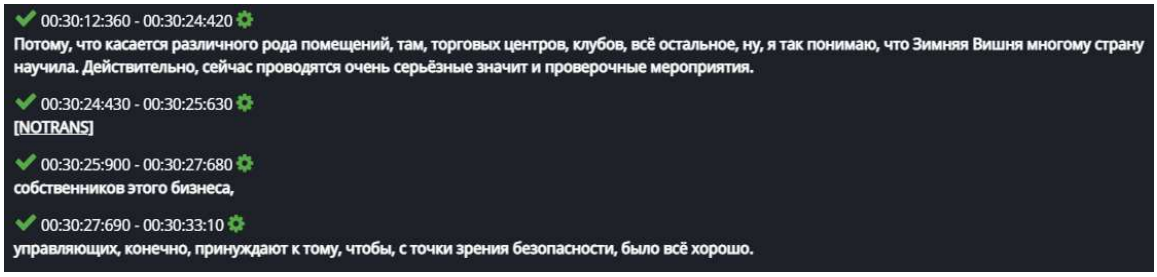
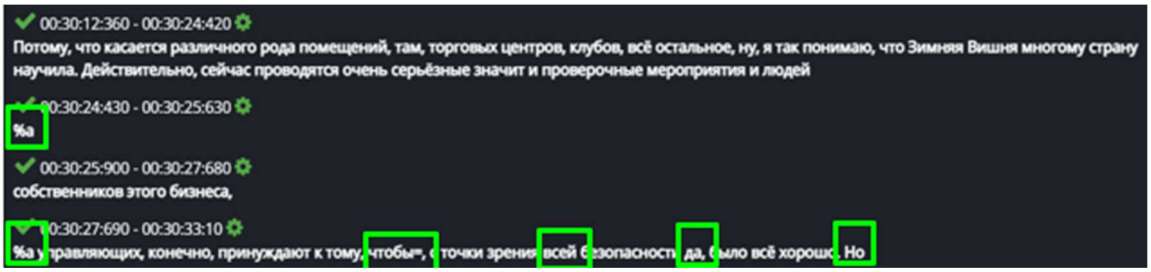


Figura 3. Erros após a correção



Na Figura 2, observamos no minuto 00:30:24:430 que a pausa preenchida foi transcrita como um conteúdo impercetível - [NOTRANS] e no minuto 00:30:27:690 a pausa preenchida - %a, no início do segmento de fala, foi ignorada e não transcrita. Na Figura 3, procedeu-se à correção de todos os erros, assinalados com quadrados verdes.

Para além das pausas preenchidas em falta, podemos observar na Figura 3, no minuto 30:27:690 a palavra "чтобы" 'para' apresenta agora um símbolo "=", indicando que foi pronunciada com um prolongamento. Além disso, foram acrescentadas palavras no mesmo minuto – "всей" 'inteira', "да" 'sim' e "Но" 'mas' - uma vez que nem o sistema, nem o anotador as transcreveu.

As palavras repetidas, interjeições e palavras truncadas também fizeram parte das estruturas problemáticas para os anotadores. Os anotadores humanos tentaram obter uma transcrição "bem estruturada" e por isso, ignoraram as disfluências e os marcadores discursivos para obter as frases gramaticalmente corretas. Também foram encontrados e corrigidos erros como a falta de grafar a maiúscula, normalização ou pontuação. No futuro, será criado o manual de anotação para a língua russa com o mesmo tipo de erros mais comuns, com o objetivo de melhorar o desempenho dos anotadores humanos e do reconhecedor de fala. Até lá, esta informação será transmitida ao departamento responsável pelo contacto com os anotadores da *VoiceInteraction*, para que estes recebam *feedback* e possam aplicá-lo.

As transcrições foram avaliadas utilizado o WER (*Word Error Rate*) que foi calculado através da Equação 1.

Equação 1. Equação usada para calcular o WER

$$WER = \frac{SUBS + DEL + INS}{N}$$

Quanto mais baixo for o WER, melhor a precisão do sistema de reconhecimento. Por exemplo, um WER de 10% significa que 90% da transcrição encontra-se correta. Na equação apresentada acima, SUBS é o número de substituições que foram efetuadas no texto de *output* em comparação com o que foi reconhecido automaticamente, DEL é o número de apagamentos e INS é o número de inserções. Finalmente, N é o número total de palavras de uma transcrição.

5.4. **Léxico**

A *VoiceInteraction* dispunha de um léxico previamente criado para a língua russa, baseado em regras fonéticas da língua russa. O número de entradas do léxico era superior a 400 mil entradas. Estas foram organizadas a partir das mais frequentes nos dados de fala para as menos frequentes. De 409447 pronúncias, 10 mil palavras mais frequentes foram revistas e modificadas, para permitir a extração de padrões a utilizar no módulo estatístico de G2P (*Grapheme-to-phone*). As pronúncias mais frequentes foram usadas como amostra, para observar se as regularidades e regras fonéticas foram seguidas e se as pronúncias eram as pronúncias mais frequentes e fiáveis para cada palavra.

Foi possível detetar erros comuns nas 10 mil pronúncias mais frequentes, como se exemplifica:

— letra "e" em posição final da palavra tinha, em alguns casos, uma pronúncia de [ə] em vez de [e], como em *основное* [ʊsnʌvnójə] > [ʊsnʌvnój**e**] 'principal';

— palavras com a pronúncia de [ɪ] em posição tónica em vez de [i], como em *мило* [mʲílə] > [mʲ**í**ɬə] 'agradável' e vice-versa;

— em algumas pronúncias a partícula palatalizadora [j] antes dos fonemas [ɪ] e [i] encontrava-se em falta, como em *часы* [**t͡ɕ**ɪsɨ́] > [**t͡ɕ**ʲɪsɨ́] 'relógio', por exemplo.

Também foram encontrados erros que ocorreram de forma aleatória, apenas uma ou duas vezes na amostra analisada. Por exemplo, a letra "в", em vez de ter pronúncias correspondentes de [v] / [f] / [vɛ], tinha também uma pronúncia de [vʌstók]. Neste caso, verificamos se esta pronúncia particular - [vʌstók] - estava no léxico associado à palavra correspondente - *восток* 'Este'. A verificação foi realizada com recurso a uma função de linha de comando chamada "grep". Após a verificação, a correspondência entre "в" e [vʌstók] foi eliminada por estar incorreta.

Além disso, foram incluídas pronúncias alternativas de acordo com as variedades da língua russa. Por exemplo, a partícula palatalizadora, de acordo com a variedade, é realizada ou não, como em *следовательно* [**s**ʲlʲédəvətʲɪlʲnə] 'portanto', a primeira consoante [s] é pronunciada com palatalização, ou [**s**lʲédəvətʲɪlʲnə] em que o [s] é pronunciado sem palatalização. A língua russa também apresenta palavras homógrafas. Normalmente, o léxico continha apenas um dos significados. O significado e a pronúncia alternativos foram acrescentados. Por exemplo, *пропасть*, se pronunciado como [**pró**pəstʲ] com o acento na primeira sílaba, tem o significado de 'abismo'. No entanto, se a palavra for pronunciada com o acento na segunda sílaba, como em [prʌ**pástʲ**], significa 'desaparecer'.

Para contagem de número de substituições e modificações ao léxico, foi utilizada uma linha de comando chamada "grep". A expressão regular utilizada para comparar os dois léxicos, antes e depois das modificações, é apresentada na Figura 4:

Figura 4. Expressão regular utilizada para comparar os dois léxicos

```
fgrep -vf lexicon-antes.txt lexicon-depois.txt
```

5.5. **Pausas preenchidas**

A análise das transcrições permitiu identificar as pausas preenchidas mais frequentemente utilizadas pelos falantes de russo, já descritas na literatura e indicadas na Tabela 7.

Foi realizada uma análise acústica e espectrográfica de cada uma das pausas preenchidas. Estas foram extraídas manualmente da amostra dos áudios, utilizando o *software Praat* (Boersma & Weenink, 1999–2022) e guardadas em formato *.wav* para o treino do reconhecedor.

Após a extração de todas as pausas preenchidas, cada pausa foi organizada pela sua duração, de mais curta para a mais longa. Desta forma, o reconhecedor de fala foi treinado com diferentes durações de cada pausa para as reconhecer automaticamente. Além disso, cada pausa foi adicionada ao léxico de língua russa, apresentando pronúncias distintas de acordo com a duração de cada pausa. Por exemplo, uma pausa preenchida *%a* poderia ser pronunciada como [a], mas também como [aaa], com uma duração maior, ou *%ам* que poderia ser pronunciada como [am] e [aam], entre outras variações.

Na literatura, a maioria das pausas preenchidas encontradas já havia sido estudada, com exceção de duas que, até onde sabemos, não haviam sido descritas. Para categorizar as duas pausas preenchidas restantes, utilizámos o *software Praat*, que permitiu extrair o espectrograma e observar a frequência, os formantes, a duração e a intensidade de cada pausa preenchida. Estas serão apresentadas na secção de Resultados.

6. **Resultados**

O presente capítulo irá descrever os resultados da análise das transcrições; os resultados antes e depois das modificações realizadas às 10 mil entradas mais frequentes do léxico; as pausas preenchidas que foram encontradas ao longo das transcrições, juntamente com a sua frequência nas transcrições analisadas; e aplicação do presente trabalho à um projeto Europeu AIDA (Artificial Intelligence and Advanced Data Analysis for Authority Agencies).

6.1. **Análise das transcrições**

Através da análise das transcrições, foi possível verificar que o manual de anotação já existente para as outras línguas também era replicável para a língua russa. Os anotadores humanos seguiram o manual de anotação de outras línguas para transcrever a língua russa, visto que o mesmo não havia sido validado anteriormente.

Por conseguinte, estabelecemos as estruturas problemáticas para transcrever automaticamente e o seu impacto nos transcritores humanos: fala sobreposta, pausas preenchidas, palavras prolongadas, repetições, palavras truncadas, palavras soletradas, pontuação e grafia da letra maiúscula. Para a obtenção de um discurso "bem estruturado", os anotadores humanos eliminaram todos os mecanismos de fala que os falantes utilizam, criando enunciados longe da realidade e não seguindo o manual de anotação fornecido pela *VoiceInteraction*.

Na Tabela 10, podemos observar os resultados de comparação das transcrições automáticas *versus* as transcrições manuais.

Na primeira coluna encontra-se a divisão em três regiões de língua russa: Russo europeu (com cinco ficheiros correspondentes das diferentes campanhas), a região do Cáucaso e a Bielorrússia, com um ficheiro cada. Cada ficheiro apresenta o WER antes e depois da validação das transcrições e a correspondente diferença.

Tabela 10. Comparação das transcrições automáticas *versus* as transcrições manuais

	Antes	**Depois**	**Diferença (%)**
Variedade	**WER (%)**	**WER (%)**	
Rússia Europeia	18,32	18,17	-0,15
	6,86	6,88	0,02
	16,87	16,22	-0,65
	10,67	9,24	-1,43
	12,39	12,48	0,09
Cáucaso	**24,43**	**22,71**	**-1,71**
Bielorrússia	**20,57**	**18,69**	**-1,88**

Nota. A negrito encontram-se os valores de WER que apresentam a maior descida após a validação das transcrições automáticas.

Na maioria dos casos, observamos a diminuição do WER, como esperado. No entanto, também observámos um aumento do WER em dois casos, 0,02% e 0,09%. Embora tenhamos corrigido as transcrições, os modelos não foram suficientemente eficazes para identificar as palavras corretas, potencialmente problemáticas. Pode ser devido aos erros do sistema que estamos a eliminar e as novas alterações são detetadas como incorretas. Além disso, os áudios com sobreposição de fala, disfluências e eventos acústicos influenciam o WER.

Para o russo europeu, o WER mais elevado foi de 18,32%, antes da validação, e de 18,17%, após a validação. A nossa explicação para um WER tão elevado é a presença de debate político no áudio com fala sobreposta, áudio de má qualidade e disfluências, o que resulta em problemas de ASR.

Quanto as variedades do russo do Cáucaso e da Bielorrússia, o WER é superior ao da variedade da Rússia europeia - 24,43% e 20,57%. Uma vez que o ASR só foi treinado com a variedade do russo europeu, foi a primeira vez que teve de reconhecer outras variedades. Após a validação, o WER diminuiu em ambos os casos, 1,71% para a região do Cáucaso e 1,88% para a Bielorrússia. Para diminuir ainda mais o WER, seria necessário ajustar o ASR a cada variedade, por exemplo, ajustar as pronúncias, os modelos acústicos, entre outros, assim como incluir mais dados.

6.2. **Léxico**

O léxico apresentado pela *VoiceInteraction* continha 400 mil entradas, a partir das quais foram revistas e validadas 10 mil palavras mais frequentemente observadas num corpus de texto recolhido de *websites* de jornais de referência russos. A verificação do léxico, juntamente com a adição de novas pronúncias, baseou-se no estado da arte da língua russa, tal como descrito anteriormente. Após a validação do léxico, foram incluídas 145 novas entradas e 86 foram modificadas, totalizando 231 alterações. O léxico recentemente revisto, composto por 10 145 entradas, foi utilizado para avaliar a ASR após as modificações.

Foram efetuadas duas avaliações: antes e depois das modificações do léxico. Os resultados preliminares sem retreino dos modelos acústicos não demostram resultados significativos - 19,85% de WER antes das modificações e 19,97% de WER depois, com uma diferença de 0,12%. Os resultados não refletem imediatamente a melhoria devido a um vocabulário extenso de 400 mil entradas. Além disso, novos erros passaram a ser produzidos pelo ASR. Por exemplo, palavras anteriormente reconhecidas pelo reconhecedor não foram reconhecidas após as modificações efetuadas - a palavra *наверное* 'provavelmente' foi anteriormente reconhecida corretamente, mas após as modificações, foi reconhecida como *наверно* com o último grafema *е* em falta. No entanto, foi possível observar um melhor desempenho do sistema de reconhecimento em algumas pronúncias, corrigindo erros anteriormente produzidos. Algumas dessas palavras eram palavras funcionais – *в*, *и*, *за*, acrónimos – *ВМФ*, adjetivos – *главная*, verbos – *ловлю* e substantivos – *курение*, *семь*.

Embora os resultados preliminares não tenham demonstrado resultados significativos, pudemos proceder à validação do léxico com base na literatura e criar e incluir novas regras fonéticas, descritas na secção de "Metodologia":

- o grafema *е* em posição final de palavra com a pronúncia de [e] (em vez de [ə]);
- [i] sempre em posição tónica (em vez de [ɪ]);
- partícula palatalizadora [j] antes dos fonemas [ɪ] e [i];
- [ɛ] em início de palavra, quando em posição átona com as correspondentes pronúncias de [ɛ], [ɪ] e [ɨ].

Foi notório que as regras fonéticas e o conjunto de fones criados para a língua russa na *VoiceInteraction* seguiram a literatura sobre a fonética da língua russa. É essencial realçar a importância da validação do vocabulário pois, dessa forma, podemos ter um modelo de pronúncia e um modelo de léxico bem construídos,

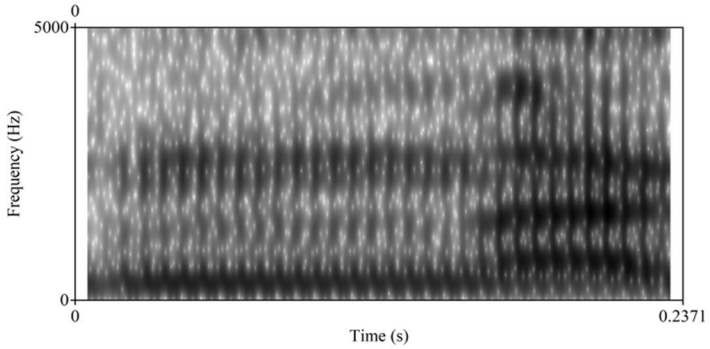
produzindo transcrições no futuro o mais próximo possível das produções reais. A qualidade do léxico não foi previamente verificada na *VoiceInteraction*. Assim, ao revermos o léxico e ao estabelecermos a criação de novas regras fonéticas a serem alargadas às 400 mil entradas garantimos a eliminação de erros sistemáticos.

6.3. **Pausas preenchidas**

Dados de fala foram recolhidos de noticiários transmitidos na Rússia europeia, Bielorrússia e na região do Cáucaso. Os dados recolhidos foram transcritos utilizando a plataforma *Calligraphus*, anotando os mecanismos de fala e as disfluências o mais próximo possível das produções reais.

Após uma análise das transcrições, identificámos duas pausas preenchidas adicionais que, tanto quanto sabemos, ainda não foram descritas na literatura - *%на* e *%мна*. Extraímos o espectrograma de cada uma destas pausas preenchidas encontradas ao longo da análise da transcrição. A primeira - *%на*, é pronunciada como [na], e o seu espectrograma é apresentado na Figura 5.

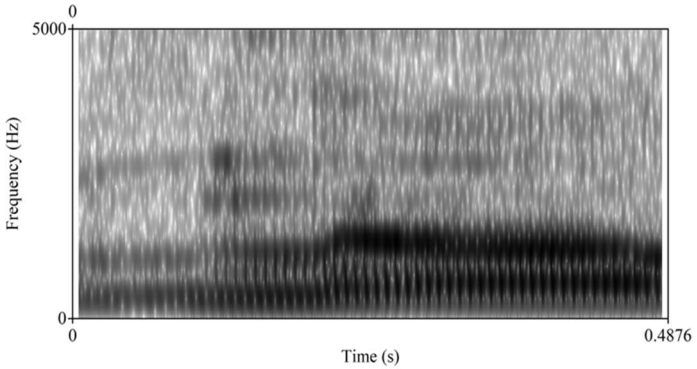
Figura 5. Espectrograma da pausa preenchida *%на*



Na Figura 4, é visível através de energia menos intensa a consoante [n] e através de formantes mais intensos, no final, é visível a vogal [a]. Para a nasal alveolar, [n], tal como apresentado em Ladefoged (2003), o F1 tende a ser baixo - 250-300 Hz e o F2 a rondar os 2500 Hz. Entre os formantes, como podemos observar, também há pouca energia (onde o som [a] tem sua F2). A vogal [a] para o F1 apresentou valores entre 640-720 Hz e 1200-1450 Hz para o F2.

Ao mesmo tempo, observamos a pausa preenchida *%мна,* pronunciada como [mna], apresentada na Figura 6.

Figura 6. Espectrograma da pausa preenchida *%мна*



No início do espetrograma, observamos um segmento com menor intensidade, que é a consoante [m], após esse segmento com maior intensidade é apresentada a consoante [n], e o último segmento apresenta visualmente maior energia, indicando que estamos na presença da vogal [a], também com maior duração. Assim como o som [n], o som [m] apresentou valores de F1 entre 250-300 Hz e F2 em torno de 2500 Hz, com energia entre os dois formantes.

A Tabela 11 apresenta a distribuição das pausas preenchidas em cada ficheiro analisado, juntamente com a distribuição das pausas preenchidas *%на* e *%мна*.

Tabela 11. Distribuição das pausas preenchidas nos ficheiros analisados

Variedade	**Pausas preenchidas (%)**	***%на* (%)**	***%мна* (%)**
Rússia Europeia	0,7	0,04	0
	0,4	0	0
	1,3	0,26	0,05
	3,5	0,31	0,1
	2,7	0,02	0
Cáucaso	**3,3**	**0**	**0**
Bielorrússia	0,7	**0**	**0**

Nota. Na segunda coluna, a negrito encontram-se os ficheiros com maior taxa de frequência das pausas preenchidas observadas; nas colunas terceira e quarta, encontram-se a negrito os valores das variedades nas quais não foram observadas pausas preenchidas %на e %мна.

Na variedade do russo europeu, os ficheiros com a maior frequência de pausas preenchidas apresentam valores de 3,5% e 2,7%. A explicação para este fenómeno deve-se ao tipo de conteúdo dos ficheiros - debate político. Uma vez que o discurso apresentado não foi planeado e foi muito emotivo, favoreceu a produção das pausas preenchidas e, posteriormente, contribuiu para a produção de outros tipos de disfluências, como repetições, truncamentos (devido à fala sobreposta), prolongamentos, erros de pronúncia, entre outros.

Na região do Cáucaso, o ficheiro analisado também apresentou uma alta frequência das pausas preenchidas - 3,3%. Neste caso, o discurso foi maioritariamente proveniente das entrevistas de rua, correspondendo a um discurso espontâneo repleto de diversas disfluências.

De acordo com os resultados, as pausas preenchidas *%на* e *%мна* só foram encontradas na variedade do russo europeu. A produção de *%на* foi mais frequente do que a de *%мна*. No entanto, estas foram produzidas apenas por falantes nativos, como políticos, repórteres, entrevistadores e jornalistas. Na região do Cáucaso e na Bielorrússia, estas pausas preenchidas não foram encontradas. Os resultados podem sugerir que *%на* e *%мна* estão a ser produzidas apenas por falantes nativos de russo. No entanto, para confirmar essa hipótese, seria necessário alargar o conjunto de dados para confirmar a hipótese e analisar outras variedades de russo.

7. **Projeto AIDA**

O projeto *Artificial Intelligence and Advanced Data Analysis for Authority Agencies* (AIDA) é um projeto europeu que visa detetar possível *cybercrime* e terrorismo. A *VoiceInteraction*, enquanto membro do projeto, contribuiu com a transcrição de dados de várias línguas, entre elas o russo, e outras informações estritamente confidenciais. Uma vez que os dados de fala eram conteúdos sensíveis fornecidos pela polícia e confidenciais, a nova versão do *Calligraphus* foi criada especialmente para o AIDA, apenas disponível para as pessoas que trabalham diretamente no projeto.

Os parceiros do projeto forneceram-nos dados no domínio da propaganda terrorista, naturalmente gravados em condições acústicas muito difíceis para as tecnologias de transcrição automática. No total, foram recolhidas 7 horas de dados de fala, e cerca de 3 horas foram transcritas e revistas. Foi um desafio transcrever estes áudios, uma vez que os oradores eram falantes nativos de árabe e por isso possuíam um sotaque forte na língua russa: as palavras tinham pronúncias diferentes do russo europeu; a concordância entre verbo e sujeito, na maioria dos casos, estava ausente; e os oradores usavam frequentemente termos, nomes ou conceitos na sua língua materna – árabe, desconhecidos para nós e inexistentes no vocabulário do sistema.

Como mencionado, foram transcritas cerca de 3 horas de fala e ocorreram 3113 substituições devido a: palavras inexistentes no vocabulário do sistema, por exemplo, *Аллах* 'Alá'; palavras incorretamente reconhecidas; e discurso estrangeiro que não é transcrito, marcado com [NOTRANS], mas que o ASR transcreveu incorretamente como sendo da língua russa.

Tabela 12. Palavras mais frequentemente substituídas nos ficheiros transcritos

Palavra	**SUBS**	**SUBS (%)**
Аллаха	112	3,6
Аллах	100	3,2
И	85	2,7
На	23	0,7
Не	22	0,7
Кафиров	20	0,6
В	20	0,6
По	17	0,5
Мы	16	0,5
И	16	0,5

As substituições mais frequentes (cf. Tabela 12) ocorreram em palavras relacionadas com a religião e em palavras funcionais. Uma vez que o modelo linguístico foi adaptado para o conteúdo noticioso, as palavras de conteúdo religioso não foram reconhecidas. As palavras funcionais também fazem parte das substituições mais frequentes, na maioria dos casos, devido à coarticulação com a palavra prévia ou adjacente. Além disso, as palavras funcionais têm uma duração curta e podem ser absorvidas pelos modelos acústicos durante a descodificação da fala, resultando numa transcrição incorreta.

8. **Conclusões**

Inicialmente, validámos o manual de anotação a ser aplicado ao russo. Através da análise das transcrições de material noticioso, pudemos confirmar as estruturas mais problemáticas tanto para o ASR como para os anotadores humanos, tais como disfluências, fala sobreposta e palavras funcionais. Em trabalhos futuros, tencionamos adaptar o manual de anotação para a língua russa que abordem as estruturas problemáticas e fornecer a sua resolução.

Os áudios com uma maior ocorrência de fala sobreposta e diferentes disfluências, como pausas preenchidas, prolongamentos e truncamentos, apresentaram um WER mais elevado - para o russo europeu o WER mais elevado foi de 18,32% antes da validação e de 18,17% após a validação. Observámos um aumento do WER após as modificações em dois ficheiros - 0,02% e 0,09%. Apesar das nossas modificações nas transcrições, foi difícil para os modelos identificarem as palavras corretas, potencialmente problemáticas. Além disso, o áudio com fala sobreposta, disfluências e eventos acústicos influenciam o WER. As variedades da região do Cáucaso e da Bielorrússia também apresentaram um WER mais elevado - 24,43% e 20,57%, porque o ASR só foi treinado com a variedade do russo europeu. Após a validação das transcrições, foi possível

diminuir o WER em ambas as regiões - 1,71% para a região do Cáucaso e 1,88% para a Bielorrússia. No futuro, o nosso objetivo é ajustar a ASR a cada variedade de língua russa, por exemplo, ajustar as pronúncias, os modelos acústicos, entre outros.

Foi confirmado que o léxico analisado das 10 mil palavras mais frequentes estava de acordo com as regularidades e regras fonéticas russas, exceto em alguns casos, como o grafema "e" em posição final da palavra com a pronúncia de [ə] (em vez de [e]) e [I] em posição tónica (em vez de [i]). No futuro, esperamos alargar as correções às 400 mil entradas, de modo a garantir a eliminação de erros sistemáticos e a obter um modelo de pronúncia e um modelo léxico bem construídos, produzindo transcrições tão próximas quanto possível das produções reais.

Embora os resultados preliminares sem retreino dos modelos acústicos não mostrem resultados significativos - 19,85% WER antes das modificações e 19,97% WER depois, com uma diferença de 0,12%, observámos uma melhoria em algumas das palavras corrigidas, tais como palavras funcionais, acrónimos, adjetivos, verbos e substantivos. No futuro, deve ser efetuada uma análise do reconhecimento de erros para diminuir o WER, como a análise dos erros, uma análise lexical mais alargada e uma análise dos modelos acústicos russos.

Além disso, validámos o conjunto de pausas preenchidas e acrescentámos duas ainda não descritas na literatura, tanto quanto sabemos, através da análise das transcrições - *%на* e *%мна*. De acordo com os resultados, estas só foram produzidas na variedade do russo europeu - *%на* com uma frequência de 0,63% e *%мна* com 0,15%. Os falantes não produziram estas pausas preenchidas na Bielorrússia e na região do Cáucaso. Assim, os resultados sugerem que estas pausas podem estar a ser produzidas apenas por falantes nativos. No entanto, seria necessário analisar um conjunto alargado de dados da Bielorrússia e da região do Cáucaso, juntamente com as outras variedades de língua russa, para confirmar a hipótese.

O trabalho realizado foi aplicado a um projeto europeu – AIDA. Através da validação do manual de anotação, da validação das transcrições, da validação das regras e regularidades fonéticas e da validação das pausas preenchidas, foi possível aplicar o trabalho ao projeto de transcrição da fala. Os dados deste projeto eram difíceis de reconhecer devido ao forte sotaque dos falantes na língua russa, os falantes não seguiam as regras gramaticais - por exemplo, a concordância entre verbo e sujeito encontrava-se em falta e os falantes usavam frequentemente termos, nomes ou conceitos na sua língua materna que não reconhecíamos.

Após a análise das transcrições, pudemos observar um elevado número de substituições. O modelo linguístico foi adaptado ao conteúdo noticioso. Por conseguinte, muitas das palavras russas foram incorretamente reconhecidas ou não foram reconhecidas de todo, uma vez que não faziam parte do vocabulário. As palavras funcionais também foram incorretamente reconhecidas, na maioria dos casos, devido à coarticulação com a palavra anterior ou seguinte, ou devido à sua curta duração - as palavras funcionais podem ser absorvidas pelos modelos acústicos durante a descodificação da fala.

Por último, a validação de sistema ASR da língua russa já contribuiu para o projeto europeu AIDA e ainda para *Young Female Researchers in Speech Workshop* (2022), um evento satélite de *Interspeech*, realizado em Incheon, Coreia do Sul. O presente trabalho reflete a identificação de duas novas pausas preenchidas, informação crucial para os sistemas ASR e para a diminuição do WER. A metodologia utilizada neste trabalho juntamente com os procedimentos, poderão ser replicados para a criação e validação de outros sistemas ASR.

Referências

Adell, Jordi, David Escudero & Antonio Bonafonte (2012) Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication* 54 (3), pp. 459–476. https://doi.org/10.1016/j.specom.2011.10.010

Allwood, Jens, Joakim Nivre & Elisabeth Ahlsèn (1990) Speech management: On the non-written life of speech. *Nordic Journal of Linguistics* 13 (1), pp. 3–48. https://doi.org/10.1017/S0332586500002092

Arnold, Arnold, Maria Fagnano & Michael Tanenhaus (2003) Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research* 32 (1), pp. 25–36. https://doi.org/10.1023/a:1021980931292
Barczewska, Katarzyna & Magdalena Igras (2012) Detection of disfluencies in speech signal. *Challenges of modern technology*, 4 (2), pp. 3–10.
Bear, John, John Dowding & Elizabeth Shriberg (1992) Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 56–63.
Betz, Simon (2020) *Hesitations in spoken dialogue systems*. Tese de doutoramento, Universitat Bielefeld.
Boersma, Paul & David Weenink (1992–2022) *Praat: Doing phonetics by computer* (Versão 6.2.14) [Programa de computador]. Disponível em https://www.praat.org
Bogdanova-Beglarian, Natalia & Ekaterina Baeva (2018) Nonverbal elements in everyday Russian speech: An attempt at categorization. In *Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018)*. Akademija nauk Respubliki Tatarstan, pp. 3–13. Disponível em http://ceur-ws.org/Vol-2303/
Borunova, Svetlana, Vera Vorontsova & Natalia Es'kova (1989) Orfoepicheskii slovar' russkogo yazyka. Proiznoshenie. Udarenie. Grammaticheskie formy [Dicionário ortoépico da língua russa. Pronúncia. Ênfase. Formas gramaticais] (5.ª ed.). Russkii yazyk Publ.
Butzberger, John, Hy Murveit, Elizabeth Shriberg & Patti Price (1992) Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*. Morgan Kaufmann, pp. 339–343.
Clark, Herbert & Jean Fox Tree (2002) Using uh and um in spontaneous speaking. *Cognition* (84), pp. 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3
Demidov, Dmitriy, Vasiliy Klepatskiy, Elena Morozova, Michail Popov, Dmitriy Rudnev, Anastasia Ryko, Dmitriy Cherdakov & Olga Cherepanova (2013) *Istoricheskaya grammatika russkogo yazyka: Khrestomatiya* [Gramática histórica da língua russa: Um livro didático]. Izdatel'stvo Sankt-Peterburgskogo universiteta.
Gabrea, Marcel & Douglas O'Shaughnessy (2000) Detection of filled pauses in spontaneous conversational speech. In *Proceedings Sixth International Conference on Spoken Language Processing (ICSLP 2000)* (Vol. 3), pp. 678–681. https://doi.org/10.21437/ICSLP.2000-626
Kharlamova, Tatiana L. (2008). Sopostavitelnyj analiz zvukov zapolnitelej pauz hezitacii, kak vyrazitelej chuvstvovanij, v russkom i anglijskom yazykah. [Análise comparativa dos sons das pausas preenchidas em russo e inglês, como expressores de sentimentos].
Heeman, Peter & James Allen (1994) Detecting and correcting speech repairs. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 295–302. Disponível em https://aclanthology.org/P94-1041
Heeman, Peter A. & Allen, James F. (2001). Improving robustness by modeling spontaneous speech events. In *Robustness in language and speech technology* (pp. 123-152). Dordrecht: Springer Netherlands.
Hough, Julian & Matthew Purver (2012) Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*. SemDial, pp. 136–144. Disponível em https://www.semdial.org/anthology/papers/Z/Z12/Z12-3018/
Jurafsky, Dan & James Martin (2021). *Speech and language processing*. [Manuscrito em preparação]
Kasatkin, Leonid (2014) *Sovremennyy russkiy yazyk. Fonetika* [Língua russa moderna. Fonética]. (3.ª ed.). Knizhnyi dom 'LIBROKOM'.
Levelt, Willem (1983) Monitoring and self-repair in speech. *Cognition* (14), pp. 41–104.
Litnevskaya, Elena (2006) *Russkiy yazyk: kratkiy teoreticheskiy kurs dlya shkol'nikov* [Língua russa: Um breve curso teórico para estudantes]. Disponível em http://www.gramota.ru/
Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf & Mary Harper (2006) Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE*

Transactions on audio, speech, and language processing, 14 (5), pp. 1526–1540. https://doi.org/10.1109/TASL.2006.878255
Mendes, Carlos, Alberto Abad, João Neto & Isabel Trancoso (2019) Recognition of Latin American Spanish Using Multi-Task Learning. In *Proceedings INTERSPEECH 2019*. ISCA, pp. 2135–2139. Disponível em https://www.isca-speech.org/archive/interspeech_2019/mendes19_interspeech.html
Moniz, Helena (2013) *Processing disfluencies in European Portuguese*. Tese de doutoramento, Universidade de Lisboa.
Nakatani, Christine & Julia Hirschberg (1994) A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95 (3), pp. 1603–1616. https://doi.org/10.1121/1.408547
Osipova, Tamara (2018) *Fonetika sovremennogo russkogo yazyka* [Fonética da língua russa moderna]. Knizhnyy dom "LIBROKOM".
Oviatt, Sharon (1995) Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9 (1), pp. 19–35. https://doi.org/10.1006/csla.1995.0002
Paulo, Sérgio (2019, 15–11 março) *Automatic subtitling: Pushing the limits of speech recognition* [Apresentação de comunicação]. Engineering and Tech Talks - JEEC 2019, Instituto Superior Técnico, Lisboa, Portugal.
Popov, Michail (2014) *Fonetika sovremennogo russkogo yazyka* [Fonética da língua russa moderna]. Filologicheskiy fakul'tet Sankt-Peterburgskogo gosudarstvennogo universiteta.
Schettino, Loredana (2022) *The role of disfluencies in Italian discourse. Modelling and speech synthesis applications*. Tese de doutoramento, Universita Degli Studi di Salerno.
Shriberg, Elizabeth (2001) To ‘errrr’is human: ecology and acoustics of speech disfluencies. *Journal of the international phonetic association* 31 (1), pp. 153–169. https://doi.org/10.1017/S0025100301001128
Sokolova, Anastasia (2021) *Fonetika i fonologiya russkogo yazyka* [Fonética e fonologia da língua russa]. Masarykova Univerzita.
Swerts, Marc (1998) Filled pauses as markers of discourse structure. *Journal of Pragmatics* (30), pp. 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9
Teng, Hai & Svetlana Androsova (2022) Osobennosti pauzacii v rodnoj i akcentnoj spontannoj rechi: akusticheskij analiz kitajskoj rodnoj, russkoj akcentnoj i russkoj rodnoj rechi [Características das pausas na fala espontânea nativa e com sotaque: uma análise acústica da fala nativa chinesa, da fala com sotaque russa e da fala nativa russa]. *Vestnik NGU* 21 (2), pp. 67–86.
Teng, Hai (2015) Universalnye i tipologicheskie cherty pauzacii v spontannoj rechi nositelej raznyx yazykov. [Características universais e típicas das pausas na fala espontânea de falantes de diferentes línguas]. *Teoreticheskaya i prikladnaya lingvistika* [Linguística teórica e aplicada] 1 (2), pp. 105–113.
Yanushevskaya, Irena & Daniel Bunčić (2015) Russian. *Journal of the International Phonetic Association* 45 (2), pp. 221–228. https://doi,org/10.1017/S0025100314000395
Yearley, Jennifer (1995) Jer vowels in Russian. In *University of Massachusetts occasional papers in linguistics* (Vol. 18). University of Massachusetts, pp. 533–571.

Estratégias e praticas de tradução do Padre Joaquim Gonçalves: Uma análise dos dados bilingues preservados na sua trilogia para o ensino-aprendizagem do chinês

Ling Li[1]

[1]Centro de Estudos Humanísticos da Universidade do Minho, Braga, Portugal

Resumo

A tradução assume um papel primordial nas missões evangelizadoras e educacionais dos missionários, sobretudo no que diz respeito ao ensino-aprendizagem das línguas indígenas, porém a sua relevância não tem sido reconhecida nem nos estudos desenvolvidos no âmbito da linguística missionária nem na história dos estudos de tradução. Joaquim Afonso Gonçalves (1781–1841) representa uma das figuras mais relevantes da sinologia europeia, sendo especialmente conhecido pela sua trilogia composta pela *Arte China* (1829), o *Diccionario Portuguez-China* (1831) e o *Diccionario China-Portuguez* (1833), um método inovador de ensino-aprendizagem da língua chinesa destinado aos seus discípulos ocidentais no Extremo Oriente. Os recursos bilingues, contidos neste corpus tripartido, permitem a realização de uma grande variedade de estudos contrastivos, entre os quais o da tradução, uma ferramenta didática de importância colossal que simultaneamente faz parte dos principais objetivos de ensino dos missionários. As estratégias de tradução do autor raramente se encontram explicitamente documentadas nas suas obras: existem apenas dois manifestos no prólogo do seu primeiro dicionário português-chinês que abordam respetivamente a técnica de adaptação à luz de domesticação e o empréstimo no âmbito de estrangeirização, duas estratégias fundamentais na teoria de tradução moderna. Procura-se interpretar estas duas observações do autor juntamente com restantes reflexões também no domínio de tradução, espalhadas nos três volumes interligados, antes de avançar para a análise das respetivas praticas de tradução, com o objetivo de perceção das escolhas e desafios do autor durante a conversão das referências interculturais. O resultado deste trabalho permitirá compreender a importância destes materiais didáticos bilingues para o estudo da tradução entre o português e o chinês.

Palavras-chave: História da tradução, linguística histórica, linguística missionária, tradução chinês-português.

Abstract

Translation plays a crucial role in the mission of evangelization and education of indigenous languages carried out by missionaries. However, its relevance has often been overlooked in studies conducted within the field of missionary linguistics as well as in the history of translation. Joaquim Afonso Gonçalves (1781–1841) stands as one of the most prominent figures in European sinology, particularly renowned for his trilogy composed of *Arte China* (1829), *Diccionario Portuguez-China* (1831), and *Diccionario China-Portuguez* (1833). This innovative language learning method for teaching Chinese to Westerners in the Far East has garnered significant recognition. The bilingual resources encompassed within this tripartite corpus enable a wide range of contrastive studies, including the study of translation, a monumental didactic tool that is also a major objective of missionary teaching. The author's translation strategies are rarely explicitly documented in his works: only two manifestos found in the preface of his first dictionary touch upon the technique of adaptation through domestication and the use of loanwords within the context of foreignization – two fundamental strategies in modern translation theory. This study will firstly interpret these two observations, along with other scattered reflections on translation found throughout the interconnected three volumes. Subsequently, an analysis of the author's translation practices will be conducted, aiming to comprehend the choices and challenges encountered by the author in the process of converting intercultural references. The outcome of this study will shed light on

the significance of these bilingual instructional materials for the study of translation between Portuguese and Chinese.

Keywords: History of translation, historical linguistics, missionary linguistics, Chinese-Portuguese translation.

1. Introdução

Os missionários assumiram funções cruciais para o estabelecimento dos primeiros diálogos interculturais a nível mundial. Durante as missões de evangelização, o ensino-aprendizagem das línguas locais foi uma das abordagens fundamentais dos evangelizadores, através do qual não só conseguiram propagar a fé cristã mediante uma comunicação direta com a população nativa, mas também desenvolveram um conjunto de obras metalinguísticas – fruto do contacto que tiveram com as línguas e culturas locais, anterior às restantes comunidades académicas. No entanto, uma grande parte destes recursos primários ainda não foram estudados da forma que mereciam, quanto menos reconhecidos na história dos estudos linguísticos. A tradução, enquanto ferramenta fundamental para a comunicação intercultural, desempenha um papel crucial na produção de gramáticas, dicionários e outros materiais linguísticos para ensinar e documentar as línguas indígenas, os quais continuam a mostrar-se importantes para pesquisas interlinguísticas uma vez que oferecem um conjunto de recursos linguísticos, que frequentemente são bilingues, de primeira mão sobre as línguas em questão.

Neste estudo, pretende-se examinar e comparar as diferentes abordagens e técnicas de tradução aplicadas aos dados bilíngues português-chinês selecionados dos trípticos do padre lazarista português Joaquim Afonso Gonçalves, dedicado ao enino-aprendizagem do chinês no Macau de inícios do século XIX: *Arte China* (1829), *Diccionario Portuguez-China* (1831) e *Diccionario China-Portuguez* (1833). Com o objetivo de criar um enquadramento teórico, será primeiramente examinada a importância da tradução nas obras missionárias, assim como o contributo destas produções para o estudo da tradução (Secção 2). Em seguida, serão brevemente apresentadas as principais características dos dados bilingues preservados nas obras de Gonçalves (Secção 3) que facilitarão a análise das estratégias e práticas de tradução deste sinólogo com base nas suas reflexões acerca da tradução, diretamente observadas nas suas obras (Secção 4). Por fim, através das considerações finais (Secções 5), espera-se despertar interesse no melhor proveito destes recursos linguísticos e referências interculturais em estudos multidimensionais e contribuir para o desenvolvimento dos estudos de tradução entre o português e o chinês.

2. Tradução e linguística missionária

Os missionários começaram a aprender as línguas indígenas dos destinos das suas missões de evangelização desde o início dos Descobrimentos. Inicialmente o seu objetivo consistia sobretudo na interação com a comunidade local. Posteriormente, com o aumento contínuo de contactos entre as sociedades de línguas e culturas distintas, surgiu também a necessidade de ensino destas línguas. Embora o principal público-alvo tenha sido os evangelizadores futuros, os métodos pedagógicos dos idiomas locais foram também aproveitados por outros elementos deste encontro transcontinental, como é o caso dos diplomatas, funcionários públicos, e tradutores oficiais que trabalhavam na China. Além disso, algumas partes das obras didáticas, fruto das atividades de ensino-aprendizagem das línguas locais, chegaram a ser transportadas de volta para o continente europeu, resultando num conjunto de estudos linguísticos das línguas exóticas realizado por académicos que nunca tiveram contacto direto nem com as línguas alvo do estudo nem com as respetivas culturas, recorrendo somente às fontes missionárias.

O método tradução-gramática, também conhecido como método tradicional, era o método principal anteriormente ao comportamentalismo marcado pela criação do método natural a partir das novas conceções linguísticas e didáticas assinaladas pela apresentação do método de ensino audiolingual por Bloomfield em 1942 (Bruscato & Baptista, 2020, p. 327). Deste modo, não é surpreendente que a tradução tenha assumido um

papel fundamental nas obras missionárias do século XIX, como é o caso de Joaquim Gonçalves, embora o seu método já apresentasse inovações[1] no que diz respeito ao ensino-aprendizagem da língua chinesa através não só dos recursos escritos, mas também dos materiais orais.

A maioria dos missionários muitas vezes já tinham passado pela formação e doutrina em línguas clássicas europeias, pelo que não é inesperado que estes tenham recorrido com ou sem adaptação aos métodos de ensino-aprendizagem dessas línguas. Isto verifica-se não só na valorização de diálogos bilingues, mas também na descrição das regras gramaticais das línguas indígenas. Para ambos os casos, não é surpreendente que a tradução assuma um papel fundamental como uma ponte translinguística. Gianninoto (2014) afirma o seguinte no seu estudo baseado nas obras gramaticais bilingues do ensino-aprendizagem do chinês realizadas por missionários, diplomatas e académicos dos séculos XVIII e XIX:

> Translation in bilingual primers and grammars can be analyzed at two different levels: (1) from the point of view of the history of the didactic ideas and practices, translation was adopted as an important learning approach; (2) from the point of view of linguistic ideas, a considerable number of concepts and terms were translated and transposed (2014, p. 232).

Ou seja, a tradução enquanto auxílio do ensino e elemento constituinte da aprendizagem, aplicava-se em ambos os sentidos, isto é, do chinês e para o chinês, tal como era comum nas teorias e práticas didáticas da altura.

Enquanto ferramenta pedagógica, é expectável que a tradução seja altamente precisa e fiel ao texto de origem, redigido na língua alvo do ensino, tal como vários autores defendem nas advertências aos leitores, sobretudo iniciantes do chinês. Isto faz com que a utilidade da tradução enquanto ferramenta de ensino ultrapasse a importância da elegância e fluidez do texto-alvo (Gianninoto, 2014, p. 235). Devido às diferenças estruturais entre o chinês e uma grande parte das línguas europeias, como é o caso do inglês e das línguas românicas, por vezes o latim é proposto como uma solução ideal devido à liberdade da ordem sintática que facilita a tradução palavra a palavra. Com ou sem apoio do latim como intermediário, era recorrente a tradução interlinear, isto é, a segmentação e tradução subdividida dos elementos sintáticos. Entretanto, em alguns casos, a tradução era incorporada também como parte do objetivo do ensino, daí serem analisadas as diferentes tendências de tradução e ser sublinhada importância de manter o equilíbrio entre a tradução demasiado literal e a tradução demasiado livre.

Embora a tradução tenha sido omnipresente nas gramáticas missionárias para o ensino de línguas locais, a mesma não tem sido valorizada da forma que merece na história das teorias de tradução, conforme explica Zwartjes (2014, p. 2): "It appears to have been felt that missionary linguists did not contribute substantially to 'Western translation theory', so that missionary sources do not deserve to be included in such manuals". Isto deve-se sobretudo aos seguintes fatores: em primeiro lugar, uma grande parte dos estudos gramaticais dos missionários está conservada em arquivos inéditos ou antigos, espalhada pelos cantos das bibliotecas ou estabelecimentos religiosos que necessitam de ser primeiramente descobertos e, posteriormente, reeditados para serem mais úteis e conhecidos; em segundo lugar, nos estudos que abordam a linguística missionária, o principal interesse consistia maioritariamente nas abordagens linguísticas e nos métodos didáticos; em terceiro lugar, nos estudos de tradução, o maior foco vê-se em traduções endógenas do que em traduções exógenas. No entanto, as práticas de tradução dos missionários merecem maior atenção, tal como Zwartjes (2016) defende:

> When translating from Spanish, Portuguese or Latin into the indigenous languages or the other way around, missionaries often came to the conclusion that the source and the target languages show asymmetries. When equivalences were difficult to find or when specific terms were untranslatable, they developed, implicitly or explicitly, different strategies in order to offer the best solution. Not many missionary

[1] O estudo de Levi (2007) apresenta uma análise robusta sobre esta abordagem de ensino inovadora.

grammars include large sections on translation theory, but their translation practices and their observations as to the problems of transmission are worth studying. (2016, p. 65)

Ou seja, as práticas e observações no âmbito de tradução dos missionários valem maior valorização, não só porque representam as primeiras tentativas de solução perante a dificuldade em encontrar correspondências exatas, no sentido matemático, isto é, simétricas e reversíveis, entre as duas línguas em contacto, mas também porque as palavras dedicadas às teorias de tradução de forma explícita são escassas nas suas obras.

Em resumo, a tradução foi amplamente utilizada como ferramenta de ensino-aprendizagem nos estudos gramaticais dos missionários do século XIX, pelo que o estilo de tradução pode variar muito consoante o foco didático. Além disso, a tradução nas obras missionárias costuma ser bidirecional, isto é, tanto da língua materna do autor para a língua alvo do estudo como no sentido contrário. As práticas de tradução dos missionários, enquanto primeiras tentativas de estabelecimento de uma ponte entre as duas línguas em contacto, são importantes não só para o estudo da tradução, mas também para o estudo das teorias de tradução dos missionários, as quais raramente se encontram documentadas de forma explícita nas obras missionárias.

3. **Características dos dados bilingues nas obras de Joaquim Gonçalves**

O núcleo da trilogia para o ensino-aprendizagem do chinês de Joaquim Gonçalves, a *Arte China*, divide-se em oito capítulos interdependentes: "Alphabeto China"; "Frases Vulgares e Sublimes"; "Grammatica"; "Syntaxe"; "Dialogos"; "Proverbios"; "Historia, e Fabula"; e "Composições Chinas", acrescidos de um prólogo que dá conta do seu método de ensino, das suas principais inovações metodológicas em relação aos costumes didáticos locais e das particularidades diatópicas, diastráticas e diafásicas da língua chinesa por si observadas e, por fim, um apêndice intitulado "Arte China sem Letras Chinas com a Pronúncia Mandarina, e de Cantão" que visa oferecer uma alternativa simplificada das duas variedades da língua chinesa, o registo vulgar e o registo sublime, alvo do seu ensino, ao substituir os caracteres chineses por representações fonológicas com letras europeias. Os oito capítulos acima identificados destinam-se ao ensino progressivo dos dois estilos literários do chinês, neste caso, o estilo vulgar, que se aplica na fala quotidiana, e o estilo sublime, que predomina nos recursos escritos, desde os caracteres chineses e os seus elementos constituintes, isto é, os componentes mais básicos desta língua, até à pragmática, competência essencial para o domínio profundo de um idioma. Este compêndio é maioritariamente bilingue, contendo exemplos elaborados paralelamente em português e em chinês; encontram-se apenas em português os prólogos ou advertências que antecedem os primeiros três capítulos e as notas do autor que precedem o oitavo capítulo e o apêndice, assim como as notas de rodapé, que ocasionalmente ocorrem ao longo da leitura, e os títulos categoriais de todos os capítulos menos o quinto.

Os recursos bilíngues português-chinês de Joaquim Gonçalves, alvo da análise do presente estudo, apresentam um conjunto de características próprias em comparação com outras obras metalinguísticas da mesma época. Em primeiro lugar, o ensino é destinado simultaneamente aos falantes de ambas as línguas envolvidas, ou seja, o compêndio não só é dirigido aos falantes portugueses que desejam aprender o chinês, mas também serve para o público-alvo chinês interessado em aprender o português, conforme exposto pelo próprio autor (Gonçalves, 1829, p. 89), diferentemente do que acontece na maioria das gramáticas missionárias ou académicas até ao final do século XIX (Cf. Gianninoto, 2014, p. 232). Esta versatilidade bidirecional faz com que os dados bilingues sejam os mais autênticos e naturais possíveis nos dois sentidos e enriquecidos de esclarecimentos socioculturais e históricos assim como referências interculturais. Em segundo lugar, sendo o método muito diretamente dirigido aos seus alunos, o compêndio não contém muitas explicações que o autor certamente lhes forneceria no decurso das aulas, conforme presume Barros (2014, p. 110), reservando espaço sobretudo para exemplos bilingues. Por este motivo, os leitores atuais têm somente acesso ao corpus bilingue, sem contar com a interpretação e análise que costumam acontecer em obras didáticas da mesma época (Cf. Morrison, 1815; Wade, 1867). Em terceiro lugar, a heterogeneidade dos dados bilingues, desde palavras isoladas a frases e textos compridos em estilos variados, permite a realização de estudos contrastivos das duas

línguas em géneros textuais vastamente distintos. Em quarto lugar, existe uma diversidade de informações interculturais difundidas por meio de exemplos, contextos, intervenções e várias outras formas de comunicação (Cf. Barros, 2014). Em termos específicos da tradução, estas características implicam que o corpus é constituído por recursos linguísticos de géneros variados para a finalidade do ensino de línguas, altamente requintados que procuram incapacitar a didática de ambas as línguas; são expectáveis, embora raros, as informações paratextuais e os esclarecimentos interculturais que aludem ao raciocínio do autor aquando da criação dos equivalentes bilingues.

Os dois dicionários constituintes do método de ensino tripartido apresentam também particularidades que devem ser consideradas para o estudo dos métodos de tradução dos equivalentes bilingues. As entradas do *Diccionario Portuguez-China* são palavras portuguesas alfabeticamente organizadas que dispõem de equivalentes chineses dos estilos oral e escrito; os usos das palavras são fornecidos através de exemplos bilingues; existem por vezes parágrafos reservados a informações interculturais como, por exemplo, as províncias da China e as suas capitais (Cf. Gonçalves, 1831, p. 474), que ocorrem como suplementos didáticos. As entradas do *Diccionario China-Portuguez* são caracteres chineses organizados na ordem do "alfabeto china" – fruto da redução dos 214 géneros a 129 (Gonçalves, 1829, p. IV) de acordo com o número de traços dos mesmos; a pronúncia de cada entrada é oferecida tanto em caracter homófono como em romanização previamente anotada; cada lema normalmente possui um equivalente em português, ou explicação em português quando a equivalência é difícil de alcançar; os usos dos caracteres são fornecidos através de exemplos bilingues; no final encontra-se anexado um *Diccionario Tonico* destinado à consulta simplificada dos caracteres chineses recorrendo à sua pronúncia. Do ponto de vista do estudo de tradução, estamos perante um corpus bilingue enciclopédico e pedagógico que complementa o ensino da gramática disponibilizado na *Arte China*.

4. **Estratégias e práticas de tradução nas obras de Joaquim Gonçalves**

Antes de mais, é imprescindível clarificar algumas noções de tradução a serem aplicadas na seguinte análise, tendo em conta a confusão geral a nível conceptual e terminológico quanto aos nomes utilizados na descrição das práticas de tradução. Debruçar-nos-emos sobre a proposta terminológica de Xiong (2014) que diferencia as estratégias, métodos e técnicas de tradução, conforme demonstrado na Tabela 1.

Tabela 1. Estratégias, métodos e técnicas de tradução na proposta de Xiong

Translation strategies	**Translation methods**		**Translation techniques**
(1) Foreignizing strategy, or Foreignization	1) Zero translation		i.Addition ii.Omission iii.Division iv.Combination v.Shift
	2) Transliteration		
	3) Word-for-word translation		
	4) Literal translation		
(2) Domesticating strategy, or Domestication	5) Liberal/free translation	5a) Paraphrase	
		5b) Idiomatic translation	
	6) Imitation		
	7) Variation translation		
	8) Recreation		

Esta classificação de três níveis segue a recomendação de Molina e Hurtado Albir no que diz respeito à necessidade de distinguir os métodos, estratégias e técnicas de tradução (2002, p. 507); propõe a discriminação das duas estratégias de tradução, isto é, a estrangeirização (foreignization) e a domesticação (domestication), e a subdivisão destas duas categorias principais em oito métodos de tradução, designadamente a tradução zero, transliteração, tradução palavra a palavra, tradução literal, tradução liberal, imitação, variação e recreação, que

correspondem aos oito termos ingleses acima identificados através da numeração 1) a 8), sendo os primeiros quatro categorizados no âmbito da estrangeirização e os últimos quatro no âmbito da domesticação. A prática de domesticação é descrita como "an ethnocentric reduction of the foreign text to receiving cultural values, bringing the author back home" enquanto a prática de estrangeirização é vista como "an ethnodeviant pressure on those values to register the linguistic and cultural differences of the foreign text, sending the reader abroad" (Venuti, 2008), ou seja, a domesticação tem como ponto de partida o público-alvo e procura transmitir os fatores culturais com base nos hábitos linguísticos e culturais da língua de chegada, ao passo que a estrangeirização pretende preservar tanto quanto possível as especificidades linguísticas e culturais do texto de partida. Quanto às técnicas de tradução, Xiong sublinha o âmbito de aplicação deste termo, defendendo que se referem às aplicações específicas dos métodos de tradução no decorrer da atividade de tradução (2014, p. 83), e especifica cinco técnicas de tradução, a saber: adição, omissão, divisão, combinação e mudança, correspondendo aos termos ingleses acima identificados com a numeração i. a v. Ainda que seja uma categorização completa e sistemática, os termos propostos por Xiong nem sempre são suficientes para discriminar as práticas de tradução mais comuns, pelo que serão igualmente tidas em consideração as técnicas de tradução na classificação de Molina e Hurtado Albir (2002, pp. 509–511), conforme exemplificadas na seguinte tabela de transcrição:

Tabela 2. Técnicas de tradução na classificação de Molina e Hurtado Albir

Adaptation	Baseball (E) ⇒ Fútbol (Sp)
Amplification	شهر رمضان (A) ⇒ Ramadan, the Muslim month of fasting (E)
Borrowing	Pure: Lobby (E) ⇒ Lobby (Sp) Naturalized: Meeting (E) ⇒ Mitin (Sp)
Calque	École normale (F) ⇒ Normal School (E)
Compensation	I was seeking thee, Flathead (E) ⇒ En vérité, c'est bien toi que je cherche, O Tête-Plate (F)
Description	Panettone (I) ⇒ The traditional Italian cake eaten on New Year's Eve (E)
Discursive creation	Rumble fish (E)⇒ La ley de la calle (Sp)
Established equivalent	They are as like as two peas (E) ⇒ Se parecen como dos gotas de agua (Sp)
Generalization	Guichet, fenêtre, devanture (F) fi Window (E)
Linguistic amplification	No way (E) ⇒ De ninguna de las maneras (Sp)
Linguistic compression	Yes, so what? (E) ⇒ ¿Y? (Sp)
Literal translation	She is reading (E) ⇒ Ella está leyendo (Sp)
Modulation	ستصير أبا (A) ⇒ You are going to have a child (Sp)
Particularization	Window (E) ⇒ Guichet, fenêtre, devanture (F)
Reduction	Ramadan, the Muslim month of fasting (Sp) ⇒ شهر رمضان (A)
Substitution (linguistic, paralinguistic)	Put your hand on your heart (A) ⇒ Thank you (E)
Transposition	He will soon be back (E) ⇒ No tardará en venir (Sp)
Variation	Introduction or change of dialectal indicators, changes of tone, etc.

Por uma questão de economia, não será discutida em pormenor a definição de cada um dos termos acima identificados, mas serão esclarecidos os necessários para a análise das estratégias e práticas de tradução observadas nas obras acima mencionadas do padre Joaquim Gonçalves, e para isso, apresentar-se-ão como ponto de partida as referências diretas no âmbito de tradução preservadas na sua trilogia, as quais servirão também como critérios de classificação temática das práticas de tradução de Joaquim Gonçalves. Será inventariada e analisada uma seleção de práticas mais representativas que correspondem ou diferem destas estratégias diretamente refletidas pelo autor, tendo em conta as teorias e terminologias de tradução convencionalmente conhecidas.

Os objetivos didáticos do padre Joaquim Gonçalves englobam o ensino da tradução, para além da leitura e composição na língua chinesa, pois no prólogo da *Arte China*, ao apresentar os três volumes interdependentes do seu método de ensino, o próprio autor explica que:

> Sendo o meu intento dar ao Estudante da Lingua China todos os meios, para entrar no seu conhecimento, e pratica, tanto na falla, como na escrita: foi-me necessario fazer tres differentes volumes, que devem andar juntos, por fazer hum todo combinado, e necessario: combinado, para naõ engrossar os volumes: necessario: porque ficaõ as suas partes dependentes, e o estudante naõ obstera o seu fim sem a posse de

> todas ellas: assim he, que a Arte he ncessaria tanto por ensinar a ler, traduzir, e compor, como por dar ideas, que facilitaõ o uso, e intelligencia doas diccionarios; o Diccionario China-Portuguez he necessario ao Portuguez-China para a pronuncia, e uso das letras neste indicadas (Gonçalves, 1829, p. II).

Neste sentido, entende-se que a arte de traduzir não é ensinada de forma direta, separada e especializada no método de Gonçalves, tal como acontece nas gramáticas dos diplomatas que tinha como a sua missão a formação de futuros intérpretes para os postos consulares onde trabalhavam, como é no caso de Thomas Wade (1867). Ainda assim, é possível identificar algumas referências no domínio da tradução na sua trilogia. Nas quinze advertências que o autor providencia no início do seu *Diccionario Portuguez-China*, os seguintes três números fazem referência à tradução:

> 8 Sendo as vezes em climas taõ distantes as coisas naturaes mui differentes das nossas, muito mais o saõ as instituições humanas; assim quando traduzo, v. g. Rabeca 胡琴 naõ pense o estudante, que este instrumento China he inteiramente como o nosso.
>
> 9 Algumas coisas ha na Europa, que nem por semelhança ha na China, e as avessas; neste caso ponho algumas vezes o som da palavra Portugueza em letras Chinas, ou as avessas: outras vezes as omitto inteiramente: muitos destes sons estando ja corrompidos pelo uso, pensei melhor seguir o costume, e assim escrevi, v. g. Ginsaõ, se bem que em China Mandarim soa *Jen xen*: logo se conhece, que a palavra naõ he classica, quando tem a pronuncia tirada do China.
>
> 10 Considerando, que ha grande difficuldade em traduzir muitas frases, e que o sentido, e uso de huma palavra se naõ pode algumas vezes exprimir bem senaõ por huma sentença; demais disto, que seria util algumas vezes mostrar a doutrina singular do China sobre algum assumpto (sem a approvar), sobrecarreguei (ao parecer) o Diccionario de frases (Gonçalves, 1831, p. II).

Nesta primeira observação acima citada, o autor salienta a distância sociocultural entre as sociedades portuguesa e chinesa e alerta para o uso comparativo dos termos para alguns objetos culturalmente específicos. Ainda no que diz respeito a referências socioculturalmente particulares, o autor recorre também ao estrangeirismo transliterado, quando o método de comparação anteriormente adotado não se encontra viável, advertindo que a conversão de certas palavras, que já haviam amplamente circulado na língua de chegada, pode afastar-se da sua representação fonética da língua de origem. Por fim, quando é difícil traduzir uma palavra isolada, o autor opta por incluir esta palavra numa frase concreta e traduzir o conjunto contextualizado, o que ao mesmo tempo contribui para o ensino dos pensamentos chineses. Entre as doze advertências que se encontram no princípio do *Diccionario China-Portuguez*, destaca-se o número dez, no qual o autor adverte que a exatidão das traduções não é garantida:

> 10 Em huma linha se achaõ frequentemente duas frases separadas por hum zero, e defronte se achaõ correspondentemente traduzidas, e separadas por hum ponto: naõ asseguro a exaçaõ da traduçaõ das letras, e frases, mas advirto, que os mesmos letrados Chinas naõ concordaõ muitas vezes na intelligencia ate das explicações do Diccionario China (Gonçalves, 1833, p. II).

No *Diccionario China-Portuguez* destinado à consulta da pronúncia e do uso dos caracteres chineses registados no *Diccionario Portuguez-China*, o autor procura manter o estilo bilingue da trilogia ao oferecer equivalentes tanto para as entradas como para os exemplos de cada verbete, informando que as suas traduções podem não ser as mais exatas devido à existência de interpretações divergentes.

Tendo como ponto de partida as referências do autor no âmbito de tradução, as suas práticas de tradução serão estudadas em duas categorias: a domesticação e a estrangeirização. Para cada uma destas categorias serão analisadas, nesta tentativa de investigação, apenas uma pequena seleção de amostras mais intrínsecas às observações próprias de Joaquim Gonçalves. Em termos específicos, para a primeira serão revisitados somente

nomes de instrumentos musicais, como é no caso da tradução da rabeca "胡琴 [hú qin]", e para a segunda unicamente empréstimos correspondentes ao exemplo de "Ginsão".

4.1. **Domesticação**

O exemplo que o padre Joaquim Gonçalves utiliza para chamar a atenção à distância sociocultural entre as sociedades portuguesa e chinesa é um caso representativo de adaptação a nível lexical: a tradução da "rabeca" – um "instrumento musical em forma de viola, com quatro cordas, das quais se extrai o som através de um arco guarnecido de crinas previamente passadas por resina («rabeca (1)», s. d.)", em "胡琴 [hú qin][2]" – um "instrumento musical de corda, com um arco de bambu guarnecido de crina de cavalo que fricciona entre as duas cordas, incluindo *jinghu*, *erhu*, etc[3] («胡 [hú]», 2016)". Do ponto de vista etimológico, o termo rabeca tem origem árabe, inicialmente era utilizado como sinónimo de "rabel" ou "arrabil" que designavam o *rebab* dos árabes e, posteriormente, com o aperfeiçoamento das violas de arco, passou a ser o nome vulgar do violino (Vieira, 1899, pp. 437–438), enquanto o termo chinês *huqin***,** cuja tradução literal seria "instrumento de corda (*qin)* [4] dos bárbaros (*hu*)", data da Dinastia Tang (618-907 d. C) e teria sido utilizado para designar primeiramente instrumentos de cordas dedilhadas e depois instrumentos de arco (Stock, 1993, pp. 88–90). Portanto, considerando as respetivas características físicas, o posicionamento do arco e o desenvolvimento histórico, é possível afirmar que os dois substantivos utilizados como equivalentes nos dicionários de Gonçalves se referem na verdade a instrumentos musicais significativamente distintos, cada um com as suas características culturais próprias, tal como o autor adverte aos seus alunos, apesar das semelhanças compartilhadas: instrumentos de corda e arco e possíveis origens estrangeiras. Em termos de estratégia de tradução, não há dúvida de que se trate de uma prática de domesticação, pois as características técnicas e culturais do texto de partida estão minimamente presentes no texto de chegada. Em termos específicos, este exemplo de tradução enquadra-se no método de adaptação ("Adaptation" na Tabela 2), também conhecida como equivalente cultural, isto é, a substituição de um elemento cultural por um equivalente da cultura de chegada (Molina & Hurtado Albir, 2002, p. 509), o qual, na teoria de tradução moderna, costuma ser considerado como uma técnica controversa e por vezes problemática, que só deve ser utilizada quando a distância intercultural é tão grande que uma tradução precisa poderá dificultar a compreensão do texto traduzido (Fawcett, 2003, p. 39).

A nível lexical, esta mesma adaptação acontece frequentemente na tradução de uma grande parte dos nomes de instrumento musicais, conforme se exemplifica na Tabela 3.

[2] Entre parênteses retos são transcrições fonéticas dos caracteres chineses em *Pinyin*, adicionadas pela autora deste artigo de acordo com a pronúncia moderna do mandarim padrão.
[3] Tradução da autora. Texto original: "弦乐器，在竹弓上系马尾毛，放在两弦之间拉动。有京胡、二胡等。".
[4] Designação genérica de um conjunto de instrumentos musicais (Cf. «琴 [qín]», 2016).

Tabela 3. Nomes de instrumentos musicais traduzidos através de adaptação lexical

N.º	Entrada	Equivalente(s)	Gonçalves (1831)
(1)	Alaúde	琵琶 [pí pa]	p. 31
(2)	Clarinete	笙 [shēng]	p. 163
(3)	Flauta	太平簫[tàipíng xiāo]	p. 376
(4)	Gaita	簫[xiāo]。[a] 龠[yuè]	p. 397
(5)	Guitarra	月琴 [yuèqín]	p. 416
(6)	Trombeta	畫角[huà jiǎo]	p. 827

[a] O pequeno círculo, correspondente ao ponto final no sistema de pontuação chinês, que ocorre no meio da linha, é utilizado para separar sentidos diferentes (Gonçalves, 1831, p. III).

O termo alaúde, segundo Ernesto Vieira, refere-se a um instrumento de cordas dedilhadas, com o tampo harmónico em forma de pera, e o cravelhame achatado e inclinado para trás em ângulo com o braço, tendo sido trazido para a Europa pelos cavaleiros cruzados quando regressaram do Oriente no século XII. O seu nome é de origem árabe *èud* e teria passado por muitas mudanças na Europa, sendo as primeiras: *leut, leuth, luit, lut, luc, lucs, luc,* e *luz* (Vieira, 1899, pp. 41–43), enquanto que o termo "琵琶 [pí pa]", atualmente conhecido como *pipa*, mas também descrito como *Pear-shaped bowl lute*[5] ou *Chinese lute,*[6] teria sido importado na China há dois milénios, por via de comércios terrestres, e designava de forma genérica qualquer instrumento de cordas dedilhadas antes da sua denotação moderna (Myers, 1992, pp. 1, 6–7) de um "instrumento musical de cordas, fabricado de madeira, com quatro cordas, cuja parte inferior é uma caixa achatada, com o formato de semente de melão, e a parte de cima é um pescoço comprido com a extremidade superior curvada[7] («琵琶 [pí]», 2016)". Apesar de os dois nomes não corresponderem em todos os pormenores, entende-se esta comparação, baseada possivelmente nas características físicas dos dois instrumentos e na maneira de tocar, que o padre Joaquim Gonçalves faz no seu dicionário, a qual continua a ser feita nas traduções e descrições modernas.

No caso do clarinete, identificado como um dos mais importantes instrumentos de sopro composto de um tubo usualmente de madeira e, por vezes, de metal e equipado com boquilha na extremidade superior na qual se adapta a palheta (Vieira, 1899, p. 149), já se nota uma maior distância entre o mesmo e o termo chinês "笙 [shēng]", que designa um instrumento de sopro de palheta livre que é paralelo ao *qin* e ao *se* na Antiguidade (Marks, 1932, p. 600), sendo "constituído de um conjunto de tubos de bambu aprestados de palhetas e um tubo de sopro, todos colocados numa cabaça em forma de caçarola[8] («笙 [shēng]», 2016)", pois variam substancialmente mesmo só em termos do aspeto físico. Ainda assim, o autor faz uma ligação entre os dois termos porventura pela consideração da importância que os dois instrumentos assumem nas respetivas culturas musicais das sociedades nas quais estão inseridas. Tal como o clarinete, a flauta e a gaita também se referem a instrumentos de sopro: a gaita é definida como um "nome que o vulgo dá a qualquer instrumento rustico de sopro, como a flauta ou pífano (Vieira, 1899, p. 263)", ou seja, trata-se de uma designação genérica de instrumento rudimentar de sopro; a flauta, por sua vez, é um instrumento de sopro composto de um tubo cónico ou cilíndrico, com vários orifícios e chaves, e o aspeto que a mais afasta do clarinete é o seu mecanismo de produção do som: neste caso, "o tocador sopra diretamente contra a aresta de um orifício, denominado

[5] Ver "Bilingual Chart of Common Chinese Instrument Names" (Wang & Chen, 2018).
[6] Designação adotada pelo diplomata, músico e sinólogo holandês Robert Van Gulik (1910-1967) por razões poéticas (Cf. Myers, 1992, p. 7).
[7] Tradução da autora. Texto original: "弦乐器，用木料制成，有四根弦，下部为瓜子形的盘，上部为长柄，柄端弯曲。".
[8] Tradução da autora. Texto original: "把若干根装有簧的竹管和一根吹气管装在一个锅形的座子上制成。" .

embocadura, a qual divide a columna de ar, que em parte é obrigada a retroceder estabelecendo assim o movimento vibratorio (Vieira, 1899, p. 249)". Quanto à tradução dos dois termos, a palavra chinesa "簫[xiāo]" está presente em ambas as propostas, sendo elemento constituinte e núcleo da palavra composta "太平簫[tàipíng xiāo]", modificada pelo adjetivo predicativo "太平[tàipíng]", que significa "pacífico", no equivalente da flauta, e em forma isolada no termo correspondente para a gaita. Esta palavra monossilábica, assim como a sua alternativa equivalente "龠[yuè]", denomina no chinês moderno um conjunto de instrumentos de sopro, de forma similar à denotação da gaita, embora não seja diastraticamente restrita. Entretanto, o termo "太平簫[tàipíng xiāo]" já se restringe a um tipo de *xiao* muito mais específico e associado ao grupo étnico chinês *Miao*, cuja melodia mitologicamente teria acalmado a situação bélica entre dois tribos rivais (Cf. Liu, 2010), daí a sua conotação pacificadora. Através da comparação dos nomes traduzidos destes dois instrumentos de sopro, repara-se que Joaquim Gonçalves teria tido em consideração os respetivos âmbitos de aplicação dos termos enquanto adaptavam a sua tradução, procurando discriminar os termos genéricos dos específicos.

Nos exemplos (5) e (6), em relação à tradução da guitarra "月琴 [yuèqín]" e da trombeta "畫角[huà jiǎo]", as respetivas caracterizações musicológicas e origens etimológicas não serão aqui discutidas em pormenor, uma vez que não é foco deste trabalho o estudo da musicologia, mas sim as estratégias e práticas de tradução. Os termos chineses designam dois instrumentos musicais da tradição chinesa que diferem dos objetos denominados pelos termos portugueses, apesar da existência de traços comuns, seja em termos da aparência, seja em termos da forma de tocar, tal como acontece aos exemplos anteriormente analisados.

A adaptação a nível lexical com base em semelhança ou analogia é amplamente aplicada, mas não parece ser a única solução disponível, sobretudo quando um equivalente cultural é difícil de encontrar, como é o caso dos exemplos (7), (8) e (9) da Tabela 4.

Tabela 4: Nomes de instrumentos musicais traduzidos através de descrição

N.º	Entrada	Equivalente(s)	Gonçalves (1831)
(7)	Sanfona	手琴 [shǒu qín]	p. 742
(8)	Piano-forte	洋琴[yáng qín]	p. 630
(9)	Zabumba	大鼓 [dà gǔ]	p. 871

A sanfona é registada como um instrumento antigo, muito usado na idade média, com duas cordas, uma das quais "varia de entoações por meio de um pequeno teclado que o tocador dedilha com a mão esquerda em quanto com a direita dá movimento á manivella (Vieira, 1899, p. 450)". Devido possivelmente à ausência de um instrumento estruturalmente parecido na cultura chinesa, o autor optou por uma descrição baseada na participação ativa das mãos ou porventura pelo facto de o instrumento poder ser segurado nas mãos, neste caso, "手琴 [shǒu qín]", cujos dois caracteres constituintes significam respetivamente "mão" e "*qin*". A mesma técnica ocorre na tradução do piano-forte, "instrumento de cordas percutidas com teclado (Vieira, 1899, p. 416)", referência cultural tipicamente ocidental, cujo primeiro caracter "洋[yáng]" significa literalmente "oceano", o qual enquanto predicador de substantivo implica algo oriundo do estrangeiro por via marítima. Já a zabumba, "nome popular e grotesco dado ao bombo (Vieira, 1899, p. 549)", que é um "instrumento de percussão do genero dos tambores (Vieira, 1899, p. 101)" é descrito como um "*gu*[9] grande (*da*)". Nestas situações, passa a ser utilizada a técnica de descrição, ou "Description" da Tabela 2, que implica a substituição de um termo ou expressão por uma descrição da sua forma e/ou função (Molina & Hurtado Albir, 2002, p. 510),

[9] Designação genérica de um conjunto de instrumentos de percussão em forma cilíndrica ou redonda achatada, coberto de pele numa ou em ambas as superfícies (Cf. «鼓 [gǔ]», 2016).

igualmente enquadrada no âmbito da domesticação. Baseado nos três exemplos acima analisados, pode afirmar-se que as descrições de Joaquim Gonçalves tendem a ser concisas e genéricas.

A particularização ("Particularization" da Tabela 2) refere-se ao uso de um termo mais preciso e concreto (Molina & Hurtado Albir, 2002, p. 510) na tradução, a qual se encontra aplicada no exemplo (10). Segundo Vieira (1899, p. 394), o termo órgão designava genericamente qualquer instrumento, em relação à música, e com especialidade os instrumentos de vento. Para este nome genérico, encontram-se três equivalentes alternativos, denominando o primeiro um "*qin* de vento (*feng*)" e os outros dois instrumentos musicais específicos da música tradicional da China. A mesma técnica é aplicada a mais nomes de instrumentos musicais, conforme exemplificados na Tabela 5.

Tabela 5. Nomes de instrumentos musicais traduzidos através de particularização

N.º	**Entrada**	**Equivalente(s)**	**Gonçalves (1831)**
(10)	Órgão	風琴[fēngqín]。鎖吶[suǒnà]。笙[shēng]	p. 583
(11)	Pandeiro	緊皮鼓[jǐn pí gǔ]。太平鼓[tàipíng gǔ]。單面鼓[dān miàn gǔ]	p. 594
(12)	Trompa	號頭|筒[hào tóu|tǒng]。喇叭[lǎba]△ [a] 觱篥[bìlì]	p. 827
(13)	Viola	絃[xián]。琵琶 [pípa]	p. 860

[a] O triângulo indica que as frases seguintes são sublimes, usadas só na escrita (Gonçalves, 1831, p. III).

As mesmas técnicas acima abordadas ocorrem também quando o sentido de tradução é invertido, isto é, quando se traduz do chinês para o português, conforme exemplificadas na Tabela 6.

Tabela 6. Nomes de instrumentos musicais traduzidos com a adaptação lexical

N.º	Entrada	Equivalente(s)	Exemplo	Tradução do exemplo	Gonçalves (1833)
(14)	龠 [yuè]	Gaita	管龠	Canudo, e gaita.	p. 142
(15)	鼓 [gǔ]	Tambor	一彈再鼓志足樂也	O tocar já cravo já tambor he bastante para satisfazer.	p. 203
(16)	叭 [bā]	/	喇叭	Trombeta	p. 249
(17)	琶 [pá]	/	一曲琵琶帶恨歌	Huma cantiga tocada na viola, e cantada com ar de raiva.	p. 613
(18)	琵 [pí]	Viola	猶抱琵琶馬上彈	Ainda com a viola nos braços a cavallo toca.	p. 613
(19)	琴 [qín]	Piano China	琴瑟和諧	Uniaõ do cravo, e psalterio (consorte)	p. 614
(20)	瑟 [sè]	Psalterio	琴瑟	Cravo psalterio	p. 614
(21)	笙 [shēng]	Orgãosinho, gaita	笙簧寫心	Os tubos do orgaõ representaõ o coraçaõ (do tocador)	p. 682
(22)	笛 [dí]	Flauta	牧童之短笛載犢而歸	A flauta curta do pastor he posta no novilho para voltar (noite.)	p. 682
(23)	筒 [tǒng]	/	筒簫	C. gaita	p. 684
(24)	箏 [zhēng]	Cravo	瑤箏	Cravo de vidros suspensos, que o vento toca.	p. 685
(25)	簫 [xiāo]	Gaita de capador	吹簫引鳳	Tocar a gaita para chamar a aguia.	p. 690
(26)	鈸 [bó]	Timbales	鐃鈸	Timbales	p. 970
(27)	鐃 [náo]	Timbales	鼓以始之、鐃以收之、節奏於是乎得法	O tambor he para principiar a musica, os timables para acabar, o tocar com regra, e convenientemente esta talvez \nisto.	p. 985
(28)	鑼 [luó]	Batega	鼓鑼	Tambor, e batega	p. 990

Por vezes o autor não oferece nenhum equivalente para certas entradas, como é o caso dos exemplos (16), (17) e (23), possivelmente porque só é possível entender o sentido dos mesmos quando surgem acompanhados.

Por este motivo, a Tabela 6 inclui também um exemplo bilingue selecionado de cada entrada que permite mais facilmente compreender o seu significado. A técnica de descrição está presente no exemplo (19), quando o *qin* é traduzido como o piano chinês enquanto que a técnica de particularização volta a aparecer no exemplo (25), quando o termo genérico *xiao* é traduzido como gaita de capador, uma "flauta composta de pequenos canudos de diversas dimensões, muídos numa fileira sobre os quaes o tocador sopra produzindo em cada um differente som (Vieira, 1899, p. 263) ". O exemplo (20) apresenta uma tradução curiosa do termo "瑟 [sè]": o termo psaltério que designa um "instrumento de cordas metallicas, percutidas, composto de uma caixa approximadamente triangular, sobre cujo tampo se estendem as cordasa ... usado pelos povos orientaes da antiguidade, assyrios, chaldeus e egypcios (Vieira, 1899, p. 482) " parece ser tão específico que nem se encontra registado no *Diccionario Portuguez-China* de Joaquim Gonçalves.

Por razões de economia, as restantes práticas de tradução chinês-português não serão aqui analisadas em pormenor, e passar-se-á agora ao foco de análise: a tradução dos nomes de instrumentos quando estão inseridos em textos específicos, conforme exemplificados na Tabela 7, sublinhados em negrito pela autora deste estudo.

Tabela 7. Ocorrências dos nomes de instrumentos musicais em frases

N.º	Texto em português	Texto em chinês	Gonçalves (1829)
(29)	Queimar incensos, tocar **piano** com hum som claro, e finaes pausados, nao he senão sacar bons sons, mas não chega à maravilha da harmonia do **piano, e cravo** (matrimònio.)	焚香掃**琴**聲清韻逸此不過僅善其音而不若**琴瑟**諧之妙也	p. 115
(30)	Elles julgão, que os **órgãos, e violas, instrumentos de côrdas, e baffo**, e representar comédias desabafão o coração, e eu tomando isto por dissipação da fazenda, mais me instristêço verificando-se o provérbio: O theatro de prazer para outros he o lugar da minha pena; não tenho remêdio senão tapar o nariz, e lamentar; pôr as mãos na cabêça, e dormir.	彼則以為**管弦絲竹**演戲暢情我則以為浪費貨財更加納悶正所謂他人得以場是我傷心所惟有掩鼻苦吟蒙頭酣睡而已	pp. 119–120
(31)	Toca **cravo** a burros, (a quem naõ entende da materia.)	對驢撫**琴**	p. 308

O termo *qin* ocorre simultaneamente nos exemplos (29) e (31) mas com duas traduções distintas: piano e cravo. O exemplo (29) foi retirado do segundo capítulo da *Arte China*: trata-se de uma frase do estilo clássico, destinada ao ensino dos caracteres chineses de treze traços. A primeira ocorrência do *qin* nesta frase é isolada, sendo complemento direto do verbo tocar, e a segunda, combinada com um outro instrumento musical, também da Antiguidade, que é parecido com o *qin,* e frequentemente tocado em simultâneo com este, porque os sons são melodiosos, pelo que a palavra composta *qinse* também costuma ser utilizada no sentido metafórico para se referir a relações harmónicas (Cf. «琴 [qín]», 2016). Quanto à tradução desta metáfora, encontra-se aplicada a técnica de compensação ("Compensation" da Tabela 2) que implica a introdução de um elemento informativo ou um efeito estilístico do texto de partida no local diferente no texto de chegada, porque não é possível refletir o mesmo no mesmo lugar que no texto de partida (Molina & Hurtado Albir, 2002, p. 510); neste caso, o sentido alegórico da palavra *qinse*, "matrimónio", explicita-se entre parênteses como forma compensatória, pois o sintagma nominal "piano e cravo" não possui a mesma conotação em português. Esta mesma técnica ocorre também no exemplo (31), o qual foi retirado do sexto capítulo da mesma obra, dedicado ao ensino de provérbios chineses, juntamente com a tradução literal ("Literal translation" da Tabela 2), que também é designada como tradução palavra a palavra ("Word-for-word translation" da Tabela 1), uma técnica no enquadramento da

estratégia de estrangeirização. No que diz respeito ao exemplo (30), a técnica de tradução literal, a nível frásico, é ainda mais evidente: os substantivos coletivos "管弦 [guǎnxián]" – instrumentos de cano e cordas, e "絲竹 [sīzhú]" – termo geral de um conjunto de instrumentos como o *qin*, *se*, *xiao* e *di*, foram traduzidos palavra a palavra, resultando num sintagma composto de quatro palavras, correspondendo cada uma a um dos caracteres chineses, alvo do ensino do segundo capítulo da *Arte China*.

4.2. **Estrangeirização**

Para referências culturalmente próprias que nem possuem um equivalente semelhante ou análogo na cultura de chegada, o padre Joaquim Gonçalves opta pelo método de transliteração ("Transliteration" da Tabela 1), que se traduz na "representação dos caracteres ou símbolos semióticos de uma língua recorrendo aos caracteres ou símbolos semióticos que tenham uma pronúncia similar ou idêntica noutra língua[10]" (Xiong, 2014, p. 85), correspondendo ao "Borrowing" da Tabela 2, mais concretamente o "Naturalized borrowing" (Cf. Molina & Hurtado Albir, 2002, p. 510), ou seja, um empréstimo naturalizado consoante as regras ortográficas da língua de chegada. Serão analisadas nesta subsecção somente práticas de tradução recolhidas do *Diccionario Portuguez-China* de Joaquim Gonçalves, no âmbito da estratégia de estrangeirização, uma vez que as entradas do seu dicionário chinês-português são todas monossilábicas, pelo que é raro encontrar exemplos que se enquadrem nesta categoria temática. Vale realçar que, tal como o próprio autor alerta aos seus alunos, não são clássicas as palavras portuguesas cuja pronúncia é tirada do chinês, ou seja, ainda que surjam no seu dicionário português-chinês, são empréstimo do chinês, pelo que a tradução destas palavras foneticamente emprestadas só pode ocorrer do chinês para o português.

Começando com o exemplo de Gonçalves, a palavra ginsão, também escrito como "ginsam", é registada da seguinte forma no dicionário de Raphael Bluteau (1713, p. 75):

> GINSAM. He huma raiz da China, [q[11]] tira a vermelho, mas escuro & desmayado. Lança hum talosinho branco, & lenhoso. Vendese por preço de prata; os Grandes a usaõ, cozendo pequena porçaõ della em agoa, & abebem para refazer as forças. Pao da China, Assucar, *Ginsão*. Queiròs. Vida do Irmaõ Basto, Epistol. Dedicar.

O termo "ginsão" é um empréstimo do chinês "人參 [rénshēn]", adotado diretamente da língua chinesa e adaptada ao sistema ortográfico e fonético do português. O próprio Gonçalves adverte que a representação fonética desta palavra é afastada da sua pronúncia em mandarim (1831, p. II). Para além do termo que deu origem a esta palavra estrangeira no português, o autor regista ainda um conjunto de designações alternativas: "神草 [shén cǎo]" – erva (*cao*) divina *(shen*); "高丽参 [Gāolí shēn]" – ginsão (*shen*) coreano (*Gaoli*); "洋参 [yáng shēn]" – ginsão estrangeiro; "丹参 [dān shēn]" – ginsão vermelho (*dan*) e "沙参 [shā shēn]" – ginsão de areia (*sha*) (Cf. Gonçalves, 1831, p. 404), recorrendo maioritariamente às técnicas de descrição e particularização. O nome desta planta medicinal, que teria sido incorporado na língua portuguesa antes do século XVIII, possui mais que uma variante no português moderno, nomeadamente *ginseng*, jinsém e jinsão, tendo o primeiro sido oriundo do empréstimo puro do inglês *ginseng*, apresentando um maior nível de aproximação fonética da sua pronúncia mandarina.

Verifica-se o mesmo método de tradução nos exemplos empréstimos exemplificados na Tabela 8, com transcrições fonéticas entre parênteses retos da autora:

[10] Tradução da autora. Texto original: "把一种语言的文字符号用另一种语言中与它发音相同或相近的文字符号表示出来的方法。".
[11] Trata-se da letra "q" com til, abreviatura de "que".

Tabela 8. Nomes chineses culturalmente específicos traduzidos através de empréstimo

N.º	Entrada	Equivalente(s)	Gonçalves (1831)
(32)	Lichia	荔枝 [lìzhī]	p. 485
(33)	Longan	龍眼 [lóngyǎn]	p. 491
(34)	Ocá (remédio)	阿膠 [ējiāo]	p. 571
(35)	Vom-fá, peixe	黃花鱼 [huánghuā yú]	p. 868

Os primeiros dois termos botânicos, oriundos da língua e cultura chinesas, revelam a mesma técnica de tradução utilizada na palavra "ginsão", isto é, o empréstimo; porém, não parecem ter sido naturalizados da mesma forma, ou seja, mantêm uma realização fonética semelhante à do chinês mandarim, apesar da perda do valor tónico e do acréscimo de um "a" final no caso de "lichia" (Cf. Chen, 2020; Guo, 2020). Quanto aos exemplos (34) e (35), não há muita dúvida quanto à origem chinesa dos dois empréstimos em português; contudo, as pronúncias não parecem ter sido emprestadas do chinês mandarim. Talvez por se tratar de dois nomes menos comuns para o público-alvo português, encontra-se aplicada mais uma técnica para além do empréstimo: a técnica de amplificação ("Amplification" da Tabela 2 e "Addition" da Tabela 1), que se traduz na introdução de pormenores não formulados no texto de partida de natureza informativa e/ou explicativa (Molina & Hurtado Albir, 2002, p. 510). No exemplo (34), através da anotação entre parêntese, um leitor entenderá o género categórico deste substantivo, mesmo que não tenha conhecimento do produto em si. O exemplo (35) apresenta certa variação quanto à forma de inserção da informação adicional, a qual, dependendo do ponto de vista, também pode ser entendida como uma combinação das técnicas de empréstimo e tradução literal, pois o constituinte semântico "鱼 [yú]", correspondente ao "peixe", está de facto presente na designação chinesa.

A técnica de empréstimo ocorre também na tradução das próximas duas palavras portuguesas, conforme exemplificada na Tabela 9.

Tabela 9. Nomes portugueses culturalmente específicos traduzidos através de empréstimo

N.º	Entrada	Equivalente(s)	Gonçalves (1831)
(36)	Bálsamo	巴兒撒末 [bā er sā mò]。水安息 [shuǐ ānxī]	p. 91
(37)	Café	戛啡 [jiá fēi]	p. 119

O termo português "bálsamo", é derivado da palavra persiana *Bassam*, ou da arábica *Belsan*, e designa um licor, diferentemente dos dois potenciais étimos que podem indicar qualquer óleo aromático ou goma odorífera, e os bálsamos naturais conhecidos na cultura portuguesa são de origens variadas: o bálsamo puro deitado em leite ou em água, o bálsamo de Peru, o bálsamo Tolutano ou bálsamo de Honduras, o bálsamo novo originário da Ilha de S. Domingos, e ainda o bálsamo do Brasil (Cf. Bluteau, 1712, pp. 25–26). Quanto ao primeiro equivalente oferecido por Joaquim Gonçalves, esta transliteração é constituída por três sílabas principais, designadamente, a primeira, terceira e quarta sílaba, que correspondem respetivamente às três sílabas portuguesas, e uma adição *er*, que faz parte da sílaba anterior[12] e ao mesmo tempo representa *l* final da primeira

[12] Lê-se a seguinte indicação no *Valor das Letras Europeas na Pronuncia do China*: "As addições *toū*, *olr* saõ so para encher, e nada significaõ: ellas se pronunciaõ breves, e o *olr* fica fazendo huma syllaba com a antecedente: assim *t'ou olr*, chefe, soara *t'oulr*. (Gonçalves, 1829, p. VIII)".

sílaba portuguesa, apresentando um alto grau de fidelidade à pronúncia portuguesa. O segundo equivalente chinês parece estar relacionado com uma variação de "安息香 [ānxī xiāng]", que se traduz em "benzoinum", um ingrediente vastamente utilizado nas fórmulas da medicina chinesa, cujo nome chinês é derivado do nome de Ársaces I, fundador do Império Parta ou Arsácida da Pérsia Antiga (Sun et al., 2023, p. 16), acrescido do caracter final "香 [xiāng]", que significa cheiro agradável ou incenso. O primeiro caracter da designação "水安息 [shuǐ ānxī]" significa literalmente "água", pelo que o termo pode indicar uma espécie líquida de *benzoinum* ou um tipo de *benzoinum*, importado por via marítima. Sun et al. (2023, p. 20) defende que o produto por este nome designado na realidade difere substancialmente do *benzoinum*, apesar de o mesmo ter sido traduzido como "liquid benzoinum" em *Notes on Chinese Materia Medica* de Daniel Hanbury, onde este autor documenta a origem obscura deste produto:

> This drug is a dark-brown, semi-fluid resin, having an extremely fragrant odour of storax. It is met with in small glabular wooden shells, apparently the pericarp of some fruit, about $1\frac{3}{4}$ inches in diameter, closed with wax. Its origin is very obscure. The Chinese assert that they import it from the Straits, or, in other words, by way of the Indian Archipelago; but I have not been able to trace it either there or in Siam. It is curious, moreover, that this fragrant resin, even to the shell enclosing it, is extremely like that kind of balsam of Peru which was brought to Europe long ago in the capsules of a *Lecythis*, and naturally supposed to be a product of South America. [mudança de parágrafo omitida] The *Liquid Benzoin* is very expensive, a single shell, holding perhaps half an ounce, neing worth four dollars, or 20*s* (1862, p. 39).

Esta observação torna claro que o produto designado como "水安息 [shuǐ ānxī]" é um ingrediente usado na medicina chinesa, independentemente da sua relação com o *benzoinum*, e que este produto poderia ter sido oriundo da América do Sul, embora haja discrepância entre esta hipótese de Hanbury e as informações por si recolhidas dos povos chineses locais. Portanto, pode considerar-se que se trata de uma prática de tradução com recurso à técnica de adaptação, caso o ponto de partida de Joaquim Gonçalves tenha sido as semelhanças entre os dois termos, seja em termos da potencial origem, seja no que diz respeito às proximidades físicas. De qualquer forma, tendo em conta a presença simultânea de uma tradução através de empréstimo com adaptação a nível fonético, é talvez mais lógico presumir que não havia uma correspondência exata entre os dois termos.

O exemplo (37) é um típico exemplo de empréstimo com adaptação a nível fonético. De acordo com o dicionário etimológico de Antenor Nascentes (1955, p. 87), o termo café pode ter chegado à língua portuguesa ou a partir da palavra árabe *kahwa*, que geralmente designa qualquer bebida, mas ordinariamente o café (Bluteau, 1712, p. 35), ou do nome geográfico *Kaffa*, primeiro habitat da planta. Por outro lado, normalmente considera-se que o café é uma mercadoria estrangeira importada na China e o termo correspondente no chinês moderno "咖啡 [kāfēi]" tem passado por 28 formas de tradução antes da fixação da sua ortografia atual, uma das quais é o equivalente utilizado no dicionário português-chinês de Joaquim Gonçalves "戛啡 [jiá fēi]", de origem cantonense (Zhu & Zhao, 2019, pp. 234–235). Entre as 28 traduções deste produto analisadas em Zhu e Zhao (2019), muitas partilham esta mesma pronúncia, mas a nível ortográfico Gonçalves parece ter sido o único que adotou o primeiro caracter desta palavra chinesa, o que implica que não havia consenso quanto à representação ortográfica deste produto de origem estrangeira. Portanto, pode afirmar-se que se recorreu a técnica de empréstimo naturalizado na prática desta tradução.

Em síntese, as duas estratégias de tradução, explicitamente documentadas nas obras de Joaquim Gonçalves, foram amplamente cumpridas nas suas práticas de tradução, sobretudo quando estão em causa as referências culturalmente próprias e distintas, através de um conjunto de técnicas que coincidem com as propostas em teorias de tradução moderna. Os exemplos de tradução intrinsecamente relacionados com os dois casos exemplares contidos nas observações do próprio autor, alvo da análise desta secção apresentam frequentemente mistura de técnicas e até cruzamento das duas estratégias. Os termos traduzidos nesta abordagem, apesar do desuso de vários no contexto linguístico moderno, continuam a ser importantes para o estudo das referências sociolinguísticas, histórias e contrastivas.

5. Considerações finais

Nesta pesquisa, foram estudadas as estratégias e práticas de tradução do sinólogo e missionário português Joaquim Gonçalves através da recolha e análise das suas observações e exercícios de tradução que se encontram preservados nos dados bilingues da sua trilogia para o ensino-aprendizagem das línguas e culturas chinesas. A tradução assume um papel fundamental nas obras metalinguísticas dos missionários, sendo não só uma ferramenta didática tradicional, mas também um dos principais objetivos de ensinos dos missionários. Tal como a maior parte das obras missionárias, no tríptico de Gonçalves, repleto de exemplos contrastivos chinês-português, raramente se encontram palavras dedicadas à explicitação dos princípios orientadores das suas atividades de tradução, razão pela qual o estudo dos equivalentes bilingues concretos se torna imprescindível para a perceção da sua teoria de tradução. Nesta atividade intercultural, a transferência de informações ocorre não só entre os dois idiomas em causa, mas também, frequentemente, entre as respetivas culturas, permitindo estabelecer contacto entre culturas que estariam tão distantes quanto o comprimento do diâmetro da Terra. Além disso, a traduzibilidade dos elementos culturalmente específicos, quando estes ultrapassam os limites linguísticos, traz sempre desafio aos tradutores, pelo que as práticas de tradução dos missionários, enquanto as primeiras tentativas de criar uma ligação direta entre línguas e culturas distintas, merecem maior valorização tanto nos estudos modernos de tradução como na história das teorias de tradução.

Nas obras de Joaquim Gonçalves, foram observadas duas estratégias de tradução, isto é, a domesticação e a estrangeirização, que sobressaem especialmente quando as referências culturais próprias ou distintas de cada sociedade estão em causa, oferecendo uma orientação geral sem delimitar as escolhas de tradução. O método de Gonçalves foi concebido não só para servir ao ensino do chinês para europeus, mas também para chineses que desejem aprender o português. Esta utilidade bidirecional obriga o autor a priorizar a naturalização dos textos redigidos em ambas as línguas, pelo que não se admira que a domesticação tenha predominado nas suas práticas de tradução, recorrendo a uma grande variedade de técnicas de tradução como a adaptação, seja a nível lexical seja a nível fonético, a compensação, a descrição e a particularização, com o objetivo de garantir a fluidez da tradução na língua de chegada. Por outro lado, quando não é possível estabelecer uma equivalência translinguística baseada em semelhanças interculturais, isto é, quando se trata de referências culturalmente particulares, o autor abraça também a estratégia de estrangeirização através das técnicas de amplificação, empréstimo naturalizado e tradução literal, procurando maximizar sempre que possível a acessibilidade dos conceitos introduzidos na língua de chegada. Por vezes, a finalidade da tradução também parece ter influenciado a escolha entre uma estratégia e outra, sobretudo em termos do uso da técnica de tradução palavra a palavra, no âmbito da estrangeirização. Em termos específicos, quando o objetivo é o ensino de cada um dos principais elementos constituintes da frase, torna-se necessário oferecer uma tradução para cada palavra de modo separado. O estudo da aplicação das duas estratégias de tradução nos textos bilingues compilados nas obras de Gonçalves, seja de forma isolada seja de forma combinada, permite ainda indagar sistematicamente as informações histórias e sociolinguísticas inerentes nos termos traduzidos.

Em resumo, as práticas de tradução do padre Joaquim Gonçalves, seja no sentido endógeno, seja no sentido exógeno, permitem estabelecer uma visão única sobre as técnicas de tradução entre o português e o chinês, já que em muitos casos foram cuidadosamente concebidas para serem, simultaneamente, tão naturais quanto possível, e serem os primeiros ensaios de tradução. Este trabalho foi apenas uma tentativa de inspecionar este tesouro gramatical e lexicográfico, muito limitado às poucas reflexões de tradução de Gonçalves disponíveis nas suas obras, pelo que será de muito interesse e relevância abordar as demais práticas em futuros estudos temáticos.

Financiamento

Este trabalho foi financiado no âmbito da bolsa de investigação para doutoramento com a referência 2021.05393.BD da Fundação para a Ciências e a Tecnologia (FCT), com verbas do Orçamento de Estado e com verbas do Fundo Social

Europeu, a disponibilizar ao abrigo do PORTUGAL2020, através do Programa Operacional Regional do Norte (NORTE 2020).

Referências

Barros, Anabela (2014) Referências interculturais oitocentistas nas obras metalinguísticas em Português e Chinês do Pe Joaquim Gonçalves. *Diacrítica* 28 (1), pp. 103–139.

Bluteau, Rafael (1712) *Vocabulario portuguez e latino, aulico, anatomico, architectonico, bellico, botanico, brasilico, comico, critico, chimico.../pelo Padre D. Raphael Bluteau* (Vol. 2). Collegio das Artes da Companhoa de Jesu.

Bluteau, Rafael (1713) *Vocabulario portuguez e latino, aulico, anatomico, architectonico, bellico, botanico, brasilico, comico, critico, chimico.../pelo Padre D. Raphael Bluteau* (Vol. 4). Collegio das Artes da Companhoa de Jesu.

Bruscato, Amanda Maraschin & Baptista, Jorge (2020) Relações entre teorias linguísticas, teorias da aprendizagem e métodos de ensino de línguas. *Revista Educação, Cultura e Sociedade*, 10 (2), pp. 324–338. https://doi.org/10.30681/ecs.v10i2.3939

Chen, Chao (2020) *Empréstimos do cantonês no patuá de Macau*. Dissertação de mestrado, Universidade do Minho.

Fawcett, Peter (2003) *Translation and language: Linguistic theories explained*. St. Jerome Pub.

Gianninoto, Mariarosaria (2014) Translation in Chinese Grammars: Bilingual works by Western missionaries, diplomats and academics in the 18th and 19th centuries. In Otto Zwartjes, Klaus Zimmermann & Martina Schrader-Kniffki (eds.), *Studies in the history of the language sciences* (Vol. 122). John Benjamins Publishing Company, pp. 231–250. https://doi.org/10.1075/sihols.122.08gia

Gonçalves, Joaquim Afonso (1829) *Arte China, constante de alphabeto e grammatica comprehendendo modelos das differentes composiçoens*. Real collegio de S. Jose.

Gonçalves, Joaquim Afonso (1831) *Diccionario Portuguez-China no estilo vulgar mandarim e classico geral*. Real collegio de S. Jose.

Gonçalves, Joaquim Afonso (1833) *Diccionario China-Portuguez*. Real collegio de S. Jose.

Guo, Binyu (2020) *Estudo linguístico de palavras de origem chinesa no português.* Dissertação de mestrado, Universidade de Aveiro.

Hanbury, Daniel (1862) *Notes on Chinese Materia Medica*. The Pharmaceutical Journal and Transactions.

Levi, Joseph Abraham (2007) Padre Joaquim Afonso Gonçalves (1781-1834) and the Arte China (1829): An innovative linguistic approach to teaching Chinese grammar. In Otto Zwartjes, Gregory James & Emilio Ridruejo Alonso (eds.), *Missionary linguistics III/Lingüística misionera III*. John Benjamins, pp. 211–231.

Liu, Guohai 刘国海 (2010) 苗族太平箫 [Taiping Xiao da etnia Miao]. *今日民族 Ethnic Today 7*, 34.

Marks, Robert W. (1932) The music and musical instruments of ancient china. *The Musical Quarterly 18* (4), pp. 593–607.

Molina, Lucía & Amparo Hurtado Albir (2002) Translation techniques revisited: A dynamic and functionalist approach. *Meta 47* (4), pp. 498–512. https://doi.org/10.7202/008033ar

Morrison, Robert (1815) *A grammar of the Chinese language*. Mission Press.

Myers, John E. (1992) *The way of the Pipa structure and imagery in Chinese lute music*. The Kent State University Press.

Nascentes, Antenor (1955) *Dicionário etimológico da língua portuguesa*.

Rabeca (1). (s. d.). In *Dicionário da língua portuguesa*. Academia das Ciências de Lisboa. Disponível em https://dicionario.acad-ciencias.pt/pesquisa/?word=rabeca [consultado em 28 de junho de 2023].

Stock, Jonathan (1993) A historical account of the chinese two-stringed fiddle erhu. *The Galpin Society Journal* 46, pp. 83–113. https://doi.org/10.2307/842349

Sun, Yuanyuan 孙园园, Jian Feng 冯剑, Mingsong Wu 吴明松, & Yangyang Liu 刘洋洋. (2023) 经典名方中安息香的本草考 [Herbal Textual Research on Benzoinum in Famous Classical Formulas]. *中国实验方剂学杂志 [Chinese Journal of Experimental Traditional Medical Formulae]* 29 (6), pp. 14–24.
Venuti, Lawrence (2008) *The translator's invisibility: A history of translation* (2.ª ed.). Routledge.
Vieira, Ernesto (1899) *Diccionario musical ornado com gravuras e exemplos de música* (2.ª ed.). Lambertini.
Wade, Thomas Francis (1867) *Yü-yen Tzŭ-erh Chi, a progressive course designed to assist the student of colloquial Chinese, as spoken in the capital and the Metropolitan Department: In eight parts, with key, syllabary, and writing exercises*. Trübner.
Wang, Chenwei 王辰威 & Mingzhi Chen 陈明志 (eds.) (2018) *常用中国乐器中英文对照表 [Bilingual chart of common Chinese instrument names]*. 中国乐器缩略名称编撰委员会 [Committee for the Abbreviation of Chinese Musical Instrument Names].
Xiong, Bing 熊兵 (2014) 翻译研究中的概念混淆——以"翻译策略"、"翻译方法"和"翻译技巧"为例 [Confusão de conceitos nos estudos de tradução – das estratégias, métodos e técnicas]. *中国翻译 Chinese Translators Journal* 3, pp. 82–88.
Zhu, Xue 祝雪 & Xu Zhao 赵旭 (2019) 试论外来词coffee的汉译特征 [On Chinese translation features of Loanword Coffee]. *沈阳大学学报（社会科学版 [Journal of Shenyang University (Social Science)]* 21, pp. 234–239.
Zwartjes, Otto (2014) The Missionaries' contribution to translation studies in the Spanish Colonial period: The *mise en page* of translated texts and its functions in foreign language teaching. In Otto Zwartjes, Klaus Zimmermann & Martina Schrader-Kniffki (eds.), *Studies in the history of the language sciences* (Vol. 122). John Benjamins Publishing Company, pp. 1–50. https://doi.org/10.1075/sihols.122.01zwa
Zwartjes, Otto (2016) Colonial missionaries' translation concepts and practices: Semantics and grammar. In Sabine Dedenbach Salazar-Sáenz (ed.), *La transmisión de conceptos cristianos a las lenguas amerindias: Estudios sobre textos y contextos de la época colonial* (Vol. 48). Academia Verlag, pp. 43–76.
琴 [qín] (2016) In *Xiàndài hànyǔ cídiǎn 现代汉语词典 (Dicionário do Chinês Moderno)* (7.ª ed.). The Commercial Press, p. 1059.
琵 [pí] (2016) In *Xiàndài hànyǔ cídiǎn 现代汉语词典 (Dicionário do Chinês Moderno)* (7.ª ed.). The Commercial Press, p. 994.
笙 [shēng] (2016) In *Xiàndài hànyǔ cídiǎn 现代汉语词典 (Dicionário do Chinês Moderno)* (7.ª ed.). The Commercial Press, p. 1173.
胡 [hú] (2016) In *Xiàndài hànyǔ cídiǎn 现代汉语词典 (Dicionário do Chinês Moderno)* (7.ª ed.). The Commercial Press, p. 550.
鼓 [gǔ] (2016) In *Xiàndài hànyǔ cídiǎn 现代汉语词典 (Dicionário do Chinês Moderno)* (7.ª ed.). The Commercial Press, p. 469.

Clíticos reflexos e inerentes na variedade de português de São Tomé e Príncipe

Raquel Teixeira Madureira[1]

[1]Universidade de Lisboa, CLUL, Lisboa, Portugal

Resumo

Este artigo foca-se em construções verbais que selecionam clíticos reflexos e inerentes em português de São Tomé e Príncipe. Alguns estudos têm apontado uma tendência para a omissão dos clíticos do paradigma reflexivo nas variedades africanas do português. Este estudo incidirá sobre a variedade falada em São Tomé e Príncipe, onde o português se encontra em fase de nativização. Terá uma primeira parte dedicada à revisão de literatura sobre estas estruturas em português europeu. Numa segunda parte, serão analisados dados orais de um corpus, com base nas variáveis *escolaridade* e *idade*. A análise quantitativa de ocorrências permitirá constatar qual é a relevância da estratégia de omissão na língua falada. Os dados recolhidos parecem, de facto, indicar que esta variedade do português apresenta uma forte tendência para a supressão do clítico. Os clíticos inerentes sofrem maior omissão do que os reflexos, embora estes também sejam suprimidos. Além disso, a variável *escolaridade* parece ter algum peso na estratégia privilegiada pelos falantes, ao contrário da variável *idade*. O estudo procurará ainda discutir se o contacto com o crioulo forro tem um papel relevante na estratégia de omissão ou se, pelo contrário, esta se deve a uma mudança linguística resultante da aquisição (histórica) do português como L2.

Palavras-chave: português de São Tomé e Príncipe, clíticos reflexos e inerentes, variação e mudança linguística, aquisição L1/L2, contacto de línguas.

Abstract

This article focuses on verb constructions that select reflexive and inherent clitics in São Tomean Portuguese. Some studies have pointed out a tendency towards the omission of clitics from the reflexive paradigm in African varieties of Portuguese. The present study will deal with the variety of Portuguese spoken in São Tomé and Príncipe, that is undergoing a process of nativization. The first part offers a brief literature review on reflexive and inherent clitics in European Portuguese. The second part analyses oral data from a corpus, focusing on the variables *education* and *age*. The occurrences will be analyzed from a quantitative perspective, in order to verify the relevance of the omission strategy in spoken language in São Tomé. The data seems, in fact, to indicate that this variety of Portuguese presents a strong tendency towards the suppression of clitics. Inherent clitics suffer greater omission than reflexive clitics, which, nevertheless, are also suppressed. Furthermore, the variable *education* seems to have some weight in the choice of strategy by speakers, unlike the variable *age*. The article will also discuss if the omission strategy present in São Tomean Portuguese is a product of contact with Santome Creole or the result of a process of (historical) acquisition of Portuguese as L2.

Keywords: São Tomean Portuguese, reflexive and inherent clitics, linguistic change and variation, L1/L2 acquisition, language contact.

1. Introdução

As propriedades e a colocação dos pronomes clíticos em português europeu têm sido alvo de estudo, estando amplamente descritas em Duarte (1983), Duarte e Matos (2000), Brito et al. (2003) e Martins (2013), entre outros. A investigação sobre as variedades africanas do português, por outro lado, tem incidido sobre

outras questões sintáticas (e.g., duplos objetos, regência verbal), não havendo, por isso, estudos aprofundados sobre o tópico em análise, sobretudo para a variedade de São Tomé e Príncipe. Além disso, o estudo dessas variedades tem-se desenvolvido principalmente nas últimas décadas, o que reflete a recente difusão do português nas ex-colónias. Vários estudos têm procurado descrever as características dessas variedades. No entanto, não tem existido um investimento uniforme nas variedades do português analisadas (Hagemeijer, 2016).

Ainda assim, já foram elaborados alguns estudos centrados nos clíticos com ênfase nas variedades africanas do português, principalmente para o português de Moçambique (Gonçalves, 1990; Matsinhe, 1998; Mapasse, 2005). Alguns têm apontado uma tendência geral de supressão de pronomes reflexos em todas as variedades africanas do português (Hagemeijer, 2016; Mendes & Estrela, 2008). No entanto, a omissão dos clíticos do paradigma reflexivo carece ainda de uma análise aprofundada e de uma descrição detalhada. Este artigo procurará contribuir para a expansão do conhecimento da variedade de São Tomé e Príncipe. A análise quantitativa de ocorrências em dados orais terá como alvo constatar qual é a relevância da estratégia de omissão na oralidade. Essa análise será feita com o objetivo de dar resposta às seguintes questões: (i) Os clíticos inerentes, considerados não argumentais, sofrem maior omissão do que os reflexos, considerados argumentais? (ii) Será que informantes com um nível de escolaridade mais elevado apresentam uma maior percentagem de realização do clítico? (iii) Qual é a relevância da variável *idade* na estratégia da omissão do clítico? (iv) Qual é o papel do contacto com o forro (crioulo dominante em São Tomé e Príncipe) na estratégia de omissão do clítico em português de São Tomé e Príncipe?

Assim, este artigo fará: (i) uma análise da variação interfalante e intrafalante nas estratégias de realização e omissão do clítico, com base nas variáveis *idade* e *escolaridade*; e (ii) uma avaliação da relevância do contacto com o crioulo forro. Com estes objetivos em mente, entre a secção 2 e 4, faz-se uma revisão de literatura sobre estas estruturas que selecionam clíticos reflexos e inerentes em português europeu. Na secção 5, abordam-se as conclusões a que chegaram alguns estudos sobre a supressão do clítico nas variedades africanas do português. Nas secções 6 e 7, descreve-se o contexto linguístico de São Tomé e Príncipe, com especial ênfase no forro. Finalmente na secção 8, serão analisados dados orais de um corpus, com base nas variáveis *escolaridade* e *idade*.

2. **Aspetos gerais dos clíticos especiais**

Os pronomes clíticos são unidades lexicais átonas e, como tal, são dependentes de um hospedeiro, isto é, de uma unidade lexical com acentuação própria. Designados como clíticos especiais (Zwicky, 1977), os pronomes clíticos têm algumas propriedades em comum. Em primeiro lugar, eles ocorrem em adjacência estrita ao verbo, mesmo quando essa não é a posição canónica da sua função sintática, como é visível em (1). Além disso, apenas cliticizam em verbos, e não em qualquer outra classe de palavras, como mostra (2). Isso acontece, mesmo quando o clítico está estruturalmente relacionado com um elemento não verbal, como no caso do dativo de posse. Finalmente, eles não têm uma posição fixa em relação ao verbo que lhes serve como hospedeiro, podendo ocorrer em ênclise (1a), próclise (3a) ou mesóclise (3b).

(1a) Ele enviou-*lhes* os livros.
(1b) Ele enviou os livros [aos pais].

(2a) O Tiago viu a ferida do gato.
(2b) O Tiago viu-lhe a ferida.
(2c) * O Tiago viu a ferida-*lhe*.

(3a) Ele não *lhes* enviou os livros.
(3b) Ele enviar-*lhes*-á os livros.

Estas características distinguem os clíticos especiais da classe dos simples, a que pertencem os artigos e as preposições.

Os clíticos especiais apresentam também propriedades fonológicas próprias, como descrito em Vigário (1999). Quando o verbo termina em nasal, o pronome clítico acusativo *o(s)/a(s)* surge como *no(s)/na(s)*, conforme é visível em (4c). O clítico pode ainda surgir como *lo(s)/la(s),* quando o verbo ou o clítico que o precede termina com as consoantes *s* ou *r*, como mostra (5c). Nesse caso, a consoante final do verbo ou do clítico (*nos* ou *vos*) é apagada. A consoante final é igualmente apagada quando o verbo se encontra conjugado na primeira pessoa do plural e é seguido pelos clíticos dativos *nos* ou *vos* (conforme exemplificado em (6c)). No entanto, estes fenómenos fonológicos não se verificam no caso dos clíticos simples, nem mesmo com o artigo definido *o(s)/a(s)*, homónimo do clítico acusativo. Isso mesmo é visível nas alíneas (b) dos exemplos (4) a (6). Note-se que, em (4b), *no filme* é interpretado como um sintagma preposicional introduzido pela preposição *em*, não como um sintagma nominal.

(4a) Eles viram [*o* filme].
(4b) * Eles viram no filme.
(4c) Eles viram-*no*.

(5a) Eles costumam ver [*os* filmes da Marvel].
(5b) * Eles costumam vê los filmes da Marvel.
(5c) Eles costumam vê-*los*.

(6a) Nós encontramos os melhores filmes nos cinemas.
(6b) * Nós encontramo nos melhores filmes nos cinemas.
(6c) Nós encontramo-*nos* todos os dias.

Ainda assim, os clíticos especiais não exibem um comportamento uniforme, podendo distinguir-se diferentes tipos. Brito et al. (2003) apresentam uma divisão desta classe em cinco tipos de clíticos: o clítico com conteúdo gramatical, o clítico argumental proposicional ou predicativo, o clítico quase-argumental, o clítico com comportamento de afixo derivacional e o clítico sem conteúdo semântico ou morfossintático.

Esta divisão é feita com base nos seguintes critérios apontados pelas autoras: (i) o potencial referencial ou predicativo; (ii) a referência específica ou arbitrária; (iii) a possibilidade de lhes ser atribuído um papel temático; (iv) a possibilidade de surgirem em construções de redobro do clítico e de extração simultânea de clítico; e (v) a capacidade de modificarem a estrutura argumental de um verbo. Os clíticos reflexos e inerentes encontram-se precisamente em pontos opostos deste espetro.[1] Os clíticos reflexos são incluídos nos clíticos com conteúdo argumental. Quanto aos clíticos inerentes, considera-se que não têm conteúdo semântico ou morfossintático.

3. **Os clíticos reflexos**

O clítico reflexo é um pronome anafórico, retomando um sintagma nominal que ocorre na mesma oração, considerado o antecedente, e que tem a função de Sujeito. Estabelece-se um processo de correferência entre o clítico e o seu antecedente, já que ambos se referem à mesma entidade extralinguística. Assim sendo, seguindo os critérios mencionados em Brito et al. (2003), o clítico reflexo (i) tem potencial referencial e (ii) tem uma referência específica ou definida.

[1] Brito et al. (2003) afirmam que esta classificação reflete fases diferentes do processo de gramaticalização, visível do ponto de vista formal e do ponto de vista semântico. As autoras (Brito et al., 2003, p. 845) constatam, a respeito deste último ponto: "Os clíticos em português repartem-se por diferentes subtipos que vão desde aqueles que exibem um conteúdo substantivo pleno (argumental ou predicativo), aos clíticos que apresentam (parcial ou exclusivamente) propriedades de afixos, ou até aos clíticos que estão desprovidos de qualquer conteúdo atribuível, ocorrendo como vestígios fossilizados de estádios anteriores da língua."

3.1. **Anáfora ligada**

O clítico reflexo não tem qualquer valor referencial por si só, dependendo totalmente do seu antecedente para que o seu valor seja definido. Trata-se de uma anáfora ligada. Visando esta relação, Chomsky (1981, p. 188) apresenta o seguinte princípio: "An anaphor is bound in its governing category." Brito et al. (2003, p. 811) enumeram três propriedades das anáforas ligadas.

Em primeiro lugar, o antecedente e a anáfora devem pertencer à mesma oração, o que não acontece em (7a).[2] De facto, o clítico reflexo tem de estar 'ligado' dentro de um domínio local a que também pertence o antecedente. Segundo Otero (1999, p. 1446), o domínio de uma expressão anafórica reflexiva é a unidade mínima que contém a anáfora e um possível 'ligador'. Brito et al. (2003, p. 812) identificam esse domínio mínimo como a frase finita simples.

Em segundo lugar, a anáfora não pode ocupar a posição de sujeito, como demonstrado em (8). Finalmente, nenhuma expressão nominal com função de sujeito pode interferir entre o antecedente e a anáfora. Isso é exemplificado em (7a), onde o sintagma nominal *a mãe* está posicionado entre o antecedente e a anáfora. Lobo (2013, p. 2210) acrescenta ainda que o antecedente deverá ser mais proeminente do que o clítico, não podendo estar demasiado encaixado na estrutura da oração. O clítico reflexo tem de estar no escopo do elemento 'ligador', isto é, o seu antecedente. Por outras palavras, o antecedente tem de c-comandar a expressão anafórica. Em (9), o sintagma nominal *o menino* não pode ser o antecedente do clítico, porque está contido no SN sujeito e, como tal, encontra-se demasiado encaixado e não c-comanda o clítico.

(7a) * O menino disse [que a mãe *se* penteou *a si próprio*].
(7b) A mãe disse [que o menino *se* penteou *a si próprio*].

(8) * *Se a si mesma* penteou.

(9) * [A mãe do menino] penteou-*se a si próprio*.

Nos exemplos de (7) a (9), as expressões *a si próprio/a* e *a si mesmo/a* são um reforço pleonástico do complemento. Estas expressões apresentam traços de pessoa e número iguais aos do clítico. Assim, nos exemplos (7a) e (9), a expressão *a si próprio*, ao contrário do clítico, tem traços de género, para além de pessoa e número. Assim, permite forçar a interpretação pretendida: demonstrar que o clítico não está sob o escopo do antecedente com os traços de pessoa e número correspondentes aos da expressão pleonástica.

3.2. **Funções gramaticais e papéis temáticos**

Apesar de o antecedente e o clítico reflexo serem correferentes, desempenham funções gramaticais diferentes e recebem papéis temáticos distintos. Segundo Lobo (2013, pp. 2213–2214): "Numa construção reflexa, o referente designado pelo antecedente e pela expressão reflexa participa numa situação na qual tem simultaneamente dois papéis diferentes, i.e., realiza uma ação que incide sobre ele próprio."

Assim, o antecedente é um SN sujeito,[3] que recebe o papel temático de agente ou experienciador. Já o clítico reflexo pode ter a função de objeto direto, sendo-lhe atribuído o papel temático de paciente, ou de objeto indireto, caso em que recebe o papel de recipiente. Para a autora, o clítico reflexo pode receber um papel

[2] Maria Lobo (2013, p. 2219) menciona que o antecedente, em certos casos, deverá pertencer ao mesmo sintagma nominal. Nesse caso, o pronome é um constituinte preposicionado e ocorre na forma forte *si*, não na forma de clítico, como exemplificado abaixo:

(i) [A opinião da Maria sobre *si mesma*] é pouco realista.

[3] Lobo (2013) afirma que o antecedente do pronome reflexo nem sempre é o sujeito. No entanto, nesses casos, o pronome não ocorre na forma de clítico, mas na forma forte *si* e é um constituinte preposicionado. No exemplo abaixo, o antecedente pode corresponder ao sujeito ou ao objeto direto, havendo ambiguidade entre as duas interpretações.

(i) O médico protegeu o doente de *si mesmo*.

temático, o que é precisamente um dos critérios que justificam a classificação de Brito et al. (2003) dos clíticos.[4] O clítico tem ainda um estatuto argumental.

Para Brito et al. (2003, p. 807), o clítico reflexo também é tipicamente o objeto direto, podendo igualmente ser objeto indireto. As autoras apresentam dois testes que procuram comprovar que o clítico está associado a posições argumentais. Esses testes são a construção de redobro do clítico e a omissão do clítico em construções de extração simultânea. Como as autoras afirmam, essas operações "operam fundamentalmente sobre clíticos com conteúdo substantivo, argumental ou predicativo: em qualquer dos casos se pressupõe a existência de uma posição com conteúdo semântico que o clítico redobra ou recupera" (Brito et al. 2003, p. 834).

A construção de redobro do clítico consiste na expressão de um argumento através de um clítico e de um pronome forte preposicionado (*a si mesmo/a* ou *a si próprio/a*), com traços de pessoa e número correspondentes aos do clítico. Este constituinte redobrado assinala a posição argumental associada ao clítico, como é exemplificado em (10).

(10) A mãe penteou-*se a si mesma*.

As construções de extração simultânea, por sua vez, envolvem frases coordenadas, onde uma única forma clítica recupera os argumentos associados ao verbo de cada termo coordenado. Conforme é visível no exemplo (11a), o clítico reflexo pode ocorrer nestas construções. Para Brito et al. (2003, p. 836), essas frases não são sentidas como casos de objeto nulo, uma vez que o clítico c-comanda as categorias vazias a que está associado, inclusive no segundo membro coordenado. Note-se que, nos casos de ênclise, se o hospedeiro do clítico for um verbo no interior da estrutura coordenada, não é possível recuperar o conteúdo do clítico no segundo termo coordenado (Matos, 2000). Por isso, em (11b), a categoria vazia é interpretada como um caso de objeto nulo.

(11a) Ele nunca *se* penteava [-] nem barbeava [-].
(11b) # Ele penteou-se e barbeou [-].

Confirma-se, assim, que o clítico reflexo aceita construções de extração simultânea e de redobro do clítico, um dos critérios apontados por Brito et al. (2003) para considerar estes clíticos argumentais.

Duarte (2013, p. 448) refere também que o clítico reflexo surge com verbos transitivos diretos não estativos que denotam mudança de estado, de lugar e de posse. Os verbos que descrevem estados psicológicos ou ações sobre o corpo também aceitam a diátese reflexa. Todos estes verbos aceitam igualmente um objeto direto não correferente com o sujeito, como exemplificado em (12).

(12a) A mãe penteou-*se*.
(12b) A mãe penteou *a filha*.

Este facto parece confirmar que o clítico reflexo surge numa posição argumental associada ao argumento interno com função de objeto direto. Da mesma forma, certos verbos que selecionam objetos indiretos também aceitam a diátese reflexa, como mostra (13). Nesse caso, o clítico reflexo tem a função de objeto indireto.

(13a) O polícia perguntou-*se* se deveria multar o condutor.
(13b) O polícia perguntou *ao condutor* se tinha bebido.

[4] Ver Secção 2.

4. O clítico inerente e o seu papel enquanto partícula destransitivizadora

O clítico inerente ou pseudorreflexo é um pronome não anafórico, que parece não estar associado a qualquer posição argumental ou papel temático. Uma prova nesse sentido é a impossibilidade de o clítico ser redobrado, ao contrário do que acontece com os clíticos reflexos. Conforme demonstra o exemplo (14), o clítico inerente não admite as expressões *a si próprio/a* e *a si mesmo/a*.

(14) * A Maria apaixonou-*se a si mesma*.

Note-se ainda que o clítico inerente só pode ocorrer em construções de extração simultânea quando o seu hospedeiro não pertence à estrutura coordenada, como é visível em (15a), onde os elementos coordenados são os VPs ou AspPs selecionados pelo verbo semiauxiliar *estar a*. Quando isso não acontece, a frase é marginal, como é visível no contraste apresentado em (15b).

(15a) Ela estava-*se* a levantar ou a esticar, quando caiu.
(15b) ?? Ela não só *se* estava a levantar como esticou.

Finalmente, o clítico inerente surge principalmente com verbos de movimento e de postura corporal (como *dirigir-se*) e com verbos que denotam processos subjetivos de natureza psicológica ou emotiva (como *apaixonar-se* e *queixar se*). Conforme Fonseca (2010, p. 46) afirma, o sintagma nominal com função de sujeito recebe sobretudo o papel temático de experienciador, apesar de poder apresentar alguns indícios de agentividade. Referindo-se a orações pseudorreflexas, Duarte (2013, p. 449) aponta para "o facto de o constituinte com função de sujeito agir volitivamente". Isso é comprovado através do uso do advérbio *propositadamente* ou de uma oração final, como demonstrado em (16a) e (16c). Note-se ainda que o uso do advérbio *involuntariamente* em (16b) resulta numa oração agramatical.

(16a) Dirigi-me *propositadamente* ao banco.
(16b) * Zanguei-me *involuntariamente* com o meu colega.
(16c) Queixei-me, [para conseguir o reembolso do pagamento].

Assim, segundo Brito et al. (2003), o clítico inerente não tem conteúdo semântico ou morfossintático. Seguindo os critérios definidos pelas autoras, estes clíticos (i) não têm potencial referencial ou predicativo; (ii) não têm uma referência específica; (iii) não podem receber um papel temático; e (iv) não surgem em construções de redobro do clítico, admitindo a construção de extração simultânea do clítico apenas em circunstâncias específicas. As autoras consideram ainda que os clíticos inerentes (v) não têm capacidade de modificarem a estrutura argumental de um verbo.

De facto, de acordo com Brito et al. (2003), o clítico inerente não pode ser interpretado como um morfema destransitivizador, como o clítico anticausativo. O exemplo (17b) mostra que o verbo *queixar-se* não admite uma alternativa transitiva sem o clítico e com um argumento externo com o papel temático de causa. O mesmo não acontece em (18a), com o verbo inacusativo *divertir-se*. Neste caso, o clítico anticausativo pode ser considerado um morfema associado à perda do papel temático do argumento externo do verbo (Gonçalves, 1990). Por esse motivo, o verbo apresenta um alternativa transitiva com o argumento externo expresso, como é visível em (18b).

(17a) A professora queixou-se do barulho.
(17b) * O barulho queixou a professora.
(17c) * A professora queixou o barulho.

(18a) As crianças divertiram-se (com o palhaço).
(18b) O palhaço divertiu as crianças.

Já no caso do clítico inerente, para que a causa esteja explícita, é necessário recorrer a outras estruturas, como, por exemplo, a estrutura causativa de marcação excecional de caso, com o verbo *fazer* (Brito et al., 2003, p. 843). Isso é exemplificado em (19).

(19) O barulho fez a professora queixar-se.

Esta perspetiva não é inteiramente partilhada por Fonseca (2010), que defende que, em certos casos, o clítico inerente pode estar relacionado com a destransitivização de predicados, como demonstrado em (20). Quando há um clítico inerente, o verbo perde o objeto direto e, em vez disso, passa a selecionar um sintagma preposicional. Assim, o clítico é um afixo com consequências sintáticas, bem como semânticas, já que, em (20a), a leitura em que *lembrar* seleciona um objeto indireto nulo com a função de recipiente é possível. Note-se, contudo, que, no caso do clítico anticausativo, o argumento interno do verbo é realizado na posição de sujeito.

(20a) Ele lembra a matéria para o teste.
(20b) Ele lembra-se da matéria para o teste.

Alboiu et al. (2004) também consideram que o reflexo inerente é não argumental, mas que permite que o verbo selecione um sintagma preposicional. Os autores afirmam: "SE is non-argumental, and cabaple (sic) of accounting for the obligatory PP reading associated with these pronominal verbs" (Albiou et al., 2004, p. 24).

Note-se, no entanto, que as construções semelhantes a (20a), em que o verbo seleciona um objeto direto e em que não ocorre o clítico inerente, são privilegiadas, sobretudo, no português do Brasil. Como Fonseca (2010, p. 73) menciona: "A situação de desuso da Prep e, globalmente, a perda de pronominalização com verbos como: *lembrar*, *recordar*, e *rir*, entre outros, é visível sobretudo em PB. […] Pode concluir-se que os falantes de PE continuam a selecionar certos verbos com os pronomes clíticos pseudo-reflexos e o complemento preposicional respectivo."

5. A supressão do clítico nas variedades africanas do português

Um dos traços gerais apontado por alguns estudos para as variedades africanas do português é a omissão dos clíticos do paradigma reflexivo (Gonçalves, 2013; Hagemeijer, 2016). Como Hagemaijer (2016, p. 59) afirma, "em todas elas [as variedades africanas do português], observa-se uma tendência para a supressão dos pronomes reflexos".

Mendes e Estrela (2008) focam-se exatamente nesse aspeto, recorrendo a dados da oralidade das cinco variedades africanas do português, integrados no Corpus África do CLUL.[5] No total, as autoras encontraram 129 casos de omissão do clítico.[6] Com base nessas ocorrências, concluem que a variedade de São Tomé e Príncipe é aquela que apresenta um maior desvio em relação à norma do português europeu (41 das ocorrências). Além disso, o estudo demonstra que os clíticos intrinsecamente pronominais (103 ocorrências), que são não argumentais, sofrem maior omissão do que os clíticos reflexos (5 ocorrências).

Esta tendência para a supressão de clíticos não argumentais já tinha sido apontada por Gonçalves (1990) para o português de Moçambique. A autora afirma que essa variedade segue a norma europeia quanto aos

[5] O Corpus África encontra-se disponível para pesquisa na plataforma CQPweb (http://alfclul.clul.ul.pt/CQPweb/ca/). As características do corpus encontram-se descritas em Bacelar do Nascimento et al. (2006).
[6] Mendes e Estrela (2008) não apresentam dados sobre os casos de realização do clítico, isto é, que convergem com a norma do português europeu. Assim, não é claro qual é a relevância da estratégia da omissão no corpus utilizado pelas autoras.

clíticos reflexos argumentais, mas suprime pronomes clíticos pseudorreflexos e anticausativos. Propõe ainda duas hipóteses para explicar este fenómeno: (i) um processo de mudança interna do português, exacerbado pela aquisição do português como língua segunda e a perda de material funcional; ou (ii) a influência das línguas bantu em contacto com o português.

As línguas bantu não têm uma estratégia reflexiva morfológica correspondente aos clíticos inerentes do português. No entanto, no que diz respeito aos pronomes reflexos argumentais, estas línguas dispõem de um prefixo verbal com uma forma invariável (Gonçalves, 1990). Como Hagemeijer (2016, p. 60) afirma, habitualmente, a reflexividade é "marcada por verbos transitivos, […] sob a forma de um prefixo invariável, que, à semelhança dos prefixos de objeto, ocorre na posição pré-radical". Os exemplos em (21) apresentam dados do quimbundo, em que a reflexividade é marcada pelo prefixo verbal *di*, com uma forma invariável.

(21a) ngi-di-sukula. (Hagemeijer, 2016)
'Eu lavo-me.'
(21b) u-di-sukula.
'Tu lavas-te.'

Note-se, no entanto, que a variedade do português falada na Guiné-Bissau não sofre a influência de línguas bantu. Na verdade, na Guiné-Bissau, o kriol, um crioulo de base lexical portuguesa, poderá ser a principal influência por detrás das estruturas encontradas no português de Guiné-Bissau, já que é amplamente falado pela população.[7] Mesmo sem a influência de línguas bantu, Mendes e Estrela (2008) assinalam casos de omissão do clítico naquela variedade (33 das 129 ocorrências). O mesmo se passa nos arquipélagos de Cabo Verde e de São Tomé e Príncipe, onde o português também se encontra em contacto com crioulos. Assim, não parece provável que haja influência das línguas bantu. No ponto 7, será explorado o papel do contacto com os crioulos, mais especificamente, com o forro.

A pluralidade de influências linguísticas poderá ser argumentos a favor de uma mudança interna do português, já que a omissão do clítico é comum a todas as variedades (embora a relevância dessa estratégia não seja igual em todas elas). É de realçar que, em alguns dialetos do português do Brasil, também existe uma tendência para a omissão do clítico reflexo. Como Pereira (2006, pp. 32–33) afirma, "a literatura específica sobre cliticização no PB aponta o mineiro como um dos dialetos brasileiros que menos apresentam clíticos". A autora salienta sobretudo a supressão de clíticos não argumentais. Fonseca (2010) refere que, em português europeu, as construções com o clítico continuam a ser preferidas. Contudo, os verbos *esquecer*, *recordar* e *lembrar* suscitam, no caso de alguns falantes, a supressão do clítico inerente.

Estes argumentos poderão indicar que a supressão do clítico é resultado de uma mudança interna da língua, que se reflete em duas gramáticas em competição. Esta tendência poderá ser mais acentuada nas variedades africanas do português, devido ao seu passado recente como língua segunda. Com efeito, a realização dos clíticos inerentes não é regular em português, o que poderá ter contribuído para dificuldades na aquisição das regras da construção.

Neste sentido, Rosário (2005) analisa a aquisição dos clíticos por falantes de português como língua não materna (falantes nativos de francês e de inglês). O estudo chega à conclusão de que "o facto de os clíticos não-argumentais não serem argumentos de verbos e não possuírem um papel semântico parece ter dificultado o seu reconhecimento e aplicação com correcção" (2005, p. 559). Assim, a supressão dos clíticos, sobretudo dos não argumentais, pode ser um efeito da aquisição do português como L2.

Com base nos estudos referidos nesta secção, surgem quatro questões de investigação que nortearão a análise dos dados neste artigo: (i) Os clíticos inerentes, considerados não argumentais, sofrem maior omissão do que os reflexos, considerados argumentais? (ii) Será que informantes com um nível de escolaridade mais elevado apresentam uma maior percentagem de realização do clítico? (iii) Qual é a relevância da variável *idade*

[7] Quando o último censo foi realizado, em 1979, cerca de 44,3% da população falava kriol (Scantamburlo 1999, p. 58).

na estratégia da omissão do clítico? (iv) Qual é o papel do contacto com o forro na estratégia de omissão do clítico em português de São Tomé e Príncipe?

6. **Contexto linguístico de São Tomé e Príncipe**

O atual panorama linguístico em São Tomé e Príncipe reflete dois períodos distintos de colonização, que se desenrolaram ao longo dos cerca de cinco séculos de história do arquipélago. O primeiro período tem início com a fase de habitação da ilha de São Tomé (1485–1515), datando a primeira tentativa bem-sucedida de povoamento permanente de 1493. Esta fase foi particularmente favorável à crioulização, já que o contacto entre os portugueses e a mão-de-obra africana escravizada era intenso. Apesar de esta se encontrar em maioria, eram os falantes nativos de português que tinham uma posição privilegiada. Este contexto favoreceu, como Hagemeijer (2009, p. 2) afirma, "uma aproximação, por parte dos escravos, ao código linguístico utilizado pelos povoadores portugueses". Provenientes sobretudo do delta do Níger, mais especificamente do antigo Reino do Benim, os escravos eram maioritariamente falantes de línguas edóides, com particular relevância para o edo. Deste modo, essa língua teve um efeito fundador e um papel importante enquanto substrato do Proto-Crioulo do Golfo da Guiné, de base lexical portuguesa (Hagemeijer, 2011).

Essa língua tornou-se a língua alvo dos escravos que chegaram na fase de plantação (1515–1600), marcada pela introdução da cultura do açúcar. Estes escravos eram já provenientes de zonas onde se falavam línguas bantu, que desempenharam, assim, um papel de adstrato (Hagemeijer, 2011). O Proto-Crioulo do Golfo da Guiné difundiu-se no espaço, ramificando-se, com o tempo, em quatro crioulos distintos. Três desses são ainda atualmente falados em São Tomé e Príncipe: o forro, o angolar (falados na ilha de São Tomé) e o lung'ie (falado na ilha do Príncipe).[8] Estes crioulos foram dominantes até ao final do século XIX, sendo o português apenas falado pela elite local e pelos colonizadores portugueses (Hagemeijer, 2018).

No início do século XIX, uma segunda fase de colonização viria alterar o contexto linguístico de São Tomé e Príncipe. Esta fase é marcada pela introdução da cultura do café e do cacau, que geraram a necessidade de um grande contingente de mão-de-obra. Em sequência da abolição da escravatura em 1869 e da recusa dos escravos libertos em continuar a trabalhar nas plantações, os proprietários destas passam a recorrer a contratados provenientes, na sua maioria, de Angola, Cabo Verde e Moçambique (Hagemeijer, 2018) e que trazem consigo as suas línguas maternas. Em resultado, falam-se ainda atualmente no arquipélago o crioulo de Cabo Verde e, em menor grau, algumas línguas bantu, como o quimbundo e o umbundo. Como Hagemeijer (2009, p. 17) constata, "ao contrário do crioulo de Cabo Verde, as línguas do continente africano tendem a desaparecer rapidamente das ilhas, porque muitas vezes não houve transmissão de geração em geração".

Por outro lado, o português, não o forro, torna-se a língua-alvo dos contratados, que a adquirem como língua segunda. De facto, os proprietários das plantações são maioritariamente portugueses. Além disso, os antigos escravos, denominados "forros", mantêm-se deliberadamente segregados dos novos trabalhadores das plantações, que consideram de menor estatuto. Nas maiores plantações, onde muitos trabalhadores eram falantes nativos de línguas bantu, surgiu ainda uma nova variedade de contacto, o Português dos Tonga. No entanto, a sua dispersão espacial e grande variação levou a que, na atualidade, tenha praticamente desaparecido e sofrido fusão com o português (Hagemeijer, 2009).

Os contratados contribuíram, assim, para a expansão do português no arquipélago. É de salientar que o seu número chegou mesmo a ultrapassar o dos autóctones, segundo Hagemeijer (2018). Como Gonçalves (2016, p. 24) afirma, "se […] o período da primeira colonização foi favorável à crioulização, o período da segunda colonização foi determinante para a hegemonia do português, uma vez que era esta, e não o forro, a língua-alvo dos contratados." A política de repressão linguística dos crioulos por parte do Estado Novo contribuiu para este processo. Assim, no século XX, o português passou a ser usado como língua franca no arquipélago, sendo

[8] Um quarto crioulo, o Fa d'Ambô, é falado na ilha de Annobón, pertencente atualmente à Guiné Equatorial. Esta ilha pertenceu até 1778 a Portugal.

língua segunda de um número crescente de falantes bilingues. O período até à independência foi marcado por uma diglossia estável, em que o português era utilizado nos domínios altos e os crioulos nos domínios baixos.

Foi sobretudo após 1975 que a situação se alterou. De facto, o português conheceu uma expansão expressiva quer no número de falantes (como língua materna) quer nos contextos de utilização. Este fenómeno resultou num contexto de diglossia instável e de *shift*, que envolve o abandono gradual e, eventualmente, total da língua nativa de uma comunidade, em prol da língua-alvo. Como Winford (2003, p. 27) afirma "when stable bilingualism collapses, through either the erosion of ethnolinguistic boundaries or the resolution of diglossia or some other cause, the result is language shift".

Depois da sua independência, São Tomé e Príncipe adotou o português como língua oficial, de forma a garantir a "unidade nacional" num contexto multilingue. A ampliação das redes escolares, a mobilidade social e os meios de comunicação na língua oficial, entre outros fatores, favoreceram a difusão do português. Segundo Hagemeijer (2018), o português é atualmente a língua com maior expressividade no país, sendo falada por cerca de 98,4% da população. Esta expansão do português nas últimas décadas desencadeou um processo de nativização, dando origem a uma nova variedade do português (Gonçalves, 2016). O passado recente desta variedade como L2 reflete-se num espetro amplo de variação.

Por outro lado, tem-se assistido à erosão linguística e ao desaparecimento das línguas maternas, designadas frequentemente de 'línguas nacionais', cujo papel social é ocupado gradualmente pela língua oficial (Thomason, 1997; Winford, 2003). Esta tendência é reforçada pela ausência de políticas linguísticas relevantes, que protejam os crioulos. Em resultado, de acordo com o INE (2012), menos de 50% da população falava, em 2012, alguma das 'línguas nacionais'.[9] Note-se que o forro é o crioulo falado por uma maior fatia da população (34%). Assim sendo, será com base neste crioulo que se procurará perceber até que ponto o português de São Tomé e Príncipe reflete o contacto com línguas locais.

7. **Os clíticos reflexos e inerentes no forro**

Os crioulos falados em São Tomé e Príncipe, incluindo o forro, não apresentam um paradigma reflexivo que corresponda ao do português (Hagemeijer, 2016). De facto, Ferraz (1979, p. 72) afirma que o forro não recorre ao clítico. Esta estratégia ocorre em construções que em português europeu selecionam não só clíticos inerentes (22), como também clíticos reflexos (23).

(22) Bô li Ø. (Hagemeijer, 2016, p. 60)
2SG rir
'Tu riste-te.'

(23) N ga ba kenta Ø. (Hagemeijer, 2007, p. 40)
1SG ASP ir aquecer
'Eu vou aquecer-me.'

Por outro lado, Hagemeijer (2011) menciona que o forro apresenta também uma estratégia reflexiva nominal (*body-part reflexives*), recorrendo à palavra *ubwê* 'corpo', como exemplificado em (24). Esta estratégia é partilhada pelos restantes crioulos do Golfo da Guiné (Hagemeijer, 2009).

(24) Ê mat-*ubwê* dê buta. (Hagemeijer, 2011, p. 129)
3SG matar-corpo POSS PRT

[9] Os dados do censo realizado em 2012 pelo Instituto Nacional de Estatística não devem, contudo, ser interpretados sem as devidas ressalvas. Como Gonçalves (2010, p. 16) afirma, eles "não nos permitem extrair conclusões quanto à possibilidade de as línguas identificadas serem adquiridas como L1, L2 ou LE".

'Ele/a suicidou-se.'

A estratégia parece estar relacionada com uma das línguas de substrato deste crioulo: o edo. Conforme demonstrado em (25), esta língua utiliza uma estrutura semelhante e recorre igualmente à palavra correspondente a 'corpo', isto é, *ègbé*. Note-se ainda que, tanto no forro como no edo, a palavra 'corpo' é tipicamente seguida de um pronome possessivo (Hagemeijer, 2011, p. 129).

(25) Ọ gbé-ègbé èré ruà. (Ogie, 2004, p. 5)
3SG matar-corpo POSS PRT
'Ele/a suicidou-se.'

Assim, o contacto com o forro poderá refletir-se numa maior tendência para omitir o clítico na variedade de português de São Tomé e Príncipe. Em vista dos dados analisados nesta secção, levanta-se ainda a hipótese de essa variedade do português utilizar uma estratégia reflexiva nominal, recorrendo à palavra *corpo*. Na secção seguinte, serão abordadas esta e outras hipóteses, com base na análise de dados orais.

8. **Análise de dados da variedade de português de São Tomé e Príncipe**

8.1. **Metodologia**

A análise realizada nesta secção recorre a dados empíricos, tendo por base um corpus de discurso oral espontâneo, elaborado no âmbito do projeto *Posse e Localização: Microvariação em variedades africanas do português* (PALMA) do CLUL. Toma-se como base entrevistas semiestruturadas realizadas com falantes da variedade de português de São Tomé e Príncipe, recolhidas entre 2008 e 2012. O corpus é constituído por 77 entrevistas, realizadas com informantes distribuídos de acordo com o seu nível de escolaridade e idade. Relativamente ao género, 40 dos informantes são do sexo masculino e 37 do sexo feminino (cf. Tabela 1).

A idade dos informantes varia entre os 17 e os 71 anos, com uma média de 37 anos. Seguindo o modelo do Corpus PALMA, os informantes foram ainda distribuídos por quatro grupos etários e quatro níveis de escolaridade, como é visível na Tabela 1. Finalmente, todos os informantes incluídos no corpus têm o português como língua materna ou primária, embora alguns apresentem proficiência noutra língua (forro ou caboverdiano).

Tabela 1. Distribuição dos informantes por nível de escolaridade, idade e género

Variáveis		M	F	Total
Idade	17-25	9	10	77
	26-35	10	11	
	36-45	10	6	
	46-71	11	10	
Escolaridade	0 – 4.º ano	5	8	77
	5.º ano – 9.º ano	11	12	
	10.º ano – 12.º ano	19	10	
	Ensino superior (ou a frequentar)	5	7	

A extração dos dados foi feita através da leitura do corpus, de forma a identificar a supressão de clíticos. Posteriormente, fez-se a classificação dos dados, distinguindo-se as ocorrências de clíticos reflexos e de clíticos inerentes, bem com a estratégia utilizada, isto é, a realização ou a supressão do clítico. No total, foram

identificados 727 contextos onde o clítico seria realizado segundo o padrão do português europeu. Das 77 entrevistas pertencentes ao corpus, 72 apresentam contextos relevantes para a investigação em questão.

8.2. **Relevância da estratégia de supressão do clítico**

Dos 727 contextos relevantes, 449 correspondem a casos de supressão do clítico, como demonstra a Tabela 2. Os informantes parecem, assim, demostrar uma preferência pela estratégia de supressão, que corresponde a 61,76% das ocorrências. Salienta-se a variação entre as duas estratégias, o que poderá estar relacionado com o passado recente desta variedade como L2 (Hagemeijer, 2016, p. 49). Ainda assim, evidencia-se um grau de divergência considerável relativamente ao português europeu. Estes dados parecem estar de acordo com a conclusão de Hagemeijer (2016), isto é, de que a tendência de supressão é particularmente acentuada no português de São Tomé e Príncipe. Note-se ainda que, de acordo com os dados do Corpus África, analisados por Mendes e Estrela (2008, p. 103), a variedade de São Tomé e Príncipe foi a que apresentou mais contextos de omissão do clítico.

Tabela 2. Supressão e realização dos clíticos (dados gerais)

	Supressão	**Realização**	**Total**
Ocorrências	449 (61,76%)	278 (38,24%)	727 (100%)

A estratégia da omissão abrange não só casos de clíticos inerentes, considerados não argumentais, mas também clíticos reflexos. Vejam-se os exemplos em (26), relativos à supressão do clítico reflexo, e em (27), relativos à supressão do clítico inerente. Apresentam-se ainda exemplos de realização do clítico reflexo em (28) e do clítico inerente em (29).

(26a) Eu *vesti* Ø, fui.
(26b) Levanta cedo, *prepara* Ø para a escola.
(26c) Esse mais velho já Ø *matriculou*.

(27a) E também Ø *habituei* já com clima cá na cidade.
(27b) Quando a minha mãe Ø *concentrou* e tal, para ajudar.
(27c) *Queixar* Ø aí para a violência doméstica.

(28a) Eu tinha que *inscrever*-***me*** naquela disciplina.
(28b) Para ir-***me*** *lavar*, preparar.
(28c) Não quer dizer que eu ***me*** *exclua* desses assuntos.

(29a) Por isso, eu não ***me*** *lembro* quem era.
(29b) *Esqueci*-***me*** do nome.
(29c) Eu acho que devemos ***nos*** *esforçar*.

Tabela 3. Supressão e realização dos clíticos reflexos e inerentes (dados gerais)

Ocorrências	**Reflexo**	**Inerente**
Supressão	125 (54,35%)	324 (65,19%)
Realização	105 (45,65%)	173 (34,81%)
Total	230 (100%)	497 (100%)

Vários estudos têm apontado para uma probabilidade maior de supressão do clítico inerente relativamente ao clítico reflexo, já que este está associado a uma posição argumental (Gonçalves, 1990; Mendes & Estrela, 2008; Pereira, 2006). Os dados orais analisados neste estudo parecem, de facto, confirmar essa tendência, conforme plasmado na Tabela 3. O clítico inerente é suprimido 65,19% das vezes, enquanto o clítico reflexo é omitido em 54,35% dos contextos. A omissão do clítico reflexo é expressiva, apesar de não ser tão acentuada como nos clíticos inerentes.

8.3. **A relevância de variáveis sociolinguísticas**

A variedade de português falada em São Tomé e Príncipe tem um passado recente como variedade L2, pelo que apresenta um espetro amplo de variação, para que muito contribuem variáveis sociolinguísticas, como o nível de escolaridade, idade e o grau de exposição à língua. Hagemeijer (2009, p. 19) afirma que, em São Tomé, existem vários registos de português, com maior ou menor influência dos crioulos, o que é "muitas vezes determinado por factores tais como o nível de escolaridade, nível económico e o ambiente de inserção social (urbano/rural)". Com isso em mente, analisar-se-á de seguida os dados recolhidos, à luz das variáveis *escolaridade* e *idade*.

8.3.1. **A escolaridade**

Como mencionado anteriormente, o corpus analisado é constituído por quatro grupos com níveis de escolaridade distintos. A escolaridade tem sido apontada como um fator de aproximação à norma do português europeu. De acordo com Gonçalves e Chimbutane (2004, p.7):

> A distribuição e frequência dos traços não-padrão no discurso dos falantes não é idêntica para todos os membros desta comunidade. Com efeito, à semelhança do que acontece com outras línguas ex-coloniais, o português moçambicano apresenta um amplo espectro de variação que inclui desde as subvariedades 'basilectais', mais distantes do padrão europeu, dos falantes com pouco contacto com a língua-alvo, até às subvariedades mais próximas deste padrão, dos falantes mais instruídos.

De facto, conforme mostra a Tabela 4, os dados do corpus analisado parecem indicar que, quanto mais elevado o nível de escolaridade, maior a convergência com a norma. Por exemplo, os informantes com o ensino superior privilegiam a estratégia de realização do clítico (que corresponde a 63,57% das ocorrências), ao contrário do que sucede com os informantes menos escolarizados. Os informantes que frequentaram até ao 4.º ano apresentam a percentagem mais elevada de supressão do clítico (80,9%). Note-se ainda que, neste grupo, dez das dezassete ocorrências de realização de clíticos são produzidas por um só informante (58,82%). Este é o que apresenta uma atividade profissional mais especializada, sendo diretor de um centro de cultura e de um arquivo. Este facto pode indicar que o grau de convergência com a norma poderá estar associado a outros fatores sociolinguísticos, não necessariamente relacionados com a escolaridade.

Tabela 4. Supressão e realização dos clíticos (por escolaridade)

Escolaridade	Supressão	Realização	Total
0 – 4.º ano	72 (80,9%)	17 (19,1%)	89 (100%)
5.º – 9.º ano	140 (78,21%)	39 (21,79%)	179 (100%)
10.º – 12.º ano	190 (57,58%)	140 (42,42%)	330 (100%)
Ensino superior	47 (36,43%)	82 (63,57%)	129 (100%)

É ainda possível verificar que os informantes menos escolarizados apresentam percentagens elevadas de omissão tanto do clítico inerente como do clítico reflexo, como mostra a Tabela 5. Contudo, o clítico inerente

caracteriza-se por uma maior tendência para a omissão. O grupo que abrange os informantes dos 0 aos 4 anos de escolarização omite o clítico inerente 86,67% das vezes e o clítico reflexo 68,97%. Esta diferença entre clíticos argumentais e não argumentais parece esbater-se nos informantes dos níveis de escolaridade seguintes.

Tabela 5. Supressão e realização dos clíticos reflexos e inerentes (por escolaridade)

Escolaridade	Clítico	Supressão	Realização	Total
0 – 4.º ano	Inerente	52 (86,67%)	8 (13,33%)	60 (100%)
	Reflexo	20 (68,97%)	9 (31,03%)	29 (100%)
5.º – 9.º ano	Inerente	104 (76,47%)	32 (23,53%)	136 (100%)
	Reflexo	36 (83,72%)	7 (16,28%)	43 (100%)
10.º – 12.º ano	Inerente	142 (60,42%)	93 (39,57%)	235 (100%)
	Reflexo	48 (50,52%)	47 (49,47%)	95 (100%)
Ensino superior	Inerente	26 (39,39%)	40 (60,61%)	66 (100%)
	Reflexo	21 (33,33%)	42 (66,67%)	63 (100%)

Finalmente, faz-se uma distinção entre os informantes com base nas estratégias que utilizam, tendo sido identificados três grupos distintos, conforme representado na Tabela 6. A maioria dos informantes, isto é, 55 falantes, apresenta variação entre supressão e omissão do clítico, independentemente do seu nível de escolarização (grupo II). Assim, utilizam diferentes estratégias, o que pode refletir gramáticas em competição a nível individual. Como Lightfoot (2006, p. 164) constata, "there are coexisting grammars within [these] speech communities and even within the brains of some individuals, who have multiple competencies". Por outro lado, apenas um informante com o ensino superior realiza o clítico em todos os contextos, seguindo a norma do português europeu (grupo III). Assim, a estratégia da realização parece ser pouco relevante, enquanto estratégia única, já que a grande maioria dos informantes omite o clítico em alguma instância. Finalmente, nenhum informante com o ensino superior opta pela estratégia de omissão em todos os contextos.

Tabela 6. Distribuição da supressão e realização do clítico

Escolaridade	Grupo I (supressão)	Grupo II (supressão e realização)	Grupo III (realização)	Total
0 – 4.º ano	7	5	0	12
5.º – 9.º ano	6	16	0	22
10.º – 12.º ano	3	24	0	27
Ensino superior	0	10	1	11
Total	16	55	1	72

8.3.2. **A idade**

As variáveis sociolinguísticas, como mencionado anteriormente, têm uma relevância acrescida nas variedades africanas do português que se encontram em fase de nativização. Assim, procurou-se identificar a relevância da variável *idade* na estratégia da omissão do clítico.

De forma a investigar esse aspeto, agrupou-se os informantes em quatro conjuntos, tendo por base a idade dos mesmos. Para tal, seguiu-se a divisão prévia estabelecida no Corpus PALMA, como explanado na secção 8.1. Os dados da Tabela 7 mostram que informantes mais novos apresentam percentagens de omissão do clítico semelhantes às dos informantes mais velhos. Por exemplo, os dois grupos de informantes mais novos (dos 18 aos 35 anos) apresentam taxas de omissão semelhantes à do grupo de informantes mais velhos (dos 46 aos 7 anos). Apenas o grupo dos 36 aos 45 anos apresenta uma taxa de supressão mais elevada. Isto poderá indicar que a variável *idade* é menos importante do que outros fatores, como a *escolarização*, na estratégia de omissão.

Tabela 7. Supressão e realização dos clíticos (por idade)

Idade	Supressão	Realização	Total
18–25 anos	70 (60,87%)	45 (39,13%)	115 (100%)
26–35 anos	163 (58,63%)	115 (41,37%)	278 (100%)
36–45 anos	93 (75,61%)	30 (24,39%)	123 (100%)
46–71 anos	123 (58,29%)	88 (41,71%)	211 (100%)

8.4. **O papel do contacto com o forro**

A forte tendência para a supressão do clítico poderá ser um reflexo do contacto com o forro, como mencionado na Secção 7. O forro não tem um paradigma reflexivo correspondente ao do português europeu, não dispondo de clíticos reflexos e inerentes. Pode, assim, ter ocorrido *transfer* desta propriedade do crioulo. Convém salientar, contudo, que não foram encontradas ocorrências de uma estratégia reflexiva nominal, que também se encontra atestada no forro (conforme o exemplo (24)). Além disso, o português é, na atualidade, hegemónico em São Tomé e Príncipe e a sua transmissão dá-se sobretudo como L1, como enunciado na Secção 6. O papel do forro como língua de contacto é, assim, primariamente histórico.

É de salientar também que a tendência de supressão converge com o que é observado em outras variedades do português em África e até em alguns dialetos do português do Brasil, conforme explanado no ponto 5. Deste modo, a mudança linguística poderá resultar de princípios gerais da gramática, relacionados com a aquisição (histórica) de L2. A variedade do português em São Tomé e Príncipe foi, pelo menos numa fase inicial, adquirida como língua segunda num contexto de pouca exposição à norma-padrão do português. Assim, a falta de *input* sólido poderá ter conduzido a uma reanálise de traços gramaticais, dando origem ao fenómeno de supressão do clítico. Note-se ainda que o clítico corresponde a uma forma átona, sem acentuação própria e foneticamente fraca, o que poderá dificultar a sua aquisição num contexto de L2. Recorde-se Rosário (2005), que concluiu que os falantes de português como língua não materna têm mais dificuldade em reconhecer e aplicar com correção os clíticos não argumentais. De facto, o clítico inerente não é argumental e a sua realização é, por vezes, opcional, o que poderá contribuir para a sua supressão. Isso seguiria uma tendência já documentada em português de São Tomé e Príncipe para a perda de material funcional, como clíticos acusativos (Gonçalves, 2016) e preposições mais fracas e opacas, principalmente *a*, *de* e *em* (Gonçalves, 2010).

Em conclusão, a tendência para a supressão do clítico no português falado em São Tomé e Príncipe parece não se dever a apenas um fator. Como Alexandre et al. (2011, p. 32) realçam, há "um enredo complexo de factores de natureza sociolinguística e extra-linguística que ditou a emergência de uma variedade L1 em S.

Tomé e Príncipe que se apresenta mais afastada da norma europeia", quando comparada com outras variedades do português.

9. Conclusão

A tendência para a supressão de pronomes do paradigma reflexivo tem sido observada para as variedades do português em África em diversos estudos (e.g., Mendes & Estrela, 2008). Este artigo procurou contribuir para uma melhor compreensão desse fenómeno, bem como para a expansão do conhecimento sobre a variedade do português de São Tomé e Príncipe. A análise quantitativa de dados orais permitiu lançar luz sobre as quatro questões basilares deste estudo: (i) Os clíticos inerentes, considerados não argumentais, sofrem maior omissão do que os reflexos, considerados argumentais? (ii) Será que informantes com um nível de escolaridade mais elevado apresentam uma maior percentagem de realização do clítico? (iii) Qual é a relevância da variável *idade* na estratégia da omissão do clítico? (iv) Qual é o papel do contacto com o forro na estratégia de omissão do clítico em português de São Tomé e Príncipe?

Os dados recolhidos parecem, de facto, indicar que esta variedade apresenta uma forte tendência para a supressão do clítico. Os clíticos inerentes sofrem maior omissão do que os reflexos, embora estes também sejam suprimidos. Além disso, a variável sociolinguística *escolaridade* parece ter algum peso na estratégia privilegiada pelos falantes. Este não parece ser o caso da variável *idade*, uma vez que os informantes mais novos apresentam percentagens de omissão do clítico semelhantes às dos informantes mais velhos.

Por outro lado, o contacto com o forro poderá ter influência na variedade do português que está em formação em São Tomé e Príncipe, uma vez que não tem um paradigma reflexivo correspondente ao do português europeu. Ainda assim, por si só, o contacto linguístico poderá ser insuficiente para explicar a tendência para a supressão, já que não são atestadas outras estratégias reflexivas do forro em português de São Tomé e Príncipe. Assim sendo, parecem estar envolvidos vários fatores, incluindo o passado recente desta variedade como língua segunda.

Finalmente, este artigo permitiu esboçar novas questões tendo em vista investigação futura, mais especificamente: (i) A omissão dos clíticos reflexos e inerentes é condicionada pela classe de verbos? (ii) Construções com verbos de experiência psicológica favorecem a omissão do clítico, conforme sugerido para o português europeu em Fonseca (2010)?

Trabalhos futuros procurarão responder a estas questões com base numa análise mais fina dos dados. A investigação nesse sentido permitirá aprofundar o conhecimento sobre processos de nativização, bem como sobre as variedades que estão em formação nas ex-colónias portuguesas em África.

Agradecimentos

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do Projeto Estratégico UIDB/00214/2020 e UI/BD/152294/2021.

Referências

Alboiu, Gabriela, Michael Barrie & Chiara Frigeni (2004) SE and the unaccusative-unergative paradox. *Antwerp Papers in Linguistics* 107, pp. 109–141.

Alexandre, Nélia, Rita Gonçalves & Tjerk Hagemeijer (2011) A formação de frases relativas de PP no português de Cabo Verde e de São Tomé. In *Actas do XXVI Encontro da Associação Portuguesa de Linguística*. APL, pp. 17–34.

Bacelar do Nascimento, Maria Fernanda, Luísa Alice Santos Pereira, Antónia Estrela, José Bettencourt Gonçalves, Sancho Oliveira & Rui Santos (2006) The African varieties of Portuguese: Compiling comparable corpora and analyzing data-derived lexicon. In *Proceedings of the Fifth International*

Conference on Language Resources and Evaluation. ELRA, pp. 1791–1794. Disponível em http://www.lrec-conf.org/proceedings/lrec2006/pdf/654_pdf.pdf
Brito, Ana Maria, Inês Duarte & Gabriela Matos (2003) Tipologia e distribuição das expressões nominais. Tipologia dos pronomes clíticos. In Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte, Isabel Hub Faria, Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário & Alina Villalva (orgs.), *Gramática da língua portuguesa*. Caminho, pp. 79–867.
Chomsky, Noam (1981) *Lectures on government and binding* (5.ª ed.). Foris.
Duarte, Inês (1983) Variação paramétrica e ordem dos clíticos. *Revista da Faculdade de Letras*, n.º especial comemorativo do 50.º aniversário da revista, pp. 158–176.
Duarte, Inês, & Gabriela Matos (2000) Romance clitics and the minimalist program. In João Costa (org.), *Portuguese syntax. New comparative studies.* Oxford University Press, pp. 116–142.
Duarte, Inês (2013) Construções ativas, passivas, incoativas e médias. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do português* (Vol. 1). Fundação Calouste Gulbenkian, pp. 429–458.
Ferraz, Luiz Ivens (1979) *The Creole of São Tomé*. Witwatersrand University Press.
Fonseca, Paula (2010) *Os verbos pseudo-reflexos em Português Europeu*. Dissertação de mestrado, Universidade do Porto.
Gonçalves, Perpétua (1990) *A construção da gramática do português em Moçambique: Aspectos da estrutura argumental dos verbos*. Dissertação de doutoramento, Universidade de Lisboa.
Gonçalves, Perpétua (2012, 7–10 fevereiro) *Lusofonia em Moçambique: Com ou sem glotofagia?*. [Apresentação de comunicação]. II Congresso Internacional de Linguística Histórica, São Paulo, Brasil. Disponível em https://catedra.itcom.co.mz/storage/app/media/bibliografia/Goncalves_Ataliba2012.pdf
Gonçalves, Perpétua (2013) O português em África. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do português* (Vol. 1). Fundação Calouste Gulbenkian, pp. 157–168.
Gonçalves, Perpétua, & Feliciano Chimbutane (2004) O papel das línguas bantu na génese do português de Moçambique: O comportamento sintáctico de constituintes locativos e direcionais. *Papia* 14, pp. 7–30.
Gonçalves, Rita (2010) *Propriedades de subcategorização verbal no português de São Tomé*. Dissertação de mestrado, Universidade de Lisboa.
Gonçalves, Rita (2016) *Construções ditransitivas no português de São Tomé.* Dissertação de doutoramento, Universidade de Lisboa.
Hagemeijer, Tjerk (2007) *Clause structure in Santome*. Dissertação de doutoramento, Universidade de Lisboa.
Hagemeijer, Tjerk (2009) As línguas de S. Tomé e Príncipe. *Revista de Crioulos de Base Lexical Portuguesa e Espanhola* 1 (1), pp. 1-27.
Hagemeijer, Tjerk (2011). The Gulf of Guinea Creoles. *Journal of Pidgin and Creole Languages* 26 (1), pp. 111–154. https://doi.org/10.1075/jpcl.26.1.05hag
Hagemeijer, Tjerk (2016) O português em contacto em África. In Ana Maria Martins & Ernestina Carrilho (orgs.), *Manual de linguística portuguesa*. Mouton de Gruyter, pp. 43–67.
Hagemeijer, Tjerk (2018) From creoles to Portuguese. Language shift in São Tomé and Príncipe. In Laura Álvarez López, Perpétua Gonçalves & Juanito Ornelas de Avelar (orgs.), *The Portuguese language continuum in Africa and Brazil*. John Benjamins, pp.169–184.
Lightfoot, David (2006) *How new languages emerge.* Cambridge University Press.
Lobo, Maria (2013) Dependências referenciais. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do português* (Vol. 2, 2.ª ed.). Fundação Calouste Gulbenkian, pp. 2177–2227.
Mapasse, Ermelinda (2005) *Clíticos pronominais em português de Moçambique.* Dissertação de doutoramento, Universidade de Lisboa.

Martins, Ana Maria (2013) Posição dos pronomes pessoais clíticos. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do ortuguês* (Vol. 2, 2.ª ed.). Fundação Calouste Gulbenkian, pp. 2231–2302.
Matos, Gabriela (2000) Across-the-Board clitic placement in Romance languages. *Probus* 12 (2), pp. 229–259.
Matsinhe, Sozinho (1998) *Pronominal clitics in Tsonga and Mozambican Portuguese: A comparative study*. Dissertação de doutoramento, School of Oriental and African Studies.
Mendes, Amália, & Antónia Estrela (2008) Constructions with SE in African varieties of Portuguese. *Phrasis* 2, pp. 83–107.
Otero, Carlos Peregrín (1999) Pronombres reflexivos y recíprocos. In Ignacio Bosque & Violeta Demonte (orgs.), *Gramática descriptiva de la lengua española* (Vol. 1, 3.ª ed.). Real Academia Española & Espasa Calpe, pp. 1426–1517.
Pereira, Ana (2006) *Os pronomes clíticos do PB contemporâneo na perspectiva teórica da Morfologia Distribuída*. Dissertação de doutoramento, Universidade Federal de Santa Catarina.
Rosário, Joana (2005) Aquisição dos clíticos por falantes de português língua não materna. In Dulce Carvalho, Vila Maior Dionísio & Rui de Azevedo Teixeira (orgs.), *Des(a)fiando discursos: Homenagem a Maria Emília Ricardo Marques*. Universidade Aberta, pp. 553–562.
Scantamburlo, Luigi (1999) *Dicionário do guineense* (Vol. 1). Edições Colibri & FASPEBI.
Thomason, Sarah (1997) Typology of contact languages. In Arthur K. Spears & Donald Winford (orgs.), *The structure and status of pidgins and creoles*. John Benjamins, pp. 71–88.
Vigário, Mariana (1999) Pronominal cliticization in European Portuguese: A postlexical operation. *Catalan Working Papers in Linguistics* 7, 219–237.
Winford, Donald (2003) *An introduction to contact linguistics*. Blackwell Publishing.
Zwicky, Arnold (1977) *On clitics*. Indiana University Linguistics Club.

Modo verbal em descrições de recusa de factos

Rui Marques[1]

[1]Universidade de Lisboa, FLUL/CLUL

Resumo

O significado de frases que expressam a recusa de um facto envolve duas proposições: uma proposição completiva que descreve um facto e a proposição, matriz, que expressa a não aceitação da primeira como verdadeira. Em português, o modo que ocorre na oração completiva não é o mesmo em todas as construções que veiculam este significado. Nalgumas construções, embora se verifique uma alternância entre o Indicativo e o Conjuntivo na oração completiva, só com Indicativo se tem a interpretação de que esta proposição descreve um facto da realidade. Noutras construções, o modo da oração completiva é selecionado pelo predicado da frase matriz, não existindo alternância de modo, e existem predicados factivos que regem Indicativo e outros que regem Conjuntivo. Em todas estas construções o modo da oração completiva é explicável por análises de modo em português disponíveis na literatura. No entanto, há dois tipos de construção do português que expressam a recusa do facto descrito pela oração completiva e em que o modo desta oração é o inverso do que se esperaria. Propõe-se neste texto uma descrição do significado dessas construções e uma análise dos modos Indicativo e Conjuntivo em português que explica de uma forma integrada o modo observado em todas estas construções que expressam uma atitude de rejeição do facto descrito pela oração completiva.

Palavras-chave: Modo, indicativo, conjuntivo, atitudes proposicionais, veridicidade.

Abstract

The meaning of sentences that express the refusal of a fact involves two propositions: a complement clause that describes a fact, and the matrix proposition, that expresses the non-acceptance of the truth of the embedded proposition. In Portuguese, the mood that occurs in the complement clause is not the same in all constructions that convey this meaning. In some constructions, although there is an alternation between Indicative and Subjunctive in the complement clause, only with Indicative the interpretation that this proposition describes a fact of reality is observed. In other constructions, the mood of the complement clause is a matter of lexical selection, no alternation of mood being observed. Some factive predicates rule the Indicative, others rule the Subjunctive. In all these constructions the mood of the complement clause is explainable by mood analyzes in Portuguese available in the literature. However, there are two types of construction in Portuguese that express the refusal of the fact described by the complement clause and in which the mood of this clause is the opposite of the expected one. This text proposes a description of the meaning of these constructions, as well as an analysis of the Indicative and Subjunctive moods in Portuguese that explains in an integrated way the mood observed in all these constructions that express an attitude of rejection of the fact described by the complement clause.

Keywords: Mood, indicative, subjunctive, propositional attitudes, veridicality.

1. Introdução

Frases como as que se seguem descrevem uma atitude de não aceitação como verdadeira da oração completiva:

(1a) Ele recusa-se a acreditar que *perdeu* as eleições.

(1b) Ele recusa-se a acreditar que *tenha perdido* as eleições.

(2a) Ele não aceita que *perdeu* as eleições.
(2b) Ele não aceita que *tenha perdido* as eleições.

(3a) (Na altura,) eu não acreditava que ele *era* espanhol.
(3b) (Já na altura,) eu não acreditava que ele *fosse* espanhol.

Esta é a única informação que é veiculada pelas frases das alíneas (b), com Conjuntivo na oração completiva, mas não pelas frases das alíneas (a), em que na oração completiva o verbo ocorre no Indicativo e que indicam também que a proposição completiva é tida como verdadeira pelo enunciador no tempo de enunciação. Ou seja, nestas construções, com Indicativo na oração completiva as estruturas descrevem um contraste de crenças (de diferentes entidades, em (1a) e (2a), ou da mesma entidade em diferentes alturas, em (3a)), enquanto que com Conjuntivo na oração completiva as frases descrevem apenas a atitude da entidade identificada pelo sujeito da frase matriz em relação à proposição completiva. Por outras palavras, nas alíneas (a), com Indicativo na oração completiva, a proposição completiva descreve um facto da realidade, mas não nas alíneas (b), com Conjuntivo.

Estes dados são explicáveis pelas análises do modo verbal em português disponíveis na literatura, mas o mesmo não se pode dizer acerca de dois tipos de construção que expressam igualmente uma atitude de não aceitação do facto descrito pela oração completiva. O primeiro, exemplificado por (4), é a construção formada pelo verbo *acreditar* sob escopo da negação, flexionado na 1.ª pessoa do singular e Indicativo na oração completiva. O segundo, exemplificado por (5), é a construção, também negativa, com a sequência *querer saber* e com Conjuntivo na oração completiva:

(4a) [contexto: o enunciador acaba de constatar que a loja está fechada]
Não acredito que a loja *está* fechada!
(4b) «Ainda não acredito que *venci* depois de ter perdido tantas corridas em cima da meta.»
(CETEMPÚBLICO, *par=ext84460-des-91a-1*)

(5a) Eu vou sair; não quero saber que *esteja* a chover!

(5a) «Considero-o um cobarde, não quero saber que *tenha sido* um grande dançarino."
(CETEMPÚBLICO, *par=ext286064-clt-93a-2*)

O presente texto pretende oferecer uma explicação para o modo verbal usado em português em orações completivas que descrevem factos, com particular foco nos casos em que se expressa em relação a essa proposição uma atitude de recusa da aceitação da proposição como verdadeira.

2. **Conjuntivo como o modo complementar**

Genericamente, a distribuição dos modos Indicativo e Conjuntivo em português parece poder explicar-se pela proposta de Marques (1995), segundo a qual o Indicativo ocorre em contextos verídicos e epistémicos e o Conjuntivo nos outros contextos. Ou seja, o Indicativo ocorre nas frases em relação às quais se verifiquem conjuntamente duas condições: (i) serem verdadeiras no modelo relativamente ao qual são avaliadas, (ii) ser expressa para com as mesmas uma atitude de crença ou conhecimento (i.e., uma atitude do plano epistémico, no sentido lato do termo, que abrange as noções de conhecimento e crença). Em frases independentes ou adverbiais, o modelo relativamente ao qual as frases são avaliadas é o modelo que representa as crenças do enunciador, enquanto orações completivas são avaliadas relativamente ao modelo introduzido pelo predicado da frase matriz. Por exemplo, a oração completiva do verbo *sonhar* é avaliada relativamente ao modelo que representa o sonho, a oração completiva de *achar* é avaliada relativamente ao modelo que representa as crenças da entidade

identificada pelo sujeito da frase matriz, etc. Saliente-se que o conceito de veridicidade e o de factualidade são distintos. Uma frase que ocorra num contexto factivo descreve um facto do mundo real. Uma frase que ocorra num contexto verídico é tida como verdadeira no modelo relativamente ao qual é avaliada, pelo que, no caso de o modelo em questão não representar o mundo real, o facto de a frase ser verídica não implica que seja verdadeira no mundo real. Ou seja, factividade implica veridicidade, mas não o inverso. Veja-se, por exemplo, que as frases (6) e (7) indicam que as respetivas orações completivas são verdadeiras no modelo relativamente ao qual são avaliadas (o sonho do Pedro, no caso de (6) e as crenças do mesmo, no caso de (7)), independentemente do valor de verdade destas no mundo real. Trata-se, portanto, de orações verídicas, mas não factivas. Já em (8), a oração completiva é factiva, descrevendo um facto da realidade. Por último, em (9) a oração completiva não é factiva – não descreve um facto – nem verídica – a construção não permite inferir que a proposição completiva é tida como verdadeira por alguma entidade:

(6) O Pedro sonhou que estava a passear numa floresta.

(7) O Pedro acha que o Jorge está em Espanha.

(8) O Pedro sabe que a Ana está em casa.

(9) O Pedro espera que o filho dele estude Medicina.

Se em diferentes línguas, como o Grego Moderno, de acordo com Giannakidou (1994) e vários outros textos, a veridicidade é a propriedade que determina a distribuição dos modos Indicativo e Conjuntivo (simplificadamente, o Conjuntivo ocorre em contextos não verídicos e o Indicativo em contextos verídicos), em português, tal como noutras línguas românicas, a veridicidade é condição necessária para a legitimação do Indicativo, mas não suficiente. De facto, tanto o Indicativo como o Conjuntivo podem ocorrer em contextos verídicos, como mostra, por exemplo, o par de frases que se segue:

(10) O Pedro descobriu que a Ana *nasceu* / **tenha nascido* em Évora.

(11) Lamento que o prazo **foi* / *tenha sido* antecipado.

Qualquer destas orações completivas descreve um facto, pelo que ocorre num contexto verídico. No entanto, na primeira o verbo flexiona obrigatoriamente no Indicativo e na segunda no Conjuntivo. A razão para que alguns predicados factivos, como *descobrir*, sejam regentes de Indicativo e outros, como *lamentar*, sejam regentes de Conjuntivo deve-se ao tipo de atitude proposicional que expressam. O verbo *descobrir* expressa uma atitude do plano epistémico, enquanto *lamentar* expressa uma atitude avaliativa (não epistémica, portanto). Mais genericamente, observa-se que em português o Indicativo ocorre apenas em contextos epistémicos e verídicos (i.e., o verbo de uma frase flexiona no Indicativo apenas se for expressa para com essa frase uma atitude do plano epistémico e a frase for tida como verdadeira no modelo relativamente ao qual é avaliada), ocorrendo o Conjuntivo nos outros contextos. Nesse sentido, o Conjuntivo não assinalará nenhum valor semântico particular, sendo antes o modo que ocorre por defeito nos contextos em que não se verifiquem as duas condições conjuntas que legitimam o Indicativo, este sim o modo que assinalará um valor semântico particular. A Tabela 1 resume esta proposta:

Tabela 1. Fatores determinantes dos modos Indicativo e Conjuntivo

Epistémico	Verídico		Exemplos
+	+	Indicativo	*descobrir, saber, achar, ...*
+	–	Conjuntivo	*duvidar, talvez, ...*
–	+	Conjuntivo	*lamentar, surpreender, ...*
–	–	Conjuntivo	*querer, pedir, tentar, ...*

Posto isto, retomem-se as frases (1) – (3), em que o verbo *acreditar* ocorre sob o escopo da negação, seguidamente repetidas e renumeradas:

(12a) Ele recusa-se a acreditar que *perdeu* as eleições.
(12b) Ele recusa-se a acreditar que *tenha perdido* as eleições.

(13a) Ele não aceita que *perdeu* as eleições.
(13b) Ele não aceita que *tenha perdido* as eleições.

(14a) (Na altura,) eu não acreditava que ele *era* espanhol.
(14b) (Já na altura,) eu não acreditava que ele *fosse* espanhol.

Nestas construções, o verbo da oração completiva pode flexionar no Indicativo ou no Conjuntivo, sendo a opção por um ou outro modo acompanhada por uma diferença de significado. Com Indicativo, veicula-se a informação de que a oração completiva é tida como verdadeira pelo enunciador no tempo de enunciação, contrariamente ao que se verifica com o Conjuntivo, caso em que as frases descrevem apenas uma atitude de não aceitação como verdadeira da proposição completiva por parte da entidade identificada pelo sujeito da frase matriz no intervalo de tempo a que esta frase faz referência. Tanto nas alíneas (a) como nas alíneas (b), cada uma destas frases expressa uma atitude do plano epistémico (uma atitude de não crença). Assim, no caso das alíneas (b), como a atitude expressa para com a proposição completiva é do plano epistémico, mas a mesma não é tida como verdadeira, o seu verbo flexiona no Conjuntivo, como previsto. Diferentemente, nos casos das alíneas (a) destes exemplos, as frases expressam, além disso, a informação de que a proposição completiva é tida como verdadeira pelo enunciador (no tempo de enunciação). Por outras palavras, nas alíneas (b), a proposição completiva é avaliada relativamente ao modelo que representa as crenças da entidade identificada pelo sujeito da frase matriz (no intervalo de tempo referido), enquanto nas alíneas (a) é avaliada também relativamente ao modelo que representa as crenças do enunciador (no tempo de enunciação), sendo apresentada como verdadeira nesse modelo (ver também Quer (1998), que faz a mesma observação a respeito de casos análogos do espanhol e do catalão). Assim, como a proposição completiva é verídica e é expressa em relação à mesma uma atitude do plano epistémico, o seu verbo flexiona no Indicativo.

3. **Casos problemáticos**

Dado que em frases que descrevem a ausência de crença na verdade da proposição completiva o Indicativo decorre do facto de o enunciador expressar a sua própria crença na verdade dessa proposição, é de esperar que não seja possível usar este modo na oração completiva deste tipo de construção quando o sujeito da frase matriz identifica o enunciador. De facto, em tal situação, a frase veicularia a informação contraditória de que a mesma pessoa acredita que a proposição completiva é verdadeira e acredita que não o é. Essa previsão confirma-se, como mostra o seguinte contraste:

(15) #Eu não acredito que *ganharei* o Euromilhões.

(16) Eu não acredito que *ganhe* o Euromilhões.

No entanto, exemplos como os seguidamente repetidos e renumerados são gramaticais, contrariamente ao esperado:

(17a) [contexto: o enunciador acaba de constatar que a loja está fechada]
Não acredito que a loja *está* fechada!
(17b) «Ainda não acredito que *venci* depois de ter perdido tantas corridas em cima da meta.»
(CETEMPÚBLICO, *par=ext84460-des-91a-1*)

A proposta acima resumida, de que o Indicativo ocorre em contextos epistémicos e verídicos, não explica esta construção, que, como dito, devia ser impossível por expressar a contradição de que a mesma entidade acredita que a proposição completiva é verdadeira e que é falsa. No entanto, a construção existe e não causa estranheza.

Outro tipo de construção em que na oração completiva ocorre o modo inverso ao que é esperado pela observação de que o Indicativo ocorre em contextos simultaneamente epistémicos e verídicos, sendo o Conjuntivo o modo complementar é ilustrado pelas seguintes frases:

(18a) Eu vou sair; não quero saber que *esteja* a chover!
(18b) «Considero-o um cobarde, não quero saber que *tenha sido* um grande dançarino»
(CETEMPÚBLICO, *par=ext286064-clt-93a-2*)

Também nesta construção a proposição completiva descreve um facto[1] e, como se expressa perante a mesma uma atitude epistémica, codificada pelo verbo *saber*, prevê-se que o Indicativo, mas não o Conjuntivo, possa ocorrer na oração completiva. No entanto, essa previsão não se confirma.

Veja-se que, noutras construções, com os mesmos verbos na frase matriz, igualmente negativa, o modo que ocorre na oração subordinada é o esperado. Nos exemplos de (19), a oração completiva ocorre num contexto não verídico, pelo que o verbo flexiona no Conjuntivo, como esperado. Nos exemplos de (20), a oração completiva descreve um facto (ou seja, é verídica) e ocorre num contexto epistémico, pelo que o verbo flexiona no Indicativo, como previsto:

(19a) Não acredito que a loja *esteja* fechada.
(19b) Ainda não acredito que *tenha vencido*.

(20a) Se ele não quisesse saber que *está* a chover, não teria perguntado.
(20b) Eles não devem querer saber que ele *foi* um grande dançarino.

Assim, o facto de na proposição completiva das frases em (17) ocorrer o Indicativo e em (18) ocorrer o Conjuntivo permanece inexplicável.

Uma hipótese para explicar esses dados é a de que, nessas construções, o predicado matriz é interpretado como um predicado complexo, sendo a sequência *não acreditar* interpretada como uma expressão equivalente a *surpreender* e a sequência *não querer saber* como uma expressão equivalente a *ser indiferente*. Ou seja, nos exemplos de (19) e (20), as frases têm uma interpretação composicional, sendo o modo da oração completiva o que decorre da conjugação dos fatores [epistémico] e [verídico], enquanto em (17) e (18) as sequências *não acreditar* e *não querer saber* são interpretadas como um predicado complexo, cujo significado não resulta da combinação do significado das palavras que o compõem.

[1] Pelo menos no caso de (18a), é possível também a interpretação de que a proposição completiva não descreve um facto, mas uma hipótese. Isto é, a frase tem uma leitura em que é equivalente a *não quero saber se estiver / está (ou não) a chover*.

Em favor desta hipótese aponta o facto de em construções com a sequência *não acreditar* o Indicativo só ser possível se a frase expressar surpresa (e, eventualmente, desilusão ou contentamento) e de construções com a sequência *não querer saber* indicarem que a entidade identificada pelo sujeito da frase matriz tem conhecimento do facto descrito pela oração completiva, mas não lhe dar importância, tendo a sequência a mesma interpretação que o verbo *ignorar* em frases como as de (21):

(21a) Ele foi avisado, mas *ignorou* o aviso.
(21b) Decidiu *ignorar* que as crianças estavam cansadas e continuo a caminhada.

Aliás, veja-se que o verbo *ignorar* pode ter quer a interpretação que tem em (21) quer a que tem em (22), caso em que é equivalente a *desconhecer*, do mesmo modo que *não querer saber* pode ter a mesma duplicidade de interpretação, como mostram os exemplos em (18) e em (20):

(22) Ele *ignorava* que a lei tinha sido revogada.

No entanto, esta hipótese de que, nalgumas frases, *não acreditar* e *não querer saber* são interpretados como predicados complexos não explica o modo na oração completiva dessas construções. De facto, a respeito de *não acreditar*, se o seu significado é o mesmo que o de *surpreender* (ou de outro predicado que expresse surpresa, como *admirar* ou *espantar*), seria de esperar que na oração completiva ocorresse o Conjuntivo e não o Indicativo, tal como se verifica com estes predicados:

(23) surpreende-me que *esteja* / **está* a chover

Analogamente, se *não querer saber* tem o mesmo significado que *ser indiferente*, coloca-se a questão de que, com este predicado, se tem o Conjuntivo independentemente de a frase matriz ser afirmativa ou negativa, mas não no caso de (*não*) *querer saber*, em que o Conjuntivo na oração completiva só é possível se a frase matriz for negativa:

(24) (Não) é indiferente que *esteja* a chover.

(25a) A funcionária continua a falar ao telemóvel, não quer saber que a loja **está* / *esteja* cheia de clientes!
(25b) Achas que eles quereriam saber que já *chegámos* / **tenhamos chegado*?

Descartando a hipótese de as construções em causa envolverem um predicado complexo, resta assumir que têm uma análise composicional e os predicados que nelas ocorrem têm a mesma interpretação que em qualquer outra construção. Como se mostrará de seguida, esta assunção permite chegar à conclusão de que o modo da oração completiva, também nessas construções, é o esperado, dado o significado da construção, sendo apenas aparente a excecionalidade do modo na oração completiva destas construções.

4. Semântica formal e modo verbal

Acima as condições para ser usado o modo Indicativo ou o modo Conjuntivo em português foram descritas como decorrendo da conjugação dos valores [epistémico] e [verídico]. As condições que regulam a distribuição destes dois modos em português podem ser descritas com mais rigor, no quadro da semântica formal. Para se perceber melhor as condições que determinam o uso do Indicativo ou do Conjuntivo, é útil colocar a questão "O que assinala o modo (verbal)?". Ou seja, se, por exemplo, o morfema de número assinala pluralidade, e os morfemas de género assinalam a oposição masculino/feminino, o que assinalam os morfemas de modo Indicativo e de modo Conjuntivo?

Vários autores, como Godard (2012) ou Giannakidou e Mari (2021), defendem que o Conjuntivo ocorre numa proposição *p* se o contexto em que *p* ocorre leve a que se considerem tanto mundos possíveis em que a proposição é verdadeira (mundos-*p*) como mundos possíveis em que é falsa (mundos não-*p*), enquanto o Indicativo ocorre nos contextos em que o significado leva a que se considerem apenas mundos possíveis em que a frase é verdadeira ou então apenas mundos possíveis em que é falsa. Mais detalhadamente, sendo um modelo relativamente ao qual uma proposição é avaliada concebido como um conjunto de mundos possíveis, o verbo de uma proposição *p* flexionará no Indicativo se o modelo (ou Base Modal) relativamente ao qual a proposição é avaliada for homogéneo, contendo apenas mundos-*p* ou apenas mundos não-*p*, e flexionará no Conjuntivo se for avaliada relativamente a um modelo heterogéneo, que contém mundos-*p* e mundos não-*p*. Por outras palavras, para estas autoras, o Conjuntivo assinala a presença de uma Base Modal heterogénea e o Indicativo a presença de uma Base Modal homogénea. Para ilustração, considerem-se as seguintes frases:

(26a) *Está* alguém em casa.
(26b) Não *está* a chover.

(27) Talvez amanhã *chova*.

Tratando-se de frases declarativas não subordinadas, a enunciação de qualquer destas frases corresponde à realização de um ato de fala assertivo. Ou seja, ao enunciá-las, o enunciador expressa a sua crença de que as mesmas são verdadeiras, pelo que as frases são interpretadas relativamente a um modelo que representa as crenças do enunciador, o seu estado epistémico. Este modelo é um conjunto de mundos possíveis: o conjunto de mundos possíveis que corresponde à interseção de todas as proposições que o enunciador tem como verdadeiras (uma proposição denota um conjunto de mundos possíveis – o conjunto de mundos possíveis em que a mesma é verdadeira). Alternativamente, pode considerar-se que o modelo relativamente ao qual as frases são interpretadas modela a informação partilhada pelos participantes na interação comunicativa. Ou seja, pode igualmente assumir-se que as frases são interpretadas relativamente a um *context set*, que inclui o *common ground* – o conjunto de proposições (tacitamente) aceites como verdadeiras pelos participantes na interação comunicativa para efeitos dessa interação (Stalnaker, 1979) – e as possibilidades compatíveis com o mesmo (ver, e.g., Portner, 2009). Também neste caso o modelo relativamente ao qual a frase é interpretada é um conjunto de mundos possíveis: o conjunto de mundos possíveis em que são verdadeiras todas as proposições que pertencem ao *common ground*. O efeito da enunciação de uma frase declarativa relativamente a um contexto *c* é, se a frase for aceite como verdadeira pelos interlocutores, um novo contexto, *c'*, que contém apenas mundos possíveis em que a frase é verdadeira (i.e., $c + p = c \cap p$). Por simplificação, e porque a questão é irrelevante para os propósitos deste texto, assumirei que o modelo relativamente ao qual as frases em (26) e (27) são interpretadas é o modelo epistémico do enunciador, embora, pelo menos na questão de que se ocupa este texto, nada de fundamental mude se for assumido que as frases são avaliadas relativamente a um modelo que representa o contexto conversacional e não simplesmente as crenças do enunciador.

Posto isto, ao enunciar (26a), o enunciador expressa a sua convicção de que a frase enunciada é verdadeira; ou seja, indica que o seu estado epistémico contém apenas mundos possíveis em que (26a) é uma frase verdadeira, e, ao enunciar (26b) indica que o seu estado epistémico contém apenas mundos possíveis em que a frase *está a chover* é falsa. Em qualquer dos casos, a frase é interpretada relativamente a um modelo (ou Base Modal) homogéneo, que contém ou apenas mundos-*p* ou apenas mundos não-*p*, o que, de acordo com Giannakidou e Mari (2021), e.o., leva a que numa e noutra frase ocorra o Indicativo, sendo o Conjuntivo bloqueado. Pelo contrário, em (27), a enunciação da frase veicula a informação de que o enunciador aceita a possibilidade de chover no dia seguinte e a de não chover; ou seja, o modelo que representa as crenças do enunciador, relativamente ao qual a frase é avaliada, contém mundos-*p*, mundos em que chove no dia seguinte, e mundos não-*p*. Tratando-se de um modelo heterogéneo (i.e., que contém mundos-*p* e mundos não-*p*), é o Conjuntivo o modo que ocorre, sendo o indicativo bloqueado.

Em síntese, de acordo com estas autoras, o Indicativo e o Conjuntivo assinalam que a proposição é interpretada relativamente a uma Base Modal homogénea e relativamente a uma Base Modal heterogénea, respetivamente.

Na minha opinião, a proposta de que o Indicativo assinala que a Base Modal é homogénea e o Conjuntivo assinala que é heterogénea é questionável. De facto, em frases negativas como as que se seguem, ocorre o Conjuntivo, estando bloqueado o Indicativo:

(28a) O mau tempo impediu que o avião **descolou / descolasse*.
(28b) Saiu sem que se **deu / desse* por isso.

Nestes casos, são considerados apenas mundos possíveis em que a oração subordinada é falsa. Ou seja, o modelo relativamente ao qual esta oração é avaliada é uma Base Modal homogénea, formada por mundos não-*p*, mas o modo da oração subordinada é o Conjuntivo.

A minha proposta é antes a de que o Conjuntivo indica que são considerados mundos possíveis em que a frase é falsa, independentemente de a Base Modal conter também mundos possíveis em que a frase é verdadeira ou não conter, e o Indicativo indica que são considerados apenas mundos possíveis em que a frase é verdadeira. Por outras palavras, o Indicativo assinala que são considerados apenas mundos-*p* (mundos possíveis em que a proposição é verdadeira) e o Conjuntivo indica que são considerados mundos não-*p*. Em (26b) – *não está a chover* –, a frase, negativa, *não está a chover* é dada como verdadeira. Ou seja, todos os mundos considerados são mundos em que esta frase, negativa, é verdadeira. Por contraste, nos exemplos de (28), o operador de negação é exterior à oração com Conjuntivo e esta oração é apresentada como falsa.

Dito de outro modo, em (28) são considerados mundos possíveis em que a frase completiva é falsa, pelo que o verbo desta frase flexiona no Conjuntivo; em (26b), são considerados apenas mundos possíveis em que a frase – *não está a chover* – é verdadeira, pelo que o verbo desta frase flexiona no Indicativo.

Em suma, a proposta que defendo é a de que os modos Indicativo e Conjuntivo assinalam, respetivamente, que o modelo relativamente ao qual a frase é avaliada contém apenas mundos possíveis em que a frase é verdadeira e que o modelo contém mundos possíveis em que a frase é falsa. Por outras palavras, e informalmente, o Indicativo assinala que se considera apenas a possibilidade de a frase ser verdadeira e o Conjuntivo assinala que se considera a possibilidade de a frase ser falsa.

Esta proposta explica naturalmente os casos de Conjuntivo em frases negativas como as de (28) bem como noutras frases não verídicas, como, e.g., orações completivas de verbos desiderativos, como *querer* ou *preferir*, de verbos diretivos, como *pedir* ou *mandar*, ou de operadores modais como *talvez*. Com qualquer destes operadores a proposição sob o seu escopo é interpretada relativamente a um modelo que contém mundos possíveis em que é falsa (i.e., o significado de qualquer destes operadores leva a que se considere a possibilidade de não se verificar a situação descrita pela frase que introduzem). É mais questionável que a proposta seja sustentável quando se consideram casos de Conjuntivo em frases que descrevem factos, como orações completivas de predicados factivos não epistémicos. No entanto, também nestes casos o significado da construção leva a que se considerem mundos possíveis em que a proposição é falsa, o que explica que o verbo da mesma flexione no Conjuntivo.

Como defendido em Marques (2022), em português podem observar-se quatro tipos de predicados factivos, tendo por base a estrutura argumental associada a cada predicado e o modo que regem, decorrendo este modo do significado do predicado, em concordância com a proposta de que o Indicativo assinala a consideração de apenas mundos-*p* e o Conjuntivo assinala a consideração de mundos não-*p* (cf. Tabela 2).

Tabela 2. Classes de predicados factivos em português

A	B	C	D
A oração subordinada é argumento interno		A oração subordinada é argumento externo	
saber, *descobrir*, *verificar*, …	*lamentar*, *gostar*, …	*{ser / achar} (a)normal / (in)justo / curioso / intrigante / …*	*surpreender, irritar, alegrar, …*
Regentes de Indicativo		Regentes de Conjuntivo	

Os predicados das classes A e B são verbos de atitude proposicional, expressam uma relação entre a entidade identificada pelo seu argumento externo e a proposição completiva. De entre estes, os predicados da classe A expressam uma atitude epistémica: indicam que a entidade identificada pelo argumento externo do predicado tem a proposição completiva como verdadeira. Por outras palavras, estes predicados indicam que o modelo epistémico do utente da atitude proposicional contém apenas mundos possíveis em que a proposição completiva é verdadeira. Por conseguinte, dado que o Indicativo assinala que são considerados apenas mundos-*p*, estes predicados regem o Indicativo. Quanto aos predicados da classe B, expressam uma atitude avaliativa e o seu significado envolve raciocínio contrafactual, como proposto inicialmente por Heim (1992). Simplificadamente, a frase *o Pedro lamenta que não possa vir* indica que o Pedro preferiria poder vir (i.e., no mundo real, o Pedro não pode vir e, para o Pedro, um mundo em que tudo fosse igual ao mundo real, exceto que ele podia vir, seria preferível) e a frase *gostei que o Pedro tivesse dito o que disse* indica que, para o enunciador, o facto de o Pedro ter dito o que disse é preferível a não o ter feito. Dado que o significado destes predicados envolve considerar mundos possíveis em que a proposição completiva é falsa, os mesmos regem o Conjuntivo.

Quanto aos predicados, adjetivais, da classe C, estes expressam uma classificação da situação descrita pela proposição subordinada. Classificar uma situação como, e.g., justa ou injusta implica compará-la com outras situações e considerar uma escala de justeza de situações (ver também Giannakidou & Mari, 2021; Villalta, 2008). Ou seja, se, e.g., numa classificação de situações em função de padrões éticos, uma dada situação é classificada como justa, a sua contraditória será injusta, e vice-versa. Assim, o significado destes predicados envolve a consideração de uma escala, denotada pelo adjetivo. Esta escala tem uma linha divisória, que separa as partes positiva e negativa (ver, e.g., Kennedy, 2007), sendo a situação identificada pela proposição completiva classificada como estando associada a uma destas partes, por comparação com a sua contraditória, que estará associada à parte inversa da escala. Por exemplo, classificar uma situação como rara implica compará-la com a situação inversa, que será habitual. Esta dimensão escalar dos predicados da classe C explica o facto de serem regentes de Conjuntivo. De uma forma simples, dizer, por exemplo, que *é estranho que esteja a chover* significa o mesmo que *seria normal que não estivesse a chover*. Ou seja, o significado da construção envolve considerar a proposição contraditória à proposição encaixada. Assim, como o significado do predicado envolve considerar mundos possíveis em que a frase encaixada é falsa, o verbo desta flexiona no Conjuntivo.

Por último, quanto aos predicados da Classe D, foi defendido em Marques (2022) que os mesmos expressam uma relação de causalidade entre os seus argumentos, o que justifica que sejam regentes de Conjuntivo. Simplificadamente, uma frase como *surpreende-me que esteja a chover* significa que o facto de estar a chover causa um estado de surpresa no enunciador, *irritou-o que o estivessem sempre a interromper* significa que o facto de o interromperem com frequência lhe causou um estado de irritação, etc. Como defendido por vários autores desde Lewis (1973), a causalidade envolve raciocínio contrafactual. Isto é, dizer que A foi a causa de B significa que se A não tivesse ocorrido, sendo tudo o resto igual, B também não teria ocorrido. Assim, explica-se porque é que tanto os predicados da classe D como outros predicados que expressam uma relação causal ou próxima (como *ser preciso* ou *bastar*, predicados que expressam, respetivamente, uma relação de condição necessária e de condição suficiente) são regentes de Conjuntivo.

Em suma, em português o Indicativo assinala que são considerados apenas mundos possíveis em que a frase é verdadeira e o Conjuntivo assinala que são considerados mundos possíveis em que a frase é falsa. Isto não implica que uma frase com Conjuntivo seja tida como falsa no mundo real ou que uma frase com Indicativo

seja tida como verdadeira na realidade. O que cada um destes modos verbais assinala é que o significado da construção leva a que se considerem mundos não-*p*, no caso do Conjuntivo, ou apenas mundos-*p*, no caso do Indicativo.

Posto isto, retomem-se agora as construções problemáticas, que descrevem factos mas em que o modo verbal não é o que se esperaria.

5. **Recusa de aceitação de factos e modo verbal**

Comecemos por considerar a construção com o verbo *acreditar* sob escopo da negação e Indicativo na oração completiva. Como se viu acima, em casos como (29a) a construção indica que para o enunciador a oração completiva é verdadeira, informação que não é veiculada por (29b), com Conjuntivo na oração subordinada:

(29a) A Ana não acredita que *está* a chover.
(29b) A Ana não acredita que *esteja* a chover.

Por outras palavras, ao usar o Indicativo, o enunciador expressa a sua própria crença de que a proposição completiva é verdadeira. A frase expressa, portanto, um contraste entre as crenças da Ana e as do enunciador, enquanto que a frase (29b) descreve apenas uma crença da Ana. Como observado, entre outros, por Quer (1998), em (29a) a proposição completiva é interpretada relativamente a um modelo que representa as crenças do enunciador e em (29b) é interpretada relativamente a um modelo que representa as crenças da Ana. Como para o enunciador a frase é verdadeira (i.e., o seu estado epistémico contém apenas mundos possíveis em que a frase é verdadeira), explica-se o Indicativo em (29a). O Conjuntivo em (29b) decorre igualmente da proposta de que o Conjuntivo assinala a consideração de mundos não-*p*: a frase indica que a Ana não tem a oração completiva como verdadeira; ou seja, indica que o seu modelo epistémico inclui mundos possíveis em que não está a chover.

O problema, como se viu acima, reside em explicar o Indicativo nos casos em que o verbo da frase matriz flexiona na 1.ª pessoa do singular, já que o enunciador e a entidade identificada pelo sujeito da frase matriz são a mesma pessoa, pelo que, aparentemente, a frase expressa a contradição de que a mesma pessoa acredita que a oração completiva é falsa e acredita que é verdadeira (i.e., a frase parece dar a indicação contraditória de que o estado epistémico do enunciador tem mundos não-*p* e não tem mundos não-*p*, só mundos-*p*).

De facto, é precisamente uma contradição que a frase «*x* não acredita que *p*» expressa: uma contradição entre as crenças de *x* e a proposição *p.* Há duas possibilidades de resolver esta contradição. Uma possibilidade é *x* rejeitar a proposição *p*, por ser incompatível com o seu sistema de crenças. A outra possibilidade é *x* rever o seu sistema de crenças de forma a acomodar a aceitação da proposição *p* como verdadeira. Creio que são precisamente estas duas possibilidades que são expressas pelas frases (30a) e (30b):

(30a) Não acredito que a loja *esteja* fechada.
(30b) Não acredito que a loja *está* fechada!

Ambas estas frases expressam uma incompatibilidade entre o sistema de crenças do enunciador e a proposição completiva. A frase (30a) indica que o sistema de crenças do enunciador contém mundos possíveis em que a proposição completiva é falsa, pelo que o verbo desta proposição flexiona no Conjuntivo. Ou seja, (30a) expressa a informação de que a contradição entre o sistema de crenças do enunciador e a proposição *p* é resolvido pela recusa de *p*. Quanto à frase (30b), expressa a informação de que a contradição entre o sistema de crenças do enunciador e a proposição *p* é resolvida pela revisão do sistema de crenças de forma a acomodar *p*. Isto é, a frase indica que o enunciador reconhece que a oração completiva é verdadeira, pelo que indica que estão acessíveis apenas mundos possíveis em que essa frase é verdadeira (i.e., que descarta a possibilidade de a proposição completiva ser falsa), o que explica que o verbo flexione no Indicativo; ao mesmo tempo, o enunciador indica que ainda não procedeu à revisão do seu sistema de crenças de forma a acomodar a verdade da

proposição. Dito de outro modo, a frase (30b) corresponde ao caso em que a contradição entre um sistema de crenças e uma proposição é resolvida pela revisão do sistema de crenças e a frase descreve o estado epistémico que precede essa revisão.[2] Ou seja, a frase indica que o falante já reconheceu que a proposição completiva é verdadeira, pelo que tem de rever o seu sistema de crenças, mas ainda não procedeu a essa revisão para acomodar a verdade da proposição. Veja-se que no exemplo (31), do *corpus* CETEMPÚBLICO, é bastante sugestiva a presença de *ainda*, cuja presença indica que o estado de não crença é temporário. Este exemplo mostra mais claramente que nestas construções, com Indicativo na oração completiva, está em causa a necessidade de rever o sistema de crenças do falante:

(31) «Ainda não acredito que *venci* depois de ter perdido tantas corridas em cima da meta.»
(CETEMPÚBLICO, *par=ext84460-des-91a-1*)

Considere-se agora a outra construção problemática, com Conjuntivo na oração completiva de *saber*, como (32):

(32) Eu vou sair; não quero saber que *esteja* a chover!

Creio que também frases como esta têm uma interpretação literal e o facto de ser o Conjuntivo o modo que ocorre na oração completiva decorre da proposta de que este modo assinala a consideração de mundos não-*p*. Interpretada literalmente, esta construção expressa a rejeição de um estado epistémico que só contém mundos em que a proposição completiva é verdadeira. Isto é, '*não querer saber que* p' significa '*querer não saber que* p'. A proposição 'x *saber que* p' indica que o estado epistémico de *x* contém apenas mundos-*p*. Ao negar-se esta frase – obtendo-se 'x *não saber que* p' – nega-se que o estado epistémico de *x* contenha apenas mundos-*p*¸ ou, equivalentemente, indica-se que contém mundos não-*p*. Dado serem considerados mundos não-*p*, é o Conjuntivo o modo selecionado. Por fim, o verbo *querer* indica que é esse o estado epistémico que o enunciador deseja. Ou seja, interpretada literalmente, a frase (32) expressa o desejo do enunciador de ter um estado epistémico que contém mundos possíveis em que não chove. A presença do Conjuntivo na oração completiva de *saber*, um verbo que, em princípio, só deveria aceitar o Indicativo, decorre assim, naturalmente, do significado composicional da construção e é coerente com a ideia de que o Conjuntivo é uma marca que assinala a consideração de mundos não-*p* e o Indicativo assinala a consideração de apenas mundos-*p*.

Como observado acima, o verbo *ignorar* pode também significar o mesmo que *não saber*, em casos como *eles ignoram que está a chover*, equivalente a *eles não sabem que está a chover*, ou ter a mesma interpretação que *não querer saber*. Num exemplo como *ele foi avisado, mas ignorou os avisos*, a interpretação é basicamente a mesma que a de *ele agiu como se não tivesse sido avisado*. Do mesmo modo, a frase (32) indica que o enunciador pode saber que está a chover, mas opta por ignorar esse facto e agir como agiria se não soubesse. Por outras palavras, nessa frase *não quero saber* tem a mesma interpretação que tem o verbo *ignorar* em casos como *ele ignorou os avisos*. Curiosamente, com o verbo *ignorar* também é possível ocorrer quer o Indicativo quer o Conjuntivo na oração completiva, como mostram os seguintes exemplos:

(33a) «Os antigos romanos *ignoravam* que o Vesúvio *era* um vulcão – nem sequer tinha cratera.»
(CETEMPÚBLICO, *par=ext440425-nd-91a-2*)
(33b) «*Ignoro* que *tenha*, alguma vez, assinado disco seu, em seu nome, de corpo inteiro.»
(CETEMPÚBLICO, *par=ext1164810-clt-95b-2*)

[2] Como observado por um(a) revisor(a), há um paralelismo entre o significado da construção e o de frases como *sei que ele morreu, mas ainda não acredito.*

(33c) «Nem se pode *ignorar* que ministros do seu Governo *tenham assinado* despachos para que esta lei não se aplique aos seus ministérios, como se não fizessem parte do Conselho de Ministros que a aprovou.» (CETEMPÚBLICO, *par=ext53027-opi-98a-2*)

A explicação para o modo na oração completiva nestes casos é a mesma que para as construções *não acreditar* e *não querer saber*, analisadas acima. Nos dois primeiros excertos, i.e., em (33a) e (33b), o verbo *ignorar* significa o mesmo que *não saber* ou *não ter conhecimento*. No primeiro exemplo, ocorre o Indicativo porque o falante apresenta a oração completiva como descrevendo um facto, do mesmo modo que em frases como *eles não acreditam que está a chover* ou *eles não sabem que está a chover*. Já no segundo exemplo, é o Conjuntivo o modo que ocorre na oração completiva porque essa oração não é dada como descrevendo um facto, do mesmo modo que em frases como *eles não acreditam que esteja a chover* ou *não tenho conhecimento de que esteja a chover*. Ou seja, nestes casos ocorre o Conjuntivo porque se descreve um estado de crenças que contém mundos possíveis em que a frase é falsa. Em ambos os exemplos, o verbo *ignorar* significa o mesmo que *ter conhecimento* ou *saber*. Finalmente no último exemplo, *ignorar* tem a mesma interpretação que *não querer saber* e o modo que ocorre na oração completiva é também o Conjuntivo. Neste exemplo, a frase é equivalente a *não se pode não querer saber que*; ou seja, pode ser parafraseada por *não se pode fazer de conta que se tem um estado de crenças que inclui mundos possíveis em que a oração completiva é falsa*. Daí o Conjuntivo na oração completiva. O significado da frase envolve a consideração de mundos possíveis em que a oração completiva é falsa.

6. **Conclusão**

Em português, o modo verbal em frases que descrevem factos da realidade não é sempre o mesmo. Nalgumas construções, tanto o Indicativo como o Conjuntivo podem ocorrer, sendo que a primeira opção indica que a proposição descreve um facto e a segunda não o indica. É o caso de construções como *ele não acredita que {está / esteja} a chover* ou de *não tive conhecimento de que {estava / estivesse} a chover*. Noutras construções, só um dos dois modos pode ocorrer, consoante o predicado, factivo, que introduz a proposição seja regente de um ou de outro modo. É o que se verifica, por exemplo, em construções como *ele sabe que {está / *esteja} a chover* e *é surpreendente que {*está / esteja} a chover*. Estes casos são explicáveis pela observação de que, em português, o Indicativo ocorre em contextos simultaneamente verídicos e epistémicos, enquanto o Conjuntivo é o modo complementar, que ocorre nos outros contextos, não assinalando nenhum valor semântico particular. No entanto, esta proposta não explica o modo em orações completivas de dois tipos de construção: (i) frases como *não acredito que está a chover!*, com o verbo *acreditar* flexionado na 1.ª pessoa do singular, sob o escopo da negação e com Indicativo na oração subordinada (uma construção que é possível apenas em frases exclamativas, que expressam surpresa pelo facto descrito pela oração completiva), (ii) frases como *não quero saber que esteja a chover!*, com Conjuntivo na oração completiva de *saber*, um verbo factivo regente de Indicativo.

A proposta de que em português as desinências de modo são marcas que assinalam se o significado da construção envolve a consideração de apenas mundos possíveis em que a frase é verdadeira, no caso do Indicativo, ou de mundos possíveis em que a frase é falsa, no caso do Conjuntivo, permite explicar o modo verbal que ocorre em todos os tipos de construção analisados, sem assumir acerca de alguma das construções que é um caso excecional em que o modo que é usado não decorre das condições que, em português, regulam o uso de Indicativo ou Conjuntivo. Em todos os tipos de construção considerados, a ocorrência dos modos Indicativo ou Conjuntivo decorre naturalmente da proposta de que o verbo flexiona no primeiro destes modos quando o significado da construção leva a que sejam considerados apenas mundos possíveis em que é verdadeira a frase a que esse verbo pertence e flexiona no Conjuntivo quando o significado da construção leva a que sejam considerados mundos possíveis em que a frase a que o verbo pertence é falsa.

Referências

Giannakidou, Anastasia (1994) The semantic licencing of NPIs and the Modern Greek subjunctive. In Ale de Boer, Helen de Hoop & Henriette de Swart (orgs.), *Language and Cognition 4, yearbook of the Research Group for Theoretical and Experimental Linguistics.* University of Groningen, pp. 55–68.

Giannakidou, Anastasia & Alda Mari (2021) *Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes*. Chicago University Press.

Godard, Danièle (2012) Indicative and subjunctive mood in complement clauses: From formal semantics to grammar writing. In Christopher Piñón (org.), *Empirical Issues in Syntax and Semantics 9*, pp. 129–148. Disponível em http://www.cssp.cnrs.fr/eiss9/

Heim, Irene (1992) Presupposition projection and the semantics of attitude verbs. *Journal of Semantics* 9 (3), pp. 183–221. https://doi.org/10.1093/jos/9.3.183

Lewis, David (1973) Causation. *The Journal of Philosophy* 70 (17), pp. 556–567.

Kennedy, Christopher (2007). The grammar of vagueness. *Linguistics and Philosophy* 30, pp. 1–45.

Marques, Rui (1995), *Sobre o valor dos modos conjuntivo e indicativo em português*. Dissertação de mestrado, Faculdade de Letras da Universidade de Lisboa.

Marques, Rui (2022, 21–23 abril) *Explaining the subjunctive in factive contexts* [Apresentação de comunicação]. LSRL52, University of Winsconsin-Madison, Madison, EUA.

Portner, Paul (2009). *Modality*. Oxford University Press.

Quer, Josep (1998). *Mood at the interface*. LOT.

Stalnaker, Robert (1979) Assertion. In Peter Cole (org.), *Syntax and Semantics 9*. New York Academic Press, pp. 315–332.

Villalta, Elisabeth (2008) Mood and gradability: An investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31 (4), pp. 467–522. https://doi.org/10.1007/s10988-008-9046-x

Marcadores discursivos com funções modais: Uma análise contrastiva de *claro* e seus equivalentes funcionais em francês

Amália Mendes[1], *Pierre Lejeune*[1]

[1]Faculdade de Letras da Universidade de Lisboa, Centro de Linguística da Universidade de Lisboa

Resumo

As funções modais desempenhadas por alguns marcadores discursivos têm sido alvo de análises para várias línguas. Este artigo analisa os diferentes valores veiculados pelo marcador discursivo *claro*, tomando como ponto de partida uma perspetiva contrastiva com o francês, com base em corpora escritos e orais monolíngues e paralelos, de forma a identificar marcadores discursivos que podem ser equivalentes funcionais de *claro* em diferentes contextos. Os contextos de *claro* e dos seus equivalentes em francês nos corpora permitem identificar um valor de concordância e conhecimento partilhado com o interlocutor e de eliminação da alteridade (*bien sûr*), de factualidade (*de fait*, *c'est un fait*), de oposição e até ironia (*bien entendu*) e a inserção em estruturas argumentativas precedendo segmentos contrastivos ou concessivos (*certes*). Mesmo nos contextos em que *claro* não responde ou comenta o discurso de um interlocutor direto, estabelece um relação entre o discurso do locutor e um outro discurso, quer o discurso de uma comunidade epistémica em que o locutor se insere, quer um discurso implícito de um alocutor real ou virtual.

Palavras-chave: marcadores discursivos, polifuncionalidade, modalidade, análise contrastiva, corpora.

Abstract

The modal functions performed by some discourse markers have been the subject of analysis for several languages. This paper analyzes the different values conveyed by the Portuguese discourse marker *claro*, and takes a contrastive perspective with French, based on monolingual and parallel, written and spoken, corpora, in order to identify discourse markers that are functional equivalents of *claro* in different contexts. Contexts of *claro* and its French equivalents in corpora show a meaning of agreement and shared knowledge, as well as the elimination of alterity (*bien sûr*), factuality (*de fait*, *c'est un fait*), opposition and even irony (*bien entendu*) and the insertion in argumentative structures preceding contrastive or concessive segments (*certes*). Even in contexts in which *claro* does not responds to or comments on the discourse of a direct interlocutor, it establishes a relationship between the speaker's discourse and another discourse, either the discourse of an epistemic community that includes the speaker or the implicit discourse of a real or virtual speaker.

Keywords: discourse markers, polyfunctionality, modality, contrastive analysis, corpora.

1. Introdução

Os marcadores discursivos constituem uma categoria cujas unidades podem assumir várias funções na estruturação semântica e pragmática do discurso, nos níveis proposicional, estrutural e modal/interpessoal (entre outros, Cuenca & Marín, 2009; Halliday & Matthiessen, 2004; Pons Bordería, 1998). Os marcadores discursivos podem desempenhar funções ao nível proposicional, ligando entre si duas proposições e especificando a natureza semântica dessa ligação, como por exemplo, um valor causal, contrastivo ou condicional, sendo nesse caso caracterizados como conetores discursivos. Por outro lado, podem desempenhar funções ao nível da organização textual, sinalizando, por exemplo, uma relação elaborativa ou resumativa, ou uma mudança de

tópico, entre dois segmentos textuais. Finalmente, os marcadores discursivos podem ter funções pragmáticas, associadas a valores interpessoais e valores modais, sendo por vezes designados como marcadores pragmáticos quando desempenham essas funções que não requerem a conexão entre dois segmentos. Não é possível distinguir subcategorias precisas dentro da categoria dos marcadores discursivos, uma vez que um mesmo marcador pode desempenhar várias funções, em contextos diferenciados (um exemplo dessa polifuncionalidade é o caso de *pois:* Costa, 2013; Lejeune & Mendes, 2020; Lima, 2002; Lopes, 2012). No caso da língua alemã, está ainda descrita a categoria das partículas modais, que veícula valores modais e apresenta propriedades morfossintáticas distintivas (Diewald, 2013; Waltereit, 2001). A relação entre marcadores discursivos e partículas modais, no alemão e também nas línguas românicas, tem sido objeto de discussão (Degand et al., 2013). A questão é pertinente para a nossa análise de *claro* e iremos discutir na próxima secção análises que por vezes refletem sobre a categorização de equivalentes funcionais de *claro* enquanto marcador discursivo e/ou partícula modal.

O nosso objetivo é refletir sobre o comportamento de *claro*, que apresenta funções discursivas e, também, pelo significado do adjetivo que lhe dá origem, valores modais. A análise terá em consideração possíveis equivalentes funcionais de *claro* em francês. Uma abordagem contrastiva apresenta, na nossa opinião, várias vantagens: por um lado, o resultado pode contribuir para estudos de tradução, ao integrar soluções para unidades frequentemente polissémicas e polifuncionais (Crible & Degand, 2019), e, por outro lado, configura-se como uma metodologia eficaz na identificação dos valores das unidades e da sua identidade semântica (Franckel, 2005), ao revelar diferenças de significado que poderiam não ser evidentes numa análise não contrastiva. Para levar a cabo este objetivo, iremos recorrer a contextos de *claro* em dois corpora paralelos: o corpus Europarl (Koehn, 2005), que permite observar os contextos de *claro* e seus equivalentes em francês em textos de um género muito específico, as atas das sessões do Parlamento Europeu nas várias línguas do espaço comunitário; e o romance *Balada da Praia dos Cães* de Cardoso Pires e sua tradução para francês, escolhido por a obra ter várias sequências de diálogos entre as personagens (pela frequente representação de diálogos informais, esta obra de ficção deu já resultados interessantes na análise de marcadores discursivos (Duarte, 1989; Duarte, 2005; Lejeune & Mendes, no prelo). Consideramos os corpora paralelos como um recurso fundamental para estudos contrastivos, pois permitem identificar diferenças significativas de forma e sentido entre diferentes línguas (Mauranen, 1999, 2016; Noël, 2003), sendo que estes corpora "should be recognized as the normal part of a natural language that they are" (Mauranen, 1999. p. 161). Notamos aqui a dificuldade em determinar no corpus Europarl a língua original a partir do qual foram elaboradas as versões portuguesa e francesa, podendo nalguns casos a tradução ter sido elaborada a partir de uma língua pivô, i.e., o inglês. Independentemente da fonte de cada texto, consideramos que as opções de transposição são igualmente válidas para um trabalho contrastivo e contribuem para o objetivo final de identificar variações de significado no marcador discursivo *claro*. Para complementar a análise e confirmar valores de *claro* e seus equivalentes em francês, recorremos a dois corpora monolíngues, o *Corpus de Referência do Português Contemporâneo*, subcorpus escrito online (Généreux et al., 2012)[1], e subcorpus oral[2], e o Corpus de Français Parlé Parisien des années 2000 (CFPP2000)[3].

A análise de *claro* insere-se no trabalho em curso sobre marcadores discursivos que veículam valores modais, como o caso de *de facto*, que assinala, de forma geral, um comprometimento do locutor baseado num valor de factualidade. Pretendemos analisar a relação entre *claro* e os seus equivalentes funcionais em francês, como *bien sûr, evidemment, il est évident que, bien entendu, de fait, certainement, naturellement, il va de soi, certes*, (Amiot & Flaux, 2007; Anscombre, 2013; Dostie, 2001; Péroz, 1992), sendo muito limitados os contextos em que a forma etimológica mais próxima *clair* pode ocorrer como equivalente funcional de *claro*, mostrando processos de gramaticalização distintos nas duas línguas.

[1] gamma.clul.ul.pt/CQPweb/crpc
[2] teitok.clul.ul.pt/crpcoral
[3] http://cfpp2000.univ-paris3.fr/search-transcription/#10

2. Valores de *claro* e seus equivalentes funcionais descritos na literatura

A análise dos marcadores *claro*, em castelhano, e *(és) clar*, em catalão, que derivam de formas adjetivais, identifica a sua polifuncionalidade (Cuenca, 2012; Cuenca & Marín, 2012; Fuentes, 1993; Pons, 2003). Estes marcadores, que ocorrem tipicamente em contextos interacionais, são descritos como marcando valores de concordância em relação a outro turno de fala e um conhecimento partilhado. No caso do catalão, c*lar* é descrito como podendo apontar para um valor de contraste ou concessão e, na posição inicial de turno, como podendo ainda amenizar a tomada do turno de fala ao apresentá-la como alinhada com a intervenção precedente do interlocutor (Cuenca, 2013). Esta análise aponta para uma função modal de marcação de certeza e uma função estrutural de introdução ou reorientação de um turno de fala. A autora discute a categorização desta unidade, tendo em consideração as propriedades que caracterizam as partículas modais:

> It [(és) clar] exhibits the main features that Waltereit (2001) identifies as distinctive of MPs: it operates at a speech-act level modifying the speech-act conditions and has a non-modal counterpart – the adjective clar –, to which it is related metonymically. (…) However, it also introduces turns and utterances, as conjunctions, parenthetical connectives and pragmatic connectives do, and it is parenthetical and combines with conjunctions, just like parenthetical and pragmatic connectives. It can be thus considered a form occupying the fuzzy space between pragmatic connectives and modal particles (Cuenca, 2013, p. 204).

Várias análises consideram válido aplicar a categoria de partícula modal às línguas românicas (Coniglio, 2008; Favaro, 2021; Remberger, 2020). Com base nos trabalhos para o alemão, são identificadas algumas propriedades desta categoria: não terem flexão; terem um homónimo noutra classe morfossintática; não serem acentuadas; não poderem ocorrer em posição inicial de frase, nem poderem ser coordenadas ou interrogadas; estarem frequentemente associadas a um tipo de ato de fala; não poderem constituir uma resposta curta a perguntas; serem sintaticamente integradas (não serem um elemento parentético) (Coniglio, 2008; Diewald, 2013). Algumas destas propriedades são partilhadas pelos marcadores discursivos, como a ausência de flexão. Além disso, os marcadores discursivos e as partículas modais focam ambos valores ao nível das atitudes e dos atos de fala. No entanto, outras propriedades das partículas modais, como não ocorrerem em posição inicial de frase/enunciado ou estarem sintaticamente integradas, são distintivas (Diewald, 2013). Para o português, um pequeno conjunto de elementos, como *lá*, tem sido considerado como tendo usos que reúnem as propriedades desta categoria (Franco, 1990; Meisnitzer, 2012).

Embora não seja nosso objetivo analisar equivalentes funcionais do inglês, vale a pena referir o trabalho de Aijmer (2013) sobre *of course*, uma vez que a autora identifica um uso como marcador discursivo que tem um valor de contraste ou concessão**,** podendo ocorrer com conjunção adversativa e aditiva (*but of course, and of course*), e um valor de organização do discurso (no início de frase/turno de fala, seguido de vírgula). Para além desses usos, identifica contextos em que tem a função de partícula modal**,** assinalando informação de tipo *common ground***,** marcando consenso e consolidando uma relação harmoniosa com o ouvinte. No entanto, utiliza a categoria "partícula modal" com base em critérios funcionais, não aplicando vários critérios formais, como, por exemplo, a impossibilidade de ocorrência em posição inicial de frase/enunciado, que caracteriza as partículas modais.

O marcador *claro* do português europeu foi objeto de análise em Lopes (2021), que distingue usos que constituem ou não diálogos. Nos usos em diálogos, *claro* é descrito como configurando uma resposta afirmativa enfática a atos discursivos de pergunta ou pedido, em que o caráter marcado de *claro* se opõe ao uso de *sim*, tal como referido para o castelhano e para o catalão (Cuenca, 2013; Pons, 2012), por denotar que o conteúdo proposicional sobre o qual *claro* tem escopo é conhecimento partilhado; poderá ainda ter um valor de marcador de acordo, e um valor que sinaliza uma atitude de atenção cooperante na interação. Nos restantes usos, descrevem-se um valor de modalizador epistémico e valor concessivo.

Enquanto o português, o espanhol e o catalão têm um marcador discursivo com origem no adjetivo *claro* e seus equivalentes etimológicos, o mesmo não ocorre em francês. Em (1), *claro* e *clair* são adjetivos em

construção predicativa em português (1a) e em francês (1b). No entanto, nos exemplos (2) a (4), o equivalente possível para *claro* seria o advérbio *clairement* e não o adjetivo *clair* (embora outros marcadores discursivos fossem mais adequados do que *clairement* como se verá abaixo). Em situação de diálogo, como em (2), *clair* não pode constituir uma resposta a uma pergunta e, na avaliação de uma proposição, apenas o advérbio é aceitável, quer em posição inicial de frase (3), quer em posição pré-verbal (4). Os usos de *é claro que* admitem, pelo contrário, os equivalentes *c'est clair que* e *il est clair que* como equivalentes (5), não podendo ocorrer nestes casos o advérbio *clairement* seguido de *que*. É de notar que uma pesquisa no corpus Europarl de *il est clair que* mostra que um equivalente funcional frequente em português é *é evidente que* (6). Os dados em (1)-(6) mostram um percurso de gramaticalização diferente no caso do francês, em que o adjetivo *clair* ocorre nas construções *c'est clair*, *il est clair que*, mas não assumiu funções de marcador discursivo em contextos como (2)-(4).

(1a) Precisamos, naturalmente, de navios de casco duplo. Isto é claro, mas é algo que apenas vai ter efeitos a médio e longo prazo.
(1b) Naturellement, nous avons besoin de navires à double coque. C'est clair. Cependant, c'est une chose qui ne peut être atteinte qu'à moyen ou long terme. (Europarl 2305)
(2a) A: O João vai ganhar as eleições ?
B: Claro.
(2b) A: Jean va gagner les élections?
B: *Clair / Clairement

(3a) Claro, o João vai ganhar as eleições.
(3b) *Clair / Clairement, Jean va gagner les élections.

(4a) O João, claro, vai ganhar as eleições
(4b) Jean, *clair / clairement, va gagner les élections.

(5a) É claro que o João vai ganhar as eleições
(5b) C'est clair que / Il est clair que Jean va gagner...

(6a) é evidente que acompanharemos de muito perto a forma como a legislação da UE será aplicada
(6b) il est clair que nous examinerons minutieusement la manière dont la législation de la CE sera appliquée (Europarl 1498)

Quais são então as unidades que em francês podem funcionar como equivalentes de *claro*, uma vez que o candidato mais óbvio (*clair*) não assume a totalidade dessas funções? Com frequência, o equivalente é *bien sûr*, que literalmente corresponde a 'bem certo/seguro', com valor epistémico. A análise de Anscombre (2013) salienta que o uso de *bien sûr* é uma reação a um enunciado anterior, explícito em (7) ou virtual, como em (8) (a Lia não vai chegar). A proposição é encarada como um saber partilhado ou previsível, por ser por exemplo uma regra geral, num ponto de vista que tem como fonte uma comunidade indefinida a que pertence o locutor do enunciado, fonte que Anscombre (2013) designa como um *ON-locuteur*. Note-se que, em (7) e (8), é apresentado como saber partilhado o que não era óbvio para o interlocutor.

(7) – Je me demande s'il va faire beau.
– Bien sûr qu'il va faire beau! La météo l'a dit. (Anscombre, 2013, p. 81, ex. 37)
– 'Pergunto-me se vai estar bom tempo.
– Claro que vai ! Disse-o o Instituto de meteorologia' [nossa tradução]

(8) Arrête de tourner en rond! Bien sûr que Lia va venir! (Anscombre, 2013, p. 80, ex. 33)
'Para de andar às voltas! Claro que a Lia vai chegar!' [nossa tradução]

Outro equivalente funcional de *claro* é *évidemment* (Dostie, 2001), formado a partir do adjetivo *évident*. Este marcador não aceita ser modificador do predicado, à semelhança de *claro* e *clair* e de *evidentemente* (*explicou claro a situação / *Il a expliqué clair la situation / *explicou evidentemente a situação / * Il a expliqué évidemment la situation) e ao contrário de *clairement* (Il a expliqué clairement la situation). O uso de *évidemment* pode qualificar um enunciado do próprio locutor, contrariando um potencial juízo negativo ou contra-argumento do co-enunciador, ou pode qualificar uma afirmação do co-enunciador como sendo algo com que concorda ou eventualmente algo que considera banal. Tal como no caso de *bien sûr*, *évidemment* sinaliza uma conclusão previsível.

(9) A: Tu ne peux pas venir.
B: Évidemment. (Dostie 2001, p. 71, ex. 11)
'A: Não podes vir
B: Claro.' [nossa tradução]

No exemplo (9), a resposta com *évidemment* (que pode ter em português o equivalente funcional *claro*) indica que a informação é conhecida, ou que, embora não conhecida, não é surpreendente (Dostie, 2001). Como refere Franckel:

> *Évident* est formé sur le même étymon que *voir* : *evidere* (latin) signifie 'ce qui ressort de ce que tout sujet peut voir'. À la différence de *visible*, *évident* en tant que signifiant 'issu et sorti du visible' ne convoque pas un S [sujet] particulier: ce qui est évident s'impose, se passe de justification, n'a pas à être argumenté, raisonné, justifié, ni même asserté. (Franckel et al., 2017)[4]

Por sua vez, o possível equivalente funcional *naturellement* remete para um saber partilhado, que pode ser construído durante a interação, em que sobressai a interpretação do enunciado como decorrendo de uma convenção, eventualmente uma convenção natural e inevitável (Amiot & Flaux, 2007).

Quanto à *clairement*, apesar da sua proximidade etimológica com *claro* e de poder como ele assumir valores de marcador discursivo, é funcionalmente muito menos próximo de *claro* do que *bien sûr* e *évidemment*. Mesmo quando é aceitável o uso de *clairement* numa posição sintática igual à de *claro*, os dois marcadores têm geralmente valores semânticos diferentes. Contrariamente a *bien sûr, évidemment* e *claro*, encontramos *claramente* em posição predicativa como "advérbio de modo orientado para o sujeito": Paul a distingué *clairement* (**bien sûr* , **évidemment*) deux silhouettes / O Paulo distinguiu *claramente* (**claro*) duas silhuetas (Molinier, 1990. p. 40). Num artigo já antigo, Molinier (1990) afirma peremptoriamente que *clairement* não existe como marcador discursivo modal epistémico (na sua terminologia "advérbio disjuntivo de estilo"). Recusa por exemplo um enunciado como "**Clairement*, Paul a eu tort / *Claramente* (**Claro*) o Paulo enganou-se (no sentido de não há dúvida de que…"). Ora encontramos na base Frantext alguns (muito poucos, é verdade) exemplos que mostram que não é o caso, p. ex no século XIX, em Zola (La bête humaine) :

(10) Jacques restait les yeux largement ouverts sur lui; et, sous ce regard, où il lisait une surprise croissante, il s'agitait, comme pour échapper à sa propre ressemblance; tandis que sa femme, elle aussi, suivait, glacée, le travail sourd de mémoire, exprimé par le visage du jeune homme. Clairement, celui-ci s'était étonné d'abord de certaines analogies entre Roubaud et l'assassin

[4] Tradução nossa: "*Évident* é formado sobre o mesmo étimo de *voir* 'ver': *evidere* (latim) significa 'o que emerge daquilo que todo o sujeito pode ver'. Contrariamente a *visible* 'visível', *évident*, por significar 'oriundo e saindo do visível' não convoca um S particular: o que é evidente impõe-se, não precisa de justificação, não precisa de ser argumentado, fundamentado, justificado, nem mesmo asserido."

‘Jacques continuava com os olhos muito abertos na sua direção; e, por baixo desse olhar, em que lia uma surpresa crescente, agitava-se como se quisesse fugir da sua própria semelhança; enquanto a sua mulher seguia, congelada, o trabalho surdo de memória, exprimido pela cara do jovem. Claramente / ?Claro (que) este tinha estranhado primeiro algumas analogias entre o Roubaud e o assassino’ [nossa tradução]

Esse tipo de uso de *clairement* enquanto marcador epistémico, raro na altura do artigo do Molinier, passou nos últimos anos a proliferar. Em posição sintaticamente não integrada, encontramos atualmente exemplos como os seguintes (extraídos de fontes monolíngues, para além dos corpora indicados na Secção 1, e com traduções nossas):

(11) Clairement, on va avoir une pénurie et une flambée des prix de l'énergie. (France Inter, 2023)
‘Vamos ter de certeza uma penúria e uma subida em flecha dos preços da energia’

(12) On fait de notre mieux mais là clairement on est franchement à la bourre ! (Radio Crimi 2020)
‘Estamos a fazer todos os possíveis mas nesta altura está tudo claramente (? *claro*) fora de controlo’

(13) –C'est pas mal, ce qu'ils ont fait, nota la mère de Steph.
– Ouais. –Tout était gris dans cette ville. C'était moche.
– Clairement (Nicolas Mathieu, *Leurs enfants après eux*, 2018)
‘Não está nada mal, o que fizeram, reparou a mãe do Steph – Sim – Tudo era cinzento nesta cidade. Era feio – Sem dúvida’

(14) Arturo Vidal donne son avis sur le cas Ousmane Dembélé : « Ousmane est un footballeur avec un talent fou. Quand il arrivera à maturité, il sera un joueur important du Barça et de l'équipe de France. Mais, clairement, il a besoin de mûrir. (BFM TV, 2018)

‘Arturo Vidal dá a sua opinião sobre o caso Ousmane Dembélé: «Ousmane tem um talento incrível. Quando atingir a maturidade, será um jogador importante do Barça e da equipa francesa. Mas, sem dúvida, precisa de amadurecer.’

Nesses exemplos, *bien sûr* e *evidemment* na versão francesa, *claro* na versão portuguesa, não seriam impossíveis, mas dariam aos respetivos enunciados um sentido ligeiramente diferente. Enquanto *claro, évidemment, bien sûr* remetem para uma pré-construção (com saber partilhado – ou sem – o interlocutor dúvida de p ou acha que não p – validação) da relação predicativa, o valor epistémico de *clairement* é construído diretamente no enunciado (valor de certeza: seleção de um valor e eliminação dos valores complementares). No primeiro caso, temos um efeito de sentido de tipo “como toda a gente sabe” que não se tem no segundo.

A pré-construção no caso de *claro*, *évidemment* e *bien sûr*, e a ausência da mesma com *clairement*, é visível nas seguintes (in)compatibilidades, em enunciados em que nas duas línguas *que* é marcador de pré-construção:
Bien sûr / *Evidemment* / ? *Clairement* qu’il va le faire (*Claro* que o vai fazer).

Seria interessante confrontar esses empregos de *clairement* com os de *claramente*, com os quais existe um paralelismo evidente. Por exemplo, nos três exemplos extraídos da mesma entrevista do ciclista João Almeida, *claramente* podía cada vez ser traduzido em francês por *clairement*, sem alteração de sentido.

(15) No dia em que ganhei a etapa, claramente pensei que havia a possibilidade de trazer a [camisola] rosa, mas nas etapas seguintes, que eram muito duras e em que os adversários estavam muito fortes, percebi que seria muito difícil. […] Claramente ele [Tadej Pogacar] é um ciclista que está um nível acima de mim - ou dois, quem sabe. […] Quando fiz a última preparação na Serra

Nevada, de três semanas, estava com os meus companheiros [da UAE Emirates] e um dia decidimos deixar só o bigode. Dissemos 'é só até à corrida, depois tiramos'. Depois, não tirei, e ficou o bigode. Claramente que, se calhar, chego a casa e corto o bigode", concluiu, entre risos. (Público, 29/5/2023)

Na nossa análise, iremos observar as propriedades de *claro*, tendo em consideração as propostas de categorização e valores referidos nesta secção. Interessa-nos analisar os contextos de *claro* e observar se existem fronteiras claras entre as diferentes funções desempenhadas por esta unidade ou se um valor de base modal pode configurar um significado de base (*core meaning*) desta unidade, com variações de significado em contexto, à semelhança do que propomos para *de facto*, em que o valor geral de confirmação está presente nos diferentes usos do marcador discursivo.

3. **Os valores de *claro* em português europeu**

A análise dos contextos de *claro* em português europeu será dividida em dois grandes conjuntos: por um lado, contextos de diálogo, que envolvem dois locutores, em que *claro* constitui, ou é parte de, uma resposta de um locutor a um enunciado do seu interlocutor (Secção 3.1), por outro lado, contextos em que *claro* não configura uma resposta ou comentário ao discurso do interlocutor, mas comenta o próprio discurso, embora, como veremos, esteja sempre, explícito ou implícito, o interdiscurso (Secção 3.2). Os equivalentes funcionais de *claro* em francês nesses dois tipos de contextos serão integrados na discussão, com base nos contextos dos dois corpora paralelos que utilizamos, e de contextos dos corpora monolíngues, quando pertinente, com propostas de tradução nossas.

3.1. ***Claro* em contextos de diálogo**

Claro pode constituir resposta a uma pergunta do interlocutor ou qualificar uma afirmação do interlocutor. Na verdade, *claro* envolve sempre uma reação a um enunciado outro, quer seja o enunciado do interlocutor a quem responde, quer seja uma convocação de outra perspetiva que pode estar explícita ou implícita.

O marcador *claro* pode constituir uma resposta positiva ao enunciado do seu interlocutor, sendo frequentemente resposta a interrogativas-tag, como em (16). Nesses contextos, *claro* assinala a concordância, mas além disso indica que o locutor tinha já conhecimento do que o interlocutor afirmou, isto é, que o enunciado do interlocutor configura conhecimento partilhado. Nesse aspeto, *claro* difere de *sim*, que assinala concordância sem que a informação seja considerada do conhecimento do locutor. Ao tratar-se de conhecimento partilhado, *claro* configura a eliminação de respostas alternativas, como em (17), em que a resposta com *claro* marca que outra alternativa não poderia ser considerada e o equivalente em francês no corpus é *bien sûr*. A frase que segue *claro* (em itálico) assinala essa inevitabilidade, podendo mesmo a resposta revelar surpresa em relação à pergunta.

(16) RF: Isso é já uma condicionante muito grande, e as pessoas começam a pensar na saúde, não é? Eu costumava outro dia dizer
FSS: É claro. E se conduzir não beba
RF: que o bêbedo era o grande amigo do viticultor, mas não há dúvida nenhuma que temos de combater o alcoolismo público. (CRPC-oral)

(17a) A respondente para cá, a respondente para lá, Lisboa tantos de tal na sede desta Polícia. A linhas tais e tais o chefe de brigada, interrompe-se para lembrar que qualquer correção pode ser feita em aditamento. Mena sabe. Adiante. Avisa no entanto que desta vez é definitivo, últimas declarações. Convém deixar tudo em ordem porque vai ser transferida, esclarece ele.

Transferida?, pergunta a presa.
Elias Chefe: Claro. *Não é nada que não estivesse à espera, acho eu*. (Balada,163)

(17b) Bien sûr. Il n'y a là rien qui puisse vous surprendre, je crois (Balada, 235)

A pergunta e a resposta podem ter uma única fonte enunciativa, sendo uma forma apelativa de apresentar a informação, num caso de dupla locução, uma vez que a pergunta-resposta não remete para dois locutores reais, mas sim para um "segundo locutor *virtual*" (Grésillon & Lebrave, 1984, p. 123; Mendes et al., 2020). Nesse caso, o próprio locutor sugere através da pergunta a possibilidade de várias respostas, para de seguida identificar a única válida, como em (18).

(18a) Este relatório remete-nos para a seguinte pergunta: para quê a política regional? Para reduzir as disparidades regionais, está claro.

(18b) Ce rapport nous renvoie à la question du pourquoi de la politique régionale? Pour réduire les disparités régionales bien sûr. (Europarl 1819)

Em contextos de diálogo, *claro* pode não responder a uma pergunta, mas sim qualificar uma asserção do interlocutor, como em (19), um exemplo de corpus monolíngue. Um uso equivalente é documentado no corpus monolíngue de francês, no exemplo (20), ocorrendo dois marcadores discursivos: *bien sûr* no turno do falante 1 e *c'est clair* no turno do falante 3 (neste caso, seguido ainda de *oui* 'sim'). Em (19) e (20), não existe resposta positiva, mas sim confirmação da validade do conteúdo da fala do interlocutor, a indicação de que a informação ou o posicionamento do interlocutor é partilhado pelo locutor.

(19) *MAR: / quer dizer / não &se / não / não sei se / não haveria / não deveria haver / um / certo / uma certa apreciação / casuística //$ porque também há pessoas que se eternizam nos <lugares> //$
*JOS: [<] <claro> //$ claro //$ <hhh>$. (CRPC oral)

(20a) Spk2: enfin dans Paris aussi on entend parler les étrangers tiens on s'dit
Spk1: bien sûr (CFPP2000)
'Spk2: mas em Paris também se ouvem falar os estrangeiros olha dizemos nós. Spk1: claro

(20b) Spk 4: non parce qu'le R.E.R. ça va vraiment très vite mais j'vais plus vite qu'en métro [en vélo]
Spk 3: oh ben c'est clair oui (CFPP2000)
'Spk4 : não porque o R.E.R vai mesmo depressa mas eu vou mais depressa do que de metro [de bicicleta] Spk 3 : ah sim está claro sim

É frequente *claro* ocorrer em resposta a um comentário modalizado do interlocutor, como em (21), em que elimina uma alteridade, que pode ser contemplada de forma explícita ou implícita. Em (21), a primeira fala sugere que ser uma insinuação terrível é uma opinião do interlocutor (e implicitamente não é opinião de quem fala), sendo essa posição enunciativa eliminada pela resposta (*claro que é*), que estabelece uma inevitabilidade e reposiciona a interação do domínio da opinião para o da factualidade. São estes enunciados tipicamente exclamativos, podendo ocorrer *claro* / *claro que é* / *claro que sim*, e sendo os possíveis equivalentes funcionais em francês *évidemment*, *bien sûr*, *bien sûr que oui*. *Claro* pode qualificar um enunciado anterior que tem a mesma fonte enunciativa, havendo eliminação de outras possíveis interpretações do próprio locutor, tal como (18) era uma resposta a uma pergunta da mesma fonte enunciativa. Veja-se (22), em que a inevitabilidade do comentário é marcado por *claro* após um processo de eliminação interior de outras alternativas (em itálico), sendo *claro* traduzido por *bien sûr*.

(21) Se acha que é uma insinuação terrível
O Sr. Miguel Macedo (PSD) : - Claro que é ! (CRPC-escrito)

(22a) Feitos os cálculos pelo provável, Elias Chefe determina que chegaram ali de madrugada. De táxi, não podia ter sido doutra maneira. Claro, de táxi. (Balada, 16)

(22b) En taxi, *cela ne pouvait être autrement*. Bien sûr, en taxi. (Balada, 31)

Esse processo de reflexão sobre possíveis alternativas e escolha de uma única válida é frequentemente expresso através de *claro* precedido, e sobretudo seguido, do verbo *estar* (*está claro/ claro está*) e com equivalentes em francês como *bien sûr, c'est clair, bien entendu*. A expressão *está claro* parece remeter para o momento da enunciação e para a confirmação ancorada nesse momento, pelo que pode ocorrer em contextos em que o locutor está a considera alternativas e, durante esse processo de pensar/dizer, elimina todas menos uma. Pela proximidade com o momento de enunciação, também apresenta maior proximidade semântica com o adjetivo (o que dizes está claro, o que dizes não pode levantar dúvidas). A expressão ocorre em posição apositiva, com orientação para o segmento precedente, como em (23) ou para o segmento seguinte, como em (24), ou introduzindo uma completiva (*está claro que / claro está que*), como em (25). Embora o verbo *estar* remeta para o momento da enunciação, a expressão está lexicalizada e gramaticalizada, não admitindo variação da flexão verbal. O equivalente em francês em (25), *la chose est claire* 'a coisa é clara', apresenta o verbo *être* em construção copulativa com sujeito feminino singular e concordância do adjetivo *clair*.

(23a) Agradeço à Comissão que o tenha feito, ainda que não concorde, claro está, com tudo o que a Comissão aqui disse.
(23b) je remercie la Commission de l'avoir fait, même si, bien entendu, je ne partage pas tout ce qu'a elle a dit. (Europarl 9669)

(24a) A questão reside em saber se podemos - e sobretudo se queremos - assegurar que os nossos cidadãos gozem de um descanso nocturno e, claro está, de um descanso diurno saudáveis.
(24b) La question qui se pose est la suivante : sommes-nous capables d'assurer à nos citoyens une nuit, et même une journée saine de tranquillité et, surtout, en avons-nous la volonté politique ? (Europarl 50027)

(25a) Senhor Presidente, claro está que todos têm direito ao silêncio.
(25b) Monsieur le Président, la chose est claire : tout le monde a droit au silence. (Europarl 36164)

O uso de *claro está* em posição apositiva em final de enunciado tem frequentemente valor de avaliação negativa em relação a outra posição enunciativa, como em (26), que segue um contexto com modalização autonímica de empréstimo (segmentos entre aspas) que reconfigura, negativamente, o discurso do outro e pode ser expresso em francês por *bien entendu*. Ou um posicionamento em relação ao seu próprio discurso, como no exemplo (27) com *bien entendu*.

(26) Exprime as suas exigências pela boca do presidente da CIP, advoga a tranquilidade dos espíritos e a ordem nas ruas e nas empresas, a « sua tranquilidade » e a « sua ordem », claro está. (CRPC-escrito)

(27) quand je vais voilà quand je vais m'acheter de la nourriture + quand j'ai envie d'une chose tant que c'est pas de m'acheter un diamant de dix carats place Vendôme hein bien entendu (CFPP2000)

'quando vou lá está quando vou comprar comida + quando tenho vontade de uma coisa desde que não seja comprar um diamante de dez quilates na praça Vendôme hum claro está'

Em certos contextos, *claro* não assinala necessariamente que o locutor também partilha o conhecimento do seu interlocutor, mas sim que a informação está de acordo com o conhecimento que tem de outras situações, que tornam p inevitável. *Claro* indica um valor factual, inevitável de acordo com as circunstâncias e, portanto, incontestável. Os equivalentes funcionais em francês nestes contextos apontam aliás para o estatuto factual, pela presença do elemento *fait* 'facto' nalgumas unidades: *évidemment, de fait, c'est un fait*. É o caso de (28), em que *claro* assinala que o comportamento da polícia era previsível e inevitável, numa construção antiteleonímica, com avaliação negativa e construção retroativa da previsibilidade, que poderia ter como equivalente em francês *évidemment*. A natureza incontestável de p pode ser expressa no enunciado do interlocutor, como em (29), em que o interlocutor usa marcadores de certeza (*não há dúvida nenhuma*), e que poderia ser expressa em francês por *de fait*, *c'est un fait*.

(28) *PED: [<] <e> depois <a polícia foi-se embora durante a noite> //$
*SAN: [<] <hhh>$
*NUN: <hhh> //$
*AMA: [<] <claro> //$ (CRPC-Oral)

(29) RF: (...) Ora, *não há dúvida nenhuma* que do ponto de vista cultural o vinho não é a mesma coisa que o whisky.
FSS: É claro. Ah... há um outro dado: a... o consumo de vinho em Portugal tem vindo a decrescer acentuadamente nos últimos anos. Concorda com, com esta, com esta análise? (CRPC-oral)

Claro pode assinalar apenas que o interlocutor mantém a sua atenção ao discurso do outro, sem forçosamente garantir concordância com o que é dito. No entanto, em comparação com o marcador discursivo *pois*, *claro* favorece uma leitura, mesmo fraca, de concordância (que pode ser verídica ou apenas assegurar a continuação da interação). No exemplo (30), a informação apresentada pelo primeiro falante (GRA) não é conhecimento partilhado por LUC antes da interação, que inicia o turno de fala com *pois*, que marca a sua atenção à troca comunicativa. Num segundo momento do seu turno de fala, LUC integra a informação como conhecimento próprio (há uma pausa longa, com fim de enunciado após o segundo *pois*), e usa *claro*, dando aliás de seguida uma justificação para este posicionamento enunciativo (*claro porque*), podendo ser equivalentes funcionais em francês *bien sûr*, *en effet*. Pelo contrário, no exemplo (31), em que dois interlocutores diferentes reagem respetivamente com *pois* e *claro* ao turno de fala de MAR, não é possível determinar se *claro* envolve concordância ou apenas a confirmação da atenção.

(30) *GRA: eu acho / eu só tenho um termo em / francês / para definir um tipo destes //$ É um emmerdeur //$
*LUC: [<] <hum //$ pois //$ pois> //$ hhh / <claro //$ claro / *porque* não resolve a vida dele nem a das pessoas também> //$ (CRPC-Oral)

(31) *MAR: &eh / &u / uma avaliação / por exemplo //$ para j· / devia ser casuístico //$ portanto / a &casuisi / a casuisticidade / implicaria / uma avaliação / e a possibilidade de prolongar ou <não> //$
*JOS: [<] <hum> hum //$
*MAR: / incentivos //$ conhecimentos / etecetera //$ e não é com aquelas / &eh / com aqueles / cursozinhos de nada / <e com aqueles> /$
*FER: [<] <pois> //$

*JOS: [<] <claro> //$ (CRPC-oral)

O exemplo seguinte, mais extenso, confirma a disponibilidade de *claro* para assegurar um valor de concordância (e conhecimento partilhado) e um valor de confirmação de atenção (*monitoring*): o uso de *claro*, juntamente com *exacto* e *isso mesmo* na fala de B aponta para um valor de concordância, mas o uso de *claro* pelo falante A parece apenas assinalar a atenção ao discurso de B e incentivar a sua continuação.

(32) A - eh, eh, essas pessoas, que portanto agora q[...] há muita gente que, de formação no exterior, isso significará que talvez de, dentro dum, dum, dum futuro mais ou menos próximo, quem quiser fazer... a sua investigação terá mais condições até para a fazer dentro de Portugal. Porque *já há p[...] pessoas que vieram de fora com boa preparação e que podem orientar, não*?
B - exacto, claro, isso mesmo. eh... eh... porque são pessoas que estão muito mais disponíveis realmente para o trabalho científico, para orientar os outros, eh... para, para, para criar equipas de trabalho, eh, vão precisar é que lhe também dêem meios técnicos, não é.
A - claro.
B - mas desde que isso exista e desde que as escolas apostem também na investigação, que era uma velha pecha, as universidades portugueses estavam exclusivamente voltadas para o ensino, eram exclusivamente máquinas de avaliação, em que... se davam as aulas e se avaliavam os alunos e mais nada e não se procurava realmente favorecer nem... incentivar a investigação. não... não se procuravam realmente, que é uma componente que tem que ser forte e que tem de ser acarinhada, não é, tem que se dar condições de trabalho aos docentes, dar-lhe gabinetes, dar-lhe meios técnicos, computadores laboratórios, essas coisas todas e... e criar-lhes condições de trabalho. [...] tem que se fixar as pessoas assim. porque senão também eles começam a... dispersar-se, a ficar divididos, a procurar um múltiplo emprego,
A - claro.
B - também depende um pouco do, da visão das escolas de que, do, de quem estiver à frente das escolas e de que é, da forma de gerir das escolas e também um pouco também de... da visão do, de, de quem pode às vezes... repartir os bolos e os dinheiros, não é, eh... mas... penso que essa componente não pode ser descurada, sob pena de, de realmente a universidade não, não, não... estar realmente a... a fa[...], a... a funcionar da maneira que, que, do meu ponto de vista... devia ser a sua, não é, funcionar m[...] realmente a universidade tem que ser um local essencialmente de investigação, do aprofundamento do saber. também ensinar, também é, isso é verdade.
A - claro. (CRPC-oral)

3.2. *Claro* em contextos de dialogismo interdiscursivo

Nos exemplos anteriores, *claro* responde ou qualifica, direta ou indiretamente, o discurso do interlocutor (podendo ser um locutor virtual, como em (18)). *Claro* pode ainda ocorrer em contextos que não configuram um diálogo e em que o locutor qualifica uma proposição do seu próprio enunciado como expectável, inevitável, decorrendo de convenções conhecidas do interlocutor e dos participantes na comunicação. Em (33), *claro* tem como equivalente funcional em francês o advérbio *naturellement* e em (34) *bien évidemment*.

(33a) Os fundos de pensões, a que foram impostas regras estipulando em que é que podiam ou não investir, tal como acontecia com as nossas obrigações do Tesouro, tiveram uma rentabilidade anual de 5,2%, ou seja, menos de metade da rentabilidade anual de 9,5% dos fundos livres, entre 1984 e 1996. Claro que estamos a falar dos resultados obtidos depois de terem sido deduzidas todas as perdas.

(33b) Les fonds de pension qui ont souffert de cette seconde approche et qui n' ont pas pu investir où bon leur semblait pour respecter les exigences de nos bons du Trésor, ont obtenu un rendement de 5,2 % par an, à peine la moitié des 9,5 % de rendement annuel obtenu par les fonds libres entre 1994 et 1996. Il s' agit naturellement du rendement après déduction de toutes les pertes. (Europarl 14751)

(34a) Aquilo que agora esperam as populações sinistradas, aqueles que perderam tudo, nomeadamente entre os que trabalham no mar ou no turismo, aqueles cuja actividade ficou comprometida por vários anos, é não só que os poluidores reparem os danos que cometeram, mas que sejam envidados todos os esforços para que a sua desgraça actual sirva, de futuro, aos outros, a fim de impedir que voltem a verificar-se crimes semelhantes.
Estamos a pagar, é claro, o preço do nosso desleixo. (Europarl 2221)

(34b) Ce qu'attendent maintenant les populations sinistrées, ceux qui ont tout perdu, notamment parmi les professionnels de la mer et du tourisme, ceux dont l'activité est compromise pour plusieurs années, c'est non seulement que les pollueurs réparent les dégâts qu'ils ont commis, mais que tout soit mis en oeuvre pour que leur malheur actuel serve demain aux autres afin d'empêcher le renouvellement de pareils crimes.
Nous payons bien évidemment le prix de nos abandons.

É frequente o uso de *claro* indicar uma reação a uma outra interpretação, explícita ou implícita, de um interlocutor real ou virtual. Por exemplo, em (35), poderia haver alguma dúvida por parte do interlocutor sobre a posição da comissão a que pertence o locutor. Nestes contextos, os equivalentes em francês apontam para uma fonte enunciativa que é uma comunidade não especificada a que pertence o locutor (*ON-locuteur*, na proposta de Anscombre (2013)), que atribui um valor epistémico consensual e inevitável à proposição: *il est évident que* 'é evidente que'*, nul doute que* 'nenhuma dúvida de que'*, naturellement* 'naturalmente', sendo que *nul doute que* verbaliza a eliminação da alteridade.

(35a) A Comissão partilha da preocupação crescente com o desaparecimento do jornalista russo Andrei Babitsky e com a questão da liberdade de imprensa em geral na Chechénia. Esta questão veio agravar a nossa grande preocupação com a situação da população civil da Chechénia em geral, e já não é a primeira vez que oiço o senhor deputado Posselt referir-se ao assunto. Chegam-nos notícias alarmantes sobre violações dos direitos humanos. Claro que o recurso à força neste conflito é desproporcionado. A delegação da Comissão em Moscovo participou numa démarche da tróica da União Europeia em Moscovo, a onze deste mês. Sublinhámos a nossa grande preocupação com a questão da liberdade de imprensa e da liberdade de expressão.

(35b) La Commission partage l'inquiétude croissante due à la disparition du journaliste russe, M. Babitsky, et aux conditions des médias indépendants de Tchétchénie en général. Ce sujet est au sommet de nos principales inquiétudes face à la situation critique de la population civile de Tchétchénie, dont M. Posselt vient de parler. Des nouvelles alarmantes parlent de violations des droits de l'homme. Il est évident qu'il y a eu un recours disproportionné à la force dans ce conflit. La délégation de la Commission à Moscou a participé à une démarche de la Troïka de l'Union européenne le 11 de ce mois. Elle a souligné ses vives inquiétudes en matière de liberté de la presse et de liberté d'expression (Europarl 16859)

No caso de (36), a fala de X, aqui encurtada, mostra que este falante se posiciona contra a introdução de um imposto e implicita que esse imposto trará menos clareza. O falante Y considera inquestionável a necessidade de clareza (é claro que...) e de seguida responde ao implícito (o imposto Tobin vai...).

(36a) X: O mercado de capitais não precisa de insegurança provocada por nós, *o mercado de capitais precisa é de clareza da nossa parte*. (...)
Y: Estamos à procura de soluções. É claro que os mercados de capitais precisam de clareza – é de facto essa a ideia. O imposto Tobin vai trazer uma certa transparência a uma questão que é muito obscura.

(36b) X : Le marché des capitaux n'attend pas de nous que nous amenions l'incertitude, il a besoin de clarté.
Y : Nous cherchons des solutions. Nul doute qu'il faille plus de clarté sur les marchés de capitaux : c'est bien là l'idée. La taxe Tobin mettra un peu de transparence dans cette question très obscure. (Europarl 6294)

É frequente *claro* ocorrer numa sequência argumentativa, em que estabelece a factualidade e inevitabilidade de p, e é seguido por uma relação de oposição, marcada explícita ou implicitamente: *claro que p, mas*. Em (37), *claro* indica que o locutor e o seu interlocutor (leitor virtual) sabem ambos que Mena não o poderia ter dito, mas mesmo tendo em conta esse facto, o locutor pode fazer a afirmação sobre o estado de Mena e do major. Em (38), *claro* assinala a integração de um contra-argumento (houve progressos) seguido de um relação concessiva (contrariamente ao que p sugere, é necessário continuar os esforços). Esta estrutura é mantida na tradução francesa, com um movimento de concordância seguido de uma oposição marcada em ambos os casos por *mais*. Em (37), *claro* tem como equivalente *évidemment*, mas em (38) ocorre um marcador que é específico destes contextos de oposição, *certes* (que em português poderia ser transposto por *é certo que*, *é verdade que*).

(37a). «Nem se vestiu.» Palavras da própria. Nua em pêlo como saiu dos braços do major, a lavar paredes no grande desespero. E o major a dormir o sono dos desgastes, mais que pacificado no corpo dela. Não o disse, claro, *mas* esses segredos lê-lhos Elias na raça que ela tem, não carecem de ser mencionados. (Balada, 128)

(37b) «Elle ne s'est même pas habillée.» Ses propres paroles. Complètement nue, comme elle était sortie des bras du major, lavant les murs dans le grand désespoir. Et le major qui dormait d'un sommeil repu, tout pacifié par le corps de Mena. Elle ne l'a pas dit, évidemment, *mais* ces secrets, Elias les lit sur la nature de ce corps, ils n'ont pas besoin d'être mentionnés (Balada, 186)

(38a) Os actos racistas e xenófobos são completamente inaceitáveis na nossa Comunidade, seja qual for o local em que ocorram. São contrários aos princípios que estão na base da fundação da União Europeia, como o disse ontem o presidente Havel: os princípios da liberdade, da democracia e do respeito pelos direitos humanos. Claro que nestes últimos anos se fizeram progressos, *mas* temos de continuar a esforçar-nos juntos por criar um clima de tolerância, em que o racismo e a xenofobia sejam considerados totalmente reprováveis e inaceitáveis, ao mesmo tempo que tratamos com severidade incidentes como aqueles de que estamos a falar aqui esta tarde.

(38b) Les actes racistes et xénophobes sont totalement inacceptables au sein de notre Communauté, quel que soit l'endroit où ils se manifestent. Ils sont contraires aux principes mêmes sur lesquels a été fondée l'Union européenne, comme M. Havel le disait hier : les principes de la liberté, de la démocratie et du respect des droits de l'homme. Certes, des progrès ont été accomplis ces dernières années, *mais* nous devons encore oeuvrer ensemble pour aboutir à un climat de tolérance où le racisme et la xénophobie sont inacceptables, tout à fait inadmissibles, tout en répondant avec une extrême vigueur aux incidents tels que ceux dont nous parlons cet après-midi. (Europarl 16487)

O exemplo (39) evidencia uma oposição marcada implicitamente nas duas línguas. *Claro* assinala a concordância com o que é asserido e a relação de oposição é assegurada pela estrutura contrastiva (não é aí que está o mal (...) o mal está...).

(39a) «Claro», insiste o major, «o homem tem todo o direito de olhar». (Mena continua sob a mão de Dantas C; percorrida, divagada.) «Olhar para onde ele quiser. As vezes que quiser. Olhar à vontade, não é aí que está o mal. Mal nenhum», repetiu. «*O mal está no ar sorna do gajo*, na maneira como o gajo anda a «mudar o telefone.» (Balada, 40)

(39b) «Bien sûr», insiste le major. «il a bien le droit de regarder» (Mena est encore sous la main de Dantas C. ; parcourue, sillonnée). «De regarder où il veut. Autant de fois qu'il le veut. Regarder à son aise, le mal n'est pas là. Pas du tout», répéta-t-il. «*Le mal est dans l'air sournois qu'il a*, dans sa manière de surveiller le téléphone.» (Balada, 64)

Mesmo nos contextos em que *claro* não responde ou comenta o discurso de um interlocutor direto, estabelece um relação entre o discurso do locutor e um outro discurso, quer o discurso de uma comunidade epistémica em que o locutor se insere, quer um discurso implícito de um alocutor real ou virtual.

A Tabela 1 sistematiza os valores de *claro* e os seus equivalentes funcionais em francês.

Tabela 1. Valores de *claro* e equivalentes funcionais em francês

Valores de *claro*	**Equivalentes funcionais em francês**
Contextos de diálogo	
Resposta como eliminação de alternativa e conhecimento partilhado (≠ sim) claro / é claro / é claro que p	bien sûr (que oui/non)
Qualificação de uma asserção: concordância e conhecimento partilhado claro / é claro / é claro que sim / está claro	bien sûr, c'est clair
Eliminação de outra posição enunciativa oposição; avaliação negativa; ironia claro / é claro / claro que	évidemment, bien sûr (que oui) bien entendu
Qualificação do conteúdo enunciativo como factual, incontestável	évidemment, de fait, c'est un fait, bien sûr
Concordância, mo*nitoring*	bien sûr, oui
Contextos de dialogismo interdiscursivo	
Qualificação do conteúdo enunciativo como inevitável, expectável (decorre de convenções)	naturellement
Qualificação do conteúdo enunciativo como inevitável; enunciado como reação a outra posição enunciativa	il est évident que, nul doute que
Relação de oposição, concessão	certes

Analisados os contextos de *claro* e seus equivalentes funcionais em francês, interessa-nos voltar à discussão inicial sobre a categorização desta unidade. Como vimos, o equivalente funcional *of course* em inglês foi categorizado como partícula modal nalguns dos contextos analisados em Aijmer (2013). No caso de *claro*, e à semelhança da discussão em Cuenca (2013), vemos que, embora tenha um valor epistémico, não apresenta algumas das propriedades essenciais para poder ser considerado uma partícula modal. Assim, como mostra o exemplo (17a), *claro* pode constituir, de forma isolada, uma resposta a uma pergunta e ocorre como elemento

parentético, precedido ou não de verbo *ser* ou *estar*, em posição final de frase, como em (18a), inicial (22a), medial (37a). Estas propriedades excluem desde logo a sua possível categorização como partícula modal, se respeitarmos os critérios estabelecidos para as línguas germânicas. Estas características levam, aliás, a uma proposta de categorização do catalão *clar* como "marcador modal", uma categoria que em Cuenca (2013) inclui também interjeições e advérbios modais. Uma vez que a definição de marcador discursivo usada na nossa análise é abrangente e envolve várias funções, entre as quais valores modais, analisamos *claro* como um marcador discursivo nos vários contextos aqui considerados.

4. Comentários finais

A análise dos contextos de *claro* e seus equivalentes funcionais em francês mostra que esta unidade desempenha funções discursivas com valores modais e também estruturais. A análise dos equivalentes em francês do marcador *claro* aponta para um conjunto restrito de contextos em que *clair* pode ocorrer: são construções com verbo copulativo em que *clair* ainda tem, possivelmente não totalmente, um comportamento adjetival. Assim, o processo de gramaticalização de adjetivo em marcador discursivo que é descrito para o português, o castelhano e o catalão, não ocorreu em francês. Na generalidade dos contextos em que ocorre *claro*, os corpora paralelos consultados mostram que *bien sûr* pode ser um equivalente funcional, mas alguns contextos apontam para outros equivalentes preferenciais, como mostra a Tabela 1. A análise de *claro* insere-se no trabalho em curso sobre marcadores discursivos que veículam valores modais, como o caso de *de facto*, que assinala, de forma geral, um comprometimento do locutor baseado num valor de factualidade (Mendes & Lejeune, no prelo). Os marcadores *de facto* e *claro* aproximam-se pelos seus valores de confirmação / concordância e de marcação de um conhecimento prévio: *de facto* tem primeiramente um valor de confirmação da perspetiva de outra entidade (em contexto dialógico), ou da perspetiva do próprio locutor quando esta é inicialmente modalizada. Diferem, nestes contextos, por *claro* assinalar a posição enunciativa como inevitável, resultado de convenções partilhadas por uma comunidade enunciativa mais lata. Tanto *de facto* como *claro* podem ocorrer em estruturas argumentativas, mas com diferenças significativas. O marcador *de facto* é nesses contextos tipicamente precedido de *mas* e introduz um segmento com valor de oposição, podendo ser interpretado com valor adversativo mesmo sem a presença da conjunção. O marcador *claro* marca um primeiro movimento de integração da perspetiva do outro e dos seus contra-argumentos, sendo seguido de um movimento de oposição tipicamente marcado pela adversativa *mas*, ou marcado implicitamente. Ambos os marcadores partilham ainda funções estruturais, mas enquanto *de facto* é usado nesses contextos para fazer progredir a informação e introduz um segmento elaborativo, *claro* é usado para marcar a concordância com o outro e/ou a atenção ao interlocutor. Os equivalentes de *claro* em francês realçam a concordância e o conhecimento partilhado (*bien sûr*), a factualidade (*de fait*, *c'est un fait*), a oposição e até ironia (*bien entendu*) e a inserção em estruturas argumentativas precedendo segmentos contrastivos ou concessivos (*certes*).

Sobressai da análise de *claro* a sua natureza dialógica e polifónica: configura uma resposta ou comentário a um enunciado anterior ou virtual e também configura um contexto polifónico, que convoca um conhecimento partilhado entre locutor, interlocutor e a comunidade discursiva a que pertence o locutor (*ON-locuteur* (Anscombre, 2013)).

Agradecimentos

Este trabalho foi realizado com o apoio do financiamento da Fundação para a Ciência e a Tecnologia, no âmbito do projeto UIDP/00214/2020.

Referências

Aijmer, Karin (2013) Analyzing modal adverbs as modal particles and discourse markers. In Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (eds.), *Discourse markers and modal particles. Categorization and description*. John Benjamins, pp. 89–106.

Amiot, Dany & Nelly Flaux (2007) Naturellement en position détachée. In Nelly Flaux & Dejan Stosic (eds.), *Les constructions détachées : Entre langue et discours*. Artois Presses Université, pp. 58–102.

Anscombre, Jean-Claude (2013) Entité lexicale : Bien sûr. In Jean-Claude Anscombre, María Luisa Donaire & Pierre Patrick Haillet (eds.), *Opérateurs discursifs du français. Éléments de description sémantique et pragmatique*. Peter Lang, pp. 73–82.

Coniglio, Marco (2008) Modal particles in Italian. *University of Venice Working Papers in Linguistics* (Vol. 18). University of Venice, pp. 91–129.

Crible, Ludivine & Liesbeth Degand (2019) Domains and functions: A two-dimensional account of discourse markers. *Discours* 24. https://doi.org/10.4000/discours.9997

Costa, Ana Luísa (2013) Um pois estruturador. In *Textos Selecionados, XXIX Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 199–211.

Cuenca, Maria Josep & Maria Josep Marín (2009) Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics* 41 (5), pp. 899–914. https://doi.org/10.1016/j.pragma.2008.08.010

Cuenca, Maria Josep & Maria Josep Marín (2012) Discourse markers and modality in spoken Catalan: The case of *(és) clar*. *Journal of Pragmatics* 44 (15), pp. 2211–2225. https://doi.org/10.1016/j.pragma.2012.09.006

Cuenca, Maria Josep (2013) The fuzzy boundaries between discourse marking and modal marking. In Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (eds.), *Discourse markers and modal particles. Categorization and description*. John Benjamins, pp. 181–216.

Degand, Liesbeth, Bert Cornillie & Paola Pietrandrea (eds.) (2013) *Discourse markers and modal particles. Categorization and description*. John Benjamins.

Diewald, Gabriele (2013) "Same same but different" – Modal particles, discourse markers and the art (and purpose) of categorization. In Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (eds.), *Discourse markers and modal particles. Categorization and description*. John Benjamins, pp. 19–45.

Dostie, Gaétane (2001) L'ambiguïté, la synonymie et l'implicite en lexicographie. Quelques observations à partir du champ sémantique 'évidence'. In Paul Bogaards, Johan Rooryck & Paul J. Smith (eds.), *Quitte ou Double sens : Articles sur l'ambiguïté offerts à Ronald Landheer*. Rodopi, pp. 65–85.

Duarte, Isabel Margarida (1989) *Alguns operadores de agulhagem comunicativa (na prosa narrativa de Eça de Queirós e José Cardoso Pires*). Tese de mestrado, Universidade do Porto.

Duarte, Isabel Margarida (2005) Palavras do falar desataviado de todos os dias em Balada da Praia dos Cães. *Semear* 11, pp. 97–125.

Favaro, Marco (2021) *Pragmatic markers in Italian. Four case studies on illocutive functions of adverbs and sociolinguistic variation*. Tese de Doutoramento, Universidade de Turim & Universidade Humboldt de Berlim.

Franckel, Jean.-Jacques (2005) De l'interprétation à la glose: Vers une méthodologie de la reformulation. In D. Lebaud (ed.), *Actes du colloque D'une langue à l'autre*. Presses Universitaires de Franche-Comté, pp. 51–78.

Franckel, Jean-Jacques, Dar Non & Sophie Rose (2017) Étude de certains marqueurs discursifs « perception » en français, russe et khmer. *Langages* 207, pp. 49–64. https://doi.org/10.3917/lang.207.0049

Franco, António C. (1990) Partículas modais do português. *Revista da Faculdade de Letras: Línguas e Literaturas* 7, pp. 175–196.

Fuentes Rodríguez, Catalina (1993) *Claro*: modalización y conexión. In Pedro Carbonero Cano & Catalina Fuentes Rodríguez (eds.), *Sociolingüistica andaluza: Estudios sobre el lenguaje oral* (Vol. 8). Secretariado de Publicaciones de la Universidad de Sevilla, pp. 99–126.

Généreux, Michel, Iris Hendrickx & Amália Mendes (2012) A large Portuguese corpus on-line: Cleaning and preprocessing. In Helena Caseli et al. (eds.), *Proceedings of the 10th International Conference PROPOR1012*. Springer-Verlag, pp. 113–120.
Grésillon, Almuth & Jean-Louis Lebrave (1984) Qui interroge qui et pourquoi? In Almuth Gresillon & Jean-Louis Lebrave (eds.), *La langue au ras du texte*. Presses Universitaires de Lille, pp. 57–132.
Halliday, Michael A. K., & Christian M. Matthiessen (2004) *An introduction to functional grammar* (3.ª ed.). Arnold.
Koehn, Philipp (2005) Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings of the tenth Machine Translation Summit*. Asia-Pacific Association for Machine Translation, pp. 79–86. Disponível em https://aclanthology.org/2005.mtsummit-papers
Lejeune, Pierre & Amália Mendes (2020) Le marqueur discursif du portugais européen *pois* et ses principaux équivalents fonctionnels en français : Analyse contrastive. In Isabel Duarte & Rogélio Ponce de León (eds.), *Marcadores discursivos. O português como referência contrastiva*. Peter Lang, pp. 99–120.
Lejeune, Pierre & Amália Mendes (no prelo). *Functional French equivalents of some modal uses of European Portuguese discourse particle* lá. Peter Lang.
Lima, José Pinto de (2002) Grammaticalization, subjectification and the origin of phatic markers. In Ilse Wischer & Gabriele Diewald (eds.), *New reflections on grammaticalization*. John Benjamins, pp. 363–378.
Lopes, Ana Cristina Macário (2012) Contributos para uma análise semântico-pragmática das causais de enunciação no português europeu contemporâneo. *Alfa* 56 (2), pp. 451–468. https://doi.org/10.1590/s1981-57942012000200005
Lopes, Ana Cristina Macário (2021) Contributos para o estudo do marcador discursivo "claro" em português europeu. *Revista Galega de Filoloxía* 14, pp. 71–83.
Mauranen, Anna (1999) Will 'translationese' ruin a contrastive study?. *Languages in Contrast* 2 (2), pp. 161–85. https://doi.org/10.1075/lic.2.2.03mau
Mauranen, Anna (2016) Corpora, universals and interference. In Anna Mauranen & Pekka Kujamäki (eds.), *Translation universals: Do they exist?*. John Benjamins, pp. 65–82.
Meisnitzer, Benjamin (2012) Modality in the Romance Languages: Modal verbs and modal particles. In Werner Abraham & Elisabeth E. Leiss (eds.), *Modality and theory of mind elements across languages*. De Gruyter, pp. 335–359.
Mendes, Amália & Pierre Lejeune (no prelo) Discourse markers in Portuguese. In Maj-Britt Mosegaard Hansen & Jacqueline Visconti (eds.), *Manual of discourse markers in Romance*. De Gruyter.
Mendes, Amália, Pierre Lejeune & Carolina Nunes (2020) Perguntas-respostas em textos escritos: Uma análise no âmbito das relações discursivas. *Revista da Associação Portuguesa de Linguística* 7, pp. 226–241. https://doi.org/10.26334/2183-9077/rapln7ano2020a14
Molinier, Christian (1990) Une classification des adverbes en -ment. *Langue française* 88, pp. 28–40.
Noël, Dirk (2003) Translations as evidence for semantics: An illustration. *Linguistics* 41 (4), pp. 75–785. https://doi.org/10.1515/ling.2003.024
Péroz, Pierre (1992) *Systématique des valeurs de* bien *en français contemporain*. Droz.
Pons Bordería, Salvador (1998) *Conexión y conectores: Estudio de su relación en el registro informal de la lengua*. Universitat de Valencia.
Pons Bordería, Salvador (2003) From Agreement to Stressing and Hedging: Spanish *Bueno* and *Claro*. In Gundrun Held (ed.), *Partikeln und Höflichkeit*. Peter Lang, pp. 219–236.
Pons Bordería, Salvador (2012) Una palabra sobre los apellidos de la sintaxis. In José Bustos Tovar, Rafael Cano Aguilar, Elena García de Paredes & Araceli López Serena (eds.), *Sintaxis y análisis del discurso hablado en español. Homenaje a Antonio Narbona*. Servicio de Publicaciones de la Universidad de Sevilla, pp. 375–390.
Waltereit, Richard (2001) Modal particles and their functional equivalents: a speech-act-theoretic approach. *Catalan Journal of Linguistics* 6, pp. 61–80. https://doi.org/10.1016/S0378-2166(00)00057-6

Advérbios compostos do Português do Brasil

Izabela Müller[1,3]*, Nuno Mamede*[2,3]*, Jorge Baptista*[1,3]

[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais, Faro, Portugal
[2] Universidade de Lisboa, Instituto Superior Técnico, Lisboa, Portugal
[3] INESC-ID Lisboa - Human Language Technology Lab, Lisboa, Portugal

Resumo

Este artigo apresenta um estudo em curso que tem como objetivo identificar, classificar e descrever advérbios compostos do português brasileiro, com foco em suas propriedades sintáticas e semânticas. A meta é criar um léxico abrangente de advérbios compostos do português, que possa ser aplicado em vários domínios, como processamento de linguagem natural (PLN), tradução e ensino de idiomas. Essa classificação ajudará a estabelecer os padrões de distribuição específicos da variedade brasileira, permitindo uma distinção clara entre expressões adverbiais brasileiras e europeias. Além de suas aplicações práticas, o estudo visa investigar vários aspectos dessas expressões adverbiais. Isso inclui examinar as suas frequências, possíveis ambiguidades, *faux-amis* e diferenças gramaticais quando comparadas ao português europeu. Além disso, esta pesquisa visa contribuir para o desenvolvimento de um léxico abrangente de expressões adverbiais, com vista a uma melhor compreensão do funcionamento e uso desses elementos linguísticos.

Palavras-chave: advérbios compostos, sintaxe, português brasileiro, expressão idiomática, MWE.

Abstract

This ongoing study aims to identify, classify, and describe multiword/compound adverbs in Brazilian Portuguese, focusing on their syntactic-semantic properties. The goal is to create a comprehensive lexicon of compound adverbs in Portuguese, which can be applied in various domains such as Natural Language Processing (NLP), translation, and language teaching. This classification will help establish the distribution patterns specific to the Brazilian variety, allowing for a clear distinction between Brazilian and European Portuguese adverbial expressions. In addition to their practical applications, the study aims to investigate various aspects of these adverbial expressions. This includes examining their frequency, potential ambiguities, *faux-amis*, and grammatical differences when compared to European Portuguese. Furthermore, this research seeks to contribute to the development of a comprehensive lexicon of adverbial expressions, in view of enhancing the understanding of the functioning and usage of these linguistic elements.

Keywords*:* compound adverbs, syntax, Brazilian Portuguese, idioms, MWE.

1. Introdução

Os *advérbios* são considerados uma classe gramatical complexa e heterogênea[1] de palavras, sem variação flexional e classificações semânticas diversas (Bechara, 2009; Costa, 2008; Cunha & Cintra, 2017; Raposo, 2013). Podem funcionar como modificadores de verbos, adjetivos e outros advérbios e podem, ainda, modificar frases. Semanticamente, fornecem sobretudo informações adicionais sobre circunstâncias de tempo, modo, lugar, quantidade, intensidade, entre outros aspectos.

[1] Este texto foi escrito na variedade do português do Brasil.

Os *advérbios compostos* (ou *expressões multipalavra* adverbiais, do inglês multiword expressions (MWE); ou, ainda, *locuções adverbiais*, de acordo com a terminologia gramatical do português)[2] são expressões formadas por duas ou mais palavras que funcionam nessa combinação como uma só unidade lexical, cuja constituição é fixa, ou seja, seus elementos ocorrem em uma determinada ordem, fortemente restrita, geralmente não sendo permitida a permutação, a inserção, a transposição nem a redução (elipse) de seus elementos (Gross, 1982, 1986; Guimier, 1996; Mel'čuk, 2023). Têm como característica a não-composicionalidade semântica, ou seja, a interpretação global não pode ser derivada do significado individual de seus elementos, ou seja, em sua maioria, são idiomáticos. Os exemplos abaixo mostram primeiramente, (1) expressões adverbiais idiomáticas em português brasileiro (PB); (2) uma expressão similar, em português europeu (PE); e (3) um advérbio derivado terminado em *-mente*, equivalente às anteriores (3)[3]:

(1) [O Leo fez a prova] *na maciota*; *com o pé nas costas* (PB)

(2) [O Leo fez a prova] *com uma perna às costas* (PT)

(3) [O Leo fez a prova] *facilmente*

Este artigo tem como objetivo apresentar um estudo em andamento sobre as construções adverbiais no português brasileiro, mais especificamente os advérbios compostos/multipalavra. Nos últimos anos, houve um considerável aumento nas pesquisas dedicadas ao estudo de expressões multipalavra (MWE), expressões idiomáticas, fraseologia, e outras construções de natureza não composicional (Ramisch, 2015) que constituem um problema para as aplicações de processamento de língua natural, principalmente no que tange a sua identificação e reconhecimento automático em textos. Os advérbios compostos são formados por elementos lexicais (preposições, nomes, adjetivos, etc.) que em sua maioria são detectados pelos sistemas como unidades individuais. Experiências realizadas (Gonçalves et al., 2020) demonstram os limites dos léxicos computacionais disponíveis para o processamento desta classe.

A pesquisa tem como principal propósito identificar, classificar e descrever os advérbios compostos, levando em consideração suas propriedades sintáticas e semânticas, com o intuito de enriquecer um léxico de advérbios compostos do português. Além disso, busca-se fornecer uma descrição formal e sintática destes advérbios, bem como uma descrição linguística voltada para o Processamento de Linguagem Natural (PLN) e outras aplicações. Por fim, pretende-se também determinar as fronteiras lexicais entre o português brasileiro e europeu, no que se refere às diferenças idiomáticas, gramaticais e respectivas frequências destas expressões.[4]

No âmbito deste estudo, realizou-se, até ao momento, um levantamento de aproximadamente 3.300 expressões adverbiais em português. Dentro deste conjunto, muitas destas expressões são comuns a ambas as variedades linguísticas, enquanto outras são exclusivas de cada uma das variedades.

Este artigo está organizado do seguinte modo: Começamos esta breve introdução por apresentar sucintamente os conceitos aqui empregues de advérbio e advérbio composto, bem como apresentar as motivações subjacentes a este estudo. Em seguida, na seção 2, são apresentados alguns estudos similares sobre construções de advérbios compostos em diferentes línguas. Na seção 3, descrevemos genericamente a metodologia empregada para a construção do léxico de expressões adverbiais, juntamente com os critérios

[2] Neste artigo, usamos o termo *advérbio composto* no sentido em que este é definido no texto. Tal corresponde, em grande medida, ao conceito do termo mais tradicional *locução adverbial*, bem como ao do termo mais recente (e genérico) de *expressão multipalavra* (adverbial). Não atribuímos importância a esta diferença terminológica.

[3] Nos exemplos, a frase sobre a qual o advérbio opera é representada entre parênteses retos. No caso das expressões (mais) idiomáticas, acrescenta-se uma glosa entre aspas simples '...', sobretudo quando uma dada interpretação é exclusiva de uma das variedades do português. Quando relevante, essas variedades são assinaladas, em expoente, pelas siglas PT (=português europeu) e PB (= português brasileiro). Os exemplos retirados do corpus *TenTen2018* do Sketch Engine (Wagner et al., 2018), https://app.sketchengine.eu, vêm assinalados com a indicação [SE].

[4] O estudo das construções adverbiais noutras variedades do português, nomeadamente de Angola e de Moçambique, deverá ser prosseguido noutros trabalhos, para os quais este recurso linguístico certamente será útil.

estabelecidos para inclusão e exclusão de determinadas expressões. A seção 4 apresenta o enquadramento teórico-metodológico adotado neste estudo e prossegue com a classificação formal das expressões adverbiais. Por fim, na seção 5, são apresentados os resultados provisórios deste estudo e indicados os próximos passos.

2. Trabalhos relacionados

Embora os advérbios simples mais frequentes, precisamente aqueles terminados em *-mente*, no contexto do português brasileiro, tenham já sido previamente estudados e classificados de acordo com suas características sintáticas e semânticas (Fernandes, 2011), os advérbios compostos ainda não receberam a mesma atenção quando se trata de análise detalhada e classificação descritiva de suas propriedades sintáticas e semânticas.

No que diz respeito ao português europeu, Palma (2009) desenvolveu um léxico para as construções adverbiais, tendo descrito suas propriedades sintáticas com o objetivo de compará-las a estruturas semelhantes em espanhol. A autora coletou e classificou cerca de 1.800 advérbios compostos, que foram posteriormente contrastados com construções equivalentes em espanhol. Mais tarde, estes advérbios foram adicionados ao léxico da cadeia de PLN para português, STRING (Mamede et al., 2012). Este léxico computacional de advérbios compostos nos serve de base para este projeto. Um trabalho mais recente (Català et al., 2020) aborda os desafios das traduções dos advérbios compostos entre o português e o espanhol.

Encontramos estudos semelhantes sobre a construção de léxicos de advérbios compostos em diferentes línguas, além do português. Por exemplo, Gross (1996) desenvolveu um léxico de advérbios compostos em francês, contendo cerca de 6.800 entradas, com suas respectivas descrições sintáticas e semânticas. Este mesmo dicionário foi mais tarde incorporado em sistemas de Processamento de Linguagem Natural (Laporte et al., 2008; Laporte & Voyatzi, 2008). Para a língua japonesa, o *Dictionary of Multi-Word Expressions* (JDMWE) (Shudo et al., 2011) possui aproximadamente 104.000 entradas de expressões multipalavra, das quais 6.000 são expressões adverbiais. Este léxico fornece informações sobre as funções sintáticas, estrutura, mobilidade nas frases e frequências destas expressões. Em relação à língua Tcheca, um estudo realizado por Žižková (2018) analisou cerca de 470 construções adverbiais e, dentre estas, 103 expressões adverbiais foram acrescentadas a um dicionário morfológico de MWE já existente naquela língua. Quanto à língua espanhola, Català (2003) construiu um léxico de expressões adverbiais fixas, contendo cerca de 6.000 entradas. Essas expressões foram formalizadas e classificadas em 11 classes, seguindo os critérios de M. Gross (1986). Adicionalmente, para o espanhol, o léxico *SentiText*, (Moreno-Ortiz et al., 2013), criado especificamente para a análise de sentimentos, inclui 2.255 construções adverbiais. Para a língua inglesa, (Shigeto et al., 2013) marcaram aproximadamente 1.500 MWE adverbiais do Wiktionary e 468 do Penn Treebank.

3. Metodologia

Esta seção apresenta a metodologia para a construção de um léxico de advérbios compostos, contendo tanto as expressões do português europeu como da variedade do Brasil.

3.1. Recursos linguísticos para a construção do léxico

O foco principal do estudo consistiu na expansão de um recurso lexical computacional já existente, originalmente desenvolvido para o português europeu, o qual compreende aproximadamente 2.750 advérbios, incluindo advérbios simples, derivados em *-mente* e advérbios compostos. Os advérbios compostos deste léxico são baseados no trabalho de Català et al. (2020), que foram inicialmente estudados e descritos por Palma (2009). O léxico inclui as seguintes informações:

1. cerca de 1.000 advérbios derivados e terminados em *-mente* acompanhados de sua base adjetival correspondente e sua classificação sintático-semântica (Fernandes, 2011), feita de acordo com propriedades apresentadas por Molinier e Levrier (2000);
2. aproximadamente 1.750 advérbios compostos (Català et al., 2020; Palma, 2009), classificados com base em suas estruturas formais (Gross, 1986) e também de acordo com propriedades sintático-semânticas de Molinier e Levrier (2000);
3. aproximadamente 6.000 advérbios derivados terminados em *-mente* com informação sobre o adjetivo de base, porém sem informação complementar sobre a classe sintática-semântica.

Muitos advérbios dos indicados em 2., acima, existem igualmente no português do Brasil, o que foi verificado manualmente por um dos autores, que é falante nativa, e validado em *corpora* ou em textos da internet (ver adiante).

Para completar e aprimorar os recursos lexicais existentes com advérbios compostos do português do Brasil, examinamos algumas fontes para a coleta de novas expressões, mais precisamente o *Dicionário de Locuções e Expressões da Língua Portuguesa* (Rocha & Rocha, 2011), que contém por volta de 18.000 expressões, das quais, aproximadamente, 3.000 adverbiais. Neste dicionário, as expressões adverbiais estão em meio a outras construções multipalavra, tanto adjetivais quanto verbais, pelo que é necessária uma consulta manual e meticulosa para extrair somente as que fazem parte deste estudo. Embora o dicionário não especifique a variedade do português à qual essas expressões pertencem, uma razoável porcentagem se alinha com as encontradas no português europeu, portanto algumas já estavam integradas ao léxico computacional inicial. Houve de se ter o cuidado em verificar que, mesmo quando as expressões são aparentemente as mesmas, poderia haver diferenças distribucionais, morfológicas ou morfossintáticas (por exemplo, em que uma variedade use a expressão mas com diferentes preposições; ou só admita algumas flexões) ou até mesmo a ocorrência de *faux-amis*, os chamados falsos amigos. Dentre os demais recursos para a coleta das expressões, utilizamos ainda o livro *Locuções Adverbiais* (Schwab, 1985), e o *Dicionário Brasileiro de Fraseologia* (Silva, 2013), cuja análise ainda está em curso.

3.2. **Critérios para o recenseamento dos advérbios compostos**

Esta pesquisa envolve, como se disse, o recenseamento, a classificação formal e a descrição das propriedades sintáticas e semânticas de advérbios compostos no português brasileiro. Com o intuito de desenvolver um léxico abrangente de advérbios compostos para o português, partimos de um trabalho previamente elaborado para o português europeu, (Palma, 2009), o qual foi examinado minuciosamente para identificar aquelas expressões que fossem, a princípio, comuns às duas variedades da língua, assinalando-as como tal. Logo nessa etapa, foi possível detectar alguns casos em que cada variedade apresenta pequenas diferenças para uma expressão de significado idêntico (e.g., *com uma perna às costas*^{PT}/*com o pé nas costas*^{PB}). Em seguida fizemos o recenseamento das expressões nos recursos citados acima. Estabelecemos alguns critérios a fim de determinar a inclusão de expressões que fossem relevantes ao nosso estudo ou a sua exclusão, devidamente fundamentada.

Decidimos, assim, incluir os seguintes tipos de construções adverbiais:

1 - Expressões adverbiais idiomáticas, e.g., *com a cara e a coragem*

(4) [O Leo explorou as ruas] *com a cara e a coragem*

Estas expressões adverbiais são semanticamente não-composicionais, ou seja, exibem um certo grau de fixidez formal interna, apresentando restrições quanto a:

(i) *permutação* dos elementos coordenados: *[O Leo explorou as ruas] *com a coragem e a cara*;[5]
(ii) a *variação* de gênero e/ou número de seus elementos: *[O Leo explorou as ruas] *com a**s** cara**s** e a**s** coragen**s***;
(iii) a *substituição* de seus elementos por sinônimos/antônimos: *[O Leo explorou as ruas] *com a cara e **o coração***;
(iv) *inserções* de determinantes e/ou modificadores livres ou outros elementos lexicais, mesmo que tais determinantes e/ou modificadores possam modificar os elementos do composto em outro lugar: *[O Leo explorou as ruas] *com a cara **limpa** e a **destemida** coragem* e, finalmente
(v) a *exclusão* de alguns de seus elementos: *[O Leo explorou as ruas] *com a cara*

2 - Construções adverbiais multipalavra equivalentes a uma única palavra, nomeadamente os advérbios derivados com o sufixo *-mente* e.g., *em geral = geralmente*, *de súbito = subitamente*.

3 - Construções adverbiais que permitem algum grau de variação em seus componentes: e.g., *a certa altura*, a qual permite variação dos pronomes demonstrativos usados, como, por exemplo: *a esta* / *essa* / *aquela altura*; ou a variação de outro elemento da construção, neste caso, o adjetivo *em_um (num) determinado / certo / dado / momento*

(5) *A essa altura*, [o Leo já desistiu de ir ao show]

(6) *Em um determinado momento*, [tudo mudou na vida do Leo]

Seria pouco interessante representar num léxico (uma listagem) todas estas famílias de expressões adverbiais, com variação de elementos gramaticais relativamente previsíveis. Estas expressões deverão ser representadas por meio de gramáticas locais (GL), como sugerem, para as expressões temporais, Baptista (1999), Hagège et al. (2010) e Maurício (2011). Nestes casos, somente uma entrada lexical é registrada.

4 - Expressões temporais idiomáticas que denotam referências temporais como *no tempo das vacas gordas/magras*, *à hora das galinhas*[PB], *ao toque das ave-marias*.

(7) [Se a obra não foi realizada] *no tempo das vacas gordas* [, também não será feita agora] *no tempo das vacas magras* [SE]

(8) [O Leo se deita] *à hora das galinhas* ‘muito cedo’

(9) [Todos se recolhiam] *ao toque das ave-marias* ‘ao entardecer’

5 - Construções comparativas fixas idiomáticas, semelhantes às já estudadas por Ranchhod (1991) para o português, a partir dos trabalhos homólogos para o francês de Gross (1984, 1986), mas que sejam exclusivas do português brasileiro, como por exemplo:

(10) [… os candidatos falam como santos, mas se comportam] *como o diabo gosta* [SE]

(11) [O menino correu] *como uma flecha* [ao sentir que estava em perigo]

(12) [O Leo e a irmã vivem] *como cão e gato* (cf. *dão-se como o cão e o gato*)

[5] Os símbolos ‘*’ e ‘?’ indicam a inaceitabilidade da expressão ou a sua aceitabilidade duvidosa, respectivamente. O símbolo ‘°’ indica que a expressão é aceitável mas que o significado é diferente do que está em discussão.

(13) [O Leo vestiu-se impecavelmente], *como dita o figurino* (cf. *como manda o figurino*)

(14) [O Leo fuma] *como uma caipora*

(15) [A saudade amarga] *que nem jiló*

Resolvemos, porém, excluir os seguintes tipos de construções:

1 - Construções preposicionais e conjuncionais: algumas destas expressões têm valor adverbial, porém selecionam elementos livres e variáveis, por exemplo, *aos cuidados de*, *ao som de, em harmonia com.*

(16) [O Leo deixou os filhos] *aos cuidados de_*a enfermeira/a mãe/a avó/a tia

(17) [O Leo acordou] *ao som de_as* ondas/a viola/o violino/o vento

(18) [O Leo fez algo] *em harmonia com* a lei/a natureza/o universo

2 - Construções adverbiais associadas a frases com nomes predicativos e o verbo-suporte *estar*. Excluímos, assim, expressões como [estar] *com a corda no pescoço*, dentre outras, muitas delas já previamente estudadas por Ranchhod (1990) para o português europeu, na medida em que são analisadas como construções com verbo-suporte. Efetivamente, estas expressões dão frequentemente origem a modificadores adverbiais por redução do verbo-suporte (Ranchhod 1983; Ranchhod, 1990, pp. 90 ss.):

(19) [O Zé estava] *com a corda no pescoço*. [classe EPC; Ranchhod, 1990]

(20) [Depois de ter jogado] *com a corda no pescoço* [até abril, o Arsenal simplesmente perdeu o fôlego][6]

Do mesmo modo, foram excluídas as construções adverbiais com nomes predicativos que selecionam o verbo-suporte *ter* (e que por vezes apresentam construções equivalentes com *estar com*):

(21) [Sou português] *com muita honra*!

(22) Tenho muita honra de/em ser português. /É uma honra para mim ser português.

3.3. **Metodologia da descrição linguística**

Para fazer a descrição das expressões e aferição das propriedades linguísticas das construções adverbiais nas variedades europeia e brasileira do português, recorremos a dois tipos de fontes: (i) a intuição linguística de falantes nativos de cada variedade; juntamente com (ii) consultas em *corpora*. Efetivamente, a elicitação de juízos de aceitabilidade quanto à boa formação e interpretabilidade de exemplos construídos desempenha um papel essencial na descrição linguística (Laporte, 2015), enquanto o recurso de forma crítica a dados textuais obtidos de *corpora* de dimensões apreciáveis pode servir não só para validar essas intuições, como também suscitar novas direções de pesquisa.

Inicialmente, consultamos o CETEMFolha/NILC (Pinheiro & Aluísio, 2003),[7] um *corpus* com aproximadamente 24 milhões de palavras em textos jornalísticos do periódico Folha de São Paulo (1999), coletados em 1994 para o português brasileiro, e disponível na plataforma da Linguateca. Consultamos também

[6] https://www.flashscore.pt/noticias/futebol-premier-league-arsenal-perdeu-o-titulo-devido-aos-jogos-mentais-de-guardiola-e-a-falta-de-profundidade/YVh9iWjJ/ [consultado em 31/05/2023].
[7] https://www.linguateca.pt/cetenfolha/index_info.html

o CETEMPúblico (Rocha & Santos, 2000),[8] um *corpus* de aproximadamente 180 milhões de palavras em português europeu, também disponível na plataforma Linguateca. Ao analisarmos os dados disponíveis, principalmente no NILC, notamos que somente algumas das expressões daquelas que coletamos apareciam nas buscas, o que nos leva a crer que o volume deste *corpus* não era suficiente para o escopo da nossa pesquisa.

Tivemos depois acesso ao *corpus ptTenTen18* (Kilgarrif et al., 2014; Wagner et al., 2018), um *corpus* de grandes dimensões (precisamente 5,542,074,775 tokens para o português brasileiro, e 206,869,477 tokens para o português europeu), disponível na plataforma do Sketch Engine. Aí verificamos que praticamente todas as buscas de expressões apresentavam resultados positivos, o que nos levou a considerar que este recurso já respondia à necessidade de validar em textos as expressões aqui estudadas, tornando-se, assim, uma ferramenta indispensável para o desenvolvimento deste trabalho. Outro ponto positivo do *corpus ptTenTen18*, além da abrangência e cobertura das variedades europeia e brasileira do português, é o fato de ser um *corpus* relativamente recente, que nos permite explorar expressões contemporâneas e observar alguns fenômenos interessantes, como, por exemplo, o uso de expressões tipicamente brasileiras em textos das mídias portuguesas e vice-versa.

Na seção seguinte, apresentamos os critérios de classificação formal, de classificação sintático-semântica e de representação das propriedades linguísticas das construções adverbiais recenseadas neste estudo.

4. **Descrição linguística**

Adotamos a perspectiva teórico-metodológica do Léxico-Gramática proposta por Maurice Gross (1975, 1981) e fundada na gramática transformacional de operadores de Zellig S. Harris (1991). De acordo com essa perspectiva, a *frase elementar*, constituída pelo elemento predicativo e seus argumentos – sujeito e eventuais complementos essenciais, é considerada a unidade mínima de análise linguística. Gross (1984, p. 275) afirma que somente numa frase elementar se pode determinar claramente as propriedades sintáticas e o significado preciso das expressões linguísticas. Consequentemente, a classificação dos advérbios, tanto simples como compostos, é estabelecida com base nas propriedades sintáticas dessas unidades lexicais, dentro do contexto da frase elementar.

4.1. **Classificação formal**

Gross (1986, p. 12) introduz o conceito de *advérbio generalizado* (em francês *adverbe généralisé*), que engloba tanto expressões composicionais e sintaticamente analisáveis quanto expressões fixas, cristalizadas, não-composicionais e idiomáticas. O conceito de advérbio generalizado inclui as seguintes estruturas:

(i) advérbios simples, incluindo os advérbios derivados (em *-mente*):

(23) [O Leo trabalha] *diariamente*

(ii) complementos preposicionais circunstanciais (adjuntos), geralmente de estrutura morfossintática livre:

(24) [O Leo trabalha] *com afinco*/*nos feriados*/*em casa*/*à beça*[PB]

(iii) orações subordinadas circunstanciais:

[8] https://www.linguateca.pt/CETEMPublico/

(25) [O Leo trabalha] *enquanto eu descanso*/*enquanto o diabo esfrega um olho*

Gross (1986) também propõe uma organização taxonômica das expressões adverbiais compostas que é determinada com base em sua sequência interna de categorias gramaticais. Com base nessa estrutura, Gross (1986) classifica as expressões adverbiais do francês em 16 classes, dentre as quais utilizamos 10 para a formalização das expressões em português, dada a semelhança entre as estruturas das expressões do francês e do português. Palma (2009) também formalizou os advérbios compostos do português europeu fazendo uso destes critérios. A Tabela 1 apresenta de forma compacta esta classificação.

As classes são identificadas por siglas que indicam a estrutura interna de cada tipo de advérbio composto. Por exemplo, **PAC** é uma expressão composta tipicamente por uma *preposição*, um *adjetivo* e um *nome*, como se indica na coluna da estrutura interna do composto; e.g., *de má vontade*: [*O Pedro fez isso*] *de má vontade*. Cada classe é ilustrada por um exemplo. Os valores da coluna **PT**, indicam advérbios que são exclusivos da variedade europeia, seguidos da porcentagem em cada uma das classes. Igualmente, os valores da coluna **BR** indicam expressões exclusivas da variedade brasileira, e sua respectiva porcentagem nessa classe. Finalmente, a coluna **PTBR** indica o número de expressões comuns às duas variedades e a porcentagem dessa classe. Na coluna **Total** apresenta-se a soma dos valores PT+BR+PTBR e a porcentagem do total das expressões desta classe relativamente ao total de entradas do léxico. A última linha apresenta o número de expressões e porcentagem do léxico de cada variedade (ou comum a ambas as variedades).

Tabela 1. Classificação dos advérbios compostos em português europeu e brasileiro. As classes são indicadas por códigos convencionais

Classe	Estrutura Interna	Exemplos	PT	%	BR	%	PTBR	%	Total	% total
PC	*Prep C*	*em vão*	79	8%	451	46%	441	46%	971	29%
PDETC	*Prep Det C*	*pelo menos*	112	16%	236	32%	381	52%	729	22%
PAC	*Prep Adj C*	*de má vontade*	27	11%	109	44%	113	45%	249	7%
PCA	*Prep C Adj*	*por maioria absoluta*	52	15%	104	33%	158	52%	314	9%
PCDC	*Prep C1 de C2*	*por conta da casa*	54	21%	105	41%	98	38%	257	8%
PCPC	*Prep C1 Prep C2*	*da cabeça aos pés*	60	16%	140	39%	166	45%	366	11%
PCONJ	*Prep C1 Conj C2*	*em verso e prosa*	14	6%	87	36%	138	58%	239	7%
PF	*Frase fixa*	*dito isso*	5	5%	63	67%	26	28%	94	3%
PV	*Prep V W*	*até dizer chega*	1	4%	10	40%	14	56%	25	1%
PJC	*Conj C*	*e por aí vai*	3	3%	61	69%	24	28%	88	3%
		Total	**407**	**12%**	**1.366**	**41%**	**1.559**	**47%**	**3.332**	

Nota. Na representação da estrutura interna, usa-se a notação seguinte: Adj – adjetivo, Conj – conjunção, Det – determinante, Prep – preposição; C – indica um elemento nominal fixo da combinação; W – indica qualquer sequência (não especificada, eventualmente nula) de complementos.

Para fins desta pesquisa, determinamos que excluiríamos certas classes propostas por Gross (1986), nomeadamente, as classes **PCDN**, que inclui construções como *à margem de*, *por ocasião de*, e a classe **PCPN**, que inclui construções como d*e acordo com*, *em relação a*, por entendermos que se trata de expressões preposicionais e conjuncionais, na medida em que introduzem um elemento distribucionalmente livre, as primeiras exclusivamente um grupo nominal (e.g., *à margem de*_as negociações) e as segundas podendo introduzir uma oração subordinada ou um grupo nominal de conteúdo proposicional/nome predicativo (e.g., *Em relação a essa tarefa/a comprar os medicamentos, o Leo sabe o que fazer*), apesar de se reconhecer o estatuto adverbial de todo o constituinte que estas locuções introduzem.

Além destas, também deixamos de fora as classes de construções comparativas (Gross, 1984): **PECO**, que incluem construções comparativas modificadoras de um adjetivo, como, por exemplo, [*ser magro*] *como um palito*; **PVCO**, que são construções comparativas que modificam verbos, como, por exemplo, [*fumar*] *como uma chaminé*; e, finalmente, a classe **PPCO**, formada por expressões comparativas introduzidas por preposição, e.g., [*surgir*] *como* (*que*) *por encanto.* Estas expressões já foram previamente descritas para o português europeu (Ranchhod, 1991) e também revisadas e formalizadas no desenvolvimento do sistema STRING, como mencionado anteriormente. Consideramos e acrescentamos ao léxico somente expressões que são exclusivas da variedade brasileira, como, por exemplo, [*fumar*] *como uma caipora* 'muito' (a *caipora* é uma entidade mitológica da cultura brasileira).

Finalmente, tratando-se de um processo de recenseamento em curso, os números apresentados na Tabela 1 devem considerar-se em permanente atualização.

4.2. **Propriedades sintáticas e semânticas**

Para apresentar um exame detalhado do funcionamento dos advérbios compostos quando integrados em frases, adotamos, genericamente, a classificação sintático-semântica proposta por Molinier e Levrier (2000). O estudo destes autores concentrou-se na classificação de advérbios derivados terminados em *-ment* em francês. Acreditamos que esta estrutura de classificação não só se aplica genericamente aos advérbios correspondentes, derivados e terminados em *-mente* em português (já previamente estudados por Fernandes (2011), embora apenas para um conjunto de advérbios do português do Brasil); mas também pode ser estendida e adaptada para a descrição das construções das expressões adverbiais compostas.

Em essência, Molinier e Levrier (2000) estabeleceram um conjunto de critérios para distinguir diferentes classes de construções adverbiais com base na relação que estes elementos estabelecem com a frase de base (ou frase elementar) em que se encontram e/ou com constituintes dessa frase. Define-se, assim, os dois tipos principais de construções adverbiais, (P) e (M):

(P) advérbios modificadores *externos* das frases; e
(M) advérbios modificadores dos elementos *internos* das frases.

Seguindo os princípios metodológicos do Léxico-Gramática (Gross 1975, 1981), a determinação das propriedades linguísticas (sintático-semânticas) dos elementos lexicais (neste caso, dos advérbios) só se pode fazer adequadamente no quadro sintático das frases de base (ou frases elementares) da gramática. As frases elementares são concebidas, nesta perspetiva, como a expressão linguística de um predicado semântico reduzido à sua expressão mais simples (com os seus elementos essenciais expressos). Como os advérbios constituem (geralmente) *operadores de segunda ordem* -- no sentido de Harris (1991), ou seja, operam sobre o resultado de outro operador) -- o contexto sintático adequado para determinar as suas propriedades é uma frase elementar. Vejamos os exemplos a seguir, para ilustrar estes conceitos:

(26) *No entanto*, [o Leo acredita que possa ganhar o jogo]

(27) [O Leo chegou à festa] *de mãos abanando*/*de mãos a abanar*

Para determinar se os advérbios são modificadores *internos* ou *externos* à proposição, Molinier e Levrier (2000) propõem as seguintes duas propriedades, (A) e (B):

A. Topicalização do advérbio e negação do predicado principal:

(26a) *No entanto*, [o Leo **não** acredita que possa ganhar o jogo]

(27a) **De mãos abanando* / *de mãos a abanar*, [o Leo **não** chegou à festa]

B. Clivagem (ou extração *ser...que*):

(26b) ***É** *no entanto* **que** [o Leo acredita que possa ganhar o jogo]

(27b) **Foi** *de mãos abanando* / *de mãos a abanar* **que** [o Leo chegou à festa]

De modo geral, os modificadores externos das frases podem ser *topicalizados* quando a frase é colocada na forma negativa, pois eles operam sobre a frase inteira e não são por isso afetados pela negação do predicado dessa frase. No entanto, eles não podem sofrer a *clivagem*, que é uma operação reservada aos constituintes internos de uma frase. Por outro lado, os advérbios internos à frase geralmente não podem ser topicalizados quando esse predicado se encontra sob uma negação; mas podem ser extraídos. Estas duas propriedades são, em grande medida, independentes. A classificação das construções adverbiais isola, em primeiro lugar, os advérbios modificadores externos à frase (P), por apresentarem *coordenação* da propriedade (A) com a *inaceitabilidade* da propriedade (B): P = A+¬B. Todas as outras possibilidades são classificadas como advérbios *internos* da frase (M).

Assim, nos exemplos acima, com o advérbio *no entanto* (26a), a topicalização não afeta a relação entre este advérbio e o predicado, pelo que a modificação que o advérbio exerce é independente da negação. Por outro lado, em (26b), a clivagem não pode ser aplicada a esse advérbio, o que demonstra não se tratar de um modificador/complemento do verbo *acreditar* ou de outro elemento dessa frase de base. O advérbio funciona, portanto, como um modificador externo da frase.

Em contrapartida, a expressão adverbial *de mãos abanando* / *de mãos a abanar* não admite a *topicalização* com a frase na forma negativa, como visto em (27a) mas permite a *clivagem* (ou extração *ser…que*), como se vê em (27b). Trata-se, portanto, de um advérbio modificador interno à frase.

Na classificação proposta pelos autores, a esta divisão de ordem mais geral (P ou M), vem juntar-se uma subclassificação mais fina, que distingue diferentes classes de construções adverbiais. Quanto aos advérbios que operam como modificadores externos da frase, consideram-se três classes principais:

(i) Advérbios conjuntivos (**PC**)

Estes advérbios têm como característica principal a função de conectar a frase em que se encontram à frase anterior e, em consequência disso, nunca ocorrem em início absoluto de discurso, por exemplo, *consequentemente, portanto, a propósito, em suma*, *não obstante, pelo contrário*, entre outros:

(28) *Consequentemente*/*portanto*/*a propósito*/*em suma*/*não obstante*/*pelo contrário*, [isso não é assim]

Esta classe reúne alguns subtipos de advérbios, agrupados por seus valores semânticos, e tendo como critério principal o fato de serem conectores, portanto não existe uma subclassificação ou organização especial. Na proposta de Molinier e Levrier (2000), alguns dos subtipos conjuntivos são: (a) enumerativos (*primeiramente*, *em seguida*); (b) justificadores (*pois*, *com efeito*, *aliás*); (c) transicionais (*a propósito*, *de resto*, *por esta ordem de ideias*) (d) reformulativos (*em suma*, *em conclusão*); (e) apositivos (*a saber*, *quer dizer*, *isto é*, *por exemplo*), entre outros.

Por oposição aos advérbios conjuntivos, Molinier e Levrier (2000) definem outro conjunto de advérbios de frase a que chamam *disjuntivos* (termo sem qualquer relação com a operação lógica). Neles distinguem em primeiro lugar a classe dos:

(ii) Advérbios disjuntivos de estilo (**PS**)

Esta classe de advérbios define-se por operar sobre o operador metalinguístico harrissiano *eu digo* (Harris 1991, p. 141), e que corresponde àquilo a que Molinier e Levrier (2000, p. 65) designam por "un verbe du type « dire » placé dans une phrase supérieure" (um verbo do tipo "dizer" colocado numa frase superior). Estes advérbios são, pois, modificadores de modo particularmente apropriados a verbos *de dizer* «*verba dicendi*», (Baptista, 2010) ou de expressões nominais como *termos*, *palavras*, podendo ser identificados pela paráfrase com uma construção desse tipo:

(29) Sinceramente/numa palavra/ por outras palavras, [é um erro fazer isso]

(30) Eu digo sinceramente/numa palavra/ por outras palavras que [é um erro fazer isso]

Trata-se, portanto, de advérbios que estabelecem algum tipo de relação entre o locutor e o destinatário ou o próprio enunciado. Molinier e Levrier (2000, p. 65 ss.) identificam ainda, neste conjunto, vários conjuntos, de que se destacam: (a) os de atitude do locutor relativamente ao destinatário (*sinceramente*, *com toda a sinceridade*); (b) os que caracterizam a forma do enunciado (*concretamente*, *em concreto*); (c) a origem do enunciado (*oficialmente*, s*egundo N*, *de acordo com N*); (d) expressão do ponto de vista do locutor e implicação no próprio enunciado (*pessoalmente*, *no que me diz respeito*, *da*/*pela minha parte*); entre outros.

Para os restantes advérbios disjuntivos, a que chamam *de atitude* (**PA**), os autores procedem a uma subclassificação, sem que, contudo, indiquem um critério que os distinga, no seu conjunto, destes advérbios disjuntivos de estilo (**PS**). Haveria, talvez, que considerar uma classificação que opusesse os advérbios *conjuntivos* (**PC**) dos restantes ("disjuntivos", não seria um termo muito motivado) e, entre estes, considerar uma única classe, com critérios particulares para a identificação de cada subclasse. No entanto, por uma questão de clareza, mantemos aqui a estrutura taxonômica e a nomenclatura proposta por Molinier e Levrier (2000). Assim, na grande classe dos:

(iii) Advérbios disjuntivos de atitude (**PA**), que se dividem em quatro subclasses:

— Advérbios *de hábito* (**PAh**)

Podem ser definidos pela seguinte propriedade: eles só são compatíveis com o presente e o imperfeito do indicativo, em sua interpretação aspectual **habitual**. Como se sabe, o presente e o imperfeito se prestam a duas grandes interpretações aspectuais: a interpretação referencial (ou de evento) e a interpretação habitual. Os advérbios de *hábito*, de que é modelo o advérbio *habitualmente*, não devem ser confundidos com os advérbios de *frequência* (uma das subclasses dos advérbios de tempo), de que é modelo o advérbio *frequentemente*, nomeadamente por serem modificadores de toda a frase e não poderem, por isso ocorrer destacados no início de frase com o predicado desta na negativa:

(31) *Habitualmente*, [o Leo (não) bebe/bebia álcool]

(32) **Frequentemente*, [o Leo (*não) bebe/bebia álcool]

Estes advérbios de hábito são, pois, incompatíveis com os tempos-modos verbais que denotam um aspecto pontual/perfectivo, nomeadamente os pretéritos perfeito, mais-que-perfeito e perfeito composto:

(33) *Habitualmente, [o Leo bebeu/tinha bebido/bebera álcool]

Alguns exemplos de advérbios compostos deste tipo:

(34) *Por norma* / *por hábito* /*em regra* (*geral*) [o Leo (não) bebe/bebia álcool]

— Advérbios avaliativos (PAa)

Estes advérbios, como o nome indica, exprimem uma avaliação subjetiva por parte do enunciador relativamente ao conteúdo do enunciado e em particular a percepção do caráter favorável ou desfavorável desse conteúdo. Alguns exemplos incluem *felizmente*, *curiosamente* e *surpreendentemente*; entre os compostos, *por sorte*, *por estranho que pareça*, *por um acaso*. Dado exprimirem um juízo do falante sobre uma proposição, estes advérbios não podem ser usados em frases interrogativas nem nas imperativas, como nos exemplos:

(35) *Curiosamente*/*por estranho que pareça*, [o Leo (não) bebe álcool]

(35a) **É curiosamente*/*por estranho que pareça*, que [o Leo não bebe álcool]
(35b) Eu digo que *curiosamente*/*por estranho que pareça* [o Leo não bebe álcool]
(35c) **Curiosamente*/*por estranho que pareça*, [o Leo não bebe álcool?]
(35d) **Curiosamente*/*por estranho que pareça*, [não beba álcool, Leo!]

Alguns destes advérbios aceitam a paráfrase por um verbo dito de *percepção mental* (Baptista, 2013; Baptista & Mamede, 2020; classe 06) como *achar*, *considerar*, etc

(35e) *Eu acho* curioso/estranho *que* [o Leo (não) beba álcool]

— Advérbios modais (PAm)

Os advérbios desse grupo, também chamados de advérbios *assertivos* ou de *modalização da asserção* (Borillo, 1976) constituem uma classe relativamente heterogênea mas, como o nome indica, modificam globalmente uma frase conferindo-lhe uma determinada modalidade, concretamente, a atitude/perspectiva do locutor relativamente ao conteúdo da proposição, como, por exemplo, a verossimilhança, probabilidade, dúvida, etc. São exemplos típicos desta classe:

(36) *Certamente* / *fatalmente* / *indubitavelmente* / *provavelmente* /*presumivelmente,* [o Leo chegará hoje].

(37) *Salvo melhor opinião*[PT]*/juízo*[PB] / *até prova em contrário*[PT] / *sem dúvida* (*alguma*) / *em hipótese alguma* [o Leo fará isso].

Estes advérbios podem ser caracterizados por não aceitarem que a frase seja colocada nos modos imperativo (a) ou exclamativo (b) ou numa interrogativa (c), dado já se encontrar modalizada pelo advérbio:

(36a) **Certamente / fatalmente / indubitavelmente* / *provavelmente* /*presumivelmente,* [faz/faça isso, Leo]!

(37a) **Salvo melhor opinião*[PT]*/juízo*[PB] */ até prova em contrário*[PT] */ sem dúvida* (*alguma*) / *em hipótese alguma,* [faz/faça isso, Leo]!

(36b) **Certamente / fatalmente / indubitavelmente* / *provavelmente* / *presumivelmente,* [Leu fez isso]!

(37b) **Salvo melhor opinião*[PT]*/juízo*[PB] */ até prova em contrário*[PT] */ sem dúvida* (*alguma*) / *em hipótese alguma,* [Leu fez isso]!

(36c) **Certamente / fatalmente / indubitavelmente* / *provavelmente* /*presumivelmente,* [Leo chegará hoje]?

(37c) **Salvo melhor opinião*[PT]*/juízo*[PB] */ até prova em contrário*[PT] */ sem dúvida* (*alguma*) / *em hipótese alguma,* [Leo fará isso]?

— Advérbios de frase orientados para o sujeito (PAs)

Trata-se de um conjunto de expressões, definido para os advérbios franceses terminados em *-ment*, que, além de se comportarem como a generalidade dos advérbios de frase (possibilidade de topicalização do advérbio com a frase na negativa e inaceitabilidade da extração *ser…que*), apresentam ainda um escopo sobre o sujeito, geralmente parafraseadas pela construção adjetival associada ao advérbio (38b). Em português, podemos referir, por exemplo, advérbios como *inteligentemente*:

(38) *Inteligentemente*, [o Leo (não) instalou o computador]
(38a) ºFoi *inteligentemente* que [o Leo (não) instalou o computador]
(38b) O Leo foi *inteligente* em [(não) instalar o computador]

A frase do exemplo (38a) é, obviamente, aceitável, mas o seu significado é subtilmente diferente (o que se assinala com o símbolo 'º', tratando-se, então, de um modificador do verbo (advérbio de modo orientado para o sujeito; ver adiante). Nesta construção, a frase de base não pode ser negada, já que o advérbio indica o modo como se fez a ação (a instalação do computador). No significado que nos interessa aqui, o locutor exprime uma opinião sobre a ação descrita na frase e, simultaneamente, faz uma apreciação sobre o respectivo sujeito. Nesse sentido, as frases adjetivais associadas às construções destes advérbios admitem ser encaixadas sob um verbo de **opinião**/**percepção mental**, tais como *achar*, *considerar*, etc.

(38c) Eu acho que Leo foi *inteligente* [em (não) instalar o computador]

Por essa razão, estas construções não admitem que a frase seja colocada nos modos imperativo (38d) ou exclamativo (38e) ou numa interrogativa (38f):

(38d) **Inteligentemente*, [instala/instale o computador, Leo]!
(38e) **Inteligentemente*, [o Leo (não) instalou o computador]!
(38f) **Inteligentemente*, [o Leo (não) instalou o computador]?

dado que as modalidades imperativa, exclamativa e a formação da interrogativa são incompatíveis com a expressão de uma opinião sobre todo o conteúdo da frase. Até ao momento, não se encontrou exemplos de advérbios compostos desta classe.

Quanto aos advérbios modificadores internos da proposição, estes podem apresentar qualquer outra configuração das propriedades definitórias -- exceto a que determina os advérbios de frase, ou seja, apresentam comportamentos sintáticos variados relativamente à topicalização com a frase na negativa ou a extração *ser…que*.

Estes advérbios são classificados segundo Molinier e Levrier (2000) em seis grandes classes:

(i) os advérbios de modo (**MV**), e.g., *à moda antiga*, *lentamente*, *de mala e cuia*[PB]

Essa classe de advérbios é, sem dúvidas, a mais numerosa, e estes respondem adequadamente à interrogativa *como*? São definidos por meio de quatro propriedades, nomeadamente: (1) a impossibilidade de ocorrer destacado na posição inicial de uma frase na negativa (39); (2) a possibilidade de aplicar a extração *ser…que* (39a); (3) a possibilidade de associar o adverbial ao advérbio interrogativo *como?* (39b); (4) no caso dos advérbios simples terminados em *-mente*, a impossibilidade de associar à frase em que estes aparecem uma frase predicativa na qual o adjetivo de que é derivado o advérbio aparece a qualificar o sujeito, quando este é humano (O *Pedro procurou diligentemente o livro*# *O Pedro foi diligente* (*a procurar o livro*). Esta propriedade permite distinguir os advérbios de modo modificadores do verbo (classe **MV**) dos advérbios de modo orientados para o sujeito (classe **MS**, ver adiante); esta propriedade não se aplica portanto diretamente à classificação dos advérbios compostos; (5) impossibilidade de associação com um adverbial de quantidade, quer de completude (e.g., *completamente*) quer de extensão qualitativa (e.g., *predominantemente*); esta propriedade permite diferenciar os advérbios de modo dos de quantidade (39c).

(39) **De mala e cuia*,[o Leo não se mudou para o Japão]
(39a) **Foi** *de mala e cuia* **que** [o Leo se mudou para o Japão] 'definitivamente, com todos os seus pertences'

(39b) P: **Como** foi que [o Leo se mudou para o Japão]? R: de mala e cuia
(39c) [*O Leo se mudou para o Japão*] ?*/º*completamente*/**predominantemente*

(ii) os advérbios orientados para o sujeito (**MS**), e.g., *de bom grado*, *aos berros*, *do fundo do coração*

Estes advérbios são definidos através de três propriedades, nomeadamente: (1) a inaceitabilidade na posição inicial e destacada de uma frase negativa (40); (2) a possibilidade de extração com *ser…que* (40a); (3) como os restantes advérbios de modo, estes podem também ser associados ao advérbio interrogativo *como*? (40b);

(40) **Aos berros*, [o Leo não entrou no carro]
(40a) ***Foi** aos berros **que*** [o Leo entrou no carro]
(40b) P: **Como** foi que [o Leo entrou no carro]? R: *aos berros*

Tal como se disse atrás, no caso dos advérbios simples terminados em *-mente*, distingue-se no seio dos advérbios de modo, um subconjunto, que autoriza uma paráfrase adjetival com sujeito humano (e.g., [O Pedro entrou no carro] *cuidadosamente # O Pedro foi cuidadoso* [a entrar no carro]): esta propriedade não se aplica diretamente, portanto, à classificação dos advérbios compostos. Contudo, é possível conceber uma adaptação e, assim, generalizar este critério a outras formas linguísticas, incluindo as expressões fixas idiomáticas aqui tratadas. Assim, no caso de o advérbio envolver um nome predicativo (e.g., *berro*), é possível construir a frase de base desse nome em que se reencontra o sujeito (humano) da construção analisada, e.g., [*O Pedro entrou no carro*] *aos berros # O Pedro deu* (*uns*) *berros* (sobre as construções com verbo-suporte *dar*, ver, entre outros, Baptista, 1997; Rassi, 2015; Vaza, 1988). Noutros casos, a possibilidade de inserção de um pronome possessivo demonstra a correferência restrita entre este e o sujeito da construção, e.g., [Nós perdoamos o João] *do fundo do nosso*/**seu coração*.

(iii) os advérbios de tempo **(MT**), e.g., *na calada da noite*, *todo santo dia*

Os advérbios de tempo podem ser verificados a partir das seguintes propriedades: (1) a aceitabilidade em início destacado de uma frase negativa (41); (2) a possibilidade de extração *ser…que* (41a); (3) a possibilidade de responderem adequadamente à interrogativa com o advérbio interrogativo *quando?* (41b):

(41) *No tempo das vacas magras*, [o Leo não investiu nessa empresa]
(41a) ***Foi** no tempo das vacas magras **que*** [o Leo investiu nessa empresa]
(41b) P: **Quando** (é que) [o Leo investiu nessa empresa]? R: *No tempo das vacas magras*

Os advérbios de tempo subclassificam-se, por sua vez, em expressões de **data** (MTd, como o exemplo acima, *no tempo das vacas magras*), de **duração** (MTu, *o dia todo*) ou de **frequência** (MTf, *todo santo dia*[PB]).

(iv) os advérbios de ponto de vista (**MP**), e.g., *em teoria, de direito*, *em tese*

Estes advérbios podem ser verificados através das seguintes propriedades: (1) a possibilidade de ser aceito em início destacado de uma frase negativa (42); (2) a possibilidade de se aplicar, em geral, a extração *ser…que* (42a); (3) no caso dos advérbios derivados e terminados em *-mente*, a possibilidade de paráfrase do advérbio por um adjetivo a modificar o nome *ponto de vista* (42b); *mutatis mutandis*, esta propriedade pode ser estendida a várias locuções, incluindo algumas expressões idiomáticas aqui tratadas:

(42) *Em teoria*, [o Leo (não) fez o trabalho]
(42a) ?***Foi** em teoria **que*** [o Leo fez o trabalho]
(42b) *Em teoria*/*do ponto de vista teórico/teoricamente*, [o Leo fez o trabalho]

Como definem Molinier e Levrier (200, pp. 222) estes advérbios "restringem o domínio sobre o qual uma afirmação é válida ou verdadeira". Por esse motivo, estas construções admitem uma continuação de significado contraditório, desde que essa seja restringida por um advérbio de ponto de vista diferente:

(42c) *Em teoria*, [o Leo (não) cometeu o crime] mas, *na prática*, [foi ele quem tudo fez].

Neste sentido, dado só se combinarem com proposições, as construções com MP não aceitam o imperativo:

(42d) *Teoricamente/em teoria/de um ponto de vista teórico, [Faz/Faça o trabalho, Leo]!

Esta adaptação nos permite incluir nesta classe advérbios quase-sinônimos como *em tese*, mesmo que estes não possam ser associados a *ponto de vista.*

(vi) os advérbios quantificadores (**MQ**), e.g., *à beça*[PB], *para dar e vender*

Estes advérbios funcionam como quantificadores sobre um predicado e essa construção pode ser verificada a partir de três propriedades: (1) inaceitabilidade de ocorrerem destacados em início de frase na negativa (43a); (2) a possibilidade de, na grande maioria, admitirem a extração *ser…que* (43b); (3) a possibilidade de associar o adverbial a uma frase interrogativa com o advérbio *muito* (43c):

(43) [O Leo chorou] *à beça*
(43a) **À beça*, [o Leo não chorou]
(43b) ***Foi** à beça **que*** [o Leo chorou]
(43c) P: [O Leo chorou **muito**]? R: Sim, [ele chourou] à beça

Os advérbios quantificadores foram organizados por Molinier e Levrier (2000) em três subclasses: (a) os advérbios de *intensidade* (MQi), como *à beça*, acima; (b) os advérbios de *completude* (MQc), *por completo*, *de todo, em parte*); e (c) os advérbios de *extensão qualitativa* (*tipicamente, verdadeiramente*). Dada a sua diversidade e complexidade de comportamentos sintáticos, uma descrição pormenorizada desta classe terá de ficar para outro momento.

(vii) advérbios focalizadores (**MF**), e.g., *em especial*, *especialmente*

Estes advérbios formam uma classe caracterizada pelas seguintes propriedades: (1) inaceitabilidade de ocorrerem destacados em início de uma frase na negativa (44a); (2) a impossibilidade de se aplicar a extração *ser…que* (44b); e (3) a possibilidade de extração do advérbio na companhia de um grupo nominal ou outro constituinte maior de uma frase (sujeito, complemento) (44c):

(44) [O Leo (não) gostou] *em especial/especialmente* [da sopa]
(44a) **Em especial/especialmente*, [o Leo (não) gostou da sopa]
(44b) *Foi *em especial/especialmente* que [o Leo (não) gostou da sopa]
(44c) Foi ?*em especial/especialmente* da sopa que [o Leo (não) gostou]

A estas classes, acrescentamos a proposta de uma classe para:

(viii) os advérbios locativos (**ML**), e.g., *à beira-mar, onde o vento faz a curva*[PB]

Apesar da natureza sistemática da classificação dos advérbios derivados terminados em *-ment* para o francês de Molinier e Levrier (2000), verificamos que estes autores não constituem uma classe específica para acolher os advérbios locativos. Ora, tendo os advérbios locativos sido mencionados em vários trabalhos

anteriores (Bechara, 2009; Costa, 2008, p. 44; Raposo, 2013), propomos, como parte desta pesquisa, a constituição de uma classe de *advérbios locativos* compostos, com o intuito de preencher esta lacuna na classificação das construções adverbiais que nos serviu de base. De um modo geral, ainda que marcadamente idiomáticos, estes advérbios podem ser parafraseados por *onde?* numa interrogativa, envolvendo construções verbais locativas, isto é, com um complemento essencial locativo:

(45) [Leo vive] em Lisboa / *onde o vento faz a curva*
(45a) P: **Onde** vive o Leo? R: em Lisboa / *onde o vento faz a curva*

A Tabela 2 apresenta as classes sintático-semânticas brevemente identificadas acima e o estado atual do seu recenseamento (e respectiva porcentagem):

Tabela 2. Classes semântico-sintáticas dos advérbios compostos em português

Classe	**Exemplos**	**Total**	**%**
PC (conjuntivos)	*afinal de contas*	213	0,065
PS (disjuntivos de estilo)	*com toda a franqueza*	52	0,015
PA (disjuntivos de atitude)	*em geral*	55	0,016
MV (modo)	*por amor à camisa*	2.071	0,633
MS (modo orient. sujeito)	*de boa vontade*	108	0,033
MT (tempo)	*ao romper do dia*	429	0,131
MP (ponto de vista)	*na prática*	6	0,001
MQ (quantitativos)	*aos montes*	162	0,049
MF (focalizadores)	*em especial*	19	0,005
ML (locativos)	*nos confins do mundo*	155	0,047
Total		**3.270**	

5. **Considerações finais e trabalhos futuros**

Este estudo é parte de uma pesquisa mais ampla que tem como objetivo descrever os advérbios compostos do português brasileiro quanto às suas propriedades sintáticas e semânticas. Após o recenseamento de expressões adverbiais, utilizando vários recursos (sobretudo dicionários), com o intuito de construirmos um léxico abrangente destes advérbios, começamos por descrever suas propriedades de acordo com os princípios teóricos e metodológicos do Léxico-Gramática. Tal consistiu em (a) atribuir-lhes uma classificação formal, com base na estrutura interna das expressões, definida a partir da sequência das categorias morfossintáticas por que são formadas; (b) classificá-las com base nas propriedades que estes advérbios apresentam relativamente às frases (elementares) que modificam; e (c) identificar a variedade do português (europeu ou brasileiro) em que ocorrem, recorrendo sobretudo a *corpora* de dimensões apreciáveis.

Até o momento, examinamos aproximadamente 3.300 advérbios compostos. Os exemplos abaixo (Figura 1) exemplificam a representação de algumas destas expressões na construção do léxico.

Figura 1. Algumas entradas do léxico dos advérbios compostos do português, no formato DELA (Paumier et al. 2021).

de ora em diante, .ADV+**PCPC+MTd+PT+BR+EN**= "from now on"
de orelha a orelha, .**ADV+PCPC+MV+PT+BR+EN**= "from ear to ear"
de pai para filho, .**ADV+PCPC+MV+PT+BR+EN**= "from father to son"
de papel passado, .**ADV+PCPC+MV+PT+BR+EN**= "on paper"
de passagem, .**ADV+PC+MV+PT+BR+EN**= "in passing"
de par em par, .**ADV+PCPC+MV+PT+BR+EN**= "(open) wide"
de pé em pé, .**ADV+PCPC+MV+BR+EN**= "slowly, carefuly"
de ponta a ponta, .**ADV+PCPC+MV+PT+BR+EN**= "end to end"
de porta em porta, .**ADV+PCPC+MV+PT+BR+EN**= "door to door"
de preferência, .**ADV+PC+PC+PT+BR+EN**= "preferably"
de propósito, .**ADV+PC+MV+PT+BR+EN**= "on purpose"
de quando em quando, .**ADV+PCPC+MTf+PT+BR+EN**= "from time to time"
de quando em vez, .**ADV+PCPC+MTf+PT+BR+EN**= "from time to time"

Os resultados deste recenseamento permitem desde já construir um mapa que dá uma dimensão aproximada das diferentes estruturas formais por que são constituídos os advérbios compostos (Tabela 1), a dimensão das diferentes classes sintático-semânticas (Tabela 2), bem como a distribuição destas expressões pelas duas variedades (portuguesa e brasileira) do português.

Algumas das limitações deste trabalho deverão, porém, ser referidas. Muitos advérbios compostos são idiomáticos e relevam de situações comunicativas próprias da oralidade. O recurso a *corpora* de textos escritos, nomeadamente para (a) determinar as propriedades sintático-semânticas dos advérbios e (b) quantificar a frequência da ocorrência destas expressões e daí tirar conclusões quanto à sua distribuição nas variedades do português, poderá, de algum modo, estar enviesado. Contudo, consideramos que o uso do *corpus TenTen18* como fonte poderá mitigar esta limitação, dado a sua extensão e abrangência de conteúdos.

Naturalmente, as restrições distribucionais que se observa entre os advérbios e os predicados que eles modificam, embora tenham sido parcialmente levadas em conta na descrição das entradas do léxico-gramática, nomeadamente pela sua ilustração com um exemplo característico e claro (isto é, não ambíguo), seja ele retirado do *corpus* ou de exemplos da internet, seja ele construído ou adaptado a partir de exemplos reais, é um trabalho que deverá ser desenvolvido de forma sistemática em futuros trabalhos.

Os próximos passos deste trabalho incluem, assim, a descrição mais precisa e detalhada das propriedades de cada uma das classes sintático-semânticas propostas por Molinier e Levrier (2000). Este trabalho já foi iniciado a partir da proposta de uma classe, que não tinha sido considerada no trabalho de Molinier e Levrier (2000), que descreve os advérbios locativos (ML). Seguiremos com a descrição minunciosa das demais classes, incluindo o desdobramento lexical de algumas entradas, uma vez que, ao decorrer do estudo, identificamos que algumas expressões podem, em determinados casos, ser classificadas em mais de uma classe. Por exemplo, a expressão adverbial *no tempo* possui valor temporal (classe MT), e.g., [O estudante terminou a prova] *no tempo* 'dentro do tempo previsto/destinado'; mas também um valor locativo (classe ML), como no exemplo [As crianças brincam] *no tempo*[PB] 'ao ar livre', exclusivo da variedade brasileira. Esse desdobramento de sentidos é essencial para diversas aplicações.

A recolha das expressões poderá ser prosseguida sistematicamente em outros recursos, já identificados, como por exemplo o dicionário de da Silva (2013), e que ainda não foram integralmente examinados.

Para algumas construções relativamente produtivas e que, por isso, se prestam mal a uma representação por listagem, e.g., *em DET circunstância*: *nesta/nessa/naquela circunstância*, *nestas/nessas/naquelas circunstâncias*, será necessário construir um conjunto bastante apreciável de gramáticas locais, por forma a representá-las adequadamente nos léxicos computacionais existentes, especificamente nos léxicos do sistema

de processamento computacional do português STRING (Hagège et al., 2010; Mamede et al. 2012; Maurício, 2011). Neste momento, estas expressões produtivas são representadas por uma única entrada no léxico-gramática. Pelo contrário, casos já identificados de forte fixidez morfossintática (e.g., *em circunstância alguma/nenhuma* vs. **em circunstâncias algumas/nenhumas*) foram representados no léxico-gramática como entradas independentes, e consideradas construções autônomas daquelas expressões produtivas. A análise dessas "famílias" de expressões e das condições que presidem às variações observadas será objeto de outro trabalho.

Pretendemos ainda determinar a frequência e a distribuição das expressões pelas duas variedades, brasileira e europeia, através de consulta em *corpora* e, se necessário, validação dessas observações com falantes nativos de cada variedade.

Esperamos, com o avanço deste estudo, contribuir para um maior conhecimento dos advérbios compostos em português, de forma a construir uma base sólida para futuros estudos, eventualmente para outras variedades do português.

Agradecimentos

A pesquisa para este trabalho foi parcialmente apoiada pelo programa de doutoramento em Ciências da Linguagem da Faculdade de Ciências Humanas e Sociais da Universidade do Algarve e através da Fundação para a Ciência e a Tecnologia, pelo INESC-ID Lisboa, Human Language Technology Laboratory (Ref. 50021/2021).

Referências

Baptista, Jorge (1997) *Sermão*, *tareia* e *facada*. Uma classificação das construções conversas *dar* - *levar*. *Seminários de Linguística* 1, pp. 5–37.

Baptista, Jorge (1999) *Manhã*, *tarde*, *noite*. Analysis of temporal adverbs using local grammars. *Seminários de Linguística* 3, pp. 1–27.

Baptista, Jorge (2010) *Verba dicendi*: *A structure looking for verbs*. In Takuya Nakamura, Éric Laporte, Anne Dister & Cédrick Fairon (orgs.), *Les tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*. Université Catholique de Louvain & Presses Universitaires de Louvain, pp. 11–20.

Baptista, Jorge (2013) ViPEr: Uma base de dados de construções léxico-sintáticas de verbos do Português Europeu. In *Actas do XXVIII Encontro da APL - Textos Selecionados*. APL & Colibri, pp. 111–129.

Baptista, Jorge & Nuno Mamede (2020) *Dicionário gramatical de verbos do português*. Universidade do Algarve Editora.

Bechara, Evanildo (2009) *Moderna gramática da língua portuguesa*. (37.ª ed.). Nova Fronteira.

Borillo, Andrée (1976) Adverbs and the modalization of assertion. *French Language* 30, pp. 74–89.

Català, Dolors (2003) *Les adverbes composés. Approches contrastives en linguistique appliquée*. Tese de doutoramento, Universitat Autònoma de Barcelona.

Català, Dolors, Jorge Baptista & Cristina Palma (2020) Problèmes formels concernant la traduction des adverbes composés (espagnol/portugais). *Langues & Parole* 5, pp. 67–82. https://doi.org/10.5565/rev/languesparole.64

Costa, João (2008) *O advérbio em português europeu.* Colibri.

Cunha, Celso & Luís Filipe Lindley Cintra (2017) *Nova gramática do português contemporâneo. de acordo com a nova ortografia* (7.ª ed.). Lexikon Editora Digital.

da Silva, José (2013) *Dicionário brasileiro de fraseologia* [em linha]. Disponível em http://www.josepereira.com.br/_/DBF_2013.pdf [consultado em 31/05/2023].

Fernandes, Gaia (2011) *Automatic disambiguation of* -mente *ending adverbs in Brazilian Portuguese*. Tese de mestrado, Universidade do Algarve / Universitat Autònoma de Barcelona.

Folha de São Paulo (1999) *CD-ROM Folha - Edição 99* [Texto integral da Folha de S. Paulo de 1994 a 1998]. Publifolha. Disponível em http://bd.folha.uol.com.br/cdrom.html
Gonçalves, Matilde, Luísa Coheur, Jorge Baptista & Ana Mineiro (2020) Avaliação de recursos computacionais em Português. *Linguamática* 12 (2), pp. 51–68. https://doi.org/10.21814/lm.12.2.331
Gross, Maurice (1975) *Méthodes en syntaxe*. Hermann.
Gross, Maurice (1981) Les bases empiriques de la notion de prédicat sémantique. *Langages* 63, pp. 7–52.
Gross, Maurice (1982) Une classification des phrases figées du français. *Revue québécoise de linguistique* 11 (2), pp. 151–185. Presses de l'Université du Québec à Montréal. https://doi.org/10.7202/602492ar
Gross, Maurice (1984) A linguistic environment for comparative romance syntax. In *Papers from the XIIth Linguistic Symposium on Romance Languages*, (4) 26, pp. 373–446.
Gross, Maurice (1986) *Grammaire transformationnelle du français : 3 - Syntaxe de l'adverbe*. ASSTRIL.
Gross, Maurice (1996) Lexicon-Grammar. In Keith Brown & Jim Miller (eds.), *Concise encyclopedia of syntactic theories*. Pergamon, pp. 244–259.
Guimier, Claude (1996) *Les adverbes du français : Le cas des adverbes en* -ment. Editions Ophrys.
Hagège, Carolina, Jorge Baptista & Nuno Mamede (2010) Caracterização e processamento de expressões temporais em português. *Linguamática* 2 (1), pp. 63–76.
Harris, Zellig S. (1991) *Theory of language and information: A mathematical approach*. Clarendon Press.
Kilgarriff, Adam, Miloš Jakubíček, Jan Pomikalek, Tony Berber Sardinha & Pete Whitelock (2014) PtTenTen: a *corpus* for Portuguese lexicography. In Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese corpora*, pp. 111–130. https://doi.org/10.5040/9781472593641.ch-006
Laporte, Éric, Takuya Nakamura & Stravoula Voyatzi (2008) A French *corpus* annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pp. 48–51. Disponível em https://shs.hal.science/halshs-00286541
Laporte, Éric, & Voyatzi, Stravoula (2008) An electronic dictionary of French multiword adverbs. In *Language Resources and Evaluation Conference (LREC). Workshop towards a shared task for multiword expressions*, pp. 31–34 Disponível em https://shs.hal.science/halshs-00286546
Laporte, Éric (2015) The science of Linguistics. *Inference: International Review of Science* 1 (2). https://inference-review.com/article/the-science-of-linguistics
Mamede, Nuno, Jorge Baptista, Cláudio Diniz & Vera Cabarrão (2012) STRING: A hybrid, statistical and rule-based natural language processing chain for Portuguese. In *Proceedings of the 10th International Conference on Computational Processing of Portuguese (PROPOR'12)*. Springer-Verlag, Disponível em http://www.inesc-id.pt/ficheiros/publicacoes/8578.pdf
Maurício, Andreia (2011) *Identificação, classificação e normalização de expressões temporais*. Tese de mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa.
Mel'čuk, Igor (2023) *General phraseology: Theory and practice*. John Benjamins Pub. Co.
Molinier, Christian, & Françoise Levrier (2000) *Grammaire des adverbes : Description des formes en* -ment. Librairie Droz.
Moreno-Ortiz, Antonio, Chantal Pérez-Hernández & Maria Del-Olmo (2013) Managing multiword expressions in a lexicon-based sentiment analysis system for Spanish. In *Proceedings of the 9th Workshop on Multiword Expressions*, pp. 1–10. Disponível em https://aclanthology.org/W13-1001.pdf
Palma, Cristina (2009). *Estudo contrastivo português-espanhol de expressões fixas adverbiais.* Tese de mestrado, Universidade do Algarve.
Paumier, Sébastien, Franz Guenthner, Eric Laporte, Friederike Malchok, Clemens Marschner, Claude Martineau, Cristian Martínez, Denis Maurel, Sebastian Nagel, Alexis Neme, Maxime Petit, Johannes Stiehler & Gilles Vollant (2021) *United 3.3. User Manual*. Paris: Unversité de Marne-la-Vallée. Disponível em https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf

Pinheiro, Gisele & Sandra Aluísio (2003) *Corpus NILC: Descrição e análise crítica com vistas ao projeto Lacio-Web*. São Paulo: Instituto de Ciências Matemáticas e de Computação – ICMC/USP. Disponível em http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Ramisch, Carlos (2015) *Multiword expressions acquisition. A generic and open framework.* Springer International Publishing.

Ranchhod, Elisabete (1983) On the support verbs *ser* and *estar* in Portuguese. *Lingvisticae Investigationes* 7 (2), pp. 317–353. https://doi.org/10.1075/li.7.2.07ran

Ranchhod, Elisabete (1990) *Sintaxe dos predicados nominais com* estar. INIC – Instituto Nacional de Investigação Científica.

Ranchhod, Elisabete (1991) Frozen adverbs—Comparative forms *Como C* in Portuguese. *Lingvisticæ Investigationes* 15 (1), pp. 141–170. https://doi.org/10.1075/li.15.1.07ran

Raposo, Eduardo (2013) Advérbio e sintagma adverbial. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do português* (Vol. 2). Fundação Calouste Gulbenkian, pp. 1569–1675.

Rassi, Amanda (2015) *Descrição, classificação e processamento automático das construções com o verbo DAR em Português Brasileiro*. Tese de doutoramento, Universidade Federal de São Carlos.

Rocha, Paulo Alexandre & Diana Santos (2000) CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *V Encontro para o processamento computacional da língua portuguesa escrita e falada* (*PROPOR 2000*). ICMC/USP, pp. 131–140. Disponível em http://www.linguateca.pt/documentos/RochaSantosPROPOR2000.pdf

Rocha, Carlos Alberto & Carlos Eduardo Rocha (2011) *Dicionário de locuções e expressões da língua portuguesa*. LEXIKON Editora.

Schwab, Artur (1985) *Locuções adverbiais* (2.ª ed). Fundação da Universidade Federal do Paraná.

Shigeto, Yutaro, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung & Yuji Matsumoto (2013) Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions*, pp. 139–144. Disponível em https://aclanthology.org/W13-1021.pdf

Shudo, Kosho, Akira Kurahone & Toshifumi Tanabe (2011) A Comprehensive Dictionary of Multiword Expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (HLT'11), pp.161–170. Disponível em https://aclanthology.org/P11-1017.pdf

Vaza, Aldina (1988) *Estruturas com nomes predicativos e o verbo-suporte* dar. Dissertação de mestrado, Universidade de Lisboa.

Wagner Filho, Jorge, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio (2018) The brWaC corpus: a new open resource for Brazilian Portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC 2018), pp. 4339–4344. Disponível em https://aclanthology.org/L18-1686.pdf

Žižková, Hana (2018) Improving compound adverb tagging. In *Recent Advances in Slavonic Natural Language Processing (RASLAN 2018)*, pp. 103–109. Disponível em https://nlp.fi.muni.cz/raslan/raslan18.pdf

(In)aceitabilidade do padrão de colocação dos pronomes clíticos no português brasileiro por falantes escolarizados

Ronan Pereira[1]

[1]Centro de Linguística da Universidade Nova de Lisboa (CLUNL)

Resumo

O português brasileiro prioriza a colocação pré-verbal (próclise) dos seus pronomes clíticos. No entanto, a colocação pós-verbal (ênclise) não foi excluída por completo dessa variedade, sendo adquirida por meio da escolarização. Este estudo objetivou avaliar a aceitabilidade das diferentes possibilidades de colocação dos clíticos em frases com e sem proclisadores – elementos que, segundo a tradição normativa (baseada no português europeu, cujo padrão de colocação observa restrições sintáticas), desencadeiam a próclise (ao passo que se obtém a ênclise na sua ausência) – por meio de uma tarefa de juízos de aceitabilidade com uma escala de seis pontos aplicado a falantes nativos do português brasileiro que tivessem concluído pelo menos o Ensino Médio. Os resultados demonstraram que, ainda que os itens com próclise tenham sido mais bem avaliados do que os que continham a ênclise, a colocação enclítica foi avaliada acima do ponto central da escala usada no estudo, indicando que é, também, uma colocação aceitável pelos falantes. Conclui-se que, embora o ensino da ênclise nas salas de aula apresente aos alunos regras similares àquelas que se observam na sintaxe do português europeu, não parece desenvolver neles o conhecimento sintático restrito a cada possibilidade de colocação dos clíticos, muito provavelmente pelo facto de o português brasileiro ser a variedade oral em uso nas escolas e as regras (e a sua aplicação) ficarem restritas à modalidade escrita.

Palavras-chave: português brasileiro, pronomes clíticos, colocação pronominal, teste de juízos de aceitabilidade.

Abstract

Brazilian Portuguese prioritizes pre-verbal placement (proclisis) of its clitic pronouns. However, post-verbal placement (enclisis) has not been completely excluded from this variety, being acquired by means of schooling. This study aimed at evaluating the acceptability of both these possible clitic placements in sentences with and without proclisis triggers – elements that, according to the normative tradition (based on European Portuguese, whose clitic placement pattern is syntacticly restrained), give rise to proclisis (whereas enclisis obtains at their absence) – by means of an acceptability judgement test with a six-point scale carried out by 79 Brazilian Portuguese native speakers who had completed, at least, Secondary School. The results showed that, even though the items with proclisis have been better evaluated than the ones with enclisis, the enclitic placement was evaluated above the central point of the scale used in this study, indicating that it is, as well, an acceptable placement for those speakers. It is concluded that, although the classes regarding clitic placement revolve around teaching the students similar rules to those found in the European Portuguese syntax, it does not seem to allow them to develop the syntactic knowledge specific to proclisis and enclisis, probably because Brazilian Portuguese is the oral variety used in the classrooms and the clitic placement rules (and their application) are restricted to the written modality.

Keywords: Brazilian Portuguese, clitic pronouns, clitic placement, acceptability judgement test.

1. Introdução

Uma das características mais marcantes que diferenciam o português brasileiro (PB) do português europeu (PE) tem que ver com o padrão de colocação dos pronomes clíticos em relação ao verbo da oração. Enquanto no PE um sistema complexo baseado em fatores sintáticos determina a colocação desses pronomes, o PB prioriza, em geral, a colocação pré-verbal (próclise) (Duarte, 2020; Luís & Kaiser, 2016; Morais & Ribeiro, 2005). No entanto, a colocação pós-verbal (ênclise) não foi excluída por completo do PB, sendo adquirida por meio da escolarização (Lobo, 2002). Neste sentido, a possibilidade da colocação enclítica, ainda que siga regras muito similares àquelas que descrevem o funcionamento da colocação pronominal no PE, parece emergir na modalidade escrita ou em contextos de alta monitorização da produção linguística, até mesmo em casos em que o PE não a aceitaria (*i.e.*, em contextos de próclise obrigatória) (Carneiro, 2005; Lacerda et al., 2021).

Tendo em vista a situação supracitada e considerando-se que grande parte dos estudos sobre a colocação dos clíticos em PB se debruçou sobre análises de *corpora* orais ou escritos, intencionou-se executar um estudo com vista a recolher dados empíricos acerca da aceitabilidade da ênclise em frases finitas com e sem proclisadores por falantes nativos do PB que tivessem completado, pelo menos, o Ensino Médio.[1] O contéudo deste artigo organiza-se, portanto, da seguinte forma: na Secção 2 descreve-se resumidamente o padrão de colocação dos pronomes clíticos nas variedades brasileira e europeia do português; na Secção 3 trazem-se pormenores acerca do ensino desses pronomes no Brasil e na Secção 4 expõem-se dados de alguns estudos que observaram o comportamento dos falantes nativos do PB relativamente à produção dos clíticos; na Secção 5 apresentam-se as questões de investigação e as hipóteses que nortearam este estudo; as Secções 6 e 7 trazem, respetivamente, a metodologia empregada e os resultados obtidos por meio dela; finalmente, na Secção 8 discutem-se os resultados e a Secção 9 traz as considerações finais.

2. O padrão de colocação dos pronomes clíticos em português

Martins (2013, p. 2231) ressalta que os clíticos são itens lexicais que possuem características próprias que os distinguem de afixos e palavras. Por exemplo, não podem ser coordenados e não possuem acento prosódico atribuído (diferentemente das palavras), mas têm uma posição mais livre na frase do que os afixos. Porém, apesar de não serem formas morfologicamente presas de facto, ocorrem sempre em adjacência a uma palavra hospedeira (Martins, 2013, p. 2231). No caso da língua portuguesa,[2] os clíticos pronominais ocorrem obrigatoriamente em adjacência a um verbo,[3] podendo ocorrer depois (ênclise), antes (próclise) ou entre o radical e o morfema de flexão verbal (mesóclise). Porém, constatam-se diferenças nesse padrão de colocação ao se compararem o PE e o PB (Duarte, 2020; Luís & Kaiser, 2016).[4]

Diferentemente de outras línguas românicas, que têm o padrão de colocação baseado na finitude do verbo (e.g., espanhol) (Soriano, 2015) ou que possuem a próclise generalizada, com a ênclise a ocorrer somente em formas imperativas (e.g., francês) (Kayne, 1975, p. 66), o PE possui como ordem básica em frases afirmativas neutras simples a ênclise (1).[5] A próclise emerge em certos contextos sintáticos, como, por exemplo, em frases

[1] Os 12 anos do ensino básico brasileiro estão divididos em Ensino Fundamental, correspondente aos nove primeiros anos de escolarização, e em Ensino Médio, correspondente aos últimos três anos.

[2] Embora este artigo se centre na variedade brasileira em comparação à europeia, evidências de variação no padrão de colocação dos clíticos também podem ser observadas nas variedades africanas do português. Ver Mapasse (2005) para o português de Moçambique e Mutali (2019) para o português de Angola, além de Martins (2021) para ambas as variedades.

[3] Excluem-se desta generalização os casos de interpolação detetados no PE (ocorrência de outro item entre o clítico e o verbo), considerados formas não padrão restritas a certos falantes ou variedades.

[4] Por questões de espaço, diferenças na morfologia dos pronomes objeto nas duas variedades não serão abordadas. Ver Duarte (2020), Luís e Kaiser (2016) e outros para este tópico.

[5] Por ser o foco deste estudo, a descrição apresentada ater-se-á aos casos de frases finitas. Uma descrição pormenorizada do padrão de colocação dos clíticos em frases infinitivas pode ser consultada em Martins (2013).

negativas (2), subordinadas (3), com alguns quantificadores (4) e alguns advérbios pré-verbais (5) e em interrogativas (6) ou exclamativas (7) introduzidas por uma palavra-qu:

(1) A Rita ajudou-**me** com a tarefa.[6]

(2a) A Rita não **me** ajudou com a tarefa.
(2b) Ninguém **me** ajudou com a tarefa.

(3a) Acho que a Rita **me** ajudará com a tarefa.
(3b) Quero que a Rita **me** ajude com a tarefa.
(3c) Terminei as tarefas cedo porque a Rita **me** ajudou.

(4) Todos **me** ajudaram com a tarefa.

(5) Já **me** ajudaram com a tarefa.

(6) Quem **me** ajudará com a tarefa?

(7) Como **me** ajudaram!

Já a mesóclise ocorre nos mesmos contextos de ênclise, mas está restrita a verbos conjugados no futuro do indicativo ou no condicional (Martins, 2013, 2016) (8a, b)[7]:

(8a) A Rita ajudar-**me**-á com a tarefa amanhã.
(8b) A Rita ajudar-**me**-ia com a tarefa amanhã se estivesse na cidade.

Quando comparado esse sistema complexo de colocação dos clíticos do PE[8] com o padrão de colocação do PB, observa-se uma diferença clara: nesta variedade o padrão de colocação é maioritariamente proclítico (Duarte, 2020; Luís & Kaiser, 2016). De facto, pelo menos superficialmente, a colocação nas duas variedades coincide nos contextos em que o PE exige a próclise (Exemplos 2 – 8), havendo diferenças somente em frases afirmativas simples (1) ou frases simples no futuro do indicativo (8a) ou no condicional (8b), que exigem a ênclise ou a mesóclise, respetivamente. Assim, reproduzindo-se os exemplos (1) e (8a, b) acima com o padrão de colocação do PB, somente se observa a próclise:

(9) A Rita **me** ajudou com a tarefa.

(10a) A Rita **me** ajudará com a tarefa amanhã.
(10b) A Rita **me** ajudaria com a tarefa amanhã se estivesse na cidade.

Os exemplos trazidos até aqui referem-se ao comportamento dos clíticos nas duas variedades no que tange a frases com apenas um verbo. Em complexos verbais, o PB admite a colocação do clítico em próclise ao verbo principal (11), o que parece indicar que funcionam, de certa forma, como prefixos, dada a generalização da colocação pré-verbal (Martins, 2002, p. 377); a mesma colocação é impossível no PE, no qual só pode ocorrer a ênclise (12):

[6] Optou-se por destacar em negrito os pronomes clíticos nos exemplos ao longo deste artigo.
[7] Este tipo de colocação tem vindo a diminuir, sendo preferencialmente substituída por perífrases verbais, e, diferentemente da ênclise e da próclise, só é adquirida durante a adolescência por intervenção do ensino escolar (cf. Santos, 2002).
[8] Salienta-se que o padrão de colocação também apresenta indícios de variação no PE em certos contextos sintáticos. Neste sentido, a variação dá-se, em geral, pelo uso da ênclise em contextos que exigem a próclise (cf. Martins, 2013).

(11) A Rita vai-**me** ajudar.

(12) A Rita vai ajudar-**me**.

Alternativamente, o PE também possibilita a subida do clítico,[9] caso em que o clítico se adjunge ao verbo finito do complexo. Neste caso, além da possibilidade de ênclise ao verbo no infinitivo (12), poderá ocorrer próclise (13a) ou ênclise (13b) ao verbo finito de acordo com as mesmas restrições sintáticas apresentadas anteriormente:

(13a) A Rita não **me** quer ajudar.
(13b) A Rita quer-**me** ajudar. [10]

É importante salientar que a ênclise não desapareceu de todo do PB. Segundo Lobo (2002), fatores sociais, como a escolarização, visto que o padrão enclítico é ensinado nas escolas, e linguísticos parecem ter um papel importante no surgimento desta colocação. Ainda assim, a autora concorda que a próclise é a forma normal de colocação no PB vernacular contemporâneo. Sendo a escolarização o fator que aparentemente introduz a possibilidade da ênclise aos falantes do PB, a próxima secção trará mais detalhes sobre o ensino da colocação pronominal nesta variedade.

3. **O ensino dos clíticos no português brasileiro**

A literatura demonstra que a próclise é efetivamente o padrão básico de colocação pronominal no PB, independentemente do contexto sintático, o que resulta, por exemplo, na possibilidade de um clítico iniciar uma frase ou, ainda, de estar em próclise ao verbo principal de locuções verbais (Morais & Ribeiro, 2005, pp. 24–25). No entanto, o ensino dos clíticos faz parte da matriz curricular das aulas de português no Brasil. É por meio desse ensino que os alunos tomam conhecimento do padrão de colocação desses pronomes que, em teoria, já não existem na sua gramática nuclear (Kato, 2005, 2017), ou seja, na configuração do seu sistema linguístico adquirido nos primeiros estádios do seu desenvolvimento.

No que concerne ao padrão de colocação dos clíticos, a tradição baseia-se em expor aos alunos regras que, à partida, simulam as regras do PE. Tal tradição remonta a gramáticas produzidas no Brasil já independente no século XIX, as quais moldaram a postura que as gramáticas posteriores assumiriam em relação aos clíticos (Romeo, 2019, p. 348). Ainda assim, alguns gramáticos da época já evidenciavam a existência de um padrão de colocação tipicamente brasileiro, ainda que devesse ser evitado (Romeo, 2019, p. 348).

Apesar de toda a evolução no conhecimento acerca da variedade brasileira do português, essa tradição persiste no século XX. Por exemplo, a *Moderna Gramática Brasileira*, de Celso Pedro Luft, publicada em 1987, ainda que reconheça as diferenças entre o PE e o PB (daí o título especificar *Gramática Brasileira*) e exemplifique essas diferenças com a colocação pronominal (Luft, 1987, p. 14), quando aborda efetivamente tal assunto, evoca a norma tradicional. O autor considera, portanto, que a colocação normal é a ênclise, sendo que a próclise emerge na presença de "elementos de atração" (Luft, 1987, p. 19). Não obstante, sobre a próclise, admite que o seu uso possa "imprimir um tom coloquial, intimista ou descontraído" (Luft, 1987, p. 39), o que

[9] Há evidências de casos de subida do clítico em PB, nomeadamente em construções passivas (Cruz & Namiuti, 2019; Reis, 2011) e com os clíticos acusativos *o(s)/a(s)* em tempos compostos (Nunes, 2015). Os autores alegam que a estrutura passiva não corresponde à estrutura sintática de outras perífrases verbais e que os clíticos acusativos de terceira pessoa tampouco correspondem aos demais clíticos, funcionando como marcas de concordância de objeto (*vd.* Nota 15).

[10] Esta frase diferencia-se da frase (9) pelo facto de não admitir nenhum material fonético entre o verbo e o clítico, o que demonstra que, no primeiro caso abaixo, o clítico está efetivamente em ênclise ao verbo auxiliar e, no segundo caso, em próclise ao verbo principal:

a. A Rita quer-me finalmente ajudar. (??PB) (PE)
b. A Rita quer finalmente me ajudar. (PB) (*PE)

contrasta com a ênclise, de tom mais cerimonioso e usada em linguagem objetiva, técnica, entre outros usos da língua de maior formalidade.

No século XXI, Azeredo (2008), com a sua *Gramática Houaiss da Língua Portuguesa*, já aborda a questão da colocação por outra ótica. Ainda que os dados para a constituição dessa gramática provenham da produção de brasileiros cultos (escritores, jornalistas e autores brasileiros) "desde a segunda metade do século XIX até os dias atuais, em obras literárias, técnicas, científicas e ensaísticas em geral, [...] principais jornais e revistas dos grandes centros urbanos contemporâneos" (Azeredo, 2008, p. 26), ou seja, duma aparente norma padrão brasileira, este autor distancia-se da análise tradicional para a colocação dos clíticos e considera a próclise como o padrão de colocação normal. Mais além, diz que a colocação proclítica deve ocorrer na variedade culta em situações em que um verbo no futuro do indicativo ou no condicional esteja precedido por um pronome sujeito expresso[11] e quando um advérbio ou pronome de significação negativa ocorra imediamente antes do verbo. Ademais, a próclise é preferível quando antes do verbo se encontre um conetivo de subordinação, especialmente se o verbo da oração subordinada introduzida por ele estiver no modo conjuntivo (Azeredo, 2008, p. 259). Quanto à ênclise, deriva sobretudo de fatores sociocomunicativos. Neste sentido, não estará condicionada por gramaticalidade, ocorrendo simplesmente em situações com um alto nível de monitorização, quando um tom mais formal for exigido, ainda que possa exprimir um certo artificialismo no discurso.

A *Gramática do Português Brasileiro*, de Mário Perini, publicada em 2010, traz uma inovação. O autor preocupou-se em descrever o PB falado, em oposição à norma culta escrita dessa variedade. Descreve que "as diversas variedades faladas, em conjunto, se opõem nitidamente à variedade padrão escrita, o que nos autoriza a falar do **português falado do Brasil** como uma entidade linguística razoavelmente coerente" (Perini, 2010, p. 19, grifo do autor). Destarte, admite a próclise como a regra geral no PB (Perini, 2010, p. 119) e rejeita a ênclise.

Portanto, a tradição oriunda dos gramáticos do século XIX, de definir para o PB um modelo de colocação dos clíticos similar ao do PE, vem perdendo espaço, passando-se a assumir que a próclise é o padrão de colocação típico da variedade brasileira, pelo menos nas gramáticas mais atuais. Contudo, as regras tradicionais continuam a ser, de certo modo, impostas em salas de aula do ensino regular brasileiro, principalmente quando se adota aquilo que Vieira (2008) considera uma orientação tradicional-normativa do ensino da colocação pronominal. Nessa orientação, o foco dá-se sobre as regras de colocação, especificando a presença de itens como os proclisadores (situações em que a ênclise não pode ocorrer) e debruçando-se sobre a linguagem escrita. Além dessa orientação, Vieira (2008) explica que se pode abordar os clíticos sob uma perspetiva mais progressista, em que se considera a próclise como o padrão de colocação do PB, seja na modalidade oral, seja na escrita. Por fim, apresenta uma terceira orientação, que decorre sob um viés sociolinguístico inovador, focada na "opção do aluno na concretização da norma de uso do PB, que prevê a próclise como opção preferencial (sem desconsiderar a realização da ênclise)" (Vieira, 2008, p. 143).

O livro *Português Contemporâneo: Diálogo, Reflexão e Uso* (Cereja et al., 2016), na sua versão destinada a alunos do último ano do Ensino Médio, parece fazer uso das diferentes orientações apontadas acima, com um maior foco nas orientações tradicional-normativa e sociolinguística inovadora. A lição que aborda os clíticos baseia-se, em geral, em diversos textos (de diferentes níveis de formalidade, podendo ou não simular a linguagem coloquial) os quais incluem pronomes clíticos[12] (Cereja et al., 2016, pp. 161–167). Primeiramente, trabalham-se questões de interpretação dos textos para que então sejam abordados os clíticos por meio de exemplos retirados desses textos. Os exercícios requerem a identificação pelo aluno da posição em que esses clíticos se encontram e, em seguida, uma reflexão sobre o seu idioleto, em que devem considerar se, ao produzirem os mesmos enunciados, colocariam o clítico na mesma posição.

[11] O autor refere que a mesóclise também pode ocorrer com estas formas verbais, mas trata-se de um registo *ultraformal* (Azeredo, 2008, p. 259).

[12] O termo utilizado pelos autores do livro é "pronome oblíquo átono." Os outros termos gramaticais referidos também seguem a nomenclatura usada pelos autores.

Quanto à explicitação da colocação pronominal, esta é definida como "a posição do pronome oblíquo átono em relação ao verbo que ele acompanha" (Cereja et al., 2016, p. 163). Neste momento, retoma-se a ideia de que, no PB, a colocação proclítica (14a) é a que ocorre maioritariamente, mas admite a possibilidade de outras colocações, nomeadamente a ênclise (14b) e a mesóclise (14c) em registos mais formais:

(14a) "sempre **me** predispõe"[13]
(14b) "lançou-**me** a ideia"
(14c) "ir-**me**-ia emprestar"

O texto didático segue com as regras que exigem a ênclise e a mesóclise. Quanto à primeira, deve ser usada em início de frase (15a) ou quando uma vírgula antecede o verbo (15b), em frases afirmativas imperativas (15c) e em casos de gerúndio (15d) e de infinitivo pessoal (15e):

(15a) Sentei-**me** na primeira fileira.
(15b) Se chegar cedo, sento-**me** na primeira fileira.
(15c) Amanhã sentem-**se** todos na primeira fileira.
(15d) Chegou sentando-**se** na primeira fileira.
(15e) Era meu intuito ajudá-**lo**.

Sobre a mesóclise, deve-se dar-lhe preferência quando os verbos estão no futuro do indicativo (16a) ou no condicional (16b):

(16a) Dir-**se**-á que somos loucos.
(16b) Sentar-**se**-ia na primeira fileira se tivesse chegado cedo.

Os autores sistematizam num quadro a explicação acerca das *palavras atrativas* (*i.e.*, proclisadores). Segundo a informação contida nesse quadro, essas palavras "atraem o pronome para antes do verbo e geram próclise, mesmo em contextos nos quais se teria ênclise ou mesóclise." Em tal descrição encontram-se, assim, os advérbios (17a, b), os pronomes relativos e indefinidos (17c, d), as conjunções subordinativas (17e, f) e as preposições em casos de verbos no gerúndio ou no infinitivo pessoal (17g, h):

(17a) Não **se** queixe.
(17b) Antes **me** encontrei com ela.
(17c) Tudo o que **me** propus a fazer...
(17d) Ninguém **me** avisou...
(17e) Disse que **me** deixariam ficar.
(17f) Se **me** contassem...
(17g) Em **se** tratando de...
(17h) Para **se** sentarem.

Os exercícios que surgem após a explicitação das regras de colocação mantêm o intuito de estimular uma reflexão sobre o uso dos clíticos nos textos e na linguagem falada, aliando também a comparação entre esses usos e a norma. Também se preocupam em apresentar diferentes tipos de textos, demonstrando que, quanto mais formal for o texto, mais respeitada a norma é. A última secção da lição trabalha em cima de dois anúncios, sendo que um deles inclui a seguinte frase (18):

[13] Os exemplos de (14) a (18) foram retirados do livro didático.

(18) Porque preço que está na moda não cai, **se** joga.

O quarto exercício da secção menciona que "[a] colocação pronominal tal como aparece no anúncio não é a indicada pelas regras da gramática normativa" (Cereja et al., 2016, p. 167). O livro do professor, neste momento, salienta que, apesar da norma, "se joga" é uma gíria e que não faria sentido modificá-la, pois tal colocação está cristalizada,[14] ou seja, não pode ser utilizada de outra forma.

O facto de a tradição normativa buscar no PE as regras para a colocação pronominal aumenta consideravelmente a distância entre a língua falada e o padrão escrito culto do PB (Tarallo, 1996, p. 70). Destarte, a secção a seguir apresentará alguns dados de produção conduzidos com falantes nativos do PB.

4. Os efeitos do ensino da colocação pronominal

Dadas as diferentes posições de colocação dos clíticos, diversos autores preocuparam-se em observar a posição em que os falantes os produziam no PB, tanto na modalidade oral, quanto na modalidade escrita, com populações de diferentes idades e de diferentes regiões do país. Por exemplo, Vandresen (2004) investigou a produção oral desses pronomes por adultos habitantes de Jaguarão, Chuí e Pelotas (Rio Grande do Sul). O autor constatou a ênclise em apenas 48 (1,48%) dos 3226 enunciados orais analisados, sendo que 31 deles correspondiam ao uso da expressão *ir-se embora*, e nove do clítico *se* com função reflexiva.

O estudo de Machado (2006), realizado com estudantes na cidade do Rio de Janeiro, apontou um uso de 20% de ênclise na escrita desses indivíduos (p. 91), sendo que os alunos do terceiro ano do Ensino Médio não tiveram um desempenho diferente daquele dos alunos da quarta série do Ensino Fundamental (p. 106). Contudo, a autora deteta diferenças no uso da ênclise em certos contextos. Por exemplo, em 23% das frases que incluíam o clítico *se* com função de indeterminador do sujeito, tal clítico aparecia em ênclise, percentagem apenas superada pelo uso dos clíticos acusativos *o(s)/a(s)*[15] em 43% dos casos.

Numa comparação mais direta entre diferentes modalidades na mesma população, Lacerda et al. (2021) analisaram cartas produzidas por indivíduos de Feira de Santana (Bahia) durante o século XX e encontraram uma proporção de 32% de ênclise, contra 68% de próclise, em frases simples sem proclisadores. As autoras então analisaram os dados combinando duas variáveis: o contexto sintático e o nível de escolaridade dos

[14] De facto, é importante salientar que muitos casos de ênclise e de próclise no PB podem ser fruto da cristalização. Neste sentido, o clítico perde a sua autonomia e passa a fazer parte, por meio de reanálise, da estrutura (Nunes, 2007):

a. Dane-**se**!
b. Bem-**te**-vi.

Outros fatores que também podem contribuir para essa cristalização são de ordem fonológica, como na expressão abaixo, em que a rima obtida entre a combinação do verbo e do clítico e a última palavra da expressão se perderiam caso houvesse a próclise típica do PB nesse contexto (Nunes, 2007):

c. Acabou-**se** o que era doce.

[15] Note-se que os clíticos acusativos de terceira pessoa também são adquiridos por meio da escolarização – a tendência do PB é a utilização dos pronomes fortes *ele(s)/ela(s)* no seu lugar (Duarte, 2020; Luís & Kaiser, 2016). Sobre a preferência pela colocação pós-verbal com os clíticos acusativos de terceira pessoa, Nunes (2015) considera que esses clíticos foram reinterpretados como marca de concordância de objeto, dada a sua tendência a ocupar, num verbo no infinitivo, a posição canónica para a morfologia de concordância (*i.e.*, pós-verbal):

a. A Rita quer ajudá-**lo**.

Em verbos que apresentem flexão de concordância (e.g., verbos finitos ou no infinitivo flexionado), o autor considera que essa colocação é impossibilitada e a próclise emerge:

b. A Rita **o** ajudou.
c. A Rita disse para **o** ajudarmos.

Da mesma forma, enquanto os outros clíticos no PB ocorrem em próclise a verbos participiais, tal possibilidade, segundo o autor, inexiste com os clíticos acusativos de terceira pessoa, os quais ocorrem em próclise ao verbo auxiliar, visto que formas participiais não têm marcas de concordância:

d. A Rita tem **me** ajudado.
e. A Rita **o** tem ajudado.

Ressalta-se, no entanto, que a análise acima ainda carece de estudos mais profundos.

escreventes. Assim, repararam que os escreventes cultos usaram somente a ênclise quando o verbo estava em início absoluto da frase, ao passo que os escreventes semicultos e populares a utilizaram em 66,7% e 61,9% neste contexto, respetivamente. Os números mudam drasticamente quando se analisam as ocorrências de clíticos antecedidos por outros itens[16]: os escreventes cultos utilizaram a ênclise em 55,6% das ocorrências, os semicultos em 20% e os populares em 4,3%. Por outro lado, Carneiro (2016) observou produções orais de indivíduos dessa mesma localidade e constatou somente o uso de próclise, independentemente do tipo de fala (culta, semiculta ou popular) (p. 146).

Os resultados acima assemelham-se àqueles de outros estudos conduzidos noutras regiões do Brasil, insinuando que a colocação proclítica está bem assentada em todo o território. Contudo, é importante salientar que Carneiro (2005) e Lacerda et al. (2021) apontam diversos casos de hipercorreção, em que os escreventes utilizam a ênclise em contextos que, historicamente, exigiram sempre a próclise, sobregeneralizando, aparentemente, o uso da colocação enclítica. Curiosamente, os dados de aquisição dos clíticos por crianças nativas do PE indicam que, nos primeiros estádios do desenvolvimento, elas sobregeneralizam a colocação enclítica e os contextos proclíticos são adquiridos gradativamente (cf. Costa et al., 2016). O mesmo pode ser observado por aprendentes de PE como língua segunda (L2) com diferentes línguas maternas (L1) (Gu, 2019; Madeira & Xavier, 2009; Pereira, 2022).

Note-se, no entanto, que os escreventes apresentarem um comportamento similar aos das crianças nativas do PE e dos aprendentes L2 desta variedade do português não é estranho. Kato (2005) já comentava que a "gramática da fala e a 'gramática' da escrita apresentam uma distância de tal ordem que [...] pode ter a natureza da aprendizagem de uma segunda língua" (p. 131). Embora a literatura seja ainda escassa relativamente ao processo de aquisição de estruturas morfossintáticas do PE por falantes nativos do PB, o estudo de Tomaz et al. (2019) contribuiu para trazer alguma evidência deste processo.

As autoras conduziram um estudo com alunos duma escola bilingue francês-PE em França, incluindo, além das crianças falantes de herança do PE, alguns alunos que eram falantes de herança do PB, possibilitando observar como este grupo se comportava nas suas produções linguísticas, dado estarem inseridos num ambiente escolar em que o PE predominava. Utilizando uma metodologia de produção induzida, as autoras observaram que algumas crianças falantes de herança do PB apresentavam indícios de aquisição dos pronomes clíticos do PE, mas ainda se distinguiam das crianças falantes de herança do PE relativamente ao tipo de pronome objeto produzido e ao padrão de colocação dos clíticos. Por exemplo, produziam, além dos pronomes fortes (*vd.* Nota 15), pronomes clíticos e a ênclise, comportamento diferente do percurso de desenvolvimento das crianças nativas do PB no Brasil, que apresentam apenas pronomes fortes acusativos de terceira pessoa e a próclise generalizada (cf. Casagrande, 2007, 2010; Corrêa, 1991; Moura, 2001), com a inserção dos clíticos acusativos de terceira pessoa e da ênclise por meio da escolarização.

Um processo similar pode ser observado no ensino escolar cipriota, curiosamente, concernente aos pronomes clíticos da língua grega. A variedade do grego falada no Chipre difere daquela falada na Grécia, dentre outros pontos, na colocação dos pronomes clíticos (Grohmann et al., 2017): enquanto no grego padrão os clíticos ocorrem em próclise à forma verbal (como no PB), a colocação no grego cipriota é condicionada por fatores sintáticos, sendo a ênclise o padrão em frases afirmativas simples, como no PE (embora os contextos de próclise não sejam todos os mesmos). O que Grohmann et al. (2017) observam é que as crianças em idade pré-escolar no Chipre produzem os clíticos em ênclise em frases afirmativas simples, padrão que muda no momento em que vão para a escola, passando a produzirem próclise em tal contexto. Logo, as crianças cipriotas ao iniciarem os seus estudos são expostas ao grego padrão e adquirem o padrão de colocação dessa variedade com êxito (Themistocleous, 2017), interferindo até mesmo nas suas produções em grego cipriota, tendência que se reverte após alguns anos (supostamente quando conseguem diferenciar que cada tipo de colocação está atrelado

[16] Sujeitos pronominais, DP ou com oração relativa, sintagmas preposicionais e advérbios não modais.

a uma variedade específica), ainda que interferências possam continuar a ocorrer até mesmo em idade adulta (Leivada at al., 2010).[17]

O ponto fulcral no caso cipriota está no facto de que a variedade linguística usada nas escolas é efetivamente o grego padrão, seja na oralidade, seja nos materiais didáticos, de maneira similar ao que ocorre com os participantes do estudo de Tomaz et al. (2019), os quais estão expostos ao PE. Note-se que o facto de as crianças do estudo de Tomaz et al. (2019) apresentarem mais formas do PB nos seus enunciados pode ser porque estão a adquirir o PE num ambiente bilingue, já que a suposta interferência do sistema linguístico francês também se observa nos falantes de herança do PE.[18]

O caso do ensino brasileiro difere dos dois exemplos supracitados precisamente no facto de que as crianças são escolarizadas em PB, tendo na escola, na verdade, contacto com a norma padrão da variedade brasileira, nomeadamente na forma de descrição de regras (muitas vezes complexas) e na sua aplicação em materiais escritos.[19] Embora não se possa fazer uma afirmação baseada em observações empíricas, não parece ser o caso que todos os professores do ensino regular brasileiro façam uso da colocação pronominal tal como preconizada pela norma nos seus enunciados orais. Logo, retomando Kato (2005), durante o processo de escolarização, não encontrando as crianças regras gramaticais que se assemelhem ao seu vernáculo, a aprendizagem dessas regras pode fazer com que, na realidade, experimentem um processo de aquisição duma L2. O resultado, segundo a autora, é o desenvolvimento duma "gramática do letrado", a qual compete com a gramática nuclear dos falantes e emerge, sobretudo, na modalidade escrita e em registos formais, por associação à escolarização. A partir deste breve referencial teórico, passa-se à descrição do presente estudo, iniciando-se pelos seus objetivos.

5. **Objetivos**

Os dados dos estudos maioritariamente baseados em análise de *corpora* de produção oral ou escrita, brevemente descritos anteriormente demonstraram que a próclise é a posição preferencial de colocação pronominal no PB. Porém, a tradição normativa, baseada no PE, cujo padrão de colocação observa restrições sintáticas, ainda é imposta em salas de aula, o que explica a ocorrência da ênclise nesses dados.

Tendo sido a maioria dos estudos conduzidos com vista a analisar dados de produção, propõe-se, com este estudo, averiguar a (in)aceitabilidade do padrão de colocação dos clíticos no PB em diferentes contextos sintáticos por indivíduos que tenham concluído, pelo menos, o Ensino Médio por meio de um teste de juízos de aceitabilidade.[20] As questões que se colocam são:

[17] Tais autores também referem que esses indivíduos conseguem alternar entre as duas variedades de acordo com o interlocutor.

[18] Recorde-se que o padrão de colocação no francês é maioritariamente proclítico (Kayne, 1975, p. 66).

[19] Outra questão que se levanta é que, se se compararem as regras descritas no livro *Português Contemporâneo: Diálogo, Reflexão e Uso* (e, possivelmente, de tantos outros que circulam pelas salas de aula brasileiras) com a descrição da colocação da norma do PE na Secção 2, ver-se-á que, embora similares, a norma a que os alunos brasileiros estão expostos é "artificial", no sentido em que tenta simular o padrão de colocação do PE, simplificando-o, e dando origem a regras que não são utilizadas em nenhum lado do Atlântico. Um exemplo pode ser visto quando no referido livro se exige o uso da ênclise após vírgulas (Cereja et al., 2016, p. 163). Ora, se, por exemplo, a vírgula em questão servir de limites para uma oração relativa que modifique o sujeito duma oração subordinada, a próclise deverá ocorrer, mesmo que o verbo não esteja adjacente à conjunção que introduz a oração subordinada:

a. Quero que o meu amigo, aquele que faz Medicina, **me** ajude a estudar para o exame.

Da mesma forma, as explicações insistem no facto de que o elemento proclisador deva anteceder imediatamente o verbo, quando podem estar interpostos por conteúdo lexical:

b. Só os meus amigos **se** confundiram.

[20] Segundo Myers (2017, p. 4), os testes de juízos de aceitabilidade complementam os dados obtidos em *corpora*, em especial, por serem dados experimentais (e não observacionais). Ademais, podem ser importantes para estabelecer a aceitabilidade de formas raras em dados de produção, especialmente em casos em que se tem como objetivo avaliar usos linguísticos em potencial e não só aqueles que são largamente usados pelos falantes (Gries, 2012), como é o caso deste estudo. Ainda, para Gerasimova e Lyutikova (2020), só é possível avaliar em que direção vão formas em variação num sistema aliando-se dados de produção aos de juízos de aceitabilidade. Ainda que as autoras defendam que os dados de produção e de aceitbilidade devam vir dos mesmos participantes, este estudo tenta suprir a inexistência de dados de juízos concernentes à aceitabilidade das diferentes posições de colocação pronominal no PB.

Questão 1: Existem diferenças na aceitabilidade da ênclise e da próclise entre contextos sintáticos com e sem proclisadores por falantes escolarizados nativos do PB na modalidade escrita?

Questão 2: A continuidade dos estudos para além do Ensino Médio contribui para diferenças nos resultados?

Hipotetiza-se que ambas as colocações sejam aceites pelos participantes inobstante o contexto sintático. Esta hipótese apoia-se naquilo que foi apontado por Azeredo (2008): a ênclise e a próclise no PB não têm que ver com a gramaticalidade, senão com fatores, à partida, alheios à sintaxe. Neste sentido, a presença ou ausência de proclisadores, apesar da explicitação das regras em salas de aula, não terá efeito sobre os julgamentos. Relativamente à continuidade dos estudos para além do Ensino Médio, espera-se que o aumento nos anos de escolarização tenha efeito no sentido de os falantes terem julgamentos que os aproximem da norma preconizada na escola, visto que a continuidade dos estudos os terá exposto a ela por mais tempo dentro de um ambiente que a valoriza, isto é, a universidade.[21]

6. **Metodologia**

Este estudo apoiou-se num questionário com recurso à ferramenta Google Formulários para a sua execução. Os participantes que nele se dispuseram a participar tiveram de, primeiramente, ler e concordar com os termos de realização do trabalho. Em seguida, responderam a um questionário sociolinguístico, sendo que, no total, 79 participantes realizaram a tarefa. Todos os participantes tinham mais de 18 anos de idade, sendo que 50 deles (63,3%) eram mulheres e 29 (36,7%) eram homens. Relativamente ao seu grau de escolaridade, 12 (15,2%) tinham concluído o Ensino Médio, ao passo que 35 (44,3%) obtiveram um diploma de Ensino Superior ou Técnico e 32 (40,5%) prosseguiram com estudos de pós-graduação. Provinham de todas as regiões geográficas brasileiras,[22] sendo que a maior parte, 35 participantes (44,3%), habitavam na região Sul. Quatro participantes (5,1%) habitavam na região Centro-Oeste, 16 (20,3%) na região Nordeste, oito (10,1%) na região Norte e 16 (20,3%) na região Sudeste. Ademais, dez participantes (12,7%) indicaram possuir outra L1 e 68 participantes (85,5%) indicaram ter aprendido pelo menos uma L2. Os dados recolhidos por meio do questionário estão resumidos na Tabela 1.

[21] Inclusive, a colocação pronominal é um dos tópicos abordados nos exames de admissão a universidades brasileiras nas provas de português e de redação.

[22] Estabeleceu-se que somente poderiam realizar a tarefa indivíduos que vivessem na mesma região brasileira (Norte, Nordeste, Centro-Oeste, Sul ou Sudeste) em que tivessem sido criados e que tivessem pelo menos um progenitor que também lá tivesse sido criado.

Tabela 1. Perfil sociolinguístico dos participantes

N	79
Faixa etária (em anos)	18 – 29 = 24 (30,4%) 30 – 39 = 32 (40,5%) 40 – 49 = 13 (16,5%) 50 ou mais = 10 (12,6%)
Género	Feminino = 50 (63,3%) Masculino = 29 (36,7%)
Escolaridade	Ensino Médio = 12 (15,2%) Ensino Superior/Técnico = 35 (44,3%) Ensino Pós-Graduado = 32 (40,5%)
Região de origem	Centro-Oeste = 4 (5,1%) Nordeste = 16 (20,3%) Norte = 8 (10,1%) Sudeste = 16 (20,3%) Sul = 35 (44,3%)
Outra(s) L1 além do português[23]	Não = 69 (87,3%) Alemão = 2 (2,5%) Italiano = 6 (7,6%) Espanhol = 1 (1,3%) Ucraniano = 1 (1,3%)
L2 (independentemente do nível de proficiência alcançado)	Não = 10 (12,7%) Inglês = 64 (81%) Espanhol = 31 (39,2%) Italiano = 13 (16,5%) Francês = 10 (12,6%) Alemão = 8 (10,1%) Japonês = 4 (5,1%) Russo = 2 (2,5%) Hebraico = 1 (1,8%) Não respondeu = 1 (1,8%)

Os participantes tiveram de julgar, por meio de uma escala de Likert de seis pontos, diversas frases na modalidade escrita[24] relativamente à sua aceitabilidade, de modo que os seus julgamentos obedecessem ao seguinte gradiente: 0 – péssima; 1 – muito ruim; 2 – ruim; 3 – boa; 4 – muito boa; e 5 – ótima. Segundo as instruções do teste, deviam considerar que uma frase "ótima" era uma frase que se esperava ser utilizada pelos falantes, fazia sentido, era natural e seguia a estrutura da língua, ao passo que uma frase "péssima" seria o oposto: era artificial, feria a estrutura da língua ou não fazia sentido. Além disso, não havia limite de tempo para a realização da tarefa.

No total, julgaram 40 frases que incluíam verbos simples finitos com pronomes clíticos (*me* e *se*[25]) ora em próclise, ora em ênclise, em cinco contextos sintáticos diferentes (afirmativas simples sem proclisadores, negativas, adverbiais introduzidas por *porque*, subordinadas no indicativo introduzidas por *que* e subordinadas

[23] Uma análise individual dos resultados dos falantes com outras línguas maternas não apontou diferenças entre as suas performances e a dos demais participantes, pelo que se optou por manter os seus dados nas análises. Salienta-se que não se obtiveram dados relativamente ao tipo de bilinguismo desses participantes (*i.e.*, se adquiriram o português ou a sua outra L1 em simultâneo ou sequencialmente), tampouco relativamente ao uso que fazem dela atualmente. Todavia, todos foram criados no Brasil. A pergunta feita a eles foi: *Além do português, tem outro(s) idioma(s) materno(s) (idioma adquirido sem instrução formal no âmbito familiar ou da comunidade durante a primeira infância)?*

[24] Visto que a ênclise emerge, em geral, na modalidade escrita, optou-se por utilizar esta modalidade.

[25] Optou-se por estes dois pronomes por serem os que menos sofrem variação dialetal no PB (cf. Duarte, 2020). Repare, também, que são os pronomes mais presentes nos exemplos retirados do livro *Português Contemporâneo: Diálogo, Reflexão e Uso*.

no conjuntivo introduzidas por *que*).[26] A combinação das variáveis gerou, assim, dez contextos, cada um com quatro frases (duas com *me* e duas com *se*). É importante salientar que as frases afirmativas simples sem proclisadores eram sempre iniciadas por um sujeito pronominal ou nominal expresso para que houvesse material lexical pré-verbal da mesma forma que nas frases com proclisadores (que obrigatoriamente ocorrem nessa posição). Ademais, nos itens com proclisadores, o verbo e o seu clítico ocorriam imediatamente após o proclisador (*não*, *porque* ou *que*). Recorde-se de que alguns materiais produzidos no Brasil salientam a necessidade de adjacência do proclisador ao verbo com o clítico. Ainda que, pela norma do PE, isto não se aplique, optou-se por manter a adjacência entre os elementos para evitar possíveis interferências da existência de material lexical interposto entre o proclisador e o verbo com o clítico. Além dos itens de teste, os participantes tiveram de julgar mais 44 frases que serviram como distratores e que não continham clíticos, mas possuíam estruturas possíveis na língua, estruturas impossíveis na língua ou estruturas em variação, de modo que os participantes pudessem fazer uso de toda a escala. Ademais, antes do início da tarefa, treinaram com duas frases que tampouco continham clíticos. A Tabela 2 traz exemplos de cada um dos contextos mencionados.

Tabela 2. Exemplos de itens de teste por contexto e de distratores.

Contexto	***Exemplo de Item***
AE	O professor conhece-me da faculdade.
AP	As novas alunas se comparam muito comigo.
NE	O guarda não viu-me na estação.
NP	O carteiro não se esqueceu das cartas.
QE	A enfermeira está cansada porque levantou-se cedo.
QP	Os meninos desistiram do jogo porque se irritaram.
IE	Acho que cortei-me com o papel.
IP	Disseram que me viram no aeroporto.
CE	Peço que deem-me todos os detalhes.
CP	Espero que me considerem capaz de assumir a empresa.
Distrator	Apesar do frio, não nevou.
Distrator	Eu não sabemos falar inglês.
Distrator	Ele tem chego tarde todos os dias.

Nota. A = afirmativas simples; N = negativas; Q = adverbiais introduzidas por porque; I = subordinadas no indicativo introduzidas por que; C = subordinadas no conjuntivo introduzidas por que; E = ênclise; P = próclise.

Todas as frases estavam organizadas em quatro páginas do formulário, de modo que, em cada uma, os participantes tivessem de julgar dez itens de teste (dois de cada contexto sintático, um com o clítico em ênclise e outro com o clítico e próclise). Embora as mesmas frases aparecessem sempre na mesma página, a ordem em que apareciam era aleatória, o que possibilita a redução de efeitos de *priming*.[27] Os resultados obtidos estão descritos na secção a seguir.

7. **Resultados**

A análise primária dos dados demonstrou que as frases que incluíam o clítico em próclise ao hospedeiro verbal foram mais bem avaliadas do que aquelas que o tinham em ênclise, com uma média de 4,26 (DP = 0,60) no primeiro caso e uma média de 3,39 (DP = 1,13) no segundo. No entanto, como exposto no referencial teórico

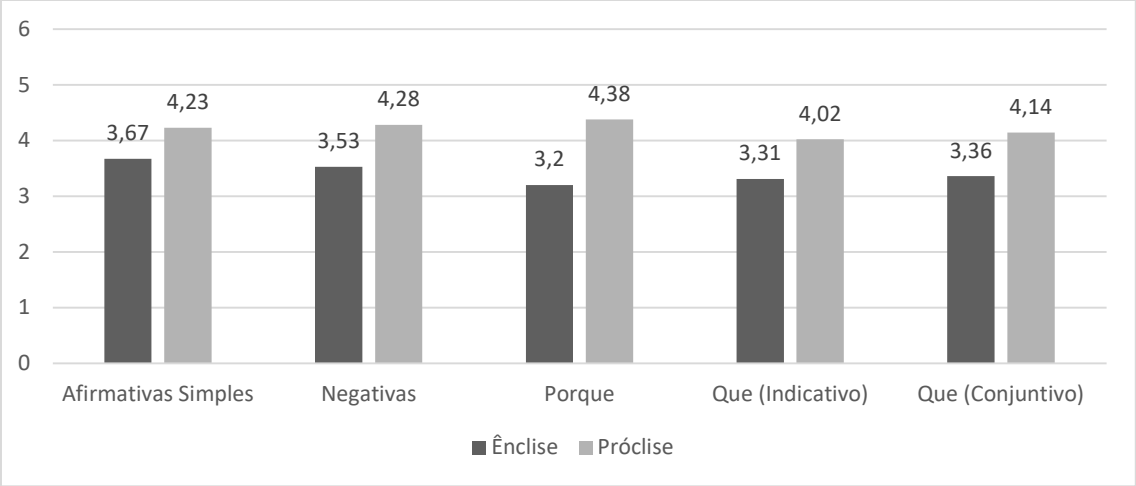
[26] Os contextos com proclisadores utilizados foram selecionados por não apresentarem a necessidade de conhecimento lexical. Recorde-se de que nem todos os advérbios ou quantificadores pré-verbais são proclisadores, ao passo que os contextos de negação e de subordinação resultam em próclise (cf. Secção 2).

[27] Isto é, o efeito facilitador que certos estímulos (linguísticos ou não) podem criar para a ativação de conceitos representados mentalmente (cf. Chartrand & Jefferis, 2004, p. 854). Neste sentido, a aleatorização torna-se importante para que os participantes não sejam todos expostos à mesma sequência de itens, o que poderia interferir sistematicamente nos julgamentos.

deste trabalho, diferentes contextos sintáticos exigem diferentes colocações do clítico segundo a tradição normativa baseada no PE. Portanto, é necessário desdobrar as médias de acordo com os contextos aqui utilizados.

Relativamente às frases afirmativas simples (que exigem a ênclise no PE), viu-se que a média das frases que possuíam o clítico em ênclise foi de 3,67 (DP = 1,07) e em próclise foi de 4,23 (DP = 0,79). Quanto ao restantes contextos, os quais exigem o uso da próclise, viu-se que, com frases negativas, as médias foram de 3,53 (DP = 1,11) e 4,28 (DP = 0,64) para a colocação enclítica e proclítica, respetivamente. Em frases introduzidas por *porque*, a média da colocação enclítica foi de 3,20 (DP = 1,34) e da proclítica foi de 4,38 (DP = 0,55). Quanto aos contextos de subordinadas introduzidas por *que*, aquelas com os verbos no indicativo tiveram uma média de 3,31 (DP = 1,33) no caso de clíticos em ênclise e de 4,02 (DP = 0,70) no caso de próclise; já no caso de frases com o verbo no conjuntivo, a média obtida para a ênclise foi de 3,36 (DP = 1,22) e para a próclise, 4,14 (DP = 0,74). Todas as médias por contexto sintático e colocação do clítico indicadas estão ilustradas no Gráfico 1.

Gráfico 1. Médias dos julgamentos por contexto sintático e colocação do clítico por falantes nativos do PB escolarizados



Como no caso das médias, para a análise estatística também é preciso analisar os contextos sintáticos levando em conta os dados individuais para cada tipo de colocação.[28] Para tanto, realizou-se um teste de Kruskal-Wallis relacionando as variáveis independentes (contextos sintáticos e a colocação) à variável dependente (média dos juízos), o que ajuda a determinar se haverá algum efeito daquelas sobre esta. O teste apontou haver diferença estatisticamente relevante (X^2 = 103.17, df = 9, p < 0,001). Para deteminar precisamente onde se encontra(m) a(s) diferença(s), foi então realizado um teste de Wilcoxon corrigido pela correção de Holm-Bonferroni, o qual comparou todos os contextos entre si.[29] Estipulando-se que as médias entre dois contextos/colocação serão estatisticamente diferentes entre si quando o resultado do teste apontava para um valor p menor do que 0,05, verificou-se que as diferenças estatisticamente relevantes se encontram quando se comparam os dois tipos de colocação dentro do mesmo contexto sintático. Os resultados estão incluídos na Tabela 3.

[28] Uma análise prévia foi feita para determinar se havia diferenças entre as frases com os clíticos *me* e *se*. Porém, não se observaram diferenças (média do pronome *me* = 3,83 (DP = 0,72), média do pronome *se* = 3,82 (DP = 0,78), p = <001).

[29] Para efeitos de objetividade na apresentação destes dados, só serão apresentados os dados relacionados às comparações entre as colocações intracontextos (ou seja, as comparações entre a próclise e a ênclise no contexto de frases afirmativas simples sem proclisadores, entre a próclise e a ênclise no contexto de frases negativas, etc.), visto que a comparação intercontextos relativamente à próclise ou à ênclise não apresentou diferenças estatísticas.

Tabela 3. Resultados do teste de Wilcoxon

Comparação	*p*
AE – AP	0,002*
NE – NP	<0,001*
QE – QP	<0,001*
IE – IP	0,003*
CE – CP	<0,001*

Nota. A = afirmativas simples; N = negativas; Q = adverbiais introduzidas por porque; I = subordinadas no indicativo introduzidas por que; C= subordinadas no conjuntivo introduzidas por que; E = ênclise; P = próclise.
* = p estatisticamente diferente.

A última análise realizada relacionou os resultados obtidos de cada contexto/colocação ao nível de escolarização dos falantes.[30] Contudo, o teste de Dunn corrigido pela correção de Šidák não apontou diferença estatisticamente relevante[31] entre a variável escolarização, os contextos sintáticos e as possíveis colocações do clítico (Tabela 4). Os resultados apresentados serão discutidos na secção a seguir.

Tabela 4. Resultado do teste de Dun

	P									
	AE	AP	NE	NP	QE	QP	IE	IP	CE	CP
EM – ST	0,865	0,667	0,624	0,863	0,685	0,820	0,584	0,735	0,656	0,645
EM – PG	0,389	0,820	0,426	0,847	0,440	0,609	0,579	0,452	0,501	0,568
ST – PG	0,229	0,450	0,049	0,815	0,078	0,560	0,096	0,113	0,093	0,126

Nota. A = afirmativas simples; N = negativas; Q = adverbiais introduzidas por porque; I = subordinadas no indicativo introduzidas por que; C = subordinadas no conjuntivo introduzidas por que; EM = Ensino Médio; PG = Ensino Pós-Graduado; ST = Ensino Superior/Técnico.

8. **Discussão**

Este estudo foi conduzido com vista a observar a perceção que os falantes escolarizados nativos do PB têm relativamente às diferentes possibilidades de colocação pronominal, visto que a colocação proclítica é a colocação padrão do PB vernacular e a ênclise é adquirida por meio da escolarização (cf. Secções 2 e 3). Deste modo, o que se observou por meio dos dados da tarefa de julgamento de aceitabilidade é que, apesar de os itens que continham a próclise aparecerem sempre mais bem avaliados do que aqueles que continham a ênclíse, a média de julgamentos que incluíam este tipo de colocação foi de 3,39 (DP = 1,13; contra 4,26, DP = 0,60, dos itens com próclise), estando, portanto, acima do limiar de 2,5 – o ponto central da escala.

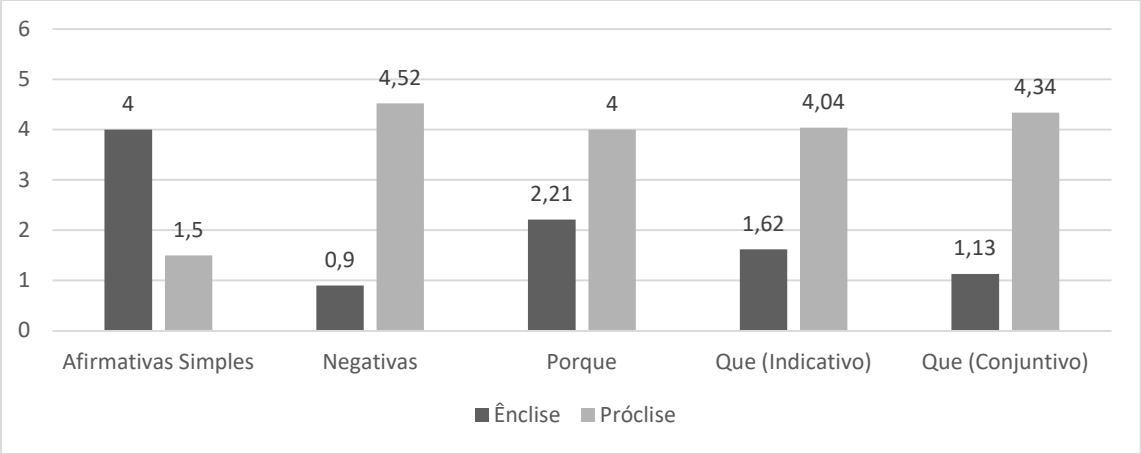
Quando se analisam os dados de acordo com os contextos sintáticos deste estudo (cf. Secção 4), vê-se que o mesmo se repete: independentemente do contexto sintático analisado, os itens com o clítico em próclise foram mais bem avaliados do que aqueles com o clítico em ênclise (cf. Gráfico 1). Esta constatação, em teoria, só seria importante para o contexto de frases afirmativas simples sem proclisadores, pois, segundo as gramáticas

[30] Além da análise relativamente ao nível de escolarização dos participantes, por haver dados disponíveis oriundos do questionário sociolinguístico, também se aproveitou para cruzar os dados com a origem geográfica dos falantes. Porém, não se obtiveram diferenças significativas na análise. É de se salientar a discrepância na quantidade de participantes oriundos de cada região brasileira (cf. Tabela 1), pelo que tais dados devem ser analisados com cautela. Ainda assim, a literatura demonstra que o padrão de colocação proclítico está generalizado no território brasileiro, sendo possível que resultados semelhantes sejam obtidos mesmo com um número amostral mais representativo de cada região. Da mesma forma, cruzaram-se os dados com o género dos participantes e nenhuma diferença foi observada.
[31] Neste caso, para que houvesse significância estatística, o valor de *p* deveria ser menor do que 0,025 (0,05/alfa).

normativas, são contextos que exigem a ênclise,[32] ao passo que os contextos restantes exigem a próclise. Contudo, o que se viu foi, de maneira geral, uma maior aceitabilidade da próclise, mas sem uma rejeição da ênclise, independentemente do contexto sintático (ainda que com diferenças estatísticas entre os julgamentos de cada posição de colocação do clítico).

A título de curiosidade, observe-se que os resultados obtidos neste estudo diferem daqueles obtidos com falantes nativos do PE. O Gráfico 5 reúne os dados duma pesquisa em andamento[33] que faz uso da mesma metologia descrita neste estudo.

Gráfico 2. Médias dos julgamentos por contexto sintático e colocação do clítico por falantes nativos do PE



No caso dos falantes do PE, é clara a preferência pela ênclise nas frases afirmativas simples e pela próclise nos demais casos.[34] Ao mesmo tempo a próclise é rejeitada em frases afirmativas simples e a ênclise é rejeitada nos outros contextos, coincidindo com aquilo que se espera dos falantes nativos do PE, de acordo com a descrição desse sistema linguístico (ainda que nem todos os contextos se comportem da mesma maneira, havendo mais variação em certos contextos do que noutros, cf. Martins, 2013, 2016) e pelos resultados de estudos experimentais que utilizaram essa população como grupo de controlo (cf. Costa et al., 2016; Gu, 2019; Madeira & Xavier, 2009; Pereira, 2022; Tomaz et al., 2019; entre outros).

Portanto, a pergunta de investigação 1 (*Existem diferenças na aceitabilidade da ênclise e da próclise entre contextos sintáticos com e sem proclisadores por falantes escolarizados nativos do PB na modalidade escrita?*) pode ser respondida de maneira negativa. Apesar da diferença estatística observada entre os julgamentos dos itens com próclise e dos itens com ênclise, ambas as colocações foram julgadas acima do limiar de aceitabilidade. Ou seja, os participantes tiveram, globalmente, o mesmo comportamento: julgaram os itens com próclise com valores mais altos (acima de 4), ao mesmo tempo que também consideraram a ênclise uma colocação possível, mesmo na presença de proclisadores (com médias acima de 3). Este resultado vai

[32] Ainda que algumas considerem que, existindo um sujeito expresso, a próclise é aceite; note-se que tal afirmação não se aplica ao PE, o qual, em geral, só admite a próclise na presença de proclisadores. Neste sentido, pode ser interessante em estudos futuros investigar se há um efeito da presença do sujeito, incluindo frases com e sem sujeito. Da mesma forma, pode-se investigar também se há um efeito da adjacência ou não ao proclisador em frases que o contenham.

[33] A pesquisa em progresso difere do estudo descrito neste artigo no léxico utilizado ao conter palavras típicas da variedade europeia. Os dados prévios aqui expostos são baseados nas respostas de 52 adultos nativos do PE (18 homens e 34 mulheres), habitantes, na sua maioria, da região de Lisboa e Vale do Tejo de diferentes faixas etárias com, pelo menos, o Ensino Secundário (equivalente ao Ensino Médio brasileiro) completo.

[34] Estes resultados, aliados aos distratores claramente agramaticais que foram incluídos na tarefa e que tiveram julgamentos abaixo de 2, ajudam a refutar o argumento de que a metodologia usada possibilitou aos participantes falantes do PB emitirem juízos globalmente altos. Logo, os resultados aqui obtidos parecem efetivamente indicar que eles permitem ambas as colocações.

parcialmente ao encontro da hipótese inicialmente aventada, pois, embora os participantes tenham aceitado ambas as colocações, ainda aceitaram mais as frases com próclise do que as frases com ênclise (independentemente do contexto), o que se confirmou pelos testes estatísticos. Neste sentido, a presença de proclisadores não produziu efeito nos julgamentos, já que as frases com o clítico em ênclise nos contextos que os apresentavam foram avaliadas no mesmo patamar que as frases na ausência deles.[35]

Mesmo que a hipótese não se tenha confirmado por completo, as diferenças observadas nos julgamentos não surpreendem. Como Kato (2005) propõe, os falantes escolarizados do PB lidam com a competição de duas gramáticas. Assim, ter sido a próclise bem avaliada demonstra que os falantes seguiram aquilo que consideravam natural na sua variedade do português, ou seja, seguiram a configuração da sua gramática nuclear. De facto, as instruções do teste de julgamento explicitavam-lhes que deviam julgar o quão boas as frases eram, sendo que uma frase "ótima" seria uma frase que se espera ser utilizada pelos falantes, faz sentido, é natural e segue a estrutura da língua. Neste sentido, os participantes provavelmente julgaram as frases com próclise de acordo com a descrição de uma frase "ótima."

No caso das frases com ênclise, por terem sido avaliadas acima do ponto central da escala, o que ocorreu deve ter sido a influência da escolarização, que introduz a ênclise como uma das possíveis posições de colocação pronominal, mas ainda menos bem avaliadas do que as frases com próclise por não serem utilizadas (tanto quanto a próclise) pelos falantes, o que possivelmente as torna menos naturais, ainda que façam sentido. Ademais, a ênclise teve julgamentos acima de 3 em todos os contextos sintáticos, mesmo com a presença dos proclisadores. Neste sentido, pode-se considerar que o ensino da ênclise tem êxito na introdução da sua possibilidade, mas não desenvolve nos falantes (tal como o efeito do *input* a que estão expostos), aparentemente, o conhecimento sintático restrito a cada possibilidade de colocação dos clíticos (como ocorre com os falantes de PE). O que se viu com os dados obtidos neste estudo parece evidenciar tal afirmação, visto que, se tivessem efetivamente adquirido as restrições sintáticas de colocação, as frases com ênclise, à partida, não teriam sido tão bem avaliadas em contextos com proclisadores[36] como no caso dos falantes do PE (cf. Gráfico 2). Ao mesmo tempo, a aprendizagem[37] das regras tampouco parece ter ocorrido com sucesso: se tivessem usado o seu conhecimento metalinguístico (pressupondo-se que o têm por terem todos terminado o Ensino Médio),[38] o resultado deveria ser o mesmo.

De facto, estes resultados também permitem responder negativamente à segunda questão de investigação (*A continuidade dos estudos para além do Ensino Médio contribui para diferenças nos resultados?*). Hipotetizava-se que haveria efeitos nos resultados no sentido de que os falantes que tivessem prosseguido com os seus estudos seguiriam mais a norma, observando a presença ou ausência de proclisadores, visto que teriam mais contacto com ela devido ao ambiente académico (e, quem sabe, pelos círculos sociais que frequentam). No entanto, salienta-se que somente 12 participantes (15,2% da amostra) tinham concluído apenas o Ensino Médio. Este número é notadamente menor do que o dos outros participantes (35 participantes com o Ensino Superior/Técnico, ou 44,3% da amostra, e 32 com o Ensino Pós-Graduado, ou 40,5% da amostra). Assim, é preferível abster-se de tirar conclusões acerca destes dados devido ao menor poder estatístico causado pela disparidade na distribuição dos grupos.

Logo, mesmo com a explicitação das regras em que cada tipo de colocação deve ocorrer, os falantes mantêm o seu padrão de colocação "padrão" (a próclise). A ênclise, por outro lado, poderá ser uma estratégia governada não por fatores sintáticos (como no PE), senão, por fatores alheios à sintaxe (Azeredo, 2008).

[35] Um contexto em que a ênclise poderia ter sido mais bem avaliada seria em início absoluto de frases, visto que este é um dos contextos em que, pelo menos na escrita, a norma parece ter mais sucesso em ser assimilada.

[36] Note-se que tampouco houve diferenças na aceitabilidade da ênclise quando se compararam os contextos sintáticos entre si, não se podendo dizer, então, que um contexto esteja mais atrelado a uma posição de colocação do que outros (diferentemente do que ocorre no PE).

[37] Assumindo-se aqui a dicotomia aquisição/aprendizagem de Krashen (1981), em que, resumidamente, a aquisição ocorre naturalmente por exposição à língua e a aprendizagem é um processo consciente que desenvolve o conhecimento metalinguístico dos indivíduos.

[38] Recorde-se que o teste não envolvia pressão de tempo, pelo que os participantes podiam pensar acerca dos seus julgamentos pela quantidade de tempo que lhes conviesse.

Ademais, poderá ser, também, uma questão de registo, visto que a ênclise está associada à norma culta. Neste sentido, a sua aceitabilidade dá-se sem que os falantes tenham consciência das restrições gramaticais (já que não fazem parte da sua gramática, mas sim da gramática do PE), e o seu uso em contextos formais ocorre, provavelmente, no sentido de evidenciar um (aparente) conhecimento da norma,[39] facto que é corroborado, por exemplo, pelos dados de Carneiro (2005) e de Lacerda et al. (2021), os quais mostram uma sobregeneralização da ênclise em dados de produção escrita, mesmo em contextos de próclise obrigatória.

9. **Considerações finais**

Apesar de os estudos acerca do uso da ênclise demonstrarem que ela é quase inexistente na oralidade, em registos escritos o seu uso perdura devido a uma tradição normativa, baseada na sintaxe do PE, na qual a ênclise ocorre consistentemente em oposição aos contextos em que a próclise deve ocorrer (cf. Secção 2). Além disso, já que o PB vernacular continua a ser utilizado na oralidade no ambiente escolar, os falantes tendem a associar a ênclise à modalidade escrita, facto reforçado por materiais didáticos que preconizam certas estruturas nessa modalidade. Ademais, o seu uso emerge mais consistemente na modalidade escrita, especialmente em registos mais formais, muito provavelmente devido ao seu contexto de ensino no ambiente escolar, sem que a configuração sintático-discursiva da oração tenha um papel central na colocação pronominal.

Como referido neste artigo, este estudo fez-se valer duma metodologia alternativa às tradicionais análises de *corpora* de produções orais e escritas e pôde atingir os objetivos propostos. É importante salientar que este estudo não objetivava aferir o impacto de diferentes abordagens pedagógicas de ensino no padrão de colocação dos clíticos. Para tal, um desenho experimental específico deveria ser proposto tendo como foco crianças em fase de escolarização. Ademais, não se preocupou em abordar questões relativas a possíveis alterações na gramática nuclear dos falantes do PB (cf. Kato, 2017), visto que a tarefa de juízos de aceitabilidade sem pressão de tempo possibilitava o acesso ao seu conhecimento metalinguístico, isto é, às regras gramaticais explícitas que, à partida, aprenderam na escola – o que também pode ser uma perspetiva de estudo futura.

Ainda assim, os dados obtidos neste estudo aliam-se a outros que já abordaram o mesmo assunto, podendo ser úteis para proporcionar reflexões em relação ao estatuto dos diferentes padrões da colocação pronominal no sistema linguístico do PB (e, também, para comparar este sistema linguístico a outras variedades do português). No âmbito do ensino, como bem pontua Machado (2018), em contextos de variação, é importante que os professores tenham conhecimento dessas variações para "desenvolver procedimentos que façam do aluno um eficiente usuário da língua nos diferentes contextos comunicativos a que diariamente é exposto" (p. 179). Assim, dados como estes instigam a discussões benéficas sobre a determinação do que efetivamente compõe a variedade padrão no PB, bem como sobre diferentes estratégias para aliar as variações observadas nesse sistema ao processo de ensino da língua portuguesa em salas de aula e aos objetivos que esse ensino deve alcançar.

Agradecimentos

Agradece-se à Professor Doutora Ana Madeira e à Profesora Doutora Alexandra Fiéis do Centro de Linguística da Universidade NOVA de Lisboa pelos comentários acerca deste estudo, à Doutora Luísa Pilz de Charité - Universitätsmedizin Berlin pelo auxílio nos testes estatísticos e a todos que se dispuseram a participar neste estudo ou que partilharam o link de acesso a ele. Este trabalho foi financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito da bolsa de doutoramento 2021.05667.BD e do projeto UIDB/03213/2020 e UIDP/03213/2020 – Centro de Linguística da Universidade NOVA de Lisboa (CLUNL).

[39] Agradece-se a um revisor anónimo por esta sugestão.

Referências

Azeredo, José C. (2008) *Gramática Houaiss da língua portuguesa.* Publifolha.

Carneiro, Zenaide O. (2005) Cartas brasileiras: Um estudo linguístico-filológico. Tese de doutoramento, Universidade Estadual de Campinas.

Carneiro, Zenaide O. (2016) Colocação de clíticos em orações finitas em duas vertentes do português oral feirense: um contexto não variável. In Norma L. F. Almeida, Silvana S. F. Araújo & Eliana P. Teixeira (orgs.), *Variação linguística em Feira de Santana – Bahia.* UEFS Editora, pp. 141–174.

Casagrande, Sabrina (2007). *A aquisição do objeto direto anafórico em português brasileiro*. Dissertação de mestrado, Universidade Federal de Santa Catarina.

Casagrande, Sabrina (2010) *A correlação entre aspecto e objeto no PB: uma análise sintático-aquisicionista*. Tese de doutoramento, Universidade Estadual de Campinas.

Cereja, William, Carolina Dias-Vianna & Christiane Damien (2016) *Português contemporâneo: Diálogo, reflexão e uso* (livro do professor). Editora Saraiva.

Chartrand, Tanya L. & Valerie E. Jefferis (2004). Priming. In Michael S. Lewis-Beck, Alan Bryman & Tim Futing Liao (orgs.), *The SAGE encyclopedia of social science research methods*. Sage Publications, pp. 854–855. https://doi.org/10.4135/9781412950589.n747

Corrêa, Vilma R. (1991) *Objeto direto nulo no português do Brasil*. Dissertação de mestrado, Universidade Estadual de Campinas.

Costa, João, Alexandra Fiéis & Maria Lobo (2016) A aquisição dos pronomes clíticos no português L1. In Ana M. Martins & Ernestina Carrilho (orgs.), *Manual de linguística portuguesa*. De Gruyter, pp. 365–386.

Cruz, Raiana C. D. & Cristiane Namiuti (2019) A subida de clítico no português brasileiro: O caso das passivas. *ID on line. Revista Multidisciplinar e de Psicologia* 13 (44), pp. 393–403. https://doi.org/10.14295/idonline.vl3i44.1626

Duarte, Maria E. (2020) Aspetos contrastivos entre o português do Brasil e o português europeu. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do Português* (Vol. 3). Fundação Calouste Gulbenkian, pp. 2732–2779.

Gerasimova, Anastasia & Ekaterina Lyutikova (2020) Intralingual variation in acceptability judgments and production: Three case studies in Russian grammar. *Frontiers in Psychology* 11 (348). https://doi.org/10.3389/fpsyg.2020.00348

Gries, Stefan T. (2012) Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges. In Joybrato Mukherjee & Magnus Huber (eds.), *Corpus linguistics and variation in English*. Brill, pp. 41–63. https://doi.org/10.1163/9789401207713_006

Grohmann, Kleanthes K., Elena Papadopoulou & Charalambos Themistocleous (2017) Acquiring clitic placement in bilectal settings: Interactions between social factors. *Frontiers in Communication* 2. https://doi.org/10.3389/fcomm.2017.00005

Gu, Wenjun (2019) Aquisição de pronomes clíticos de português europeu por falantes de chinês: Dados sobre a colocação. *Revista da Associação Portuguesa de Linguística* 5, pp. 190–206. https://doi.org/10.26334/2183-9077/rapln5ano2019a14

Kato, Mary A. (2005) A gramática do letrado: questões para a teoria gramatical. In Maria A. Marques, Erwin Koller, José Teixeira & Aida S. Lemos (orgs.), *Ciências da linguagem: 30 anos de investigação e ensino*. Centro de Estudos Humanísticos da Universidade do Minho, pp. 131–145.

Kato, Mary A. (2017) A variação no domínio dos clíticos no português brasileiro. *Linguística* 33 (1), pp. 135–152. https://doi.org/10.5935/2079-312x.20170009

Kayne, Richard S. (1975) *French syntax: The transformational cycle.* MIT Press.

Krashen, Stephen D. (1981) *Second language acquistion and second language learning.* Pergamon.

Lacerda, Mariana F. O., Maiara S. Lemos & Zenaide O. Carneiro (2021) A colocação dos clíticos em sentenças finitas: Um estudo sócio-histórico das vertentes do PB em cartas do sertão baiano (século XX). *Confluência* 61, pp. 229–334. https://doi.org/10.18364/rc.2021n61.426

Leivada, Evelina, Paraskevi Mavroudi & Anna Epistithiou (2010) Metalanguage or bidialectism acquisition of clitic placement by Hellenic Greeks, Greek Cypriots and binationals in the diglossic context of Cyprus. In *Proceedings of the 3rd ISCA Workshop ExLing 2010*. ISCA & University of Athens, pp. 97–100.
Lobo, Tânia (2002) A sintaxe dos clíticos: O século XVI, o século XX e a constituição da norma padrão. In Rosa V. Mattos e Silva & Américo V. L. Machado Filho (orgs.), *O português quinhentista: Estudos linguísticos.* EDUFBA, pp. 8–101.
Luft, Celso P. (1987) *Moderna Gramática Brasileira.* Editora Globo.
Luís, Ana R. & Georg A. Kaiser (2016) Clitic pronouns. In W. Leo Wetzels, João Costa & Sérgio Menuzzi (orgs.), *The handbook of Portuguese linguistics*. Wiley-Blackwell, pp. 210–233. https://doi.org/10.1002/9781118791844.ch12
Machado, Ana C. (2006) *O uso e a ordem dos clíticos na escrita de estudantes da cidade do Rio de Janeiro*. Dissertação de mestrado, Universidade Federal do Rio de Janeiro.
Machado, Ana C. (2018) O uso e a ordem dos clíticos na escrita de estudantes da cidade do Rio de Janeiro. In Alessandra de Paula, Danielle Kely Gomes, Eliete Figueira Batista da Silveira, Marcia dos Santos Machado Vieira & Silvia Rodrigues Vieira (orgs.), *Uma história de investigações sobre a língua portuguesa: Homenagem a Silvia Brandão*. Blucher, pp. 167–182. https://doi.org/10.5151/9788580393088-11
Madeira, Ana M. & Maria F. Xavier (2009) The acquisition of clitic pronouns in L2 European Portuguese. In Acrísio Pires & Jason Rothman (orgs.), *Minimalist inquiries into child and adult language acquisition*. Mouton de Gruyter, pp. 273–299.
Mapasse, Ermelinda L. A. (2005) *Clíticos pronominais em português de Moçambique*. Dissertação de mestrado, Universidade de Lisboa.
Martins, Ana M. (2002) Tipologia e mudança linguísticas: Os pronomes pessoais do português e do espanhol. *Santa Barbara Portuguese Studies VI*, pp. 340–386.
Martins, Ana M. (2013) A posição dos pronomes pessoais clíticosIn Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do português* (Vol. 2). Fundação Calouste Gulbenkian, pp. 2231–2302.
Martins, Ana M. (2016) A colocação dos pronomes clíticos em sincronia e diacronia. In Ana M. Martins & Ernestina Carrilho (orgs.), *Manual de linguística portuguesa*. De Gruyter, pp. 401–430.
Martins, Ana M. (2021) A "língua desportuguesa". Próclise no português angolano e no português moçambicano. *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*, Volume especial in honorem Ana Maria Barros de Brito, pp. 71–97.
Morais, Maria A. T. & Ilza Ribeiro (2005) Contraste da sintaxe dos clíticos no português europeu e português brasileiro. *Linha D'Água* 19, pp. 19–47. https://doi.org/10.11606/issn.2236-4242.v0i17p19-47
Moura, Ana C. C. (2001) Como as crianças usam o clítico em frases imperativas no discurso direto? *Revista entreideias: Educação, cultura e sociedade* 6 (5), pp. 141–151. https://doi.org/10.9771/2317-1219rf.v6i5.2844
Mutali, Henrique S. (2019) *A colocação dos pronomes clíticos no português angolano escrito*. Dissertação de mestrado, Universidade de Lisboa.
Myers, James (2017) Acceptability judgment. In *Oxford Research Encyclopedia of Linguistics*, pp. 1–72. Disponível em https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-333
Nunes, Jairo (2007) Triangulismos e a sintaxe do português brasileiro. In Ataliba T. Castilho, Maria A. Morais, Ruth E. V. Lopes & Sônia M. L. Cyrino (orgs.), *Descrição, aquisição e história do português brasileiro*. Pontes & FAPESP, pp. 25–33.
Nunes, Jairo (2015) De clítico a concordância: O caso dos acusativos de terceira pessoa em português brasileiro. *Cadernos de Estudos Linguísticos* 57 (1), pp. 61–84. https://doi.org/10.20396/cel.v57i1.8641472
Pereira, Ronan (2022) A aquisição de clíticos em português europeu L2 e a Hipótese de Reconfiguração dos Traços. *Diacrítica* 36 (1), pp. 108–132. https://doi.org/10.21814/diacritica.698

Perini, Mário A. (2010) *Gramática do português brasileiro.* Parábola.
Reis, Fernanda E. B. (2011) *A perda da subida de clítico no português brasileiro: séculos XIX e XX*. Dissertação de mestrado, Universidade Estadual de Campinas.
Romeo, Rogelio P. L. (2019) Critérios descritivos e prescritivos na colocação dos pronomes pessoais átonos na gramaticografia da língua portuguesa durante o século XIX. In Clarinda A. Maia & Isabel A. Santos (orgs.), *Estudos de linguística histórica: Mudança e estandardização*. Imprensa da Universidade de Coimbra, pp. 329–352. https://doi.org/10.14195/978-989-26-1756-5_10
Santos, Maria F. N. (2002) *Os pronomes pessoais átonos no português europeu. Descrição de problemas que ocorrem no 3º ciclo e proposta de actividades didáticas*. Dissertação de mestrado, Universidade de Lisboa.
Soriano, Olga F. (2015) Clíticos. In Javier Gutiérrez-Rexach (org.), *Enciclopedia de Lingüística Española*. Routledge, pp. 423–436.
Tarallo, Fernando A. (1996) Diagnosticando uma gramática brasileira: o português d'aquém e d'além mar no final do século XIX. In Ian Roberts & Mary A. Kato (orgs.), *Português brasileiro: Uma viagem diacrônica*. Editora da UNICAMP, pp. 69–105.
Tomaz, Margarida, Maria Lobo, Ana Madeira, Carla Soares-Jesel & Stéphanie Vaz (2019) Omissão e colocação de clíticos por crianças bilingues Português-Francês. *Revista da Associação Portuguesa de Linguística* 5, pp. 385–412. https://doi.org/10.26334/21839077
Vandresen, Paulino (2004) Os clíticos no português da fronteira gaúcha: Chuí, Jaguarão e Pelotas. *Anais da XX Jornada - GELNE*. Idéia, pp. 2083–2090.
Vieira, Sílvia R. (2008) Colocação pronominal. In Sílvia R. Vieira & Sílvia F. Brandão (orgs.), *Ensino de gramática: Descrição e uso*. Contexto, pp. 121–146.

Influências estilísticas e sociais na variação dos possessivos em 3.ª pessoa no *corpus* D&G Natal

Mariana Lorena dos Santos Silva[1]

[1]Universidade Nova de Lisboa (UNL-FCSH)

Resumo

Este estudo é resultado da análise da variação linguística entre as formas possessivas SEU e DELE (e suas flexões) em Natal/RN, no final do século XX, com ênfase para os aspectos sociais e estilísticos, enquanto fatores extralinguísticos condicionantes de uso dessas variantes. Foram selecionados 40 textos orais e suas respectivas versões escritas, extraídos do *Corpus* Discurso e Gramática (D&G), e esses dados estão igualmente estratificados em modalidade da língua, gênero/sequência textual, idade/escolaridade e sexo. Verificou-se, por meio de análises quantitativas, fornecidas pelo Programa Estatístico Goldvarb, que a forma DELE(A)(S) se destacou na oralidade, nos gêneros/sequências de esfera narrativa, entre os indivíduos de menor idade/menos escolarizados, e no grupo de homens. Dentre esses fatores, o de maior influência foi com relação à modalidade da língua. E a partir das generalizações sociolinguísticas apresentadas, pode-se concluir que o fenômeno de variação entre as formas possessivas de 3.ª pessoa do singular, quanto aos fatores analisados, indicava, àquela altura, uma *variação estável* entre as variantes.

Palavras-chave: Variação sociolinguística, pronomes possessivos, terceira pessoa do singular, condicionamentos estilísticos e sociais.

Abstract

This study is the result of an analysis of the linguistic variation between the possessive forms SEU and DELE (and their inflections) in Natal/RN, at the end of the 20th century, with emphasis on social and stylistic aspects, as extralinguistic factors that condition usage. 40 oral texts and their respective written versions were selected, extracted from the *Corpus* Discurso e Gramática (D&G), and these data are equally stratified in language modality, gender/textual sequence, age/education and gender. It was verified, through quantitative analyzes provided by the Statistical Program Goldvarb, that the DELE(A)(S) form stands out in orality, in genres/sequences of the narrative sphere, among younger/less educated individuals, and among men. Among these factors, the one with the greatest influence was related to language modality. And from the sociolinguistic generalizations, it could be concluded that the phenomenon of variation between the possessive forms of the 3rd person singular, regarding the analyzed factors, indicated, at that time, a *stable variation* between the variants.

Keywords: Variationist sociolinguistics, possessive pronouns, third person singular, social and stylistic constrains.

1. **Introdução**

A presente produção é resultado da análise[1] da variação entre as formas possessivas SEU/SUA(S) e DELE/DELA na indicação de posse em 3.ª pessoa do singular,[2] na cidade de Natal/RN-Brasil,[3] com ênfase para os aspectos sociais e estilísticos, e baseada em um banco de dados com textos orais e escritos, do final do século XX: o *corpus* Discurso e Gramática (conhecido como D&G Natal).

A discussão acerca das formas possessivas de 3.ª pessoa ainda não se findou, visto que a literatura aponta dados de diferentes regiões, em distintos períodos e a partir de *corpus* de estruturações diferentes. No Brasil, há trabalhos relativos ao uso de SEU e DELE em: Rio de Janeiro/RN (Silva, 1991), Belo Horizonte/MG (Rocha, 2009), Curitiba, Irati, Pato Branco e Londrina/PR (Soares, 1999), Florianópolis/SC (Sbalqueiro, 2005), Natal/RN (Silva, 2016), Alagoas/SE (Freitag, 2017), além de outras importantes contribuições como as de Castro (2006), Cunha (2007), Morais (2019), e Lopes & Guedes (2020) relevantes para a construção de atualização do estágio de variação das formas possessivas em 3.ª pessoa.

Com relação à ideia de posse no paradigma dos pronomes possessivos, tanto no português europeu (PE), quanto no português brasileiro (PB), tem-se a forma simples e a preposicionada:

(1a) o **seu** filho
(1b) o filho **dela**

Outro fator observado quanto ao uso das variantes SEU ou DELE é a posição da posse relativamente ao nome, pois a forma preposicionada só ocorre posposta ao referente, enquanto a forma simples pode ocorrer depois (em muito menor frequência) ou antes do nome.

Embora DELE seja considerada a forma inovadora, os itens SEU e DELE já estão presentes em textos do século XIII e XIV (Cunha, 2007). Porém, no século XVIII, a forma VOCÊ passou a ser utilizada como pronome pessoal de 2.ª pessoa (Cerqueira, 1996; Kato, 1985; Perini, 1985) e esse processo linguístico alterou não apenas a classe dos pronomes pessoais, mas também a de possessivos (Silva, 1991). Assim, sem haver elementos do contexto para deixar as relações de posse bem definidas, a substituição da forma simples pela preposicionada surge como uma estratégia de desambiguação (Menon, 1995):

(2a) Esse advogado estava chegando ao **seu** escritório.[4]
(2b) Esse advogado estava chegando ao escritório **dele**.

SEU/DELE e suas flexões são consideradas, nesse sentido, "formas variantes", por serem normalmente utilizadas para a codificação do mesmo significado ou da mesma função em uma dada comunidade de fala e em um dado período de tempo (cf. Tagliamonte, 2006, 2012). E a esse conjunto de formas variantes (SEU, SEUS, SUA, SUAS, DELE, DELA) em 3.ª pessoa do singular, atribui-se o rótulo de "variável linguística".

Conforme exemplo de oralidade a seguir, presente no D&G Natal, a variável linguística da expressão de posse em 3.ª pessoa do singular, com referência ao mesmo item lexical (chácara), alterna entre as formas simples e preposicionada:

(5) então logo após o fuzilamento ... né ... dos dez caras lá ... houve uma retomada da França ... então a ocupação nazista foi retirada da França ... então esse cara ficou na rua ... quando ele voltou ... ele não tinha mais nada ... ele não tinha mais escritório ... ele não tinha mais

[1] Sistematização do trabalho de Mestrado da autora (Silva, 2016).
[2] As flexões de plural DELES/DELAS não foram consideradas, pois o referente de 3.ª pessoa do plural não gera ambiguidade com relação a 2.ª pessoa, como ocorre na variação entre SEU e DELE. Nesta pesquisa foram analisadas as formas em 3.ª pessoa do singular.
[3] Natal é a capital do estado do Rio Grande do Norte (RN), localizado na região nordeste do Brasil.
[4] Dado extraído do *Corpus* Discurso & Gramática (D&G): A língua falada e escrita na cidade de Natal (D&G Natal).

> profissão ... ele não tinha mais fazenda ... a chácara **dele** lá no subúrbio ... então ele ficou desesperado ... até que ele teve uma ideia de retornar à **sua** chácara ... para ver como estava ... (Texto da oralidade, informante do sexo feminino, gênero narrativa recontada).

A partir da sistematização das variáveis em estudo, por meio de análises quantitativas, pode-se observar os padrões de distribuição em diferentes contextos de uso, ao identificar influências de naturezas linguísticas e extralinguísticas que estejam subjacentes à utilização de tais formas. Os fatores linguísticos apontam conclusões que acrescentam à discussão da variação dos possessivos. No entanto, é preciso destacar que não há cruzamento de dados na ordem linguística sobre a seleção das formas variantes neste trabalho, o que se deve à decisão de centrar e aprofundar a análise das influências estilísticas (modalidade da língua e gênero textual) e sociais (sexo, idade e escolaridade) para o fenômeno. A utilização de formas variantes pode ser influenciada não somente por fatores linguísticos, mas também e, às vezes, especialmente, por fatores sociais, como sexo, idade, escolaridade, etnia e/ou pelo estilo (cf. Labov, 2008/1972).

2. **Enquadramento teórico-metodológico**

Há trabalhos pioneiros, que avaliam quais os fatores linguísticos e sociais que determinam o emprego das variantes SEU e DELE (Almeida, 1993; Silva, 1982, 1984, 1991, 1998a, 1998b). Outros desconsideram esses fatores e defendem que há especialização de uso das formas (Negrão & Müller, 1996). E, ainda, trabalhos que discutem a posição estrutural das formas de possessivo (Cerqueira, 1993, 1996; Müller, 1997).

SEU é o pronome possessivo canônico na sua forma simples, na 3.ª pessoa do singular; enquanto DELE é a forma preposicionada também indicativa de posse na 3.ª pessoa do singular, porém, para alguns contextos de uso especializados: há especialização na medida em que as formas em uso ocorrem em contextos linguísticos específicos e diferentes (Hopper; Traugott, 2003). No conceito de especialização, os possessivos SEU e DELE não são equivalentes. SEU é a forma com comportamento de variável ligada à retomada de antecedentes não referenciais, entre eles, quantificadores, genéricos, indefinidos, enquanto DELE é a forma escolhida para retomar antecedentes referenciais e expressar a correferência. (Cf. Menuzzi 2003a, 2003b; Müller, 1997; Negrão & Müller, 1996).

De maneira geral, a forma DELE/DELA não é considerada como *pronome possessivo*. Alguns autores usam as nomenclaturas "construções possessivas perifrásticas" (Marcotulio et al., 2015), "de-possessivos ou possessivos preposicionados" (Castro, 2006), "forma analítica" (Rocha, 2009), "forma genitiva" (Coelho, 2020), ou sequência "de+ele" enquanto item lexical (Cunha, 2007), por exemplo.

Para Perini (1985), a substituição de *tu* e *vós* pelas formas *você* e *vocês* tornou ambíguo o pronome possessivo de 3.ª pessoa SEU. Depois, Perini (2010) e outros autores acrescentaram que DELE/DELA(S) passou a exercer a função de posse em 3.ª pessoa como estratégia de desambiguação. (cf. Cerqueira, 1996a; Galves, 1985; Kato, 1994; Silva, 1984). Da mesma maneira, a explicação para a presença de SEU/SUA(S) na 2.ª pessoa do plural carece de ser explicada pela substituição do pronome pessoal *vós* por *vocês*, e consequentemente, na alteração do possessivo *vosso* por SEUS/SUAS.

A concretização de uma mudança linguística deve levar em consideração os casos de *mudança em tempo real* ou de *mudança em tempo aparente*. Para saber se há uma mudança em tempo real, o pesquisador faz o rastreamento do processo histórico de mudança em períodos distintos da língua, e consegue averiguar, por exemplo, se a forma variante inovadora teve sua taxa de ocorrência aumentada com o passar do tempo. Já a verificação de indícios de mudança em tempo aparente envolve avaliar possíveis diferenças nos padrões de uso das formas variantes entre gerações distintas de falantes de uma mesma comunidade (cf. Labov, 1994). Os estudos sobre a *mudança em tempo aparente* são, portanto, de grande importância para a pesquisa linguística, pois deles deriva uma crescente ampliação do conhecimento quanto às motivações e os mecanismos da mudança (Bailey, 2004).

Todavia, há, naturalmente, muitos casos de variação em que não se identifica mudança; esses casos são considerados estando em *variação estável*, e revelam como a variabilidade é inerente à língua, ao invés de ser simplesmente uma transição de um estado do sistema linguístico para o outro (Tagliamonte, 2012, p. 55).

Com base em padrões de distribuição de formas variantes recorrentes em pesquisas linguísticas ao longo do tempo, alguns fatores sociais vão condicionar o uso das formas variáveis (cf. Labov, 2001; Tagliamonte, 2012):

- indivíduos mais escolarizados costumam fazer maior uso de variantes mais formais devido ao maior contato com a cultura letrada nos bancos escolares – e mesmo fora deles;
- indivíduos mais velhos costumam fazer maior uso de variantes mais formais, normalmente, por elas serem mais antigas na língua do que as variantes mais informais;
- quanto ao sexo, as mulheres tendem a fazer maior uso de variantes mais formais, que, em geral, são as mais prestigiadas na comunidade de fala. E ainda que a mudança se dê na direção de estruturas menos formais, são elas que a lideram.

Sendo assim, estudar a diversidade linguística prevê não apenas explicar as motivações de utilização da linguagem, mas também destacar a importância do contexto social onde as pessoas que fazem uso dela estão inseridas, pois das interações sociais é que essa diversidade se dá (Le Page, 1998). À luz disso, os fatores sociais e estilísticos têm enfoque neste trabalho.

2.1. **A questão estilística**

A expressão "variação estilística" corresponde à alternância dos estilos adotados pelo falante em uma dada situação, devido a alguma modificação em um ou mais dos fatores que podem influenciar a troca de estilo (como mudança de tópico/assunto, de gênero textual, do grau de envolvimento emocional do falante relativamente ao que diz, entre outros) (Görski & Valle, 2014, p. 70).

Nesta discussão, destarte, três critérios terão influência na análise de variação linguística quanto à questão estilística: o nível de formalidade da situação comunicativa; os gêneros textuais; e a modalidade da língua que os indivíduos se comunicam.

Quanto à noção de formalidade, Labov propôs uma escala de estilos que tem como ponto de partida o vernáculo ou fala casual, com menor nível de formalidade, "o estilo em que se presta o mínimo de atenção ao monitoramento da fala" (Labov, 2008/1972, p. 244), ou seja, o estilo que traz a manifestação mais espontânea da língua. Já o ponto de chegada dessa escala são os estilos mais formais, em que o falante tende a monitorar com mais atenção o modo como diz, e assim, há uma maior recorrência ao emprego de formas variantes prestigiadas pela comunidade.

Os usuários da língua costumam ajustar a sua fala e a sua escrita ao grau de formalidade requerido por contextos de comunicação distintos, como a familiaridade do falante com o(s) ouvinte(s); as características socioculturais dos interlocutores (idade, sexo, etnia, classe social, nível de escolaridade, profissão etc.); o tópico/assunto tratado (política, religião, família, infância, esporte, namoro, economia, lazer, etc.); o domínio em que se dá a prática social (lar, trabalho, escola, clube, igreja, bar, shopping, praia, tribunal, audiência pública, fila de banco etc.); os papéis socioculturais assumidos no momento da interação (amiga-amiga, esposa-marido, mãe-filha, patroa empregada, professora-aluna, entrevistadora-entrevistada, etc.); o maior ou menor envolvimento emocional do falante com o que diz; o gênero textual (Tavares, 2014, p. 207).

No que se refere aos gêneros textuais, eles costumam ser caracterizados especialmente em suas funções comunicativas, cognitivas e institucionais. Então, ao descrevermos e analisarmos um gênero textual, não estamos tratando de uma forma linguística; ao invés, estamos abordando uma forma de

concretizar linguisticamente objetivos específicos em determinados contextos de comunicação (cf. Marcuschi, 2002; 2003). Além disso, a narrativa de experiência pessoal tende a ser um dos gêneros textuais mais marcadamente informais, pois envolve a narração de fatos emocionantes, assustadores ou ao menos interessantes pelos quais passou o próprio indivíduo que narra (cf. Labov, 2004).

A narrativa de experiência pessoal, a narrativa recontada e relato de opinião podem ser considerados gêneros textuais, enquanto a descrição de local não é um gênero textual em si, mas a aplicação da sequência textual descritiva à descrição de um lugar, o que, na prática, pode acontecer em gêneros textuais variados. O relato de procedimentos engloba um conjunto de gêneros instrucionais: alguns informantes produziram receitas culinárias, mas outros relataram diferentes tipos de procedimentos, como os envolvidos na pintura de um quadro e no tratamento da água em uma estação (Silva, 2016).

Em virtude dessas questões, nesta pesquisa, não apenas o gênero foi visto como um possível elemento influenciador da variação, mas também a modalidade da língua em que se os informantes produziram os textos. Ademais, linguistas históricos assumem que as formas inovadoras chegam à modalidade escrita da língua segundo a ordem em que apareceram na oralidade (Pintzuk, 2003). E além disso, a "linearidade nas línguas deve ser revista", dada a complexidade inerente a cada tempo (cf. Mattos e Silva, 2008, p. 41) de utilização da língua pelos falantes.

3. **Questões de investigação e hipóteses**

Questão 1: No que se refere à variação estilística, em relação à distribuição da indicação de posse na 3.ª pessoa do singular SEU(A)(S) e DELE(A), a) qual variante tende a ser utilizada com maior frequência, a depender da modalidade escrita ou oral da língua; b) qual variante tende a ser utilizada com maior frequência, a depender dos gêneros/sequências em que estão classificados?

Hipótese 1: SEU(A)(S), por seu caráter mais formal, predomina na escrita e em gêneros/sequências textuais da esfera não narrativa, ao passo que DELE(A)(S), por seu caráter mais informal, destaca-se na oralidade e em gêneros/sequências da esfera narrativa. Fundamenta-se esta hipótese no fato de que a modalidade oral, em contextos de interação cotidiana, é geralmente favorecedora de formas mais informais, em contraponto à modalidade escrita, mesmo quando a escrita ocorre em contextos de menor monitoramento, como nos textos escritos do *Corpus* D&G, que contrastam, por exemplo, com textos escritos dos domínios jornalístico e acadêmico, predominantes em contextos de maior monitoramento. Relativamente ao gênero/sequência textual, estudos feitos sob a égide da sociolinguística variacionista têm mostrado que gêneros/sequências da esfera narrativa tendem a favorecer o emprego de formas mais informais, ao passo que gêneros/sequências de esfera não narrativa tendem a favorecer o emprego de formas mais formais (cf. Görski, Coelho & Souza, 2014; cf. Labov, 2001b, 2004).

Questão 2: Quanto à variação social, em relação à distribuição dos pronomes possessivos de 3.ª pessoa do singular SEU(A)(S) e DELE(A), a) qual variante é influenciada em maior frequência pela variável idade/escolaridade?; b) qual variante é influenciada em maior frequência pela variável sexo?

Hipótese 2: SEU(A)(S), por seu caráter mais formal, predomina entre os indivíduos de mais idade e maior escolarização e entre as mulheres, ao passo que DELE(A), por seu caráter mais informal, destaca-se entre os indivíduos de menos idade e menor escolarização e entre os homens. Esta hipótese se baseia em padrões de distribuição de formas variantes que vêm sendo encontrados por estudos variacionistas ao longo do tempo: (i) indivíduos mais escolarizados costumam fazer maior uso de variantes mais formais devido ao maior contato com a cultura letrada nos bancos escolares – e mesmo fora deles –, e (ii) indivíduos mais velhos costumam fazer maior uso de variantes mais formais normalmente por elas serem mais antigas na língua do que as variantes mais informais, e, em razão disso, predominam na fala dos indivíduos mais jovens, especialmente em situações de mudança geracional em progresso direcionadas para o aumento de uso de formas inovadoras. Quanto ao sexo, as mulheres tendem a fazer maior uso de variantes mais formais, que, em geral, são as mais prestigiadas

na comunidade de fala. Já em situações de mudança linguística em progresso, as mulheres tendem a liderar a mudança independentemente de esta se dar na direção de estruturas mais formais ou menos formais (que, em geral, ainda que não sejam estigmatizadas, são as menos prestigiadas na comunidade de fala) (cf. Labov, 2001; Tagliamonte, 2012).

Questão 3: Em Natal, na última década do século XX, o fenômeno de variação entre os pronomes possessivos de terceira pessoa do singular SEU(A)(S) e DELE(A) pode ser caracterizado como um caso de variação estável ou mudança em tempo aparente?

Hipótese 3: O fenômeno de variação entre os pronomes possessivos de terceira pessoa do singular SEU(A)(S) e DELE(A) em Natal pode ser caracterizado como um caso de mudança em tempo aparente que se reflete no maior uso da forma mais recente, DELE(A), pelos informantes mais jovens (de 18 a 20 anos), e o maior uso da forma canônica, SEU(A)(S), pelos informantes mais velhos (acima de 23 anos) (cf. Labov, 1994).

4. **Procedimentos metodológicos**

4.1. **Descrição do corpus**

O D&G Natal – *Corpus* Discurso & Gramática: A língua falada e escrita na cidade de Natal (Furtado da Cunha, 1998) é composto por entrevistas feitas com 20 informantes dessa cidade, capital do estado do Rio Grande do Norte, no nordeste do Brasil (cf. Tabela 1). A escolha desse *corpus* se justifica por se tratar de um banco de dados completo e totalmente transcrito, sob o recorte temporal do final do século XX, ao passo que os dados são estratificados homogeneamente de acordo com a idade, o nível escolaridade e o sexo:

Tabela 1. Distribuição dos informantes do *corpus* D&G Natal de acordo com as células sociais (Furtado da Cunha, 1998)

Idade	Escolaridade	Sexo
5 a 8 anos	Alfabetização Infantil	2 homens e 2 mulheres
9 a 11 anos	Ensino Fundamental I	2 homens e 2 mulheres
13 a 16 anos	Ensino Fundamental II	2 homens e 2 mulheres
18 a 20 anos	Ensino Médio	2 homens e 2 mulheres
Acima de 23 anos	Ensino Superior	2 homens e 2 mulheres

Cada informante produziu cinco textos orais e suas respectivas versões escritas. No total, o *corpus* é formado por 200 registros (100 produções orais, gravadas e transcritas), e de acordo com os seguintes gêneros textuais: narrativa de experiência pessoal, narrativa recontada, descrição de local, relato de procedimento e relato de opinião. Para esta pesquisa, no entanto, foram selecionados 40 textos orais e seus 40 textos escritos correspondentes, sendo considerados apenas o grupo de homens e mulheres dos dois mais altos níveis de escolaridade (Ensino Médio e Ensino Superior).

Pode-se notar que nesse *corpus* existe uma correlação estreita entre a idade e a escolaridade: classe de Alfabetização Infantil = de 5 a 8 anos; Ensino fundamental I = de 9 a 11 anos; Ensino fundamental II = de 13 a 16 anos; Ensino Médio = de 18 a 20 anos; Ensino Superior = acima de 23 anos – sempre com informantes dos anos finais desses níveis. Porém, tem-se ciência de que o formato de distribuição impede que, em caso de haver condicionamento de uso das variantes no fator idade, esteja claro que a influência não foi exercida, na verdade, pela escolaridade.

4.2. **Procedimentos para a recolha de dados**

Foram selecionados para esta análise 8 participantes, os quais produziram 5 gêneros textuais, por meio de entrevistas orais, e suas respectivas versões escritas. Os informantes foram orientados sobre como as entrevistas ocorreriam e prepararam previamente os tópicos que iriam abordar em relação a cada gênero textual solicitado

pelo entrevistador. No que se refere à versão escrita, por sua vez, a composição foi ainda mais facilitada por já haver uma versão prévia oral relatada ao entrevistador.

É fato que as versões orais das entrevistas são relativamente mais longas que as versões escritas, até porque, na versão oral, ocorre a participação do entrevistador com intervenções e constante estímulo para receber mais informações do informante (há pausas, hesitações) – e isso não acontece nas versões escritas.

Após a exclusão dos dados que não correspondiam ao objeto de estudo das formas possessiva e 3.ª pessoa do singular, foi verificado um total de 390 dados, dos quais 146 (37%) representaram o uso da variante SEU(A)(S), enquanto 244 (63%) indicaram o número de ocorrências da variante DELE(A), conforme o Gráfico 1.

Gráfico 1. Percentual de ocorrências de SEU(A)(S) E DELE(A) no D&G Natal

Com o auxílio do Programa Estatístico Goldvarb X (cf. Sankoff et al., 2005), o qual fornece frequências e pesos relativos referentes a cada contexto de uso das variantes, averiguou-se a influência de cada um dos fatores controlados sobre o uso de cada uma das formas variantes. Ademais, o Goldvarb X apresenta a ordem de significância dos grupos de fatores controlados, ao mostrar quais desses grupos são mais relevantes.

Os grupos de fatores foram selecionados na seguinte ordem de relevância: em primeiro lugar, a modalidade da língua; em segundo lugar, o gênero textual; em terceiro lugar, a idade/escolaridade; e, em quarto e último lugar, o sexo.

O peso relativo (P.R.), oferecido pelo programa, é uma medida multidimensional derivada do controle simultâneo de vários grupos de fatores condicionadores da variação linguística. Em uma análise multivariada, assim, como a efetuada pelo Programa Estatístico Goldvarb X, "cada efeito de um fator na análise é calculado enquanto são controlados, até o máximo possível, os outros fatores" (Guy & Zilles, 2007, p. 100), O peso relativo varia entre 0.000 a 1.000. Quanto mais próximo de 0.000, menos influente é o fator que o recebeu; quanto mais próximo de 1.000, mais influente é o fator que o recebeu. Um peso que gira em torno de 0.500 tende a ser indiferente.

5. **Análise dos dados**

A divisão para a análise se deu em dois grupos de natureza estilística (modalidade da língua e gênero/sequência textual) e dois grupos de natureza social (idade/escolaridade e sexo).

5.1. **Influência da modalidade da língua**

Há ocorrências de SEU(A)(S) e DELE(A) tanto na modalidade oral da língua, quanto na escrita:

(6) eu posso até dizer assim ... é como se ele visse ... ele olhasse pra um lado ... olhasse pra outro e visse tá aqui a solu/ a solução ... tá nas minhas mãos ... a solução do país tá nas minhas mãos

... a solução dos meus filhos futuramente tá nas minhas mãos ... mas ele tem medo de enfrentar ... de encarar a realidade ... de pegar o **seu** direito de voto e dizer assim ... "eu vou usar essa arma" ... (Informante 3, oralidade, feminino, relato de opinião).

(7) o coitado também sofre demais ... sete anos que ele trabalha lá ... sete nada ... sete faz uma sobrinha **dele** ... faz doze anos que trabalha com ele ... (Informante 7, oralidade, feminino, relato de procedimento).

(8) A madre que já estava se tornando amiga da freira, convocou todas do convento e ela entraram na galeria conseguindo enganar o traficante e **seus** capangas, a polícia então conseguiu matá-los. (Informante 8, escrita, feminino, narrativa recontada).

(9) Quando foi à noite foi dormir pensando e no meio da noite acordou assustado com uma voz repetindo a mesma frase, abrio os olhos e lá estava aquele menino que havia morrido nos braços **dele**, na porta do quarto. Foi dormir de novo e quando acordou ficou pensando e saiu para trabalhar. (Informante 1, escrita, masculino, narrativa recontada).

A proximidade entre os textos orais e escritos (uma vez que foram produzidos pelos mesmos indivíduos), permite fazer inferências sobre os resultados obtidos, a respeito da influência da modalidade da língua, com maior confiança. Os resultados obtidos estão expostos na Tabela 2.

Tabela 2. Influência da modalidade da língua sobre o uso de SEU(A)(S) E DELE(A)

	SEU(A)(S)			***DELE(A)***		
MODALIDADE DA LÍNGUA	Apl./Total	%	PR	Apl./Total	%	PR
Oralidade	38/274	14	0.148	236/274	86	0.852
Escrita	108/116	93	0.984	8/116	7	0.016
TOTAL	146/390	37	----	244/390	63	----
	Input: 0.374	Sig: 0.000		Input: 0.626	Sig: 0.000	

A hipótese de que os informantes fariam maior uso da variante mais formal SEU(S)(A) na escrita foi confirmada: houve largo favorecimento do uso de DELE(A) na modalidade ora da língua (com frequência de 86% e P.R. de 0.852) e um intenso desfavorecimento na escrita (com frequência de 7% e P.R. de 0.016). Em contraponto, SEU(A)(S*)* tem sua utilização bastante favorecida na modalidade escrita da língua (com frequência de 93% e P.R. de 0.984) e pouca presença na modalidade oral (com frequência de 14% e P.R. de 0.148). Infere-se, portanto, que na comunidade de fala de Natal/RN, no final do século XX, possivelmente havia forte especialização de DELE(A) na modalidade oral da língua, ao lado de uma forte especialização de SEU(A)(S*)* para a modalidade escrita da língua.

Outros trabalhos atestaram resultados similares desse fenômeno de variação, a exemplo de Sbalqueiro (2005), que observou uma grande diferença na proporção de ocorrência dos possessivos em narrativas escritas produzidas por alunos do Ensino Fundamental de Florianópolis: de um total de 637 ocorrências, 575 foram de SEU(A)(S) (90%) e 62 (10%) de DELE(A). E também Rocha (2009), que em um *corpus* oral de Minas Gerais, obteve, de um total de 329 dados na 3.ª terceira pessoa do singular, 257 dados (78%) da forma DELE(A) e 72 (21%) da forma SEU(A)(S). Ambos indicam a predominância da forma DELE em textos orais.

5.2. **A influência do gênero/sequência textual**

A narrativa de experiência pessoal é um relato no qual o informante conta um ou mais fatos alegres ou tristes que se passaram em determinado tempo e lugar, envolvendo a si mesmo e a outros indivíduos, com grande presença de verbos no pretérito perfeito:

(10) Fiquei sabendo que ela era filha de imigrantes alemães e **seus** pais moravam no campo. Ela, como a maioria dos jovens de **suas** redondezas, trabalhava numa cidade maior, Novo Hamburgo. (Informante 4, escrita, masculino).

(11) foram chamar Vilma ... lá vem Vilma super preocupada ... "o que aconteceu ... o que aconteceu? vamos chamar meu pai" ... e telefonou pro pai **dela** ... se a gente saiu ... não ... "a gente vai andando devagar .. porque o pai da menina trabalha logo do outro lado" ... aí ... e o diretor me olhava assim ... ele sabia que eu tava mentindo ... (Informante 2, oralidade, feminino).

A narrativa recontada (narrativa vicária) se trata de uma narração de um ou mais fatos acontecidos com alguém, mas não testemunhados pelo informante, a qual pode envolver fatos reais ou ficcionais (romances, filmes, novelas etc.).

(12) Certo dia o homem ficou só em casa e a mulher **dele** e os filhos saíram para a cidade, o gato foi encontrado morto na beira da estrada, sujo, mais sem ferimentos nenhum,[5] o homem pegou o gato e enterrou no cemitério. (Informante 1, escrita, masculino).

(13) a ocupação pegava transeuntes e levava pra confinamento ... né ... pra aquela prisão ... e chegando lá ... eles saíam é ... eliminando determinados indivíduos ... nesse dia ... esse advogado estava chegando ao **seu** escritório quando foi pego por essa ... pela ocupação ... né ... chegando lá nesse ... na cela ... (Informante 2, oralidade, feminino).

O relato de procedimento se refere à descrição das etapas necessárias à realização de alguma tarefa ou processo, geralmente de conhecimento do informante, caracterizando-se por apresentar ordem cronológica e voltar-se para ações.

(14) I: a gente bota no ... leva ao fogo ... a mesma quantidade de leite ... açúcar ... e maisena ... sabe? aí prepara o mingau ... e no caso de flocos ... passa no liquidificador com claras ... tem o de creme que passa com creme de leite ... o de ameixa que no caso no fogo bota ... a calda de ameixa né ... no mingau ... prepara ... pronto basicamente é essa a receita ...

E: você costuma fazer de quê?

I: de ... morango ... morango e ... é ... é o que eu costumo fazer mais ... até quando eu tinha a essência **dele** ... geralmente de frutas ... natural ... (Informante 8, oralidade, feminino – onde, I: informante, e E: entrevistador).

(15) Um trabalho monocromático pode, por exemplo, recorrer ao uso do grafite ou do nanquim. Este tipo de trabalho, como podemos imaginar, necessita somente de um bom papel, de preferência um de textura média ou correspondente, onde o grafite ou o nanquim deslizem sobre **sua** superfície sem muita alteração da **sua** trama. (Informante 4, escrita, masculino).

Já na descrição de local, o informante apresenta características detalhadas de um lugar em que aprecie estar, e tende a haver presença de adjetivos:

[5] Os trechos foram retirados do *corpus* tal e qual os informantes escreveram.

(16) cidade do interior ... tem um galinheiro próximo ... né ... onde meu tio-avô que gosta de criar galinha ... de ter **sua** criação de galinha ... meu ... meu ... o filho **dele** cria pássaros ... né ... gosta também ... e ... é basicamente isso ... tem ... fica numa estradazinha de barro ... né ... e ... é uma casa meio isolada da ... do centro da ... da cidade ... né ... e fica próxima ao pé da serra (Informante 6, oralidade, masculino).

(17) Falésias de um colorido espetacular que variam do amarelo acre ao terra avermelhado. Mais à frente conseguimos localizar um lugar fantástico. Fica a uns cem metros da pista. Deixamos o carro e subimos uma duna, com vegetação, até o **seu** topo. Qual não foi a nossa surpresa quando olhamos para baixo e encontramos um bosque de árvores muito altas, algumas sem folhas. (Informante 4, escrita, masculino).

E no relato de opinião, o informante tece considerações a respeito de determinado assunto, manifestando sua opinião sobre ele:

(18) por exemplo ... uma Assembleia de Deus totalmente restrita ... que corta assim ... todas as asas do indivíduo pensante ... sabe a religião da Assembleia de Deus ... principalmente ... ela não deixa o indivíduo raciocinar ... ela lhe joga aquele pensamento **dela** e você não:: você apenas aceita ... sem fazer questionamentos... (Informante 2, oralidade, feminino).

(19) No assassinato de Daniela, eu fiquei horrorizada quando fiquei sabendo, logo porque, o principal suspeito era o **seu** companheiro de trabalho que contracenava com ela, na novela de corpo e alma. (Informante 7, escrita, feminino).

Eis os resultados dessa análise na Tabela 3.

Tabela 3. Influência do gênero/sequência textual sobre o uso de SEU(A)(S) E DELE(A)

	SEU(A)(S)			***DELE(A)***		
GÊNERO TEXTUAL	Apl./Total	%	PR	Apl./Total	%	PR
Narrativa de experiência pessoal	13/39	33	0.435	26/39	67	0.565
Narrativa recontada	75/249	30	0.334	174/249	70	0.666
Relato de procedimentos	07/15	47	0.765	08/15	53	0.235
Descrição de local	08/31	26	0.790	23/31	74	0.210
Relato de opinião	43/56	77	0.901	13/56	23	0.099
TOTAL	146/390	37	----	244/390	63	----
	Input: 0.374		Sig: 0.000	Input: 0.626		Sig: 0.000

E a hipótese de que os informantes fariam uso maior uso de SEU(S)(A) em gêneros/sequências textuais não narrativos foi confirmada: DELE(A) predominou nos textos de gêneros/sequências narrativos: narrativa de experiência pessoal e narrativa recontada (com frequências de 67% e 70% e P.R. de 0.565 e 0.666, respectivamente) e não recebeu destaque nos textos de gêneros/sequências não narrativos: relato de procedimentos (com frequência de 53% e P.R. de 0.235), descrição de local (com frequência de 74%, mas P.R. 0.210) e relato de opinião (com frequência de 23% e P.R. 0.099).

É importante notar que ambas as formas tiveram seu maior número de ocorrências nos textos do gênero narrativa recontada: SEU(A)(S) contou com 75 ocorrências, e DELE(A) contou com 174 ocorrências, pois, nesse gênero, o falante relata uma história que se passou com outra pessoa, ou seja, as "narrativas de experiência vicária são apresentadas em terceira pessoa (*ele, ela, eles, elas*) em contraste com a primeira pessoa *eu* característica das histórias de experiência pessoal" (Norrick, 2013, p. 385). Nesse tipo de gênero textual, pois,

em consonância com a centração no sujeito de terceira pessoa, as formas possessivas mais frequentes tendem a ser as de terceira pessoa.

5.3. **A influência do grupo de fator idade/escolaridade**

No D&G, o grupo de fator idade está diretamente relacionado à escolaridade, e embora esteja estratificado em todos diferentes níveis, foram selecionados apenas os informantes de maior idade/escolaridade, uma vez que os outros grupos apresentaram mínima ou nenhuma ocorrência da forma simples SEU ou flexões. Talvez isso se deva ao fato de que esses indivíduos, em especial os adolescentes e os pré-adolescentes, tendem a ser os mais inovadores em uma comunidade de fala (cf. Labov, 2001). Além disso, os indivíduos mais escolarizados podem sofrer pressão do mercado de trabalho para a utilização de formas prestigiadas da língua.

Seguem exemplos de ocorrências de SEU(A)(S) e DELE(A) encontradas na fala e na escrita de informantes dos dois níveis de escolaridade e das duas faixas etárias aqui levadas em conta, seguidos, da Tabela 4, com as ocorrências verificadas:

(20) O Ricardo consegue ver a letra de Isabel quando ela, num ato de desespero total, tenta um suicídio e deixa um bilhete para Ricardo. Como ele já tinha acabado o namoro fica louco e corre para salvar Isabel, depois de ter comparado a letra da carta e das poesias. Ele se declara para ela, já recuperada, e ela também confessa **sua** paixão por ele e finalmente acabam juntos. (Informante 2, escrita, masculino, narrativa recontada – Ensino Médio).

(21) então ... ali tinha uma árvore muito bonita ... uma árvore antiga que hoje em dia só existe ... pedacinhos **dela** ... porque ela foi corroída pelo tempo ... e eu acho que deu cupim rápido [...] rapaz que pássaro lindo ... ele cantou um canto ... que me parecia um lamento ou sei lá um ... uma alerta à natureza de que ele ... de que ele tava sendo ameaçado na sua vida, no seu processo de sobrevivência ... eu parei a minha corrida e pedi desculpas a ele por todos nós. (Informante 4, oralidade, masculino, descrição de local – Ensino Superior).

Tabela 4. Influência da idade/escolaridade sobre o uso de SEU(A)(S) E DELE(A)

	SEU(A)(S)			***DELE(A)***		
IDADE/ESCOLARIDADE	Apl./Total	%	PR	Apl./Total	%	PR
18 a 20 anos/Ensino Médio	45/197	23	0.240	152/197	77	0.760
+ de 23 anos/Ensino Superior	101/390	52	0.764	92/193	48	0.236
TOTAL	146/390	37	----	244/390	63	----
	Input: 0.374		Sig: 0.000	Input: 0.626		Sig: 0.000

A hipótese de que os informantes de 18 a 20 anos e do Ensino Médio fariam maior uso de DELE(A) por ser a variante mais recente e marcada estilisticamente como informal foi confirmada com ressalvas. Isso porque esse grupo está enviesado no D&G, ou seja, idade e escolaridade são controladas em conjunto, o que significa que os resultados obtidos podem se dever mais à influência de um desses fatores do que do outro.

No entanto, conforme o esperado, há mais ocorrências de DELE(A), enquanto forma mais recente e menos formal, entre os informantes mais jovens e menos escolarizados (com frequência de 77% e P.R. de 0.760), e menos ocorrências entre os informantes mais velhos e mais escolarizados (com frequência de 48% e P.R. de 0.236).

Pessoas menos escolarizadas costumam utilizar com mais frequência variantes estigmatizadas ou avaliadas como informais pela comunidade de fala, enquanto as mais escolarizadas tendem a utilizar com mais frequência variantes tidas como de prestígio. (Labov (2001). Porém, quanto à idade, os resultados parecem

diagnosticar mudança em progresso, visto que os informantes mais jovens recorrem mais a DELE(A) para a indicação de posse de terceira pessoa do singular.

A exemplo de estudos que resultaram em constatações semelhantes, tem-se o trabalho de Silva (1991), com base em dados orais do NURC[6] do Rio de Janeiro e do Competência Básica (MOBRAL), e do Rio de Janeiro, houve bastante influência da escolaridade sobre a utilização de SEU(A)(S): Ensino Superior (45,8%, P.R. .88) > Ensino Médio (14,2%, P.R. .59) > Ensino Fundamental I (5,7%, P.R. .34) > Ensino Fundamental I > Alfabetizandos (3,9%, P.R. .26). No estudo de Soares (1999), DELE(A) foi favorecido pelos informantes de até 50 anos (com frequência de 95% e P.R. de .55) e SEU(A)(S) pelos informantes de mais de 50 anos (com frequência de 10% e P.R. de .57), em textos orais do Paraná. E Sbalqueiro (2005) utilizou como fonte de dados textos narrativos escritos produzidos por alunos das quatro últimas séries do Ensino Fundamental, nos quais a variante de maior frequência foi SEU(A)(S), que representou 90% dos dados. Constatam, esses trabalhos, a predominância da forma DELE em produções realizadas por informantes de menor idade/escolaridade.

5.4. **Influência do sexo dos informantes**

Seguem exemplos de ocorrências de SEU(A)(S) e DELE(A) encontradas na fala e na escrita de informantes dos dois sexos controlados no D&G:

(22) O professor era simplesmente louco, louco, daquele de jogar pedra na lua, aí um dia eu não tava muito afim de assistir aula **dele**. (Informante 2, escrita, feminino, narrativa de experiência pessoal).

(23) eu muito encabulado ... meu Deus ... o que que essa garota pode pensar ... se a minha mão deslizar ... e cair sobre **sua** perna? [...] Mas enquanto eu pensava isso a cabeça **dela** já derreava no meu ombro e o braço **dela** já passava no meu pescoço (Informante 4, oralidade, masculino, narrativa de experiência pessoal).

A Tabela 5 aponta os resultados obtidos:

Tabela 5. Influência do sexo sobre o uso de SEU(A)(S) E DELE(A)

	SEU(A)(S)			***DELE(A)***		
SEXO	Apl./Total	%	PR	Apl./Total	%	PR
Feminino	92/225	41	0.665	133/225	59	0.335
Masculino	54/165	33	0.282	111/165	67	0.718
TOTAL	146/390	37	----	244/390	63	----
	Input:0.374	Sig: 0.000		Input:0.626	Sig: 0.000	

A hipótese de que a forma DELE(A), por seu caráter mais informal e menos prestigiado, fosse condicionada positivamente entre os homens não foi confirmada: observa-se que os homens se inclinam ao uso de DELE(A) (com frequência de 67% e P.R. de 0.718) e não ao uso de SEU(A)(S) (com frequência de 33% e P.R. de 0.282). Em contraste, as mulheres tendem a privilegiar SEU(A)(S) (com frequência de 41% e P.R. de 0.665) em detrimento de DELE(A), que, embora tenha alcançado frequência de 59% entre as mulheres, recebeu peso relativo de 0.335.

Tais resultados poderiam ser tomados como diagnósticos de uma situação de *variação estável*, em conformidade com a generalização sociolinguística que prevê que, nesse tipo de situação, as mulheres tendem

[6] O Projeto de Estudo da Norma Linguística Urbana Culta (NURC) organizou um banco de dados de fala culta do português brasileiro através de gravações realizadas na década de 1970, em cinco capitais brasileiras: Rio de Janeiro, São Paulo, Recife, Salvador e Porto Alegre.

a optar com mais frequência pelas variantes melhor conceituadas na comunidade. Porém, na análise referente aos resultados obtidos para o grupo de fatores idade/escolaridade, verificou-se que os informantes mais jovens fizeram maior uso de DELE(A) e os informantes mais velhos fizeram maior uso de SEU(A)(S). Uma vez que podemos interpretar essa distribuição das variantes como passível de ser um reflexo de *mudança em progresso*, poderíamos considerar os resultados tangentes ao grupo de fatores sexo como indicadores de mudança em progresso liderada por homens.

Se uma mudança estivesse em progresso na direção de um aumento do uso de DELE(A), por sua vez, ela estaria sendo liderada pelos homens. Entretanto, mudanças linguísticas capitaneadas por homens são pouco frequentes, pois as mulheres lideram 90% das inovações da língua onde quer que elas tenham origem (Tagliamonte, 2012, p. 63). Já de acordo com a generalização sociolinguística relativa a situações de variação estável, os homens, em tais situações, utilizam uma frequência maior de formas estigmatizadas ou marcadamente informais do que as mulheres, que tendem a preferir formas socialmente valorizadas (cf. Labov, 1990; Chambers, 1995).

Quanto aos já mencionados trabalhos semelhantes a este, tem-se Silva (1991), onde a taxa de ocorrência da variante DELE(A) foi maior entre os homens (19,6%) e a taxa de ocorrência da variante SEU(A)(S) foi maior entre as mulheres (8,9%); Soares (1999), onde os homens favoreceram o aparecimento de SEU(A)(S) (com frequência de 10% e P.R. de .60), e as mulheres favoreceram o aparecimento de DELE(A) (com frequência de 95% e P.R. de .57); e Sbalqueiro (2005), que observou DELE(A), sendo mais frequente entre os homens (com frequência de 13% e P.R. de .62), em comparação às mulheres (com frequência de 7 % e P.R. de .41). Assim, o favorecimento de DELE alterna em preferência entre homens e mulheres e isso indica que o fator sexo não é um forte condicionante da variação dos possessivos de 3.ª pessoa.

Logo, neste estudo os grupos de fatores de natureza estilística (modalidade da língua e gênero/sequência textual) foram os mais significativos para a variação entre SEU(A)(S) e DELE(A), em detrimento dos grupos de fatores de natureza social (idade/escolaridade e sexo). E por isso, no que diz respeito à comunidade de fala de Natal, a variação das formas possessivas em 3.ª pessoa indica um fenômeno de ordem principalmente estilística e minoritariamente de ordem social.

6. **Considerações finais**

No *Corpus* D&G Natal, elaborado no final do século XX, foi possível verificar um maior número de ocorrências da forma DELE(A)(S), na **oralidade**, em gêneros/sequências de esfera **narrativa**, entre indivíduos de **menor idade**/**menos escolarizados**, e entre os **homens**. E de acordo com as generalizações sociolinguísticas, apresentadas na seção relativa ao enquadramento teórico-metodológico, a variação dos possessivos sofre forte influência dos aspectos extralinguísticos, sobretudo, estilísticos.

As hipóteses confirmadas foram:

1: A variante canônica SEU(A)(S), de caráter mais formal, ocorre em maior frequência em textos na modalidade escrita da língua, assim como em gêneros textuais de esfera não narrativa, em detrimento da variante preposicionada DELE(A)(S), de caráter mais informal, que se destaca em textos de modalidade oral da língua e em gêneros/sequências de esfera narrativa.

2: A variante mais formal SEU(A)(S) predomina na oralidade dos indivíduos mais escolarizados/de maior idade e entre as mulheres, enquanto a variante DELE(A)(S), tida como mais informal, destaca-se nos textos orais dos indivíduos menos escolarizados/de menor idade e entre os homens.

Já a hipótese não confirmada foi:

3: Em Natal, na última década do século XX, o fenômeno de variação entre os pronomes possessivos de terceira pessoa do singular SEU(A)(S) e DELE(A) pode ser caracterizado como um caso de mudança em tempo aparente que se reflete no maior uso da forma mais recente, DELE(A), pelos informantes mais jovens (de 18 a 20 anos), e o maior uso da forma mais antiga, SEU(A)(S), pelos informantes mais velhos (acima de 23 anos).

Os dados referentes à influência da modalidade da língua indicam que a variação entre os possessivos de terceira pessoa do singular se trata de uma *variação estável*. Todavia, a análise poderia indicar *mudança em progresso*, pois há favorecimento da forma DELE(A)(S) entre os mais jovens, se esse grupo não estivesse "enviesado" pelo fator escolarização. Além disso, não é comum que *mudança em progresso* seja liderada por homens. Logo, para concluir isto de forma mais robusta, é necessária a expansão do *corpus*.

Vale destacar que, o número de ocorrências da forma simples SEU(S)(S) entre os informantes de baixa idade/escolaridade foi mínimo ou nulo (razão pela qual não apareceram na rodagem dos dados: na tentativa de não enviesar ainda mais as relações entre os grupos de fatores), e isso também pode ser um indicativo de que o desaparecimento da forma canônica entre os mais jovens estivesse em curso.

Por fim, ao contrastar textos de mesmo gênero textual em suas versões orais e escritas, observou-se, que na escrita, o possessivo SEU pode corresponder a diferentes pronomes pessoais (você, ele, vocês, eles), e nem sempre o uso da forma perifrástica *de+ele* corresponderá a alcançar maior especificação ou desfazer ambiguidade, porque, por mais que em alguns casos haja emprego preferencial de uma das variantes, em vários casos a utilização de uma ou de outra dessas formas parece ser indiferente (Neves, 2002).

Referências

Almeida, Adriana B. (1993) *Pronomes possessivos de 3ª pessoa no português falado de São Paulo*. [Manuscrito não publicado]

Bailey, Guy (2004) Real and apparent time. In Jack Chambers, Peter Trudgill & Natalie Shilling Estes (eds.), *The handbook of language variation and change*. Blackwell. pp. 312–332.

Castro, Ana (2006) *On possessives in Portuguese*. Dissertação de doutoramento, Universidade Nova de Lisboa.

Cerqueira, Vicente Cruz (1996) A forma genitiva "dele" e a categoria de concordância (AGR) no português brasileiro. In Ian Roberts e Maria Kato (orgs.), *Português brasileiro*: *Uma viagem diacrônica* (2.ª ed.). Editora da UNICAMP, pp. 129–161.

Chambers, Jack (1995) *Sociolinguistic theory: Linguistic variation and its social significance*. Blackwell.

Cunha, Patrícia (2007) *Possessivos de terceira pessoa na língua portuguesa nos séculos XIII e XIV*. Tese de doutorado, Universidade Federal de Minas Gerais.

Furtado da Cunha, Maria Angélica (1998) (org.) *Corpus Discurso & Gramática: A língua falada e escrita na cidade do Natal*. EDUFRN.

Görski, Edair Maria, Izete Lehmkuhl Coelho & Christiane Souza (orgs.) (2014) *Variação estilística*: Reflexões teórico-metodológicas e propostas de análise. Insular.

Guy, Gregory & Ana Zilles (2007) *Sociolinguística quantitativa: Instrumental de análise*. Parábola Editorial.

Hopper, Paul & Elizabeth Traugott (2003) *Grammaticalization* (2.ª ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139165525

Kato, Maria (1985) A complementaridade dos possessivos e das construções genitivas no português coloquial: Réplica a Perini. *DELTA* 1 (1/2), pp. 107–120.

Kato, Maria (1994) Raízes não-finitas na criança e a construção do sujeito. *Cadernos de Estudos Linguísticos* 29, 119–136.

Labov, William (1990) The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, pp. 205–254. https://doi.org/10.1017/S0954394500000338

Labov, William (1994) *Principles of linguistic change: Internal factors*. Blackwell.

Labov, William (2001) *Principles of linguistic change: Social factors*. Blackwell.

Labov, William (2001b) The anatomy of style-shifting. In Penelope Eckert & John R. Rickford (eds.), *Style and sociolinguistic variation*. Cambridge University Press, pp. 85–108.
Labov, William (2004) Ordinary events. In Carmen Fought (ed.), *Sociolinguistic variation: Critical reflections*. Oxford University Press, pp. 31–43.
Labov, William (2008) *Padrões sociolinguísticos* (Marcos Bagno, Maria Marta Pereira Scherre e Caroline Rodrigues de Oliveira, Trads). Parábola. [primeira edição em 1972]
Le Page, Robert (1998) The evolution of a sociolinguistic theory of language. In Florian Coulmas (ed.), *The handbook of sociolinguistics*. Blackwell Publishing. pp. 13–32.
Lopes, Célia & Dailane Guedes (2020) Formas possessivas de terceira pessoa: Confrontando seu e dele a partir da abordagem experimental. *Confluência* (58), pp. 82–105. https://doi.org/10.18364/rc.v1i58.353
Marcotulio, Leonardo, Dalila de Assis & Rafaela Guedes (2015) De-possessivos de 2ª pessoa na história do português brasileiro. *Diacrítica*, pp. 203–232.
Marcuschi, Luiz (2002) Gêneros textuais: Definição e funcionalidade. In Angela Dionísio, Anna Machado & Maria Bezerra (orgs.), *Gêneros textuais e ensino*. Lucerna, pp. 19–36.
Marcuschi, Luiz (2003) *Da fala para a escrita: Atividades de retextualização* (4.ª ed.). Cortez.
Mattos e Silva, Rosa Virgínia (2008) Teorias da mudança linguística e a sua relação com a(s) história(s) da língua(s). *Linguística - Revista de Estudos Linguísticos da Universidade do Porto*, 3, pp. 39–53.
Menon, Odete (1995) O sistema pronominal do português do Brasil. *Letras*, 44, pp. 91–106. https://doi.org/10.5380/rel.v44i0.19069
Menuzzi Sergio (2003a) Sobre as opções anafóricas para antecedentes genéricos e para variáveis ligadas: Comentários a Ana Müller. *Letras de Hoje*, 38 (1), 125–144.
Menuzzi Sergio (2003b) Escopo e "varáveis ligadas típicas" do português brasileiro. *Revista Letras* 61, p. 213–248. https://doi.org/10.5380/rel.v61i0.2888
Morais, Maria (2019) Possessivos de terceira pessoa no flos sanctorum e no português brasileiro contemporâneo. *História, Histórias*, 7 (4), pp. 59–86. https://doi.org/10.26512/hh.v7i14.26600
Müller, Ana Lúcia (1997) *A gramática das formas possessivas no português do Brasil*. Tese de doutorado, UNICAMP.
Negrão, Esmeralda & Ana Lúcia Müller (1996) As mudanças no sistema pronominal do português brasileiro: substituição ou especialização de forma? *DELTA*, 12 (1), pp. 153–171.
Neves, Maria Helena (2002) Possessivos. In Ataliba Castilho (org.), *Gramática do português falado* (Vol. 3). Editora da UNICAMP, pp. 149–211.
Coelho, Olga (2020) 50 anos do GEL: Caminhos da linguística no Brasil. *Estudos Linguísticos (São Paulo. 1978)* 49 (1), pp. 22–35. https://doi.org/10.21165/el.v49i1.2508
Perini, Mário (1985) O surgimento do sistema de possessivo do português coloquial: uma interpretação funcional. *DELTA* 1 (1/2), pp.1-–15.
Perini, Mário (2010) *Gramática do português brasileiro*. Parábola.
Pintzuk, Susan (2003) Variationist approaches to syntatic change. In Brian Johnson & Richard Janda (eds.), *The handbook of historical linguistics*. Blackwell, pp. 509–528.
Rocha, Fernanda (2009) *A alternância dos pronomes pessoais e possessivos do português de Belo Horizonte*. Dissertação de mestrado, Pontifícia Universidade Católica de Minas Gerais.
Sankoff, David, Sali A. Tagliamonte & Eric Smith (2005) *Golvarb X: A multivariate analysis application*. Departament of Linguistics.
Sbalqueiro, Arnaldo (2005) *A variação dos pronomes possessivos de 2ª e 3ª pessoas em redações dos alunos de uma escola pública de Curitiba*. Dissertação de mestrado, Universidade Federal de Santa Catarina.
Silva, Giselle Machline de Oliveira (1982) Estudo da regularidade na variação dos possessivos no português do Rio de Janeiro. Tese de doutorado, Universidade Federal do Rio de Janeiro.
Silva, Giselle Machline de Oliveira (1984) Variação no sistema possessivo de terceira pessoa. *Tempo Brasileiro* 78/79, pp. 54–72.

Silva, Giselle Machline de Oliveira (1991) Um caso de definitude. *Organon* 19, pp. 90–108.
Silva, Giselle Machline de Oliveira (1998a) Os estertores da forma seu de terceira pessoa na língua oral. In: Giselle Silva & Marta Scherre (orgs.), *Padrões sociolingüísticos: Análise de fenômenos variáveis do oortuguês falado na cidade do Rio de Janeiro.* . Tempo Brasileiro, pp. 169–182.
Silva, Giselle Machline de Oliveira (1998b) Estertores da forma seu de terceira pessoa na língua oral: Resultados sociais. In Giselle Silva & Marta Scherre (orgs.), *Padrões sociolingüísticos: Análise de fenômenos variáveis do oortuguês falado na cidade do Rio de Janeiro*. Tempo Brasileiro, pp. 295–308.
Silva, Mariana Lorena (2016) *Variação dos pronomes possessivos de terceira pessoa do singular seu(a)(s)/dele(a) em Natal-RN: Aspectos sociais e estilísticos*. Dissertação de mestrado, Universidade Federal do Rio Grande do Norte.
Soares, Alexandre (1999) *Segunda e terceira pessoa*: *O possessivo em questão – Uma análise variacionista*. Dissertação de mestrado, Universidade Federal do Paraná.
Tagliamonte, Sali A. (2006) *Analysing sociolinguistic variation*. Cambridge University Press.
Tagliamonte, Sali A. (2012) *Variationist sociolinguistics*: *Change, observation, interpretation*. Wiley-Blackweel.
Tavares, Maria Alice (2014) Variação estilística e gênero textual: O caso dos gêneros textuais produzidos no macrogênero entrevista sociolinguística. In Christiane Souza, Edair Maria Gorski & Izete Lehmkuhl Coelho (orgs.), *Variação estilística*: *Reflexões teórico-metodológicas e propostas de análise*. Insular, pp. 203–223.

Ensino de escrita e avaliação de manuais didáticos de Português L2

Juliano Sippel[1]

[1]Universidade NOVA de Lisboa, CLUNL, Lisboa, Portugal

Resumo

Embora a escrita seja um conhecimento formal e estruturado, que requer instruções explícitas e detalhadas para o seu desenvolvimento adequado (Barbeiro & Pereira, 2007; Grabe & Kaplan, 1996), a literatura tem mostrado que seu ensino tem ocupado uma posição marginalizada em relação ao desenvolvimento de outras competências da L2 (Cumming, 2001; Silva, 1990). Com base nessa constatação e considerando que os manuais didáticos são fundamentais para o desenvolvimento das competências comunicativas (Tomlinson & Masuhara, 2018), pretendemos apresentar os resultados de uma análise de manuais de ensino de português L2, no tocante ao desenvolvimento da competência de escrita. A seleção dos manuais teve por base as respostas a um inquérito preenchido por professores de português L2, a partir do qual analisamos os manuais mais utilizados ou recomendados. Os resultados indicam que os manuais adotam uma abordagem comunicativa fraca e seguem um modelo de ensino da escrita como um produto, que não prevê instruções, planejamento e revisão. Verifica-se também um desequilíbrio na distribuição das competências comunicativas e falta de interligação entre a escrita e as outras competências, bem como ausência de orientações metodológicas para docentes.

Palavras-chave: ensino de escrita, português L2, avaliação de manuais didáticos.

Abstract

Although writing is a formal and structured skill that requires explicit and detailed instructions for its proper development (Barbeiro & Pereira, 2007; Grabe & Kaplan, 1996), the literature has shown that its teaching has occupied a marginalized position in relation to the development of other L2 skills (Cumming, 2001; Silva, 1990). Based on this observation and considering that textbooks are fundamental for the development of communicative competencies (Tomlinson & Masuhara, 2018), we intend to present the results of an analysis of L2 Portuguese textbooks regarding the development of writing competence. The selection of textbooks was based on the responses of a survey completed by L2 Portuguese teachers, from which we analyzed the most used or recommended textbooks. The results indicate that the textbooks adopt a weak communicative approach and follow a model of teaching writing as a product, which does not provide instructions, planning, or revision. There is also an imbalance in the distribution of language skills and a lack of interconnection between writing and other skills, as well as a lack of methodological guidelines for teachers.

Keywords: writing instruction, Portuguese L2, textbook evaluation.

1. Introdução

Há um consenso na literatura produzida relativamente ao desenvolvimento da competência escrita em L2 em afirmar que seu ensino costuma ser menos exercitado que o de outras competências. De forma paradoxal, é também comum associar ao bom desenvolvimento do texto escrito seu ensino explícito, centrado em cada uma das etapas tradicionalmente descritas do processo de produção textual: planejamento, textualização e revisão.

Diante desta constatação, e por os manuais didáticos serem fundamentais para o desenvolvimento das competências comunicativas, o objetivo deste artigo é investigar, através de uma análise de conteúdo, os

modelos de escrita presentes em manuais de português L2 (PL2) e a forma como mobilizam o desenvolvimento dessa competência.

O texto está dividido em 4 seções. Na seção 2, apresentamos uma breve síntese do estado da arte sobre o ensino da competência escrita em L2, sobre a avaliação de manuais didáticos (de forma geral e especificamente de L2) e apresentamos os documentos norteadores do ensino de PL2 em Portugal e no Brasil, que consultamos para elaborar uma grelha de análise. Na seção 3, detalhamos a metodologia de investigação utilizada neste estudo. Finalmente, na seção 4, discorremos nossa análise e discussão dos materiais didáticos avaliados. Encerramos o estudo na seção 5, expondo nossas conclusões.

2. Enquadramento teórico

2.1. Ensino de escrita em L2

Comparada às demais competências comunicativas, a escrita é uma habilidade linguística mais complexa – em termos gramaticais e em comparação com a oralidade, um bom texto exige o uso de orações completas, com ordem de palavras mais rígida, além de alguma densidade lexical com utilização constante de sinônimos. A competência escrita está relacionada com a objetividade e o explícito, ademais da capacidade de reflexão metalinguística, porque o texto escrito possui uma organização previamente estabelecida, pelo que se relaciona com a modalidade linguística estândar e com os registros formais e objetivos (Cassany, 1999, 2017).

Barbeiro e Pereira (2007, p. 15) evidenciam a complexidade do processo de escrita, que "exige a capacidade de selecionar e combinar as expressões linguísticas, organizando-as numa unidade de nível superior, para construir uma representação do conhecimento, correspondente aos conteúdos que se quer expressar". Para atingir o que chamam *competência compositiva*, o escritor deverá ativar os conteúdos, decidir se esses se integram ou não à composição, articulá-los com os demais elementos do texto e dar-lhes forma, materializando-os no texto e respeitando exigências de coesão e de coerência. Isso significa que a competência compositiva é ativada em um nível global (ou de macroestrutura, que corresponde à organização das grandes unidades do texto) e em um nível específico (ou de microestrutura, que corresponde à combinação das expressões linguísticas).

Para escrever é necessário ativar os conhecimentos prévios sobre o tópico da escrita, enquadrá-lo em um gênero, redigir o texto e avaliar o que foi escrito para reformular, se assim se fizer necessário. Escrever é uma atividade controlada e seu processo tem sido descrito, de modo tradicional, conforme mostram Grabe e Kaplan (1996), em três componentes ativos, que correspondem às operações mentais de planejar, escrever (ou textualizar) e revisar. Um escritor competente deve gerir esses três componentes com autonomia, já que esses processos não são lineares, mas interativos e simultâneos.

Ao examinar a tradição do ensino de escrita, deparamo-nos com três abordagens e implicações pedagógicas, que focalizam o processo da produção no texto, no escritor ou no leitor. Expomos as principais características de cada uma delas, conforme a sistematização de Hyland (2003, 2009).

(1) *Text-oriented research and teaching* (escrita como produto) – nessa abordagem, o texto é concebido e estruturado em sistemas e arranjos gramaticais, resultando em um objeto autônomo, que pode ser analisado e descrito independentemente de contextos e de escritores específicos, e em um ensino centrado em precisão gramatical e ausência de contexto.

(2) *Writer-oriented research and teaching* (escrita como processo) – essa concepção considera o texto como meio para solucionar problemas, o que exige do aluno o levantamento de objetivos em um processo não linear, que prevê constantes reformulações. Modelos processuais ainda muito influentes são os propostos por Hayes e Flower (1980), Flower e Hayes (1981), Hayes (1996) e Scardamalia e Bareiter (1987) e as implicações pedagógicas resultantes dessa abordagem são especialmente úteis à prática em L2 porque da análise de

estratégias adotadas pode-se observar comportamentos diferentes e semelhantes dos alunos, uma vez que os textos posicionam leitores e escritores de maneiras específicas, evocando esquemas compartilhados.

(3) *Reader-oriented research and teaching* (escrita como atividade social) – essa abordagem dá ao texto uma dimensão política e ideológica uma vez que prevê a escrita como interação, como antecipação e confirmação de propósitos que são definidos segundo os públicos aos quais se destina. Os trabalhos de Fairclough (1989) ganham destaque nessa abordagem, porque dão conta de vincular à língua as atividades que a cercam e mostram que o discurso é mediador da vida social e por ela condicionado.

Graham (2018) mostra como ao longo dos anos a escrita tem sido vista através de diferentes lentes (comportamentais, cognitivas e sociais). Sua meta-análise revela que os modelos atualmente praticados se assentam em bases cognitivas – por meio da descrição de mecanismos complexos envolvidos e acionados no ato de escrever – e sociais – pelo desenvolvimento da pedagogia do gênero, que concebe a escrita como atividade essencialmente social.

Em L2, paradoxal e historicamente, a escrita assume uma posição marginalizada em relação ao ensino de outras competências. Segundo Silva (1990), somente após a Segunda Guerra Mundial, quando estudantes internacionais começam a se matricular nas instituições de ensino superior norte-americanas, as instruções para a escrita em L2 (nesse caso, o inglês) começam a receber atenção. Desse cenário, e das diferenças que emergiam de produções desses alunos, os professores passaram a dedicar tempo a essa competência, até então ignorada. Inicialmente tida como uma habilidade que poderia ser adquirida pelo domínio de determinadas estruturas em ambientes controlados, o ensino da escrita centrava-se no nível da frase, com ênfase na gramática. Surgem depois outras concepções como as de contrastes retóricos e da escrita de inglês com finalidade acadêmica.

Cumming (2001) mostra que as pesquisas que se debruçam sobre o ensino de escrita em L2, nos três modelos de ensino referidos, produziram algumas considerações, a saber: (i) da escrita como produto, as investigações evidenciam a necessidade do domínio das competências morfossintáticas; mostram, além disso, uma relação direta entre proficiência e desempenho, na medida em que quanto mais conhecimentos léxico-gramaticais os alunos têm sobre a língua, mais bem elaborada é sua produção escrita; (ii) da escrita como processo, estudos que se debruçaram sobre processos mentais de episódios de pensamentos e tomadas de decisões encontraram comportamentos salientes, do que se pode entrelaçar uma relação entre proficiência e planejamento na medida em que quanto maior é o conhecimento de um aluno sobre a língua, maior é seu grau de controle de manipulação do texto, com mais tempo destinado às etapas de planejamento e de revisão; e (iii) da escrita como atividade social, as pesquisas revelam a importância da atenção aos gêneros textuais para a inserção do aluno em comunidades discursivas, capaz de dar-lhes versatilidade em negociar relações de poder e significado.

Sabe-se também que a escrita em L2 é afetada pela L1 e pelo contexto educacional dos alunos, que fornece os conhecimentos prévios necessários à produção textual. Reconhecendo essa importância, a síntese de investigações realizada por Leki et al. (2008) mostra alguns experimentos que analisaram as interferências da L1 na produção escrita, dos quais há que se destacar aqueles que compararam padrões retóricos das diferentes línguas e analisaram as dificuldades enfrentadas pelos alunos na sua manipulação. A instrução revela-se fundamental nesse processo porque influencia o reconhecimento e o uso dos padrões retóricos da L2, levando à produção textual coerente e com melhores esquemas organizacionais. Segundo os autores, reconhece-se nos escritores com maior experiência mais atenção aos processos hierárquicos de composição e de planejamento textual, à gramática, à escolha dos padrões e das formas retóricas adequadas ao público-alvo, e aos processos de revisão de seus textos.

De acordo com Rinnert e Kobayashi (2009), se os processos de composição não forem ensinados na L2, é pouco provável que os escritores os utilizem em seus textos. Afirmam, por isso, que sem instrução não se desenvolve conhecimento explícito sobre a escrita e que o metaconhecimento sobre a escrita influencia a qualidade dos textos, inclusive em sua progressão lógica, afetando os níveis sintáticos e semânticos.

Ainda que não trate especificamente da área de escrita, outro campo de investigação bastante profícuo no âmbito do ensino de L2 é o da análise de erros e do *feedback* corretivo, definido por Ellis (2006, p. 28) como "responses to learner utterances containing an error". Embora os manuais didáticos não sejam intervenientes diretos nesse processo, o *feedback* que se dá ao estudante de língua a respeito dos erros que comete ao produzir determinadas estruturas é uma forma de focar a atenção na forma, como destaca Lightbrown (1998), e de incentivar o processo de revisão textual.

A natureza e os diferentes tipos de *feedback* têm sido amplamente debatidos nas últimas décadas. Lyster e Ranta (1997) inicialmente descreveram seis tipos de *feedbacks* e, mais recentemente (Ranta & Lister, 2007), reorganizaram sua classificação em duas grandes categorias: reformulações e *prompts*, sendo a primeira o fornecimento da correção explícita ao aluno e, a segunda, o uso de estratégias que conduzem o aprendente à autocorreção. Ellis et al. (2009), em classificação semelhante, distinguem os *feedbacks* em explícito (operacionalizado por meio da correção explícita dos erros ou da explicação de regras gramaticais) e implícito (fornecido por meio de reformulações, ou *recasts*).

Em termos de benefícios, meta-análises como a de Lyster e Saito (2010) revelam que o uso de *feedback* corretivo tem se mostrado eficiente para a aprendizagem de L2. Relativamente ao ensino de PL2, merece destaque o estudo de Rauber (2017), que afirma ser o *feedback* explícito o mais usado pelos professores observados em sua investigação, e o de Siopa (2015), que, embora trate do contexto universitário, defende o uso sistemático de estratégias de tratamento e correção de erros, associadas à revisão da produção textual para o desenvolvimento da competência escrita.

Investigações sobre a escrita em L2 como as de Berman (1994), Matsumoto (1995), Silva (1993), Yu et al. (2023), Han e Hilver (2018), Li (2023), além das já mencionadas, sugerem, em suma, uma remodelação para o ensino que inclua não apenas as formas textuais, mas também a atenção aos processos de composição dos alunos. Isso deve incluir instruções gramaticais adequadas e uma ênfase na importância da seleção de informações, com revisões e edições de rascunhos, além de uma abordagem voltada para o processo e para o desenvolvimento dos gêneros, para manter os estudantes motivados e potenciar o papel da memória de trabalho na produção textual.

Ademais, destacamos uma discussão recente surgida na literatura, que deriva do questionamento: o ensino da competência escrita deve ser efetivamente ensinar a escrever na língua-alvo ou a escrita deve ser usada para o aprendizado da língua? A esse respeito, Bhowmik (2021) defende que as abordagens de aprender a escrever e escrever para aprender L2 não são excludentes, mas complementares, tendo como foco principal as necessidades dos aprendentes.

Estudos no âmbito do PL2 como o de Lopes e Pinto (2022) mostram como a densidade de ideias – um critério que pode ser associado à precisão gramatical – também é melhorada com o uso de instruções adequadas, e que textos pouco densos derivam da pouca familiaridade dos estudantes com a modalidade escrita. Destacamos também a investigação de Barbosa e Bizarro (2015), que revela a importância dada à competência escrita por docentes de PL2, que reconhecem a escola como espaço privilegiado, mas não destinam à escrita o primeiro nível de importância de ensino. Os autores apresentam uma consistente defesa de uma pedagogia da escrita em L2 e de manuais didáticos adequados, que forneçam aos estudantes "um claro entendimento dos processos de escrita, na sua relação com um determinado produto, com opções retóricas, e com os contextos culturais nos e para os quais os textos são produzidos e editados" (Barbosa & Bizarro, 2015, p. 4).

Podemos assinalar, com base no enquadramento teórico até aqui apresentado, que o ensino da escrita em L2 deve providenciar: (i) sensibilização da necessidade de conhecer o contexto cultural e social onde se insere a produção escrita; (ii) instrução a respeito dos marcadores discursivos da L2 para produção de textos com coesão e coerência; (iii) instrução de esquemas e padrões composicionais da L2, principalmente dos padrões hierárquicos de composição, que deve gerar um planejamento de composição; (iv) estímulo à prática constante e intensiva de planejamento de escrita; (v) incentivo à prática constante de releitura, revisão geral e global e reescrita; (vi) fornecimento de *feedbacks* corretivos e de análise de erros cometidos pelos estudantes; e (vii) fomento ao uso de dicionários, o que também resulta numa maior atenção ao planejamento.

Para ensinar a escrever é necessário promover práticas que levem as pessoas a converterem suas competências adquiridas em L2 em um desempenho controlado e hábil. Tais práticas exigem tempo disponível para reflexão e revisão, explicitação de objetivos e necessidade de avaliar hipóteses sobre a língua antes de colocá-las na produção escrita. À medida que os alunos aprendem a escrever em L2, adquirem maior controle sobre sua própria habilidade de planejar, de revisar e de editar seus textos, procurando formas cada vez melhores e mais adequadas.

2.2. **Avaliação de manuais didáticos**

Embora não haja uniformidade no conceito de manual didático, devido às múltiplas dimensões em que esse produto se encontra (linguística, política, educacional, cultural etc.), há características que lhe são próprias: (i) intencionalidade de produção para o ensino escolar: (ii) sistematicidade de exposição de conteúdos; (iii) sequencialidade e gradação de níveis de dificuldade; (iv) uso de tipologia textual expositiva; (v) uso combinado de textos e imagens; e (vi) regulamentação de conteúdos de acordo com programas oficiais nacionais (Ossenbach, 2010). É sobretudo por conta desse último aspecto que os manuais desempenham, de certa forma, o papel de divulgação e legitimação do conhecimento produzido a respeito do que se deve ensinar em uma L2, sendo parte constitutiva da construção do saber docente, como assinalou Tardif (2001).

Devido a essa diversidade de categorização e importância atribuída, a avaliação de manuais didáticos também não é tarefa uniforme e não possui um método específico. Seu desenvolvimento tem ocorrido de forma processual na área denominada *manualística* – termo cunhado pelo historiador Augustín Escolano Benito na ocasião da criação de diferentes centros dedicados à investigação de manuais. A avaliação de livros escolares percorreu diversos caminhos ao longo dos anos, conforme também mudavam os motivos e as perspectivas do que se avaliava (Aran, 2007).

Assim como a perspectiva, também varia a função de uma avaliação de manual didático, sendo as mais comuns: (i) proporcionar juízos de valor sobre o aproveitamento dos *inputs* fornecidos no processo de ensino; (ii) analisar a coerência entre objetivos e resultados; (iii) estabelecer relações entre elementos que intervêm no processo educacional; e (iv) proporcionar informação e subsídios para tomada de decisões. Com a avaliação de manuais é possível tomar decisões sobre informações e recursos e a forma de usá-los em sala de aula, contudo, a avaliação requer antes a conscientização sobre sua importância e a aprendizagem de como realizá-la (Aran, 2001, 2007).

Aran (2007) agrupa os critérios da análise de manuais em quatro âmbitos, a saber: (i) intenções educacionais; (ii) requisitos para a aprendizagem; (iii) atenção à diversidade do alunado; e (iv) aspectos formais. Essas características fazem da avaliação de manuais um requisito imprescindível porque, por meio dela, tomamos decisões sobre os recursos que são mais e menos adequados e a forma de usá-los em sala de aula. Como se disse, é uma atividade que requer uma aprendizagem sobre como realizá-la, mas que, também, por ser um procedimento, somente se aprende a fazer, fazendo-a em diversas ocasiões, até que se atinja seu domínio.

2.3. **Avaliação de manuais didáticos de L2**

Ao estabelecer critérios para a seleção de um manual didático de ensino de L2, Cunningsworth (1995) propõe *análise* e *avaliação* como dois processos diferentes e interdependentes. O primeiro trataria propriamente da descrição do manual, com suas características e objetivos, o que envolve sua metodologia e seus princípios de organização, sem julgamento de valor – sendo esse um aspecto do segundo processo, pois a avaliação tem essa característica de fins de seleção com base em critérios e objetivos da situação que a gerou. Segundo o autor, esses processos são interdependentes porque a avaliação de livros didáticos envolve quatro etapas, que são: análise, interpretação (essa etapa engloba as experiências dos profissionais), avaliação e seleção (estágio em que se busca verificar a adequação do manual em processo de avaliação ao contexto específico de ensino e aprendizagem).

Cunningsworth (1995) elenca três momentos diferentes em uma avaliação de manual didático: (i) pré-uso, para avaliar a possibilidade de se usar o manual conforme as necessidades; (ii) em uso, para julgar a eficácia do livro na prática docente; e (iii) pós-uso, para compreender a eficácia do manual e sua adequação ao contexto de utilização. Parece-nos importante retomar essas distinções do autor para afirmar que nossas análises se centram na etapa de pré-uso.

Relativamente ao âmbito de avaliação em função dos requisitos para a aprendizagem, Tomlinson e Masuhara (2018) mostram que os manuais de L2 produzidos entre as décadas de 1970 e 80 apresentavam um domínio de ensino das formas gramaticais (o que condiz com a abordagem audiolingual, amplamente difundida à época). O modelo praticado era o das práticas controladas, com uso de padrões a serem seguidos (modelo conhecido como *Presentation, Practice, Production*). Com o predomínio da abordagem comunicativa, os materiais passam a incluir atividades de conscientização nas quais os alunos são orientados a desencadear descobertas por meio de usos. Pode-se citar aqui o modelo *Focus on Form (FoF)*, no qual as formas da L2 surgem incidentalmente e seu foco privilegia a significação, opondo-se ao modelo *Focus on Forms (FoFs)*, cuja abordagem é a tradicional ênfase nas estruturas em forma de unidades isoladas de contextos de uso.

Ainda segundo os autores, o destaque à gramática pode ser observado em manuais atuais, que priorizam práticas controladas de uso de língua, com modelos a serem seguidos pelos estudantes – o que os faz afirmar ser o modelo *Presentation, Practice, Production* dominante na área comercial de manuais de L2. Tomlinson e Masuhara (2018) sustentam um posicionamento crítico quanto a essa abordagem já que esse modelo só poderia atingir áreas superficiais e gerar uma aprendizagem efêmera por causa do foco estreito, da falta de contextualização, envolvimento afetivo e cognitivo e uma oferta inadequada para um uso de língua moldado por uma comunicação autêntica. Uma alternativa ao modelo seria a aprendizagem pela descoberta (*Discovery approaches*), que vai ao encontro do modelo de foco na forma, ou a aprendizagem por meio do envolvimento das estruturas gramaticais em tarefas de comunicação, as chamadas *task-based language teaching (TBLT)* (Long, 1991, 2015).

As avaliações de manuais disponíveis na literatura que vem sendo produzida por Tomlinson (2010a, 2010b, 2011, 2016) e Tomlinson e Masuhara (2018), novamente no âmbito dos requisitos para a aprendizagem, têm sugerido que a aprendizagem de uma L2 é facilitada quando os estudantes estão engajados na interação e em situações de comunicação que lhes são significativas. Por conta disso, tarefas e exercícios eficazes em sala de aula, desenvolvidos por meio dos manuais de ensino, devem oferecer a oportunidade de negociar significados, ampliando os recursos de língua que os estudantes possuem e levando-os a observar como as formas linguísticas são usadas em trocas interpessoais. Ao docente caberia a função de ser facilitador e criar a sala de aula um ambiente propício para essa negociação de significados.

Ainda que não avaliem propriamente o desenvolvimento da competência escrita nos livros didáticos de L2, Tomlinson (2010a, 2010b, 2014), Tomlinson e Masuhara (2018) e Pinnard (2016) fornecem orientações para o trabalho com a produção textual, que podem ser resumidas em uma proposta que estimule os estudantes à leitura de um texto envolvente com o objetivo de produzirem outro texto, com conexões ao lido. Para tal, seria necessário explorar o texto principal para que os alunos façam descobertas sobre algum recurso específico, procurem textos em que se utilize o mesmo tipo de recurso e articulem generalizações. Essa metodologia de trabalho propiciaria uma produção textual semelhante à oferecida como modelo, sendo necessário para tanto certificar-se de que os estudantes teriam várias oportunidades de leitura dentro e fora da sala de aula.

A operacionalização dessas recomendações nos manuais didáticos seria feita pelo oferecimento de diferentes textos autênticos, de exercícios guiados para a posterior produção escrita e, sobretudo, da orientação detalhada para condução desse trabalho pelos docentes da L2.

2.4. **Documentos norteadores do ensino de português L2**

Em Portugal, a institucionalização da língua portuguesa para estrangeiros é recente e acompanhou as necessidades sociais e políticas no âmbito do ensino da língua para filhos de imigrantes que passaram a

frequentar as escolas do país. Dessa demanda, surge o *Decreto-Lei n.º 6/2001 de 18 de janeiro*, um documento elaborado pelo Ministério da Educação de Portugal para gerir o objetivo estratégico de garantir educação para todos, estabelecendo que as escolas devem proporcionar atividades curriculares específicas para a aprendizagem do PL2 aos alunos cuja L1 não é a língua portuguesa.

A integração do PL2 ao currículo nacional português ocorre como medida de acolhimento e em 2005 passam a vigorar as orientações do *Português Língua Não Materna no Currículo Nacional: Documento Orientador*, que fornece orientações gerais para o ensino de filhos de imigrantes. O documento elenca as necessidades desse alunado, que podem ser linguísticas, curriculares e de integração, e reconhece a igualdade e a interculturalidade como princípio básico da igualdade de direitos, além de determinar como objetivo específico o domínio oral e escrito da língua portuguesa como língua veicular. A organização do processo individual e escolar dos alunos também é contemplada no documento, bem como o reconhecimento de testes diagnósticos da língua (as escolas são incentivadas a recorrer a testes elaborados pelo Centro de Avaliação de Português Língua Estrangeira (CAPLE) para diagnosticar e avaliar o nível de proficiência linguística de seus alunos). Finalmente, o documento afirma a necessidade de adoção de metodologias próprias para a aprendizagem de PL2, que deverão ser elaboradas como orientações nacionais para ensino da língua nas escolas básicas e secundárias do país (Perdigão, 2005).

As *Orientações Programáticas de Português Língua Não Materna (PLNM): Ensino Secundário* são homologadas pelo Estado português em 2008, com o intuito principal de definir as especificidades do PLNM e de suas práticas pedagógicas. É um documento importante para nossa análise porque traz o enfoque prático-metodológico para os diferentes níveis de ensino, que estão de acordo aos grupos previstos no *Quadro Europeu Comum de Referência para as Línguas* (QECR). O desenvolvimento de competências pode ser considerado como ideia central das atividades pedagógicas, que devem dar ênfase às gramaticais e às lexicais e, sobretudo no nível intermediário, às metadiscursivas e metalinguísticas. Relativamente à competência discursiva, o documento a refere como o domínio de modelos, adquiridos pela leitura, que permite extrair conteúdos nucleares do texto lido, possibilitando a construção de um sumário da informação veiculada. Dessa forma, sugere que a compreensão da escrita é facilitada quando o texto é precedido de uma curta síntese ou apresentação: "o conhecimento prévio da sua temática e dos seus referentes ajudará consideravelmente a compreensão dos conteúdos" (Leiria, 2008, p. 14).

Outra recomendação fornecida pelo documento, no tocante à elaboração de materiais para ensino, diz respeito ao uso de textos autênticos, ainda que apresentem um grau maior de dificuldade aos alunos. Além disso, há a indicação de que as atividades propostas para o desenvolvimento da competência discursiva deverão privilegiar textos diversificados, e de uma lista de gêneros recomendados para o utilizador independente como notícia, relato de experiências pessoais, reconto, diário, memórias, autobiografia, carta, resumo, artigos de apreciação crítica, cartaz, publicidade televisiva, na Internet, além dos gêneros literários conto, poema e teatro. Tendo como base as recomendações dadas no QECR, o documento orienta que, relativamente à produção escrita, o estudante deverá ser capaz de, por exemplo, produzir pequenos textos escritos, a propósito de sequências ouvidas, lidas, ou a partir de imagens, em contextos variados, escrever sobre assuntos do quotidiano em pequenos textos estruturados, a partir de um tópico específico, além de ser capaz de justificar por escrito as razões de uma escolha (Leiria, 2008).

Em 2011, o governo português publicou o *Quadro de Referência para o Ensino de Português no Estrangeiro. Documento Orientador* (QuaREPE) (Grosso, 2011), que é um instrumento de política linguística para a difusão internacional da língua portuguesa e que remonta à necessidade de ensinar a língua a filhos de portugueses que emigraram. Ancorado também no QECR, o documento descreve as competências linguísticas esperadas para cada nível em que o estudante se encontra, além de tratar de competências relacionadas com outras áreas curriculares, que englobam aspectos da sociedade portuguesa relacionados com elementos simbólicos do país. Pode-se observar que o documento também privilegia as dimensões gramatical e lexical das competências comunicativas no decorrer das orientações que fornece para a elaboração de atividades. Relativamente ao uso de textos, o QuaREPE adota uma abordagem orientada para ação, em que os estudantes

são os atores que executam as tarefas em diferentes domínios em que ocorre a comunicação, e recomenda para o utilizador independente gêneros como: instruções, banda desenhada, publicidade, cartas e postais, mensagens de correio eletrônico e telemóvel, folhetos, verbetes de dicionários, notícias e gêneros literários (Grosso, 2011).

Em consonância com o QECR, o QuaREPE fornece descritores para os diferentes níveis de competência, dos quais interessa-nos verificar os relativos à competência escrita. A diferença entre os níveis B1 e B2 consiste basicamente entre textos mais simples (B1) e pormenorizados (B2).

Mencionamos que os documentos analisados utilizam como apoio o QECR e esse documento também fornece orientações para a seleção e para a produção de recursos de ensino da competência escrita. Ao conceber texto como atividade linguística em decurso de realização de tarefa (ação para atingir resultados na resolução de um problema), o documento entrelaça uma relação entre estratégia, tarefa e texto. Essa relação é o que dá suporte às habilidades de escrita que se pretende ser adquiridas ao longo dos diferentes níveis que tornam um aluno um utilizador elementar, independente ou proficiente da L2. Entre os níveis de competências, o desenvolvimento da produção escrita deve dar-se de forma a resultar em textos organizados, lineares, com uso de recursos adequados ao tipo de texto, contemplando diversos gêneros textuais para os níveis B1/B2, como: manuais de instruções, bandas desenhadas, brochuras e prospectos, folhetos, material publicitário, sinalizações e avisos públicos, letreiros nos supermercados e nas lojas, embalagens e etiquetas de produtos, bilhetes, formulários e questionários, verbetes de dicionários, cartas de negócios e profissionais, faxes, cartas pessoais, memorandos, relatórios e ensaios etc. (Conselho da Europa, 2001).

Relativamente ao nível intermediário e sobre a produção escrita de forma geral, o QECR diferencia as competências entre B1, que "É capaz de escrever textos coesos e simples acerca de um leque de temas que lhe são familiares, relativos aos seus interesses, ligando uma série de elementos pequenos e discretos para formar uma sequência linear", e B2, que "É capaz de escrever textos pormenorizados, com clareza, acerca de vários assuntos relacionados com os seus interesses, sintetizando e avaliando informações e argumentos recolhidos em diversas fontes" (Conselho da Europa, 2001, p. 95).

No contexto brasileiro, há um aumento da procura e da produção científica nas últimas décadas por conta dos movimentos de internacionalização dos programas das universidades e pelo interesse estratégico de algumas empresas internacionais, que resulta em incentivos à procura de cursos da língua. Entretanto, como demonstram Schoffen e Martins (2016), ainda não há no Brasil parâmetros e orientações oficiais produzidas exclusivamente para o ensino do português brasileiro para estrangeiros. Por conta disso, dois documentos oficiais vêm servindo como parâmetro paro o ensino da língua nessa modalidade: os *Parâmetros Curriculares Nacionais para o Ensino de Língua Portuguesa*, publicados em 1997, e as diretrizes do *Certificado de Proficiência em Língua Portuguesa para Estrangeiros*, conhecido como *Exame Celpe-Bras*.

Os parâmetros curriculares da educação brasileira foram publicados em 1997 e sofreram duas importantes atualizações: uma em 1998 direcionada aos anos finais da educação básica (Ensino Fundamental), e outra em 2000 dirigida à educação secundária (Ensino Médio). A concepção de língua subjacente aos documentos é a visão discursiva da linguagem, propagada pelas ideias linguísticas difundidas pelo Círculo de Bakhtin (Bakhtin, 2011). Os documentos assumem como elemento central do objeto de ensino os gêneros do discurso, que são os textos aqui compreendidos como produto da atividade discursiva humana. Configuram-se também em dois eixos norteadores de ensino, o do uso (incluindo a leitura e a produção escrita) e o da reflexão (que inclui o conhecimento metalinguístico). A compreensão dos modos de organização desses discursos e de sua organização em diferentes gêneros textuais com léxico e formas apropriadas para construção de significação passa a ser a base do ensino da língua portuguesa no país (Ministério da Educação e Cultura do Brasil & Secretaria de Educação Fundamental do Brasil, 1998; Ministério da Educação e Cultura do Brasil & Secretaria de Educação Médica e Tecnológica do Brasil, 2000; Secretaria de Educação Fundamental do Brasil, 1997).

As diretrizes do Exame Celpe-Bras operacionalizam esse enquadramento teórico presente na base dos parâmetros curriculares nacionais. O exame certifica quatro níveis, Intermediário, Intermediário Superior, Avançado e Avançado Superior, e a avaliação é feita em uma única prova, que contém uma parte escrita (com tarefas integradas de compreensão oral e escrita e produção escrita) e uma parte oral (com tarefas de

compreensão e produção oral). Tanto a produção oral quanto a escrita são compreendidas como tarefas, que exigem o uso da linguagem como propósito social – uma ação que envolve uma intenção dirigida a um interlocutor. A ênfase do exame está no uso efetivo da língua em textos autênticos e na avaliação integrada entre compreensão e produção. É por essa razão que a prova é a mesma para os diferentes níveis – parte-se da premissa de que "examinandos de todos os níveis são capazes de desempenhar ações em língua portuguesa. O que pode variar é a qualidade desse desempenho, dependendo do nível de proficiência do examinando" (Ministério da Educação e Cultura do Brasil & Instituto Nacional de Estudos e Pesquisas Anísio Teixeira, 2013, p. 5).

Por avaliar a língua em uso, os gêneros do discurso (ou textuais) são também elemento central dos parâmetros de avaliação do Celpe-Bras. Relativamente ao que se espera do nível de proficiência do examinando do Nível Intermediário no tocante à produção escrita, o mesmo documento orienta que esse deve ser capaz de produzir textos escritos sobre assuntos variados que, com dificuldade, podem ser reconhecidos como pertencentes a determinados gêneros discursivos.

Schoffen e Martins (2016) sistematizam as diferenças entre as perspectivas portuguesa e brasileira (sobretudo comparando as recomendações do QECR e QuaREPE e as dos Parâmetros Curriculares Nacionais e Exame Celpe-Bras) e mostram que a organização curricular do ensino de PL2 no contexto português se dá em torno das competências (principalmente as lexicais e gramaticais), que devem ser desenvolvidas por meio de tarefas e, no contexto brasileiro, os gêneros do discurso é que devem organizar a progressão do desenvolvimento das competências, por meio de simulações de situações reais de uso de língua (que são as propostas de produção oral e escrita de gêneros com suporte e interlocutores determinados).

A análise das diretrizes que encontramos nesses documentos oficiais que orientam o ensino de PL2 ajudou-nos a elaborar uma grelha de análise de manuais. Reforçamos que tais diretrizes não tratam efetivamente da produção editorial e não trazem recomendações específicas para esse domínio, mas as analisamos por serem documentos oficiais que veiculam políticas linguísticas para o ensino de PL2.

3. **Metodologia**

Nossa metodologia possui bases na investigação qualitativa (Minayo, 2012), com o emprego de métodos indutivos (Flick, 2002) para observar e descrever padrões, e métodos interpretativos para analisar conteúdo de manuais e de documentos. Através desses métodos de recolha de dados, ocupamo-nos da análise de conteúdo, como método de pesquisa central (Bell, 1993; Bodgan e Byklen, 1994).

Feito esse enquadramento, procedemos à formulação das questões que pretendemos avaliar nos manuais didáticos:

1. Como os documentos norteadores para o ensino de L2 orientam o trabalho a ser conduzido para o desenvolvimento da competência escrita nos manuais didáticos?
2. Que concepções e modelos de escrita podem ser encontrados em manuais didáticos de PL2?
3. Como o manual didático prepara o aluno para a produção escrita solicitada?

Fizemos uma seleção de manuais didáticos com base no critério de uso e podemos afirmar que o *corpus* constituído para a análise é sincrônico, porque não tivemos o objetivo de analisar a evolução nem buscar historicidade da produção editorial; estático, pois analisamos manuais finalizados e destinados à venda para públicos que os consomem; e especial, na medida em que se trata de uma produção de saber especializada (Sinclair, 1996).

Relativamente ao tipo de recurso selecionado, embora tenhamos uma grande produção de componentes que utilizam as outras tecnologias como as móveis e digitais (*podcasts, sites, blogs* etc.), escolhemos centralizar a análise em manuais impressos porque esse continua a ser o recurso principal do ensino de L2.

Quanto ao nível de aprendizagem, os manuais selecionados destinam-se aos níveis B1 e B2 para Portugal e intermediário para Brasil, porque nessa etapa exige-se composições textuais mais bem elaboradas nos níveis lexicais, sintáticos, semânticos e pragmáticos-discursivos. Dessa exigência emerge a necessidade de instruções detalhadas para o desenvolvimento da competência compositiva, além do ensino dos padrões retóricos e organizacionais dos gêneros a serem produzidos.

Em relação ao público-alvo e contexto de aprendizagem, os manuais selecionados são produzidos para adultos e para o ensino de língua em espaços formais (cursos de língua de institutos, escolas, universidades etc.). Analisamos manuais de PL2 produzidos em Portugal e no Brasil.

Produzimos um inquérito por questionário para ser respondido por professores de PL2 de instituições portuguesas e de uma instituição brasileira. Das respostas dadas a esse inquérito, selecionamos o manual didático mais utilizado ou recomendado por professores da FCSH-UNL e o mais usado ou referenciado por docentes da Universidade Tecnológica Federal do Paraná (porque iremos desenvolver um estudo experimental nessas universidades em um segundo momento).

A análise dos manuais selecionados desenvolveu-se, conforme os critérios estabelecidos por Aran (2001, 2007), nos âmbitos que atuam em função das intenções educacionais (para cotejar as propostas de produção textuais *versus* recomendações gerais dos documentos norteadores para o ensino de PL2), e em função dos requisitos para a aprendizagem (para comparar as propostas de produção textuais *versus* recomendações gerais para o ensino da competência escrita em L2).

Com base nas recomendações fornecidas pela revisão do estado da arte sobre o desenvolvimento da escrita em L2 e pela análise dos documentos norteadores para o ensino de PL2, elaboramos uma grelha de análise com os principais critérios que estabelecemos para avaliar os manuais, que são a abordagem metodológica, as concepções/modelos de escrita identificados, a distribuição das quatro competências do ensino de L2 entre as atividades propostas, o mapeamento dos gêneros solicitados nas tarefas de produção escrita, e as orientações para docentes.

4. Resultados e discussão

4.1. Observações dos docentes quanto aos manuais didáticos e seu uso

Responderam ao inquérito 32 docentes de PL2 distribuídos entre as seguintes instituições de ensino: 10 docentes da FCSH-UNL; 5 docentes da Universidade de Coimbra; 2 docentes da Universidade de Aveiro; 2 docentes da Universidade do Algarve; 3 docentes da Universidade de Lisboa; 3 docentes da Universidade Tecnológica Federal do Paraná-BR; e 7 docentes que não identificaram a universidade onde atuam.

O inquérito empregado foi um questionário enviado aos docentes por meio de um *link* de acesso ao *Google Docs* e seu objetivo principal foi conhecer o manual didático de PL2 mais utilizado ou recomendado por esses professores, além de saber sua apreciação sobre o uso de manuais de ensino e sobre a forma como desenvolvem a competência escrita.

Relativamente à formação desses profissionais, 28% possuem licenciatura em ensino de língua, 28% têm especialização em ensino de português língua estrangeira, 15,6% não têm formação específica nesse âmbito, 9,4% são mestres em ensino de português língua estrangeira, 6,3% são doutores em ensino de português língua estrangeira e 12,4% são formados em outras especialidades da área de Letras.

Quanto aos níveis de atuação, a maior parte dos docentes atua nos iniciais, com 78% no A1, 59,4% no A2, seguido do nível intermediário com 56,3% atuando no B1 e 40,6% no B2 e, por último, nos níveis avançados com 37,5% no C1 e 9,4% no C2. Dessa amostra, 78% utilizam manual didático de PL2 e outros materiais, 18,8% não usam livro didático, mas somente outros materiais e 3% fazem uso somente de um manual.

Em relação à importância do uso de manuais didáticos para o ensino da língua, 75% dos docentes consideram que são importantes por atuarem como guião para os conteúdos que são dados em sala de aula; 46,9% também assim o consideram por uniformizarem aspectos que são essenciais no ensino da L2; 15,6%

creditam a relevância desses materiais ao fato de serem produzidos levando em consideração investigações científicas; e 3% consideram que têm valor especialmente para professores iniciantes. Em contrapartida, 28% pensam que os manuais não têm importância porque não abrangem as especificidades de grupos diversificados; e 6,3% não veem relevância no seu uso porque podem conduzir os docentes à passividade.

Solicitamos aos participantes que elencassem, de forma geral e conforme sua apreciação na prática docente, aspectos que consideram positivos e negativos do manual que utilizam, os quais destacamos na Tabela 1.

Tabela 1. Avaliação de aspectos positivos e negativos dos manuais didáticos

Aspectos positivos	**Aspectos negativos**
Uso de temas atuais.	Produções escritas irreais.
Desenvolvimento de todas as competências.	Carência de textos que abordam elementos culturais.
Uso de diálogos vivos e naturais.\	Desatualização de conteúdos.
Apropriado ao público-alvo.	Inadequação de vocabulário e gramática.
Textos para compreensão escrita.	Ausências de textos para o trabalho com unidades gramaticais.
Sistematização de atos de fala.	Unidades muito longas.
Variedade de temas, léxico e expressões idiomáticas.	Falta de exercícios comunicativos.
Esquematização da gramática.	Inexistência de textos literários com figuras de estilo.
Uniformização de conteúdos.	Pouca variedade entre os exercícios.
Diversidade de exercícios para o trabalho com diferentes competências.	Falta de textos autênticos e ensino explícito da competência escrita.

Também investigamos como os docentes avaliam a distribuição das competências comunicativas no(s) manual(is) que utilizam: 42,3% dizem que não é equitativa, pois algumas são menos contempladas; 30,8% afirmam que a promoção das competências é feita de forma integrada; 15,4% assinalam que o manual em questão não desenvolve todas as competências comunicativas; e para 11,5% essas são desenvolvidas, mas não de forma integrada.

Relativamente ao desenvolvimento da competência escrita no manual utilizado, 80,8% dos docentes respondem de forma positiva, ou seja, afirmam que o livro desenvolve a escrita nos estudantes, enquanto 19,2% dizem o oposto. De forma pormenorizada, segundo 85,7% dos participantes, o livro apresenta propostas de produção textual adequadas ao público-alvo; para 57,1%, o manual fornece modelos variados para uso como exemplo; 47,6% dizem que o manual ensina, por meio de exemplos, padrões de composição de diferentes gêneros. Apesar disso, apenas 14,3% dos docentes afirmam que o manual didático que utilizam estimula a prática quanto ao planejamento da escrita, à releitura e à revisão dos textos.

4.2. **Análise dos manuais selecionados**

Iniciamos a discussão com a apresentação da análise dos manuais da coleção *Português em Foco*, volumes 2 e 3, na Tabela 2, que consiste na grelha de análise elaborada com base no que discutimos em nosso enquadramento teórico.

Tabela 2. Análise da coleção *Português em Foco* (Livro do aluno e Caderno de exercícios, Vols. 2 e 3)

Descrição externa do manual didático	
Título:	*Português em Foco 2*. Livro do aluno; *Português em Foco 2*. Caderno de exercícios; *Português em Foco 3*. Livro do aluno; *Português em Foco 3*. Caderno de exercícios.
Autores:	Luísa Coelho, Carla Oliveira, coordenação: João Malaca Casteleiro.
Editora:	Lidel Edições Técnicas, Lisboa, 2021.
Design e layout:	Livro do aluno: colorido, com fotografias, ilustrações, esquemas didáticos, quadros e tabelas. Caderno de exercícios: a preto e branco, com fotografias, ilustrações, esquemas didáticos, quadros e tabelas.
Sinopse:	*Português em Foco 2*: é destinado a estudantes que desejam aprofundar os conhecimentos na língua portuguesa, pretendendo atingir o nível B1, conforme os critérios do QECR. É construído seguindo princípios metodológicos da abordagem comunicativa, os temas são trabalhados ao longo de 12 unidades (e uma inicial, de revisão) de forma progressiva, para propiciar o desenvolvimento de estruturas gramaticais e vocabulário esperados à conclusão do nível B1. Pretende dar aos estudantes um domínio satisfatório das capacidades de compreensão e de produção oral e das competências de leitura e de escrita. É acompanhado de Caderno de exercícios (com tarefas adicionais para cada unidade) e de Manual do professor (com indicações para o desenvolvimento das atividades na sala de aula). O Caderno de exercícios 2 é um componente do conjunto oferecido pela editora e possui relação direta com o Livro do Aluno. Assim, também é destinado a estudantes que querem aprofundar os conhecimentos na língua portuguesa, pretendendo atingir o nível B1. Os temas são trabalhados de acordo com as mesmas 12 unidades do Livro do Aluno, seguindo a mesma progressão, e com tarefas que são consideradas adicionais às do livro principal. *Português em Foco 3*: é destinado a estudantes que querem aprofundar os conhecimentos na língua portuguesa, pretendendo atingir o nível B2. Também segue princípios metodológicos da abordagem comunicativa, ao longo de 12 unidades (e uma inicial de revisão), de forma progressiva, para propiciar o desenvolvimento de estruturas gramaticais e vocabulário esperados à conclusão do nível B2. Os conteúdos programáticos desse volume privilegiam áreas culturais da sociedade portuguesa, com o objetivo de estreitar a relação entre língua e cultura. Os arquivos de áudio são gravados por falantes nativos e respeitam a norma-padrão do português europeu. Também é acompanhado de Caderno de exercícios e de Manual do professor. Os temas do Caderno de exercícios 3 são trabalhados de acordo com as mesmas 12 unidades do Livro do aluno, seguindo a mesma progressão, e com tarefas que são consideradas adicionais às do livro principal.
Descrição interna do manual didático	
Público-alvo / nível:	Jovens e adultos, nível B1 e B2.
Componentes:	Livro do aluno com ficheiro de arquivos de áudio, acessados via *web*. Caderno de exercícios consumível.
Distribuição dos conteúdos:	Livro do aluno: Unidades; Textos áudio não transcritos; Soluções dos testes de revisão; Glossário; e Listas de faixas áudio. Caderno de exercícios: Unidades; e Soluções.

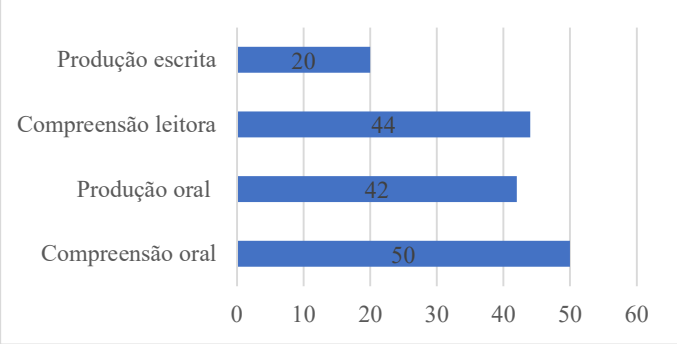
Aspectos didáticos e metodológicos do manual didático		
Análise do manual didático	**Sim**	**Não**
Identifica-se a abordagem e metodologia de ensino do manual?	X	
Abordagem (s): Abordagem comunicativa.		
O manual trabalha o desenvolvimento das quatro competências?		X

As quatro competências são trabalhadas de forma equilibrada?		X
Comentário: O Caderno de exercícios 2 não contém exercícios de desenvolvimento de compreensão oral e de produção oral e o Caderno de exercícios 3 não possui exercícios de desenvolvimento da competência oral.		
Identifica-se uma (ou mais) concepção/modelo de escrita no material?	X	
Concepção/Modelo (s): Escrita como produto – apenas em um exercício do Livro do Aluno 3 identifica-se o modelo de escrita como processo, quando se solicita a escrita de um artigo de opinião.		
O material utiliza textos autênticos?		X
O manual fornece uma síntese curta ou apresentação do que será exposto nos textos?		X
Há diversidade de gêneros textuais nas tarefas de produção?	X	
Gêneros solicitados: *E-mail*; mensagem SMS; continuação de narração; comentário; notícia; currículo; carta de apresentação; convite formal e convite informal; descrição; artigo de opinião; receita; ementa; lenda; anúncio de vaga de emprego; resumo; textos sem gênero definido.		
As propostas de produção textual possuem suporte(s) e interlocutor(s) definidos?		X
Os gêneros textuais estão de acordo com os recomendados pelos documentos norteadores, segundo a distribuição dos níveis?	X	
Há instrução para a produção dos gêneros solicitados?		X
Há espaço/instrução para o planejamento da produção escrita?		X
Há espaço/instrução para a revisão do texto?		X
O manual fornece instruções para docentes?	X	
As instruções orientam o trabalho a ser feito para a produção textual?		X
Dentre as instruções dadas, há fornecimento de modelos de textos e critérios para correção das produções dos estudantes?		X

A coleção *Português em Foco* é construída utilizando os pressupostos da abordagem comunicativa, mas não deixa de lançar mão da abordagem estrutural, que se reflete sobretudo na seção *Gramática*, por tratar essencialmente do desenvolvimento de estruturas gramaticais. As unidades da coleção são elaboradas de forma a desenvolver as competências comunicativas, inseridas em situações reais de usos de língua, mas trazem seções essencialmente estruturais, seguidas de exercícios desvinculados de contextos, seguindo o modelo de *Presentation, Practice, Production* (Howatt, 1984), o que caracteriza uma abordagem comunicativa fraca e corrobora as análises de Castro (2015) e de Tomlison e Masuhara (2018).

No tocante à distribuição das competências nas atividades e exercícios propostos nos manuais, as mais desenvolvidas nas atividades propostas pelos manuais são as de compreensão oral, que totalizam 50, seguidas das de compreensão leitora, com 44 exercícios e tarefas, de produção oral, com 42 e de produção escrita, com apenas 20 exercícios, ocupando o último lugar, conforme vemos na Figura 1.

Figura 1. Distribuição das competências nas atividades da coleção *Português em Foco* (livro do aluno, Vols. 1 e 2)



Relativamente ao modelo de escrita identificado nas propostas de produção da coleção, observamos a predominância de escrita como produto (Hyland, 2003, 2009), na qual o texto independe de contextos reais de

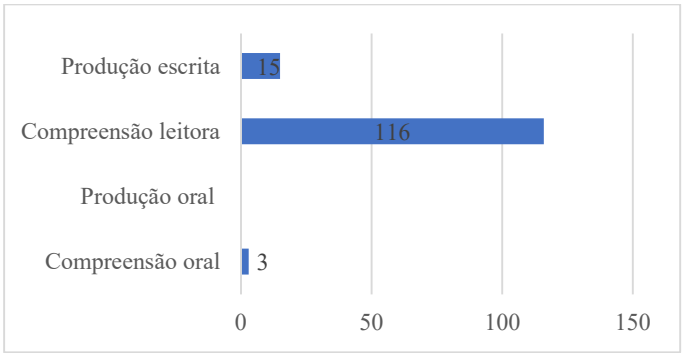
produção. Muitas das tarefas propostas não especificam gênero e interlocutor, o que se afasta dos modelos que deveriam ser utilizados por uma abordagem de ensino efetivamente comunicativa e se aproxima às metodologias estruturalistas, nas quais o foco da aprendizagem é a precisão gramatical na formulação de frases (ou textos) isolados das situações reais e complexas de comunicação.

Os manuais não utilizam textos autênticos e não fornecem pequenas introduções ou sínteses dos temas que serão tratados – o que seriam recomendações de alguns documentos norteadores do ensino de PL2 em Portugal. Tendo em conta as orientações de gêneros que os usuários devem ser capazes de produzir para atingirem a proficiência nos níveis B1 e B2, podemos dizer que há relativa diversidade, já que o manual pede produções de textos eletrônicos (*e-mails* e SMS, por exemplo) e outros diversos como cartas de apresentação, currículo, anúncios de vaga de emprego, ainda que na grande maioria das vezes não indique interlocutores nem suportes onde deveriam circular a produção.

Em relação àquilo que se recomenda para o processo de ensino da produção escrita, destacamos que as propostas do manual carecem de explicitação dos objetivos do público-alvo a quem se dirige o texto, de instrução de esquemas e padrões composicionais e hierarquias de composição (que se traduzem em um planejamento), de incentivo à releitura, à revisão da produção e ao uso do dicionário, ou mesmo de listas de léxico que auxiliem o planejamento da produção dos textos.

A abordagem do Caderno de exercícios é a mesma do Livro do aluno, ou seja, é comunicativa, mas a consideramos ainda mais fraca, pois a maior parte dos exercícios são puramente estruturais e trabalham sobretudo aspectos gramaticais da língua. Além disso, não contemplam o desenvolvimento de todas as competências, como demonstramos no gráfico que se apresenta na Figura 2.

Figura 2. Distribuição das competências nas atividades da coleção "Português em Foco" (Caderno de exercícios, Vols. 1 e 2)



Observamos que a competência de produção oral não é contemplada no Caderno de exercícios, que propõe 116 exercícios de compreensão leitora e somente 15 deles são destinados a propostas de produção escrita, seguidos de 2% de tarefas destinadas à compreensão oral.

Também como no Livro de Aluno, o modelo de ensino de escrita utilizado nas propostas de produção é o da escrita como produto, apresentando as mesmas fragilidades de ausência de contexto e interlocutor, e muitas vezes de um gênero textual definido.

Relativamente às orientações metodológicas oferecidas no Manual do professor, no tocante ao desenvolvimento e à correção das propostas de produção textual, observamos que há em maioria a repetição explícita e literal dos objetivos da proposta. Esperar-se-ia que o manual dos docentes oferecesse estratégias para o desenvolvimento dos objetivos, modelos e critérios de correção e instruções para o desenvolvimento metodológico das propostas.

A grelha que se apresenta na Tabela 3 a seguir mostra nossa análise do manual *Novo Avenida Brasil*, com nossas observações sobre o Livro do aluno (ou Livro-texto, como é chamado).

Tabela 3. Análise do livro *Novo Avenida Brasil 3. Curso Básico de Português para Estrangeiros* (Livro-texto e Caderno de exercícios)

Descrição externa do manual didático	
Título:	*Novo Avenida Brasil 3. Curso Básico de Português para Estrangeiros*.
Autores:	Emma Lima, Lutz Rohrmann, Tokiko Ishihara, Samira Iunes, Cristián Bergweilwer.
Editora:	E.P.U., Rio de Janeiro, 2022.
Design e layout:	Colorido, com fotografias, ilustrações, esquemas didáticos, quadros e tabelas.
Sinopse:	O manual é destinado a estudantes que pretendem aprofundar os conhecimentos no português brasileiro, atingindo o nível intermediário (do início ao fim), conforme os critérios das diretrizes do Exame Celpe-Bras. É construído seguindo princípios metodológicos da abordagem comunicativa e estruturalista, os temas são trabalhados ao longo de 8 unidades, que são chamadas de *lições* (e duas de revisão), de forma progressiva, para propiciar o desenvolvimento de estruturas gramaticais e vocabulário esperados à conclusão do nível intermediário. Pretende dar aos estudantes um domínio satisfatório das capacidades de produção oral (tida como a principal) e de produção escrita (que se afirma ser mais desenvolvida nos exercícios). O manual integra em um só volume o Livro do Aluno e o Caderno de Exercícios (que devem ser trabalhados em casa, segundo indicação). O livro não possui manual do professor, mas traz uma seção com explicação de uso de ícones que direcionam a forma de serem desenvolvidas habilidades em cada seção, além de dividir os conteúdos em "blocos de cadência", que esquematizam a ordem sequencial dos temas. As tarefas do caderno de exercícios dão mais destaque ao desenvolvimento das competências de leitura e de escrita, quando comparadas às presentes no livro-texto, mas também trabalham habilidades gramaticais por meio de exercícios estruturais, como preenchimento de lacunas.
Descrição interna do manual didático	
Público-alvo / nível:	Jovens e adultos, nível intermediário completo.
Componentes:	Livro do aluno (Livro-texto) com Caderno de exercícios (Livro de exercícios) consumível e ficheiros de áudio disponíveis na *web*.
Distribuição dos conteúdos:	Lições; Revisão 1; Lições; Revisão 2; Fonética; Apêndice gramatical; Textos gravados; Soluções; Vocabulário alfabético; Créditos.

Aspectos didáticos e metodológicos do manual didático		
Análise do manual didático	**Sim**	**Não**
Identifica-se a abordagem e metodologia de ensino do manual?	X	
Abordagem (s): Abordagem comunicativo-estrutural.		
O manual trabalha o desenvolvimento das quatro competências?	X	
As quatro competências são trabalhadas de forma equilibrada?		X
Identifica-se uma (ou mais) concepção/modelo de escrita no material?	X	
Concepção/Modelo (s): Escrita como produto.		
O material utiliza textos autênticos?	X	
O manual fornece uma síntese curta ou apresentação do que será exposto nos textos?		X
Há diversidade de gêneros textuais nas tarefas de produção?	X	
Gêneros solicitados: *E-mail*; comentário em rede social; lenda; formulário; mensagem de *whatsapp*; carta formal; resumo; texto sem gênero definido.		
As propostas de produção textual possuem suporte(s) e interlocutor(es) definidos?		X
Os gêneros textuais estão de acordo com os recomendados pelos documentos norteadores, segundo a distribuição dos níveis?	X	
Há instrução para a produção dos gêneros solicitados?		X
Há espaço/instrução para o planejamento da produção escrita?		X
Há espaço/instrução para a revisão do texto?		X

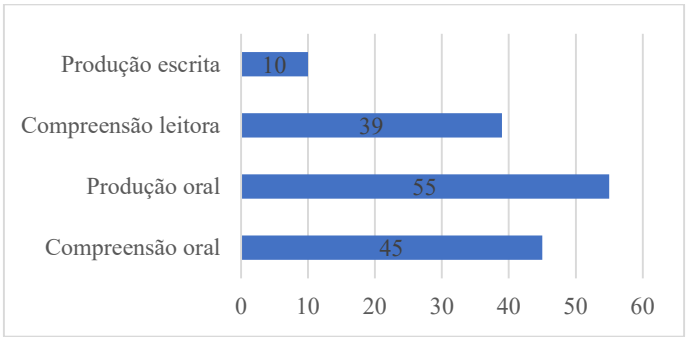
O manual fornece instruções para docentes?	X	
As instruções orientam o trabalho a ser feito para a produção textual?		X
Dentre as instruções dadas, há fornecimento de modelos de textos e critérios para correção das produções dos estudantes?		X

Os autores do manual afirmam que a coleção *Novo Avenida Brasil* é construída seguindo os pressupostos do que chamam *abordagem comunicativo-estrutural*, isso porque o método se propõe comunicativo, mas utiliza a abordagem estrutural, que se reflete sobretudo em uma seção específica, denominada *B*, que trata essencialmente do desenvolvimento de estruturas gramaticais isoladas de contextos amplos de comunicação. Nesse sentido, a abordagem do manual é bastante semelhante à da série *Português em Foco*, porque desenvolve as competências comunicativas, mas faz uso do modelo de *Presentation, Practice, Production* (Howatt, 1984).

O modelo de escrita utilizado nas atividades do manual é também o de escrita como produto (Hyland 2003, 2009), que se traduz em propostas de produção textual isoladas de contextos, interlocutores, suportes e, muitas vezes, de gêneros específicos. Esse último aspecto é relativamente problemático quando retomamos as diretrizes dadas nos *Parâmetros Curriculares Nacionais* e para o *Exame Celpe*-Bras, que apresentam uma concepção de linguagem dialógica e, portanto, pautada na organização das atividades de produção linguística em torno dos gêneros discursivos.

Relativamente à distribuição das competências comunicativas, o manual privilegia o desenvolvimento da oralidade com 55 atividades propostas para a produção oral, 45 para compreensão oral, 39 para compreensão leitora e somente 10 para a produção escrita, que ocupa uma posição muito inferior, conforme evidenciamos na Figura 3.

Figura 3. Distribuição das competências nas atividades do manual *Novo Avenida Brasil* (Livro do aluno)

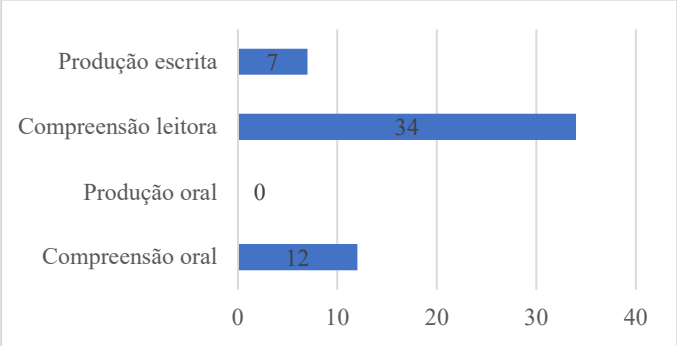


Há a presença de alguns textos autênticos no manual, mas não introduções nem sínteses dos temas que serão tratados (ainda que essa orientação seja dada somente nos documentos norteadores de Portugal). O manual traz pouca diversidade de gêneros nas propostas de produção: *e-mail* e cartas são gêneros que se repetem e quase não há relação entre os temas da escrita com os das unidades. Entretanto, tendo em conta as recomendações do que os usuários devem ser capazes de produzir, podemos dizer que atua em consonância com os documentos norteadores. Por utilizar um modelo de escrita como produto, o manual não fornece instruções, modelos, estruturas e léxico específico para a produção solicitada, nem incentiva o planejamento de escrita, a revisão da produção textual e a consulta ao dicionário.

A edição recente unificou livro do aluno e de exercícios em um só volume, dessa forma, o Livro-texto segue os mesmos pressupostos teórico-metodológicos da abordagem comunicativo-estrutural. Há muitos exercícios de compreensão leitora e de desenvolvimento de estruturas gramaticais com preenchimentos de lacunas, que seguem os pressupostos da abordagem estruturalista. O modelo de escrita é o mesmo utilizado no Livro do Aluno, que é o da escrita como produto, apresentando os mesmos problemas de propostas sem contexto, interlocutor, suporte e, algumas vezes, sem gênero definido.

As competências comunicativas não são distribuídas de forma equilibrada. A produção oral não é contemplada, 34 exercícios são destinados ao desenvolvimento da compreensão leitora, 12 à compreensão oral e somente 7 ao desenvolvimento da produção escrita, novamente ocupando uma posição muito inferior, conforme observamos na Figura 4.

Figura 4. Distribuição das competências nas atividades do manual *Novo Avenida Brasil* (Livro-texto)



Como dissemos, o Livro-texto também utiliza um modelo de escrita como produto e, como consequência, não fornece instruções, modelos, estruturas e léxico específico para a produção textual e não incentiva ou destina espaço ao planejamento e à revisão da escrita. Além disso, não há uso de textos autênticos, há pouca diversidade de gêneros textuais solicitados para produção e esses não apresentam contexto, interlocutor e suporte definidos.

A análise de nosso *corpus* corrobora as observações aventadas na revisão da literatura sobre o ensino da competência escrita em L2, que é tratada com pouca ênfase, seja pelo número reduzido de tarefas propostas, quando comparadas às das outras competências, seja pela ausência de instrução de processos composicionais.

5. Considerações finais

Nesse estudo, nos perguntamos que concepções e modelos de escrita podemos encontrar em manuais didáticos de PL2 e de que forma esses preparam os estudantes para a produção escrita.

Para tanto, analisamos os documentos norteadores do ensino de PL2 e vimos que orientam o desenvolvimento de um trabalho direcionado à ação, isto é, compreendendo a produção escrita como a execução de uma tarefa que requer o uso de estratégias para atingir objetivos previamente levantados e inseridos em contextos reais de uso de língua, com interlocutores, suporte e gênero textual definido.

Apesar disso, observamos que os manuais didáticos de PL2 analisados utilizam, de forma majoritária, o modelo da escrita como produto. As principais fragilidades encontradas nos manuais foram: (i) não utilização de um modelo de escrita que conceba a produção textual como a execução de uma tarefa que requer o estabelecimento de estratégias; (ii) ausência de instruções para a produção textual, como modelos de textos de exemplos, listas de vocabulário temáticos e exemplos de padrões textuais utilizados nos diferentes gêneros textuais; (iii) inexistência de incentivo ou seções destinadas ao planejamento e à revisão da produção escrita; (iv) carência de textos autênticos; (v) falta de síntese ou apresentação de temas que serão tratados nos textos; (vi) distribuição desigual entre as tarefas de desenvolvimento das competências comunicativas; (vii) não interligação da competência escrita às demais; e (viii) insuficiência de orientações para docentes com modelos e critérios para correção e desenvolvimento metodológico das propostas de produção textual.

Sendo a escrita um processo individual e social, julgamos adequado um modelo de ensino dessa competência que privilegie aspectos cognitivos e de natureza social e discursiva, que deveria focar-se em processos individuais de levantamento de objetivos e emprego de estratégias para alcançá-los, sempre inseridos

em contextos sociais relevantes. Os modelos cognitivos são úteis para diferenciar produções de escritores menos e mais experientes e para que dê a devida ênfase ao processo de composição de textos, com suas diferentes etapas. O ensino sistemático e estruturado de gêneros textuais é fundamental para a transposição de modelos na escrita, visto que é por meio do gênero que o escritor aciona as formas retóricas tipificadas e socialmente compartilhadas para produzir sentido na língua-alvo. Como evidencia a análise de nosso *corpus*, tal modelo não vem sendo explorado em manuais didáticos de PL2.

Apesar do recorte limitado, justificamos nossa análise sobretudo pela ausência de investigações que tratam do ensino da competência escrita de PL2 e sua produção editorial. Acreditamos que esse estudo pode fornecer subsídios a futuras pesquisas nessa agenda de investigação.

Agradecimentos

Agradecemos aos docentes da área de PL2 da FCSH-UNL, da Universidade Tecnológica Federal do Paraná, da Universidade de Coimbra, da Universidade de Aveiro, da Universidade do Algarve e da Universidade de Lisboa pelas contribuições fornecidas no inquérito utilizado nesse estudo.

Referências

Aran, Artur (2001) ¿Servir al material o servirse del material? Evaluar los materiales curriculares para mejorar su uso. *Kirikiri. Cooperación Educativa* 61, pp. 44–49.

Aran, Artur (2007) *Materiales curriculares. Cómo elaborarlos, seleccionarlos y usarlos*. Editorial GRAÓ de IRIF S.L.

Bakhtin, Mikhail (2011) *Estética da criação verbal*. Martins Fontes.

Barbeiro, Luís (2000) Profundidade do processo de escrita. *Educação e Comunicação* 5, pp. 64–76.

Barbeiro, Luís & Luísa Pereira (2007) *O ensino da escrita*: *A dimensão textual*. Ministério da Educação & Direcção-geral de Inovação e de Desenvolvimento Curricular.

Barbosa, Gabriela & Rosa Bizarro (2015, 16–17 outubro) *O ensino da escrita em português língua não materna: Conceções e práticas de professores africanos* [Apresentação de comunicação]. Simpósio SIPLE 2015. O português em espaços multilíngues, Santiago de Compostela, Galiza.

Bell, Judith (1993) *Como realizar um projeto de investigação*. Gradiva.

Berman, Robert (1994) Learners' transfer of writing skills between languages. *TESL Canada Journal* 12 (l), pp. 29–46. https://doi.org/10.18806/tesl.v12i1.642

Bhowmik, Subrata (2021) Writing instruction in an EFL context: Learning to write or writing to learn language? *Belta Journal* 5 (1), pp. 30–42. https://doi.org/10.36832/beltaj.2021.0501.03

Bodgan, Robert & Sari Biklen (1994) *Investigação qualitativa em educação. Uma introdução à teoria e aos métodos*. Porto Editora.

Cassany, Daniel (1999) *Construir la escritura*. Ediciones Paidós Ibérica S.A.

Cassany, Daniel (2017) *Describir el escribir. Cómo se aprende a escribir*. Espasa Libros S.L.U.

Castro, Catarina (2015) Existem razões para se continuar a usar manuais no ensino de línguas? Algumas considerações sobre o seu papel atual e funcionalidade. *Agália. Revista de Estudos de Cultura* 111, pp. 155–172.

Conselho da Europa (2001) *Quadro Europeu Comum de Referência para as Línguas. Aprendizagem, ensino, avaliação (QECR)*. Asa.

Cumming, Alister (2001) Learning to write in a second language: Two decades of research. *International Journal of English Studies* 1 (2), pp. 1–23. https://doi.org/10.6018/ijes.1.2.48331

Cunningsworth, Alan (1995) *Choosing your coursebook*. Heineman.

Decreto-Lei n. º 6, de 18 de janeiro de 2001 (2001) Diário da República n.º 15/2001, Série I-A de 2001-01-18.

Ellis, Rod (2006) Researching the effects of form-focused instruction on L2 acquisition. *AILA Review* 19, pp. 18–41. https://doi.org/10.1075/aila.19.04ell
Ellis, Rod, Sahwn Loewen & Jenefer Philp (2009) Implicit and explicit corrective feedback and the acquisition of L2 grammar. In Rod Ellis et al. (eds.), *Implicit and explicit knowledge in second language learning, testing and teaching*. Multilingual matters, pp. 303–332.
Fairclough, Norman (1989) *Language and power*. Longman.
Flick, Uwe (2002) Qualitative research - state of the art. *Social Science Information* 41 (1), pp. 5–24. https://doi.org/10.1177/0539018402041001001
Flower, Linda & John Hayes (1981) A cognitive process theory of writing. *CCC* 32, pp. 365–387. https://doi.org/10.2307/356600
Grabe, William & Robert Kaplan (1996) *Theory and practice of writing*. Longman.
Graham, Steve (2018) Introduction to conceptualizing writing. *Educational Psychologist* 53 (4), pp. 217–219. https://doi.org/10.1080/00461520.2018.1514303
Grosso, Maria José (coord.) (2011) *Quadro de referência para o ensino de português no estrangeiro: Documento orientador*. DGIDC. Disponível em https://www.dge.mec.pt/sites/default/files/EEstrangeiro/2012_quarepe_docorientador.pdf
Han, Jimon & Phil Hiver (2018) Genre-based L2 writing instruction and writing-specific psychological factors: The dynamics of change. *Journal of Second Language Writing* 40, pp. 44–59. https://doi.org/10.1016/j.jslw.2018.03.001
Hayes, John (1996) A new framework for understanding cognition and affect in writing. In C. Michael Levy & Sarah Ransdell (eds.), *The science of writing: Theories, methods, individual differences, and applications*. Routledge, pp. 1–27.
Hayes, John & Linda Flower (1980) *Identifying the organization of writing processes. Cognitive processes in writing*. Lawrence Erlbaum Associates.
Hyland, Ken (2003) *Second language writing*. Cambridge University Press.
Hyland, Ken (2009) *Teaching and researching* writing. Pearson Education Limited.
Leiria, Isabel (coord.) (2008) *Orientações programáticas de português língua não materna (PLNM): Ensino secundário*. DGIDC. Disponível em https://www.dge.mec.pt/sites/default/files/ficheiros/eb_orient_programat_plnm_versaofinalabril08.pdf
Leki, Ilona, Alistair Cumming & Tony Silva (2008) *A synthesis of research on second language writing in English*. Routledge Taylor & Francis Group.
Li, Shaofeng (2023) Working memory and second language writing: A systematic review. *Studies in Second Language Acquisition* 45 (3), pp. 647–679. https://doi.org/ 0.1017/S0272263123000189
Lightbrown, Patsy (1998) The importance of timing in focus on form. In Catherine Doughty & Jessica Williams (eds.) *Focus on form in classroom second language acquisition*. Cambridge University Press, pp. 177–196.
Long, Michael (1991) Focus on form: A design feature in language teaching methodology. In Kees de Bot, Ralph B. Ginsberg & Claire Kramsch (eds.), *Foreign language research in cross-cultural perspective*. John Benjamins, pp. 39–52.
Long, Michael (2015) *Second language acquisition and task-based language teaching*. Wiley Blackwell.
Lopes, Ângela & Maria da Graça Pinto (2022) Assessing L2 Portuguese writing: Idea density and sentence complexity. *Signo* 47 (88), pp. 72–85. https://doi.org/10.17058/signo.v47i88.17384
Lyster, Roy & Leila Ranta (1997) Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition* 20, pp. 37–66. https://doi.org/10.1017/S0272263197001034
Lyster, Roy & Kazuya Saito (2010) Effects of oral feedback in SLA classroom research: A meta-analysis. *Studies in Second Language Acquisition* 32, pp. 265–302. https://doi.org/10.1017/S0272263109990520
Matsumoto, Kazuko (1995) Research paper writing strategies of professional Japanese EFL Writers. *TESL Canada Journal* 13 (1), pp. 17–27. https://doi.org/10.18806/tesl.v13i1.658

Minayo, Maria Cecília (2012) *Pesquisa social. Teoria, método e criatividade*. Editora Vozes.
Ministério da Educação e Cultura do Brasil & Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (2013) *Guia do participante*: *Tarefas comentadas que compõem a edição de abril de 2013 do exame*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
Ministério da Educação e Cultura do Brasil & Secretaria de Educação Fundamental do Brasil (1998) *Parâmetros Curriculares Nacionais*: *terceiro e quarto ciclos do Ensino Fundamental: Língua portuguesa*. MEC/SEF.
Ministério da Educação e Cultura do Brasil & Secretaria de Educação Médica e Tecnológica do Brasil (2000) *Parâmetros Curriculares Nacionais*: *Ensino médio*. MEC/SEF.
Ossenbach, Gabriela (2010) Manuales escolares y patrimonio histórico-educativo. *Educatio Siglo XXI* 28 (2), pp. 115–132.
Perdigão, Manuela (coord.) (2005) *Português língua não materna no currículo nacional*: *Documento orientador*. DGIDC. Disponível em https://www.dge.mec.pt/sites/default/files/Basico/Documentos/plnmdoc_orientador.pdf
Ranta, Leila & Roy Lyster (2007) A cognitive approach to improving immersion students oral language abilities: The Awareness–Practice–Feedback sequence. In Robert M. DeKeyser (ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press, pp. 141–160.
Rauber, Bárbara (2017) O papel do feedback corretivo na sala de aula de português como língua estrangeira. *Revista EntreLínguas* 3 (10), pp. 6–18. https://doi.org/10.29051/rel.v3.n1.jan-jun.2017.9015
Rinnert, Carol & Hiroe Kobayashi (2009) Situed writing practices in foreign language settings: the role of previous experience and instruction. In Rosa M. Manchón (ed.), *Writing in foreign language contexts. Learning, teaching and research*. Multilingual Matters, pp. 23–48.
Scardamalia, Marlene & Carl Bereiter (1987) Knowledge telling and knowledge transforming in written composition. In Sheldon Rosenberg (ed.), *Advances in applied psycholinguistics*. Cambridge University Press, pp. 142–175.
Schoffen, Juliana & Alexandre Martins (2016) Políticas linguísticas e definição de parâmetros para o ensino de português como língua adicional: perspectivas brasileira e portuguesa. *ReVEL* 14 (26), pp. 271–306.
Secretaria de Educação Fundamental do Brasil (1997) *Parâmetros curriculares nacionais*: *Introdução aos parâmetros curriculares nacionais*. MEC/SEF.
Silva, Tony (1990) *Second language writing: Research insights for the classroom*. Cambridge University Press.
Silva, Tony (1993) Toward an understanding of the distinct nature of L2 writing. *TESOL Quarterly* 27 (4), pp. 657–677. https://doi.org/10.2307/3587400
Sinclair, John (1996) *Preliminary recommendations on Corpus Typology*. EAGLES (Expert Advisory Group on Language Engineering Standards).
Siopa, Conceição (2015) Competências de escrita no ensino superior e o tratamento do erro em português L2. In Mónica Bastos, José Marques, Ana Catarina Monteiro & Conceição Siopa (orgs.), *Ensinar a língua portuguesa em Moçambique: Textos seleccionados das VII Jornadas da Língua Portuguesa*. Porto Editora, pp. 99–117.
Tardif, Maurice (2001) *Saberes docentes e formação profissional*. Vozes.
Tomlinson, Brian (1994) Pragmatic awareness activities. *Language Awareness* 3 (3/4), pp. 119–129. https://doi.org/10.1080/09658416.1994.9959850
Tomlinson, Brian (2001) Materials development. In Ronald Carter & David Nunan (eds.), *The Cambridge guide to TESOL*. Cambridge University Press, pp. 66–71.
Tomlinson, Brian (2010a) Helping learners to fill the gaps in their learning. In Freda Mishan & Angela Chambers (eds.), *Perspectives on language learning materials developmen*t. Peter Lang, pp. 87–108.
Tomlinson, Brian (2010b) Principles and procedures of material development. In Nigel Harwood (ed.), *Materials in ELT: Theory and practice*. Cambridge University Press, pp. 1–10.

Tomlinson, Brian (2011) *Material development in language teaching.* Cambridge University Press.
Tomlinson, Brian (2014) Looking out for English. *Folio* 16 (1), pp. 5–8.
Tomlinson, Brian (2016) The importance of materials development for language learning. In Maryam Azarnoosh, Mitra Zeraatpishe, Akram Faravani & Hamid Reza Kargozari (eds.), *Issues in materials development*. Sense Publishers, pp. 1–9.
Tomlinson, Brian & Hitomi Masuhara (2018) *The complete guide to the theory and practice of materials development for language learning*. John Wiley & Sons Inc.
Yu, Shulin Lianjiang Jiang & Nan Zhou (2023) The impact of L2 writing instructional approaches on student writing motivation and engagement. *Language Teaching Research* 27 (4), pp. 958–973. https://doi.org/10.1177/1362168820957024

Desenvolvimento de pronomes clíticos em produções escritas iniciais

Vasiliki Vraka[1]*, Maria Lobo*[2]*, Joana Batalha*[2]

[1]NOVA-FCSH
[2]NOVA-FCSH / CLUNL

Resumo

Este trabalho investiga o desenvolvimento na produção de pronomes clíticos nas fases iniciais da escrita de crianças do primeiro ciclo. A partir de um corpus de 272 textos narrativos produzidos por 136 crianças no início do 2.º ano de escolaridade e no início do 3.º ano de escolaridade, com base numa tarefa de escrita que integra o instrumento de diagnóstico do Projeto de Intervenção Preventiva para a Aprendizagem da Leitura e da Escrita (PIPALE), compara-se a produção de clíticos em cada um dos momentos. Conclui-se que há desenvolvimento da produção de clíticos entre os dois momentos, podendo a presença de clíticos, a sua distribuição e função, nas produções das crianças ser tomada como um indicador de desenvolvimento nas competências de escrita compositiva.

Palavras-chave: pronomes clíticos, escrita, português europeu, primeiro ciclo.

Abstract

This study investigates the development, in production, of clitic pronouns in the initial phases of writing by primary school children. Based on a corpus of 272 narrative texts produced by 136 children at the beginning of second grade and at the beginning of third grade, as part of a writing task which integrates the diagnostic instrument of the Preventive Intervention Project for Learning to Read and Write (PIPALE), the written production of clitics in each moment is compared. We conclude that there is development in clitic production between these two moments. The presence of clitics in the children's written productions, as well as its distribution and function, can, thus, be taken as a developmental marker of competencies of writing composition.

Keywords: clitic pronouns, writing, European Portuguese, primary school.

1. Introdução

Neste estudo, investigamos o desenvolvimento da produção de clíticos nas fases iniciais da escrita de crianças de primeiro ciclo.[1] Como tem sido investigado noutros trabalhos (Costa, 2010; Costa et al., 2017; Lobo et al., 2022), interessa-nos investigar em que medida as produções escritas das crianças refletem etapas características do desenvolvimento oral, especificamente quanto à produção de clíticos e aos padrões de colocação de clíticos. Será importante perceber: i) se há conformidade à gramática-alvo relativamente a este fenómeno sintático nas produções escritas das crianças; ii) caso não haja produção de clítico, que estratégias são usadas; iii) se há desenvolvimento no decorrer dos primeiros anos de escolaridade. Através de um estudo longitudinal que contempla textos produzidos por um grupo de crianças com um intervalo de cerca de um ano,

[1] Este trabalho resulta da dissertação de mestrado da primeira autora, Vasiliki Vraka (Vraka, 2022), realizada sob orientação de Maria Lobo e Joana Batalha. O artigo foi escrito conjuntamente pelas três autoras. Este trabalho foi parcialmente financiado por fundos nacionais através da FCT – Fundação para a Ciência e Tecnologia, I.P., no âmbito do projeto UIDB/LIN/03213/2020 e UIDP/LIN/03213/2020 – Centro de Linguística da Universidade NOVA de Lisboa (CLUNL).

procuramos determinar: i) se existe desenvolvimento na produção de clíticos entre os dois momentos, quer ao nível da quantidade e diversidade de clíticos produzidos, quer ao nível dos padrões de colocação de clíticos; e ii) se os padrões de desenvolvimento dos clíticos na escrita refletem padrões de desenvolvimento que se encontram também na oralidade.

Na Secção 1.1., revemos alguns dos estudos que investigaram o desenvolvimento de clíticos na aquisição do português e, na Secção 1.2., referimos trabalhos que se debruçaram sobre o desenvolvimento da escrita compositiva, e, em particular, sobre o desenvolvimento sintático na escrita. Na Secção 2.1., apresentamos a metodologia seguida no nosso estudo e, na Secção 2.2., os resultados obtidos. Na Secção 3, discutimos os resultados e apresentamos as principais conclusões do nosso trabalho.

1.1. **Desenvolvimento de clíticos na linguagem oral**

Os estudos sobre a aquisição da linguagem oral do português europeu têm mostrado que existe desenvolvimento quer ao nível da produção de clíticos (Costa & Lobo, 2006, 2007; Silva, 2008), quer ao nível dos padrões de colocação de clíticos (Costa et al., 2015; Vitorino & Lobo, 2018).

Os dados de produção espontânea mostram que as crianças produzem clíticos desde cedo, ainda que com uma fraca produtividade, sendo os clíticos menos especificados - *se* e *me* - os mais frequentes (Santos et al., 2014). Comparativamente com outras línguas, as crianças falantes do português europeu têm um desenvolvimento dos clíticos mais lento, omitindo clíticos em taxas mais elevadas e até idades mais tardias (Varlokosta et al., 2016). As taxas elevadas de omissão têm sido explicadas como sendo o resultado de uma generalização da construção de objeto nulo, possível no português europeu em determinados contextos sintáticos em que o objeto nulo é recuperável do contexto linguístico ou situacional. A hipótese de que a omissão de clíticos na aquisição do português corresponde a uma generalização da construção de objeto nulo é suportada por evidência de estudos de compreensão que mostram que as crianças aceitam leituras transitivas para construções com verbos sem complemento realizado (Costa & Lobo, 2008). A investigação também tem mostrado que as crianças têm taxas de omissão variáveis consoante o tipo de clítico (Silva, 2008): os clíticos reflexos estão entre os clíticos que têm menores taxas de omissão e os clíticos acusativos de terceira pessoa não reflexos estão entre aqueles que apresentam taxas mais elevadas de omissão. A omissão é tanto maior quanto mais facilmente o clítico é omitido na gramática adulta.

Também no que diz respeito à colocação de clíticos existe um desenvolvimento mais lento no português comparativamente com outras línguas. Enquanto na maioria das línguas os clíticos são colocados de acordo com a gramática-alvo desde muito cedo (Guasti, 1993), no português isso não acontece (Costa et al., 2015; Duarte et al., 1995). Em português europeu, ao contrário de outras línguas, os clíticos podem ocorrer em posição pós-verbal (ênclise), em posição pré-verbal (próclise) ou no meio do verbo (mesóclise) em função de fatores que não dependem da finitude da oração (Martins, 2013), como ilustrado em (1):

(1a) O avô barbeou-**se**. (ênclise)
(1b) O avô não **se** barbeou. (próclise)
(1c) O avô barbear-**se**-á. (mesóclise)

Costa et al. (2015) mostram que o desenvolvimento da colocação de clíticos é progressivo e é sensível à especificidade dos contextos sintáticos: o contexto em que a próclise é adquirida mais cedo é o contexto de negação, podendo estabelecer-se a seguinte escala de desenvolvimento a partir dos contextos considerados pelos autores:

(2) negação > sujeitos negativos / completivas finitas com conjuntivo > advérbio 'já' > orações adverbiais com 'porque' > sujeitos quantificados com 'todos'

Assim, o desenvolvimento mais lento da colocação de clíticos no português europeu é atribuível ao facto de a variação entre ênclise e próclise estar dependente de uma multiplicidade de fatores, requerendo desenvolvimento sintático e lexical.

O português também se distingue de outras línguas por permitir que, em determinadas estruturas com complexos verbais, um clítico selecionado por um verbo de uma estrutura não finita encaixada ocorra adjacente a um verbo finito de um domínio superior, fenómeno conhecido como "subida de clítico", ilustrado em (3b) (Gonçalves, 2002; Martins, 2013):

(3a) O avô vai barbear-**se**.
(3b) O avô vai-**se** barbear.

A subida de clítico é obrigatória quando a forma verbal não finita é um particípio (4), facultativa com alguns verbos (semi)auxiliares que selecionam o infinitivo e com alguns verbos de controlo (5) e geralmente rejeitada com outros verbos que selecionam infinitivo (6), ainda que se encontre alguma variação entre os falantes quanto à aceitabilidade da subida:

(4a) *Foi dado-lhe um prémio.
(4b) Foi-lhe dado um prémio.

(5a) O Presidente quer dar-lhe um prémio.
(5b) O Presidente quer-lhe dar um prémio.

(6a) O Presidente decidiu dar-lhe um prémio.
(6b) *O Presidente decidiu-lhe dar um prémio.

Estudos anteriores sobre o desenvolvimento da subida de clítico na aquisição do português europeu (Lobo & Vitorino, 2021; Vitorino, 2017) mostraram que as crianças produzem desde muito cedo construções com subida de clítico, mas que há desenvolvimento, que se prolonga até idade escolar, dos contextos em que a subida de clítico é possível. Em geral, em tarefas de produção induzida, as crianças produzem mais facilmente do que os adultos construções com subida de clítico, havendo, contudo, sensibilidade ao tipo de verbo e à presença de proclisadores.

1.2. **Desenvolvimento sintático na escrita**

Vários estudos têm procurado compreender até que ponto o desenvolvimento da linguagem escrita reflete o desenvolvimento da linguagem oral. A investigação tem mostrado que, nas fases iniciais da escrita compositiva, aproximadamente por volta do 3.º ano de escolaridade (Barbeiro & Pereira, 2007; Martins & Niza, 1998), as produções escritas das crianças revelam, em geral, uma complexidade sintática menor do que a que se encontra nas suas produções orais, o que é explicável pelo facto de a criança ainda não ter automatizado os processos mais básicos relativos às dimensões gráfica e ortográfica (Pinto et al., 2015).

No entanto, a relação entre oralidade e escrita pode não ser linear. Por um lado, quando as crianças começam a dominar padrões de escrita compositiva, podem não possuir ainda um conjunto suficientemente amplificado de estruturas sintáticas necessário para enfrentar os desafios que a escola coloca relativamente ao uso da linguagem escrita. Por outro lado, e embora se considere que o desenvolvimento da escrita dependerá, em larga medida, da estabilização do conhecimento linguístico da criança, vários autores (Costa et al., 2017; Pereira & Azevedo, 2005) têm notado que o conhecimento implícito de uma dada estrutura e o conhecimento linguístico consciente que dela se possui podem não ser sempre entendidos como uma condição prévia para o desenvolvimento da escrita, já que algumas estruturas menos frequentes e mais complexas surgem precisamente em géneros discursivos requeridos pela escolarização.

Ainda assim, do que sabemos, nomeadamente a partir de trabalhos que têm investigado processos de articulação de frases, a sequência de desenvolvimento sintático que se encontra na escrita acompanha, de modo geral, a sequência de desenvolvimento que se encontra na oralidade. Nas primeiras produções escritas, as crianças fazem um uso mais precoce de processos de coordenação do que de processos de subordinação, identificando-se uma escala de emergência de conetores próxima da que se encontra nas produções orais, como mostra, para o português, o trabalho de Costa et al. (2017). A partir da análise das estratégias de coesão interfrásica em textos de crianças de 2.º e 4.º anos de escolaridade, as autoras concluem que os conetores argumentativos mais usados em textos de opinião são os mesmos que as crianças usam precocemente, na oralidade. Também o trabalho de Lobo et al. (2022), que analisou os processos de articulação de orações em textos narrativos de crianças de 1.º ciclo produzidos em dois momentos distintos (2.º e 3.º anos de escolaridade), mostra que a coordenação é o processo predominante presente nos textos produzidos, sendo *e* e *mas* os conetores que mais vezes ocorrem nos textos de 2.º ano. Contudo, no 3.º ano, os conetores típicos de estruturas subordinadas tornam-se mais frequentes e também mais diversificados, aumentando de forma expressiva entre os dois momentos o uso de estruturas sintáticas complexas, como as orações relativas. Por sua vez, Costa e Gonçalves (2010), num estudo que contemplou textos narrativos e não narrativos de crianças entre o 3.º e o 6.º ano, observaram uma progressão de competências de escrita em ambos os tipos de texto, com um aumento da complexidade de estruturas sintáticas produzidas, em particular de orações subordinadas. Pereira e Azevedo (2005) referem que entre os oito e os dez anos, a proporção das orações subordinadas na oralidade e na escrita inverte-se: aos oito anos, as crianças usam mais a subordinação na oralidade e aos dez este processo é mais usado na escrita.

Relativamente aos clíticos, e embora a investigação sobre o uso destas estruturas na produção escrita seja bastante escassa, há alguma evidência (cf. Costa & Gonçalves, 2010) de que os clíticos são de desenvolvimento tardio nas produções escritas de crianças entre o 3.º e o 6.º anos, observando-se um aumento da produção nos anos mais avançados. Torna-se, pois, necessário perceber melhor em que momento emergem os clíticos na escrita das crianças e se estas estruturas podem ser tomadas como um indicador de desenvolvimento da escrita, caso o seu uso evidencie, à semelhança do que se tem observado com as estruturas de subordinação, uma marca da transição para uma dimensão compositiva da escrita.

2. **Estudo: desenvolvimento de clíticos em produções escrita**

Tendo em conta a investigação anterior sobre o desenvolvimento de clíticos na produção oral e sobre o desenvolvimento sintático nas produções escritas, pretende-se no presente estudo responder às seguintes questões:

i) Há desenvolvimento da produção de clíticos em fases iniciais da escrita? Este parâmetro pode ser tomado como indicador de desenvolvimento da escrita?

ii) Há conformidade à gramática-alvo, no que diz respeito à produção e à colocação, relativamente a este fenómeno sintático nas produções escritas das crianças?

iii) As estratégias alternativas à produção de clíticos e os desvios encontrados são semelhantes aos que encontramos na produção oral?

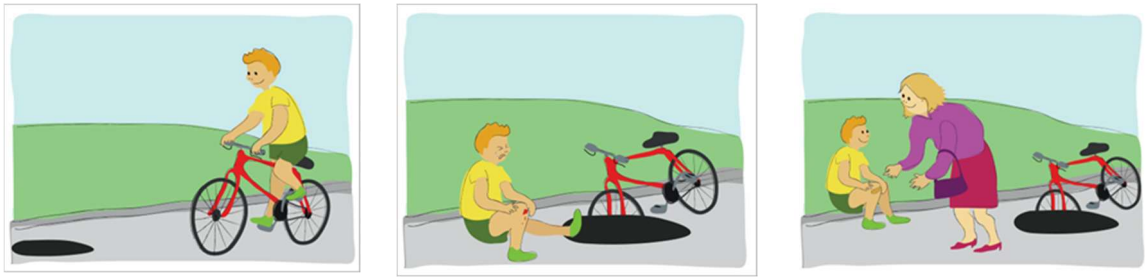
2.1. **Metodologia**

Usou-se um corpus constituído por 272 textos escritos produzidos como resposta a um item de escrita compositiva dos instrumentos de diagnóstico 1 e 2 do PIPALE – Projeto de Intervenção Preventiva para a Aprendizagem da Leitura e da Escrita. A tarefa foi realizada em contexto de sala de aula em dois momentos distintos: no início do 2.º ano de escolaridade e no início do 3.º ano de escolaridade. Os textos foram produzidos

a partir de uma sequência de três imagens que pressupunham uma situação problemática e a respetiva resolução (Ver Figuras 1 e 2). Foi fornecido material para a escrita dos textos, com um total de dez linhas e uma linha adicional para o título. A instrução dada às crianças, no material fornecido e seguida da sequência de imagens, foi a seguinte: "Observa as imagens. Escreve uma história a partir da sequência de imagens. Dá um título à tua história." Não foram dadas indicações sobre o limite de tempo para a escrita dos textos.

Figura 1. Sequência de imagens do Instrumento 1

Figura 2. Sequência de imagens do Instrumento 2



Para o presente estudo, foram consideradas as produções de 136 crianças, falantes de PE como língua materna, sem medidas seletivas ou adicionais de suporte à aprendizagem, e que participaram nos dois momentos de recolha. Foram excluídos os textos ilegíveis ou em branco. Depois de feita a transcrição do corpus, foi efetuado o levantamento de todas as ocorrências de pronomes clíticos em cada um dos momentos, tendo estas ocorrências sido posteriormente analisadas quanto a diferentes parâmetros: tipo de clítico, conformidade ao alvo, contexto sintático, posição.

2.2. **Resultados**

Descrevem-se, na Secção 2.2.1, os resultados da análise relativa à produção de clíticos e, na Secção 2.2.2., os resultados relativos à colocação de clíticos.

2.2.1. **Produção de clíticos**

A análise quantitativa dos clíticos produzidos em cada um dos momentos, reportada na Tabela 1, permite verificar que há um aumento muito considerável quer do número de clíticos globalmente produzidos (59 no primeiro momento vs. 190 no segundo momento), quer do número de crianças que produzem pelo menos um clítico (36 no primeiro momento vs. 100 no segundo momento). A média de clíticos produzidos em cada

momento confirma esse aumento, encontrando-se em média 0,44, no primeiro momento, e 1,40 no segundo momento.

Tabela 1. Dados globais de produção de clíticos nos dois momentos, considerando-se todas as ocorrências de clíticos

	N.º de crianças que produziram clíticos	**N.º global de clíticos produzidos**	**N.º mínimo de clíticos produzidos**	**N.º máximo de clíticos produzidos**	**Média de clíticos produzidos**	**Desvio padrão**
Instrumento 1	36/136 26,6%	59	0	4	0,44	0,87
Instrumento 2	100/136 74%	190	0	6	1,40	1,25

Para se comparar o número médio de clíticos produzidos nos dois momentos de testagem, procedeu-se à realização de um teste estatístico paramétrico, o teste t com amostras emparelhadas, já que se observou uma distribuição normal. Os valores obtidos ($t = 8,67$, $df = 135$, p-value < 0.0001) confirmam as diferenças estaticamente significativas entre as médias de clíticos produzidos em cada um dos momentos.

Quando consideramos a produção ou ausência de produção de clíticos em cada um dos instrumentos, verificamos (Tabela 2) que a maioria das crianças da amostra só produziu clíticos no instrumento 2.

Tabela 2. Comparação entre produção de clíticos nos instrumentos 1 e 2

Crianças que nunca produziram clíticos	**Crianças que produziram clíticos apenas no Instrumento 1**	**Crianças que produziram clíticos em ambos os instrumentos**	**Crianças que produziram clíticos apenas no Instrumento 2**
31/136	5/136	31/136	69/136
22,8%	3,7%	22,8%	50,7%

Estes resultados mostram que, ainda que haja variação entre as crianças na produção de clíticos, é claramente observável um aumento nas taxas de produção no segundo instrumento, que é revelador de um desenvolvimento da complexidade sintática da escrita.

Os clíticos distribuem-se pelas categorias de clíticos reflexos, acusativos e dativos, como se observa nos exemplos seguintes do Instrumento 1 (7) e do Instrumento 2 (8):[2]

(7a) e ela não **se** escesseu do aniverssariu do seu irmão Pedro (A0549)
(7b) O Gabriel largo o balão mas antes de ele ir para o ceu a mãe agarrou-**o**. (A0565)
(7c) e de pois a Mãe do João deo-**lhe** um balão (A0385)

[2] Mantemos, nos exemplos, a escrita das crianças não corrigindo os erros de ortografia e de pontuação. Apenas destacamos o clítico ou o pronome a negrito. Entre parênteses indicamos o código do participante.

(8a) depois o João aleijo**se** no joelho (A0391)
(8b) A Maria vio o Afonso maguado e foi a correr para **o** ajudar (A0395)
(8c) e a senhora meteu-**lhe** um penso e ele parou de churar (A0460)

A distribuição dos clíticos por cada tipo encontra-se na Tabela 3.

Tabela 3. Distribuição por tipo de clítico

	Instrumento 1	**Instrumento 2**
Reflexo	13	94
Acusativo	36	57
Dativo	10	33

Como se pode observar, há um predomínio de clíticos acusativos não reflexos no Instrumento 1, passando os clíticos reflexos a ser maioritários no Instrumento 2. Esta distribuição é, em parte, condicionada pelas imagens que servem de estímulo à produção do texto escrito, com usos frequentes de *magoar-se* ou *aleijar-se*, não sendo claro que resultem de desenvolvimento dos diferentes contextos. Observamos ainda que os clíticos encontrados são maioritariamente produzidos de acordo com a forma alvo. Registam-se alguns usos de pronome forte em vez de clítico (9) – 7 ocorrências de pronome forte em posição de complemento direto na totalidade do *corpus* – e alguns casos de 'lheísmo' (cf. (10)) – 6 casos de 'lheísmo' na totalidade do corpus –, maioritariamente no Instrumento 2, 2 casos de alteração morfofonológica, como a produção de *lo* em vez de *o* (11), e ainda casos de omissão do argumento (12): [3]

(9a) Ele caiu e vaio uma descoisida ajudou **ele**. (A0536)
(9b) Uma sinhora muito simpatica viu **ele** e pôs-lhe um penso. (A0459)

(10a) Uma senhora que passeava no campo foi **lhe** ajudar o gustavo. (A0396)
(10b) A mãe dele chegou e acalmou-**lhe** e pos-lhe um penço. (A0427)
(10c) ...a tia foi lá e ajudou-**lhe** a corar-se e ficou melhor e a tia disse-lhe:… (A0462)

(11a) a cinhora diçe olá e ajudo**lo** posle um penço na perna (A0446)
(11b) Mas um dia ele caio e a sua avó foi**lo** ajudar (A0575)

(12) Era uma vez um menino que chamavase João a sua mãe deu **[ ]** um balão (A0459)

Quando não há produção de clíticos, encontram-se predominantemente duas estratégias: i) a omissão de clíticos (13); ii) a produção de DP (14):

(13) ...era uma vez um menino chamado João, estava a aprender a andar de bicicleta, ele estava a comceguir e um buraco pequeno estava à frente e o joão caiu no chão e a ferida estava a deitar sangue e o vizinho veio e rápida-mente o vizinho pos **[ ]** um penso fim. (A0554)

(14) O João amdava na biciquenta e a senhora alice estava setada no baco e o João estava am dar de biciquénta e depois caiu e a sinhora alice ajudou **o João** e depois de ajudar **o João** de-se obrigado **o joão**. (A0533)

Esta segunda opção é largamente predominante no Instrumento 1 (62 textos), reduzindo-se de forma acentuada no Instrumento 2 (10 textos). Isto sugere dificuldades no uso de mecanismos de coesão referencial e

[3] Há, naturalmente, ainda dificuldades no domínio das regras ortográficas relativas à hifenização de clíticos, encontrando-se clíticos amalgamados com a forma verbal, clíticos em ênclise sem hífen, entre outros problemas:
i) Como ia ali ao pé uma senhora ela foi **ajudálo** a **levantarse** e a tratar da frida (A0383)
ii) mas o pedro a paiou o balão e **deu o** ao avô. (A0536)

no estabelecimento de cadeias referenciais interfrásicas associadas à fase emergente de produção escrita, em que as crianças constroem textos recorrendo ainda a processos de justaposição de frases simples, não ligadas entre si, como se pode observar no exemplo em (15):

(15) A mãe do João copor um balão azul um dia o João largou o balão e o balão voa no dia seguimte a mãe copor outro balão azul e o João numca largou o balão. (A0381)

No segundo momento, há claramente um desenvolvimento dos processos de construção textual, que se manifesta também no maior recurso aos clíticos como estratégia de manutenção da coesão referencial. Este fenómeno não se verifica nos mesmos moldes na produção oral, em que a estratégia de evitação predominante na aquisição do português é a omissão.

2.2.2. **Colocação de clíticos**

Quando consideramos o parâmetro colocação de clítico, verificamos que, em domínios verbais simples com formas verbais finitas, há maioritariamente colocação-alvo do clítico, quer em contextos de ênclise, quer em contextos de próclise, com casos residuais de redobro (16a), de ênclise em contexto de próclise (16b), mas também de próclise em contexto de ênclise (16c). Não se registaram no corpus contextos de mesóclise.

(16a) (...) e caiu na coisa estranha e **se** aleijou**se** (…) (A0523)
(16b) Era uma vez um menino que chamava**se** João a sua mãe deu um balão (A0459)
(16c) A Mãe do Afonso foi á loja e **lhe** compou um balão. (A0391)

As taxas de colocação-alvo em cada contexto são apresentadas na Tabela 4:

Tabela 4. Colocação-alvo de clítico em domínios verbais simples com formas verbais finitas

	Instrumento 1		**Instrumento 2**		**Total**	
Contextos de ênclise	44/48[a]	91,7%	93/107[b]	86,9%	137/155	88,4%
Contextos de próclise	6/7[c]	85,7%	28/31[d]	90,3%	34/38	89,5%

[a] Nas ocorrências não alvo, 3 foram casos de próclise e 1 foi um caso de redobro.
[b] Nas ocorrências não alvo, 13 foram casos de próclise e 1 foi um caso de redobro.
[c] A única ocorrência não alvo foi um caso de ênclise.
[d] Nas ocorrências não alvo, 2 foram casos de ênclise e 1 foi um caso de redobro.

Em orações infinitivas com verbo simples, registaram-se 11 ocorrências, todas elas no Instrumento 2. Quando se trata de uma oração adverbial infinitiva, introduzida por preposição (4 ocorrências), o clítico ocorre sempre em próclise:

(17a) A Maria vio o Afonso maguado e foi a correr para **o** ajudar (A0395)
(17b) Passeou passeou passeou, até **se** fartar (A0503)

As restantes 7 ocorrências são com a construção *ajudar x a + infinitivo*. Nesta construção, um clítico selecionado pelo verbo infinitivo ocorre em ênclise ao verbo infinitivo (3/7) ou em próclise ao verbo infinitivo seguindo a preposição *a* (4/7):

(18a) viu o Miguel no chão e foi ajudar o Miguel a levantar**se** do chão (A0407)
(18b) mas um simpatica senhora foila e ajudou ele a **se** levantar (A0522)

Considerando agora os clíticos produzidos em complexos verbais, verificamos que estes contextos não são em número muito expressivo, encontrando-se no conjunto dos dois instrumentos 43 ocorrências de clíticos

em complexos verbais (3 no Instrumento 1 e 40 no Instrumento 2). Encontram-se diferentes tipos de contextos com diferentes verbos (semi)auxiliares e alguns verbos de controlo: *ter* + particípio passado (6 ocorrências); *ir* + infinitivo (22 ocorrências); *vir* + infinitivo (6 ocorrências); *poder* + infinitivo (1 ocorrência); *começar a* + infinitivo (1 ocorrência); *ter de/que* + infinitivo (2 ocorrências); *tentar* + infinitivo (3 ocorrências); *conseguir* + infinitivo (1 ocorrência); *lembrar-se de* + infinitivo (1 ocorrência).

Os contextos são diversificados, apresentando variação no tipo de verbo, no tipo de clítico e na presença/ausência de proclisador. Apesar desta diversidade, que dificulta a identificação de padrões, podemos observar que existe variação entre subida e não subida de clítico. Nas ocorrências com o verbo semiauxiliar *ir*, um pouco mais numerosas, encontram-se 12 casos de ênclise ao verbo infinitivo (19), 9 casos de ocorrência entre verbo auxiliar e verbo infinitivo, geralmente identificáveis como ênclise ao primeiro verbo (20) e uma única ocorrência de subida de clítico em próclise ao verbo auxiliar (21):

(19) estava a Carolina opé e vio o Rui aleijado e foi ajoda**lo** (A0397)

(20) Mas um dia ele caio e a sua avó foi**lo** ajudar (A0575)

(21) estava a daita muito sangue e a mãe **o** foi ajodar a se corar (A0485)

Podemos também observar que, sempre que a oração contém um proclisador (que pode ser um complementador ou uma preposição), existe subida de clítico com próclise ao primeiro verbo:

(22a) E estava a chorar tão alto **que** toda a gente **o** conseguiu ouvir (A0559)
(22b) A avó podes ligar á mãe **para me** vir boscar. (A0513)

Nos complexos verbais com particípio passado, que na variedade europeia do português não permitem cliticização ao particípio passado, encontramos apenas um caso desviante (23a), estando os restantes casos de acordo com o padrão-alvo (23b-c):

(23a) A edosa ajudou-lhe a João disse obrigado minha selhora porter ajudado-**me**. (A0449)
(23b) Ele tinha-**se** aleijado no joelho e começou a chorar. (A0501)
(23c) o menino disse que estava tudo bem que o senhor **lhe** tinha dado um penso (A0566)

3. **Discussão**

Retomando as questões de investigação deste trabalho, discutimos nesta secção os resultados obtidos. A nossa primeira questão procura determinar se há desenvolvimento da produção de clíticos e se este parâmetro pode ser tomado como indicador de desenvolvimento da escrita em fases iniciais da escrita compositiva. Como vimos, o nosso estudo permitiu identificar um desenvolvimento da produção de clíticos em textos narrativos entre o início do 2.º ano de escolaridade e o início do 3.º ano de escolaridade, com um aumento significativo quer do número global de clíticos produzidos, quer do número de crianças que produzem clíticos. Estes resultados mostram que este pode ser um indicador relevante de desenvolvimento das competências da escrita compositiva, correspondendo a uma escrita mais amadurecida, com maiores níveis de complexidade sintática, na linha do que sugerem Costa e Gonçalves (2010), e à semelhança do que tem vindo a ser encontrado relativamente a outras estruturas, como conetores típicos de estruturas subordinadas (Costa et al., 2017; Lobo et al., 2022). Contudo, reconhece-se que o aumento da produção de clíticos do primeiro para o segundo momento avaliado poderá não ser imputável apenas ao efeito do desenvolvimento, mas também à natureza das

representações visuais usadas na tarefa para a elicitação da produção das narrativas, aspeto que poderá ser tido em conta em trabalho futuro.

A segunda questão pretende verificar se há conformidade à gramática-alvo relativamente a este fenómeno sintático nas produções escritas das crianças. Globalmente, verificamos que as produções de clíticos pelas crianças são maioritariamente conformes à gramática-alvo. Encontram-se, contudo, algumas estratégias não canónicas, incluindo a produção de pronomes fortes em posição de complemento direto, casos de 'lheísmo' e alguns casos de alteração na forma morfofonológica do clítico (produção de *lo* em vez de *o*, por exemplo). No que diz respeito à colocação, os padrões seguem maioritariamente os padrões esperados na variedade europeia do português. Contudo, também relativamente a este parâmetro podemos encontrar algumas produções desviantes, incluindo casos de cliticização a formas não esperadas (por exemplo, particípio passado), casos esporádicos de redobro, ênclise em contexto de próclise e próclise em contexto de ênclise.

Procurámos também determinar até que ponto as estratégias alternativas à produção de clíticos e os desvios encontrados são semelhantes aos que encontramos na produção oral. Nos dados de que dispomos sobre o desenvolvimento de clíticos na oralidade para o português europeu, a estratégia alternativa à produção de clíticos predominante é a omissão, explicável como um caso de sobregeneralização da construção de objeto nulo (Costa & Lobo, 2007, 2009). Nos nossos dados de produções escritas, em contrapartida, não encontramos um uso predominante dessa estratégia. Em vez disso, encontramos um uso predominante de retomas por DP no primeiro momento, com um desenvolvimento claro, no segundo momento, dos processos de construção textual, que se manifesta também no maior recurso aos clíticos como estratégia de manutenção da coesão referencial. Isto vai ao encontro de investigação anterior, que tem mostrado desenvolvimento nestes mecanismos de coesão referencial (Batoréo & Costa, 1997) e um domínio gradual da escala de acessibilidade (Ariel, 1996; Flores et al., 2020).

No que diz respeito à colocação, a investigação sobre o desenvolvimento da linguagem oral mostrou que há colocação alvo dos clíticos em contexto de ênclise e que há um desenvolvimento gradual dos contextos de próclise, com tendência para a generalização da ênclise (Costa et al., 2015). Nas produções escritas que analisámos, não se encontra um padrão de generalização de ênclise em contexto de próclise como foi encontrado na produção oral em estudos anteriores. Encontramos maioritariamente colocação alvo dos clíticos, com desvios ocasionais quer em contextos de próclise, quer em contextos de ênclise. Ainda que, por se tratar de dados em que os diferentes contextos não são controlados de forma criteriosa, não possamos ter dados mais robustos e conclusivos sobre os padrões de colocação, estes dados apontam para uma diferença entre as duas modalidades. Será necessário, de futuro, explorar mais detalhadamente estas diferenças, que podem eventualmente ser atribuídas a diferentes fatores, entre os quais podem estar a tomada de consciência de que existem diferentes posições possíveis para os clíticos em português através do confronto com a exposição a textos escritos.

Nos contextos com complexos verbais, Vitorino (2017) e Lobo e Vitorino (2021) mostram que, na produção oral, as crianças adquirem cedo o fenómeno de subida de clítico e que mostram alguma preferência pela subida em detrimento da não subida, havendo, contudo, desenvolvimento dos contextos em que a subida é permitida. No nosso corpus, ainda que as ocorrências de clíticos em complexos verbais não sejam numerosas, verificamos oscilação entre subida e não subida de clítico, tal como esperado na gramática-alvo, sem que haja um padrão claro de preferência pela subida. Em trabalho futuro, importará investigar de forma mais aprofundada eventuais diferenças entre modalidade oral e escrita relativamente aos padrões de colocação de clíticos.

4. **Conclusões**

Tendo em conta a investigação anterior sobre o desenvolvimento de clíticos na produção oral e sobre o desenvolvimento sintático nas produções escritas, o presente estudo pretendeu investigar o desenvolvimento de pronomes clíticos nas fases iniciais da escrita de crianças do primeiro ciclo, a partir de um corpus de 272 textos narrativos produzidos por 136 crianças no início do 2.º ano de escolaridade e no início do 3.º ano de escolaridade.

Os dados obtidos evidenciam que há desenvolvimento na produção de clíticos por crianças falantes de PE em textos narrativos escritos nos anos iniciais de escolaridade. Globalmente, concluímos que a presença de clíticos nas produções das crianças pode ser tomada como um indicador de desenvolvimento nas competências de escrita compositiva, à semelhança do que tem vindo a ser encontrado relativamente a outras estruturas, como conetores típicos de estruturas subordinadas. Apesar de os dados sugerirem que a produção de clíticos na escrita está sujeita a desenvolvimento, importa perceber melhor até que ponto esse desenvolvimento reflete etapas do desenvolvimento oral (por exemplo quanto à estratégia de evitação de clítico e ao tipo de desvios na colocação do clítico), uma vez que se encontram diferenças nas estratégias predominantes.

Apesar destas limitações, cremos que o estudo poderá fornecer um contributo importante para a caracterização do desenvolvimento sintático na escrita, com implicações também ao nível das práticas de ensino e aprendizagem da escrita, nomeadamente ao nível do uso de mecanismos de coesão que possam ser usados pelas crianças, com uma progressiva consciencialização, na produção dos textos.

Referências

Ariel, Mira (1996) Referring expressions and the +/- coreference distinction. In Jeanette Gundel & Thorstein Fretheim (eds.), *Referent and referent accessibility*. John Benjamins, pp. 13–35.

Barbeiro, Luís & Luísa Álvares Pereira (2007). *O ensino da escrita: A dimensão textual*. Ministério da Educação & PNEP.

Batoréo, Hanna & Maria Armanda Costa (1997) Referência nominal na narrativa oral e escrita aos dez anos de idade. In *Atas do XIII Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 137–149.

Costa, Ana Luísa (2010) *Estruturas contrastivas: Desenvolvimento do conhecimento explícito e da competência de escrita*. Dissertação de doutoramento, Faculdade de Letras, Universidade de Lisboa.

Costa, Ana Luísa, Sónia Cerqueira & Vanessa Carreto (2017) 'E essa é a minha opinião': Para o estudo da emergência da escrita argumentativa. *Revista da Associação Portuguesa de Linguística* (3), pp. 51–73. https://doi.org/10.26334/2183-9077/rapln3ano2017a24

Costa, Armanda & Anabela Gonçalves (2010) Progressão e complexidade na escrita do 3.º ao 6.º ano de escolaridade. In Armanda Costa, Sofia Vasconcelos & Vitória de Sousa (eds.), *Muitas ideias, um mar de palavras. Propostas para o ensino da escrita*. Fundação Calouste Gulbenkian, pp. 283–318.

Costa, João & Maria Lobo (2006) A aquisição de clíticos em PE: Omissão de clíticos ou objecto nulo? In *Textos selecionados do XXI Encontro Nacional da Associação Portuguesa de Linguística.* APL, pp. 285–293.

Costa, João & Maria Lobo (2007) Clitic omission, null objects or both in the acquisition of European Portuguese? In Sergio Baauw, Frank Drijkoningen e Manuela Pinto (eds.) *Romance languages and linguistic theory 2005.* John Benjamins, pp. 59–71.

Costa, João & Maria Lobo (2008) Omissão de clíticos na aquisição do português europeu: dados da compreensão. In *Textos selecionados do XXIII Encontro Nacional da Associação Portuguesa de Linguística*. APL, pp. 143–156.

Costa, João & Maria Lobo (2009) Clitic omission in the acquisition of European Portuguese: Data from comprehension. In Acrísio Pires & Jason Rothman (eds.), *Minimalist inquiries into child and adult language acquisition: Case studies across Portuguese*. Mouton de Gruyter, pp. 63–84.

Costa, João, Alexandra Fiéis & Maria Lobo (2015) Input variability and late acquisition: Clitic misplacement in European Portuguese. *Lingua 161*, pp. 10–26. https://doi.org/10.1016/j.lingua.2014.05.009

Duarte, Inês, Gabriela Matos & Isabel Hub Faria (1995) Specificity of European Portuguese clitics in Romance. In Isabel Hub Faria & Maria João Freitas (eds.), *Studies on the acquisition of Portuguese*. APL & Colibri, pp. 129–154.

Flores, Cristina, Esther Rinke & Aldona Sopata (2020) Acquiring the distribution of null and overt direct objects in European Portuguese. *Journal of Portuguese Linguistics* 19 (1). https://doi.org/10.5334/jpl.239

Gonçalves, Anabela (2002) Verbos auxiliares e verbos de reestruturação do português europeu. In Isabel Duarte, Joaquim Barbosa, Sérgio Matos & Thomas Hüsgen (eds.), *Actas do Encontro Comemorativo dos 25 anos de Centro de Linguística da Universidade do Porto* (Vol. 2). Centro de Linguística da Universidade do Porto, pp. 45–57.

Guasti, Maria-Teresa (1993) Verb syntax in Italian child grammar: Finite and non-finite verbs. *Language Acquisition*, 3, pp. 1–40. https://doi.org/10.1207/s15327817la0301_1

Lobo, Maria & Inês Vitorino (2021) Acquisition of clitic climbing by European Portuguese children. In Larisa Avram, Anca Sevcenco & Veronica Tomescu (eds.), *L1 acquisition and L2 learning: The view from Romance*. John Benjamins, pp. 13–38.

Lobo, Maria, Joana Batalha, Antónia Estrela & Bruna Bragança (2022). Desenvolvimento sintático em produções escritas de crianças de 1.º ciclo. *Revista da Associação Portuguesa de Linguística* (9), pp. 150–163. https://doi.org/10.26334/2183-9077/rapln9ano2022a11

Martins, Ana Maria (2013) Posição dos pronomes pessoais clíticos. In Eduardo Raposo, Maria F. B. Nascimento, Maria A. C. Mota, Luísa Segura & Amália Mendes (orgs.), *Gramática do Português* (Vol. 2). Fundação Calouste Gulbenkian, pp. 2231–2304.

Martins, Margarida & Ivone Niza (1998) *Psicologia da aprendizagem da linguagem escrita*. Universidade Aberta.

Pereira, Luísa Álvares & Flora Azevedo (2005) *Como abordar… a escrita no 1.º ciclo do ensino básico*. Areal Editores.

Pinto, Giuliana, Christian Tarchi & Lucia Bigozzi (2015) The relationship between oral and written narratives: A three-year longitudinal study of narrative cohesion, coherence, and structure. *British Journal of Educational Psychology* 85 (4), pp. 551–569. https://doi.org/10.1111/bjep.12091

Silva, Carolina (2008) *Assimetrias na aquisição de clíticos diferenciados em Português Europeu*. Dissertação de Mestrado, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa.

Varlokosta, Spyridoula et al. (2016) A cross-linguistic study of the acquisition of clitic and pronoun production. *Language Acquisition* 23 (1), pp. 1–26.

Vitorino, Inês (2017) *Aquisição de estruturas com subida de clítico em português europeu*. Dissertação de Mestrado, FCSH-UNL.

Vitorino, Inês & Maria Lobo (2018) Aquisição de estruturas com subida de clítico em português europeu. *Revista da Associação Portuguesa de Linguística* (4), pp. 276–294. https://doi.org/10.26334/2183-9077/rapln4ano2018a45

Vraka, Vasiliki (2022) *Desenvolvimento do uso de pronomes clíticos na escrita de crianças do 2º ano de escolaridade*. Dissertação de Mestrado, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa.