Fréquences et répartition des mots dans un corpus de littérature brésilienne

Xuan Luong et Carlos Maciel

UPRESA 6039 – CNRS – Institut de Linguistique Française – Nice – France[1]

Abstract

After examining a table of frequencies and realizing that only 93 works (out of 108329) were present in all of the 81 texts making up the reference corpus (Brazilian literature) we devoted our attention to the distribution of words among the separate texts, bearing in mind the presence v. absence criterion with a view to establishing a typology.

Résumé

Après avoir constaté, à partir d'un tableau de fréquences, que 93 mots seulement (sur 108329) étaient présents dans les 81 textes faisant partie d'un corpus de référence (littérature brésilienne), nous nous sommes intéressé à la répartition des mots entre les différents textes, en tenant compte de l'opposition présence absence et dans le but de définir une typologie.

Mots-clés : corpus, littérature brésilienne, distribution, répartition

1. Introduction

1.1. Le corpus

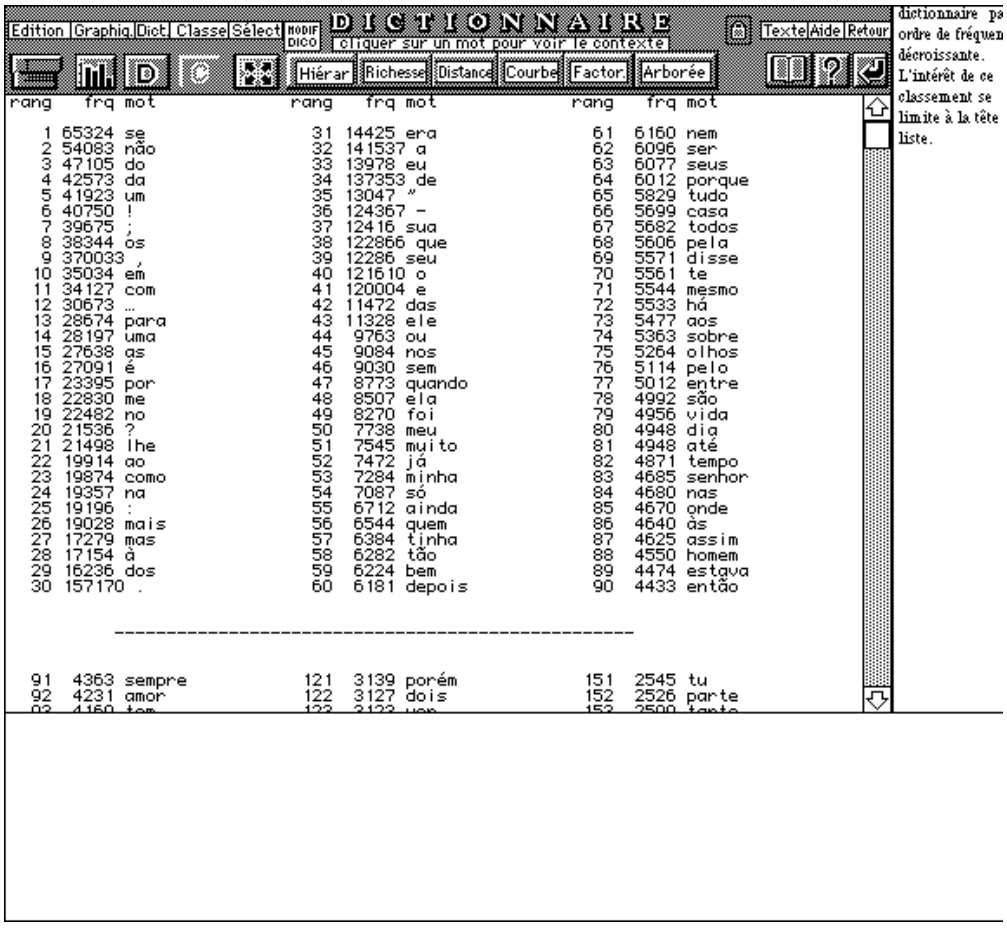
L'expérience ici décrite porte sur un ensemble de 81 textes de littérature brésilienne issus de la base de données textuelles PORTEXT[2]. Cet ensemble, avec 4620146 occurrences pour 108329 formes différentes, couvre quatre siècles de littérature brésilienne; tous les principaux auteurs y sont représentés[3], ainsi que tous les principaux genres (théâtre, roman, poésie…).

[1] Cette communication fait suite aux travaux déjà réalisés au sein de l'UPRESA 6039 (notamment PORTEXT – base de données textuelles en langue portugaise sous la responsabilité scientifique de Carlos MACIEL – et les publications de Xuan LUONG sur les analyses arborées) , auxquels nous renvoyons.

[2]Les principales caractéristiques de cette base ont été exposées aux JADT de Nice.

[3]Plus de trente différents noms: 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, etc.

Tableau I: les mots les plus fréquents



rang	frq	mot	rang	frq	mot	rang	frq	mot
1	65324	se	31	14425	era	61	6160	nem
2	54083	não	32	141537	a	62	6096	ser
3	47105	do	33	13978	eu	63	6077	seus
4	42573	da	34	137353	de	64	6012	porque
5	41923	um	35	13047	"	65	5829	tudo
6	40750	!	36	124367	-	66	5699	casa
7	39675	;	37	12416	sua	67	5682	todos
8	38344	os	38	122866	que	68	5606	pela
9	370033	,	39	12286	seu	69	5571	disse
10	35034	em	40	121610	o	70	5561	te
11	34127	com	41	120004	e	71	5544	mesmo
12	30673	…	42	11472	das	72	5533	há
13	28674	para	43	11328	ele	73	5477	aos
14	28197	uma	44	9763	ou	74	5363	sobre
15	27638	as	45	9084	nos	75	5264	olhos
16	27091	é	46	9030	sem	76	5114	pelo
17	23395	por	47	8773	quando	77	5012	entre
18	22830	me	48	8507	ela	78	4992	são
19	22482	no	49	8270	foi	79	4956	vida
20	21536	?	50	7738	meu	80	4948	dia
21	21498	lhe	51	7545	muito	81	4948	até
22	19914	ao	52	7472	já	82	4871	tempo
23	19874	como	53	7284	minha	83	4685	senhor
24	19357	na	54	7087	só	84	4680	nas
25	19196	:	55	6712	ainda	85	4670	onde
26	19028	mais	56	6544	quem	86	4640	às
27	17279	mas	57	6384	tinha	87	4625	assim
28	17154	à	58	6282	tão	88	4550	homem
29	16236	dos	59	6224	bem	89	4474	estava
30	157170	.	60	6181	depois	90	4433	então
91	4363	sempre	121	3139	porém	151	2545	tu
92	4231	amor	122	3127	dois	152	2526	parte

Le corpus ainsi constitué a d'abord été soumis à un traitement par le logiciel Hyperbase (d'Étienne BRUNET – Tableau I)[4], qui nous a fourni le dictionnaire général et le dictionnaire des fréquences. L'observation des résultats obtenus nous permet de mettre rapidement en évidence des faits dont l'importance linguistique n'est pas négligeable. Remarquons qu'Hyperbase – d'après des critères sur lesquels nous ne reviendrons pas ici – nous propose une liste de 162 formes, classées d'après l'ordre décroissant des fréquences.[5] Il s'agit là des 162 " mots " les plus fréquents dans le corpus en référence. Nous n'avons toutefois à ce niveau aucune information sur la répartition de ces formes dans les différents textes qui intègrent le corpus (des informations sur la distribution, sous la forme de données brutes, figurent toutefois ailleurs – et les usagers d'Hyperbase les connaissent bien). Au-delà donc du tableau de distribution des mots – avec leurs sous-fréquences dans les 81 textes - nous chercherons à analyser ici les différences qui surgissent au niveau de la répartition. Le tableau des répartitions, comme nous le verrons ci-dessous, tient compte, pour chaque ligne du tableau, de la présence et de l'absence de la forme considérée dans chacune des 81 colonnes[6].

[4]Voir cédérom PORTEXT - Littérature brésilienne. UPRESA 6039 - Nice.

[5]Remarquons toutefois à titre d'exemple que les dix formes les plus fréquentes, avec un total de 441814 occurrences, couvrent à elles seules environ 10% de notre corpus.

[6]Une première approche nous a conduit à supprimer des textes, de façon aléatoire, par paliers successifs (-5%, -10% etc). Cette approche présente toutefois un inconvénient qui résulte du caractère aléatoire même de la

1.2. La " répartition "

Le mot "répartition" est en conséquence utilisé ici pour désigner essentiellement l'opposition présence/absence d'une forme quelconque dans un texte ou dans un ensemble donné. Nous sommes amenés à étudier des tableaux de contingences de très grande taille et, dans une première approche, la simple présence (ou absence) d'une forme donnée dans un sous-ensemble quelconque constitué par rapport au reste des éléments qui intègrent le corpus peut être significative. Cette répartition peut se confondre, pour ce qui concerne certaines applications, avec la notion déjà très répandue d'exclusivité lexicale et elle intègre sans doute le vaste champ de la distribution ; mais, comme nous le verrons par la suite, elle garde ses spécificités.

1.2.1. A la recherche d'un seuil

Les premières questions soulevées concernent les seuils ou fractures susceptibles d'apparaître dans le tableau des répartitions, si nous tenons compte de toutes les formes présentes dans tous les textes, distribuées par paliers successifs – à partir de la valeur *n* (=81). Notre hypothèse de départ était simple : au fur et à mesure que l'on restreint le nombre de textes en jeu et compte tenu des seules présences (*n* -10, *n* - 20, *n* - *x*), un seuil devra voir le jour, donnant quelque part à la courbe une allure irrégulière. Les différences de genre et, notamment, les différences de taille entre les textes pouvaient en effet justifier à elles seules cette atttente. Avant de montrer quelques résultats des investigations, nous indiquons d'abord notre manière de procéder.

1.2.1. Données de base et programmation informatique modulaire

Le tableau II est un extrait du fichier de la distribution des formes dans les 81 textes. Par exemple, le mot *bacalhau* est de fréquence totale 23 dans le corpus, de fréquence 2 dans le texte n° 1, de fréquence 1 dans le n° 47, ..., de 6 dans le n° 56 , etc... Le tableau de présence/absence, de très grande taille, est composé de 108329 chaînes de caractères associées chacune à 81 valeurs logiques 0/1 . En numérisant ces chaînes et en codant chaque valeur logique par un seul bit, nous envoyons l'information directement dans la mémoire centrale de la machine et constituons ainsi le stock de nos données de base. A chaque ligne de la répartition correspond une lecture statistique de ce tableau. Le temps de réponse est raisonnable lorsque l'on traite ce très grand ensemble de données en mémoire centrale.

Tableau II: le fichier de la distribution des formes dans les textes

```
.....
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 des données du tableau, il nous était possible de procéder de plusieurs manières. La plus connue est l'écriture d'un programme informatique pour chaque question posée – qui sera alors modifiée pour traiter des sujets assez proches. La meilleure méthode mais aussi la plus difficile à réaliser est la construction d'un logiciel permettant toutes les différentes et

---

démarche ; la méthode " la plus sûre " (*n*-1, *n*-2, etc), utilisée en parallèle et décrite ci-dessous nous permet ainsi d'avancer avec un peu plus de certitude, les résultats étant naturellement plus fins.

nombreuses interrogations que ce genre de recherche suppose. Ce logiciel exige un travail considérable pour son élaboration, et il n'est " rentable " que s'il est destiné à un usage courant.

Nous avons ici adopté une position intermédiaire que nous appelons " méthode de programmation modulaire ". Le langage utilisé est Java 2.1 de *Sun Microsystems*. Le programme en référence est composé d'une plate-forme centrale et de plusieurs modules indépendants. Un module s'exécute à partir de la plate-forme et, une fois sa tâche terminée, les résultats s'affichent, se fixent dans une zone et l'on revient sur la plate-forme. Quelques exemples de modules : charger les données à partir d'un fichier, corriger ou modifier des données, sauvegarder les résultats dans un fichier, sélectionner un ensemble A, sélectionner un ensemble B suivant divers critères, et, enfin, réaliser les listes du vocabulaire commun des textes A et B et du vocabulaire de A \ B . Chaque fois que l'on a à poser une question spécifique, la construction d'un module particulier s'impose: c'est le *Traitement*. Ce module, intégré dans l'environnement existant, ne nécessitera qu'un minimum de codes et manipulations informatiques.

2. Quelques questions soulevées

Quelles sont les formes présentes dans tous les textes (n=81)? Combien y a-t-il de formes présentes dans n-x textes?

L'observation de la liste générale (Tableau III) – de présence dans les 81 textes du corpus – nous permet de constater que 93 formes uniquement (sur 108329) figurent dans tous les différents textes, alors que – guidés par toutes les expériences précédentes, notamment celles qui, depuis longtemps, ont mis en lumière le "fondamental" - l'on pouvait légitimement s'attendre à en trouver davantage.

Si nous mettons en face les deux listes (Tableaux I, à savoir l'index fréquentiel décroissant fourni par Hyperbase, et III, liste des 93 formes communes à tous les textes), nous constatons que, au seuil de 93 (nombre égal de formes pour chaque liste), nous trouvons 75 formes communes aux deux listes ; autrement dit, sur les 93 formes les plus fréquentes du corpus, 18 ne sont pas communes aux deux listes. Toutefois nous trouvons 89 formes communes aux deux listes si nous abaissons le seuil jusqu'à prendre en compte les 162 formes proposées par Hyperbase. Un nombre assez important de formes semble ainsi, dans un cas comme dans l'autre, échapper à l'intuition qui tend à associer très haute fréquence et répartition dans tous les textes : en fait, une fréquence très élevée – même lorsqu'il s'agit de certains "instrumentaux" de la langue – n'est pas une garantie de présence; la répartition nous livre ainsi ses premiers résultats palpables et nous invite à poser la question relative à une typologie à construire en fonction de ce seul critère.

Tableau III: les formes communes aux 81 textes

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

Parmi les 18 formes qui manquent à l'appel au seuil de 93, nous trouvons deux formes verbales (*vai* et *vem* – c'est-à-dire *va* et *vient*) et un substantif pluriel (*horas*), qui marquent d'emblée leur spécificité.

À un niveau d'exigence moindre (*n*-1), nous trouvons 147 formes communes à 80/81 textes ; à *n*-11, nous trouvons 522 formes (pour 70 textes) ; à *n*-31, ce sont 1698 formes qui sont communes à 50 textes, quels qu'ils soient – c'est-à-dire encore moins de 1% du total disponible. Mais à *n*-76, nous ne trouvons encore que 29 300 formes communes à 5 textes, quels qu'ils soient (c'est-à-dire 27% des formes disponibles « seulement »).

Mais la courbe décrite est tout à fait régulière – voir Tableau V. Aucune fracture, aucun seuil n'apparait – alors même que, à l'intérieur de ce corpus "unitaire" (textes littéraires brésiliens uniquement) nous avons cinq genres différents représentés et des oeuvres qui, comme nous l'avons précisé ci-dessus, présentent aussi d'assez fortes différences d'étendue (moins de 20000 mots pour les plus courtes, plus de 200000 mots pour la plus longue). La régularité est parfaite, de n-1 à n-80 (tableau V-a); cette régularité est confirmée au seuil des 1000 formes (environ) communes à 60 textes ainsi qu'au seuil des 4000 formes (seulement) qui sont communes à 32 textes (tableau V, b et c). La très grande masse – ou tout le reste (environ cent quatre mille autres formes) – ne trouvera sa place que dans un nombre très réduit de textes.

3. A la recherche d'une typologie

Quelles sont les formes communes à un genre, un auteur, une période chronologique, à un ensemble quelconque de textes? Pour essayer de donner une réponse à toutes ces questions qui d'emblée s'imposent à nous – et c'est là une autre façon d'observer les différences au niveau de la " répartition " dont nous nous occupons ici – nous avons construit un modèle de représentation typologique dont nous exposons ci-après les fondements.

3.1. La typologie. Approche méthodologique

Revenons au tableau II et examinons de près ces données. *Bacabau* se trouve une seule fois dans le texte n° 76 et absent dans tous les autres. *Bacamarte* apparaît 1 fois dans les textes n° 1, 12, 44, 48, 81 ; 2 fois dans le n° 66 ; 3 fois dans le n°78 ; 5 fois dans le n° 57 ; 13 fois dans le n° 43 et 60 fois dans le n° 42 . Cette forme est absente dans un certain nombre de textes, rare dans d'autres,…, fréquente dans le n° 43 et très fréquente dans le n° 42 . On peut, par exemple, résumer cette distribution par 6 caractères : absent , rare , peu fréquent, fréquent, assez fréquent et très fréquent qui peuvent être codés par 3 bits dans la mémoire machine. Le "tableau de contingence " ainsi obtenu peut être aussi directement accessible en mémoire centrale, et cela ouvre un plus vaste champ d'investigation.

Prenons ici, et à titre d'exemple, deux mots de fréquence de 500 environ (*baixa* et *basta*). Et nous cherchons à savoir comment ils répondent aux critères considérés, à savoir "absent", "rare", "peu fréquent", "fréquent", "assez fréquent" et "très fréquent".

3.2. Calculs. Premiers résultats

Fondée, par ligne, sur la fréquence moyenne, la méthode adoptée permet de définir des seuils successifs, et rend compte d'une répartition.

Pour *basta* , par exemple, la moyenne est m = 498/70 (498 occurrences , réparties sur 70 textes).

Ainsi, si nous utilisons l'échelle qui suit,

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

nous obtenons le profil du mot *basta*, qui est absent dans 20 textes, rare dans 9, peu fréquent dans 18 et ainsi de suite.

20	9	18	24	7	3

Le profil de la forme *baixa* est 24, 11, 17, 12, 10, 7.

La méthode ici proposée nous permet, par la suite :

a) de comparer les deux résultats obtenus : la liste relative à la " répartition " sera ainsi comparée à la liste des fréquences réelles observées ;

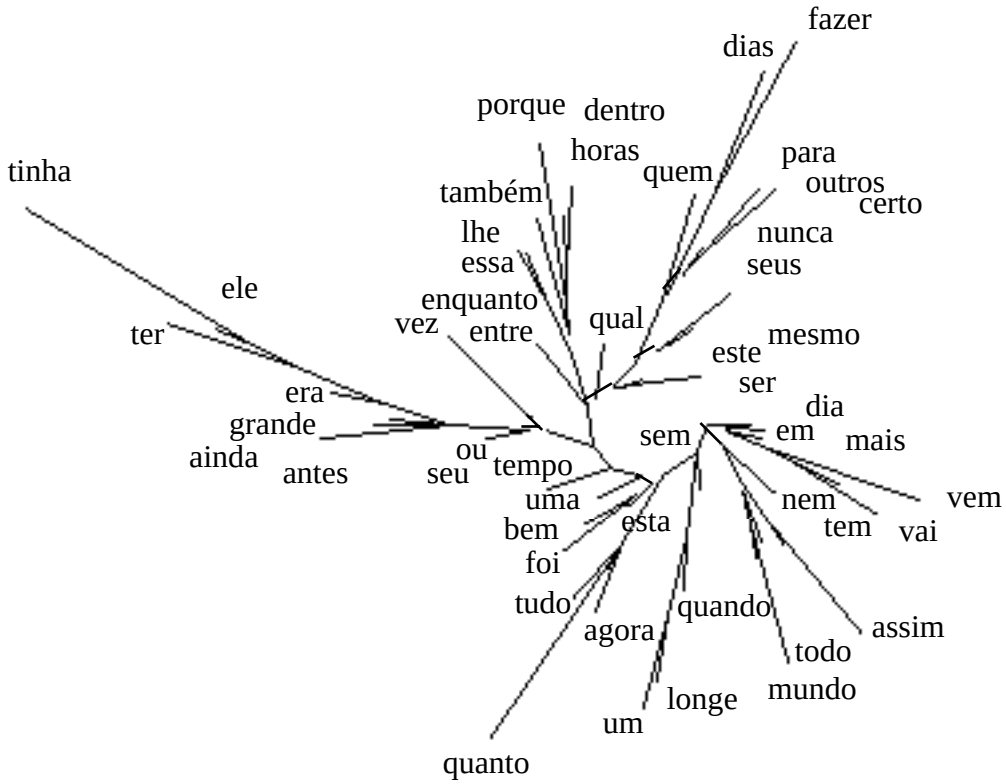
b) de soumettre éventuellement les profils constatés à une analyse multidimensionnelle (AFC ou analyse arborée).

3.3. Essai d'application : analyse arborée

Le tableau de répartition des 93 formes présentes dans tous les textes a été soumis à une analyse arborée. Pour la réalisation de cette analyse (Tableau IV) , 58 formes ont été retenues – les articles, les signes de ponctuation et les prépositions ont en effet été globalement exclus de l'expérience, à l'exception toutefois de la préposition « em » / en. Il ressort de cette analyse que trois grandes « branches » se dessinent : elles correspondent à des espaces bien délimités, qui sont ceux du verbe « ter » (avoir), du verbe « fazer » (faire) et des verbes « ir » (aller) et « vir » (venir). Nous remarquons en outre :

- que le champ du verbe « ter » est celui du passé (formes « tinha » / avait et « era » / était) et que ce champ attire dans son sillage les formes « antes » / avant et « ainda » / encore. Ce champ est également celui de la troisième personne « ele » / il-lui ;

Tableau IV



- que le champ du verbe « faire » porte aussi la marque du futur (« depois » / après) ; il comprend également les formes « horas » / heures, « dias » / jours , « quem » / qui et « para » / pour, ainsi qu'un sous-groupe conduit par « porque » / pourquoi-parce que, dans lequel nous trouvons aussi les formes « enquanto » / pendant-alors que ;
- que le champ des verbes « ir » / aller et « vir » / venir comprend les trois formes verbales qui sont au présent de l'indicatif : « vem » / vient, « vai » / va et « tem » / a. La forme « agora » / maintenant appartient à ce même champ, qui est celui dans

lequel nous trouvons également « quando » / quand, « quanto » / combien, l'indéfini tout (« tudo »), le démonstratif « esta » / celle-ci et le substantif « dia » / jour ;

- que, par ailleurs, le mot « tempo » / temps se situe à l'intersection des trois champs.

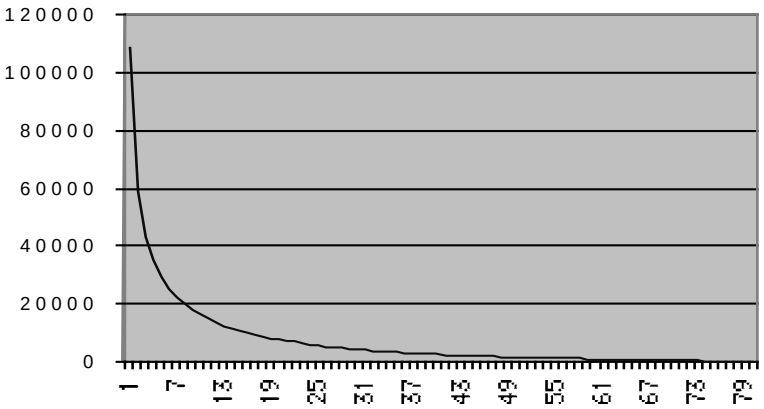
Voici le profil des 58 formes soumises à l'analyse.

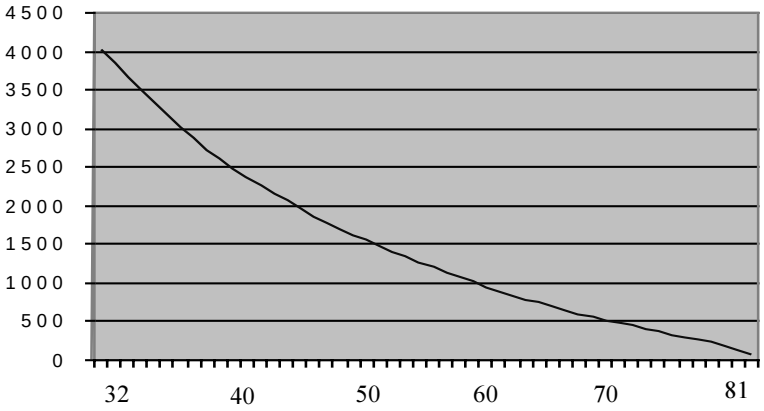
agora 10 21 32 14 4
ainda 13 14 30 22 2
antes 14 17 30 19 1
assim 7 20 44 9 1
bem 10 21 33 14 3
certo 15 12 38 13 3
dentro 15 20 30 11 5
depois 14 14 35 14 4
dia 7 18 39 15 2
dias 16 11 42 7 5
ele 19 15 23 21 3
em 8 20 37 14 2
enquanto12 15 36 14 4
entre 11 16 36 14 4
era 18 14 28 18 3
essa 14 19 29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

ter 17 16 23 24 1
tinha 25 11 21 21 3
todo 11 18 39 10 3
tudo 9 23 34 12 3
um 8 12 44 16 1
uma 10 18 34 17 2
vai 4 18 42 16 1
vem 3 24 38 14 2
vez 14 21 26 17 3

Nous remarquons que le groupe tinha/ter/antes/era répond au profil « rare » (première colonne) dans un assez grand nombre de textes ; le groupe fazer/horas/depois répond moins souvent au profil « rare », alors qu'il s'affirme avec davantage de force au niveau de la cinquième colonne (« très fréquent »). Le groupe vai/vem/tem/agora (du présent) est enfin celui qui répond le moins souvent au profil « rare », étant davantage caractérisé par les colonnes 2 et, surtout, 3 (« assez fréquent » / « fréquent »).

Tableau V : distribution des mots dans :
la totalité des textes (a), 51 textes (b) et 21 textes (c)





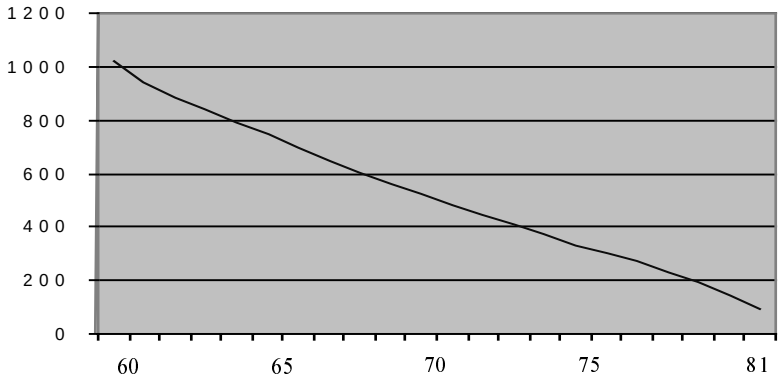


Tableau VI - extrait : 164 formes sur un total de 567 formes communes à 95% des textes

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

dessa	digo	é	era
dessas	direito	efeito	eram
desse	disse	eis	espera
desses	diz	ela	esperança
desta	dizendo	elas	esperanças
deste	dizer	ele	espírito
destino	dizia	eles	essa
deu	do	em	essas
deus	doce	embora	esse
deve	dois	enfim	esses
devia	dor	enquanto	est
dez	dos	então	esta
dia	duas	entrar	está
diante	dúvida	entre	estado
dias	e	entretanto	estão

4. Conclusion

Cet exposé de méthodologie doit être perçu avant tout comme un essai, fondé sur une approche nouvelle – celle de la répartition – dont les applications ne sont pour l'instant données qu'à titre d'exemple. Remarquons toutefois que l'opposition présence/absence permet de proposer une typologie de distribution qui tient compte également de l'absence. Par ailleurs, et ceci a tout son intérêt, les effets de seuil, légitimement attendus, ne se produisent pas – et ce fait en tant que tel mérite d'ailleurs de faire l'objet d'autres expériences.

Remarquons par ailleurs que, au fur et à mesure que l'on monte en puissance – de 1 à 81 textes – le nombre de formes communes diminue, de façon tout à fait régulière. Ce fait, en tant que tel, n'est pas nouveau. Par contre, l'opposition présence/absence nous montre bien que le noyau dur (que l'on situe souvent à mille mots environ) se montre beaucoup plus discret et concentré lorsque l'on tient compte des absences: seulement 93 formes communes aux 81 textes soumis à l'analyse alors que l'on s'attendait à en trouver bien davantage.
Ces quelques faits ou conclusions montrent que l'approche est prometteuse et qu'elle pourra sans doute s'ouvrir à d'autres applications.

Références

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

Luong X. (1988) "Using a tree model in textual analysis", in *Computers and the Humanities* ; 23; pp397-402

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'Analyse Statistique des Données Textuelles. S. Mellet et al. Ed. pp 49-71, Nice

Maciel C. (1996) *Le Projet PORTEXT*, revue CUMFID, numéro spécial, CNRS, Bases, Corpus et Langage, Nice (Edité par).