«Repartição» e «perfil das palavras»: a questão da presença/ausência nos estudos de vocabulário[1]

Carlos MACIEL
(Université de Nantes)

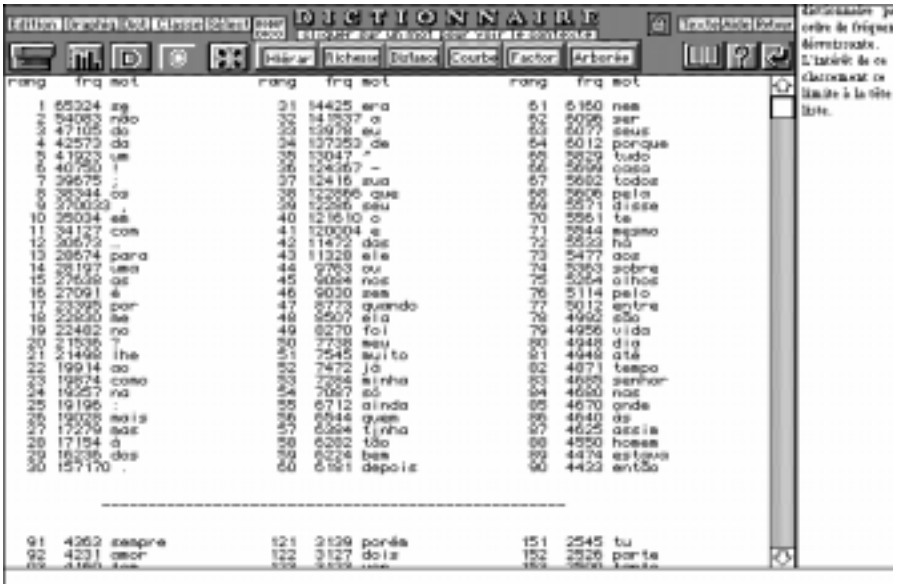
1. O *corpus*

As pesquisas feitas na área da linguística sempre consideraram, até aqui, pelo menos quando falamos dos estudos aplicados à área da lexicologia e aos corpus de grande extensão, a presença das diferentes formas (unidades de texto) e/ou dos vocábulos (unidades de vocabulário) nos textos analisados, partindo-se mais ou menos do princípio que a presença das formas mais vigorosas do sistema – ou mais «rentáveis» – estava de uma certa maneira garantida em todo e qualquer corpus, respeitada a sua extensão.

Este princípio norteou em grande parte o estudo (ensino/aprendizagem) das línguas, que obedeceu às regras daquilo a que se deu o nome de «fundamental» ou ainda, mais recentemente, de «nível limiar», por exemplo. O que se pretende aqui, em termos de uma contribuição ao conhecimento de certos fenômenos linguísticos respeitantes à distribuição, ou à relação presença/ausência de uma palavra num texto, é trazer para este debate um enfoque novo, que vai talvez surpreender alguns colegas linguistas, particularmente aqueles que se ocupam da didática das línguas e/ou do ensino-aprendizagem das línguas em geral. Também o estudioso e/ou professor de literatura brasileira poderá encontrar aqui alguns subsídios capazes de alimentar a reflexão sobre os conteúdos linguísticos, e mais propriamente lexicais, de textos literários que, diga-se de passagem, são aqueles que muitas vezes propomos como leitura aos nossos estudantes.

A experiência de que aqui se trata resulta de um trabalho feito a partir dos dados de um conjunto de 81 textos de literatura brasileira extraídos da base de dados textuais PORTEXT[2] . Este conjunto, com 4.620.146 ocorrências (ou unidades de texto) para 108.329 formas diferentes, cobre quatro séculos de literatura brasileira; todos os principais autores estão nela representados[3] , assim como todos os principais gêneros (teatro, romance, poesia…).

Quadro I: as palavras mais frequentes



O corpus assim constituído foi em primeiro lugar tratado pelo programa Hyperbase (de Étienne BRUNET – Quadro I)[4], que nos a forneceu o dicionário geral e o dicionário de frequências. A observação dos resultados obtidos fez com que fossem rapidamente postos em evidência fatos cuja importância linguística é considerável. Observemos por exemplo que o programa Hyperbase – segundo critérios já bem conhecidos pelos linguistas e de que não falaremos aqui – propõe-nos uma lista de 162 formas, classificadas segundo a ordem decrescente das frequências.[5] Trata-se neste caso das 162 "palavras" mais frequentes no corpus de referência. Não temos todavia até este momento nenhuma informação sobre a repartição destas formas nos diferentes textos que integram o corpus (embora Hyperbase forneça informações sobre a distribuição, com dados brutos gerais – e os utilizadores do programa conhecem bem esta questão). Para além no entanto do quadro de distribuição das palavras – com as suas sub-frequências nos 81 textos – nós tentaremos aqui analisar as diferenças que surgem em matéria de repartição. O quadro das repartições, como veremos a seguir, leva em consideração, para cada linha[6], a presença ou a ausência da forma considerada em cada uma das 81 colunas[7].

2. A "repartição"

A palavra "repartição" é em consequência aqui utilizada para designar essencialmente a oposição presença/ausência de uma qualquer forma num texto ou num conjunto dado de textos. Isto fez com que viéssemos a estudar quadros de contingências de grande extensão e, num primeiro enfoque, a simples presença (ou ausência) de uma forma dada num sub-conjunto qualquer constituído com relação ao resto dos elementos que integram o corpus pode ser significativa. Esta

repartição pode confundir-se, no que diz respeito a certas aplicações, com a noção já muito conhecida de exclusividade lexical e faz parte, deste ponto de vista, do vasto campo da distribuição; mas, como veremos mais adiante, ela tem as suas particularidades claramente afirmadas.

2.1. Procurando os limites — primeiros passos para a construção de um modelo

As primeiras questões levantadas dizem respeito aos limites impostos ao sistema ou ainda às fraturas que podem aparecer no quadro das repartições, se levarmos em conta todas as formas presentes em todos os textos, distribuídas por patamares sucessivos – a partir do valor *n* (=81). O nosso ponto de partida era simples: se restringíssemos aos poucos o número de textos, levando em consideração unicamente as presenças (*n* -10, *n* - 20, *n* - *x*), um desvio deveria num momento dado aparecer fazendo com que a curva apresentasse em algum lugar uma irregularidade. As diferenças de gênero e, particularmente, as diferenças de extensão entre os textos podiam com efeito por si só justificar esta expectativa. Antes de apresentar alguns resultados das pesquisas, indicaremos a seguir os procedimentos adotados.

2.2. Dados de base e programação informática modular

O quadro II foi extraído do fichário relativo à distribuição das formas nos 81 textos. Por exemplo, a palavra *bacalhau* tem uma frequência total igual a 23 no corpus, frequência 2 no texto n° 1, frequência 1 no n° 47, ..., de 6 no n° 56, etc... O quadro de presença/ausência, cuja extensão é muito grande, é composto de 108.329 sequências ou cadeias de caracteres que estão individualmente associadas a 81 valores lógicos 0/1. Ao numerizar estas sequências e ao atribuir a cada uma delas um código que corresponde ao valor lógico representado por um único bit, nós enviamos a informação diretamente para a memória central do computador e constituímos assim o estoque dos dados de base. A cada linha da repartição corresponde uma leitura estatística deste quadro. O tempo de resposta é relativamente curto quando este conjunto de dados é processado na memória central do computador.

Quadro II: o fichário de distribuição das formas nos textos

```
.....
1 bacabau 76 1
23 bacalhau 1 2 47 1 48 1 53 2 56 6 60 1 69 1 76 5
 77 1 80 2 81 1
1 bacalhoadas 60 1
88 bacamarte 1 1 12 1 42 60 43 13 44 1 48 1 57 5
 66 2 78 3 81 1
.....
```

A partir dos dados do quadro, tínhamos a possibilidade de proceder de diferentes maneiras. Com muita frequência escreve-se um programa informático para cada questão levantada – que

será então modificada para que se possa tratar outras questões, até mesmo bastante próximas. O melhor método mas também o mais difícil consiste em elaborar um único programa que permita fazer todas as diferentes e numerosas interrogações que este gênero de pesquisa supõe. Este programa, cuja elaboração exige um trabalho considerável, só é "rentável" se for destinado a um uso corrente.

Optamos aqui por uma posição intermediária que designaremos pela expressão "método de programação modular". A linguagem utilizada é Java 2.1 de *Sun Microsystems.* O programa em referência é composto de uma plataforma central e de vários módulos independentes. Cada módulo pode ser executado a partir da plataforma e, estando terminado o trabalho de execução, os resultados aparecem, são fixados numa área e volta-se então para a plataforma. Alguns exemplos de módulos: carregar os dados a partir de um fichário, corrigir ou modificar dados, salvar os resultados num repertório, selecionar um conjunto dado A, selecionar um conjunto dado B segundo diversos critérios, e, enfim, fazer as listas do vocabulário comum dos textos A e B e do vocabulário de A \ B. Cada vez que se quiser fazer uma pergunta específica a construção de um módulo particular é necessária: é o *processamento.* Este módulo, integrado ao seu espaço, vai exigir uma codificação mínima e poucas manipulações informáticas.

3. Algumas perguntas

Quais são as formas que estão presentes em todos os textos (n=81)? Quantas formas estão presentes em n-x textos?

A observação da lista geral (Quadro III) – de presença nos 81 textos do corpus – mostra que 93 formas somente (sobre 108.329) figuram em todos os diferentes textos, enquanto que – graças a todas as experiências precedentes, particularmente as que, há já bastante tempo, puseram em destaque o "fundamental" – podíamos legitimamente esperar encontrar bem mais!

Se pusermos as duas listas uma ao lado da outra (o quadro I, que diz respeito à lista decrescente das frequências fornecida pelo programa Hyperbase, e o quadro III, com a lista das 93 formas comuns a todos os textos), veremos que, respeitado o limite de 93 (número igual de formas para cada lista), encontramos 75 formas comuns às duas listas; em outras palavras, ao observar as 93 formas mais frequentes do corpus, constatamos que 18 só aparecem numa das duas listas. Vamos no entanto encontrar 89 formas comuns às duas listas se levarmos em consideração o limite das 162 formas mais frequentes processadas pelo programa Hyperbase.

Um número bastante grande de formas parece assim, nos dois casos, escapar à nossa intuição que faz com que a noção de frequência muito elevada acabe por necessariamente supor, em termos de repartição, uma presença em todos os textos: em realidade, uma frequência muito elevada – mesmo quando se trata de certos "instrumentais" da língua – não é uma garantia de presença; a repartição dá-nos assim os seus primeiros resultados palpáveis e permite que seja levantada a questão relativa a uma tipologia a ser construída em função deste único critério.

Quadro III: as formas comuns aos 81 textos

a	dentro	grande	nunca	seus
à	depois	há	o	só
agora	dia	horas	os	também
ainda	dias	já	ou	tanto
antes	do	a	outros	tão
ao	dos	lá	para	tem
aos	e	lhe	pela	tempo
aquele	é	longe	por	ter
as	ele	mais	porque	tinha
às	em	mas	qual	todo
assim	enquanto	me	quando	todos
até	então	mesmo	quanto	tudo
bem	entre	mundo	que	um
certo	era	na	quem	uma
com	essa	não	são	vai
como	esta	nas	se	vem
da	este	nem	sem	vez
das	fazer	no	ser	
De	foi	nos	seu	

Entre as 18 formas que, considerado o limite de 93, só aparecem numa das duas listas, encontramos duas formas verbais (*vai* e *vem*) e um substantivo plural (*horas*), que marcam assim de imediato o seu carácter específico.

Se formos menos exigentes com relação ao critério adotado (*n*-1), encontraremos 147 formas comuns a 80/81 textos; a *n*-11, encontraremos 522 formas (comuns a 70 textos); a *n*-31, são 1.698 formas que são comuns a 50 quaisquer textos – isto é, ainda menos de 1% do total disponível. Descendo até *n*-76, constatamos que temos ainda "só" 29.300 formas comuns a 5 textos, sejam eles quais forem, (isto é, 27% "somente" das formas disponíveis).

Mas a curva descrita é perfeitamente regular – ver Quadro V. Nenhuma fratura, nenhum limite de ruptura aparece –, o que não deixa de ser surpreendente, sobretudo se considerarmos que o corpus de referência, que é "unitário" (textos literários brasileiros unicamente) compreende cinco gêneros diferentes e obras que, como indicamos acima, apresentam fortes diferenças de extensão (menos de 20.000 palavras para as mais curtas e mais de 200.000 palavras para a mais longa). A regularidade é perfeita, de *n*-1 a *n*-80 (quadro V-a); esta regularidade confirma-se quando atingimos o patamar das 1000 formas (aproximadamente) comuns a 60 textos assim como quando chegamos ao patamar das 4000 formas (somente) que são comuns a 32 textos (quadro V, b e c). A maior quantidade – isto é, todo o resto (cerca de cento e quatro mil outras formas) – só vai aparecer num número muito reduzido de textos.

4. Em busca de uma tipologia. A definição do «perfil das palavras»

Quais são as formas comuns a um gênero, um autor, um período cronológico, a um conjunto qualquer de textos? Para tentar dar uma resposta a todas estas questões que de imediato surgem – e que representam uma outra maneira de observar as diferenças em termos desta "repartição" que aqui nos interessa – construímos um modelo de representação tipológica de que vamos a seguir expor os fundamentos.

4.1. A tipologia. Enfoque metodológico

Voltemos ao quadro II e examinemos de perto os dados. *Bacabau* aparece uma só vez no texto n° 76 e está ausente em todos os outros. *Bacamarte* aparece 1 vez nos textos n° 1, 12, 44, 48, 81; 2 vezes no n° 66; 3 vezes no n°78; 5 vezes no n° 57; 13 vezes no n° 43 e 60 vezes no n° 42. Esta forma está ausente num certo número de textos, e é rara em outros,..., frequente no n° 43 e muito frequente no n° 42. Podemos, seguindo este raciocínio, resumir esta distribuição através de seis critérios: «absent», «rare», «peu fréquent», «fréquent», «assez fréquent» e «très fréquent» que podem ser codificados com 3 bits na memória-máquina. O "quadro de contingência" que daí resulta pode ficar também diretamente acessível na memória central, o que abre um campo ainda mais vasto para a pesquisa.

Vejamos aqui, como exemplo, duas palavras que têm cerca de 500 ocorrências no nosso corpus (*baixa* e *basta*). E, para cada uma delas, vamos tentar saber como é que elas respondem aos critérios considerados, a saber "absent", "rare", "peu fréquent", "fréquent", "assez fréquent" e "très fréquent".

4.2. Cálculos. Primeiros resultados

Baseando-se, por linha, na frequência média, o método adotado permite definir patamares sucessivos, e «descreve» assim a repartição.

Para *basta*, por exemplo, a média é $m = 498/70$ (498 ocorrências, para 70 textos).

Assim, ao utilizar a escala que segue,

absent	rare	peu fréquent	fréquent	assez fréquent	très fréquent
0	0,2m	0,5	m	1,5m	3m

obtemos o perfil da palavra *basta*, que obedece ao critério «absent» em 20 textos, «rare» em 9, «peu fréquent» em 18 e assim por diante.

20	9	18	24	7	3

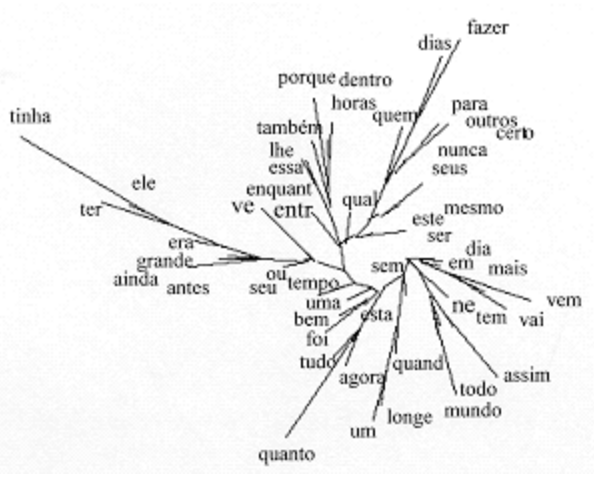
O perfil da forma *baixa* é 24, 11, 17, 12, 10, 7.

O método aqui proposto permite-nos ainda:

a) comparar os dois resultados obtidos: a lista relativa à "repartição" será assim comparada com a lista das frequências reais observadas;
b) submeter eventualmente cada resultado, cada perfil constatado, a uma análise multidimensional (AFC ou análise em árvores).

3.3 Ensaio de aplicação: análise em árvores

Quadro IV



O quadro de repartição das 93 formas presentes em todos os textos foi submetido a uma análise em árvores. Para a realização desta análise (Quadro IV), foram consideradas 58 formas – os artigos, os sinais de pontuação e as preposições foram com efeito globalmente excluídos da experiência, com exceção todavia da preposição «em». Resulta desta análise que três grandes «ramos» se destacam: eles correspondem a espaços bem delimitados, que são os do verbo «ter», do verbo «fazer» e dos verbos «ir» e «vir».

Observamos, por outro lado:

- que o campo do verbo «ter» é o do passado (formas «tinha» e «era») e que este campo atrai na sua esteira as formas «antes» e «ainda». Este campo é também o da terceira pessoa «ele»;
- que o campo do verbo «fazer» é o que carrega a marca do futuro («depois»); ele compreende igualmente as formas «horas», «dias», «quem» e «para», assim como todo um sub-grupo conduzido por «porque», no qual encontramos também a forma «enquanto»;
- que o campo dos verbos «ir» e «vir» compreende as três formas verbais que estão no presente do indicativo: «vem», «vai» e «tem». A forma «agora» pertence (naturalmente)

a este mesmo campo, que é também o das formas «quando» e «quanto», do indefinido «tudo», do demonstrativo «esta» e do substantivo «dia»;

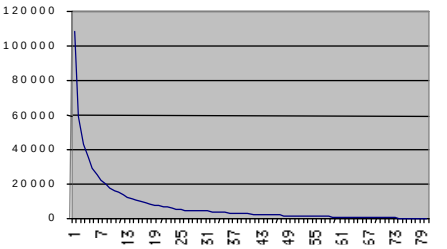
- que, por outro lado, a palavra «tempo» situa-se na intersecção dos três campos.

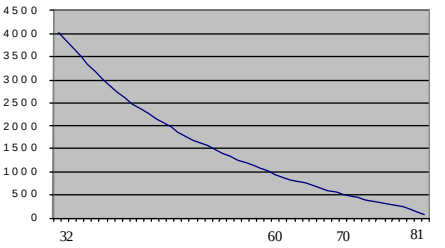
Eis aqui o perfil das 58 formas que foram submetidas à análise.

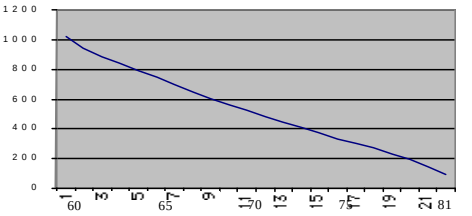
agora 10 21 32 14 4	grande 16 15 30 18 2	se 8 20 40 11 2
ainda 13 14 30 22 2	horas 13 23 29 11 5	sem 9 18 37 15 2
antes 14 17 30 19 1	lhe 14 16 31 15 5	ser 15 17 35 11 3
assim 7 20 44 9 1	longe 9 11 41 18 2	seu 13 19 30 16 3
bem 10 21 33 14 3	mais 7 20 39 13 2	seus 13 14 36 14 4
certo 15 12 38 13 3	mas 8 21 35 16 1	também 14 18 35 9 5
dentro 15 20 30 11 5	mesmo 14 16 36 12 3	tanto 10 21 37 10 3
depois 14 14 35 14 4	mundo 11 19 42 7 2	tem 7 19 39 13 3
dia 7 18 39 15 2	nem 8 19 38 13 3	tempo 12 16 35 16 2
dias 16 11 42 7 5	nunca 16 14 33 14 4	ter 17 16 23 24 1
ele 19 15 23 21 3	ou 14 18 31 16 2	tinha 25 11 21 21 3
em 8 20 37 14 2	outros 18 11 36 14 2	todo 11 18 39 10 3
enquanto12 15 36 14 4	para 17 10 39 12 3	tudo 9 23 34 12 3
entre 11 16 36 14 4	por 8 21 39 11 2	um 8 12 44 16 1
era 18 14 28 18 3	porque 18 22 28 10 3	uma 10 18 34 17 2
essa 14 19 29 14 5	qual 14 17 33 14 3	vai 4 18 42 16 1
esta 9 20 34 16 2	quando 9 16 39 16 1	vem 3 24 38 14 2
este 13 16 36 13 3	quanto 10 27 26 15 3	vez 14 21 26 17 3
fazer 20 11 34 10 6	que 5 18 43 13 2	
foi 10 20 32 18 1	quem 13 12 43 10 3	

Constatamos que o grupo tinha/ter/antes/era obedece ao critério «rare» (primeira coluna) num número bastante grande de textos; o grupo fazer/horas/depois obedece com menor frequência ao critério «rare», mas aparece com mais força na quinta coluna (critério «très fréquent»). O grupo vai/vem/tem/agora (do presente) é enfim o que obedece menos ao critério «rare» e as suas características manifestam-se sobretudo nas colunas 2 e, particularmente, 3 («assez fréquent» / «fréquent»).

Quadro V: distribuição das palavras: em todos os textos (a), em 51 textos (b) e em 21 textos (c)







Quadro VI - extrato: 100 formas sobre um total de 567 formas comuns a 95% dos textos

:	além	amigos	aqui	bom	casa	cidade	cuja
;	alguém	amor	ar	braço	caso	cima	cujo
?	algum	ano	as	braços	causa	cinco	d'
a	alguma	anos	às	branco	cedo	coisa	da
à	algumas	antes	assim	breve	certa	com	dá
acaso	alguns	ao	até	cabeça	certo	comigo	dado
agora	ali	aos	basta	cabelos	céu	como	dando
água	alma	apenas	bela	cada	chama	conta	daquela
ah	alta	apesar	beleza	cair	chão	contra	daquele
aí	alto	aquela	belo	caminho	chegar	cor	cuja
ainda	amanhã	aquelas	bem	campo	chegou	coração	cujo
alegre	ambos	aquele	boa	canto	cheia	corpo	d'
alegria	amigo	aqueles	boca	carta	cheio	creio	da

Conclusão

Esta apresentação, que diz respeito a um método de análise, deve ser vista antes de mais nada como um simples ensaio que tem como fundamento um enfoque novo – o da repartição – e de que as aplicações aqui feitas representam somente alguns exemplos. Observemos todavia que a oposição presença/ausência permite que se proponha uma tipologia da distribuição que, pela primeira vez, leva em consideração também a problemática da ausência. Por outro lado, e este ponto é de total interesse, os patamares, ou rupturas, legitimamente esperados, não surgem, não se produzem – e este fato por si só já nos permite dizer que outras experiências serão sem dúvida necessárias e úteis, com outros textos, para que possamos mais uma vez observar a natureza das regras que determinam a distribuição das diferentes formas aqui consideradas.

Observemos ainda que quando, aos poucos, aumentamos o nosso campo de observação – passando de 1 para 81 textos – o número de formas comuns diminui, e que isto se produz de modo perfeitamente regular. Este fato, por si só, não é novo. Mas a oposição presença/ausência mostra-nos que o coração do sistema (que, segundo dados já bem conhecidos e consagrados, é constituído de cerca de mil palavras) torna-se bem mais discreto e concentrado quando também as ausências são consideradas: encontramos com efeito somente 93 formas comuns aos 81 textos analisados quando, intuitivamente, pensávamos que íamos encontrar um número bem mais elevado.

Estes fatos ou conclusões mostram que este enfoque é promissor, no que se refere à descrição de certos fenômenos linguísticos e que, a partir daqui, um vasto campo se abre para muitas outras aplicações.

Notas

[1] Texto modificado da comunicação apresentada nas "JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles". Saint-Malo, 13-15 de Março de 2002. O modelo matemático aqui utilizado é proposto por Xuan LUONG (UMR6039 – ILF – CNRS – Nice).

[2] As principais características desta base foram apresentadas nas JADT de Nice, em 1998. Publicação: CNRS - Université de Nice.

[3] Mais de trinta diferentes nomes: Gregório de Matos, Padre Antônio Vieira, Basílio da Gama, Cláudio Manuel da Costa, Álvares de Azevedo, Joaquim Manuel de Macedo, José de Alencar, Machado de Assis, Joaquim Nabuco, Aluísio de Azevedo, Cruz e Sousa, Lima Barreto, etc.

[4] Ver cd-rom PORTEXT - Literatura brasileira. UPRESA 6039 - Nice.

[5] Observemos ainda a título de exemplo que as dez formas mais frequentes, com um total de 441.814 ocorrências, representam sozinhas cerca de 10% do nosso corpus de referência.

[6] Representando cada uma das 108.329 formas.

[7] Quando de um primeiro enfoque, nós suprimimos (teoricamente) textos, de forma aleatória, respeitando patamares sucessivos (-5%, -10% etc). Este enfoque tinha no entanto um inconveniente que rapidamente se fez sentir e que resultava do próprio carácter aleatório do processo; o método "mais seguro" (n-1, n-2, etc), utilizado paralelamente e descrito acima permitiu-nos avançar com um maior grau de segurança e obter mais precisão nos resultados.

Referências Bibliográficas

Barthélemy J.P. e Luong X. (1998) "Représenter les données textuelles par les arbres..." in JADT 1998, 4èmes Journées Internationales d'Análise Statistique des Données Textuelles. S. Mellet e al. Ed. pp 49-71, Nice.

Luong X. (1988) "Using a tree model in textual analysis", in *Computers and the Humanities*; 23; pp 397-402.

Maciel C. (1998) «La page Web de la base de données textuelles PORTEXT. L'outil, les textes juridiques, les aires géographiques», in JADT 1998, 4èmes Journées Internationales d'Análise Statistique des Données Textuelles. S. Mellet e al. Ed. pp 49-71, Nice.

Maciel C. (1996) *Le Projet PORTEXT*, revista CUMFID, número especial, CNRS, Bases, Corpus et Langage, Nice. (Editado por).