SYLED - CLA2T

Université de la Sorbonne Nouvelle - Paris 3

Explorations textométriques

Volume 1 : *corpus et problèmes*

Sous la direction de
André Salem et Serge Fleury

F. Abbassi, E. Née, C. Pineira-Tresmontant, A. Salem
L. Sansonetti, M. Leenhardt,
P. Couton-Wyporek, Romuald Schummer

2009

Nous avons rassemblé plusieurs compte-rendus d'expériences réalisées avec les logiciels de la famille Lexico au cours de nombreuses recherches et dans le cadre de collaborations diverses. Les navigations rassemblées ici ont été choisies pour mettre en évidence la très vaste gamme des domaines d'application des méthodes textométriques ainsi que les fonctionnalités des logiciels **Lexico3** et **mkAlign**. Elles sont publiées sous la forme de trois volumes (**volume 1** : *corpus et problèmes*, **volume 2** : *séries textuelles chronologiques*, **volume 3** : *corpus multilingues*).

Lexico3

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

mkAlign

http://tal.univ-paris3.fr/mkAlign/

Lexicometrica

http://www.cavi.univ-paris3.fr/lexicometrica/

Fonctionnalités de Lexico3

Tableau des fonctionnalités

Pour présenter les fonctionnalités mises en œuvre dans les sections suivantes, nous avons réunis ci-dessous les différentes icônes associées aux fonctionnalités visées dans leur contexte d'utilisation :

Fenêtre/barre principale

Fenêtre « *carte des sections* »

Fenêtre « *groupe de formes* »

Fenêtre PCLC

Le tableau ci-contre rassemble, nomme et présente brièvement l'ensemble des fonctionnalités du logiciel ***Lexico3*** utilisées pour mener à bien l'exploration textométrique proposée dans les sections suivantes. On pourra aussi se reporter aux différents manuels du logiciel disponibles en ligne.

N°	*Nom*	*Paramètres*	*Localisation*	*icône*
1	SEGMENTATION	Liste de délimiteurs *Par défaut* : `.,:;!?/_-\"'()[]{}§$`	Barre principale	
3	CONCORDANCE	Forme (ou Type Généralisé)	Barre principale	
4	SEGMENTS REPETES		Barre principale	
5	PCLC	Une fois la partition construite, on peut accéder au tableau présentant les Principales Caractéristiques lexicométriques de la partition.	Barre principale	
6	PARTITION	Une clé définissant une partition dans le corpus original est du type : `<CLE= « valeur »>` C'est le nom de la clé qui est donné ici pour construire la partition visée	Barre principale	
6	VENTILATION	Forme ou groupe de formes		
7	CARTE DES SECTIONS	délimiteur de section	Barre principale	
8	GROUPE DE FORMES	Cette fonctionnalité produit des listes de formes qu'il est possible de mémoriser, d'exporter ou de « projeter » sur les graphiques construits par **Lexico3**. Elle permet surtout de faire des recherches de formes ou de groupes de formes en utilisant la notion d'expression régulière.	Barre principale	
5.3	AFC		Fenêtre des PCLC	
5.1, 7.2	SPECIFICITES (POSITIVES NEGATIVES)	Partie ou section du corpus	Fenêtre des PCLC Carte des sections	 

Lexico3, Tableau des Fonctionnalités

Glossaire

segmentation - opération qui consiste à délimiter des unités minimales dans un texte. Les **unités minimales** (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent)

caractères délimiteurs / **non-délimiteurs** : distinction opérée sur l'ensemble des caractères qui entrent dans la composition du texte, permettant aux procédures informatisées de segmenter le texte en occurrences (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.

- les caractères **délimiteurs de séquences** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.

- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

forme ou "**forme graphique**" : archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence

partition (d'un corpus de textes) : division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

ventilation (des occurrences d'une unité dans les parties du corpus) : La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences de cette unité dans chacune des parties, prises dans l'ordre des parties

motif : un ensemble d'objets possédant une propriété reconnaissable.

analyse factorielle : famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

spécificité positive : pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ

spécificité négative : pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ

partie (d'un corpus de textes) : fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

section : portion de texte comprise entre deux délimiteurs de section (exemple : le paragraphe, etc.).

segment répété (ou polyforme répétée) : suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

Les expressions régulières avec *Lexico3*

Dans les sections qui suivent on utilisera à plusieurs reprises la notion d'expression régulière en particulier à travers la fonction «GROUPE DE FORMES ». Nous rappelons ci-dessous brièvement cette notion et les différents opérateurs disponibles avec ***Lexico3*** pour écrire de telles expressions. Les expressions régulières permettent de représenter de manière générique des motifs textuels : un *motif* est un ensemble d'objets possédant une propriété reconnaissable, par exemple tous les mots terminés par le suffixe « able » ou commençant par le préfixe « pré ». Les expressions régulières permettent ainsi de décrire des portions de texte à l'aide d'opérateurs particuliers. Le tableau suivant rassemble l'ensemble des opérateurs disponibles avec ***Lexico3*** pour écrire des motifs sous la forme d'expression régulière :

Opérateur	Fonction	Application
`.` (le point)	Représente n'importe quel caractère	L'expression `"m.l"` représente des séquences comme : `mal, mol`…
`*`	0 ou n occurrences du caractère qui précède	L'expression `"com*e"` représente des séquences comme : `coe, come, comme, commme`…
`+`	1 ou n occurrences du caractère qui précède	L'expression `"com+e"` représente des séquences comme : `comme, commme,`…
`\b`	Représente un début de mot	L'expression `"\bcapital"` représente des séquences comme : `capital, capitale, capitalisme`…
`\b`	Représente une fin de mot	L'expression `".*isme\b"` représente des séquences comme : `syndicalisme, capitalisme`…
`[ ]`	Représente un ensemble de caractères	L'expression `"[aeiou]"` représente des séquences comme : un des caractères de l'ensemble des voyelles minuscules. L'expression `"[a-z]"` représente un des caractères minuscules compris entre a et z.
`[^ ]`	Représente la négation du contenu de l'ensemble de caractères	L'expression `"[^aeiou]"` représente un des caractères parmi ceux qui ne sont pas ceux de l'ensemble des voyelles minuscules

Sommaire

Tutoriels pour l'analyse textométrique

[Tutoriels]

André Salem
salem@msh-paris.fr

Résumé : Ces *tutoriels* devrait permettre à l'utilisateur débutant de ***Lexico3*** (et de *mkAlign*) de se familiariser avec les différentes fonctionnalités du logiciel, à partir de corpus de recherche concrets et, au delà de cette prise en main, d'entrevoir quelques-unes des possibilités offertes par l'approche textométrique des corpus de textes.

Complétant la documentation disponible sur ***Lexico3*** :

- ***Manuel d'utilisation*** ;
- ***User's Manual***, traduction anglaise du même manuel ;
- ***Les 10 premiers pas avec Lexico3***, manuel de prise en main ;
- http://www.cavi.univ-paris3.fr/lexico3www site web *de* ***Lexico3***[1],

et sur *mkAlign* :

- ***Manuel d'utilisation en ligne*** :
  http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

ces ***Tutoriels*** devrait permettre à l'utilisateur débutant, au delà d'une simple prise en main, de se familiariser avec les différentes fonctionnalités de ces logiciels, à partir d'un corpus de recherche concret et d'entrevoir quelques-unes des possibilités offertes par l'approche textométrique des corpus de textes.

- Le corpus ***Père Duchesne*** choisi dans les deux premiers tutoriels pour servir de base à cette exploration guidée est le même que celui utilisé dans les brochures précédentes. Ce corpus à fait l'objet de plusieurs études de caractère pluridisciplinaire dont on trouvera les références dans la dernière section. La ressource textuelle ***duchn.txt*** qui sert de support à ce tutoriel est diffusée en tant que corpus d'essai sur toutes les versions du logiciel ***Lexico***. Accessible sur le CD-Rom ***Lexico3***, elle est installée automatiquement dans le dossier Lexico3 créé lors de l'installation du logiciel. Elle peut également être téléchargée directement depuis le site du logiciel.
- Le corpus ***Investiture Obama*** utilisé dans le troisième tutoriel est disponible en ligne sur le site de *mkAlign*.

On a tenté, dans ce qui suit, de trouver un compromis acceptable entre la nécessité de présenter les principales fonctionnalités du logiciel que le lecteur pourra utiliser dans d'autres

[1] Le logiciel, la documentation et les ressources textuelles (parmi lesquelles la ressource ***duchn.txt***) utilisées dans le présent manuel peuvent être téléchargées depuis ce site.

entreprises textométriques et le compte-rendu d'une recherche qui nous a conduit à agencer l'utilisation de ces méthodes en fonction des objectifs fixés au départ de l'étude, des résultats que nous avons obtenus, mais aussi des perspectives de recherche qui se sont ouvertes à cette occasion. Dans chaque cas, nous nous sommes efforcés de faire en sorte que le lecteur dispose des informations suffisantes pour reproduire par ses propres moyens les fonctionnalités décrites. Ces informations sont rassemblées, à chaque étape, en fin de paragraphe dans un encart annoncé par la séquence **=== *Lexico3* ===** ou **=== *mkAlign* ===**

On se reportera aux manuels d'utilisation pour une description plus détaillée de chacune des fonctionnalités**.**

Le ***Tutoriel n°1****, Exploration du corpus Père Duchesne,* devrait permettre à l'utilisateur de se familiariser avec les notions de *ressources numériques textuelles, de corpus textométriques,* de dépouillement d'un corpus en *unités textuelles*, de partition d'un corpus textométrique et d'acquérir quelques notions sur les principales méthodes textométriques qui permettent d'explorer ces corpus de textes.

Le ***Tutoriel n°2****, Séries textuelles chronologiques,* est consacré à l'étude d'un type de corpus particulier que l'on rencontre très souvent dans le domaine textométrique, qui est celui des corpus rassemblant une série de textes produits au cours du temps par un même émetteur. L'étude de ces corpus obéit à des règles particulières que l'on s'est efforcé de décrire dans ce tutoriel.

Le ***Tutoriel n°3****, Investiture Obama,* est consacré à l'étude d'un corpus aligné avec *mkAlign*.

.

Tutoriel n°1 :
Exploration du corpus *Père Duchesne*

Corpus, unités textuelles, partitions, méthodes textométriques

[Duchesne1]

Apprendre à :

- Construire une ressource textométrique
- Introduire des jalons textuels
- Choisir des unités d'analyse textométrique
- Utiliser les outils textométriques de base
- Conduire une exploration textométrique

1 Le corpus *Père Duchesne*

Le corpus ***Père Duchesne*** que l'on considère ici est constitué de 96 livraisons d'un journal édité par Jacques-René Hébert (1757-1794), parues entre juillet 1793 et mars 1794, durant la Révolution française, dans une période de luttes particulièrement âpres entres différentesfactions politiques. Du fait de sa reproduction et de son acheminement systématique en direction des armées, ce journal a connu une diffusion exceptionnelle pour l'époque qui lui permet de prétendre au titre de *premier media de masse de l'époque moderne*. Le corpus a été réuni dans le cadre d'une étude plus large portant sur la presse jacobine de l'époque et a donné lieu, depuis, à de nombreuses publications[2]. On peut voir sur la figure 1 une reproduction de la première et de la dernière page d'un des exemplaires du *Père Duchesne*, feuille imprimée, pliée en quatre, vendue à la fois par abonnement et à la criée dans les rues de Paris.

1.1 Etablissement de la version numérique du corpus

Lors de la saisie initiale sous forme numérique de cette ressource textuelle, quelques normalisations orthographiques mineures ont été effectuées à l'époque par les chercheurs qui ont transcrit le corpus sous forme numérique. Ainsi, les terminaisons en *oit* ont toutes été ramenées à l'orthographe moderne en *ait* (ex : *foutoit* est devenu *foutait*). Les enrichissements textuels (italiques, gras, etc.) ont été négligés. Les majuscules du texte ont été remplacées par le signe ***** suivi de la minuscule correspondante (ex : *Paris* -> **paris*)[3].

---

[2] Des recherches sur ce corpus ont été réalisées dans le cadre de l'équipe *Révolution française* de laboratoire de l'ENS de St-Cloud [Guilhaumou, 19xx], [Salem, 1993].

[3] Cette technique permet de différer la décision de savoir si les formes qui ne diffèrent que par une majuscule initiale doivent être décomptées séparément. Lors des segmentations ultérieures de la ressource on aura le choix

1.2 Balisage du corpus

Afin de permettre la comparaison entre les différents textes réunis en un même corpus, on a introduit des *jalons textuels* ou *balises* servant à délimiter des *parties*. Dans cette version de Lexico3, les *balises* qui permettent d'introduire les partitions sont du type[4] :

<type=contenu>

Chaque ***type*** particulier de balise (partie située avant le signe « = ») permet de définir une partition du corpus. Pour un type fixé, si on ignore tous les autres types, les différents *contenus* (partie située après le signe « = ») correspondent à autant de parties différentes .dans le corpus. Ainsi, par exemple, la sélection de la clé ***numero*** (<numero= xx>) permet de découper le corpus en 96 parties correspondant chacune à une des 96 livraisons qui constituent le corpus.

Les balises introduites dans le corpus ***Duchn.txt*** sont :

- <*Epg=x*> qui permettent de localiser chacune des pages à l'intérieur d'un même numéro ;
- <*numero=x*> qui permettent de délimiter chacune des 96 livraisons du corpus ;
- <*mois=x*> qui permettent d'opérer un regroupement des livraisons parues à l'intérieur de chaque période d'un mois. Ces périodes sont notées (M1, M2, …, M8) ;
- <*quinzaine=xx*> qui permettent d'opérer un regroupement de ces mêmes livraisons par quinzaines.
- <*semaine =xxxx*> qui permettent d'opérer un regroupement de ces mêmes livraisons par semaines.

2 Zones textuelles

Pour pouvoir s'appuyer sur une division du texte en paragraphes, on a fait précéder chacun des paragraphes par le caractère « **§ »**[5].

Il est également possible de réaliser un découpage correspondant approximativement à un découpage en phrase en fournissant aux outils qui assurent un tel découpage une liste de caractères délimiteurs de phrases (par exemple : « . ?! »)

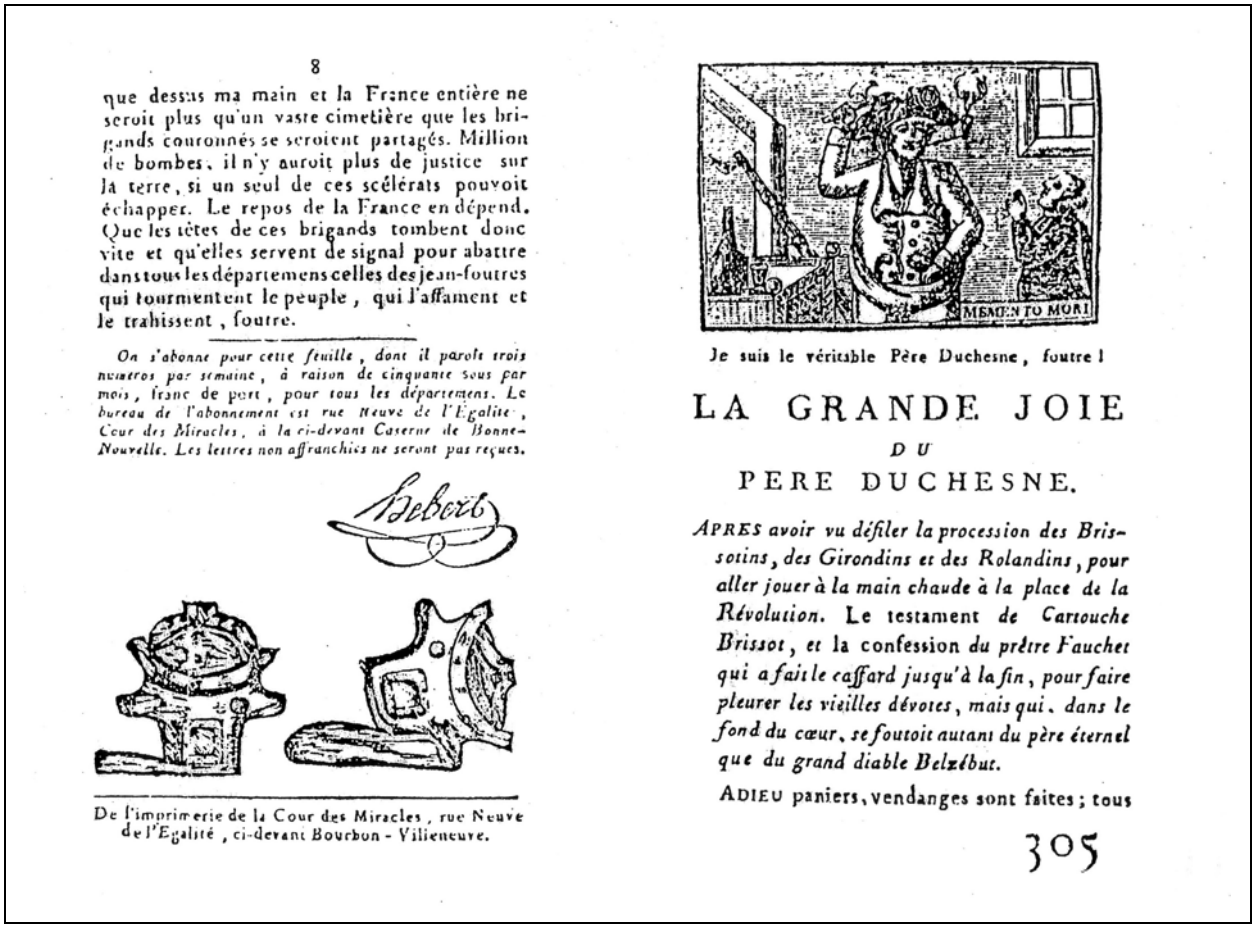
Comme on va le voir dans les sections qui suivent, les découpages en partitions constituent avec les systèmes de découpage en sections un dispositif articulé qui permet de renvoyer les constats textométriques à des zones textuelles délimités avec une précision que l'on peut faire varier.

---

entre deux options : a) on considère que le caractère * est un caractère délimiteur et les formes **abc* et *abc* seront alors considérées comme deux occurrences d'un même type (*abc*) ; b) on décide que le caractère * n'est pas un délimiteur et les formes **abc* et *abc* seront alors considérées comme des occurrences de deux types différents.

[4] Le système de balisage du texte décrit dans ce paragraphe a été élaboré avant .l'apparition de normes plus consensuelles dans la communauté des études textuelles réalisées avec l'aide de l'ordinateur. Les prochaines versions du logiciel prennent en compte les formats d'entrée des textes construits à partir de la norme XML (EXtensible Mark Up Langage). Les fonctionnalités textométriques de ces différentes formes de balisage restent cependant très voisines.

[5] Ce remplacement peut être effectué de manière générique à l'aide d'un logiciel de traitement de texte en remplaçant le caractère « retour-chariot » par la séquence « retour-chariot » suivi de « § ». Avec le logiciel Word, par exemple on utilisera les commmandes : Chercher : ^p Remplacer par :^p §.



8

que dessus ma main et la France entière ne seroit plus qu'un vaste cimetière que les brigands couronnés se seroient partagés. Million de bombes, il n'y auroit plus de justice sur la terre, si un seul de ces scélérats pouvoit échapper. Le repos de la France en dépend. Que les têtes de ces brigands tombent donc vite et qu'elles servent de signal pour abattre dans tous les départemens celles des jean-foutres qui tourmentent le peuple, qui l'affament et le trahissent, foutre.

On s'abonne pour cette feuille, dont il paroît trois numéros par semaine, à raison de cinquante sous par mois, franc de port, pour tous les départemens. Le bureau de l'abonnement est rue Neuve de l'Egalité, Cour des Miracles, à la ci-devant Caserne de Bonne-Nouvelle. Les lettres non affranchies ne seront pas reçues.

Hebert

De l'imprimerie de la Cour des Miracles, rue Neuve de l'Egalité, ci-devant Bourbon-Villeneuve.

MEMENTO MORI

Je suis le véritable Père Duchesne, foutre !

LA GRANDE JOIE
DU
PERE DUCHESNE.

APRES avoir vu défiler la procession des Brissotins, des Girondins et des Rolandins, pour aller jouer à la main chaude à la place de la Révolution. Le testament de *Cartouche Brissot, et* la confession *du prêtre Fauchet qui a fait le cafard jusqu'à la fin, pour faire pleurer les vieilles dévotes, mais qui, dans le fond du cœur, se foutoit autant du père éternel que du grand diable Belzébut.*

ADIEU paniers, vendanges sont faites; tous

305

Figure 1a :
Fac simile de l'édition originale du numéro 305 du ***Père Duchesne*** (1793)

```
<numero=305><Epg=1>
§ la grande joie, du *père *duchesne après avoir vu défiler la procession des *brissotins,
des *girondins et des rolandins, pour aller jouer à la main chaude à la place de la
révolution. le testament de *cartouche,*brissot , et la confession du prêtre *fauchet qui a
fait le cafard jusqu'à la fin,pour faire pleurer les vieilles dévotes, mais qui,dans le
fond du coeur, se foutait autant du père éternel que du grand diable *belzébut.
§ adieu paniers,vendanges sont faites ; tous <Epg=2> les châteaux en *espagne, que vous
avez bâtis, infâmes *brissotins, s'en vont tous en fumée. non, foutre, non la république
que vous aviez vendue aux brigands couronnés ne sera point déchirée. le roi *georges
*dandin, et *pitt, porte-esprit, ont tiré leur poudre aux oiseaux. nous serons républicains
malgré toutes les guinées de l'*angleterre, et tout l'or de l'*autriche, et de l'*espagne
.partout nos affaires prennent la meilleure tournure. Les brigands de la *vendée, sont
dispersés et leurs cadavres engraissent la terre qu'ils ont souillée par leurs crimes; ce
qu'il en reste est cerné de toutes parts et va bientôt tomber sous les coups des généreux
défenseurs de la république; tandis, foutre, que l'armée du nord partout victorieuse est
aux trousses des gros talons et des pieds plats que commande *cobourg ; tandis que *mons,
ouvre ses portes au brave *jourdan ,*brissot, et sa clique marchent à l'échafaud.
```

Figure 1b :
Extrait de l'édition numérisée du numéro 305 du ***Père Duchesne*** (1793)

3 Unités textuelles

Quelles sont les unités qui circulent dans un texte sociopolitique ? Quelles séquences doit-on constituer en unités insécables afin d'opérer des comptages dans les textes ? L'expérience du dépouillement informatisé des corpus de textes montre que ces interrogations constituent à chaque fois des questions centrales pour la recherche en cours et qu'elle ne peuvent être réglées une fois pour toutes et a priori.

Dans le corpus ***Duchn***, par exemple, on serait tenté de constituer en une seule unité le terme *sans-culottes,* pourvu d'une haute fréquence et qui renvoie à un référent assez clairement identifiable à l'époque. Sans doute, le tiret qui unit les deux formes graphiques n'est il pas de même nature que celui qui unit les formes dans *dit-il.* Une autre question se pose alors : Comment traiter le problème automatiquement sans être obligé de trancher au cas par cas ?

Notre expérience nous a conduit à privilégier dans un premier temps les dépouillements appuyés sur des caractères aisément automatisables (appartenance ou non de chacun des caractères à une liste préétablie délimiteurs/non-délimiteurs) et à repousser à une seconde phase l'observation d'unités plus complexes : séquences de formes, cooccurrences etc. Pour la séquence *sans-culottes* présentée plus haut, nos préfèrerons opérer dans un premier temps un dépouillement appuyé sur la segmentation en deux formes distinctes (tiret = délimiteur) laissant à d'autres procédures le soin de repérer ensuite la séquence des deux formes *sans culottes* aisément repérable du fait même de sa forte répétition dans le corpus.

Par ailleurs, au fil des recherches, est apparue la nécessité de généraliser fortement la définition du type d'unité textuelle prise en compte par les analyses textométriques. Le ***type généralisé*** ou ***Tgen*** est défini comme une sélection d'occurrences prise dans le texte. Cette définition permet de prendre en compte les types constitués à partir de critères de sélection difficiles à formaliser[6].

3.1 Le dépouillement en formes graphiques

La première phase de l'exploration textométrique est constituée par la segmentation du corpus textuel en unités qui serviront de base aux décomptes ultérieurs les ***occurrences*** (en anglais *tokens*). A l'issue de cette phase, une seconde phase d'identification constitue un dictionnaire des ***formes*** ou des ***types*** (en anglais *types*). Les *types* regroupent en une même unité chaque classe d'occurrences identiques d'après le critère d'identification retenu[7].

==== *Lexico3* ==== *Segmentation initiale*

- ✓ Lancer Lexico3
- ✓ Sélectionner l'icône ***Segmentation*** (1ère icône en haut à gauche)
- ✓ Choisir le fichier texte à segmenter (Duchn.txt)
- ✓ Accepter les délimiteurs de forme proposés « par défaut » (bouton ***OK***)

[6] Sur les types généralisés, cf. [Lamalle & Salem, 2002]

[7] Selon les études, on trouve des critères d 'identification dont la nature peut varier. Dans certains types de dépouillements, dits *dépouillement en forme graphiques*, on se base sur l'identité graphique des séquences considérées, d'autres formes de dépouillements font intervenir la nature grammaticale des occurrences isolées, voire des informations de type sémantique. On consultera sur ce sujet [Labbé xxx],

Différents outils textométriques que l'on décrira plus loin permettent d'apprécier la fréquence, la répartition, la spatialisation des occurrences relevant de chacun des types constitués à cette étape. Les résultats fournis par ces outils ne sont pas indépendants des types d'unités constitués, mais les mêmes outils s'appliquent à tous les types constitués de la sorte.

La qualité première d'une norme de dépouillement est d'être à la fois simple à énoncer et à automatiser. Le dépouillement du corpus ***Duchn*** en formes graphiques délimitées par les délimiteurs proposés par défaut conduit aux résultats suivants :

nombre des occurrences :	141 182
nombre des formes :	11 070
nombre des hapax :	5 056
forme la plus fréquente *de* :	6 130

3.2 Etude globale des types simples

Ces données sont accessibles en activant l'icône ***PCLC***, dès qu'une partition quelconque a été choisie. Sur le panneau qui apparaît alors on peut étudier l'accroissement du vocabulaire au fil du corpus en activant l'icône ***ACCV***.

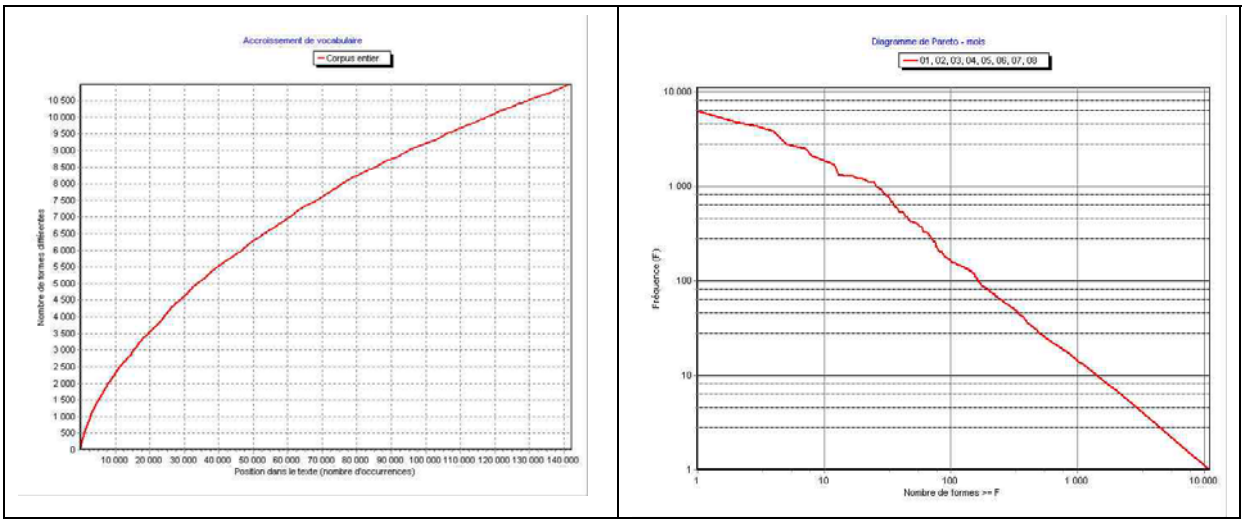


Figure 2 :

Accroissement du vocabulaire et structure de la gamme des fréquences

==== *Lexico3* ==== *Accroissement du vocabulaire*

- ✓ Sélectionner l'icône ***Statistiques par parties*** (5ème icône à partir de la gauche)
- ✓ Choisir un type de clé qui déterminera la partition active du corpus
- ✓ Sélectionner l'icône ***PCLC*** (5ème icône à partir de la gauche)
- ✓ Sélectionner, sur la droite du panneau (5ème bouton à partir du haut) le bouton ***AC (comme Accroissement du vocabulaire)***
- ✓ Le diagramme apparaît dans une fenêtre spécifique. On peut constituer le diagramme correspondant à chacune des parties, ou à un ensemble de parties en les sélectionnant l'une après l'autre et en les glissant sur la fenêtre du Diagramme d'accroissemenet.

Guide de lecture pour la figure 2

Le ***Diagramme d'accroissement du vocabulaire*** que l'on trouve sur la gauche permet d'observer l'apparition de nouvelles formes au fur et à mesure que l'on avance dans le corpus.
Comme c'est toujours le cas pour les corpus textuels, la courbe connaît une croissance rapide au début du corpus ; cette croissance ralentit à mesure que l'on avance dans le corpus. On remarque, par-delà cette caractéristique globale, des zones d'accroissement plus fort ainsi que des paliers durant lesquels l'apport de nouvelles formes est plus faible.

Le ***Diagramme de Pareto*** que l'on trouve sur la permet de visualiser la structure de la gamme des fréquences.

- L'axe vertical permet de représenter la fréquence *F* des formes du textes (laquelle varie de *1* à *Fmax*, fréquence maximale calculée pour le texte ***T***).
- Sur l'axe horizontal, on porte la quantité : *nombre de formes du texte dont la fréquence est supérieure à F*.
- Avant de tracer le Diagramme, on transforme chacune de ces quantités en son logarithme décimal.

Le Diagramme ainsi obtenu prend alors approximativement la forme droite que l'on appelle *Droite de Zipf* en l'honneur de Georges. Kingsley Zipf qui a montré que ce type de procédure réalisée à partir de larges catégories de textes permet de mettre en évidence une propriété statistique commune aux dépouillements en unités lexicales. Cette propriété est parfois présentée sous la forme excessivement simplifiée :

Rang x *fréquence* = *Constante*

3.3 Les types complexes

Les limites que l'on rencontre dès que l'on commence à explorer un corpus textuel à partir de formes isolées de leur contexte immédiat conduisent rapidement à la nécessité d'observer la répartition d'unités plus complexes.

Les segments répétés du Père Duchesne

La fonctionnalité ***Segments répétés*** permet d'établir la liste de toutes les séquences de formes répétées sans changement à différents endroits du corpus dont la fréquence totale dépasse un seuil minimal ***F*** préalablement fixé par l'utilisateur. Les segments ainsi sélectionnés peuvent ensuite être triés selon différents critères : longueur, fréquence, etc.

On retrouvera parmi les segments longs les expressions favorites du ***Père Duchesne***, comme :

```
employer le vert et le sec pour                    15
perdre le goût du pain                             12
ses bons avis aux braves sans-culottes             15
brouiller les cartes                               20
```

Parmi les segments plus courts et plus fréquents on retrouvera les unités composées évoquées plus haut comme :

```
*sans culottes                                    398
jean foutres                                      136
brigands couronnés                                 49
```

Une forme particulière de présentation des contextes du segment *tous les* qui compte 871 occurrences dans le corpus permettra de constater que cet opérateur textuel sert entre autres choses à introduire des entités présentées plutôt comme négatives et contre lesquelles le ***Père Duchesne*** propose de se mobiliser. On peut voir un extrait de cet inventaire au tableau 4.

Cependant l'ensemble constitué par la totalité de segments répétés qui se chevauchent de manière quasiment inextricable se révèle toujours d'une grande complexité et défie toute

description synthétique. En textométrie on utilise plutôt ce vaste ensemble pour en extraire des unités dont la répartition dans le corpus est particulièrement déséquilibrée. Du fait de leur longueur, ces séquences sont, dans l'ensemble, plutôt moins polysémiques que les formes simples isolées de leur contexte immédiat, ce qui facilite grandement l'interprétation des résultats.

Si l'on classe, par contre les lignes de cet inventaire d'après la fréquence de la forme qui suit, la séquence pôle, comme cela a été fait au tableau 4, on s'aperçoit que l'opérateur tous les introduit, la plupart du temps une notion appartenant à un registre négatif *(traîtres, brigands,* etc.*)* même si cette règle subit des exceptions notables[8].

Tableau 5 :

Début de l'inventaire distributionnel des segment répétés
pour la séquence *tous les* dans le corpus *Père Duchesne*.
(classement par ordre de fréquence décroissante de la forme qui suit)

```
871   --- tous les
       32 tous les hommes
       30 tous les traîtres
       29 tous les brigands
       26 tous les départements
       24 tous les ennemis
       21 tous les fripons
       20 tous les bons
       19 tous les scélérats
       15 tous les maux
       14 tous les patriotes
       13 tous les citoyens
       12 tous les bougres
       12 tous les muscadins
       12 tous les peuples
       12 tous les trônes
       11 tous les conspirateurs
       11 tous les coquins
       10 tous les jours
       10 tous les nobles
```

==== *Lexico3* ==== *Segments répétés*

- ✓ Sélectionner l'icône ***Segments répétés*** (4ème icône à partir de la gauche)
- ✓ Sélectionner un seuil de fréquence minimal pour les segments
- ✓ Les segments apparaissent dans un onglet sur la partie gauche. Ils peuvent être triés selon différents critères (longueur, fréquence, ordre lexicographique) en cliquant sur le bandeau situé au-dessus de la colonne correspondante.
- ✓ Chaque sélection, simple ou multiple, réalisée dans la fenêtre des segments peut ensuite être analysée comme un tout, en transitant éventuellement par la fenêtre ***groupe de formes*** à l'aide des différents outils disponible (concordance, histogramme, carte des sections, etc.)

[8] Actuellement, les fonctionnalités de ***Lexico3*** ne permettent pas d'obtenir directement l'état présenté au tableau 5. Cet état a été obtenu en triant, à l'aide d'un tableur (Excel), les lignes du tableau 4.

Tableau 4 :

Début de l'inventaire distributionnel des segment répétés
après la séquence *tous les* dans le corpus *Père Duchesne*.
(classement par ordre lexicographique de la forme qui suit)

```
871 ---- ---- ---- ---- ----  tous les
      2 ---- ---- ---- ----  tous les *brissotins
      7 ---- ---- ---- ----  tous les *français
      3 ---- ---- ---- ----  tous les *jacobins
          17 ---- ---- ----  tous les *sans culottes
               2 ---- ----  tous les *sans culottes à
               3 ---- ----  tous les *sans culottes de
                    2 ----  tous les *sans culottes de *paris
               2 ---- ----  tous les *sans culottes se
      3 ---- ---- ---- ----  tous les aboyeurs
      7 ---- ---- ---- ----  tous les accapareurs
      2 ---- ---- ---- ----  tous les ambitieux
      7 ---- ---- ---- ----  tous les amis
               5 ---- ----  tous les amis de la
                    4 ----  tous les amis de la liberté
      7 ---- ---- ---- ----  tous les aristocrates
               2 ---- ----  tous les aristocrates et les
                    2 ----  tous les aristocrates tous les royalistes
      6 ---- ---- ---- ----  tous les autres
      6 ---- ---- ---- ----  tous les badauds
          4 ---- ---- ----  tous les badauds de
      6 ---- ---- ---- ----  tous les bandits
          2 ---- ---- ----  tous les bandits qui
      2 ---- ---- ---- ----  tous les beaux
      3 ---- ---- ---- ----  tous les biens
     20 ---- ---- ---- ----  tous les bons
               7 ---- ----  tous les bons *sans culottes
                    2 ----  tous les bons *sans culottes se
          7 ---- ---- ----  tous les bons citoyens
          4 ---- ---- ----  tous les bons républicains
     12 ---- ---- ---- ----  tous les bougres
                    2 ----  tous les bougres à poil qui ont
          8 ---- ---- ----  tous les bougres qui
               2 ---- ----  tous les bougres qui ont
      2 ---- ---- ---- ----  tous les boutiquiers
      2 ---- ---- ---- ----  tous les bras
      3 ---- ---- ---- ----  tous les braves
          2 ---- ---- ----  tous les braves bougres
     29 ---- ---- ---- ----  tous les brigands
          19 ---- ---- ----  tous les brigands couronnés
               2 ---- ----  tous les brigands couronnés ce
               3 ---- ----  tous les brigands couronnés et
                    2 ----  tous les brigands couronnés et les
               2 ---- ----  tous les brigands couronnés qui
          2 ---- ---- ----  tous les brigands et
          3 ---- ---- ----  tous les brigands qui
                    2 ----  tous les brouillards de la *tamise se
      2 ---- ---- ---- ----  tous les bureaux
      5 ---- ---- ---- ----  tous les châteaux
               3 ---- ----  tous les châteaux en *espagne
                    2 ----  tous les châteaux en *espagne que
      3 ---- ---- ---- ----  tous les chefs
          2 ---- ---- ----  tous les chefs de
      2 ---- ---- ---- ----  tous les chiens
          4 ---- ---- ----  tous les ci devant
     13 ---- ---- ---- ----  tous les citoyens
      2 ---- ---- ---- ----  tous les coeurs
          5 ---- ---- ----  tous les coins de
      7 ---- ---- ---- ----  tous les complots
          2 ---- ---- ----  tous les complots qu
                    3 ----  tous les complots que l on
     11 ---- ---- ---- ----  tous les conspirateurs
          3 ---- ---- ----  tous les contre révolutionnaires
     11 ---- ---- ---- ----  tous les coquins
          6 ---- ---- ----  tous les coquins qui
          18 ---- ---- ----  tous les coups de
              17 ---- ----  tous les coups de chien
                    2 ----  tous les coups de chien des ennemis
                    4 ----  tous les coups de chien qu
```

Cooccurrences pour un type donné

Si l'on se donne un découpage du corpus en sections (parties, paragraphes, phrases, groupes de phrases) et une forme-pôle (nous prendrons comme ci-dessus l'exemple de la forme : ***proie***) il est possible de constituer la liste des formes et des segments répétés qui trouvent, d'après un calcul statistique particulier[9], un nombre élevé d'occurrence dans les mêmes sections que la forme-pôle. Nous avons trouvé ici : *aux, gibet, oiseaux, perd*. Le retour aux contextes nous confirmera que ces formes entrent avec le pôle choisi dans des associations récurrentes insuffisamment stéréotypées, cependant, pour constituer des segments répétés, du type : *le gibet ne perd jamais sa proie*… etc.

Les calculs de cooccurrences fournissent, de manière symétrique, des listes d'unités textuelles qui trouvent au contraire, toujours d'après le même calcul statistique, très peu d'occurrences au voisinage d'une forme-pôle donnée. On pourrait appeler ces formes des formes *anti-cooccurrentes* ou des formes *évitées* ou *repoussées* par la forme-pôle. L'étude des listes de forme dont les occurrences sont repoussées par la présence dans un contexte proche d'une unité-pôle fixée peut parfois se révéler très instructive.

==== Lexico3 ==== Cooccurrences

- ✓ Demander une carte des sections (7ème icône à partir de la gauche)
- ✓ Choisir un délimiteur de section (paragraphe ou groupe de délimiteurs de phrase . !?)
- ✓ Faire glisser une forme sur la carte à partir du dictionnaire ou de toute autre liste
- ✓ Appuyer sur l'icône des cooccurrences, à l'extrême droite de la 2eme ligne d'icônes
- ✓ Choisir une fréquence minimale et un seuil de probabilité pour les cooccurrents
  NB : si la liste des segments répétés a été préalablement demandée, on obtiendra également les segments jugés cooccurrents spécifiques pour le pôle sélectionné.

Constituer des groupes de formes

On peut constituer des groupes de formes en associant plusieurs types élémentaires, par exemple : le singulier et le pluriel d'un même substantif, les différentes flexions d'un même verbe, les différentes formes d'un adjectif *(nouveau, nouvelle, nouveaux, nouvelles)*[10]. On peut également constituer des groupes à partir de toutes sortes de critères, grammaticaux, sémantiques, etc.

3.4 Les types généralisés (TGen)

Au-delà de ces constructions simples, l'outil *groupe de formes* permet également de constituer des unités qui correspondent au codage d'un thème particulier. Nous avons utilisé cette possibilité pour coder les occurrence d'un thème important chez le *Père Duchesne*, celui de la *mise à mort*. Pour repérer les occurrences de ce thème dans le corpus ***Duchn***, nous avons du

---

[9] Nous utilisons ici un simple calcul hypergéométrique pour comparer le nombre des occurrences du candidat cooccurrent dans les sections ou est attestée la forme-pôle avec sa fréquence dans l'ensemble du corpus. Pour des compléments sur les méthodes de calcul des cooccurrences, cf. par exemple [Lafon XX] et [Heiden XX].

[10] Cette possibilité offerte à l'utilisateur n'implique pas qu'il est toujours utile de rassembler dans tous les cas le pluriel et le singulier d'un même substantif lesquels peuvent avoir des répartitions très différentes dans le corpus. D'autre part le regroupement des types correspondant à l'adjectif *nouveau* mentionné plus haut absorbera également, dans l'état actuel de la fonctionnalité ***groupe de formes***, les occurrences qui correspondent aux formes substantivales *un nouveau, une nouvelle,* etc.

relire attentivement le texte d'un bout à l'autre en nous concentrant sur les seules expressions susceptibles de renvoyer à ce thème[11]

Au delà de la mention des substantifs *guillotine,échafaud, rasoir national*, etc., le recensement des formules susceptibles de constituer des occurrences du thème de la *mise à mort* permet de sélectionner les expression suivantes :

Tableau 2 :

Exemples d'expressions renvoyant au thème de la *mise à mort* sélectionnées d'après une lecture cursive corpus ***Duchn***.

faire jouer X *à la main chaude*
avoir joué *à la main chaude*
(faire) *perdre le goût du pain* (numéro 272)
mettre la tête à la fenêtre (numéro 272)
jouer à la boule (numéro 280)
mettre la tête à la lunette (numéro 286)
(faire) *faire la bascule* (numéro 303)
faire la fatale culbute (numéro 304)
voyager dans la charrette de Samson (numéro 294)
*grimper (*ou *paraître) dans le vis-à-vis de maitre Samson* (numéro 296)
faire le voyage dans la voiture aux trente-six portières (numéro 321)
éternuer dans le sac (numéro 317)
cracher dans le sac (numéro 341)
avoir la tête dans le sac (numéro 304)
faire la grimace au pont rouge (numéro 319)

Il serait totalement déraisonnable d'espérer qu'une telle tâche puisse être confiée à une machine. Par contre, une fois repérées les séquences qui renvoient à ce thème, telle par exemple la séquence *la tête à la fenêtre* il est facile de repérer automatiquement toutes les occurrences du segment répété.

Tableau 3 :

Concordances du segment répété *la tête à la fenêtre* dans le corpus ***Duchn***.

```
 fallait , bon gré , mal gré , mettre la tête à la fenêtre , a tiré de sa manche à
ibunaux pour faire mettre promptement la tête à la fenêtre à la louve autrichienne
çoit pas d ' un pauvre bougre qui met la tête à la fenêtre . § cependant , foutre
t leurs véritables amoureux de mettre la tête à la fenêtre . § convention national
e vont dans cette semaine mettre tous la tête à la fenêtre , et six tribunaux comp
ue le dernier des *brissotins ait mis la tête à la fenêtre , foutre . § la grande
 comme son maître , va bientôt mettre la tête à la fenêtre . § il est donc vrai qu
de la convention , et il mettra aussi la tête à la fenêtre , le roi *coco . § les
punis . pas un conspirateur n ' a mis la tête à la fenêtre . le tribunal révolutio
fin à bon port . l ' ogre royal a mis la tête à la fenêtre , les *brissotins ne so
 pas échappé , et il aurait aussi mis la tête à la fenêtre . § lorsque sa foutue t
chicane , pour les empêcher de mettre la tête à la fenêtre ; mais j ' espère que t
ra pas plus à vous empêcher de mettre la tête à la fenêtre , qu ' elle n ' a pu s
 joie de voir bientôt ce butor mettre la tête à la fenêtre . ses bons avis aux bra
omme son confrère *capet , aurait mis la tête à la fenêtre , si l ' infâme *dumour
 allumer la guerre civile , aient mis la tête à la fenêtre . son grand discours au
ur qu ' elle fasse promptement mettre la tête à la lunette à l ' infâme *brissot ,
que tôt ou tard chacun d ' eux mettra la tête à la lunette comme leur confrère *ca
ps que nous aurions dû voir sa bougre de tête à la lunette . mieux vaut tard que j
```

[11] Notons qu'une bonne connaissance du corpus et de la période concernée peuvent se révéler indispensable pour repérer certaines de ces formules. Ainsi, le fait d'être informé par une source historique possiblement extérieure au corpus, que X a été exécuté dans une période précédente, permet de comprendre la formule *X a craché dans le sac* comme un équivalent de *X a été mis à mort*.

L'ensemble de ces mentions peut être rassemblé en un groupe de forme particulier dont on étudiera ensuite la variabilité au sein du corpus[12].

==== *Lexico3* ==== *Groupe de forme*

- ✓ Sélectionner l'icône ***Groupe de formes*** (8ème icône à partir de la gauche)
- ✓ Donner un nom au groupe (dans la boîte de dialogue supérieure)

Plusieurs possibilités s'offrent alors pour constituer le groupe

- ✓ Sélectionner un par un les constituants du groupe à partir du dictionnaire
- ✓ Utiliser les fonctionnalités génériques « est le début de ce que je recherche » etc.
- ✓ Sélectionner formes segments à l'aide d'une expression rationnelle[13].
- ✓ La flèche rouge située en haut à droite constitue un point d'accroche pour l'ensemble du groupe ainsi constitué. Elle peut être *traînée* vers tous les outils qui acceptent un TGen.

4 Etude la distribution d'un type

4.1 Les outils de base

L'outil concordances

L'outil ***concordances*** permet de rassembler toutes les occurrences relatives à un type donné en les munissant d'un petit fragment de contexte et de les trier selon différents critères, cf. tableau 1. En faisant varier la taille du contexte, l'ordre de présentation (ici les contextes sont triés en fonction de la forme qui suit le pôle sélectionné). A l'aide de cet outil, le chercheur peut opérer des rapprochements qu'une lecture cursive du texte ne lui aurait sans doute pas permis de saisir (ici, par exemple : *perdre sa proie* et *sa proie lui échappe*).

Tableau 1 :

Concordance de la forme *proie* dans le corpus ***Duchn***

pendant quelques instants ces oiseaux de	**proie**	avaient disparus , foutre , et depuis que
ès avoir rogné les ongles des oiseaux de	**proie**	de la finance ; après avoir détruit la mé
amée qui rugit quand on lui a arraché sa	**proie**	, elle poussait des cris affreux . " ains
fuite , mais le gibet ne perd jamais sa	**proie**	, et tôt ou tard les pigeons reviendront
e te dire que le gibet ne perd jamais sa	**proie**	? il y a plus de dix ans que tu aurais fa
er numéro que le gibet ne perd jamais sa	**proie**	. le jean - foutre est hors de la loi ,
s ' entre - déchiraient pour avoir leur	**proie**	, les *sans - culottes se fortifiaient ,
ut tout dévorer , tout engloutir ; si sa	**proie**	lui échappe , il devient enragé , et il
' examine ce tigre qui rugit de voir sa	**proie**	lui échapper . " me voilà au bout de mes
' aux tigres et aux ours de déchirer la	**proie**	qui tombe sous leurs griffes ; ils regard

[12] L'esquisse de procédure ainsi décrite ne garantit pas totalement que l'on a intégré aux comptages ***toutes*** les occurrences du textes susceptibles de relever du thème choisi. Un autre chercheur confronté au même texte disposant d'autres connaissances aurait peut-être inclus (ou exclu) d'autres occurrences susceptibles de modifier les comptages d'ensemble.

[13] Cf. sur ce point le manuel d'utilisation de ***Lexico3*** pg **xxxxxxxx**.

.

==== *Lexico3* ==== *Concordances*

- ✓ Sélectionner l'icône ***Concordances*** (3ème icône à partir de la gauche) et
- ✓ Entrer une forme dans la boite de dialogue *forme* ***(ex : proie)***
- ✓ Choisir l'ordre de présentation des contextes (Tri = après, avant, ordre du texte)
- ✓ Choisir [éventuellement] un regroupement par parties (si une partition a été sélectionnée)

L'outil statistiques par parties

L'outil ***statistiques par parties*** permet de juger de la répartition des occurrences relevant d'un même type dans les différentes parties d'une partition, cf. figure 2.

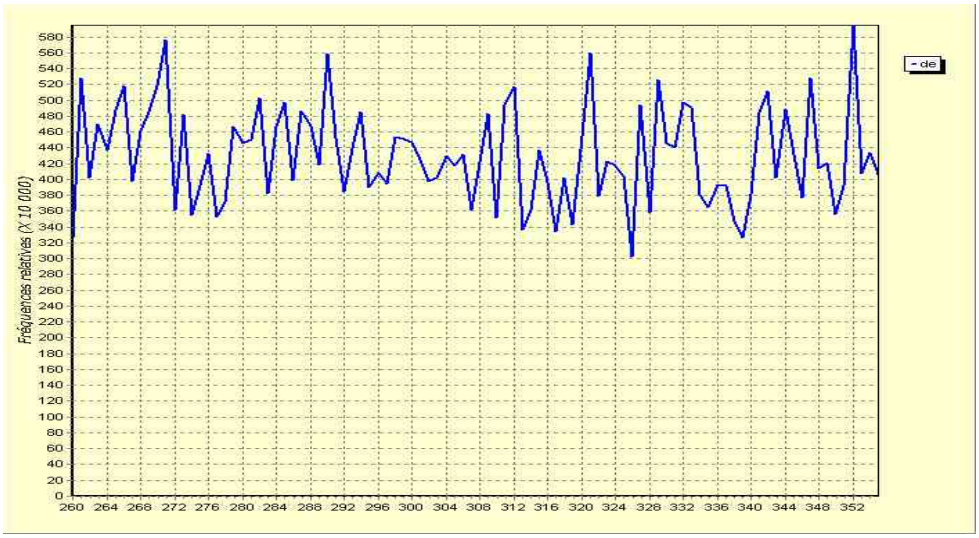


Figure 3 :
Ventilation des occurrences de la forme *de* en fréquence relative
dans les 96 numéros du corpus ***Duchn***.

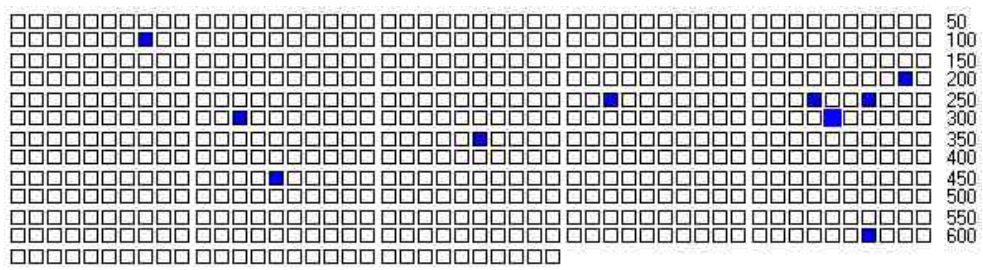
==== *Lexico3* ==== *Statistiques par parties*

- ✓ Sélectionner l'icône ***Statistiques par parties*** (5ème icône à partir de la gauche)
- ✓ Choisir le type de clé qui déterminera la partition active du corpus
- ✓ Faire glisser une forme à partir du dictionnaire ou de toute autre liste ***(ex : proie)***

L'outil carte des sections

L'outil *carte des sections* permet une visualisation globale de la répartition des occurrences qui relèvent d'un type donné dans l'ensemble du corpus. Chacun des carrés représente un élément particulier du texte découpé en sections. On a décidé, pour établir la carte présentée à la figure 4, de représenter chacun des paragraphes du texte, repérable, grâce à notre codage préalable, à ce qu'il s'ouvre sur un caractère §. La sélection à l'aide de la souris, d'un paragraphe particulier provoque son affichage dans une fenêtre située

sous la carte des sections. Comme on le verra plus loin (§ XX), il est possible, de matérialiser une partition sur ce type de carte.



[§ 295] te voilà enfin sur la fatale sellette , infâme *brissot ; quand je te prédisais que tu ferais une mauvaise fin , n ' étais je pas un bon prophète ! n ' avais je pas raison de te dire que le gibet ne perd jamais sa **proie** ? il y a plus de dix ans que tu aurais fait la grimace au pont *rouge , si l ' on t ' avait rendu justice , mais tes escroqueries d ' alors , toutes les bourses <Epg=4>que tu coupais dans l ' ancien régime , n ' étaient que des coups d ' essai des brigandages que tu as fait pendant la révolution.

Guide de lecture pour la figure 4

Chacun des carrés de la séquence du haut représente un des paragraphes du texte original. La numérotation de la colonne de droite permet de calculer le numéro de chaque paragraphe. La forme *proie* a été traînée sur la carte des sections provoquant le coloriage des sections ou elle est attestée. Le texte du paragraphe 295 sélectionné par l'utilisateur est affiché en bas de la figure. Les occurrences de la forme sélectionnée y sont mises en évidence.

Figure 4 :

Localisation des occurrences de la forme *proie*
sur une carte des sections du corpus ***Duchn***.

==== *Lexico3* ==== *Carte des sections*

- ✓ Sélectionner l'icône ***Carte des sections*** (5ème icône à partir de la gauche)
- ✓ Choisir un délimiteur de section qui servira à construire la carte
- ✓ Faire glisser une forme sur la carte à partir d'une liste ***(ex : proie)***
- ✓ Choisir [éventuellement] un regroupement par parties, si une partition a été sélectionnée

Intermède – utilisation de la partition en pages

La clé <*Epg=x*> ou x prend les valeurs 1, 2, 3, … , 8 permet de repérer les changements de page à l'intérieur de chaque numéro.[14] Comme c'est le cas pour chaque type de clé, il est possible d'utiliser la fonctionnalité ***Partition*** de *Lexico3* pour constituer, à partir de cette clé, un corpus en 8 parties. La partition réalisée à partir de la clé ***Epg*** rassemble donc en une même partie toutes les premières pages de chacun des 96 numéros, la seconde partie est composée de toutes les secondes pages et ainsi de suite jusqu'à la huitième partie qui rassemble les dernières pages de chaque numéro.

[14] Le contenu de la clé Epg : x – prend des valeurs de 1 à 8, car la publication, une grande feuille imprimée pliée en quatre par la suite est toujours composée de 8 pages.

Quel peut-être l'intérêt d'une telle partition au plan textométrique ?

Ce découpage du corpus, un peu curieux au premier abord, permet de mettre en évidence une particularité intéressante dans l'utilisation du vocabulaire. Comme on peut le voir sur la figure 5, la fréquence de la forme *foutre*, assez faible dans la première page, se maintient à un niveau stable dans les pages intérieures pour croître brutalement à l'intérieur de la dernière page. Ce déséquilibre traduit à coup sûr un procédé récurrent employé par l'auteur dans la conclusion de son périodique.

Une hypothèse explicative se présente immédiatement au vu de cette ventilation que des recherches ultérieures viendront conforter par la suite : la forme *foutre*, juron favori du ***Père Duchesne*** est utilisée assez modérément dans l'introduction de chaque livraison, sa fréquence relative reste stable dans les pages intermédiaires mais la conclusion du journal se fait sur un style plus « musclé » qui recourt largement à l'emploi de jurons et d'invectives. La visualisation des occurrences de *foutre* sur la carte des sections permet de localiser facilement des exemples de cette utilisation particulière.

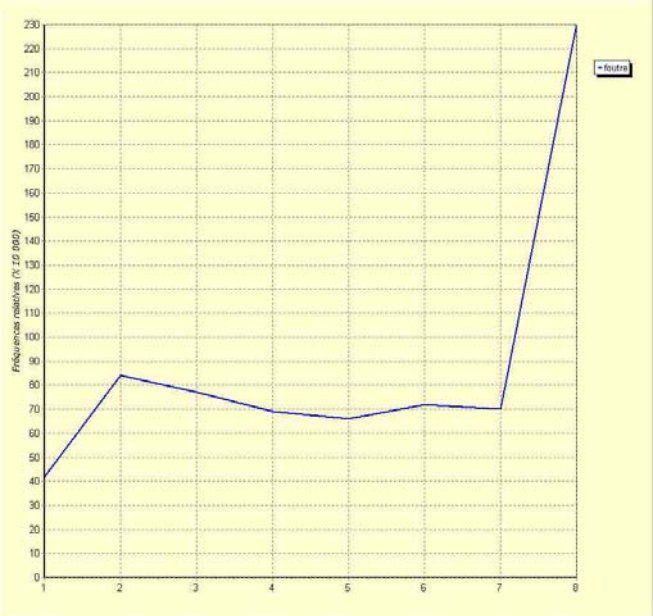


Figure 5

Ventilation des occurrences de la forme *foutre* dans les 8 pages du journal
(le numéro de page figure en abscisse sur le graphique)

On trouvera ci-dessous un exemple, parmi beaucoup d'autres possibles, d'une séquence prélevée dans la page qui clôt le *numéro 347* du corpus.

> **Numero 347** <Epg=8> imposture. ainsi donc, **foutre**, vive la raison vivent la vérité et l'humanité ! au **foutre** les prêtres, qui ne savent que mentir, tromper, voler et égorger, **foutre.**

L'analyse du vocabulaire spécifique de cette huitième partie nous permettra de dégager un ensemble de formes qui obéissent à ce même schéma d'utilisation : *vive, vos, soyez, peuple,* etc. En résumé, les résultats de cette expérience qui n'avait au départ d'autre finalité que celle de vérifier le fonctionnement correct du logiciel nous ont suggéré une possibilité d'exploration textométrique à laquelle nous n'avions pas pensé au départ. La mise en œuvre extrêmement simplifiée de la division du corpus en partie permet, on le voit, d'entreprendre à peu de frais, des expériences dont les résultats peuvent se révéler intéressants.

5 Méthodes textométriques

Plusieurs méthodes statistiques permettent d'éclairer la structure d'un corpus textuel à partir de comparaisons réalisées entre les fragments du corpus. La partition du corpus constitue une étape très importante dans l'analyse comparative des textes dans la mesure où les oppositions qu'il sera possible de mettre en évidence entre les parties soumises à comparaison dépendent étroitement du choix de la partition initiale.

Tableau 6 :

Tête du tableau lexical constitué par le décompte des 30 formes les plus fréquentes du corpus dans les 8 parties d'une partition en 8 mois

Forme	P1	P2	P3	P4	P5	P6	P7	P8
0 de	886	875	853	753	746	757	669	591
1 les	641	687	569	549	534	687	555	526
2 la	641	550	593	579	479	502	484	447
3 et	449	524	480	530	463	467	461	399
4 le	348	376	398	374	355	321	327	265
5 à	382	384	350	361	319	308	266	262
6 que	349	390	351	298	317	287	287	217
7 qui	222	276	310	261	271	268	267	204
8 des	262	262	240	201	245	243	274	217
9 il	300	233	285	199	248	229	203	128
10 l	221	260	250	236	209	201	206	187
11 pour	214	249	252	220	194	181	183	172
12 en	171	199	153	192	167	169	147	111
13 qu	184	189	225	164	184	139	115	98
14 d	170	158	170	151	162	161	152	150
15 nous	180	250	167	132	155	100	182	104
16 un	156	190	155	157	186	157	135	131
17 est	162	137	171	142	181	190	134	115
18 tous	163	176	150	140	116	156	149	145
19 ils	145	181	147	137	183	104	160	130
20 ne	159	170	168	128	181	124	129	106
21 du	164	143	161	138	145	137	123	107
22 foutre	149	141	141	103	124	126	151	165
23 pas	140	189	187	125	160	103	105	91
24 vous	157	189	140	219	86	99	135	72
25 je	111	125	97	131	146	152	132	85
26 n	129	162	146	110	124	100	101	83
27 dans	129	133	128	123	115	96	111	89
28 on	87	153	110	77	151	137	82	59
29 a	118	119	131	106	95	118	81	66
30 plus	115	99	101	77	107	94	124	81

Le Tableau lexical

On commence par constituer un tableau qui compte autant de ***colonnes*** que la partition choisie compte de parties et autant de ***lignes*** que le vocabulaire du corpus compte de formes différentes. A l'intersection de la ligne *i* et de la colonne *j*, on notera le nombre d'occurrences que la forme *i* trouve dans la partie *j*, du corpus. Le tableau 6 présente les 30 premières lignes du *tableau lexical* réalisé à partir d'une partition du corpus ***Duchn*** en 8 parties dont chacune correspond à un mois de parution du journal[15].

[15] Un fichier coran.don est créé par ***Lexico3*** qui contient le tableau lexical, précédé de quelques paramètres nécessaires aux analyses multidimensionnelles.

Cette petite partie extraite du tableau lexical (8 parties x 11 070 formes) permet d'imaginer la difficulté qu'il y aurait à essayer d'analyser un tel tableau. Cependant, plusieurs méthodes statistiques permettent d'extraire de ces tableaux des faits particulièrement remarquables sur lesquels il est pratique de concentrer son attention dans une première approche. Pour ces méthodes et pour les machines qui les mettent en œuvre, la dimension des tableaux lexicaux ne constitue pas de difficulté particulière.

La division en 96 parties, numérotées de 260 à 355 selon la numérotation originale de la publication, paraît a priori la division la plus *naturelle* du corpus ***Duchn***. La clé <*numéro=x*> introduite lors du codage du corpus permet de réaliser cette partition en 96 numeros. Nous allons étudier cette partition en combinant deux méthodes d'analyse statistiques très complémentaires et couramment utilisées en textométrie : l'analyse factorielle des correspondances (AFC)et l'analyse des spécificités.

5.1 Etude de la partition du corpus Duchn en 96 numéros

On trouve sur la figure 6, une représentation de l'ensemble des 96 numéros fournie par l'analyse factorielles des correspondances à partir du tableau (96 numéros x 1420 formes de fréquence supérieure à 10).[16]

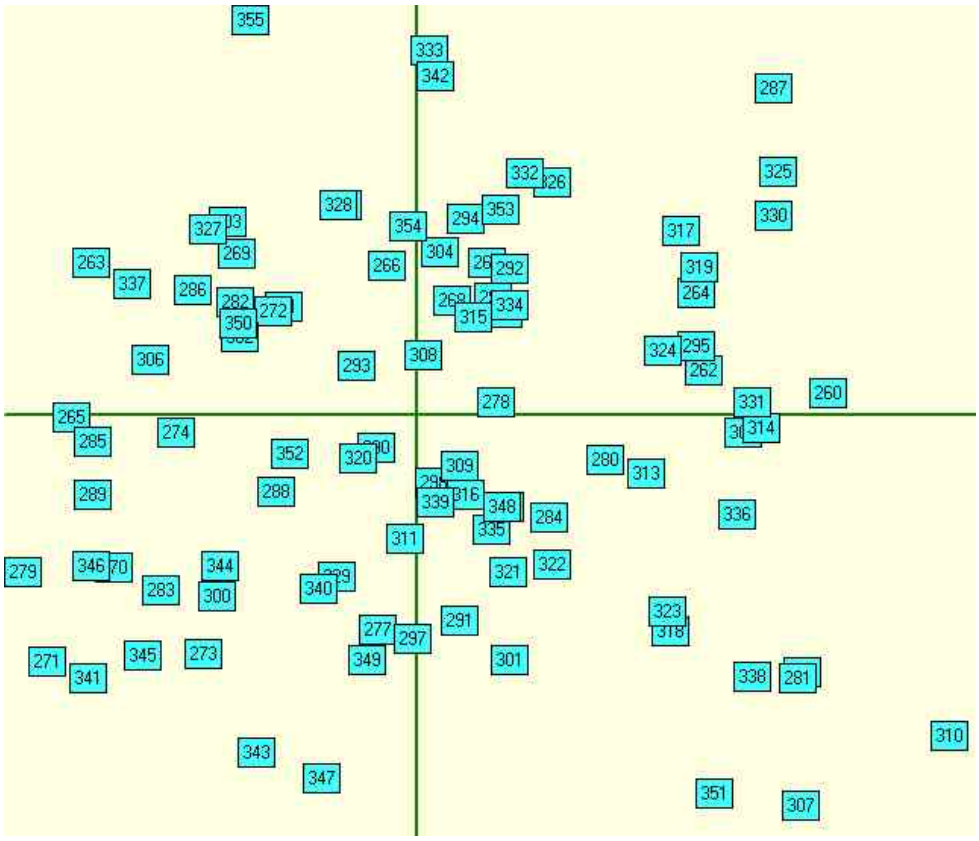


Figure 6
AFC sur le corpus ***Duchn***
96 numéros x 1420 formes de fréquence ≥10

La représentation proposée par l'AFC ne permet pas de repérer une quelconque évolution chronologique des parties. Pour tenter de comprendre les bases de l'opposition qui oppose les différents numéros opposés par le premier axe, nous pouvons consulter les longues listes de

[16] Les pourcentages d'inertie attachés aux deux premiers axes factoriels responsables de la représentation que l'on trouve au tableau 6, sont respectivement égaux à : τ_1=3%, τ_2=2%.

contributions aux facteurs fournis par les programmes d'AFC. Nous allons employer une méthode plus simple pour arriver à un résultat très proche.

==== Lexico3 ==== Analyse Factorielle des Correspondances (AFC)

- ✓ Vérifiez que vous avez opéré au moins une partition du corpus (cf. §xx)
- ✓ Sélectionner l'icône **PCLC** (5ème icône à partir de la gauche)
- ✓ Sélectionner une partition du corpus (ici : numero)
- ✓ Appuyez sur le bouton AFC ((à droite de l'écran)
- ✓ Choisissez un seuil de fréquence minimale (ou acceptez le seuil 10 proposé par défaut)
- ✓ Lancez l'analyse en appuyant sur le bouton **OK**

==== Repères méthodologiques ====

L'analyse factorielle des correspondances (AFC)

L'*analyse factorielle des correspondances* est une méthode statistique qui s'applique aux tableaux de contingence, tels par exemple les tableaux résultant du décompte de différents *types* de vocabulaire (lignes du tableau) dans les différentes *parties* (colonnes du tableau) d'un corpus de textes.

On commence par calculer une distance (dite *distance du chi-deux*) entre chacune des paires de textes qui constituent le corpus.

On décompose ensuite ces distances sur une succession hiérarchisée d'*axes factoriels*. La propriété remarquable de ce système d'axes factoriels est que les représentations limitées aux premiers axes de ce système sont celles qui déforment *le moins possible* les distances calculées entre chaque paire d'éléments. Des *pourcentages d'inertie*, dont la somme vaut 100, calculés pour chaque axe permettent d'apprécier la quantité d'information apportée par chacun des axes dans la décomposition.

Cette méthode d'obtenir des représentations synthétiques portant à la fois sur les distances calculées entre les textes et celles que l'on peut calculer entre les unités textuelles qui les composent. Les typologies obtenues sur chacun des deux ensembles mis en correspondance, sont intiment liées et peuvent être mise en relation grâce à des *représentations simultanées* sur les premiers axes factoriels.

L'intérêt principal de l'AFC réside dans sa capacité à extraire à partir de vastes tableaux de données difficilement appréhendables des structures simples qui rendent compte approximativement des grandes oppositions sous-jacentes dans un corpus de textes.

Pour en savoir plus :

Lebart, L., Salem, A. : Statistiques textuelles, Paris, Dunod, 1994.

5.2 Analyse des spécificités du corpus

L'analyse des spécificités permet de porter un diagnostic exprimé en probabilité sur l'effectif de chacune des cases d'un tableau lexical. [17]

==== Repères méthodologiques ====

La méthode des spécificités

A partir de l'effectif constaté à l'intersection de la ligne *i* et de la colonne *j* (le nombre d'occurrences de la forme *i* dans la partie *j*), étant donnés la fréquence totale de la forme F_i la longueur de la partie t_j et l'effectif total T, la méthode permet de tirer des conclusions sur l'effectif observé. Dans certains cas, la conclusion est que l'effectif observé correspond *à peu près* à ce que le modèle permettait de prévoir. On dira alors que la répartition de la forme est *banale* pour cette partie. Dans d'autres cas, le modèle amènera à conclure que l'effectif observé s'éloigne notablement des prévisions que l'on pouvait faire sous les hypothèses admises par le modèle.

On appelle *spécificités positives* les effectifs qui dépassent largement ce que le modèle laissait prévoir et *spécificités négatives* les effectifs qui se révèlent nettement inférieurs à ce que ce même modèle permettait d'espérer. On attache à ces diagnostic un *indice de spécificité* [18] qui permet de mesure les écarts constatés par rapport à ce que le modèle laissait prévoir. Plus ce diagnostic est élevé plus l'écart est jugé significatif par le modèle.

On peut étendre le calcul décrit ci-dessus pour les unités simples aux segments répétés d'un texte si l'on remarque que les occurrences d'un segment AB (ou A et B sont des formes simples) peuvent être vues comme un sous-ensemble des occurrences de la forme A pour lesquelles B succède immédiatement à A dans le texte. Le calcul simultané des spécificités sur les ensembles de formes et de segments répétés d'un même texte permet souvent de mettre en évidence des associations spécifiques composées de plusieurs formes dont les répartition particulières n'entraînent pas de diagnostic particulier.

Pour en savoir plus :

Pour un exposé et des exemples d'application de l'analyse des spécificités à l'étude des corpus de textes, on consultera par exemple :

Lafon, P. : *Dépouillements et analyses statistique en lexicométrie*, Paris, Klincksieck, 1984

Lebart, L., Salem, A. : *Statistiques textuelles*, Paris, Dunod, 1994.

[17] L'analyse des spécificités repose sur l'utilisation du modèle hypergéométrique pour l'analyse des tableaux de nombres à deux dimensions. Pour plus de détails sur le modèle des spécificités et ses applications à l'étude des corpus textuels, on consultera : [Lafon 1984] ou [Lebart et Salem 1994].

[18] Pour une spécificité *positive* et un effectif observé égal a *k*, un indice de probabilité ***x*** signifie que le modèle attache au phénomène constaté : effectif égal ou supérieur à *k*, une probabilité de l'ordre de 10^{-x}. Pour une spécificité *négative* cette probabilité s'attache à un effectif inférieur ou égal à *k*.

Pour comprendre l'opposition constatée sur le premier axe de l'AFC, on a calculé les spécificités, par rapport à l'ensemble du corpus, de deux groupes de numéros opposés par le premier facteur. Chacun des deux groupe est composé des 20 numéros les plus éloignés du centre sur la droite et sur la gauche du graphique. Les spécificités majeures pour chacun de ces groupe ont été rassemblées au tableau 6. L'analyse de ces listes nous fournira une piste pour expliquer la différence qui existe entre les deux groupes de textes.

Tableau 6 :

Formes et segments spécifiques positifs majeurs
pour les numéros opposés par l'AFC sur les 96 numéros

Spécificités positives de la partie gauche				***Spécificités positive de la partie droite***			
Forme	**Frq.Tot.**	**Partie**	**Coeff.**	**Forme**	**Frq. Tot.**	**Partie**	**Coeff.**
nous	1270	449	29	*je*	979	436	***
vous	1097	395	27	*me*	329	184	43
avez	171	94	21	*tu*	296	142	25
fermiers	28	24	13	*ma*	132	81	24
constitution	72	44	13	*m*	206	102	20
accapareurs	80	45	12	*moi*	144	80	20
est vous	24	21	12	*mon*	193	95	18
nos	348	132	12	*j*	281	123	18
vous avez	75	43	12	*ai*	202	91	14
c est vous	24	21	12	*me dit*	29	24	13
vous qui	42	28	10	*que je*	119	58	12
les	4748	1210	10	*dit*	163	72	11
subsistances	47	28	9	*j ai*	123	59	11
la constitution	40	26	9	*que j*	52	30	9
c est vous qui avez	10	10	8	**phélipotin*	13	12	8

Guide de lecture pour le tableau 6

Dans chacun des volets du tableau, on trouve les spécificités relatives à l'un des groupes de textes séparés par l'AFC.

- La première colonne du tableau indique le terme pour lequel le diagnostic de spécificité a été calculé ;
- la seconde *Frq. Tot.* donne la fréquence du terme dans l'ensemble du corpus ;
- la troisième *Partie* la fréquence de ce même terme dans la partie considérée ;
- la troisième *Coefft.* donne lle coefficient de spécificité calculé pour le terme.

Sur la partie droite du tableau 6 on trouve des formes comme *je, tu, me moi, mon* caractéristiques du dialogue, à gauche les contextes des formes comme *vous* renvoient moins au dialogue qu'a des monologues. On note également la présence de nombreux substantifs. Une analyse plus poussée de ces listes accompagnée de retours fréquents au contexte nous amènerons à la conclusion que l'écriture du *Père Duchesne* fait appel à deux types d'écritures distincts dans des proportions qui varient tout au long des huit mois sur lesquels s'étale le corpus et à l'intérieur de chaque numéro. Certains numéros relèvent plus particulièrement

d'un genre que nous appelons "parade"[19], caractérisé par la présence de nombreux effets scéniques empruntés au théâtre de foire, les autres sont de facture rhétorique plus classique. On trouve ci-dessous deux brefs extraits qui illustrent cette opposition :

Tableau 7 :
Deux extraits du corpus ***Duchn*** illustrant la différence
entre les genres *parade* et *classique*

Père Duchesne n°260 (exemple du genre « facture classique »)

```
§ *marat n'est plus, foutre. peuple, gémis, pleure ton meilleur
ami;il meurt martyr de la liberté. c'est le *calvados qui a vomi
le monstre sous les coups duquel il vient de périr. une jeune
fille, ou plutôt une furie armée par les prêtres, et pénitente,
dit on, du cafard *fauchet ,part de *caen  pour exécuter cet
horrible attentat.
```

Père Duchesne n°262 *(*exemple du genre « parade »)

```
§ voilà donc tes projets, infâme coquin; avais je tort, quand je
foutais mes fourneaux sens dessus dessous, quand je brisais ma
pipe toutes les fois que l'on m'annonçait qu'un noble avait été
nommé à quelque place importante.
tu ne savais pas en défilant ton chapelet, archi-traître, que tu
parlais au *père *duchesne? à moi mes gens, à moi mes aides de
camp /…/
```

C'est cette alternance dans le style d'écriture qui explique pour l'essentiel l'opposition constatée sur le premier axe de l'AFC. Cette opposition intéressera sans doute à la fois les spécialistes de stylistique et les historiens qui étudient de près la rhétorique du *Père Duchesne*, cependant nos préoccupations plus centrées sur l'évolution du vocabulaire dans cette période nous ont entraînés à nous intéresser à des partitions regroupant plusieurs numéros consécutifs. De tels regroupements permettent de neutraliser les différences stylistiques opposant les livraisons que nous venons d'entrevoir et d'orienter les analyses vers l'observation des changements qui surviennent au cours du temps dans l'utilisation du vocabulaire.

==== ***Lexico3*** ==== ***Liste des spécificités pour une partie (ou un groupe de parties)***

- ✓ Sélectionner l'icône ***PCLC*** (5 ème icône à partir de la gauche)
- ✓ Sélectionner une partie ou un groupe de parties
- ✓ Appuyer sur le bouton ***Spécifs*** (à droite de la fenêtre)
- ✓ Les résultats apparaissent dans une fenêtre sur la gauche
- ✓ On obtient également les segments répétés spécifiques si la liste des segments répétés a été construite avant l'appel des spécificités (cf. §2.2).
- ✓ .On peut également appeler cette fonctionnalité en sélectionnant une ou plusieurs parties sur les plans factoriels produits par l'Afc ou des zones de texte de la carte des sections.

[19] A la suite de J. Guilhaumou [Guilhaumou 19xx].

6 Conclusion

L'exploration du corpus ***Duchn***, à l'aide des méthodes textométriques met en évidence une importante évolution du vocabulaire au cours des huit mois sur lesquels s'étend le corpus.

Les analyses quantitatives sur la partition en 96 livraisons, mettent en évidence des différences stylistiques liées à une alternance de genre entretenue par l'auteur du corpus. De ce fait, elles ne permettent pas d'apprécier l'évolution lexicale du corpus.

Un regroupement des livraisons en périodes de 30 jours consécutifs permet par sa part de cerner l'évolution lexicale de manière nettement plus satisfaisante. Les méthodes quantitatives permettent alors tout à la fois : de mettre en évidence un vocabulaire offensif qui trouvera un emploi particulièrement remarquable dans la période M6. Le retour au contexte permet de préciser ces analyses.

7 Références

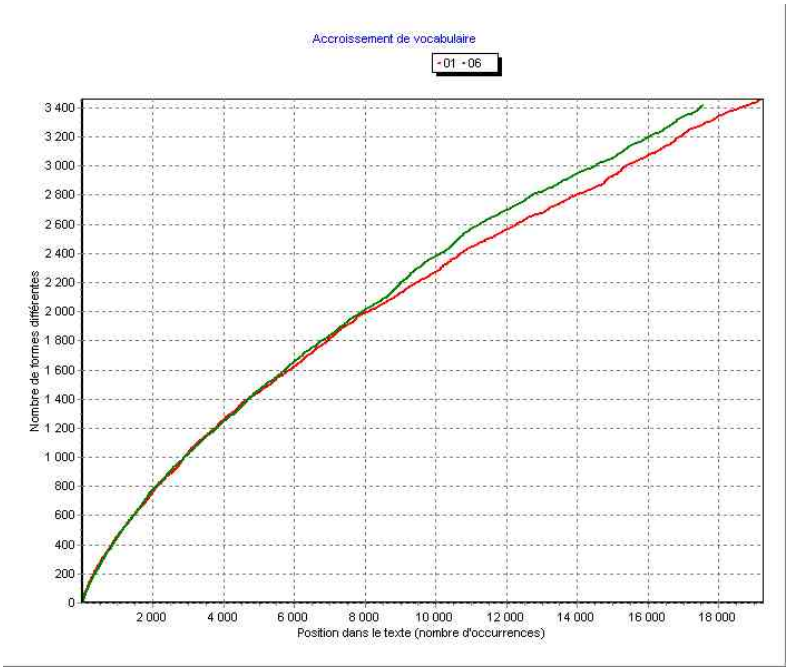
Lamalle C., Salem A., « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *actes des 6emes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002

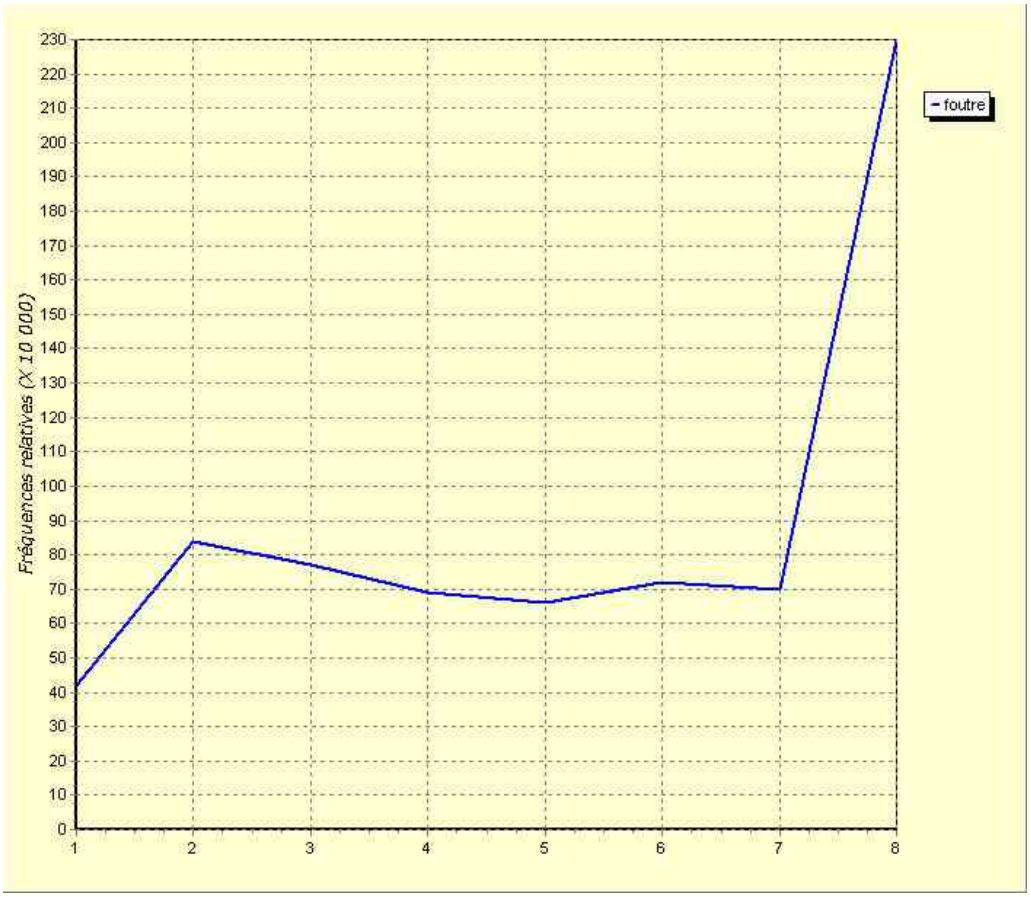
http://www.cavi.univ-paris3.fr/lexicometrica, 1997

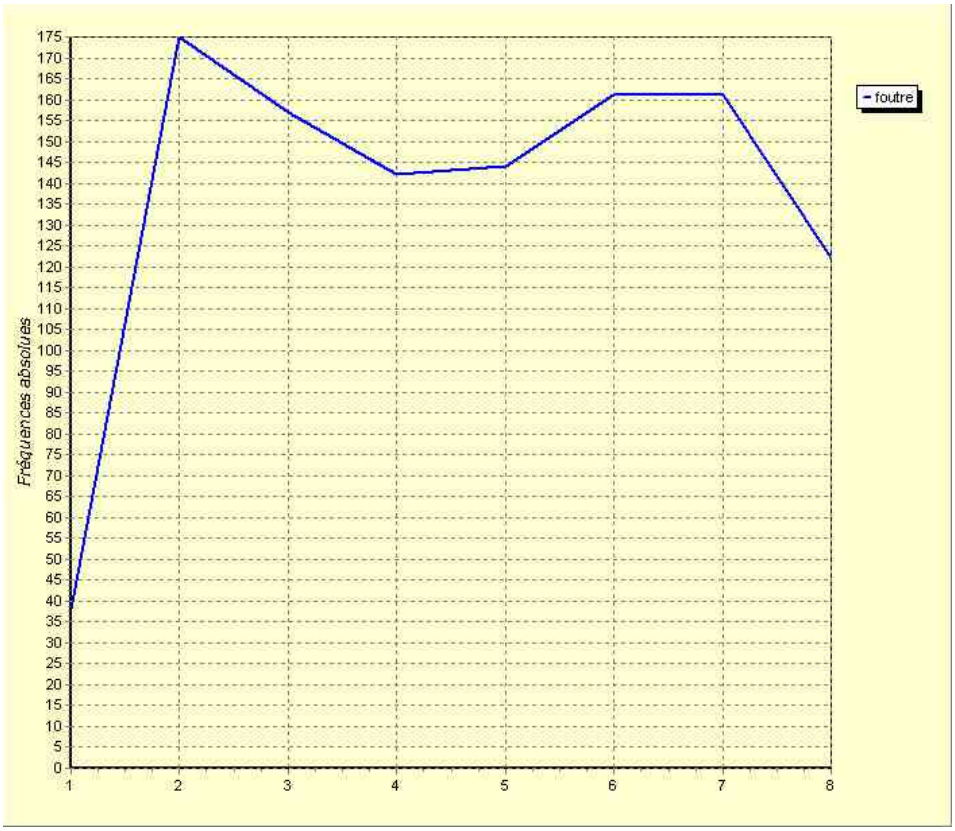
8 Principales fonctionnalités *Lexico3* utilisées

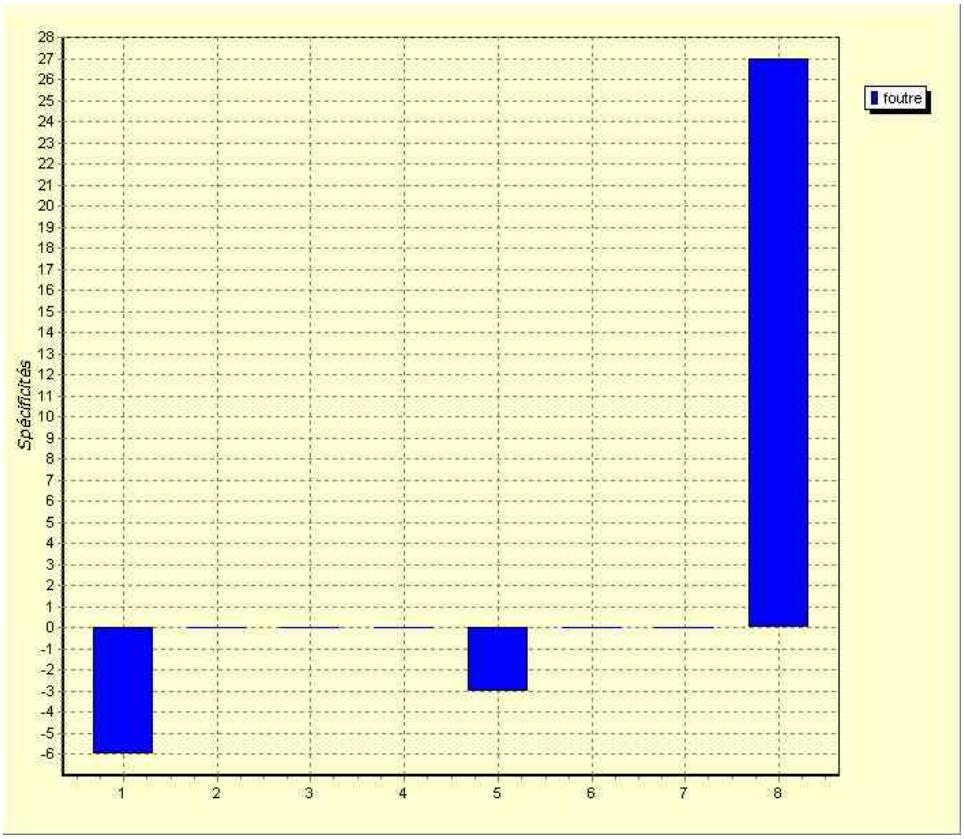
N°	*Fonctionnalité*	Résultat
2	Partition (clé a, pour année)	
5	Principales car lexicom (PCLC)	*Tableau 2*
5.6	Accroissement du vocabulaire (corpus)	*Figure 1*
5.6	Accroissement du vocabulaire (P92, P93)	*Figure 2*
4	Segments Répétés (seuil minimal =2)	
8	Sélection d'un Type (occurrence de SR long>10)	
7	Carte des sections (paragraphes, présence SR de long>10)	*Figure 3*

Annexe


Accroissement de vocabulaire
- 01 - 06
Nombre de formes différentes
3 400
3 200
3 000
2 800
2 600
2 400
2 200
2 000
1 800
1 600
1 400
1 200
1 000
800
600
400
200
0
2 000
4 000
6 000
8 000
10 000
12 000
14 000
16 000
18 000
Position dans le texte (nombre d'occurrences)


Fréquences relatives (X 10 000)
230
220
210
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
foutre





91 ---- ---- ---- ---- ---- les patriotes

```
89 ---- ---- ---- ---- ----  les hommes
88 ---- ---- ---- ---- ----  les plus
86 ---- ---- ---- ---- ----  les traîtres
75 ---- ---- ---- ---- ----  les aristocrates
66 ---- ---- ---- ---- ----  les autres
64 ---- ---- ---- ---- ----  les fripons
63 ---- ---- ---- ---- ----  les brigands
60 ---- ---- ---- ---- ----  les jean
58 ---- ---- ---- ---- ----  les ennemis
54 ---- ---- ---- ---- ----  les départements
46 ---- ---- ---- ---- ----  les bons
42 ---- ---- ---- ---- ----  les accapareurs
40 ---- ---- ---- ---- ----  les scélérats
37 ---- ---- ---- ---- ----  les uns
37 ---- ---- ---- ---- ----  les *français
37 ---- ---- ---- ---- ----  les *brissotins
35 ---- ---- ---- ---- ----  les rois
33 ---- ---- ---- ---- ----  les bougres
32 ---- ---- ---- ---- ----  les muscadins
31 ---- ---- ---- ---- ----  les riches
31 ---- ---- ---- ---- ----  les meilleurs
31 ---- ---- ---- ---- ----  les intrigants
30 ---- ---- ---- ---- ----  les prêtres
29 ---- ---- ---- ---- ----  les royalistes
```

Par page

22 foutre 38 175 157 142 144 161 161 122

Insécurité et élections présidentielles dans le journal Le Monde

[Presse]

Emilie Née

emilienee@wanadoo.fr

Résumé : En 2001-2002, pendant la campagne pour les élections présidentielles françaises, le mot *insécurité* a joué un rôle souvent dénoncé par la suite dans la structuration du débat politique. Comment analyser l'emploi de la forme *insécurité* dans le journal *Le Monde* pendant cette campagne électorale, emploi qui va d'abord se caractériser par une densification de fréquence ? Cette exploration textométrique sur un grand corpus médiatique constitué autour d'une forme-pôle permettra de repérer plusieurs phénomènes discursifs à l'œuvre dans le journal *Le Monde* et de lever le jour sur certains problèmes d'interprétation liés à la nature même de ce corpus.

1. Le corpus Monde/Insécurité

Le corpus ***Monde/Insécurité*** est composé de l'ensemble des articles publiés entre le 1er juillet 2001 et le 1er juillet 2002 qui contiennent le mot *insécurité* (965 articles). Ce corpus s'étend sur une période qui englobe la campagne électorale des présidentielles de 2002. Cette campagne qui s'achève début mai 2002 est suivie par une autre campagne pour l'élection d'un parlement qui sera élu le 16 juin 2002.

Le corpus ***Monde/Insécurité*** est d'abord divisé en 13 parties qui correspondent chacune à une période d'un mois. Ce corpus constitue donc une **série textuelle chronologique**[20]. Un balisage systématique du corpus en rubriques, articles, jours, permet d'affiner l'analyse des périodes considérées.

Tableau 1 :
Principales caractéristiques lexicométriques

Nombre des occurrences	867561
Nombre des formes	37456
Fréquence maximale	44194
Hapax	15230
Nombre d'occurrences de la forme insécurité	1705

Tableau 2 :
Extrait d'un article paru après le second tour des élections présidentielles (5 mai 2002)

Extrait du corpus ***Monde/Insécurité***

```
<mois=11-mai2002>
```

[20] Par *série textuelle chronologique*, on entend « l'échantillonnage au cours du temps d'une même source textuelle sur une période plus ou moins longue » (Lebart et Salem 1994 : 217). Voir également les récentes analyses de corpus de veille de S. Fleury (http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm).

<rub1=supplementtelevision>

<date=020511>

où est passée l'**insécurité** ?

la question de l'**insécurité** et de son traitement à la télévision avant les élections présidentielles n'a pas fini de faire parler d ' elle…

etienne mougeotte, interrogé par « le monde télévision » (daté 4 mai), se plaint d'être un bouc émissaire. je cite : « (…) si tf1, comme tous les grands médias, a longuement traité de l'**insécurité**, c'est simplement que nous nous efforçons d'être à l' écoute de nos concitoyens et de répondre à leurs attentes. ce n'est pas la télévision qui génère l'**insécurité**, c'est la montée de l'**insécurité** qui justifie que la télévision en parle. « il est probable que certains entendront ce curieux syllogisme de la façon suivante : 1. les français sont préoccupés par l'**insécurité**. 2. les médias veulent plaire aux français. 3. donc l'**insécurité** s'accroît (mais la télévision n'en est pas responsable)… etienne mougeotte peut penser ce qu'il veut et éventuellement prendre ses téléspectateurs pour des imbéciles… je ne regarde pas tf1. mais les infos de france 2 et france 3 sont, de ce même point de vue, caricaturales. avant le premier tour de l'élection présidentielle, nous y entendions chaque jour le thème de l'**insécurité** abordé sous divers aspects. à chaque journal, le thème de l'**insécurité** était énoncé en titre, abordé et développé avec des « informations » sur les banlieues, les voitures brûlées, le procès de patrick dils, les suites de la tuerie de nanterre, l'agression du « papy » d'orléans, etc. pas un journal sans que le mot « **insécurité** » soit prononcé et répété plusieurs fois. depuis le 21 avril, un calme étrange est apparu, comme si les banlieues s'étaient soudain apaisées et que les voyous avaient disparu : on n'entend plus parler d'**insécurité** dans les journaux télévisés. […]

La ventilation[21] des fréquences de la forme *insécurité* sur cette partition chronologique va mettre à jour un phénomène de densification qu'il va s'agir de décrire précisément.

2. Une densification des emplois de la forme insécurité

Les fréquences absolues de la forme

La **Figure 1** projette les fréquences absolues du mot sur les 13 parties du corpus correspondant chacune à un mois de publication.

[21] Suite des N nombres (n=nombre de parties du corpus) constituée par la succession des sous-fréquences de cette unité dans chacune des parties, prises dans l'ordre des parties (Lebart, Salem : 57, 319).

Figure 1 :
Les fréquences absolues de la forme *insécurité* (juillet 2001-juillet 2002)

Si nous rapportons les fréquences de la forme *insécurité* à l'ensemble des articles du Monde paru à cette période (dans ce cas chaque partie contient un nombre d'occurrences peu variable [≈ 1800000]), tout en conservant la même partition, nous obtenons la représentation graphique suivante (**Figure 2**):

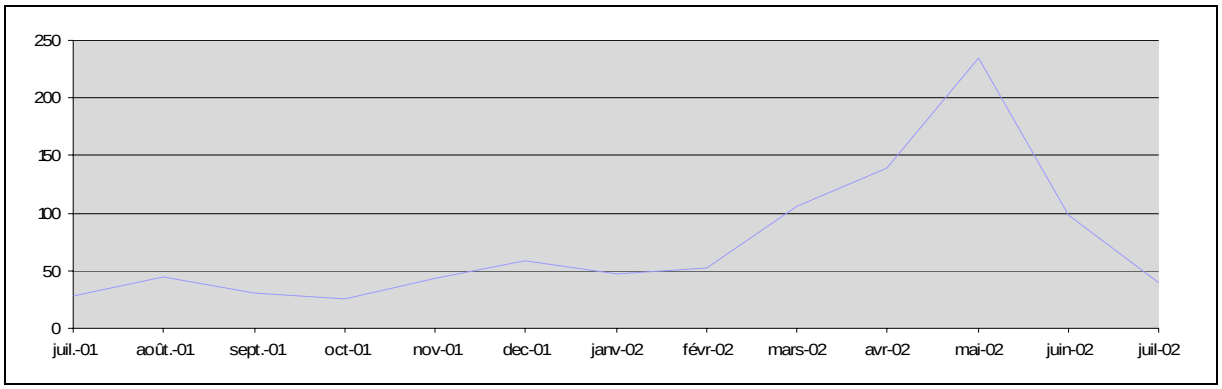


Figure 2 :
Les fréquences relatives de la forme *insécurité* dans *Le Monde* (complet, juillet 2001-juillet 2002)

Sur la **Figure 1**, nous observons une fréquence moyenne de la forme-pôle à 75 occurrences par mois, de juillet 2001 au mois de février 2002 inclus, avec deux légers pics au mois d'août et au mois de décembre. À partir du mois de mars et jusqu'au mois de mai, l'emploi d'*insécurité* s'intensifie avec des fréquences dépassant les 150 occurrences par mois. À partir de mi-avril jusqu'à la fin du mois de mai, elles dépassent le seuil de 200 occurrences mensuelles. Le mois de juin voit une réelle baisse de fréquence. Au mois de juillet 2002, le nombre d'occurrences retombe en dessous de 100, sans retrouver la fréquence de juillet 2001.

Densification de la forme *insécurité* dans les parties du corpus *Monde/Insécurité*

La **Figure 3** projette les fréquences relatives[22] du mot sur les 13 parties du corpus. Précisons ici que nous prenons en compte la fréquence d'*insécurité* à partir d'un corpus qui ne contient que les articles avec la forme et non pas à partir d'un corpus composé de tous les articles du *Monde* (cf. **Figure 2**).

[22] Dans ce cas le nombre d'occurrences du terme est rapporté à la longueur de la partie.

Figure 3 :
Les fréquences relatives de la forme *insécurité* (juillet 2001-juillet 2002)

Cette nouvelle représentation nous amène à observer plus en détail à quel type de densification est soumise la forme *insécurité*. En effet, sans pour l'instant trancher, nous pouvons poser plusieurs hypothèses quant à la manière dont la forme apparaît : soit le mot *insécurité* est employé à plusieurs reprises dans un même article et dans ce cas on observera une densification de la forme à l'échelle d'un article ainsi qu'un éventuel phénomène de « ressassement », soit un grand nombre d'articles emploient le mot et dans ce cas on observera une densification de la forme à l'échelle d'une partie, soit les deux phénomènes sont conjugués.

Figure 4 :
Ventilation de la forme *insécurité* dans les articles du corpus

La carte des sections[23] (**Figure 4**) montre d'abord que l'augmentation de la fréquence du mot est avant tout liée au nombre des articles qui emploient la forme. Dans la mesure où le volume du journal est constant, cela signifie qu'*insécurité* est de plus en plus présent dans l'espace textuel du quotidien. Cette carte confirme une des observations de la **Figure 1** : si le nombre d'occurrences baisse à partir du mois de juin 2002, et si la fréquence du terme au mois de juillet 2002 rejoint pratiquement le niveau observé au mois de juillet 2001, le nombre d'articles employant la forme *insécurité* demeure élevé.

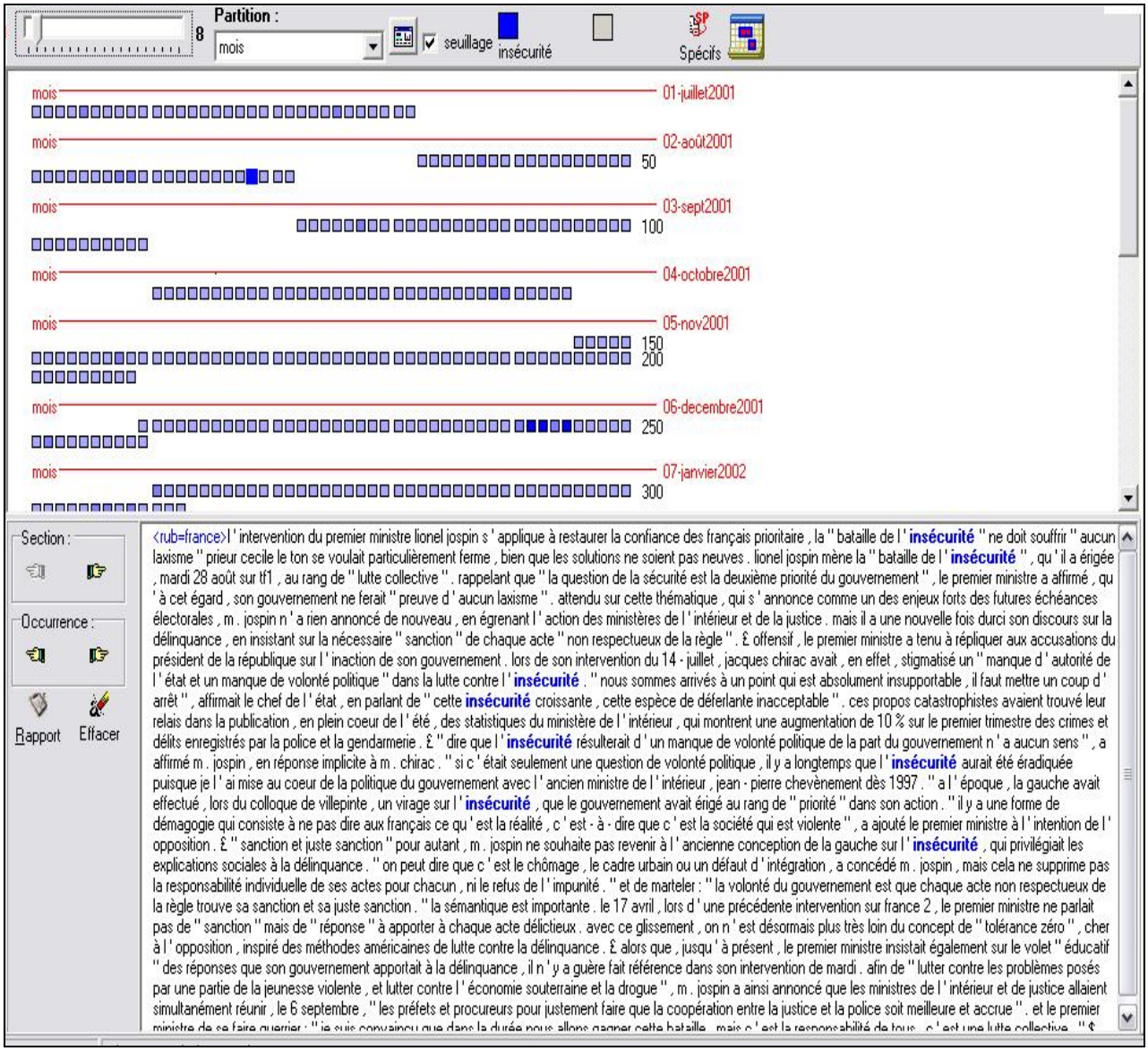
Au regard de la constitution même du corpus, cette dernière observation n'est pas de moindre importance. En effet comme nous l'avons dit plus haut, le choix des articles est exhaustif puisque nous avons rassemblé tous les articles avec le mot *insécurité*. Or certains emplois n'ont peut-être pas de liens directs avec une argumentation électorale ou des arguments post-électoraux et sans exploration ultérieure du corpus, nous ne savons même pas s'ils sont pris dans un discours politique. Pour le traitement statistique, la prise en compte de ces textes est cependant nécessaire.

[23] Chaque carré sur cette figure représente un article.

Cette carte des sections permet de décrire ensuite avec plus de précision les variations de fréquences de la **Figure 3**, et de valider les hypothèses formulées concernant les différents types de densification de la forme. Par exemple, on observe en août un pic de fréquences relatives, mais la carte des sections nous informe qu'au même moment le nombre d'articles demeure peu élevé : le mot est donc souvent employé à plusieurs reprises dans un même article comme on peut le vérifier sur le **Tableau 3**.

Tableau 3 :

Carte des sections et extrait d'un article paru au mois d'août

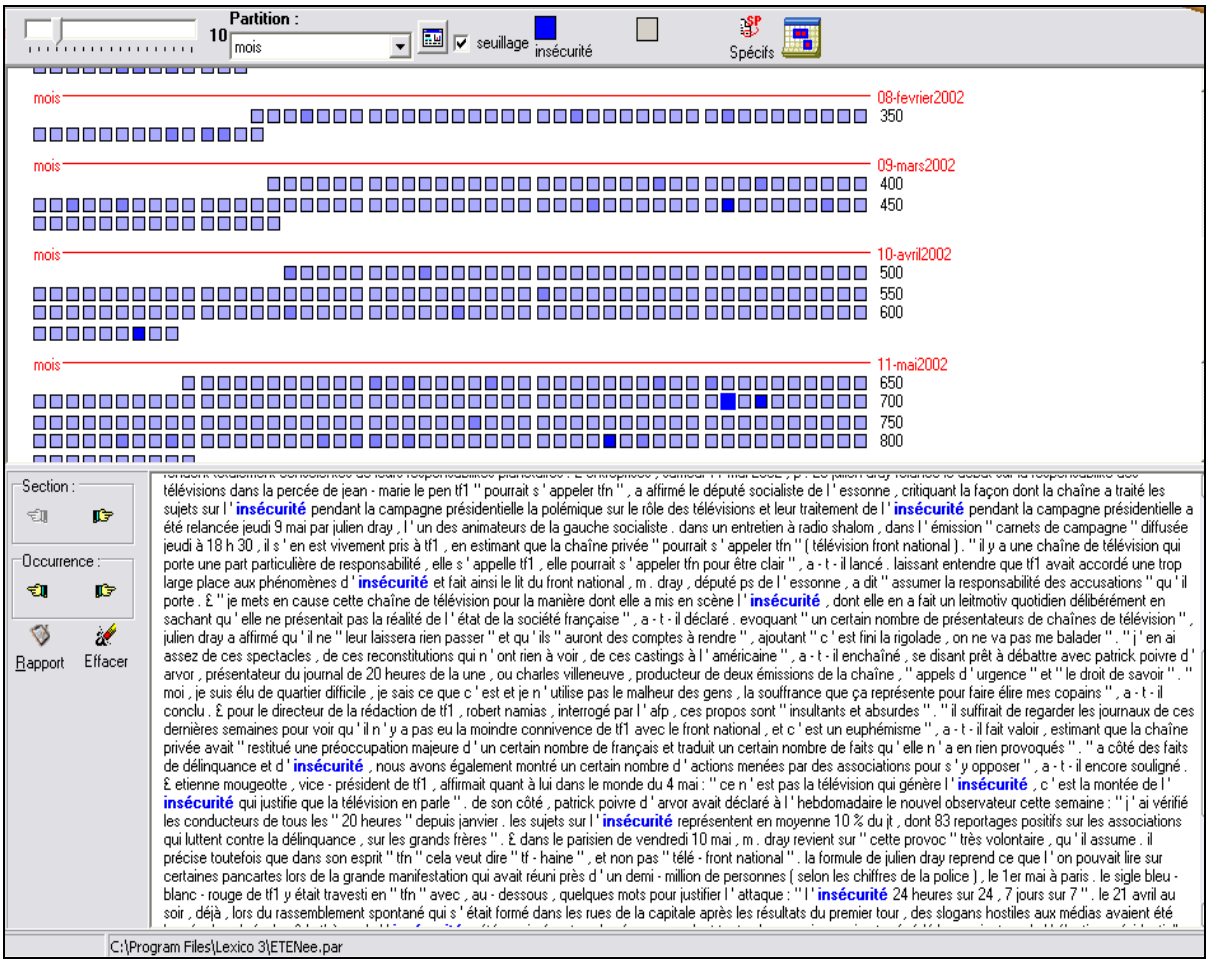


En avril, au contraire, la forme apparaît dans de nombreux articles mais la courbe des fréquences relatives montre que la forme le phénomène de densification est atténué si on considère la longueur de la partie et le nombre des articles.

Enfin, en mai, deux phénomènes sont conjugués : de nombreux articles utilisent la forme et celle-ci est répétée au sein d'un même article (voir **Tableau 4**).

Tableau 4 :

Carte des sections et extrait d'un article paru au mois de mai



Ces observations nous amènent à compléter les **Figures 3** et **4** par un graphique (**Figure 5**) représentant l'évolution du nombre d'articles avec la forme dans le corpus *Monde/Insécurité* et la fréquence moyenne de la forme dans les articles :

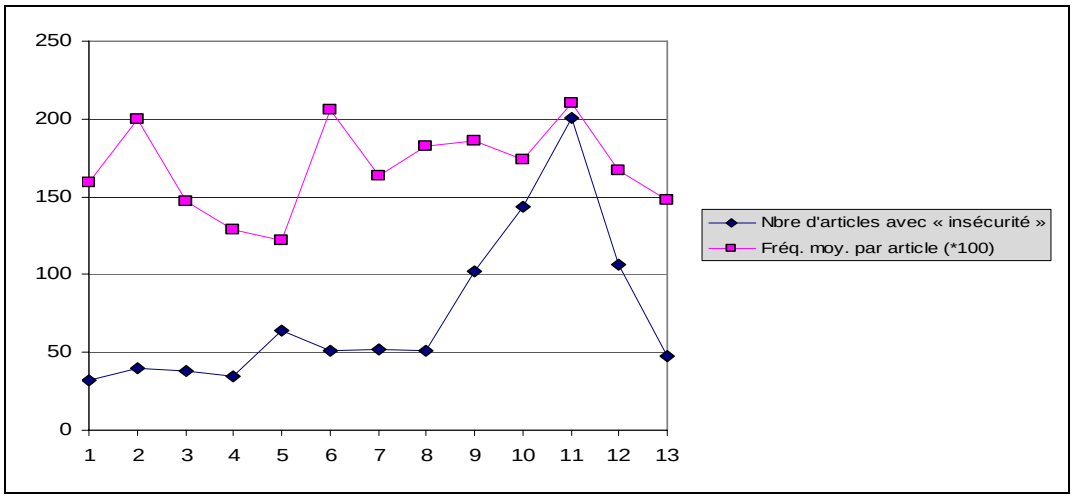


Figure 5 :

Ventilation du nombre d'articles avec *insécurité* et fréquence moyenne de la forme dans les articles (partie 1 [juillet 2001] - partie 13 [juillet 2002])

Cette dernière figure met en évidence les différents phénomènes de densification précédemment décrits et nous livre avec précision les différents modes de densification.

3. Des éléments d'explication

Comment expliquer les variations de fréquence de la forme *insécurité* et son emploi massif jusqu'au mois de mai 2002 (partie 11) ?

Ces variations naissent-elles du traitement simultané de plusieurs actualités où il est question d'« insécurité » ? Il faudra, dans ce cas, mettre en évidence les thèmes qui sont liés à la forme.

L'augmentation en fréquence n'est-elle pas directement liée à une position politique face à la campagne électorale pour les élections présidentielles, et donc à un emploi politique du mot dans *Le Monde* ?

Il serait tentant de valider la seconde interprétation sans exploration complémentaire, si on met en rapport les résultats obtenus ci-dessous avec un contexte extralinguistique, et plus précisément avec des faits concernant la politique intérieure en France entre juillet 2001 et juillet 2002. En s'appuyant sur les repères chronologiques ci-dessous (**Tableau 5**), nous pouvons par exemple faire correspondre à la première hausse de fréquence qui se situe au mois d'août 2001 (partie 02) la publication des chiffres de la délinquance en France ainsi que les premières orientations de la campagne électorale. On peut également mettre en rapport l'accroissement de la fréquence de la forme *insécurité* en mars 2002 avec un fait divers qui mobilise politiques et journalistes, « la Tuerie de Nanterre [24]» : un retour au texte en mars 2002 montre que seuls 16 articles, sur une centaine, concernent le fait divers. C'est donc de façon indirecte semble-t-il, à l'occasion d'un événement extérieur, que prolifèrent des discours sur l'insécurité.

Tableau 5 :
Repères chronologiques

6 juillet 2001 :	création par le Premier ministre L. Jospin d'une mission de réflexion sur l'élaboration d'un « nouvel instrument statistique de mesure de l'insécurité »
14 juillet 2001 :	discours du président J. Chirac qui attaque le Premier ministre sur le thème de l'insécurité.
18 juillet 2001 :	entrée en campagne de J.-M. Le Pen
1-2 août 2001 :	publication des statistiques officielles sur la délinquance (1er semestre 2001).
28 août 2001 :	intervention télévisée de L. Jospin (28 août 2001) qui répond aux attaques du président sur la gestion de l'insécurité
15 novembre 2001 :	adoption par le parlement d'une loi sur la « sécurité quotidienne »
11 février 2002 :	candidature officielle de J. Chirac à l'élection présidentielle. Le premier thème abordé est celui de la sécurité / l'insécurité
20 février 2002 :	candidature officielle de L. Jospin.
27 mars 2002 :	« tuerie de Nanterre » qui donne lieu à une polémique politique
21 avril 2002 :	premier tour des élections présidentielles. J. Chirac arrive en tête avec J.-M. Le Pen.
5 mai 2002 :	second tour des élections présidentielles, J. Chirac est réélu président.
15 mai 2002 :	création par décret d'un « Conseil de Sécurité Intérieure »
16 Juin 2002 :	élection d'une nouvelle assemblée.

[24] Un homme armé, Richard Durn, s'introduit dans le conseil municipal de la commune de Nanterre et tire sur l'ensemble des participants : la tuerie fait plusieurs morts, et la classe politique, sans distinction de courants, se sent réellement fragilisée. Une fois emmené Quai des orfèvres pour déposer, l'homme « profite » d'un moment d'inattention de la part des policiers pour se suicider, ce qui suscite une grande polémique.

On peut enfin être tenté de valider cette interprétation en comparant ces fréquences de la forme dans *Le Monde* avec les fréquences de la forme dans les discours de l'un des candidats à l'élection présidentielle, le président Jacques Chirac en 2001-2002[25] (**Figure 6**) :

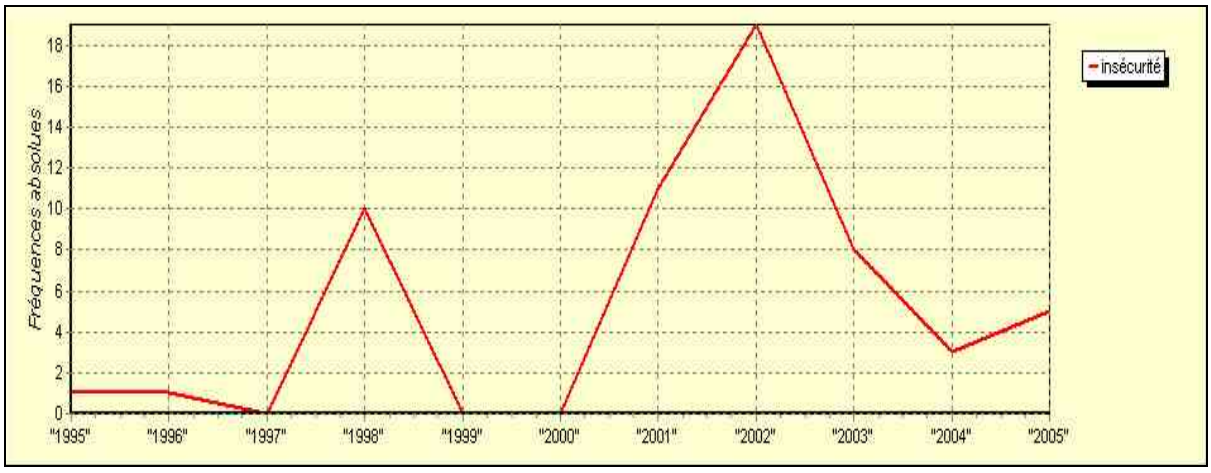


Figure 6 :

Ventilation des fréquences de la forme *insécurité* dans les interventions de J. Chirac (1995-2005)

Cependant, ce type interprétation s'appuie essentiellement sur un savoir extérieur qui ne donne aucune indication sur la manière dont *Le Monde* évoque ce même contexte. De plus, dans ce corpus qui regroupe des articles appartenant à des rubriques différentes, un évènement à portée internationale comme les attentats du 11 septembre peut avoir une incidence dans l'augmentation de fréquence du mot dans le quotidien. D'autres explorations sont donc nécessaires.

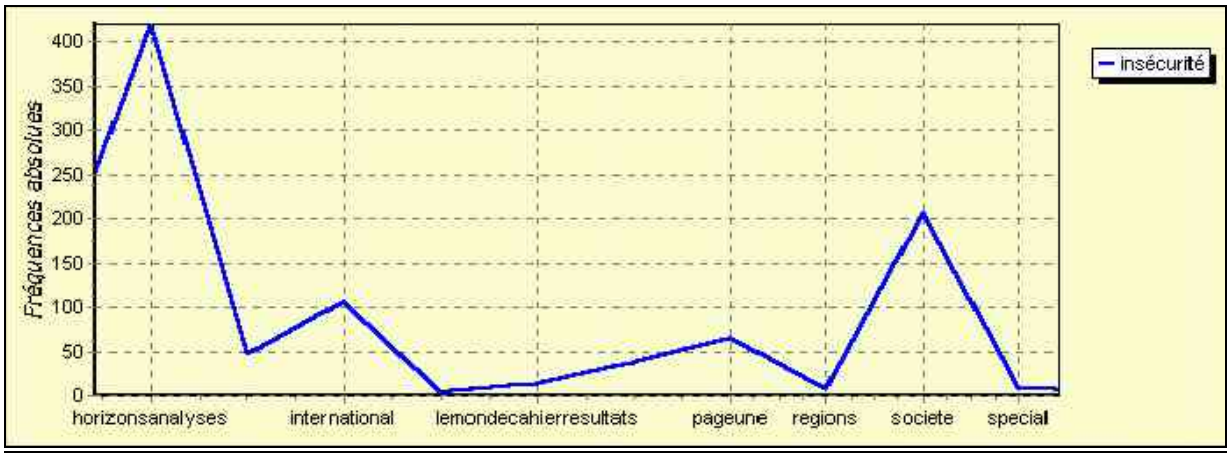
*Ventilation des occurrences d'*insécurité *dans les rubriques du journal* Le Monde

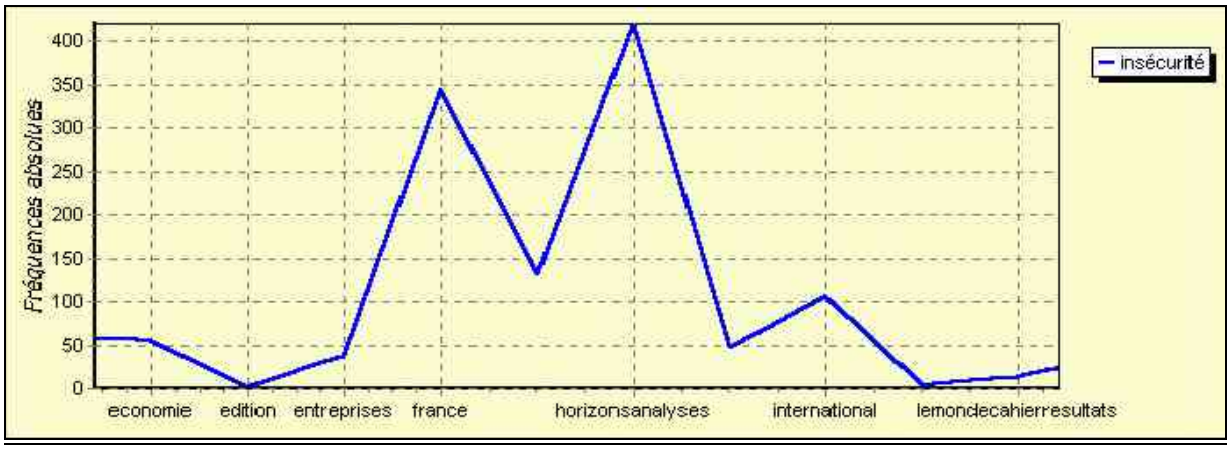
La première expérience porte sur la répartition de la forme-pôle dans les rubriques du journal. En effet, si le terme *insécurité* est majoritairement employé en rapport avec la campagne électorale, il sera surtout présent dans des articles qui traitent de l'actualité nationale.

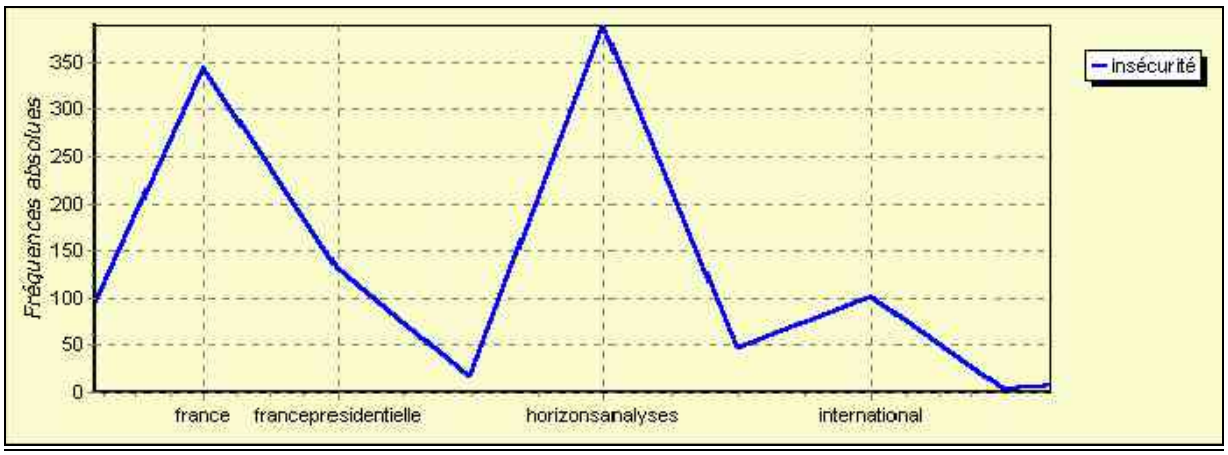
Nous avons redécoupé le corpus *Monde/Insécurité* balisant les rubriques principales qui apparaissent dans le quotidien à ce moment et avons observé la ventilation de la forme dans ces rubriques. En reprenant les classifications proposées par M. Mouillaud et J-F. Tétu, nous n'avons conservé pour cette partition que les rubriques de « niveau 1 », c'est à dire les « titres-rubriques qui figurent en haut de page intérieur et qui sont sur une page de journal « le sommet d'une arborescence qui peut contenir des nœuds à plusieurs niveaux » (J.-F. Têtu, M. Mouillaud : 118). Nous y avons ajouté les suppléments comme « Le Monde des Livres » ainsi que les pages externes du journal (« La Une » et « La Dernière ») qui certes n'ont pas le

[25] Nous nous appuyons ici sur une exploration récente effectuée par S. Fleury (*Textes présidentiels,* http://tal.univ-paris3.fr/blogtal/index.php?cat=65) sur un corpus de travail composé de toutes des textes d'interventions de Jacques Chirac récupérés sur le site de l'Elysée (http://www.elysee.fr/elysee/francais/interventions/sommaire_interventions_du_president_de_la_republique.12629.html) et préparés pour Lexico3 (balisage). « En sortie de la chaîne de traitements (aspiration+formatage+nettoyage), le corpus contient 813 textes différents (251 textes de type *Discours et Déclaration*, 10 textes de type Dialogues et Débats, 136 textes de type *Conférences et points de presse*, 108 textes de type *Interviews, Articles de presse, Interventions télévisées*, 308 textes de type *Lettres*) » (*ibid.*). Voir aussi D. Mayaffre 2004, *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République*, Paris, Champion.

même statut que les rubriques des pages intérieures, mais qui recouvrent d'autres types d'information que les rubriques principales.

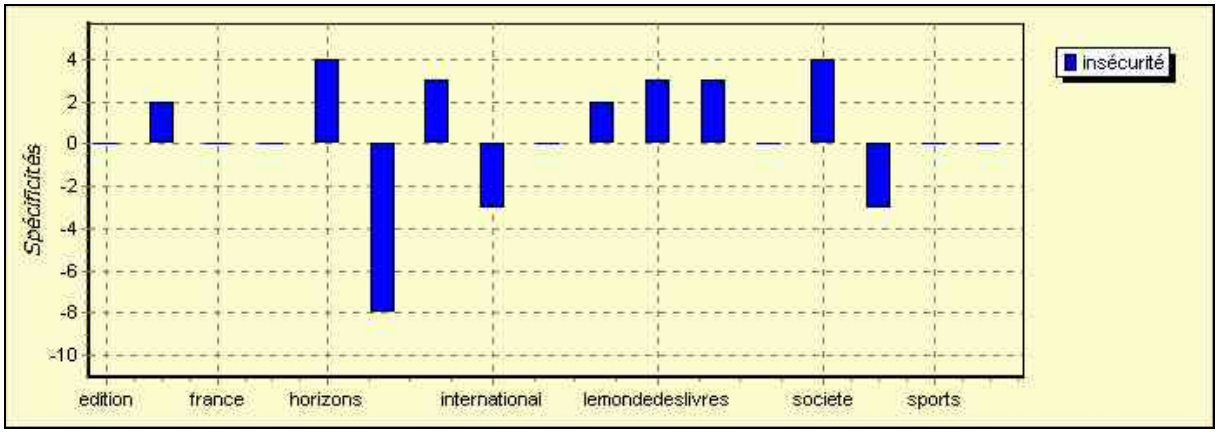


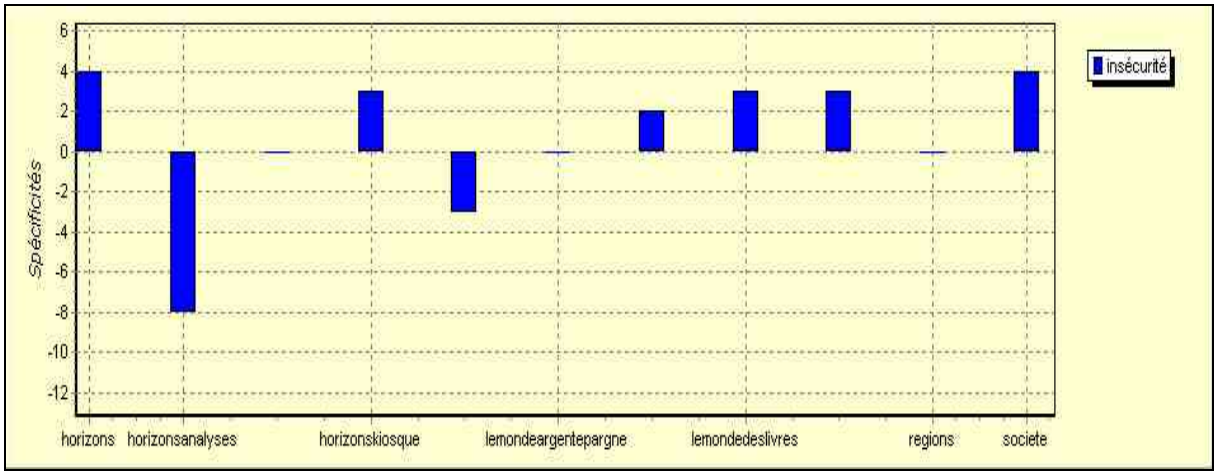




Figures 7, 8 , 9 :
Répartition des occurrences de la forme *insécurité* par rubrique

Les **Figures 7, 8 et 9** montrent que la forme est très présente dans certaines rubriques mais absente d'autres rubriques. Elle est fortement attestée dans des textes relevant des rubriques « Société », « Horizons-Analyses », « France » et de la rubrique « France-Présidentielles » qui traite de l'actualité de la campagne électorale. Elle est beaucoup moins fréquente dans la rubrique « International » (qui représente en moyenne sur l'ensemble du corpus 16,7 % du nombre total d'articles).





Figures 10 et 11 :
Spécificités de la forme *insécurité* (rubriques)

Les indices de spécificités de la forme sur la même partition (**Figure 9 et 10**) permettent d'écarter l'hypothèse d'une influence forte de l'international puisque la forme est en sous-emploi dans la rubrique « International ». Elle apporte aussi quelques précisions : la forme est en suremploi dans les rubriques « Horizons » et « Société », en sous emploi dans la rubrique « Horizons-Analyses » : l'observation du vocabulaire spécifique de ces rubriques ainsi qu'un retour au texte montre qu'il est surtout question dans la rubrique « Horizons » de point de vue sur l'actualité nationale, alors que la rubrique « Horizons-Analyses » regroupe des points de vue sur l'ensemble de l'actualité.

De manière plus générale, l'ensemble de ces visualisations montrent que la forme *insécurité* est essentiellement employée donc dans des textes traitant de l'actualité nationale. De plus, il est intéressant de noter que les rubriques « Horizons » et « Horizons-Analyses » sont des rubriques privilégiés pour l'emploi du mot : en effet celles-ci font souvent place à des tribunes où s'expriment différents points de vue de représentants politiques, sociologues, etc.

Dictionnaire du corpus et segments répétés

Une seconde observation porte sur le vocabulaire qui domine dans le corpus. Le *dictionnaire du corpus* (**Tableau 6**) range l'ensemble des termes du corpus en ordre décroissant selon leur fréquence d'apparition. Nous n'avons retenu ici que les formes pleines les plus fréquentes.

Tableau 6 :
Dictionnaire du corpus

Formes	Occurrences
France	1810
insécurité	1705
politique	1468
Chirac	1421
droite	1249
Jospin	1239
gauche	1168
sécurité	1010
Président	1070
(Le) Pen	997
tour	783
présidentielle	744
police	651
délinquance	567
élection	519
vote	496
société	484

Les mots les plus employés désignent soit l'événement politique de la période, à savoir les élections présidentielles – *campagne, vote, présidentielle* – soit des hommes politiques qui tiennent un rôle au sein de l'Etat et/ou qui sont acteurs de cet évènement – *Jospin*, *Chirac*, *Le Pen, candidat*. L'emploi du terme *insécurité* paraît donc surtout dépendant dans le quotidien d'une « masse » discursive sur le thème de la campagne électorale. A ce réseau de termes viennent s'ajouter les formes *délinquance* et *police* sur l'une desquelles nous allons revenir.

Le dictionnaire des segments répétés [26] (**Tableau 7**) donne quelques précisions supplémentaires : avec le segment *l'insécurité*, les segments les plus fréquents sont les désignants de trois acteurs politiques, *Jospin*, *Chirac* et *Le Pen*, et de deux mouvements, *la gauche* et *la droite*. Parmi les formes pleines, vient ensuite le segment *la délinquance*.

[26] Suite de formes non séparées par une ponctuation dont la fréquence est égale ou supérieure à deux. Pour plus de lisibilité nous avons ici restitué les majuscules aux noms propres.

Tableau 7 :
Segments répétés du corpus (extraits)

Formes	Occurrences
Le Pen	995
l'insécurité	1256
la France	719
Jacques Chirac	706
Lionel Jospin	640
la gauche	592
la sécurité	551
la droite	458
la délinquance	420
la république	409
la campagne	401
la police	391
le gouvernement	387
extrême droite	383
élection présidentielle	362

Une dernière expérience permet de montrer qu'il n'y pas de corrélation à l'échelle du corpus entre l'événement « attentat du 11 septembre » et le traitement de cette actualité d'une part, et entre la hausse de fréquence d'*insécurité* d'autre part.

Le « 11 septembre » et la forme insécurité

Nous avons constitué à partir du dictionnaire un type[27] particulier que nous appellerons *ATA* et qui regroupe les formes *attentats*, *terrorisme* et *terroristes* en raison de la parenté sémantique et lexicales (pour les deux dernières), formes qui apparaissent dans des textes évoquant les attentats du 11 septembre et leurs conséquences (lutte contre le terrorisme au niveau international, par exemple). Nous avons voulu comparer les indices de spécificités de ce type avec ceux des formes *délinquance* et *insécurité* sur un axe chronologique (partition « mois » du corpus), la forme *délinquance* étant l'une des formes le plus employé à l'échelle du corpus avec la forme *insécurité.*

[27] Par type, nous entendons les divers regroupements d'unités que l'on peut opérer sur la base de leur identité ou de leurs ressemblances. On peut définir « le *type généralisé* ***TGen*** comme un ensemble d'occurrences sélectionnées parmi les occurrences du texte » (C. Lamalle, A. Salem 2002 : 2).

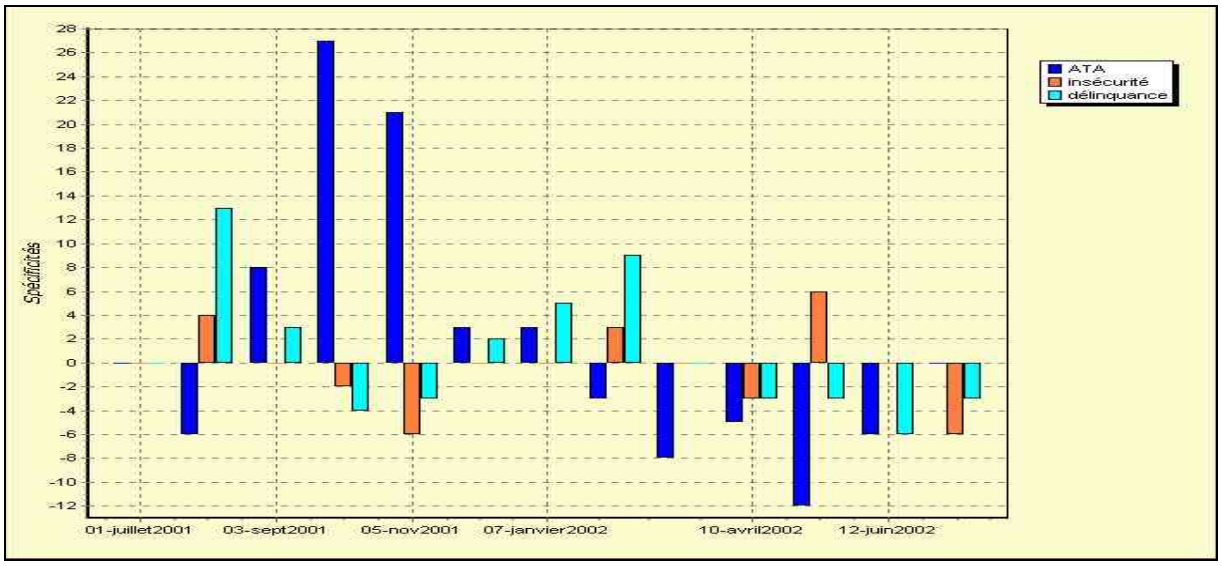


Figure 12 :
Spécificités du type *ATA* et des formes *insécurité* et *délinquance* (juillet 2001-juillet 2002)

Nous remarquons sur la **Figure 12** que les formes *insécurité* et *délinquance* sont anormalement sous-employées de manière simultanée dans les parties 4 et 5 (octobre et novembre 2001), ce qui n'est pas le cas du type *ATA* qui est en suremploi.

Ces trois ensembles d'observations nous amènent à retenir l'interprétation selon laquelle l'emploi d'*insécurité* est fortement lié au traitement de la campagne électorale par *Le Monde*. Nous souhaitons maintenant revenir sur la présence répétée de la forme *délinquance,* ce qui va nous permettre de déceler un nouveau phénomène concernant l'emploi du mot *insécurité* en 2001-2002.

4. Insécurité et délinquance, deux formes très proches

Les dictionnaires du corpus et des segments répétés et les indices de spécificité des formes *insécurité* et *délinquance* sur la partition chronologique nous ont amenée à formuler l'hypothèse que les deux formes étaient cooccurrentes à l'échelle d'un paragraphe ou d'une séquence phrastique et à identifier le lien qui les associait.

Pour affiner ces observations qui nous laissent présager l'existence d'un phénomène caractéristique des discours qui traversent *Le Monde* pendant cette période électorale, nous avons voulu mettre en évidence le vocabulaire spécifique des séquences qui contiennent le mot *insécurité,* ce qui implique d'observer la ventilation de la forme sur un nouveau découpage du corpus**.**

Le **Tableau 8** relève les formes nominales les plus employées (1ère colonne), leur fréquence dans l'ensemble du corpus (2e colonne), leur fréquence dans les séquences qui contiennent la forme *insécurité* (3e colonne), enfin leur indice de spécificité (4e colonne) : mots employés par les hommes politiques à ce moment et repris par des journalistes (*délinquance*, *chômage*, *lutte*), qui renvoient aux thèmes abordés pendant la campagne ou à des désignations qualifiantes (*déferlante*). On remarque surtout la forme *délinquance* qui fonctionne étroitement en séquences avec *insécurité* et dont on a déjà noté la forte fréquence dans le corpus.

Tableau 8 :
Extrait des spécificités relatives des séquences contenant la forme *insécurité*

Formes	Fréquence totale	Fréquence	Coeff.
sentiment	370	181	***
thème	207	108	***
lutte	305	120	***
chômage	333	88	29
immigration	294	80	28
montée	162	56	26
préoccupation	70	37	25
délinquance	567	111	24
déferlante	18	16	18
campagne	964	141	18
débat	389	74	16

Une lecture de moment de corpus montre d'une part que se croisent dans les articles des discours politiques concurrents qui vont charger le mot *insécurité* d'accents différents, d'autre part que journalistes et politiques évoquent surtout l'insécurité à travers la question de la délinquance en France. Mais jusque là nous ne pouvons parler de phénomène discursif qui prend en compte la matérialité linguistique. Pour étudier la relation entre les deux formes nous avons constitué un sous-corpus en prenant en considération les moments de suremploi de la forme (voir **Figure 12**). Une analyse plus fine sur ce corpus restreint articulant des catégories descriptives telles que la reprise ou la reformulation à la notion d'*objet de discours* telle qu'elle a été théorisée par S. Moirand et F. Sitri, révèle un jeu de reprises entre segments discursifs contenant les termes *insécurité* et *délinquance* et un paradigme de termes en relation métonymique avec *délinquance* (*vol*(s), *agression(s)*), à l'échelle de la phrase ou d'un paragraphe (**Tableau 9**), dans des séquences qui font intervenir des classes de locuteurs différentes : journalistes, hommes politiques, chercheurs, représentants de la société civile, citoyens. Une analyse plus approfondie sur corpus restreint montre que le mot va fonctionner comme une dénomination consensuelle de *délinquance*, quels que soient les locuteurs.

Tableau 9 :
Échantillon de séquences contenant les formes *insécurité* et *délinquance*

Extrait du corpus ***Monde/Insécurité***

§ adjoint chargé de « la sécurité, la prévention de la **délinquance** et la protection de l'enfance en danger », florent montillot , quarante - sept ans, tient « sa » première victoire dans sa croisade contre l'**insécurité**.

§ le chômage n'a jamais autant décru, et jamais la **délinquance** n'a autant progressé », martèle florent montillot, qui affirme vouloir affronter l'**insécurité** « sans cache - sexe , sans états d ' âme , et en même temps sans dogmatisme ».

§ la hausse sensible de la **délinquance** met l'**insécurité** au coeur du débat politique.

§ au contraire, elles ont, en confirmant une tendance à la hausse de la **délinquance** commencée en 2000, offert une assise officielle aux discours alarmistes sur la montée de l'**insécurité**.

§ de ce creuset était née la police de proximité, formule censée répondre à la fois aux nouvelles formes de **délinquance** de manière plus efficace, et

satisfaire les demandes d'une population inquiète de l'accroissement sensible des petites infractions créatrices d'un sentiment d'**insécurité**.

§ ministres en campagne sur la sécurité, le ps proclame que « le droit à la sûreté est une liberté fondamentale et l'**insécurité** une inégalité sociale » et s'engage à « apporter à tout acte d'incivilité ou de **délinquance** une réponse juste, proportionnée et rapide ».

§ dans un premier temps, l'enquête s'attache à analyser le sentiment d'**insécurité** en ile - de - france, qui se décompose entre la préoccupation générale pour la **délinquance** et « la peur du crime ».

§ tournant le dos, en octobre 1997, aux explications sociales de la **délinquance**, lionel jospin a érigé la lutte contre l'**insécurité** au rang de seconde priorité de son gouvernement, juste après l'emploi et la lutte contre le chômage.

§ alors que le thème de l'**insécurité** occupe une large place dans la campagne électorale, des magistrats, avocats, syndicalistes, éducateurs de la protection judiciaire de la jeunesse, universitaires ou sociologues multiplient les initiatives pour dénoncer les « amalgames » et la « antastique hypocrisie »des candidats en matière de lutte contre la **délinquance** des mineurs.

§ jospin ne souhaite pas revenir à l'ancienne conception de la gauche sur l'**insécurité**, qui privilégiait les explications sociales à la **délinquance**

Cette analyse nous a amenée aujourd'hui à un redécoupage du corpus en paragraphes afin de vérifier de manière plus systématique sur l'ensemble des articles la proximité des formes *insécurité* et *délinquance* (**Tableau 10**), d'une part, de créer de nouveaux types rassemblant ce paradigme de terme associé au terme *délinquance* d'autre part (**Tableau 11**).

Tableau 10 :

Cooccurrence de la forme *insécurité* et de la forme *délinquance* dans les paragraphes du corpus *Monde/Insécurité*

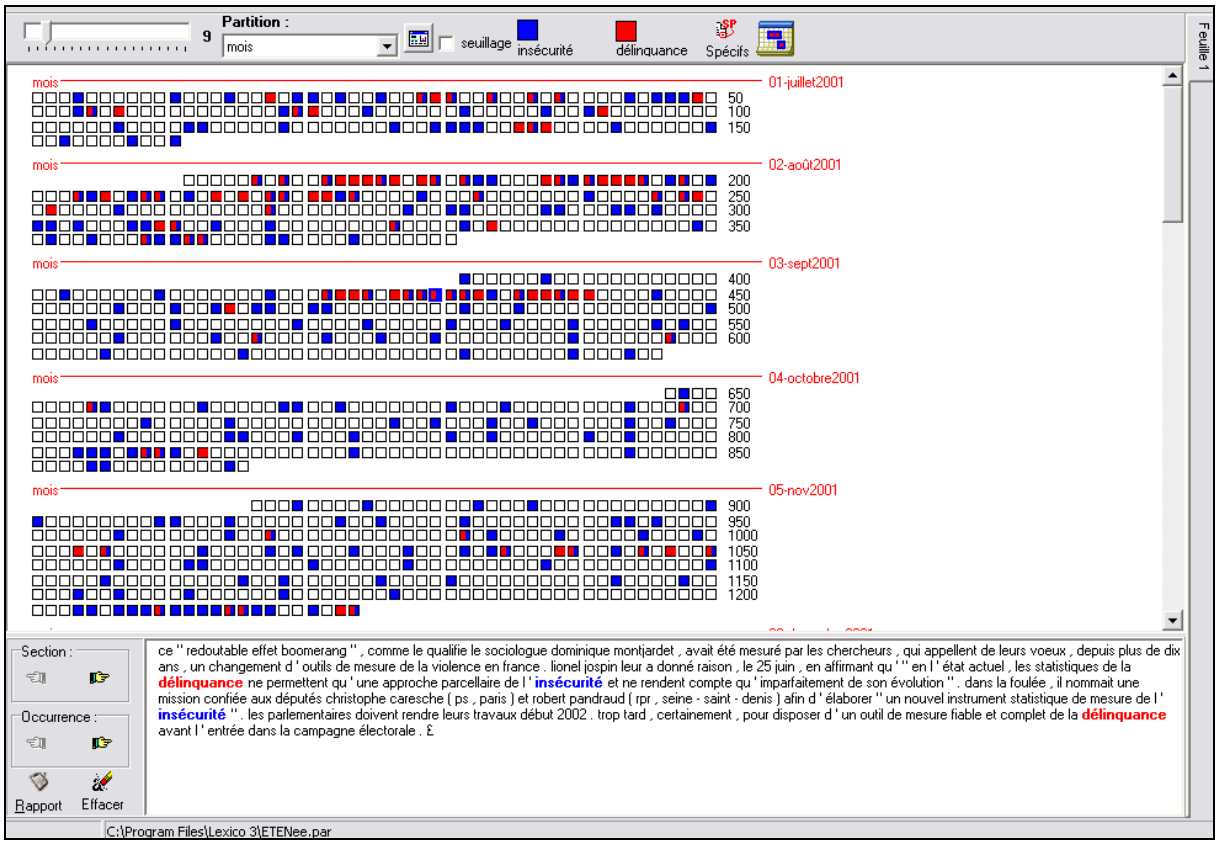
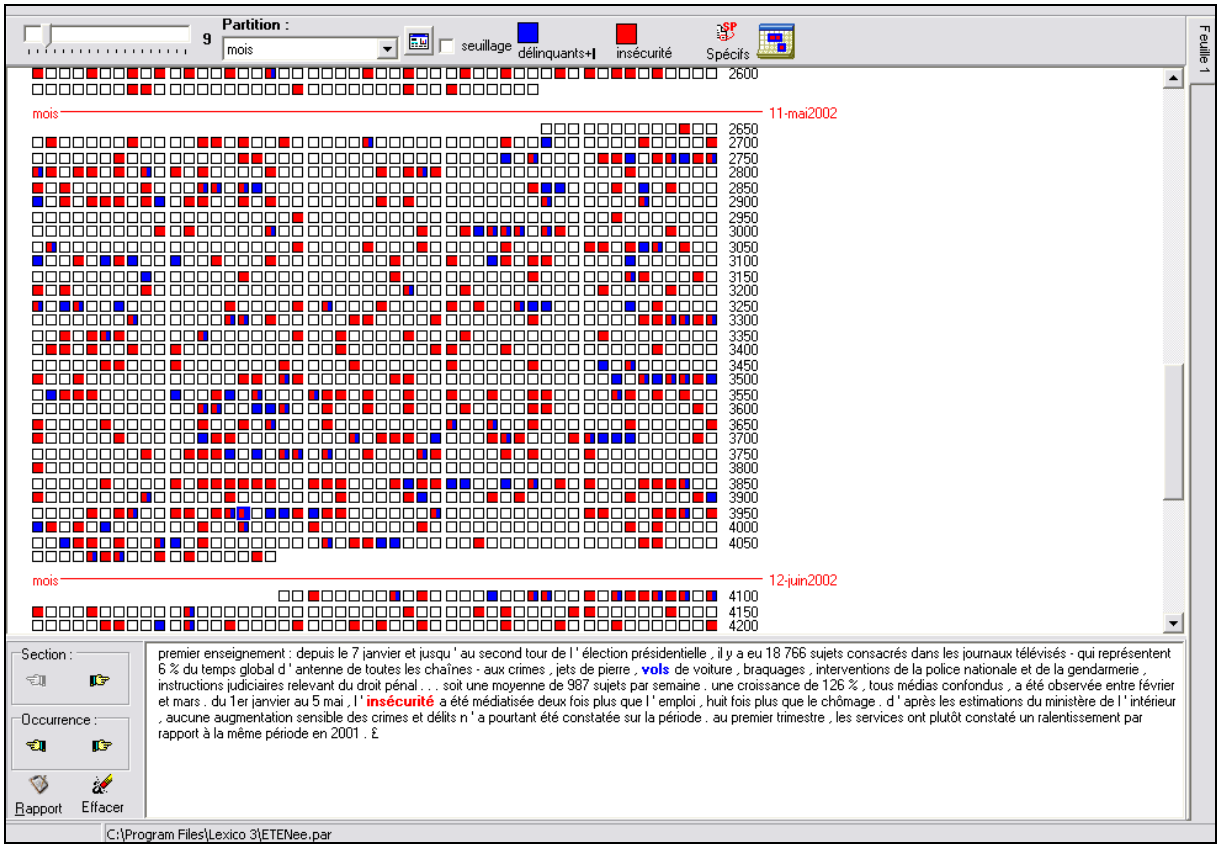


Tableau 11 :

Cooccurrence de la forme *insécurité* et du type *DELINQUANCE* (*délinquance*, *délinquant(s)*, *vol(s)*, *agression(s)*)



5. Conclusion

Cette exploration met en évidence un phénomène que tout lecteur du journal *Le Monde* pouvait pressentir sans toujours avoir les moyens de le vérifier : la densification d'emploi du mot *insécurité*. L'observation de différents types de fréquence a permis de décrire cette densification spécifique à ce corpus médiatique et d'en donner les caractéristiques complètes.

Trois types d'observations complémentaires nous ont guidée dans l'interprétation de cette densification : ainsi nous avons pu valider l'hypothèse selon laquelle il y a corrélation de cet emploi dans le journal avec un événement politique majeur, l'élection présidentielle.

L'analyse des cooccurrences telle qu'elle a été utilisée dans cette étude nous a permis de repérer un phénomène particulier : la forte proximité de deux formes. Dans le cas présent, la récurrence d'une forme cooccurrente (*délinquance*) à plusieurs échelles (corpus dans son ensemble ou séquence phrastique) nous a amenée à la sélection avertie de corpus restreints pour une analyse qui s'est appuyée cette fois-ci sur le texte dans sa linéarité. Celle-ci a pu mettre à jour un phénomène discursif, à savoir la reprise de segments discursifs avec *insécurité* par la reprise de segments avec *délinquance* et inversement.

Enfin et de manière plus générale, nous pouvons voir que sur des corpus de presse, particulièrement délicats à décrire en analyse du discours, il est nécessaire de multiplier les expériences textométriques sur différentes partitions et d'articuler analyse quantitative et analyse qualitative.

6. Indications bibliographiques

Lamalle, C., Salem, A., 2002, « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », dans *Actes des 6emes journées d'analyse statistique des données textuelles*, 2002, Inria, St Malo

[http://www.cavi.univparis3.fr/lexicometrica/jadt/jadt2002/tocJADT2002.htm].

Lebart L., Salem, A., 1994, *Statistique textuelle,* Paris, Dunod.

Moirand, S., 2003, « De la nomination au dialogisme : quelques questionnements autour de l'objet de discours et de la mémoire des mots » in Cassanas, A., Demange, A., Laurent, B., Lecler, A. *Dialogisme et nomination*, Montpellier, Praxiling Université Paul Valéry - Montpellier III, p. 27-61.

Moirand, S., 2004, « L'impossible clôture des corpus médiatiques. La mise au jour des observables entre catégorisation et contextualisation », dans *TRANEL* 40, juillet 2004, p. 72-92.

Mouillaud, M., Tétu, J.-F., 1989, *Le journal quotidien*, Presses Universitaires de Lyon.

Nee, E., 2005, *(L') insécurité ou de la fabrication d'un objet consensuel dans le discours de presse*, communication au Colloque Jeunes Chercheurs « Matérialités de l'activité de nomination » (11 mars 2005), Université Paris III- Syled EA2290 (Publication en cours).

Sitri, F., 2003, *L'objet du débat. La construction des objets de discours dans des situations argumentatives orales*, Paris, Presses de la Sorbonne Nouvelle.

Tournier, M., 1997, *Des mots en politique. Propos d'étymologie sociale 2*, Paris, Klincksieck.

7. Fonctionnalités Lexico3 utilisées dans cette exploration

N°	*Fonctionnalité*	Résultat
5	PCLC	*Tableau 1*
6	Ventilation	*Figure 1, Figure 3, Figure 6, Figures 7/8/9, Figures 10-11, Figure 12*
7	Carte des sections	*Figure 4, Tableau 3, Tableau 4, Tableau 10, Tableau 11*
4	Segments répétés	*Tableau 7*

Discours royal espagnol

[Discours gouvernementaux]

C. Pineira-Tresmontant, A. Salem

cpineirat@aol.com, salem@msh-paris.fr

Résumé : La courbe d'accroissement du vocabulaire calculée à partir d'une série de 25 allocutions adressées aux forces armées par le roi d'Espagne, (corpus ***Pascua*** 1976-2000) révèle un très faible accroissement pour l'allocution de 1993. Une suite d'opérations textométriques permet de comprendre la raison de cette anomalie. On en déduit une méthode pour repérer les passages à fort taux de répétition dans les séries textuelles du même type.

1. Contexte de la recherche

Le corpus ***Pascua*** est constitué de 25 allocutions prononcées par le roi Juan-Carlos à l'intention des forces armées espagnoles à l'occasion d'une fête annuelle, *la Pascua militar*, entre 1976, date de son accession au pouvoir, et 2000. Ce corpus a été réuni par C. Pineira-Tresmontant dans le cadre d'une étude plus large sur les stratégies de communication du monarque espagnol[28]. On trouve un exemple de ce type d'allocution au tableau 1 ci-dessous. Le corpus compte 4 731 formes pour 32 389 occurrences. La partition naturelle du corpus en 25 parties dont chacune correspond à une année amène les résultats que l'on peut voir au tableau 2.

Tableau 1

Extrait de l'allocution prononcée en 1976

Extrait du corpus ***Pascua***

```
<a=1976>
§ palabras de s.m. el rey en la celebración de la pascua militar.
§ 6 de enero de 1976
§ gracias, señor vicepresidente, por estas palabras tan cargadas de
sentimientos castrenses.§ gracias por esa lealtad y esa unión  de las
fuerzas armadas que me presentáis y que son garantía de un futuro
prometedor. § la pascua de reyes, es una fiesta de gran arraigo en
nuestra patria y es un día de ilusiones. es una fecha que nos habla de
fe, de porvenir y de esperanza. virtudes militares que son
imprescindibles para cimentar la seguridad en el triunfo, base del éxito
en los ejércitos.
§ nosotros que consagramos nuestra vida a españa, sabemos bien que la
patria necesita que todos los días le ofrezcamos algo. para cumplir este
compromiso tenemos que esforzarnos en hacer cada día mejor el servicio
encomendado.
```

[28] . On trouvera dans la dernière section les références de plusieurs articles consacrés à l'étude de ce type de corpus.

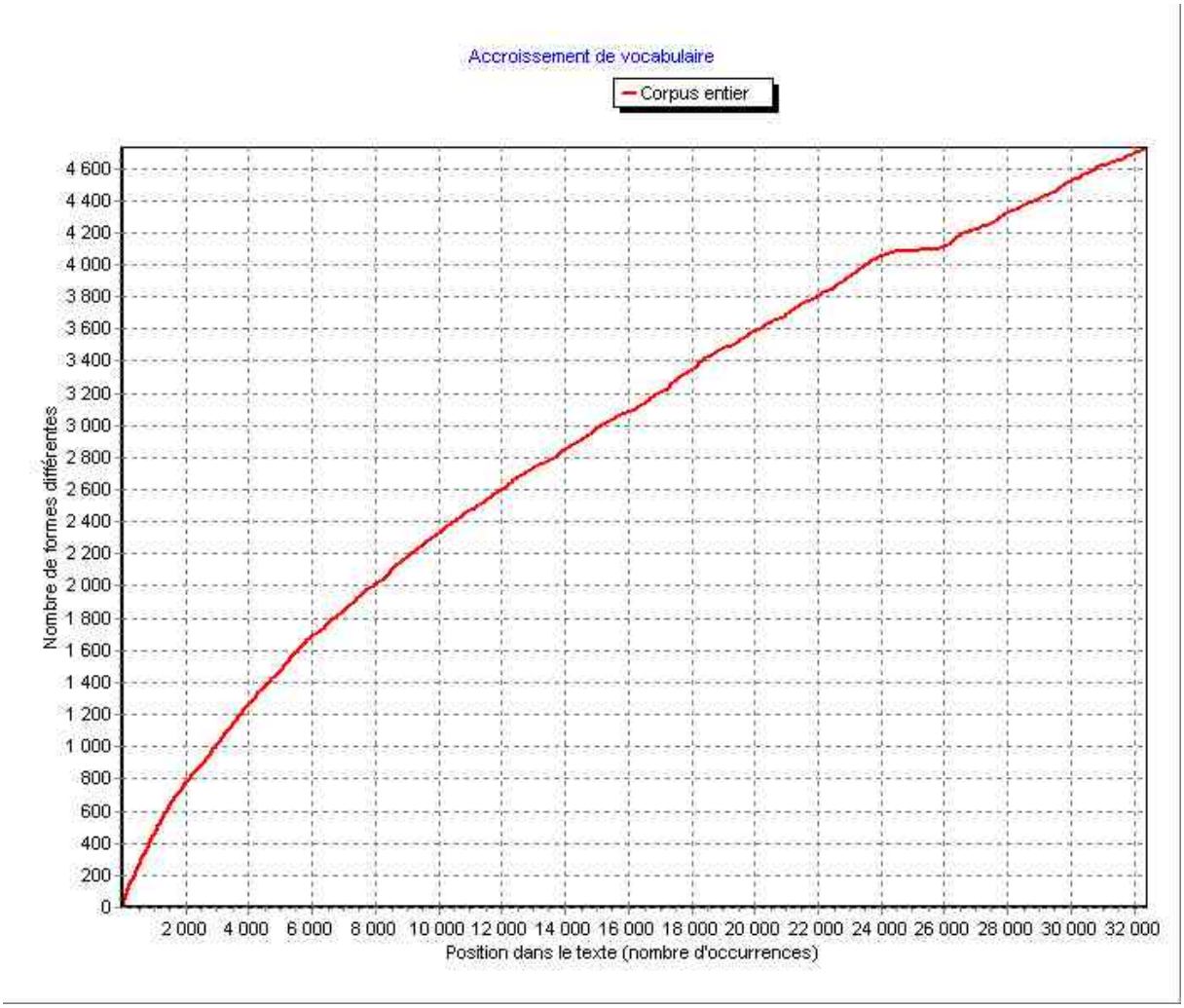


Figure 1 :
Courbe d'accroissement du vocabulaire pour la série ***Pascua***

2. Anomalies dans l'accroissement du vocabulaire

La courbe d'accroissement du vocabulaire[29], Figure 1, établie pour l'ensemble de la série ***Pascua*** révèle une particularité textométrique de ce corpus. On voit sur cette figure que cette courbe, qui ne présente pas de particularité du début de la série à l'occurrence 24 000 environ, connaît un fléchissement très net de son accroissement pour la partie du texte qui s'étend entre les occurrences 24 000 et 25 0000 environ. En se reportant au décompte cumulé des occurrences, on s'aperçoit que cette portion du texte correspond très exactement au discours prononcé à l'occasion de la fête de 1993.

Ce constat amène une question : *Comment expliquer le fait que le discours de 1993 n'apporte que très peu de formes nouvelles à la série des allocutions prononcées entre 1976 et 1993 ?*

La réponse à cette question peut être recherchée dans trois directions distinctes (sans que l'on puisse exclure, a priori, que le phénomène soit dû à une combinaison de ces trois possibilités) :

---

[29] Rappelons que la *courbe d'accroissement du vocabulaire montre* la dépendance entre *V(x)* - le nombre des formes différentes rencontrées jusqu'à l'occurrence t (ici en ordonnée) et *t*, la longueur du corpus (portée en abscisse).

H1 : le discours *P93* est *intrinsèquement pauvre* en vocabulaire, ce qui expliquerait son très faible apport à l'ensemble, du point de vue de l'accroissement.

H2 : le discours *P93* reprend systématiquement des formes lexicales déjà utilisées dans les *différentes allocutions* de la période précédente (1976-1992).

H3 : le discours *P93* reprend massivement (sous forme de recopie, de citation, etc.) des formes déjà utilisées dans *un des discours* de la période précédente, qu'il conviendra alors d'identifier.

L'hypothèse ***H1*** peut facilement être écartée si l'on considère le tableau 2 qui permet de comparer les longueurs de chacune des parties et le nombre des formes différentes qu'elles contiennent. On vérifie facilement que la partie *P93* qui compte 1 800 occurrences compte à peu près autant de formes différentes (un peu plus de 700 formes) que les parties de longueur tout à fait comparables (*P83, P90, P92*).

Tableau 2 :
Caractéristique lexicométriques pour les 25 allocutions

Année	occurrences	formes	*Année*	occurrences	formes	*Année*	occurrences	formes
1976	294	164	*1985*	131	84	*1994*	1313	593
1977	415	227	*1986*	1208	541	*1995*	1035	491
1978	1366	588	*1987*	1407	592	*1996*	924	444
1979	2333	864	*1988*	1139	500	*1997*	868	444
1980	1748	700	*1989*	1949	798	*1998*	814	416
1981	665	315	*1990*	1769	718	*1999*	660	333
1982	2601	935	*1991*	1490	632	*2000*	813	392
1983	1780	703	*1992*	1879	758			
1984	1988	757	*1993*	1800	713	*Pascua*	32 389	4 731

La figure 2, qui permet de comparer les courbes d'accroissement du vocabulaire pour les parties *P92* et *P93,* nous confirme que l'accroissement calculé pour la partie *P93* est tout à fait comparable à celui que l'on calcule pour l'allocution qui précède.

3. Résolution du problème

La procédure décrite dans les paragraphes qui suivent devrait nous permettre de trancher entre les deux hypothèses qui subsistent. Nous allons constituer un type particulier, que nous appellerons ***SegmentsLongs,*** à partir de tous les segments les plus longs que l'on peut repérer dans le texte[30]. En d'autres termes, une occurrence du corpus relève du type ***SegmentsLongs*** si la séquence composée par cette occurrence et dix occurrences autour d'elle peut être localisée à deux endroits différents du corpus.

[30] La version 3.45.1 de ***Lexico3*** permet de repérer les segments répétés composés de *onze* formes consécutives. Cette limitation n'est pas contraignante car la répétition d'une séquence aussi longue trahit en général la répétition (citation/reprise, etc.) de portions de textes beaucoup plus importantes (groupe de phrases, paragraphes, groupes de paragraphes).

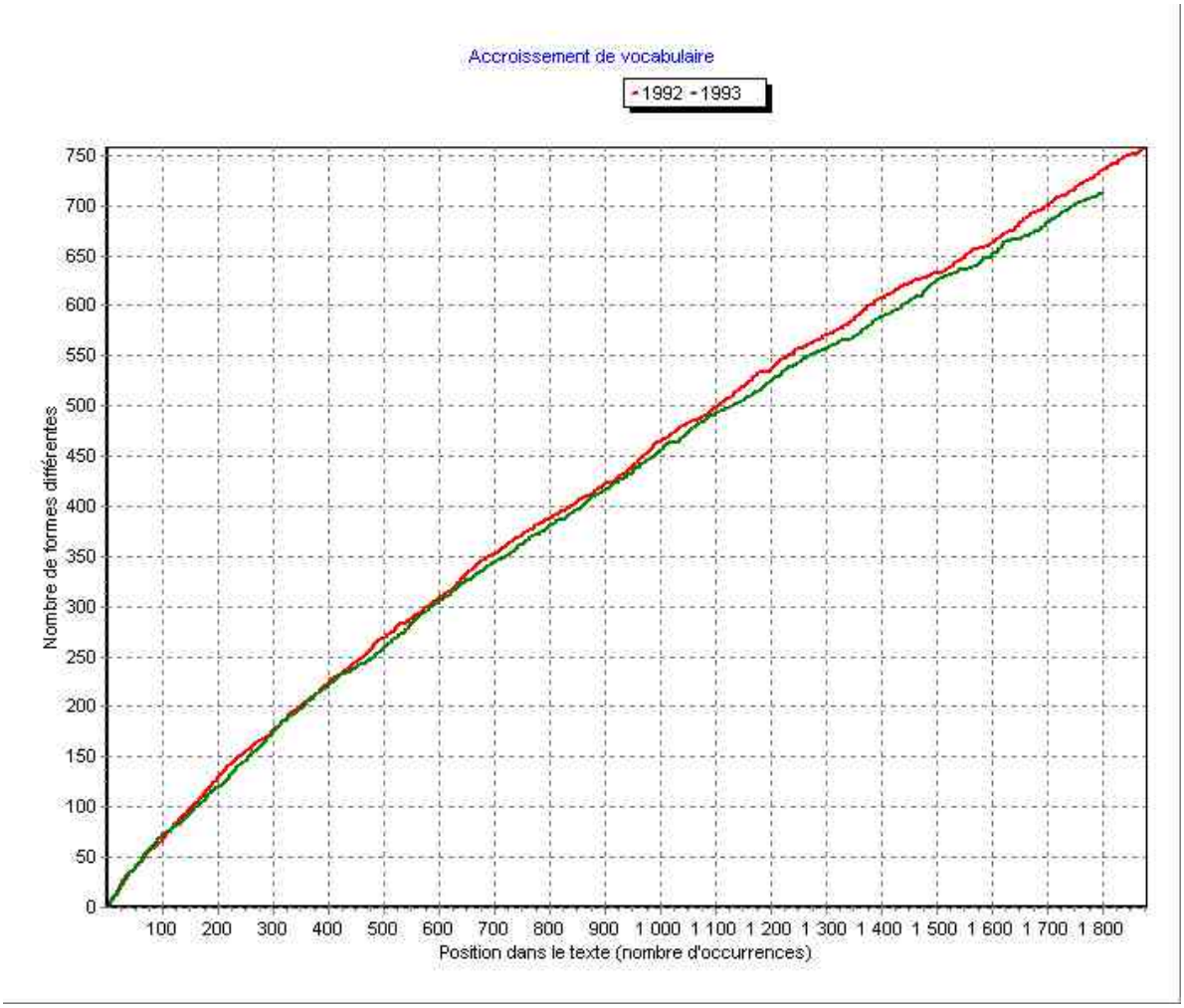


Figure 2 :
Courbes d'accroissement du vocabulaire pour les allocutions de 1992 et 1993

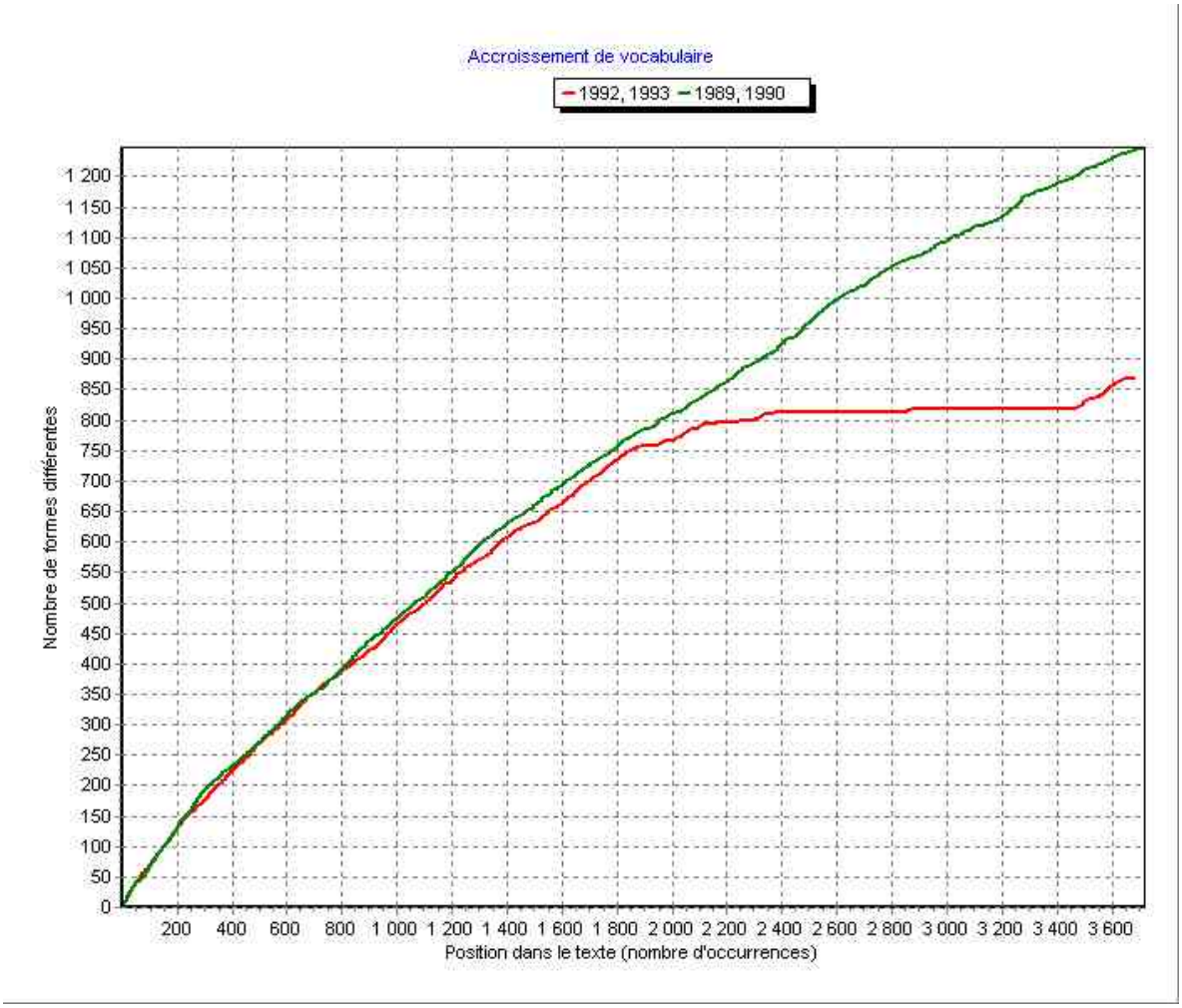


Figure 4 :
Comparaison des courbes d'accroissement pour les allocutions de [1989-1990] et [1992-1993]

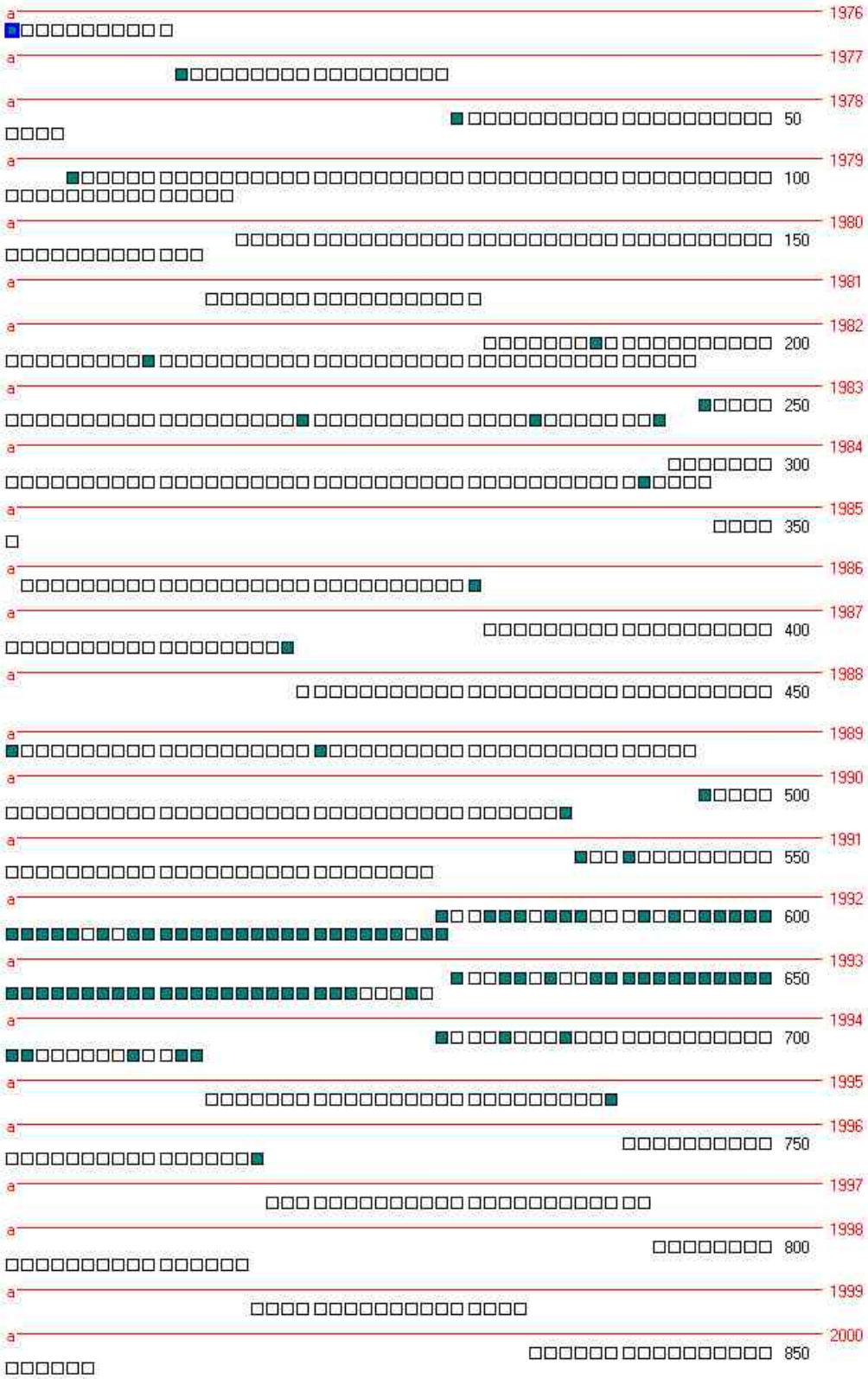


Figure 3 :
Ventilation des segments répétés de long > 11 dans les paragraphes du corpus

Tableau 3 :
Comparaison des allocutions de 1992 et 1993

<a=1992>	<a=1993>
§ discurso de s.m. el rey en la **celebracion de** la pascua militar.	§ **palabras** de s.m. el rey en la pascua militar.
§ 6 de enero de **1992**	§ 6 de enero de **1993**
§ queridos compañeros:	§ queridos compañeros:
§ aunque a través del año procuro encontrar todas las ocasiones posibles para asistir a actos, ceremonias, conmemoraciones o maniobras militares, es esta de la pascua militar la más propicia para reunirme con las representaciones de las fuerzas armadas y experimentar la satisfacción de compartir con vosotros una fiesta tan tradicional.	§ aunque a través del año procuro encontrar todas las ocasiones posibles para asistir a actos, ceremonias, conmemoraciones o maniobras militares, es esta de la pascua militar la más propicia para reunirme con las representaciones de las fuerzas armadas y experimentar la satisfacción de compartir con vosotros una fiesta tan tradicional.
	§ **por eso lamenté mucho en la celebración de 1992, verme privado -por un desafortunado accidente- de asistir a un acto como este, que encierra para mi tan profunda significación.**
§ recibid, ante todo, mi felicitación y la de mi familia, para vosotros y las vuestras, con los mejores deseos en el año que acaba de comenzar.	§ **en el de hoy,** recibid ante todo mi felicitación y la de mi familia, para vosotros y las vuestras, con los mejores deseos en el año que acaba de comenzar.
§ un año que, **si sigue la norma del pasado, puede estar** repleto de acontecimientos importantes, **imprevistos y tal vez preocupantes, que se producen en el mundo**.	§ un año que **vamos a vivir a continuación del que estuvo** repleto de acontecimientos importantes: **la conmemoración del v centenario del descubrimiento de américa, los juegos olímpicos de barcelona, la exposición universal de sevilla, los actos de madrid como capital europea de la cultura, la conferencia de jefes de estado y de gobierno iberoamericanos...**
§ lo ocurrido en 1991 está en la memoria de todos y sus consecuencias constituyen un aldabonazo a la convivencia de la humanidad. como españoles debemos sentirnos orgullosos de que nuestra nación, identificada con el ideal de la paz, que es el supremo bien de las sociedades, haya mantenido y mantenga un comportamiento vigilante, sin egoísmos ni dudas en cuanto a lo que nos corresponde hacer dentro del concierto internacional.	**§ todos ellos han sido una muestra de la vitalidad de españa, de su capacidad de organización y de su proyección en el mundo.**
§ ello supone que el análisis del periodo recientemente terminado, esté impregnado de un lógico sentimiento de optimismo y de fé hacia los tiempos venideros. porque no estamos ni solos ni aislados y se confía en nuestra capacidad para seguir cumpliendo un papel necesario y digno en europa y en el mundo.	**§ un año, el actual, que si sigue las normas de lo que viene ocurriendo en los últimos tiempos, puede caracterizarse también por novedades imprevistas y tal vez preocupantes de distinto signo que se producen en el mundo y constituyen un aldabonazo a la convivencia de la humanidad.** **.**

La figure 3 montre la ventilation du Tgen ***SegmentsLongs*** parmi l'ensemble des paragraphes du corpus. Une conclusion s'impose : Dans le corpus ***Pascua***, même si l'on peut constater des reprises de séquences longues qui concernent des parties différentes du corpus, les répétitions de séquences longues se produisent principalement entre les allocutions *P92* et *P93*. Le retour au texte assisté par la mise en évidence de ces répétitions nous permet de vérifier que l'allocution de 93 reprend effectivement de larges extraits de celle de 92.

C'est donc l'hypothèse ***H3***, que nous devons retenir pour expliquer le phénomène constaté plus haut. L'allocution de 93 reprend en grande partie celle de l'année précédente. Il reste maintenant à trouver les raisons qui peuvent expliquer ce phénomène.

Une enquête sur les publications originales qui ont servi de base à la constitution du corpus nous apprendra que l'allocution destinée à la cérémonie de 1992, bien que publiée dans les organes de presse, n'a finalement pu être prononcée par le souverain en raison d'un accident corporel dont il a été victime avant la cérémonie de la *Pascua militar* de 1992. Dans ces circonstances, la tentation a été forte pour les rédacteurs de l'allocution de l'année suivante (1993) d'utiliser le travail effectué l'année précédente tout en le modifiant pour le réactualiser.

Le tableau 3 présente une édition parallèle des paragraphes correspondant au début de chacune de ces deux allocutions. Les parties modifiées ont été signalées en caractères gras dans les deux documents. Comme on le voit, les reprises textuelles constituent de longs fragments du premier texte. Les séquences rajoutées ou supprimées dans l'allocution de 1993 vont d'une séquence de quelques occurrences au paragraphe entier.

Etait-il indispensable de mettre en œuvre une méthodologie faisant intervenir des calculs aussi compliqués pour arriver à la conclusion qu'une des allocutions reprend simplement de larges extraits de la précédente ? Cette question est plus compliquée qu'il n'y paraît au premier abord. En effet, une fois repérée, la similarité des paragraphes qui résultent de recopies totales ou partielles semble tout à fait évidente. Cependant, l'expérience montre que le rituel énonciatif propre à ce type d'intervention complique la distinction entre des tournures et des formules difficilement évitables dans ce genre d'allocution et ce qui constitue manifestement des reprises *in extenso* d'un texte antérieur.

Par ailleurs, une fois le corpus mis à disposition sur support informatisé, le repérage des séquences répétées constitue de nos jours une opération relativement facile à mettre en œuvre pour le chercheur, même si elle entraîne pour la machine un volume de calculs relativement important.

4. Une méthode de repérage du taux des reprises textuelles

Sur la figure 4 on a tracé deux courbes d'accroissement du vocabulaire. La première (en dessous sur la figure) concerne l'ensemble composé des deux allocutions *P92* et *P93* mises bout à bout. La seconde concerne deux allocutions, correspondant à d'autres périodes du corpus et choisies en raison de leurs tailles comparables.

Ce rapprochement permet de localiser des portions du texte pour lesquelles l'accroissement est très faible et dont on peut supposer qu'elles correspondent à des reprises importantes d'un texte précédemment énoncé. On voit ici que la fin du texte de 1992 semble contenir peu de redites, si l'on en juge par la reprise régulière de l'accroissement du vocabulaire.

5. Conclusion

La démarche proposée permet donc de comprendre les raisons de l'anomalie repérée sur la courbe d'accroissement du vocabulaire. La suite des opérations textométriques convoquées pour repérer les reprises textuelles d'une allocution à l'autre constitue une méthode largement applicable à d'autres séries textuelles.

A la phase de repérage direct, appuyée sur la localisation des segments répétés les plus longs, succède une phase de remise en contexte des répétitions constatées qui débouche sur une édition contrastée des textes repris et de textes originaux.

6. Références

Labbé D., Hubert P. « Vocabulary Richness », in *Lexicometrica n°0*, 1997 http://www.cavi.univ-paris3.fr/lexicometrica/article/numero0/DLVocRich.html

Lamalle C., Salem A., « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *Actes des 6emes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002 http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/lamalle_salem.pdf

Pineira-Tresmontant C., «Un pas en avant un pas en arrière » in *Le poids des mots*, *Actes des 7emes journées d'analyse statistique des données textuelles*, Presses universitaires de Louvain, Louvain-la-neuve, 2004 http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_085.pdf

Pineira-Tresmontant C., « Persuasion ou tradition, la communication du roi d'Espagne », in , actes du colloque *Argumentation, Manipulation, Persuasion : ressources linguistiques et stratégies discursives*, Université de Pau, 2005 (à paraître)

7. Fonctionnalités Lexico3 utilisées dans cette navigation

N°	*Fonctionnalité*	Résultat
6	Partition (clé a, pour année)	
5	Principales car lexicom (PCLC)	*Tableau 2*
5.5	Accroissement du vocabulaire (corpus)	*Figure 1*
5.5	Accroissement du vocabulaire (P92, P93)	*Figure 2*
5.5	Accroissement du vocabulaire ([P92,P93] et [P89,P90])	*Figure 4*
4	Segments Répétés (seuil minimal =2)	
	Sélection d'un Type (occurrence de SR long>10)	
7	Carte des sections (paragraphes, présence SR de long>10)	*Figure 3*

Qu'en pensent les Chinois ?

Essai d'exploration de l'opinion publique chinoise à travers des documents disponibles sur la toile.

[Bad karma]

Liangcai Shen, André Salem[31]
liangcaishen@gmail.com, salem@msh-paris.fr

Résumé : Les nombreux moyens d'expressions liés aux technologies du web deviennent chaque jour plus accessibles aux citoyens chinois désireux d'exprimer leurs réactions à propos de sujets d'actualité. A propos d'un incident médiatique entraîné par les propos d'une célèbre actrice américaine, après une catastrophe naturelle survenue en Chine, nous avons cherché à mettre à jour quelques-unes des dimensions de la réaction suscitée par ces propos dans l'opinion publique chinoise. Pour cette première étude, nous avons choisi de comparer quelques échantillons de textes publiés sur la toile par la presse officielle, des textes relevés sur des blogs personnels et des interventions collectées sur des forums publics. Cette première démarche, aux dimensions modestes, illustre la possibilité et l'intérêt du type d'enquête proposé.

Mots-clés : Etude d'opinion, médias, textométrie

Abstract : The expressions of many ways related to web technologies become ever more accessible to Chinese citizens wishing to express their opinions about topical issues. About an incident caused by the remarks of an american actress, after a natural disaster in China, we try to explore the dimensions of the reaction to these words in chinese public. We chosed to compare a few samples of texts published on the web by the official press, texts recorded on personal blogs and responses collected from public forums. This first approach illustrates the ability and interest of the type of proposed investigation.

Keywords : Opinion studies, media analysis, textometrics,

摘要：如今的网络科技五花八门，一日千里，中国老百姓现在可以更轻松自如，随心所欲地表达自己对时事热点的各种意见和看法。中国汶川重大自然灾害之后，美国某影星的某些个人看法在华人世界中激起了轩然大波。为此，作者分别选取了公布于官方新闻网站上的新闻与评论，记录在个人博客中的文章以及公众论坛中的回帖等内容作为比较样本，并运用词量法来分析和对比中国社会各阶层民众的语料语库，进而实现了跟进该事件在国民大众舆论中不同反响，为民众观点探究开拓了新的方法。此端倪之作，文集容量虽小，但该方法却充分证明了它在社会舆论调查上的新潜力，并且也拓展中文自然语言处理的新领域。

关键词：词量法，媒体，观点探究，民众意见，百姓观点，抵制

Comment cerner les sentiments suscités dans les différentes couches de l'opinion publique chinoise par un événement médiatique dont la répercussion a été planétaire, compte tenu des moyens modernes de circulation de l'information ? S'agissant d'un pays aussi étendu et aussi diversifié que la Chine, la question peut paraître naïve voire dénuée de sens. Le titre quelque peu provocateur que nous avons donné à cette étude souligne en fait le caractère chimérique d'une telle entreprise conçue comme une tentative d'exploration exhaustive aboutissant à des conclusions nettes et clairement formulées.

Cependant, par-delà la multiplicité et la diversité des réactions individuelles susceptibles d'être observées au sein d'un peuple qui compte plus d'un milliard d'habitants, dont les

[31] Les auteurs remercient Jean-Maxence Granier, de la société *Think-Out*, pour ses précieuses suggestions et ses encouragements dans la réalisation de cette étude.

langues, les coutumes, sont extrêmement variées, les technologies liées au web offrent désormais la possibilité d'observer, et ce quel que soit l'endroit où l'on se trouve sur la planète, des interactions entre citoyens chinois qui échangent des avis par ce biais. Il est bien entendu que cet échantillon de réactions, limité aux intervenants qui ont accès à ces nouveaux médias, ne constitue en aucun cas une photographie fidèle de l'ensemble de la société chinoise moderne. Cependant, de part le fait qu'elles aient été produites par des acteurs de la vie économique chinoise, ces réactions constituent un matériau extrêmement précieux pour ceux qui s'intéressent à l'étude de l'opinion publique chinoise.

Dans ce qui suit, nous commencerons par rappeler le contexte des évènements qui ont été au centre de notre enquête (§1). Nous présenterons ensuite les différents supports que notre stratégie de fouille de textes à permis de repérer en liaison avec notre recherche (§2). La section suivante (§3) est consacrée au dépouillement du corpus. La dernière section (§4) analyse les emplois contextuels de la forme *boycott,* particulièrement fréquente dans les forums.

1 Contexte de la recherche

S'exprimant à l'occasion du festival de cinéma de Cannes, le 24 mai 2008, quelques jours après le tremblement de terre survenue en Chine[32], l'actrice américaine Sharon Stone a tenté d'établir un parallèle entre cette catastrophe naturelle et l'action politique de l'État chinois au Tibet, possiblement responsable, selon elle d'un, d'une altération du *karma* commun aux chinois. La connotation particulièrement sensible de ce type de conclusions dans le monde sinophone, qui implique, d'une certaine manière, une punition *méritée* par ses victimes, a immédiatement suscité de très vives réactions dans l'opinion publique chinoise.

Dans cette étude nous avons tenté une première exploration des réactions à cet événement à partir des textes accessibles sur l'Internet. Dans un premier temps, nous avons interrogé deux moteurs de recherche (***Google*** et ***Baidu***) pour localiser les documents qui contenaient à la fois les termes : 莎朗－斯通 Shalang-Sitong (Sharon Stone en caractères chinois), et 四川 (Sichuan, région dans laquelle a eu lieu le tremblement de terre 汶川 Wenchuan). Au vu des résultats, nous avons constitué un premier corpus à partir de trois types de sources textuelles différentes que nous avions identifiées :

- des sites d'information en ligne (presse, agences, etc.) ;
- des textes présentés sur des blogs personnels ;
- des réactions individuelles sur des forums ouverts au public par des sites très fréquentés.

Bien entendu, dans cette première étude, qui ne porte que sur un corpus restreint, nous ne prétendrons pas rendre compte de la totalité des réactions suscitées à cette occasion dans l'ensemble de la population chinoise. Il nous semble cependant qu'elle montre la possibilité de recueillir et de classer une certaine variété de réactions qui diffèrent largement selon le media utilisé et dont certaines présentent une fréquence importante au sein d'un même média. Nous nous proposons essentiellement de montrer la possibilité de réaliser de manière relativement simple, une enquête de ce type.

[32] Une série de tremblements de terre survenus autour du 12 mai 2008, dans la région de Si Chuan (Chine), a causé la mort de plus de 69 000 personnes et entraîné de très importantes destructions dans toute la région. Ces circonstances ont été à l'origine d'une campagne nationale et internationale de solidarité avec les victimes.

Points de repères

- *Les propos incriminés (Cannes 24 mai 2008)*[33]

> *And I have been concerned about... oh… how should we deal with the Olympics because they haven't been nice to the Dalai Lama, who is a good friend of mine.. and the earthquake…and all the stuff happened… I think … is that karma ? when you're not nice that the bad things happen.*

- *Le karma*[34] *:*

(sanskrit कर्म, de la racine *kri*, *acte*, *action*) est un terme utilisé dans plusieurs religions orientales. Le karma désigne le cycle des causes et des conséquences lié à l'existence des êtres sensibles. Le *karma* est la somme de ce qu'un individu a fait, est en train de faire ou fera. Dans les religions incorporant les concepts de réincarnation, les effets de ces actes *karmiques* se répercutent sur les différentes vies d'un individu. Chaque être y est responsable de son karma et donc de sa sortie du Samsara.

- *La notion de* ***« karma négatif »*** *:*

La traduction chinoise diffusée dans les médias à partir de cette déclaration initiale : 报应 *bao ying* . (*karma négatif),* accentue peut-être, en les synthétisant, le caractère blessant des propos tenus par l'actrice. L'expression possède une connotation particulièrement négative dans le monde sinophone de *mauvaise conséquence justement méritée (punition méritée)*.

Chronologie sommaire

avril2008 : agitation suivie d'une répression dans la province chinoise du Tibet.
incidents et manifestations contre le gouvernement chinois sur le parcours de la flamme olympique dans plusieurs pays occidentaux.
tremblement de terre dans la province de Si Chuan (70 000 victimes)

mai 2008 : déclarations de Sharon Stone au festival de Cannes (cf. supra)
Vives réactions dans la presse, sur les blogs et les forums chinois
déclaration de Dior Chine se désolidarisant de l'opinion de Sharon Stone

juin 2008 : excuses officielles de Sharon Stone
la campagne de réactions se poursuit sur les forums

août 2008 : ouverture des jeux olympiques à Pékin

---

[33] Propos retranscrits d'après la vidéo enregistrée pendant l'interview et postée, entre autres, sur le site http://lionelshen.free.fr/Labo/stage/ST_Karma_interview.flv

[34] Les auteurs se sont largement inspirés de l'article *karma* de l'encyclopédie en ligne *Wikipédia*.

2 Localisation et présélection des textes

Dans un premier temps, nous avons entrepris de localiser, sur la toile, les textes susceptibles de concerner le débat créé parmi les internautes chinois. Nous avons utilisé de manière complémentaire deux moteurs de recherche : le classique ***Google***, dans sa version www.google.cn[35], mais également le moteur ***Baidu*** 百度[36], réputé plus performant pour la recherche des documents numériques rédigés en chinois.

Figure 1a

Recherche sur ***Google.cn*** à partir du mot-clé (Sharon Stone)

a) en haut les suggestions du moteur de recherche
b) les premiers résultats référencés

[35] Signalons que la version chinoise du moteur www.google.cn est placée sous le contrôle effectif des autorités chinoises.

[36] Le moteur de recherche ***Baidu*** a été créé par des chercheurs sinophones expatriés aux Etats-Unis. En Chine continentale, sa popularité dépasse largement celle de son concurrent *Google*.

Les réponses fournies par les deux moteurs présentaient une grande intersection pour ce qui concerne les sites officiels (sites de presse, etc.). Comme prévu, la couverture du moteur ***Baidu***, s'est révélée plus importante pour ce qui concerne les blogs et les forums.

Le moteur ***Google*** nous fournit de très nombreuses références concernant ce débat à partir du seul mot-clé *Sharon Stone*. La recherche promptée par ***Google***, nous propose de choisir entre :

- Sharon Stone 1 160 000 résultats indexés, 1ère suggestion
- *Sharon Stone et ses films* 15 600 résultats indexés, 2ème suggestion
- *Sharon Stone, le tremblement de terre à Sichuan cause du karma négatif des chinois* 78 700 résultats indexés, 3ème suggestion

C'est cette dernière suggestion qui correspond manifestement le mieux à notre recherche.

Figure 1b

Recherche sur le moteur ***Baidu*** à partir des mots-clé (*Sharon Stone* et *Sichuan*)
(premiers résultats référencés)

À partir des deux mots-clé : *Sharon Stone* et *Sichuan*, le moteur ***Baidu*** nous fournit une liste de références qui, si elle recoupe largement la précédente liste fournie par ***Google.cn*** en ce qui concerne les sites officiels, est beaucoup plus abondante pour ce qui concerne les sites de type « forum ».

Parmi les trois types de médias qui apparaissent dans les références (presse en ligne, blogs, forums) nous avons choisi, pour cette étude exploratoire, de sélectionner un échantillon de sites plus particulièrement référencés par les moteurs de recherche[37].

[37] Le calcul de l'indice de référencement (*ranking*) qui sert à trier les sites dans les résultats d'un moteur de recherche s'appuie en principe sur le nombre des consultations effectuées sur chacun des sites.

Nous avons choisi de retenir pour notre étude les textes publiés sur l'Internet entre le 26 mai 2008 (date des premières réactions) et le 02 octobre 2008 (clôture du fil de discussion sur ce sujet sur les forums observés). Nous avons privilégié les sites les plus fréquentés par les internautes à partir des indices de fréquentation calculés par les différents moteurs de recherche (ranking).

2.1 La presse en ligne

Comme partout ailleurs, les grands quotidiens nationaux chinois et les agences de presse entretiennent des sites informationnels sur l'Internet qui leur permettent de mettre leurs principales publications à la disposition des internautes dans des délais relativement courts. Nous avons sélectionné vingt-sept mis en ligne après leur parution par ces grands organes de la presse officielle.

Le site de l'agence 新浪 Sina (Nouvelle vague) et celui de l'agence 新华社 Xinhua (Chine nouvelle) ont constitué nos principales sources lors de cette sélection de notre volet de *Presse en ligne.* À partir des sites sélectionnés par les deux moteurs de recherche utilisés, nous avons retenu 27 articles signalés comme ayant été le plus souvent consultés par les internautes.

2.2 Les blogs

Dans le contexte chinois comme dans le contexte francophone, le concept *blog* peut recouvrir des situations très différentes : un journal intime assumé ou anonyme, un journal d'opinion tenu par un journaliste, les échanges quotidiens d'une classe de collège, une œuvre littéraire collective en construction, etc. Comme partout dans le monde, le phénomène connaît en Chine un immense succès grâce à une grande facilité de publication en ligne, une relative tolérance éditoriale et une grande capacité d'interaction avec le lectorat.

Le blog est en général édité et mis à jour par un auteur ou un groupe d'individus identifiés qui ne donnent que très rarement aux lecteurs potentiels la possibilité de s'exprimer à leur tour sur le site du blog. Le nombre de lecteurs d'un blog surpasse souvent celui des lecteurs d'une publication traditionnelle sur papier. En très peu de temps, certains blogs sont devenus extrêmement fréquentés[38] au sein de la communauté des internautes chinois en Chine et à l'étranger).

Nous avons sélectionné vingt-six blogs parmi les plus fréquentés, nous avons veillé à rassembler des opinions différentes autant qu'il se pouvait. Ces blogs nous ont fourni un matériau à peu près comparable (du point de vue du volume de texte) à la partie sélectionnée pour représenter la presse dans notre corpus.

Signalons enfin que les moteurs de recherche proposent des outils spécifiques pour la recherche des blogs (*blogsearch.google.cn, blogsearch.baidu.com)* qui nous ont permis de localiser ces derniers sans difficulté.

[38] On trouvera plus loin l'exemple d'un blog consulté par plus d'un million d'utilisateurs au cours de la semaine considérée.

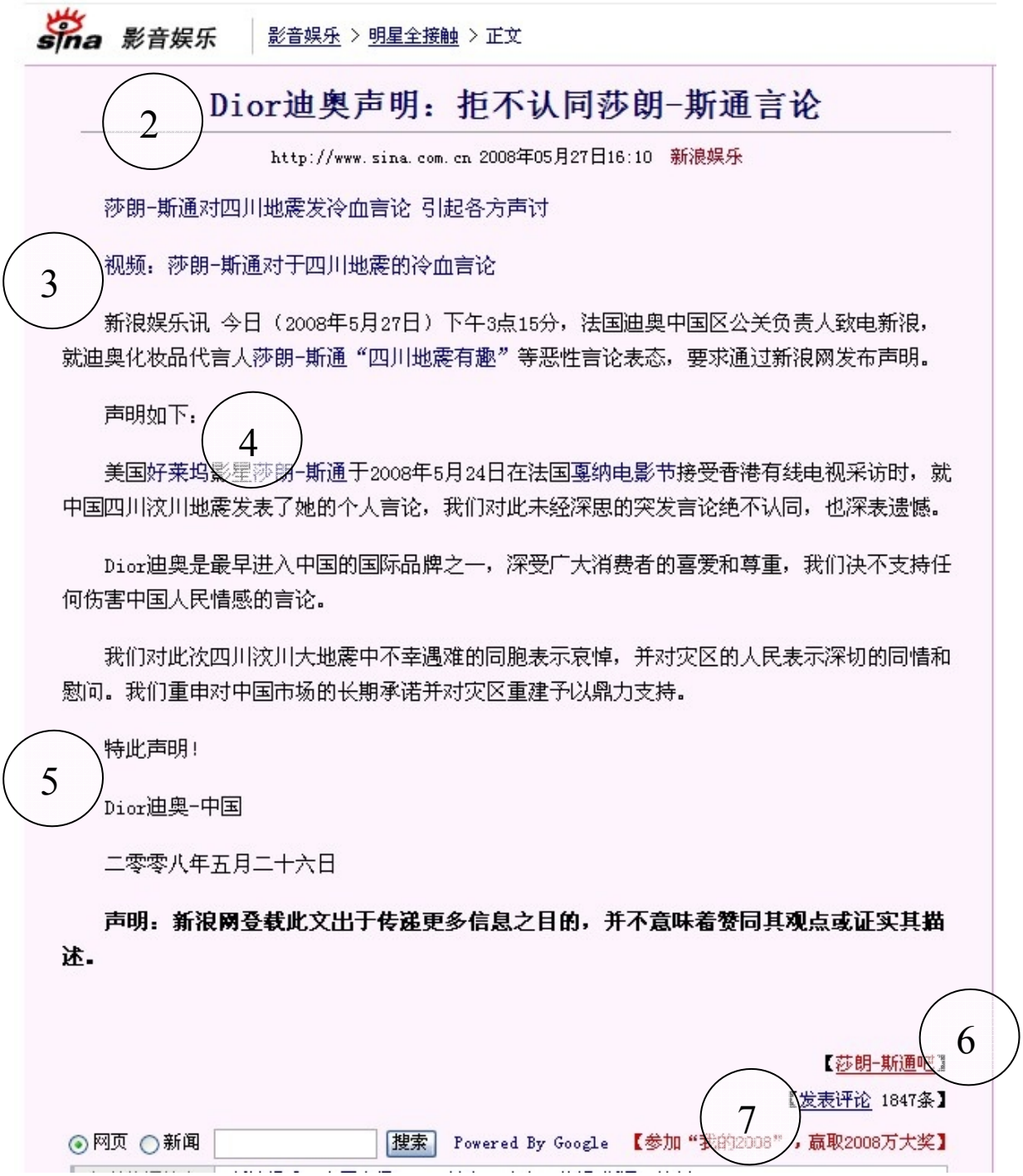


Figure 2

Volet *Presse* du corpus ***StoneKarma***
Déclaration de *Dior-Chine* reproduite par l'agence *Sina*
Retranscription intégrale de la déclaration de Dior-Chine

Figure 3

Volet ***Blogs*** du corpus ***StoneKarma***
Les propos de l’actrice reproduits sur le blog de HU Jianli

14

15

16

17

18

19

20

Traduction de la dernière intervention (IP 59.61.*.*) :
Dior, si tu ne changes pas de représentante, quitte la Chine !

Figure 4

Volet *Forum* du corpus ***StoneKarma***
Exemples de discussion sur le forum *NetEase*

Guide de lecture pour les figures 3 à 5

Principaux composants du site de l'agence *Sina*

1. Nom du site Internet (Sina)
2. Titre de l'article (Déclaration de Dior Chine)
3. Hyperlien de la Vidéo de l'interview de Sharon Stone
4. Le corps de la déclaration de Dior se désolidarisant de l'actrice
5. Datation et signature de Dior Chine
6. Proposition du thème de la discussion
7. Nombre de réactions (1847 fois à ce jour)

Principaux composants du blog de HU Jianli

(1 million de visiteurs dans cette période)

8. Nom de l'auteur du blog
9. De haut en bas : 1) Notation du blog numéro 1 (fréquentation globale) [39];
2) Notation du blog numéro 2 (mise à disposition et réactions),
3) Nombre de visites (1 067 602 dans la période)
10. Vidéo de l'interview de Sharon Stone
11. Retranscription des propos incriminés
12. Déclaration de Dior se désolidarisant de l'actrice
13. Commentaires de l'auteur du blog

Principaux composants du forum du site de *Netease*

14. Nom du forum du site (Netease – www.163.com)
15. Thème de la discussion (Les propos de ST sur le séisme à Sichuan provoquent des appels à sanctions. Avec des extraits vidéos)
16. Nombre de réponses (16 069)
17. Numéro IP de l'internaute (partiel, ne permettant pas de l'identifier totalement, signalant cependant que le site est capable de localiser l'émetteur)
18. Nombre de ceux qui ont voté « pour » cette proposition (ici : 1048 pour)
19. Nombre de ceux qui ont voté « contre » cette proposition (ici : 147 contre)
20. Citation de propos d'un internaute s'étant exprimé précédemment, reprise dans la réaction d'un internaute suivant.

[39] Cette note de *popularité* est attribuée aux blogs individuels par le gestionnaire du site *Sina* en fonction du nombre de visites reçues par chacun d'eux.

2.3 Les forums

Un *forum en ligne* est un site d'échanges entre internautes se situant au même niveau du point de vue éditorial. Les discussions y prennent place sous la forme de « fils » de messages, publication instantanée ou différée ; cette publication est souvent durable, car les messages ne sont pas effacés. Elle est par nature le fait de plusieurs auteurs. Dans certains forums à inscription, les messages sont modifiables a posteriori par leurs auteurs.

Les fonctionnalités offertes par les différents forums (citation d'un point de vue précédemment exprimé, création d'intertitres, mise en page/indentation particulière, modération des droits d'accès, a priori ou a posteriori...) peuvent varier d'un forum à l'autre : certains forums ne permettent que de contribuer de manière ponctuelle à un sujet discussion, tandis que d'autres permettent de répondre plus longuement à un message particulier, voire à un paragraphe particulier contenu dans ce message.

Pour représenter le type de média *forum*, nous avons sélectionné l'ensemble du fil de discussion *: L'actrice Sharon Stone considère que le tremblement de terre de Sichuan est une conséquence d'un mauvais karma, ce qui entraîne de vives réactions* qui se sont développé sur le site de l'agence NetEase (www.163.com)[40]. Nous avons choisi ce site, parmi d'autres parce qu'en dépit d'un thème nominal au caractère réprobateur, il présentait, à première vue, une discussion beaucoup plus ouverte à des opinions variées que des sites concurrents. Comme on le verra plus loin, l'intégrale de la discussion, au sein de laquelle nous nous sommes refusés à faire des sélections, présentait un volume beaucoup plus important que les deux autres volets du corpus.

Nous appellerons désormais ***StoneKarma*** le corpus ainsi rassemblé. Ces trois volets, prélevés sur des supports électroniques de différents types, englobent *grosso modo* trois sources qui peuvent prétendre représenter en partie l'opinion publique chinoise. Le choix de ces différents supports permet de mieux cerner l'hétérogénéité de cette opinion publique qui s'exprime sur le web. On peut supposer, a priori, que la presse représente, dans la plupart des cas, l'opinion officielle des autorités chinoises. Les blogs et les forums fournissant une approche moins contrôlée de l'opinion des citoyens.

3 Dépouillement quantitatif du corpus

L'ensemble du corpus compte 512 806 caractères chinois (balises comprises) que le segmenteur isole en 208 707 occurrences de mots chinois[41]. On peut diviser le corpus en 16 953 paragraphes. Ces paragraphes correspondent à des retours à la ligne dans les textes de presse et dans les blogs et à des successions de tours de paroles.

Les textes rassemblés dans le volet *Presse* sont au nombre de 27, les textes de blogs au nombre de 26 et le volet forum est constitué de 3 023 interventions individuelles.

---

[40] Le nom du site : ***163*** constitue un jeu de mots à partir de la forme phonétique de l'expression « tout va bien /ou avoir le vent en poupe » (一路平安, yi lu pin an).

[41] La notion de « mot chinois » et la segmentation automatique en « mots » seront précisés plus loin (§2).

Tableau 1

Exemple de réactions dans la partie *forum* du corpus

Un forum : Le forum du site Netease (163.com)

<media=forum>
网易 论坛 话题 ： 美 影星 莎朗斯 通称 四川 地震 是 报应 引 声讨
评论
网易 广西 桂林 网友 [moqingli0317] ：
2008-10-02 13 ： 11:06 发表#
莎朗斯 的 B 都 给 别人 操 烂 了 ！ **1**
网易 四川 资阳 网友 ip ： 222.213 . * . * ：
2008-08-27 23 ： 28:26 发表#
地震 的 那 一个 星期 我 根本 就 不敢 看 电视 ， 因为 看 一次 就要 哭 一次 ， 不想 第二天 上班 的 时候 红 红肿 肿 的 。 没 想到 听到 那个 老 女人 这样 称 四川 地震 有趣 ！ 天 啦 ！ 这 都 是 人 说 的 说 吗 ？ 本来 迪奥 这个 品牌 的 东东 我 一直 都 在用 的 ， 现在 开始 ！ ！ ！ ！ 从今 以 后 ！ ！ ！ ！ ！ ！ ！ 坚决 抵制 法国 迪奥 化妆品 ， 莎朗斯通 的 影片 ！ ！ **2**
网易 重庆 永川 网友 ip ： 222.181 . * . * ：
2008-06-10 12 ： 11:02 发表#
网易 加拿大 网友 （ 70.69 . * . * ） 的 原 贴 ：
大家 团结 起来 ， 让 她 知道 中国 人民 的 力量 ！ 坚决 抵制 法国 迪奥 化妆品 （ 莎朗斯通 代言 ） ， 莎朗斯通 的 影片 ！ ！
网易 辽宁 大连 网友 （ 116.3 . * . * ） 的 原 贴 ：

Sur le forum, les intervenants s'identifient par un numéro IP[42]. Les propos sont parfois exprimés avec certaine retenue, parfois avec une grande violence, n'excluant pas la vulgarité :

1. 莎朗斯 的 B 都 给 别人 操 烂 了 !

   Sharon Stone a été b... jusqu'à la destruction de son s... !

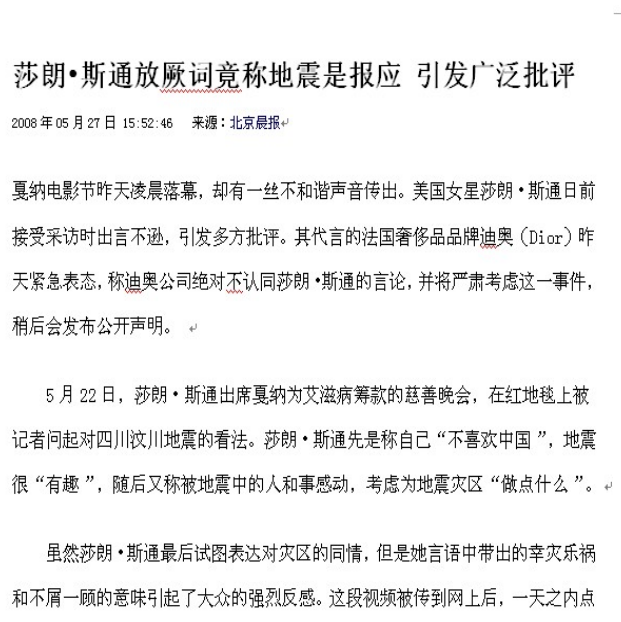
2. 本来 迪奥 这个 品牌 的 东东 我 一直 都 在用 的 , 现在 开始 ！ ！ ！ ！ 从今 以后 ！ ！ ！ ！ ！ ！ 坚决 抵制 法国 迪奥 化妆品 , 莎朗斯通 的 影片 ! !

   Jusqu'à présent, j'utilisais assez souvent les produits Dior, à partir de maintenant !!!! Et dorénavant !!!!!!! Je les boycotterai totalement ainsi que les films de Sharon Stone !!

3.1 Segmentation du texte

Les comparaisons textométriques supposent que l'on définisse des unités de décompte dont on étudie ensuite les variations de fréquence au sein des différentes parties du corpus. Pour pourvoir mettre en oeuvre des comparaisons textométriques, on utilise des outils informatiques qui permettent de découper automatiquement les unités du texte avec lesquelles on pourra s'en servir de leurs occurrences et spécificités.

[42] Notons que ce numéro d'identification, avant tout destiné à l'identification de la machine utilisée sur l'Internet pour des satisfaire des impératifs techniques de transfert de données, peut également être utilisé pour localiser l'internaute sur le web avec une précision plus ou moins élevée.



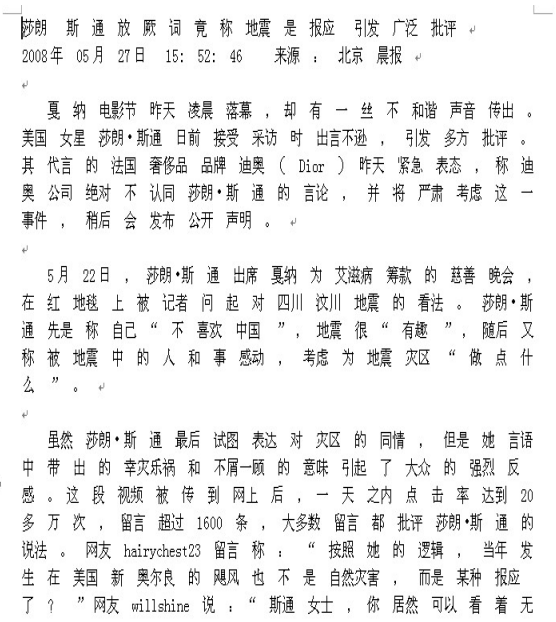
莎朗•斯通放厥词竟称地震是报应 引发广泛批评

2008年05月27日 15:52:46 来源：北京晨报

戛纳电影节昨天凌晨落幕，却有一丝不和谐声音传出。美国女星莎朗•斯通日前接受采访时出言不逊，引发多方批评。其代言的法国奢侈品品牌迪奥（Dior）昨天紧急表态，称迪奥公司绝对不认同莎朗•斯通的言论，并将严肃考虑这一事件，稍后会发布公开声明。

5月22日，莎朗•斯通出席戛纳为艾滋病筹款的慈善晚会，在红地毯上被记者问起对四川汶川地震的看法。莎朗•斯通先是称自己“不喜欢中国”，地震很“有趣”，随后又称被地震中的人和事感动，考虑为地震灾区“做点什么”。

虽然莎朗•斯通最后试图表达对灾区的同情，但是她言语中带出的幸灾乐祸和不屑一顾的意味引起了大众的强烈反感。这段视频被传到网上后，一天之内点



莎朗 斯通 放厥词 竟称 地震 是 报应 引发 广泛 批评
2008年 05月 27日 15: 52: 46 来源： 北京 晨报

戛 纳 电影节 昨天 凌晨 落幕 ， 却 有 一 丝 不 和谐 声音 传出 。 美国 女星 莎朗•斯通 日前 接受 采访 时 出言不逊 ， 引发 多方 批评 。 其 代言 的 法国 奢侈品 品牌 迪奥 （ Dior ） 昨天 紧急 表态 ， 称 迪奥 公司 绝对 不 认同 莎朗•斯 通 的 言论 ， 并 将 严肃 考虑 这 一 事件 ， 稍后 会 发布 公开 声明 。

5月 22日 ， 莎朗•斯 通 出席 戛纳 为 艾滋病 筹款 的 慈善 晚会 ， 在 红 地毯 上 被 记者 问 起 对 四川 汶川 地震 的 看法 。 莎朗•斯 通 先是 称 自己 “ 不 喜欢 中国 ” ， 地震 很 “ 有趣 ” ， 随后 又 称 被 地震 中 的 人 和 事 感动 ， 考虑 为 地震 灾区 “ 做 点 什 么 ” 。

虽然 莎朗•斯 通 最后 试图 表达 对 灾区 的 同情 ， 但是 她 言语 中 带 出 的 幸灾乐祸 和 不屑一顾 的 意味 引起 了 大众 的 强烈 反 感 。 这 段 视频 被 传 到 网上 后 ， 一 天 之内 点 击 率 达到 20 多 万 次 ， 留言 超过 1600 条 ， 大多数 留言 都 批评 莎朗•斯 通 的 说法 。 网友 hairychest23 留言 称 ： “ 按照 她 的 逻辑 ， 当年 发 生 在 美国 新 奥尔良 的 飓风 也 不 是 自然灾害 ， 而是 某种 报应 了 ？ ” 网友 willshine 说 ： “ 斯通 女士 ， 你 居然 可以 看 着 无

Traduction : Sharon Stone a mis en relation le tremblement de terre avec un mauvais Karma : Ceci a déclenché de nombreuses critiques 27 mai 2008 15:52:46 Source: Beijing Morning News

Le festival de Cannes a pris fin hier matin dans le calme. Quelques jours avant, l'actrice américaine Sharon Stone avait fait des déclarations brutales dans une interview. Ceci a déclenché de multiples critiques. La marque de luxe *Dior* dont elle est la représentante a déclaré hier que la société Dior se désolidarisait des propos tenus par Sharon Stone, et qu'elle envisageait de donner une suite sérieuse à cet incident. Dior fera une déclaration publique ultérieurement.

Le 22 mai à Cannes, Sharon Stone a participé à un gala de charité pour la lutte contre le sida. Sur le tapis rouge elle a été interviewée par un journaliste à propos du tremblement de terre à Sichuan. Tout d'abord, Sharon Stone a déclaré qu'elle « n'aime pas trop la Chine », le tremblement de terre était « intéressant », puis elle a dit qu'elle est touchée par les gens et les événements du tremblement de terre, également, elle envisage « faire quelque chose » pour la zone frappée par le séisme. /…/

Figure 5

Exemple d'un fragment de presse extrait du journal *Beijing Morning News* avant (à gauche) et après segmentation en mots par le logiciel ***Haylanda*** (à droite) suivi de la traduction en français du début de l'extrait (en bas du tableau).

Si la notion de *mot* est bien définie dans les grammaires chinoises, l'écriture chinoise n'intègre pas d'espace entre les unités lexicales. Les lecteurs chinois appréhendent les textes en découpant la chaîne textuelle en unités distinctes, à partir de leurs propres connaissances linguistiques. Les mots chinois sont composés d'un à quatre sinogrammes. Un même caractère peut avoir différentes fonctions grammaticales en fonction de son contexte.

Cette particularité constitue une difficulté spécifique pour l'exploitation textométrique des textes chinois. Pour pouvoir découper les textes en unités correspondant plus ou moins à des mots du chinois, nous avons choisi le logiciel de segmentation ***Haylanda*** [43]. On trouvera, ci-dessous les principales caractéristiques lexicométriques des trois sous-ensembles rassemblés dans le corpus ***StoneKarma***.

Tableau 2

Principales caractéristiques lexicométriques
des trois sous-ensembles du corpus ***StoneKarma***

Partie	occurrences	formes	hapax	F. Max
Blog	21538	3987	2161	1703
Forum	158132	7277	3209	7815
Presse	17937	2643	1055	1188

On se gardera d'interpréter directement des différences entre ces caractéristiques textométriques qui peuvent résulter d'artéfacts produits par des méthodes d'encodage et de stockage différents pour chacun des supports rassemblés en un même corpus.

3.2 Comparaisons entre médias

Les trois volets diffèrent, bien entendu par les types d'expression propres à chaque média. Style plutôt soutenu et tournures officielles pour la presse, style plus littéraire mais aussi plus personnel pour l'expression sur les blogs, très grande variété de modes d'expression sur les forums qui peuvent aller d'un style soutenu à des listes d'interjections et d'insultes.

Statut de la répétition dans les trois volets

Les procédures de repérages des segments répétés (suite de formes reproduites à l'identique à différents endroits du corpus) permettent de repérer des reprises de séquences plus ou moins étendues dans chacun des trois volets du corpus.

Sur le tableau 5, on peut voir la répétition de séquences localisées au sein d'articles de presse. Cette répétition vient ici de la reprise fréquente, par les différents journaux, du même texte présentant, au nom de Sharon Stone, des excuses exprimées par son agent :

> */…/ Mes propos déplacés ont blessé le peuple chinois et ont suscité sa colère. Je m'en excuse profondément./…/*

Les commentaires qui accompagnent ce texte contiennent aussi des répétitions de phrases de commentaires et de réactions souvent identiques.

[43] Ce logiciel est en développé par la société Hailanda Segmentation intelligente, version d'essai (海量智能分词研究版 http://www.hylanda.com/).

Tableau 3 :

Exemple de répétition localisée dans deux articles de presse

<p=P12> # 莎朗 - 斯通 道歉 遭 中国 网友 拒绝 外 媒 称 其 是 白痴 <ANNEE="2008"><MOIS="2008_5"><JOUR="2008_5_29"> 04 : 35 据 上海 《 青年报 》 报道 ， “ 我 的 错误 言行 ， 让 中国 人民 感到 悲伤 与 愤怒 ， 我 为 此 深感 歉意 。” 这 是 莎朗 - 斯通 于 昨日 下午 通过 其 经纪人 发给 Dior 公司 的 致歉 文 ， 对于 自己 于 5 月 24 日 在 法国 戛纳电影节 上 发表 的 关于 中国 四川 地震 的 冷血 言论 作出 忏悔 ， 并 表示 愿意 积极 参与 救援 活动 。 。 。 。 。 。 。 。 。 。 。 。 。 。 。 。 昨晚 ， 莎朗 - 斯 通通 过 经纪人 向 其 代言 的 品牌 公司 发 了 一封 致歉 信 ， 称 由于 自己 在 受访 中 的 不当 言行 ， 对 中国 人民 所 造成 的 伤害 ， 感到 非常 的 难过 和 ??歉 。 同时 ， 莎朗 - 斯 通称 ， 对于 此次 自然灾害 中 所有 遭受 不幸 的 受害者 ， 表示 最 深切 的 ??悼 与 同情 。“ 我 的 错误 言行 ， 让 中国 人民 感到 悲伤 与 愤怒 ， 我 为 此 深感 歉意 。 我 再次 强调 ， 我 愿意 积极 参与 任何 关于 中国 地震 灾害 的 援助 活动 , 并 尽全力 帮助 受灾 的 中国 人民 。” 虽然 道歉 信 的 内容 颇 显 其 忏悔 之 意 ， 然而 ， 截至 记者 截稿 时 ， 莎朗 - 斯通 本人 及 其 经纪 公司 并未 有 任何 要 为 中国 地震 捐款 或 捐物 等 所谓 的 救援 行为 ， 道歉 目前 还 只是 停留 在 书面 形式 。
<p=P12> # Les excuses de Sharon refusées par les internautes chinois. Les médias étrangers la qualifient d « imbécile ». <ANNEE="2008"><MOIS="2008_5"><JOUR="2008_5_29"> 4 : 35. Selon les reportages du « Journal des jeunes » de Shanghaï, « *Mes propos déplacés ont blessé le peuple chinois et ont suscité sa colère. Je m'en excuse profondément* » (reproduit 2-3 fois dans différents articles) /.../

On peut voir au tableau 4 une intervention due à un internaute qui manifeste son indignation en reproduisant un grand nombre de fois la même séquence d'injures : *Sharon Stone, tu es une espèce de chienne ; crève* ! Comme on le comprend aisément, ce type de répétition est facilité par les possibilités technologiques de l'écriture électronique (l'internaute a visiblement construit son message à l'aide du *copier/coller* répété plusieurs fois) qui rend ces répétitions incomparablement plus faciles à produire que lors d'un énoncé vocal, par exemple.

Tableau 4 :

Exemple de répétition réalisée sur un forum de discussion à base de copier/coller

<ANNEE="2008"><MOIS="2008_05"><JOUR="2008_05_28">22 : 12 : 17 发表# **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 死 吧 莎朗斯通 你 **莎朗斯通** 你 个 老 母狗 去 死 吧 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 个 老 母狗 去 **莎朗斯通** 你 个 老 母狗 去 死 吧 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 **个** 老 母狗 去 **死** 吧 **莎朗斯通** 你 个 老 母狗 去 死 吧 网易 美国 网友
<ANNEE="2008"><MOIS="2008_05"><JOUR="2008_05_28"> l'IP n° 22 : 12 : 17. x .x dit : Sharon Stone, tu es une espèce de chienne ; crève (répété 25 fois)

Le statut de ces répétitions dans les comptages textométriques se complique d'autant. D'un côté, il n'est pas tout à fait naturel de considérer que ces nombreuses répétitions équivalent à une seule séquence (l'internaute a dupliqué son texte 25 fois, ce qui traduit très certainement une grande irritation de sa part), d'un autre côté, la prise en compte de ces 25 occurrences

d'une même séquence ne peut être mise sur le même plan que 25 occurrences d'une même séquence produites par des individus différents (avec ou sans citation mutuelle).

4 Etude contextuelle de la forme 抵制-(di zhi boycott)

L'étude des *spécificités maximales* (mots particulièrement sur-employés) pour le volet ***forum*** met en évidence un emploi massif du terme : 抵制, (di zhi) par les internautes qui s'expriment sur ce type de support. Le terme 抵制 (di *résister*+ zhi *maîtrise*) correspond plus ou moins au terme occidental de *boycott*. Il apparaît avec une fréquence très élevée (784 occurrences) dans ce volet du corpus, immédiatement après les particules grammaticales qui sont les mots les plus fréquents du corpus.

Nous avons tenté d'analyser l'emploi de cette forme dans le corpus ***StoneKarma*** sous le triple aspect de sa répartition à l'intérieur des textes, de la comparaison de ses contextes immédiats et de la liste des formes lexicales qu'il attire dans son entourage syntagmatique.

4.1 Répartition de la forme

L'histogramme de la figure 6 montre la répartition du terme 抵制 (di zhi, boycott) au sein des trois ensembles de textes rassemblés dans le corpus. Comme nous l'avons signalé plus haut, le terme apparaît très majoritairement dans la partie *forum* du corpus (733 occurrences dans les forums sur 784 au total).

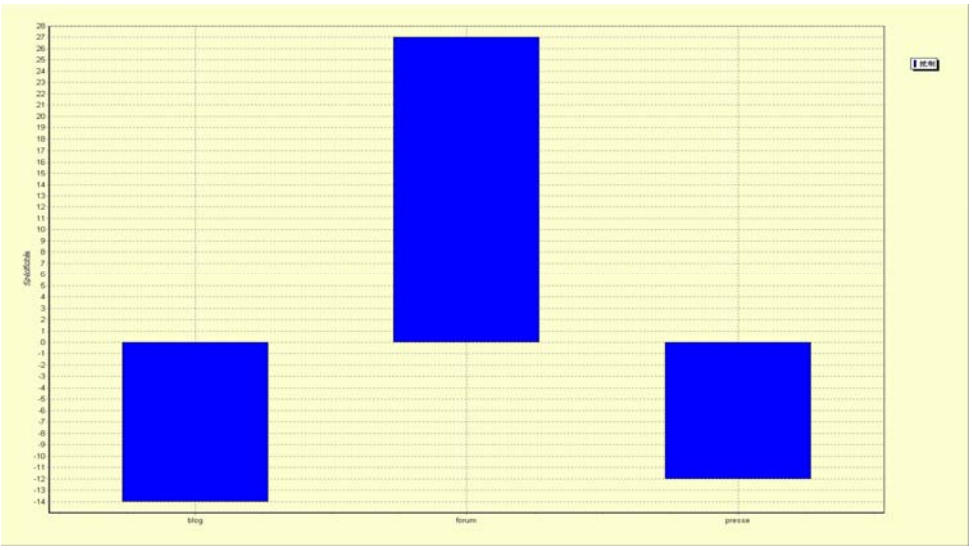


Figure 6 :
Ventilation de la forme 抵制 (boycott)
dans les 3 parties du corpus ***StoneKarma***

Cet emploi privilégié n'est pas sans rapport avec les mécanismes de saturation des messages à l'aide du *copier/coller* dont nous avons présenté un exemple ci-dessus (un même paragraphe pouvant contenir un nombre important des occurrences du terme). Cependant, la disproportion en faveur des forums nous amène à conclure que ce mot trouve une faveur particulière chez les intervenants des forums, alors que les rédacteurs de presse et de blogs, sans doute tenus à une certaine réserve évitent de l'employer trop souvent.

4.2 Contextes

La figure 6 montre, pour chaque média sélectionné, un certain nombre de contextes dans lesquels on retrouve la forme 抵制 (di zhi *boycott*). Comme on le voit, les contextes de cette forme repris dans les articles de presse proviennent le plus souvent de discours rapportés dont les auteurs sont des citoyens que l'on interroge dans le cadre du reportage :

, 并 呼吁 全国 所有 书店 、 音像店 共同 抵制 莎朗 - 斯通 。 记者 昨日 采访 的 重庆 某
Appelons au boycott des produits de Sharon Stone dans toutes les librairies et boutiques
, 引得 华人 世界 震怒 。 网友 一致 呼吁 抵制 莎朗 斯通 代言 产品 。 昨天 下午 , 其
C'est un grand choc pour le monde sinophone, les internautes appellent au boycott de Sharon Stone ainsi que des produits qu'elle représente

Les contextes prélevés sur des forums résultent au contraire de l'expression directe d'un appel au boycott, dont la cible peut varier, de la part des citoyens chinois.

果 DIOR 不 更换 代言人 , 大家 起来 坚决 抵制 DIOR, 大家 行动 起来 网易 湖北 黄石 网友
Unissons-nous pour boycotter DIOR s'il ne change pas de représentant. 我们
必须 抵制 DIOR , 只有 这样 , 才能 让 她 付出 代价
C'est du mépris pour la vie humaine, nous devons boycotter DIOR ! Elle doit le payer !
斯通代言 的 任何 产品 ! ! 山东 人民 坚决 抵制 SBST[44] 的 任何 电影 , 包括 其 代言 的 任何
Boycott de tous les films et les produits de cette conne de Sharon Stone

L'inventaire distributionnel réalisé après la même forme permet de hiérarchiser les entités que les internautes proposent de soumettre à un boycott.

Inventaire distributionnel				*Equivalent français*
415	----	----	抵制 法国	boycott France
2	----	----	抵制 和	boycott et
2	----	----	抵制 美国	boycott USA
2	----	----	抵制 你	boycott toi
8	----	----	抵制 其	boycott ce qu'elle
7	----		抵制 其 代言	boycott de tout ce qu'elle représente
6	---		抵制 其 代言 产品	boycott les produits qu'elle représente
10	----	----	抵制 莎朗	boycott Sharon
2	----	----	抵制 莎朗·斯通	boycott Sharon Stone
40	----	----	抵制 莎朗斯通	boycott Sharon Stone
9	----		抵制 莎朗斯通 代言	boycott de ce qu'elle représente
4	---		抵制 莎朗斯通 的 影片	boycott les films de ST
2	---		抵制 莎朗斯通 所有	boycott de tout ce que
2	----	----	抵制 他	boycott (elle)

Tableau 5 :
Extrait de l'inventaire distributionnel après la forme 抵制 (di zhi, *boycott*)

[44] SBST, initiales en caractères romains de *Shabi* 傻屄 *Sharone Stone*, *conne de Sharon Stone*.

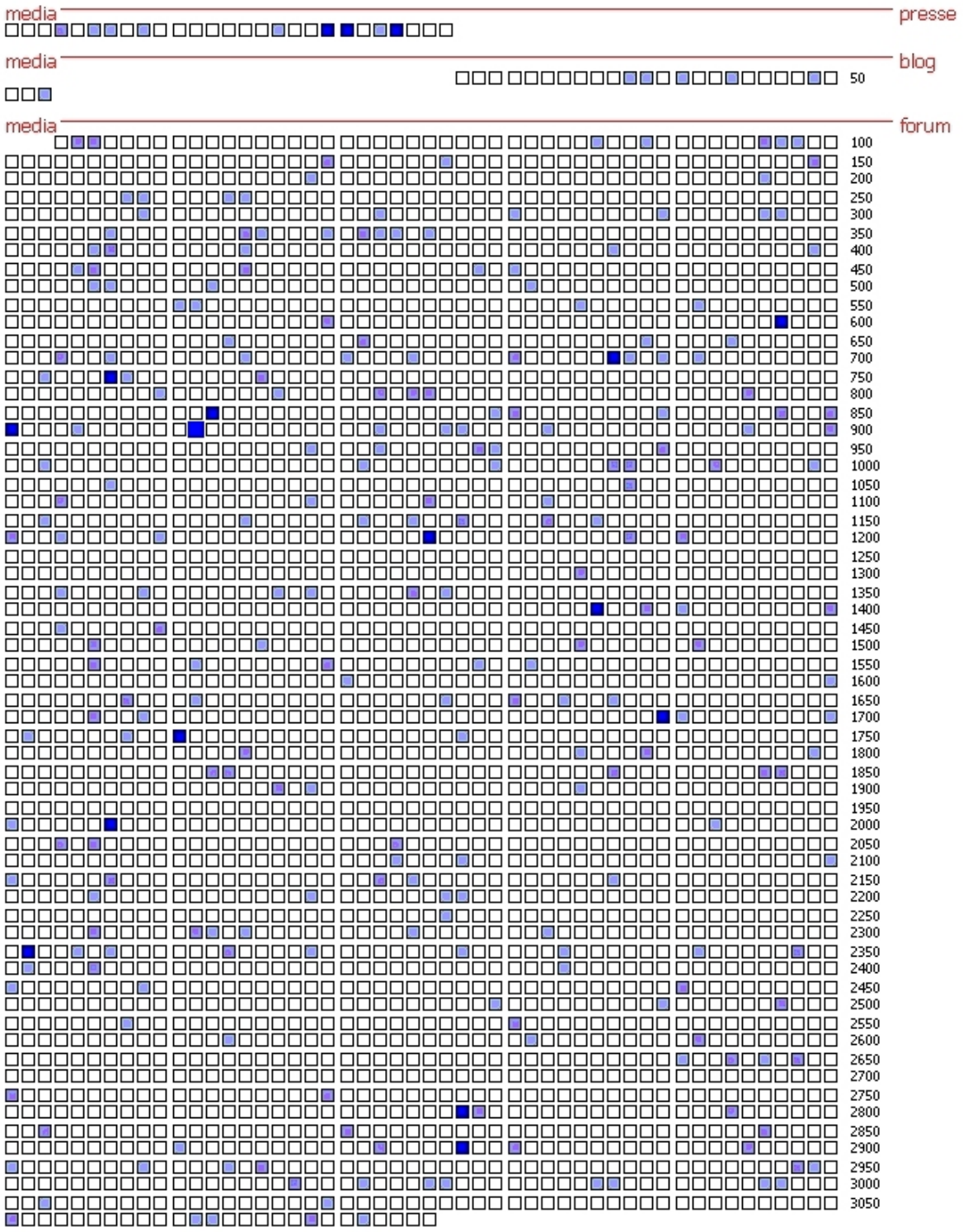


Figure 7 :

Carte des sections réalisées à partir des occurrences
de la forme 抵制 (di zhi, *boycott*)

La carte des sections qui montre la distribution de cette même unité à l'intérieur des sections (paragraphes) découpées dans le corpus permet de localiser cette vision avec une plus grande précision et de vérifier que le terme, outre ses emplois massifs par certains des internautes

repérables par la couleur foncée des paragraphes qui correspondent à leurs interventions, est largement utilisé par un grand nombre d'intervenants.

Guide de lecture pour la figure 7

Dans la carte des sections qui correspond au corpus StoneKarma, les volets correspondant à chacun des médias étudiés sont séparés par une ligne rouge.
Pour chacun des volets, les différentes sources (presse, blogs, forums) sont représentées par un carré.
Les carrés de couleur vive permettent de repérer les sections qui utilisent particulièrement le mot pour lequel la carte a été établie (ici la forme 抵制 di zhi, *boycott*).

Contextes de 抵制 (di zhi, boycott) dans la partie PRESSE (tri après)

有 网友 提出 拒 看 ' 莎 剧 **抵制** 其 代言 产品 ， 得到 了 大多数 人 的 ??
? 愤怒 ， 并 呼吁 中国 人群 起 发动 全面 **抵制** 莎朗 - 斯通 。 香港 / 文汇报 / 报道
， 并 呼吁 全国 所有 书店 、 音像店 共同 **抵制** 莎朗 - 斯通 。 记者 昨日 采访 的 重庆 某
品 重庆 下 架 据 记者 了解 ， 为了 **抵制** 莎朗 - 斯通 不 负 责任 的 讲话 ， 近日 已
， 引得 华人 世界 震怒 。 网友 一致 呼吁 **抵制** 莎朗 斯通 代言 产品 。 昨天 下午 ， 其
丛?： 网易 娱乐 ） 上海 书店 **抵制** 莎朗 - 斯通 影视 作品 全部 下 架 2008 年
UME 院线 的 老板 吴思远 就 公开 表示 要 **抵制** 莎朗 - 斯通 主演 的 电影 。 莎朗 - 斯通
UME 院线 的 老板 吴思远 就 公开 表示要 **抵制** 莎朗 - 斯通 主演 的 电影 。 莎朗 - 斯通

Contextes de 抵制(di zhi , boycott)dans la partie FORUM (tri après)

果 DIOR 不 更换 代言人 ， 大家 起来 坚决 **抵制** DIOR ， 大家 行动 起来 网易 湖北 黄石 网友
?， 是 对 人类 生命 的 漠视 . 我们 必须 **抵制** DIOR ， 只有 这样 ， 才能 让 她 付出 代价
- 27 16 : 21 : 11 发表 # 抵制 ， 坚决 **抵制** ， SB 货 网易 上海 黄浦 网友 ip : 58 . 37
. * . * ） 的 原 贴 : 河北 人民坚决 **抵制** SBST 的 任何 电影 ， 包括 其 代言 的 任何
?代言 的 任何 产品 ！ ！ 河北 人民 坚决 **抵制** SBST 的 任何 电影 ， 包括 其 代言 的 任何
?代言 的 任何 产品 ！ ！ 山东 人民 坚决 **抵制** SBST 的 任何 电影 ， 包括 其 代言 的 任何
190 . * . * ） 的 原 贴 : 辽宁人民坚决 **抵制** SBST 的 任何 电影 ， 包括 其 代言 的 任何
泄?人民 的 力量 ！ ！ ！ 辽宁 人民 坚决 **抵制** SBST 的 任何 电影 ， 包括 其 代言 的 任何
* : 2008 - 05 - 27 17 : 24 : 3 发表 **抵制** SBST 的 任何 电影 网易 江苏 泰州 网友 ip
?出来 ， 公布 到 各个 论坛 。 我们 共同 **抵制** 。 北京 、 上海 的 书店 今天 已经 把 ??
5 - 27 21 : 31 : 45 为 有 默默 的 **抵制** 。 网易 江西 新余 网友 ip : 218 . 64 .
表 # 他妈的 ， 私通 ， 政治 娼妇 坚决 **抵制** 。 网易 上海 杨浦 网友 [wang13386190]
2008 - 05 - 27 21 : 46 : 50 坚决 **抵制** " 私通 " 影片 ， 在 网上 查出 她 的 影片

Contextes de 抵制(di zhi, boycott)dans la partie BLOGS (tri après)

法国 奢侈品 品牌 迪奥 （ DIOR ） 已 被 **抵制** 。 绝大部分 的 民众 意见 都 是 希望 立刻
立即 引发 了 国人 对 莎朗 - 斯通的全面 **抵制** 。 同样 " 有趣 " 的 是 ， 莎朗 - 斯通
锟?表示 : " 她 的 音像制品 也 应 遭到 **抵制** 。" 演艺圈 愤怒 声讨 ， 要 其 道歉
代言 的 DIOR 产品 ， 昨天 不少 网友发 **抵制** ， 以 表 愤怒 。 对此 ， 记者 第一 时间
要 把 这个 产品 列入 抵制 的 黑名单， **抵制** 到 该 品牌 撤销 与 其 合同 为止 ？ ？ 这样
营销 ， 我们 也 要 把 这个 产品 列入 **抵制** 的 黑名单 ， 抵制 到 该 品牌 撤销 与 其
. 全面 封杀 莎朗斯通 的 电影 。 5 坚决 **抵制** 购买 莎朗斯通 代言 的 一切 周边 产品 ！
?莎朗斯通 ， 华人 娱乐圈 也 已经 开始 **抵制** 莎朗斯通 影视 作品 ， 香港 节目 知名 主持人
， 国内 数 以 十万 网民 们 联合发起了 **抵制** 沙朗斯通 的 大 反击 ！ 而 还 在 法国 戛
维护 人类 的 基本 良知 。我 从来不提倡 **抵制** 什么 国家 的 品牌 和 产品 。 但 此时此刻
拇?， 我们 中国 人民 不 欢迎 你 ， 坚决 **抵制** 它 代言 的 所以 产品 ， 不许 它 的 产品
国 必须 封杀 她， 有 良知 的 人 都应该 **抵制** 她。 刘威 : 这样 的 艺人 根本 不

Figure 6 :

Extraits de la concordance de la forme 抵制 (di zhi *boycott*)
dans chacune des trois parties du corpus ***StoneKarma***

On voit sur l'inventaire distributionnel réalisé après la forme 抵制 (boycott/boycotter) que l'objet de l'action de boycott envisagée est prioritairement la France ou les produits français (415 occurrences). Les USA n'apparaissent qu'occasionnellement dans ce contexte. La personne de Sharon Stone elle-même n'est visée que secondairement.

Cooccurrences

La recherche des cooccurrences (i.e. formes de vocabulaire apparaissant plus souvent qu'un modèle de répartition homogène ne le laisserait supposer) à l'intérieur des seuls textes produits sur les forums permet de préciser encore l'objet du boycott prôné par les intervenants sur les forums.

Parmi les segments les plus longs et les plus répétés dans le volet forum du corpus, on repère aisément des séquences en forme de mot d'ordre qui permettent de mieux cerner la nature de la colère exprimée par les internautes.

Forme	Equivalent français	Tot.	Fréq.
的 力量 ！ 坚决 抵制	*boycotter fermement*	379	379
坚决 抵制 法国 迪奥 化妆品	*boycotter les cosmétiques de Dior France*	413	413
大家 团结 起来,让 她 知道 中国 人民 的 力量	*unissons-nous ; pour faire connaître la force du peuple chinois*	377	377

5 Conclusion

La détection ou la fouille d'opinions est un domaine de recherche en plein essor. Ils peuvent se révéler cruciales pour les entreprises et trouve de très nombreux domaines d'applications veille technologique, marketing, concurentielle, études politiques et sociétales. La mise en relation des opinions et sentiments exprimés avec les thèmes sur lesquels ces opinions et sentiments portent est encore un domaine en cours d'exploration, dont les enjeux concernent la transformation des informations extraites des textes en informations structurées en connaissances synthétisées et exploitables.

À partir d'une vive polémique déclanchée dans l'opinion publique chinoise par les propos d'une actrice américaine, nous nous sommes intéréssés à la partie du débat accessible sur le réseau internet.

Utilisant les deux grands moteurs de recherche (Google et Baidu) nous référencer, via un certain nombre de mots-clés les textes les plus consultés par les internautes et relatifs à ce débat.

Dans un second temps, nous avons sélectionné, pour trois types de supports médiatiques identifiés (presse, blogs, forums) des échantillons de texte qui nous ont permis de constituer un corpus d'étude. Ce type d'étude, surtout lorsqu'il est pratiqué sur un échantillon restreint, ne saurait se présenter comme une synthèse des réactions repérables dans l'opinion publique chinoise.

Les traitements textométriques nous ont permis de constater la diversité des réactions exprimées par les internautes et de découvrir une hiérarchie inattendue des thèmes exprimés grâce aux observations sur la diversité des réactions.

Nous pensons avoir montré la possibilité qu'il y a d'accéder à des réactions authentiques vis-à-vis de ce qu'on croit de la liberté d'expressions, exprimées dans trois différentes couches de l'opinion publique chinoise.

6 Références

Lamalle C., Salem A., "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels", Actes *des 6èmes Journées d'analyse des données textuelles,* St Malo, 2002

Lebart L., Salem A., *Statistique textuelle*, Paris, Dunod, 1994, téléchargeable sur le site : http://www.cavi.univ-paris3.fr/lexicometrica/livre/st94/st94-tdm.html

Miao J., Salem A., Comparaisons textométriques de traductions franco-chinoises, in *Explorations textométriques*, 2008.

Shen L., http://lionelshen.free.fr/Labo/Master/Memoire_M2_LS.pdf

Blogs & environnement

[Blogs]

Patrick Couton-Wyporek
www.pcw-etudes.fr

Résumé : L'exploration textométrique d'un corpus de blogs qui abordent sur le web la question environnementale permet d'extraire une série de notions clefs du domaine et de cerner leurs usages respectifs parmi les différentes sources qui participent au débat sur ce thème. On étudie ensuite les variations dans l'emploi de ces notions au cours du temps.

Abstract : The statistical analysis of a corpus of blogs on the topic of ecology yields a series of key notions of this semantic area and their different usages by the various sources which take part in the debate around this theme. A chronological study can then be undertaken to show the evolutions in the use of these words through time.

1. Contexte de la recherche

Dans la période précédant l'élection présidentielle française de 2007, la question *environnementale* a occupé une place remarquable sur tous les médias, imposant définitivement dans l'opinion, au cours de l'année 2005-2006, l'évidence d'une crise écologique majeure. Durant cette même période, le développement exponentiel des blogs comme outil de communication privilégié sur Internet a permis de constituer un espace d'échanges particulièrement riche sur le sujet environnemental. Dans ce contexte, il nous a semblé utile, d'étudier le déploiement de notions clefs telles que *développement durable*, *énergies renouvelables,* au sein des nombreux ensembles de textes produits dans la rencontre entre ces deux tendances fortes que sont l'appropriation massive des blogs comme moyen de communication, d'une part, et l'omniprésence des questions liées à l'écologie, d'autre part.

La question initiale que l'on se pose est celle de savoir comment sont appropriés et restitués les différents concepts environnementaux véhiculés par des univers de discours a priori différenciés (scientifique, citoyen, politique…etc). Compte tenu du grand nombre de blogs disponibles (cf. la constitution du corpus ci-après) et d'une masse textuelle significative pour chacun d'eux, nous nous sommes orientés vers l'approche textométrique.

Dans un premier temps, l'outil *Lexico3* nous a permis d'identifier une série d'expressions rattachées au sujet de l'environnement et d'obtenir une vue quantitative globale sur les formes-clefs en présence. Dans un second temps, nous avons prolongé ces observations par une observation qualitative de ces formes dans leur contexte discursif afin de cerner plus précisément le sens qu'elles véhiculent dans chacun des discours mis en présence.

2. Caractéristiques du corpus

Le corpus que nous avons constitué est composé de l'ensemble des billets de chaque blog, de sa création jusqu'au mois d'octobre 2006 (date du recueil). L'exploration a été réalisée sur la base d'un corpus de 23 blogs recouvrant 5 champs discursifs identifiés comme distincts.

Le tableau 1 donne la liste des blogs réunis pour constituer le corpus ***BlogsEnvironnement***. On peut voir au tableau 2 un extrait du corpus après balisage succinct pour permettre sa prise en charge par Lexico3. Le tableau 3 donne la liste des principales caractéristiques lexicométriques du corpus.

Tableau 1 :

Les 23 blogs réunis dans le corpus ***BlogsEnvironnement***

- 9 BLOGS « ECO-CITOYENS » :
  http://ecocitoyen.over-blog.com , http://blog.toutallantvert.com/ ,
  www.changement-climatique.fr (Conseil Economique et Social) ,
  http://www.eco-echos.com/dotclear/index.php ,
  http://droitdanslemur.blogspot.com/ (gaïa)
  http ://www.criseclimatique.fr (film Al Gore) ,
  http://durable-et-responsable.hautetfort.com/ ,
  http://utopie.viabloga.com/ , http://noolithic.typepad.com
- 3 BLOGS « ENERGIE RENOUVELABLES :
  http://eole.over-blog.net/ , http://www.leblogenergie.com/ ,
  http://terre.blogs.liberation.fr/
- 3 BLOGS « SYNDICAT ENERGIE » :
  http://cfdtieglot.canalblog.com/ ,
  http://www.acspe.com/ , http://www.unsa-energie-civaux.com/
- 3 BLOGS « ALTER » :
  http://sdn49.hautetfort.com/ , http://energie.com.over-blog.com/ (pcegdf),
  http://blpwebzine.blogs.com/champg
- 5 BLOGS « POLITIQUES » :
  http://www.desirsdavenir.org/index.php , http://dsk.typepad.com ,
  http://dominiquevoynet.net/blog/ , http://blog.villepin.free.fr/ ,
  http://sarkozyblog.free.fr/index.php

Tableau 2 :

Extrait du corpus ***Blogs-Environnement***

<blog=ecocitoyen>
<date=gaout05[45]>
§ lundi 15 août 2005
§ l ' aspartame ? miam !
§ je me rappelle du dégoût ressenti en lisant un article concernant l'aspartame . l'aspartame , vous connaissez ? mais si , bien sûr , une large majorité de produits dits « light » en contiennent .
§ cet article n'est pas récent , mais il vaut le coup que je le cite pour faire le point et savoir où l'on en est dans la commercialisation de l'aspartame aujourd ' hui .
§ « coca - cola light is (no) good !
§ des milliers de g . i's , pendant la guerre du golfe (et non de l'irak) , ont été victimes d'intoxications dues au coca - cola light . ils sont édulcorés à l'aspartame . pendant les hostilités , les palettes de canettes étaient entreposées au soleil , chaud dans ces régions .
§ a partir de 33°c , l'aspartame devient du méthanol (alcool à brûler) très toxique , qui ensuite se dégrade en formaldéhyde (formol) encore plus toxique .
§ et que se pass - t - il dans l'estomac à 37°c bien tassés ?
§ bizarre , bizarre , l'aspartame a été inventé par…monsanto , dans le cadre de la guerre chimique (acésulfamine de potassium) .
§ depuis 1996 , des scientifiques et des médecins dénoncent sa dangerosité (diabètes graves et cancers du cerveau , in journal of neurology and expérimental neurology) et réclament son interdiction . cependant , l'aspartame est toujours largement consommé dans 90 pays , et notamment par les femmes par souci (erroné) de mincir . »
§ cet article a été publié par le magazine votre santé n°45 en juin 2003 .
§ et depuis 2003 ?

[45] Afin de visualiser les résultats de Lexico3 par ordre chronologique, nous avant fait précéder l'intitulé du mois de référence par une lettre, en suivant l'ordre alphabétique.

Tableau 3 :

Principales caractéristiques lexicométriques du corpus

Nombre d'occurrences	647 121
Nombre de formes	40 397
Nombre d'hapax	20 026
Fréquence maximale	32 752

3. Etude de la partition par dates

La clef *date* permet de diviser le corpus en 23 parties qui correspondent chacune à un mois. L'Analyse Factorielle des Correspondances du tableau réalisé à partir de cette partition, figure 1, permet de distinguer des groupes relativement homogènes du point de vue de la chronologie.

Le calcul des spécificités appliqué à chacun de ces groupes permet d' identifier les thèmes dominants pour chacune des périodes :

- 1er semestre 2005 : une actualité sur les énergies, notamment pour l'automobile : *carburant, diesel, voitures, hydrogène*.
- 2er semestre 2005 : le cyclone *Katrina* (survenu fin août 2005) et la dérégulation du marché énergétique en Europe.
- 1er trimestre 2006 : La production électrique et la part des énergies renouvelables : *les éoliennes, l'hydraulique*.
- D'avril à octobre 2006 : le discours des politiques à propos de la fusion GDF-Suez et la sortie du film d'Al Gore (octobre 2006).

Tableau 4 :

Exemples de spécificités lexicales d'avril à octobre 2006-

Terme	Freq.Totale	Freq.Partie	Spécif.
suez	351	302	35
gdf	325	275	30
fusion	214	176	17

Terme	Freq.Totale	Freq.Partie	Spécif.
ump	113	105	19
socialiste	83	79	16
gore	52	52	14

4. Etude de la partition par blogs

L'AFC réalisée à partir de la partition par blog permet de constituer des groupes qui corroborent en partie la typologie constitutive du corpus. Ainsi, le groupe le plus dense dans la zone inférieure de la figure 2, rassemble majoritairement les blogs sélectionnés pour leur positionnement *éco-citoyen.* On retrouve dans la partie supérieure gauche un groupe qui réunit les blogs initialement identifiés comme *syndicalistes*. On note que le blog des militants communistes d'EDF-GDF de Rouen (pcegdf) se situe à proximité de ce dernier groupe.

L'analyse isole par ailleurs certains blogs dont le discours est particulier (gaïa, eole). On note que les blogs de personnalités politiques ne sont pas rapprochés entre eux, en revanche on observe une proximité entre le groupe *eco-citoyen* et le blog de Dominique Voynet[46].

[46] Notre sujet d'étude n'étant pas centré sur le positionnement lexical des blogs, nous n'approfondirons donc pas davantage ces premières observations. Néanmoins, on relève un axe exploratoire intéressant qui consisterait à

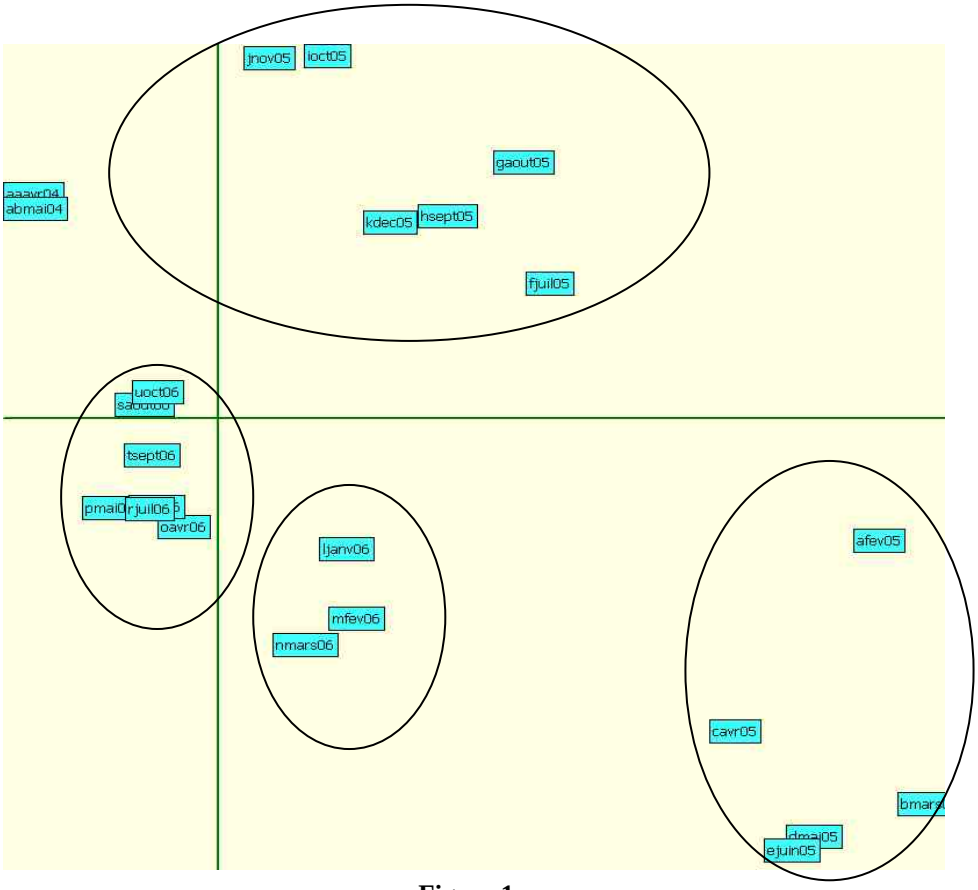


Figure 1 :
Typologie réalisée à partir de la partition chronologique en mois

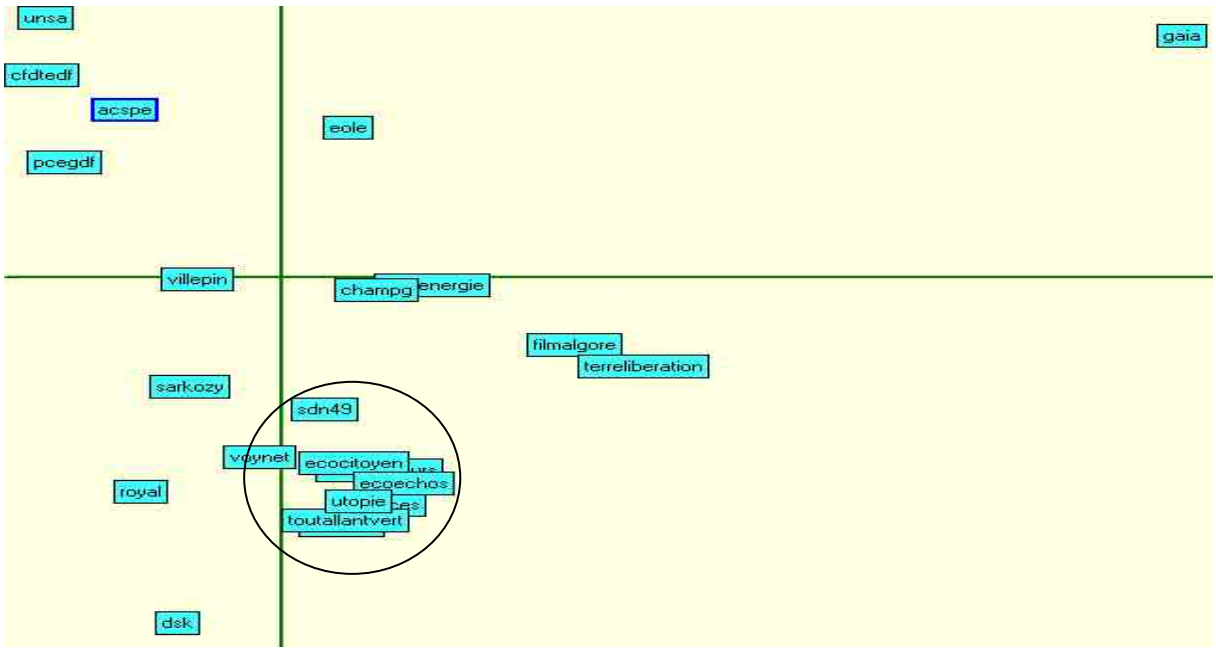


Figure 2 :
Typologie réalisée à partir de la partition par *blog*

---

cerner les récurrences de formes proches entre les blogs, par exemple entre celui de D. Voynet et les blogs *eco-citoyens*.

5. Les formes-clefs

Le tableau de fréquence du dictionnaire de formes et des segments répétés nous permet d'identifier d'emblée six termes, parmi les plus fréquents sur le thème de l'environnement[47] :

Tableau 5 :

Termes les plus fréquents, liés au thèmes de l'environnement

Termes	Fréquence
développement durable	297
énergies renouvelables	200
réchauffement climatique	141
changement climatique	91
décroissance	50
protection de l'environnement	50

6. Développement durable ou protection de l'environnement ?

L'expression ***développement durable***, avec un effectif de 297 occurrences est omniprésente et confirmée comme la notion phare en réponse au constat du réchauffement climatique. L'adjectif *durable* apparaît sémantiquement comme la forme pivot. C'est ce que révèle l'inventaire distributionnel illustré dans le tableau 3 avec quelques exemples de concordances :

Tableau 6

Extrait de la concordance autour de *durable*

en valeur de techniques d'**agriculture durable** par le don de semences traditionnelles ,
e de la part de chacun . la **consommation durable** est notamment associée à la production et
s privés dans les technologies d'**énergie durable** . § pour plus de détail , je vous invite
ource : mon ami olivier , de " **quotidien durable** " , à récemment proposé une note sur le
§ modes de consommation et de **production durable** § la publication du pnue " modes de
yer cette note § 28 novembre 2005 § **noël durable** et responsable (1) § une initiative qui
29 mai 2006 § narbonne ou l' **urbanisme durable** § dans des posts précédents je vous avais

Le terme *protection de l'environnement*, apparaît nettement moins fréquemment que *développement durable* dans le corpus de blogs (50 occurrences). Pourtant, une requête sur Google, donne plus de résultats pour *protection de l'environnement* que pour *développement durable* (3 660 000 contre 2 070 000). Bien qu'ayant été une des expressions « historique » de la cause écologique (les associations de protection de l'environnement), elle semble être tombée en désuétude.

Une exploration des spécificités par blog permet de constater que l'expression est spécifiquement employée par le blog *alternacteur* qui en fait une rubrique (d'où un nombre d'occurrence supérieur à l'usage réel de la locution). En effet, le retour au texte via le module « textploreur » (tableau 4) permet de constater que la locution fait l'objet d'une récurrence

[47] Cette liste est une sélection. D'autres formes d'intérêt pourraient faire l'objet d'explorations (biodiversité, consommation responsable…).

artificielle compte tenu de la dénomination de rubrique par opposition à une réelle récurrence d'usage dans le discours :

Tableau 7

Extrait de la concordance autour de *protection de l'environnement*

```
Partie : alternacteurs, Nombre de contextes : 40

développement durable , forums / débats , protection de l ' environnement | lien permanent |
développement durable , forums / débats , protection de l ' environnement , santé ,
développement durable , forums / débats , protection de l ' environnement | lien permanent |
dologies , blog , développement durable , protection de l ' environnement , santé | lien
dblié dans blog , développement durable , protection de l ' environnement , transports | lien
dblié dans blog , développement durable , protection de l ' environnement , santé | lien
d: 00 publié dans développement durable , protection de l ' environnement | lien permanent |
```

7. Réchauffement - changement - ou crise climatique ?

Le constat d'un *réchauffement climatique* apparaît largement partagé et côtoie l'expression *changement climatique* qui appartient au même paradigme désignationnel. Toutefois, l'exploration (figure 3) montre des spécificités fortes selon les blogs et permet de nuancer le sens de ces expressions :

Figure 3 :

Spécificités par blogs des termes *changement climatique* / *réchauffement climatique* -

Comme on le voit sur la figure 3, l'expression *réchauffement climatique* est privilégiée par le blog citoyen ***gaïa*** tandis que l'expression *changement climatique* est surreprésentée dans le blog du Conseil Economique et Social (CES), animé par des scientifiques ainsi que le blog ***ecoechos*** d'une ingénieur agronome. On peut faire l'hypothèse que l'idée de réchauffement

apparaît restrictive pour les scientifiques qui préfèrent parler de *changement*, ce qui laisse la place à d'autres analyses causales du changement climatique.

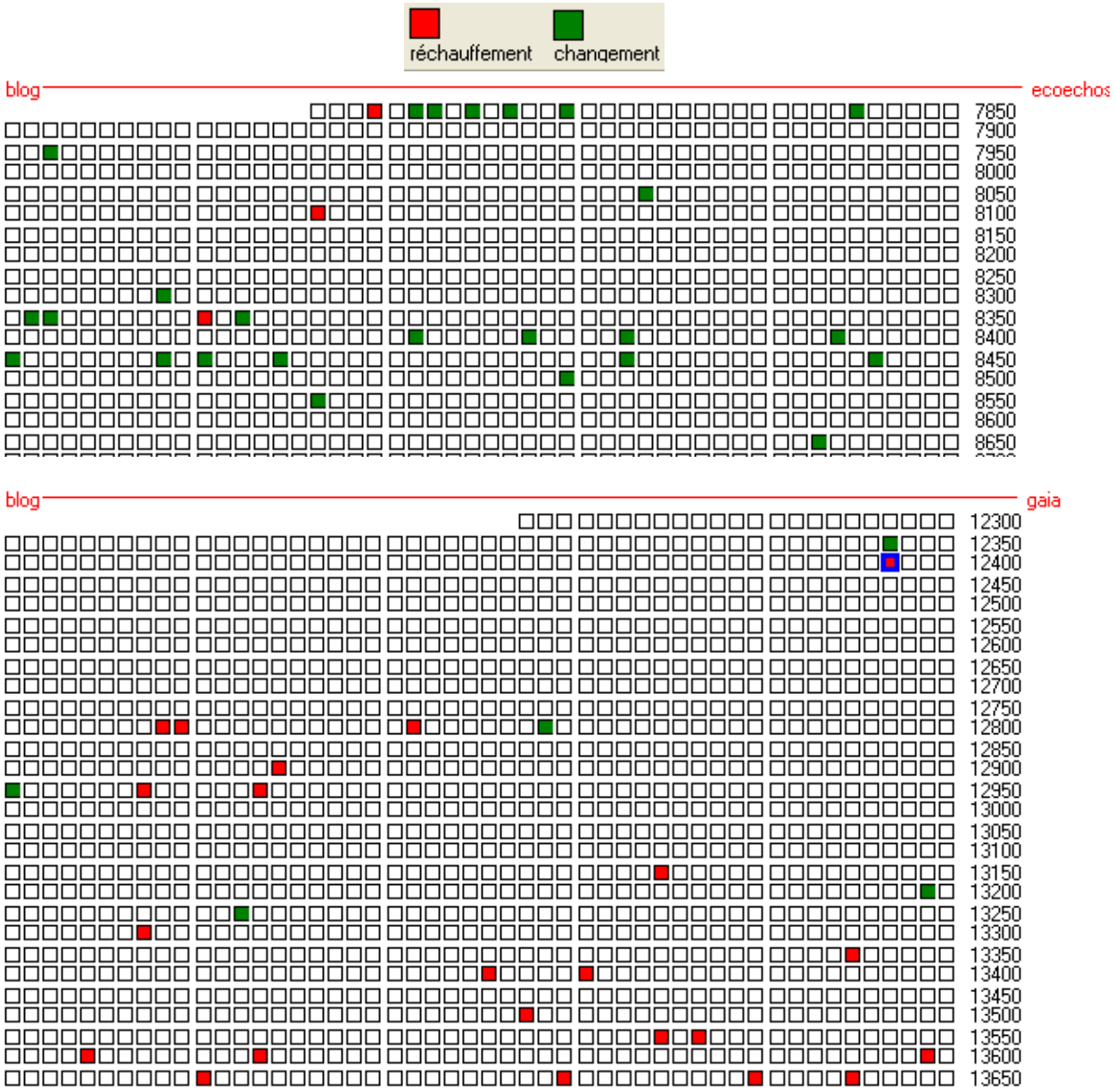


Figure 4

Répartition des formes *réchauffement* et *changement*

On constate que le parti pris éditorial énoncé à travers la bannière du blog contraint fortement la désignation notionnelle dans le discours. Ainsi, la bannière du CES installe d'emblée l'expression *changement climatique* tandis que le blog ***gaïa*** introduit son propos en parlant de *réchauffement climatique*.

Les cartes de section figure 4 permettent de visualiser les co-occurrences des deux formes. L'extrait de corpus pour chaque occurrence offre la possibilité d'identifier les constantes et les variations d'usage.

Cette approche permet de constater que le contexte sémantique de la forme *réchauffement climatique*, privilégiée par les blogs ***criseclimatique*** et ***gaïa*** s'inscrit en majeur dans un registre de sensibilisation :

> le. **réchauffement** climatique : le sommet du kilimandjaro presque sans neige (yahoo news) § (Gaïa)
>
> urgence : il faut remédier au **réchauffement** climatique ! (*criseclimatique* -filmAlgore)
>
> si l ' humanité avait besoin d ' un signal fort concernant le **réchauffement** climatique , je pense que cette saison de cyclones et de tempêtes tropicales aura été un signal quand même suffisamment dévastateur , en tous cas pour les populations directement concernées , à la nouvelle orléans et ailleurs . (Gaïa)

Dans les blogs *ecocechos* et *CES*, l'usage de l'expression *changement climatique* s'inscrirait davantage dans un registre réflexif (sur les causes, les conséquences et les enjeux du changement climatique):

> nous avons demandé à olivier godard , économiste , directeur de recherche au cnrs et professeur à l'ecole polytechnique ce que signifiait « coût du **changement** climatique » ? que prend - on en compte ? n'a t - on pas parfois considéré que les activités environnementales créaient à leur tour des productions et des richesses ? (CES)
>
> dans le passé , plusieurs civilisations très avancées se sont éteintes alors qu ' elles étaient à leur apogée . celle des maya et de l ' ile de pâques sont certainement les exemples les plus frappants . d ' autres , confrontées à des contraintes similaires ont survécu . jared diamond , grand scientifique américain* , s ' est attaché à les étudier , les comparer et à comprendre les causes de leur effondrement . il a identifié 5 facteurs : dommages environnementaux , **changement** climatique , voisins hostiles , dépendances entre partenaires commerciaux et capacités de la société à répondre à ces menaces avec ses valeurs propres . des causes qui résonnent on ne peut mieux à nos oreilles . § (ecoechos)
>
> comment le logement et l'habitat pourraient - il contribuer à lutter contre le **changement** climatique ? et comment accompagner cette évolution ? l'expérience personnelle d'une blogueuse , viviane rommelaere , montre que la société dans son ensemble - il ne s'agit pas seulement des pouvoirs publics - n'incite pas à aller dans ce sens. (CES)

On relève dans le corpus la variante ***réchauffement de la planète*** (19 occurrences) qui constitue une alternative à la désignation du réchauffement climatique (tableau 5), sans que cela trahisse selon nous une réelle nuance.

Tableau 8

Extrait de la concordance autour de *réchauffement de la planète*

climate threat (yahoo news) § 5 . changement climatique :	**réchauffement de la planète**	(che
co2 de l ' atmosphère , et donc une accélération brutale du	**réchauffement de la planète**	.
es 12 - 14° de ce mois d ' août , on va même finir douter du	**réchauffement de la planète**	. §
dant quatre mille ans . . . § pour en savoir plus : § 1 . le	**réchauffement de la planète**	fait
carotte de glace « dôme c epica » (eurekalert !) § 2 . le	**réchauffement de la planète**	(
by luc at 7 : 39 am 0 comments § jeudi , mars 24 , 2005 § le	**réchauffement de la planète**	§ la
. défi d ' une tonne § 3 . faisons vite ! (ademe) § 4 . le	**réchauffement de la planète**	(
diminuant les fameuses émissions de co2 , limitant ainsi le	**réchauffement de la planète**	. qui
mesures de température de surface océanique , bref , avec le	**réchauffement de la planète**	. et
ie et de vent sur lequel on pose nos bris de bottes . § " le	**réchauffement de la planète**	est
hui important de manifester nos préoccupations concernant le	**réchauffement de la planète**	et de
rgie nucléaire comme une des solutions pour lutter contre le	**réchauffement de la planète**	»
s qu'à ralentir l'effet de serre , c'est à dire à limiter le	**réchauffement de la planète**	. il
foundation § des études récentes viennent de montrer que le	**réchauffement de la planète**	
. peut - être échaudé par le documentaire catastrophe sur le	**réchauffement de la planète**	que
p chaud : tout ce que vous avez toujours voulu savoir sur le	**réchauffement de la planète**	et

On observe que l'expression ***crise climatique***, plus alarmiste, qui correspond à l'adresse du blog de lancement du film d'Al Gore « Une vérité qui dérange », n'est pas reprise par les

blogueurs (0 occurrence). Le discours déployé dans ce blog préfère d'ailleurs utiliser l'expression *réchauffement climatique*.

Figure 5

La page d'accueil du blog *criseclimatique.fr*

En revanche, le mot *crise* (tableau 6) est utilisé pour évoquer, la plupart du temps, la pénurie de ressources énergétiques fossiles et les tensions sur le marché du pétrole.

Tableau 9

Extrait de la concordance autour de *crise*

que énergétique du pays . § car avec la **crise** énergétique actuelle , ce n ' est plus une
demande de 50 % d ' ici là . § " dans la **crise** énergétique actuelle , on oublie souvent
rgétique pour faire face aux défis d'une **crise** énergétique annoncée , il est fort probable
ait dans les années 70 , en réponse à la **crise** énergétique de l'époque . oui , elles peuvent
la fnme - cgt . « dans un contexte de **crise** énergétique durable , il serait irresponsable
pieds dans le tapis en pleine période de **crise** énergétique . en moins d ' un an , les prix
. le président de la république , § la **crise** énergétique est profonde et durable . la
e la compétitivité . dans un contexte de **crise** énergétique grave , le gouvernement persiste
st paradoxal que face aux prémices d'une **crise** énergétique , l'europe confie encore ce
df est porteuse , qui plus est en pleine **crise** énergétique mondiale , je juge l'ouverture
orientation de notre pays à un moment de **crise** énergétique mondiale . § les objectifs affichés
e général de la cgt . « nous vivons une **crise** énergétique qui n'est pas près de se dénouer
une solution autochtone et pratique à la **crise** énergétique qui touche actuellement l '
es garnies de milliards , provoquant une **crise** énergétique sans précédent en californie
ropéenne révèle que , pour répondre à la **crise** énergétique , seuls 8 % des citoyens français
our lancer une telle opération en pleine **crise** énergétique . . . § § publié par
l ' urgente nécessité de résoudre cette **crise** environnementale . § ce passionnant
très tôt l ' ampleur et les enjeux de la **crise** environnementale ; de ce jeune sénateur
abilité face à une réalité , celle de la **crise** environnementale profonde que nous vivons

8. Energies renouvelables ou décroissance ?

Sur la période que nous avons considérée, on relève deux formes qui traduisent une certaine idée du consensus en termes de solutions environnementales et de réponse à la crise énergétique. L'expression ***énergies renouvelables***, particulièrement fréquente (200 occurrences), manifeste le concept le plus consensuel, en pleine expansion. Les concordances triées « avant » - tableau 7 -montrent que le vocabulaire associé s'inscrit dans le registre de l'essor.

Tableau 10

Extrait de la concordance autour de *énergies renouvelables*

économies d ' énergie et le passage aux	**énergies renouvelables**	: § " les événements mondiaux
ible l'après - pétrole et le passage aux	**énergies renouvelables**	? » § c ' est aussi une
ible l'après - pétrole et le passage aux	**énergies renouvelables**	? comment ne pas voir que
s parler de la nécessité de recourir aux	**énergies renouvelables**	: utopie ou véritable
s d'économie d'énergie et de recours aux	**énergies renouvelables**	faites dans l'étude des « 7
e , ensoleillement maximum , recours aux	**énergies renouvelables**	, déplacements réduits . .
efficacité énergétique et un recours aux	**énergies renouvelables**	. § a lire également leur
laires progrès en matière de soutien aux	**énergies renouvelables**	. gageons que la wallonie
ssociations et discuter de coopération ,	**énergies renouvelables**	, philosophie , etc . . .
ent de lancer un programme ambitieux d '	**énergies renouvelables**	. ils ont certes raison chacun
iques pour le choix d ' installation d '	**énergies renouvelables**	(voir ici) , et pourra vous
production d ' électricité à partir d '	**énergies renouvelables**	) . l ' éolien n ' intervient
n 2010 , 21% de l ' énergie à partir d '	**énergies renouvelables**	. des mesures incitatives
de son électricité produite à partir d '	**énergies renouvelables**	d ' ici 2010 (soit 46 twh
conomie d ' énergie et de production d '	**énergies renouvelables**	. § je fais ce que je dit
re 2005 § statistiques de production d '	**énergies renouvelables**	dans le monde § que
forte augmentation de la production d '	**énergies renouvelables**	et les activités d ' iberdrola
- ils . § certes , si la production d '	**énergies renouvelables**	a progressé de 30% , c'est
favoriser les mesures de productions d '	**énergies renouvelables**	de pair avec des économies
r découvrir six sites de productions d '	**énergies renouvelables**	situés en wallonie §
en comparaison à d ' autres sources d '	**énergies renouvelables**	, l ' avantage d ' être
compétitifs avec d ' autres sources d '	**énergies renouvelables**	. § le doe espère que la
ons sinistrées , et au développement des	**énergies renouvelables**	. et ce dans le cadre d '
iété énergétique et de développement des	**énergies renouvelables**	. § c ' est dans ce contexte
r le fonds européen de développement des	**énergies renouvelables**	(feder) , le conseil
ssociation générale de développement des	**énergies renouvelables**	, comme planète eolienne ,
e un beau potentiel de développement des	**énergies renouvelables**	! § bizarrement , les panneaux
ademe fait le bilan du développement des	**énergies renouvelables**	et des économies d ' énergie
ontarisme en faveur du développement des	**énergies renouvelables**	, il est certain que nous
élus vis - à - vis du développement des	**énergies renouvelables**	et de leur mis en place .
de france) , etc… § - développement des	**énergies renouvelables**	: en parallèle d'une hausse
re ouverte et avant le développement des	**énergies renouvelables**	. et pour cela il nous faut
cherait de financer le développement des	**énergies renouvelables**	. leur souhait : sortir à
ise de l'énergie ou le développement des	**énergies renouvelables**	et dont je déduis une évidence
rrogées plaide pour le développement des	**énergies renouvelables**	et des campagnes de maîtrise
ang des priorités , le développement des	**énergies renouvelables**	reste un poste important .
associations actives dans le domaine des	**énergies renouvelables**	. § l ' information en soit
mme énergétique destiné à développer les	**énergies renouvelables**	et à chasser le pétrole et
industries , propres , à développer les	**énergies renouvelables**	. un véhicule 100% propre
a volonté de la france de développer les	**énergies renouvelables**	améliorent la rentabilité
: maîtriser la demande , développer les	**énergies renouvelables**	, pérenniser la filière
§ il nous faudra ensuite développer les	**énergies renouvelables**	§ ces énergies sont encore
isièmement , il nous faut développer les	**énergies renouvelables**	. j'ai indiqué précédemment
res projets permettant de promouvoir les	**énergies renouvelables**	seront financés grâce à ce
européen , la volonté de promouvoir les	**énergies renouvelables**	ne paraît pas manquer
t une vitrine idéale pour promouvoir les	**énergies renouvelables**	. cette opération conduira

En revanche, le mot *décroissance*, qui, comme *énergies renouvelables,* est porteur d'une solution environnementale, apparaît beaucoup plus contesté. Un retour au texte montre que le mot *décroissance* est jugé négatif en soi et donc non porteur d'espoir :

Le mot croissance est dynamique. **Le mot décroissance est un frein**. Qui donc accepte d'être freiné ? Pour ma part, je préfère les mots sans connotation de privations à endurer. (Blog Noolithic)

« le choix du développement durable est un choix de croissance forte : le développement technologique indispensable est créateur d'emplois et fournira une base solide de la compétitivité internationale ». c'est un postulat, on le sait, qui est loin d'être partagé par tous les adeptes du développement durable, dont **certains vont jusqu'à parler d'une « décroissance » indispensable** pour que tous les pays puissent arriver à des niveaux de richesse à peu près comparables – et l'on sait à quel point les écarts sont aujourd'hui considérables. **Vous pensez bien que je ne suis pas de ceux-là** mais je m'étonne quand même qu'on puisse aujourd'hui en france traiter le problème de la croissance sans intégrer de manière plus nette l'ensemble de la réflexion aujourd'hui disponible – et urgente – sur la durabilité. (blog DSK)

La décroissance, ce mot que je n'aime pas
j'interviens régulièrement sur des billets d'autres blogs concernant la décroissance. je n'en parle presque jamais sur mon blog. parce qu' utiliser ce mot me dérange. on m'a incité à le faire : ça marche. alors, pourquoi je n'aime pas ce mot ?
1 - c'est **un mot négatif**, il **est peu enthousiasmant pour porter un projet**. or le projet que nous avons à mettre en place pour se sortir de la crise actuelle et passer "le syndrome du titanic" est lui, très enthousiasmant. j'y reviendrai plus tard.
2 - c'est **un mot flou** : "décroissance" : décroissance de quoi ? (…) (Blog eco-echos Isabelle Delannoy).

9. Conclusion

Les différentes démarches d'exploration textométrique permettent d'identifier des phénomènes quantitatifs de discours qu'une lecture cursive ou analytique ne permettrait pas d'identifier.

Le discours développé sur les blogs se prête particulièrement à une discussion sur les mots et les concepts. La réactivité induite par la mise en ligne instantanée des prises de position, la liberté de ton, et le décloisonnement des territoires (le discours scientifique côtoie le discours des citoyens), sont autant de facteurs qui favorisent la fluidité et la dynamique des idées et des discours.

Cette première exploration montre que les échanges sur la blogosphère contribuent au destin des concepts et des mots qui les portent : dans un cas, c'est l'expression *réchauffement climatique* qui est concurrencée par la variante *changement climatique*. Dans un autre cas de figure, les prises de position concordantes sur un concept comme celui de la *décroissance* concourent à sa disqualification.

10. Références

Mortureux, M-F, *Paradigmes désignationnels*, *Semen*, 08, Configurations discursives, 1993, [En ligne], URL : http://semen.revues.org/document4132.html mis en ligne le 6 juillet 2007.

Nee E., *Insécurité et élections présidentielles dans le journal Le Monde* http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/navigations/Presse3.html

11. Fonctionnalités Lexico3 utilisées dans cette exploration

N°	*Fonctionnalité*	Résultat
5	Principales caractéristiques lexicométriques (PCLC)	*Tableau 3*
5.3	AFC	*Figure 1, 2*
5.5	Concordance	*Tableau 6, 7, 8, 9, 10*
6	Ventilation dans les parties	*Figure 3*
7	Carte des sections	*Figure 4*

Interactions adulte/enfant [48]

[Interactions]

Luiggi Sansonetti
luiggi@luiggisansonetti.fr

Résumé : L'apprentissage de la langue maternelle chez l'enfant en situation dialogique avec un adulte montre à quel point l'enfant est réceptif et réactif à l'apprentissage dans le cadre de dialogues. Comment l'adulte réagit-il dans cette même situation ? L'exploration textométrique du corpus permet de repérer et de confronter les reprises et les reformulations chez les deux locuteurs. Elle permet d'observer la reprise par l'enfant des productions de l'adulte, et d'étudier la manière dont l'adulte corrige les créations enfantines.

1. L'étude des interactions adulte/enfant

La *linguistique de l'acquisition* s'intéresse, en premier lieu, à la mise en place et à l'évolution du fonctionnement cognitivo-langagier chez l'enfant. A partir de corpus d'interactions verbales entre un adulte et un enfant, recueillies en situation de parole spontanée, il est possible d'observer les changements survenus dans sa capacité d'expression au cours du temps. Les avancées du courant interactionniste (Ochs et Schieffelin, 1995) et des travaux sur le français parlé (Blanche-Benveniste, 1997) nous ont servi de point de repère pour analyser ces interactions particulières dans lesquelles l'un des sujets, l'enfant, se trouve en phase d'acquisition des moyens d'expression.

Dans un corpus longitudinal constitué de plusieurs dialogues entre un adulte et un enfant, on observe des phénomènes de reprises et de reformulations[49] de la part des deux locuteurs. Lorsque l'enfant reprend de manière inappropriée une construction employée par l'adulte et que l'adulte reformule cette construction de manière adéquate, l'adulte se trouve impliqué dans une situation de collaboration dans le processus d'énonciation entrepris par l'enfant. Il répond aux tâtonnements de l'enfant et lui fournit le moyen d'expression recherché. C'est ce que Wyatt appelle *feed-back correctif* (Wyatt, 1969), désigné aujourd'hui par *interaction ajustée* ou *adaptée*.

2. Les corpus Julien et Mathilde

Pour cette étude nous considèrerons deux corpus, chacun réalisé à partir de transcriptions de dialogues entre un adulte et un enfant[50]. Le premier corpus : ***Julien*** est constitué de trois dialogues entre un adulte et un même enfant. Le second corpus : ***Mathilde*** est également constitué de trois dialogues entre le même adulte et une petite fille.

---

[48] L'auteur remercie Emmanuelle Canut, Martine Vertalier et André Salem pour leurs lectures attentives et leurs remarques précieuses dans l'élaboration de ce travail.
[49] Nous appelons *reprise* lorsque le mot est répété à l'identique et *reformulation* lorsqu'un autre mot est proposé à la place ou s'il suit une modification morphosyntaxique.
[50] Les corpus utilisés pour cette étude ont été recueillis par Tissier (2001).

Tableau 1
Tableau de synthèse des corpus **Julien** *et* **Mathilde**

Nom	Claire	Julien	Claire	Mathilde
Age moyen	20-25 ans	5 ans 10 mois → 6 ans 4 mois	20-25 ans	4 ans 9 mois → 4 ans 11 mois
Nb énoncés[51]	145	135	141	137
Nb entretiens	3		3	

Dans les deux cas, c'est un livre illustré de Tomi Ungerer, *Crictor*[52], qui a servi de support à l'entretien. Après avoir lu le livre à l'enfant, l'avoir relu lorsque l'enfant en exprimait le désir, l'adulte a demandé à l'enfant de lui raconter à son tour l'histoire qu'il venait d'entendre, intervenant en permanence pour l'aider dans son récit. Le corpus des interactions enregistré sur support audio au cours de ces dialogues a été ensuite transcrit sous forme textuelle sur un support informatique pour tenter d'analyser les caractéristiques de ces interactions dans la co-construction d'une narration à partir d'un livre illustré[53].

Le corpus ***Julien,*** constitué des interactions entre l'adulte et le petit garçon compte 467 formes pour 2 986 occurrences. Le corpus ***Mathilde*** qui rassemble les interactions entre le même adulte et la petite fille comprend 444 formes pour 3 619 occurrences. Les corpus sont partitionnés en « dialogue ». Les énoncés sont triés par locuteur (d'abord l'adulte puis l'enfant et ainsi de suite) puis 50 par 50. Cette présentation des *tours de parole* (*énoncés* désormais) permet d'avoir sur les lignes impaires les énoncés de l'adulte et sur les lignes paires ceux de l'enfant. Dans cette représentation, deux interventions consécutives dans le temps sont situées l'une en dessous de l'autre, et les interventions de chaque locuteur sont sur une même ligne.

Les principales caractéristiques lexicométriques des corpus ainsi constitués nous conduisent à remarquer que les mots les plus fréquents ne sont pas les traditionnels mots outils comme : *de, le, la, les…* que l'on trouve à cette place dans les corpus écrits e français A leur place, nous trouvons deux pronoms (*tu*, *il*), un auxiliaire (*est*), une conjonction (*et*) et un déterminant (*un*).

[51] Selon les anciennes conventions, seuls les énoncés ne contenant que « mm » ne sont pas numérotés. C'est pourquoi, dans nos corpus informatisés, il peut y avoir un décalage dans la numérotation des énoncés.
[52] Ungerer T., *Crictor*, 1958, L'école des Loisirs pour l'édition française, 1980, Collection Lutin Poche, réédition 2000.
[53] Le corpus a été transcrit selon les conventions établies par Laurence Lentin et ses collaborateurs (Lentin, 1984-1988).

Tableau 2

Principales caractéristiques quantitatives des corpus **Julien** *et* **Mathilde**

	Formes	Occurrences	Fréquence Maximale	Mot le plus fréquent
Corpus ***Julien***	**467**	**2986**	**140**	**il**
Dialogue ***ju1***	302	1271	40	tu
Adulte	202	662	40	tu
Enfant	204	609	20	un
Dialogue ***ju2***	243	896	63	il
Adulte	122	315	19	est
Enfant	190	581	48	il
Dialogue ***ju3***	227	819	38	il
Adulte	139	413	22	qu
Enfant	169	406	23	il
Corpus ***Mathilde***	**444**	**3619**	**115**	**et**
Dialogue ***ma1***	159	572	23	et
Adulte	112	283	13	le
Enfant	119	289	13	et
Dialogue ***ma2***	301	1438	52	le
Adulte	215	682	25	Crictor
Enfant	234	756	29	le
Dialogue ***ma3***	329	1609	54	et
Adulte	239	866	31	et
Enfant	239	743	24	un

Guide de lecture du tableau 2 :

Dans ce tableau :

- *les nombres alignés à gauche et en gras correspondent à l'ensemble des dialogues réunis*
- *les nombres centrés correspondent au dialogue seul mais réunissant les locuteurs*
- *les nombres alignés à droite correspondent au locuteur seul.*

Dans la dernière colonne, le mot en gras et aligné à gauche correspond à la forme la plus fréquente dans le corpus longitudinal. Le mot centré correspond au mot le plus fréquent dans l'entretien, et le mot aligné à droite correspond au mot le plus fréquent pour chaque locuteur.

Tableau 3

Transcription du premier dialogue de **Julien** *(extrait)*

A1	Alors c'est quoi l'histoire de Crictor ?
J1	mm un jour le facteur arrive et donne un mm quand madame Bodot ouv(r)e le pa/quet mm elle va / elle va au zoo
A2	Et pourquoi elle va au zoo ?
J2	pour voir si c'est pas un / un serpent dang(e)reux (il avale sa salive)
A3	Pourquoi le serpent dang(e)reux, il est dans / il était dans l(e) paquet,, que l(e) facteur a apporté ?
J3	oui
A4	D'accord.
J4	et mm un c'était un boa constructeur alors elle l'appela Cric(tor) elle lui donna un [s] / [s] le biberon [s] elle lui apporta des palmiers
A5	Et pourquoi elle lui apporta des palmiers ?
J5	pour sa propre nature
A6	Ah, pour qu'i(l) euh se rappelle.
J6	sa nature
A7	D'accord.
J7	mm elle lui faisa un petit gilet,, elle décida de / de l'em/mener en classe il apprena,, l'alphabet
A8	Et euh, en fait elle s'occu, madame Bodot / elle s'occupe du serpent comme un / un petit enfant, en fait ?
J	mm
A9	Et c'est qui qui lui a envoyé le / le serpent ?
J8	c'est son / c'est son mari
A10	C'est son mari ?
J9	il apprena à compter
A11	Donc il a / il a appris à / à compter à l'école aussi ?

Guide de lecture du tableau 3 :

Convention de transcription[54] : (signes utilisés dans l'extrait)

- Transcription orthographique pour garder la lisibilité des énoncés et pour étudier la syntaxe
- Ponctuation dans les énoncés de l'adulte
- Pas de ponctuation dans les énoncés de l'enfant, sauf « ? » et « ! »
- Crictor : titre du livre illustré pour l'entretien
- / : marque d'hésitation ou d'interruption dans le déroulement de l'énoncé
- ouv(r)e : mise en parenthèse de syllabes non prononcées pour la lisibilité des énoncés
- [s] : transcription d'un son qu'on ne pourrait orthographier
- ,, : notation des silences avec espacements plus ou moins long selon leur durée
- mm : les énoncés contenant uniquement « mm » ne sont pas numéroté

[54] Anciennes conventions de transcription. Dans les actuelles conventions, les énoncés contenant « mm » sont numérotées, les silences sont notés *p, *pp, *ppp selon leur durée. Les conventions suivent les réflexions des chercheurs et les récentes avancées dans le traitement informatique des corpus.

Tableau 4
Corpus de **Julien** *balisé (extrait)*
énoncés de l'adulte dans la 1ère période

```
<corpus=julien>
<dial=ju1>
<part_loc-corp=adult-ju1>
<part_corp-loc=ju1-adult>
§<adult=1-001> alors c ' est quoi l ' histoire de crictor ?
§<adult=1-002> et pourquoi elle va au zoo ?
§<adult=1-003> pourquoi le serpent dangereux , il est dans / il était dans le paquet *p que le facteur a apporté ?
§<adult=1-004> d ' accord .
§<adult=1-005> et pourquoi elle lui apporta des palmiers ?
§<adult=1-006> ah , pour qu ' il euh se rappelle .
§<adult=1-007> d ' accord .
§<adult=1-008> et euh , en fait elle s ' occu , madame bodot / elle s ' occupe du serpent comme un / un petit enfant , en fait ?
§<adult=1-009> et c ' est qui qui lui a envoyé le / le serpent ?
§<adult=1-010> c ' est son mari ?
§<adult=1-011> donc il a / il a appris à / à compter à l ' école aussi ?
§<adult=1-012> il jouait à quoi avec les garçons ?
§<adult=1-013> il montrait euh à qui *p comment on faisait les noeuds ?
§<adult=1-014> à des garçons , ok .
§<adult=1-015> et euh le qui l ' a bâillonnée euh madame euh bodot ?
§<adult=1-016> le cambrioleur , d ' accord . et qu ' est - ce qui s ' est passé par la suite ?
§<adult=1-017> il attacha qui le serpent ?
§<adult=1-018> le bandit , d ' accord .
§<adult=1-019> donc il a bien un jardin qui portait son nom ?
§<adult=1-020> d ' accord . donc là c ' était l ' histoire du petit crictor
```

Guide de lecture du tableau 4 :

Dans cet extrait du corpus ***Julien***, les balises permettent de délimiter les séquences de texte produites par chaque locuteur :

- *la clé <dial> « dialogue » qui distingue les trois dialogues de Julien ;*
- *la clé <part_loc-corp=adult-ju1> « partie_locuteur-corpus » qui distingue les corpus par locuteurs pour la visibilité de certains graphques ;*
- *la clé <part_corp-loc=ju1-adult> « partie_corpus-locuteur » qui distingue les locuteurs par corpus pour la visibilité de certains graphques ;*
- *le caractère § qui matérialise les énoncés ;*
- *la clé <adult=1-001> .distingue les locuteurs (adulte : adult, enfant : child), le numéro du dialogue (1-, 2-, 3-) et les énoncés (001).*

3. pourquoi - parce que

Le centre d'intérêt des recherches menées par Lentin et ses collaborateurs concerne le développement de la syntaxe comme facteur de structuration et d'évolution du langage de l'enfant, et l'influence des interactions langagières entre un adulte et un enfant sur ce développement. Lorsque l'enfant s'approprie le langage, il s'approprie, entre autres choses, l'organisation des éléments. La syntaxe contribue à l'organisation sémantique du discours, puisque les mots prennent sens dans leur contexte énonciatif et syntaxique. L'observation porte donc sur les cheminements individuels de mise en fonctionnement du langage.

Une liste d'*Introducteurs de complexité* (IC) a été établie à partir des occurrences de formes et de constructions syntaxiques relevées dans des corpus d'enfants de 3 à 7 ans. A partir de cette liste, Lentin a recherché pour chaque corpus le ou les énoncés de l'enfant qui présentaient le maximum d'introducteurs de complexité syntaxique selon les critères adoptés (la complexité maxima) : « ce paramètre permet de comparer les apprenants entre eux et surtout chaque apprenant à lui-même, dans une observation diachronique » (Lentin, 1998, 31). C'est à partir de ces observables que l'on peut mettre en relief une appropriation par l'enfant d'une partie du système langagier de l'adulte pour l'élaboration de son propre système.

Dans la mesure où il s'agit d'analyses syntaxiques sur les transcriptions, les informations concernant la prosodie ou la phonologie ne sont pas prises en compte. Pour déceler les constructions complexes énoncées par l'enfant nous recherchons, dans les énoncés de l'adulte et de l'enfant, des mots (*quand*, *puisque*), des groupes de mots (*il faut que*, *parce que*, *pour que*), des constructions syntaxiques (*verbe + verbe infinitif*) ainsi que des tentatives de constructions syntaxiques chez l'enfant[55].

Nous recherchons, par exemple, l'emploi de la locution *parce que* (*parce qu, parce que*) et sa répartition dans le corpus longitudinal de Julien.

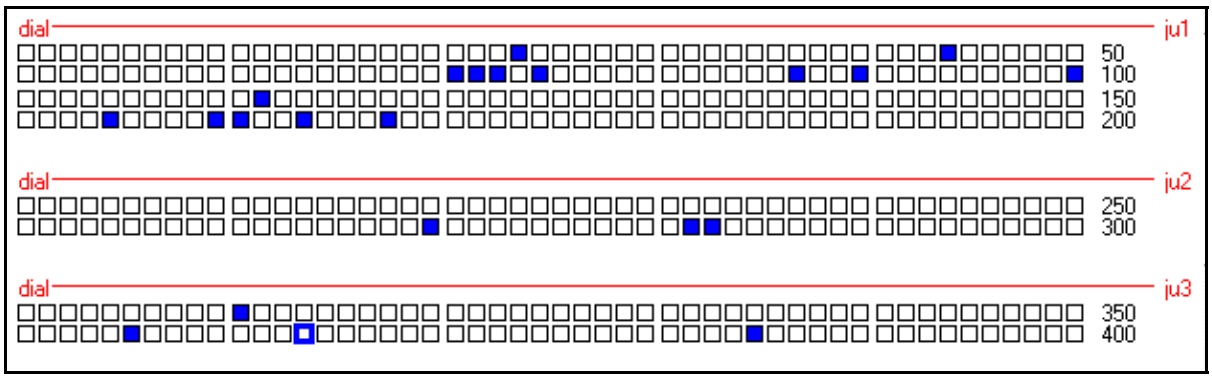


Figure 5
Localisation des parce que *dans le corpus* **Julien**

[55] Pour chaque dialogue, une grille d'analyse est remplie. Une synthèse classe les éléments et les constructions trouvés, les références des énoncés contenant ces éléments, ainsi que leurs fréquences pour chaque dialogue. L'évolution des fréquences de chaque catégorie syntaxique dans le corpus longitudinal est ensuite représentée par des courbes d'évolution.

Guide de lecture de la figure 5 :

Le corpus ***Julien*** est réparti en trois dialogues. Les lignes impaires concernent l'adulte, les lignes paires l'enfant. Chaque énoncé est représenté par un carré. Le coloriage du carré indique la présence de la forme recherchée dans le corpus.

Le premier *parce que* est produit par l'enfant dans l'énoncé 20 du premier dialogue (*J20*). Cette tentative de l'enfant n'est pas reprise par l'adulte, mais par l'enfant lui-même, à l'énoncé *J21* dans une construction syntaxique qui est, cette fois, complète :

```
A21   Et pourquoi t(u) aimes bien ce livre ?
J20   ,, parce que
A22   T(u) aimes bien les serpents ?
J21   ,, euh non mais c'est c(e) que je / c'est pa(r)ce que j'avais envie
```

Quand nous recherchons la motivation de ce *parce que* (noté en bleu) chez l'enfant, nous nous apercevons qu'il vient toujours en réponse à un *pourquoi* (noté en rouge) de l'adulte. Sur la figure 6, nous avons noté simultanément les *parce que* en rouge et les *pourquoi* en bleu pour vérifier que l'énonciation des *parce que* est induite par une question *pourquoi* :

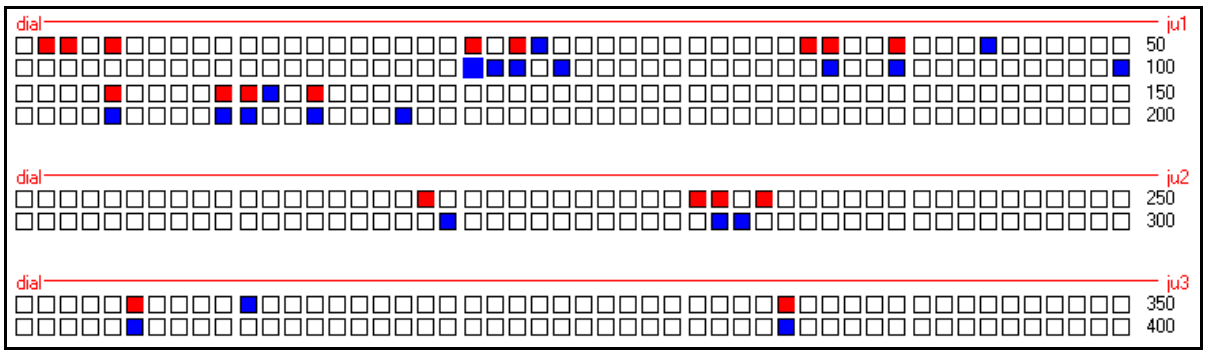
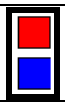
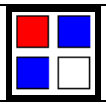


Figure 6
Localisation des pourquoi / parce que *dans le corpus* **Julien**

Guide de lecture de la figure 6 :

Les combinaisons de couleurs, sur la carte des énoncés, permettent d'identifier plusieurs situations distinctes.		
		
Question ***pourquoi*** de l'adulte et réponse immédiate de l'enfant avec ***parce que***	Question ***pourquoi*** de l'adulte, réponse immédiate de l'enfant avec *parce que* et reprise de l'adulte du ***parce que***	Présence simultanée dans le même énoncé des deux unités recherchées

La localisation des *parce que*, recherchés simultanément avec les *pourquoi*, permet de mettre en relief certains phénomènes propres à l'oral. En effet, nous cherchons les énonciations de *parce que* pour vérifier s'il s'agit d'une production en construction complète ou non. En règle générale, lorsque nous répondons à une question de type *pourquoi*, il est rare que nous reprenions la principale. L'enfant répond de façon quasi systématiquement en construction incomplète à la suite d'une question de la part de l'adulte.

Lorsque nous procédons à la même recherche dans le corpus longitudinal de ***Mathilde***, nous remarquons qu'il n'y a aucune question de type *pourquoi*.

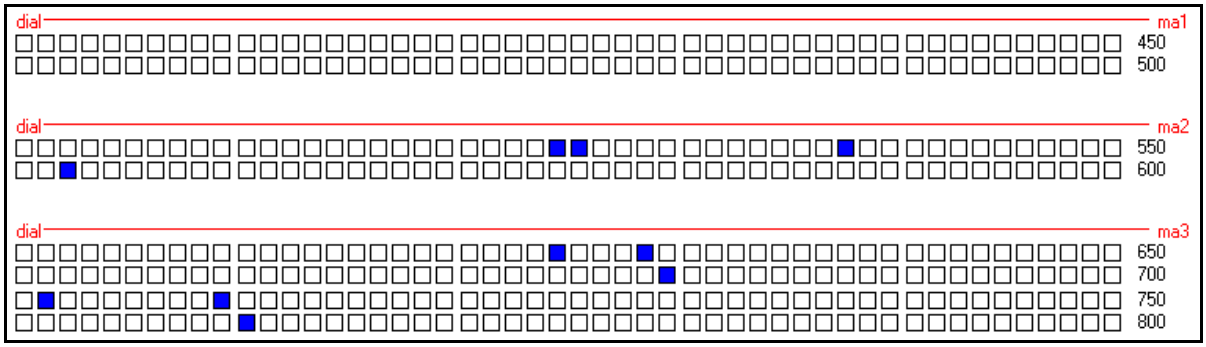


Figure 7
Localisation des pourquoi – parce que *dans le corpus* **Mathilde**

Le premier *parce que* de l'enfant, énoncé dans le deuxième dialogue, est une tentative abandonnée qui n'est pas reprise par l'adulte :

```
M3    elle va au zoo parce que le serpent trictor alors elle l'appela
      Crictor alors
```

Dans ce troisième corpus, l'adulte énonce plusieurs constructions avec *parce que* sans que l'enfant ne les reprenne immédiatement. Il n'y a que trois productions de *parce que* par Mathilde dont les deux dernières, dans le troisième dialogue, sont des reprises immédiates des énoncés de l'adulte :

```
A29   Parce que le mot néant commence par la lettre N.
M30   d'accord parce que / parce que euh dans / dans quoi ?
[…]
A60   […] on lui a fait une statue et il y a un parc qui porte son nom
      parce que il a / il a / il a arrêté le cambrioleur.
M60   d'accord parce que il a arrêté le cambrioleur (ind.)
```

Lorsque nous faisons la comparaison des deux études ***Julien*** et ***Mathilde***, nous voyons que, face aux 22 occurrences de *parce que* localisées dans le corpus de ***Julien*** (4 par l'adulte et 18 par l'enfant), 12 occurrences seulement sont présentes dans le corpus de ***Mathilde*** (8 par l'adulte et 4 par l'enfant). L'énonciation de *parce que* dans ***Julien*** s'explique par la forte présence des questions de type *pourquoi* de la part de l'adulte, mais aussi par la fréquence des questions en général. En effet, lorsque nous faisons la carte des questions dans ***Julien***, en recherchant le point d'interrogation (noté en rouge), nous remarquons que l'enfant est soumis à un questionnement serré de l'adulte tout au long de l'entretien :

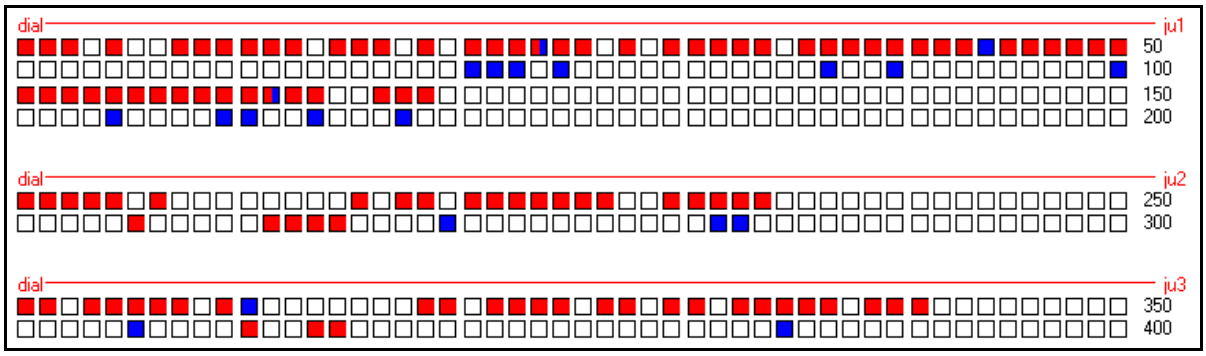


Figure 8
Localisation des ? / parce que *dans le corpus* **Julien**

A l'inverse, dans le corpus de ***Mathilde***, nous observons une dispersion des marques qui correspondent à des questions, beaucoup moins nombreuses, de la part de l'adulte, et des énonciations spontanées par l'enfant de *parce que* (notées en bleu) qui doivent être interprétées comme une tentative de mise en évidence de la cohérence du récit entendu :

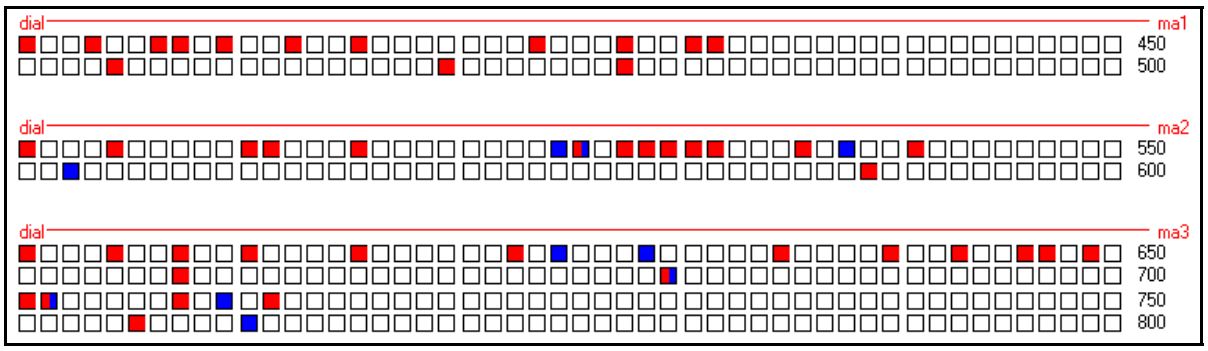


Figure 9
Localisation des ? / parce que *dans le corpus* **Mathilde**

Avec l'enfant Mathilde, l'adulte a posé moins de questions et a renoncé à l'utilisation de *pourquoi*, pour éviter de recevoir des *parce que* en construction syntaxique incomplète, sans l'énonciation de la principale. L'adulte a analysé ses dialogues avec Julien et a procédé par la

suite à un enregistrement d'un nouveau corpus longitudinal avec Mathilde, pour observer le rôle de l'adulte dans ne situation de co-construction de la narration autour du même livre illustré.

4. Acquisition de structures syntaxiques

Nous recherchons dans les énoncés de l'adulte et de l'enfant des formes lexicales (*quand, puisque*), des groupes de formes (*il faut que, parce que, pour que*), des constructions syntaxiques (*verbe + verbe infinitif*). Ces introducteurs de complexité (IC) constituent, selon Lentin, 1984, les marques les plus significatives de la progression de la complexité syntaxique en liaison avec l'articulation du raisonnement dans le langage en voie d'acquisition.

Nous observons aussi les tentatives de constructions syntaxiques chez l'enfant, car elles traduisent la mise en place de structures syntaxiques. D'autre part, nous vérifions si l'adulte fait écho à ces tentatives de l'enfant en le reprenant, en lui proposant d'autres structures.

Pour atteindre ces objectifs, nous avons étiqueté les corpus[56] ***Julien*** et ***Mathilde***. Nous utilisons désormais les corpus lemmatisés et catégorisés ***Julien-LC*** et ***Mathilde-LC***.

La lemmatisation d'un vocabulaire associe à chaque mot graphique sa forme canonique (voir tableau 10). Elle permet de rassembler les flexions d'un même verbe, la forme singulier ou pluriel d'un même nom, les formes fléchies d'un même adjectif, et de distinguer des formes graphiques correspondant aux homographes (voir tableau 11).

Tableau 10
Exemple de lemmatisation de flexions verbales

Forme graphique	Forme étiquetée	Forme lemmatisée
aimais	aimais_VIND3S	aimer_VINDI3S
aime	aime_VINDP1S	aimer_VINDP1S
aimes	aimes_VINDP2S	aimer_VINDP2S
aiment	aiment_VINDP3P	aimer_VINDP3P
aime	aime_VINDP3S	aimer_VINDP3S
aimé	aimé_VPARPMS	aimer_VPARPMS
aime	aime_VSUBP2S	aimer_VSUBP2S

[56] L'étiquetage a été réalisé avec Cordial (http://www.synapse-fr.com) puisqu'il apparaît être le plus efficace dans la reconnaissance des catégories pour le français parlé (Valli & Véronis 1999, Véronis 2000).

Tableau 11
Exemple de lemmatisation d'homographies

Forme graphique	Forme étiquetée	Forme lemmatisée
l	l_DETDMS	le_DETDMS
l	l_DETDFS	le_DETDFS
l	l_PPER3S	le_PPER3S
la	la_DETDFS	le_DETDFS
le	le_DETDMS	le_DETDMS
le	le_PPER3S	le_PPER3S
les	le_DETDPIG	le_DETDPIG
les	le_PPER3P	le_PPER3P

Tableau 12
Exemple d'étiquetage d'énoncés

Enoncé d'origine :
§<adult=1-001>alors c'est quoi l'histoire de crictor ?

Enoncé catégorisé :
§<adult=1-001>alors_ADV c'_PDS est_VINDP3S quoi_PRI l'_DETDFS histoire_NCFS de_PREP Crictor_NPI

Enoncé catégorisé et lemmatisé :
§<adult=1-001>alors_ADV ce_PDS être_VINDP3S quoi_PRI le_DETDFS histoire_NCFS de_PREP Crictor_NPI

La procédure de lemmatisation/catégorisation nous permet alors de rechercher des énoncés sur la base d'un patron syntaxique défini. Dans le corpus ***Julien-LC***, nous trouvons, à partir du patron syntaxique : *préposition suivie d'un verbe infinitif* (noté *Prep+VInf*), 21 occurrences de séquences relevant de cette construction, réparties ainsi dans le corpus :

Tableau 13
Liste des Prep+VInf *dans* **Julien-LC**

à_PREP compter_VINF
à_PREP faire_VINF
à_PREP lire_VINF
à_PREP manger_VINF
à_PREP sauter_VINF
de_PREP le_PPER3S prendre_VINF
de_PREP le_PPER3S emmener_VINF
pour_PREP voir_VINF
pour_PREP lui_PPER3S rappeler_VINF

Tableau 14
Répartition des constructions Prep+VInf *dans* **Julien-LC**

	ju-lc-1	ju-lc-2	ju-lc-3	Total Locuteur
Adulte	2	2	4	***8***
Enfant	4	4	5	***13***
Total Dialogue	***6***	***6***	***9***	**21**

A partir de ce patron, nous établissons des concordances mieux cerner la nature des prépositions et des verbes utilisés.

La préposition la plus utilisée est *à* (14 occurrences), très souvent après le verbe *apprendre*. Lorsque nous regardons la cartographie des énoncés, nous remarquons que c'est toujours l'enfant qui énonce le premier une construction de type *Prep+Vinf*. Les exemples qui suivent ont été localisés selon la procédure décrite ci-dessus. Ils sont présentés sous leur forme originale pour garder la lisibilité des énoncés.

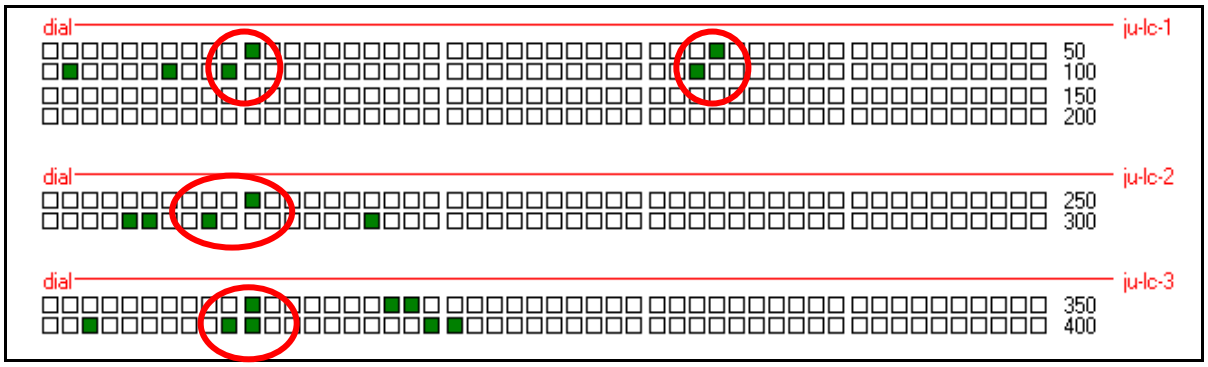


Figure 15
Localisation des Prep+VInf *dans le corpus* **Julien-LC**

Exemple 1 :
J9 il apprena **à compter**
A11 Donc il a / il a appris **à** / **à compter** à l'école aussi ?

Exemple 2 :
J32 ben j(e) lui aurais donné **à manger** j(e) l'aurais amené / j(e) lui aurais amené un lit une pe/ j(e) lui aurais mis une p(e)tite cabane pour qu'i(l) dorme dedans,, et puis euh / et puis avec sa cabane il pourrait manger
A34 Et tu lui au(rais) / tu lui aurais donné quoi **à manger** ?

Exemple 3 :
J9 elle elle / elle veut elle / veut l'emmener dans sa classe alors euh il apprend **à compter** S comme
A10 Serpent.
J10 S comme serpent E comme éléphant mm mm
A11 Donc le serpent, il apprend **à lire** et **à compter**.

Exemple 4 :
J8 il va / il va / et là il va dans la neige et il / et madame Bodot décida **de** le **prendre** dans sa classe
A11 Euh madame Bodot décide **d(e)** le **prendre** dans sa classe parce qu'elle est institutrice.
J9 mm S comme [s] il apprena **à faire** euh l'alphabet à / à sa place S comme serpent E comme éléphant c'est / c'est quoi ?

Dans l'exemple 1, l'enfant produit une construction *Prep+VInf* en tentant de construire le passé simple du verbe *apprendre*. L'adulte reformule immédiatement cette tentative en proposant un passé composé dans le même contexte lexical en reprenant la même construction.

Dans l'exemple 2, l'adulte reprend l'énonciation de l'enfant pour lui demander de préciser son raisonnement.

Dans l'exemple 3, l'adulte reprend l'énonciation de l'enfant et la complète par un autre groupe prépositionnel.

Dans l'exemple 4, l'adulte reprend l'énonciation de l'enfant et la complète par un *parce que*. L'enfant continue avec une tentative de construction au passé simple du verbe *apprendre*, que l'adulte ne reprend pas du tout dans la suite du dialogue.

En détaillant la nature des prépositions dans le corpus ***Julien-LC***, on remarque que les occurrences de la construction *de+VInf* sont toutes rassemblées dans le troisième dialogue. Il s'agit d'une énonciation spontanée par l'enfant de cette construction, reprise immédiatement par l'adulte dans le même contexte lexical. Dans ce corpus, l'enfant est le seul à produire des constructions de type *pour+VInf*.

Le nombre d'occurrences de la structure *Prep+VInf* est deux fois plus élevé dans le corpus ***Mathilde-LC*** que dans le corpus ***Julien-LC*** (21 occurrences dans ***Julien-LC***, 39 dans ***Mathilde-LC***). Là encore, la répartition des prépositions augmente au fur et à mesure des entretiens, mais l'écart entre les deux locuteurs est moindre (19 occurrences pour l'adulte et 20 pour l'enfant dans ***Mathilde-LC***, alors qu'il y avait respectivement 8 et 13 occurrences dans ***Julien-LC***).

Tableau 16
Liste des Prep+VInf *dans* **Mathilde-LC**

```
à_PREP compter_VINF
à_PREP enlever_VINF
à_PREP faire_VINF
à_PREP sauter_VINF
de_PREP prendre_VINF
de_PREP le_PPER3S emmener_VINF
pour_PREP apprendre_VINF
pour_PREP être_VINF
pour_PREP permettre_VINF
pour_PREP voir_VINF
pour_PREP lui_PPER3S rappeler_VINF
```

Tableau 17
Répartition des constructions Prep+VInf *dans* **Mathilde-LC**

	ma-lc-1	ma-lc-2	ma-lc-3	Total Locuteur
Adulte	2	7	10	***19***
Enfant	1	9	10	***20***
Total Dialogue	***3***	***16***	***20***	**39**

La cartographie (figure 18) permet de vérifier que les énoncés concernés par cette construction sont dans la plus part des cas regroupés en paires.

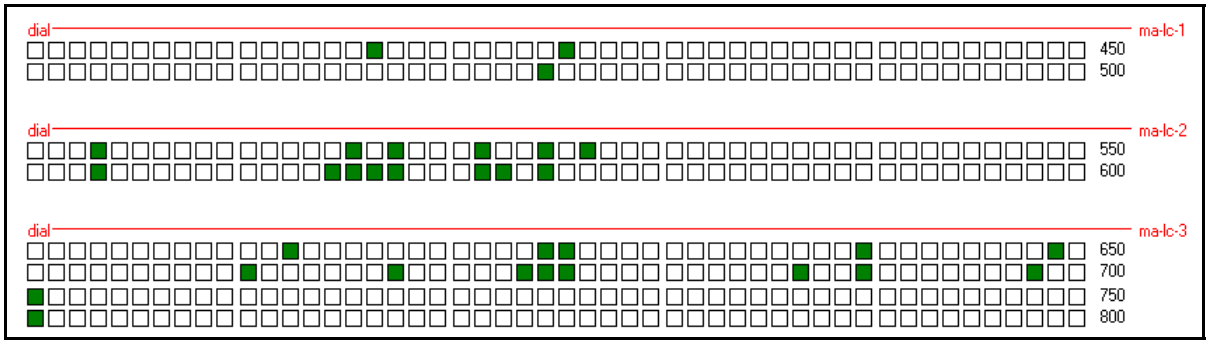


Figure 18
Localisation des Prep+VInf *dans le corpus* **Mathilde-LC**

Il n'y a aucune occurrence des constructions *de+VInf* ni *pour+VInf* dans le 1er corpus. Quant à la préposition *à+VInf*, il s'agit d'une énonciation spontanée de l'enfant, reprise immédiatement par l'adulte dans le même contexte lexical.

Dans le corpus ***Julien-LC***, la préposition *à* représente le tiers de la catégorie *Prep*, et apparaît principalement dans le contexte lexical *apprendre à compter*. Pour la préposition *de*, le contexte précédent est exclusivement *décider* suivi de *prendre* ou *emmener*. Enfin, pour la préposition *pour*, le contexte est *aller (au zoo) pour voir (si)*.

Les mêmes décomptes sur le corpus ***Mathilde-LC***, montrent que l'emploi des prépositions *à* et *pour* est équilibré (respectivement 17 et 16 occurrences). Ici encore, c'est la construction *apprendre à compter* qui est la plus utilisée. Le seul verbe qui sert à introduire la préposition *de* est le verbe *décider*. Et les deux seuls verbes infinitifs énoncés dans la même construction sont *prendre* ou *emmener*. Avec la préposition *pour*, c'est la construction *prendre (la forme) pour apprendre* qui est la plus utilisée.

Tableau 19
Répartition des prépositions dans les corpus **Julien-LC** *et* **Mathilde-LC**

Parties	Prépositions à	de	pour	Total
Julien	***14***	***3***	***4***	***21***
ju1-adult	2	0	0	2
ju1-child	2	1	1	4
ju2-adult	2	0	0	2
ju2-child	2	0	2	4
ju3-adult	3	1	0	4
ju3-child	3	1	1	5
Mathilde	***17***	***6***	***16***	***39***
ma1-adult	2	0	0	2
ma1-child	1	0	0	1
ma2-adult	3	1	3	7
ma2-child	3	2	4	9
ma3-adult	4	1	5	10
ma3-child	4	2	4	10

Dans le corpus de Julien, la construction *à+VInf* représente les deux tiers des constructions *Prep+VInf*. Dans chaque entretien, l'enfant utilise autant ces constructions que l'adulte. L'enfant est le seul à énoncer des constructions de type *pour+VInf*. Dans chaque entretien, l'enfant énonce toujours plus de constructions *Prep+VInf* que l'adulte (soit au total 13 occurrences pour l'enfant et 8 pour l'adulte).

Dans le corpus de Mathilde, les constructions *Prep+VInf* avec *à* et *pour* représentent les deux tiers des occurrences du total. Les constructions *de+VInf* et *pour+VInf* n'apparaissent qu'à partir du deuxième entretien. L'utilisation de ces constructions est équilibrée entre les deux locuteurs (19 occurrences pour l'adulte et 20 pour l'enfant).

L'adulte a plus que doublé son utilisation de ce patron *Prep+VInf* avec Mathilde (8 occurrences dans le corpus de Julien et 19 dans le corpus de Mathilde). D'autre part, il utilise seulement avec Mathilde la préposition *pour+VInf*.

5. Le rôle de l'adulte

La notion d'*interaction adaptée* de la part de l'adulte, avancée par L. Lentin et J. Bruner, s'appuie sur l'idée que les offres langagières, les reprises et les reformulations de l'adulte se produisent au moment où l'enfant cherche à verbaliser son expérience propre. Nous allons maintenant observer plus en détails certains de ces phénomènes de *feed-back correctif* à travers les créations enfantines[57].

[57] Cordial n'étiquette pas les créations enfantines, nous avons ajouté cette catégorie après relecture et repérage des éléments.

Pour améliorer la comparaison entre les différents locuteurs, et pour mieux observer le rôle de l'adulte et la réaction de l'enfant, nous avons réuni ***Julien-LC*** et ***Mathilde-LC*** en un seul et même corpus ***JuMa-LC***. La localisation des *créations enfantines* constitue une entrée particulièrement précieuse pour l'étude de l'activité que nous avons appelée *feed-back correctif*. Nous appelons *créations enfantines* les tentatives non canoniques de formation de flexions verbales comme le passé simple par exemple. L'observation porte également sur la réaction de l'adulte face à ces productions non standards de l'enfant. Si l'adulte reformule les tentatives de l'enfant en les reformulant de manière canonique, il s'agit de *feed-back correctif*. Une fois ces *feed-back correctifs* localisés, on tente d'observer la prise en compte par l'enfant, dans ces productions ultérieures, des corrections qui lui ont été proposées par l'adulte.

Lorsque nous avons vérifié et corrigé l'étiquetage et la lemmatisation du corpus par Cordial, nous avons apposé l'étiquette *CREA* pour toutes les tentatives inabouties de construction du passé simple par l'enfant. Nous avons traité de la même manière les variations sur les noms communs ou noms propres. En effet, pour ces dernières, nous avons voulu vérifier s'il s'agissait juste d'une prononciation fautive, ou un réel tâtonnement sur le mot.

Les formes étiquetées *CREA* relèvent en fait de deux grandes catégories. La première concerne des échecs qui peuvent être mis sur le compte d'une mauvaise mémorisation d'entités lexicales déjà rencontrées, comme *conscrictor*, *institutrice*, *servent* (pour *serpent*), *trictor* (pour *Crictor*), *véant* (pour *néant*) et *contistitua* pour *constata*. La seconde concerne les échecs dus à une maîtrise défectueuse des mécanismes de flexions et de conjugaisons. C'est cette dernière qui nous intéresse tout particulièrement.

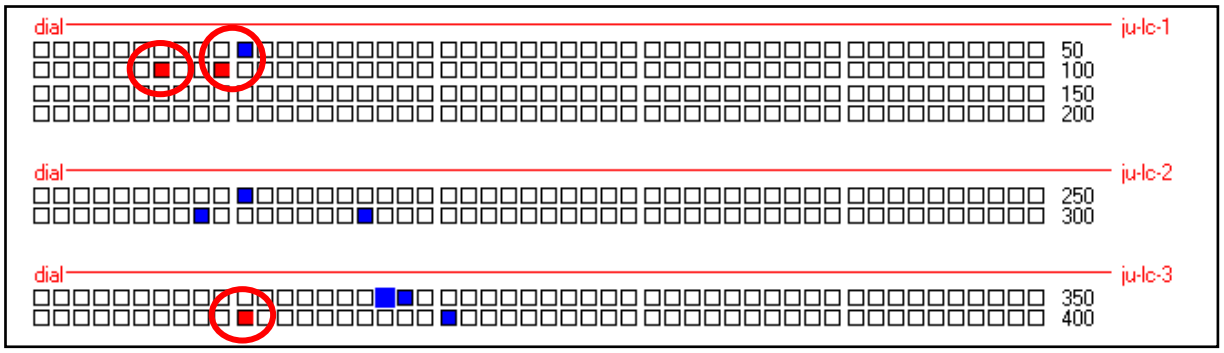


Figure 20
Localisation des apprena *en rouge et des formes canoniques du verbe* apprendre *en bleu dans le corpus* **JuMa-LC**

Dans le corpus ***JuMa-LC***, la tentative de construction de passé simple *apprena* (marquée en rouge) est comparée aux autres emplois du verbe *apprendre* (marqués en bleu). La reformulation de cette tentative par l'adulte n'aboutit jamais au passé simple *apprit*. L'enfant Julien produit une première fois une forme non canonique qui n'est ni reprise, ni reformulée par l'adulte (exemple 5). A la deuxième tentative de l'enfant, l'adulte reprend le verbe au passé composé (exemple 6). L'enfant produira par la suite un imparfait qui, lui, sera immédiatement repris par l'adulte dans le même contexte (exemple 7). Dans le dernier dialogue, l'énonciation de *apprena* n'est ni reprise ni reformulée par l'adulte. Vers la fin du dialogue, l'enfant utilise l'imparfait pour le verbe *apprendre*. L'enfant Mathilde ne produit pas de construction non canonique du verbe *apprendre*.

Exemple 5 :

J7 mm elle lui faisa un petit gilet,, elle décida de / de l'em/mener en classe il **apprena**,, l'alphabet

A8 Et euh, en fait elle s'occu, madame Bodot / elle s'occupe du serpent comme un / un petit enfant, en fait ?

Exemple 6 :

J9 il **apprena** à compter

A11 Donc il a / il **a appris** à / à compter à l'école aussi ?

Exemple 7 :

J11 à / à ça,, toboggan il était très serviable et il montrait comment on **faisait** les nœuds

A13 Il montrait euh à qui,, comment on **faisait** les nœuds ?

Exemple 8 :

J9 mm S comme [s] il **apprena** à faire euh l'alphabet à / à sa place S comme serpent E comme éléphant c'est / c'est quoi ?

A12 N.

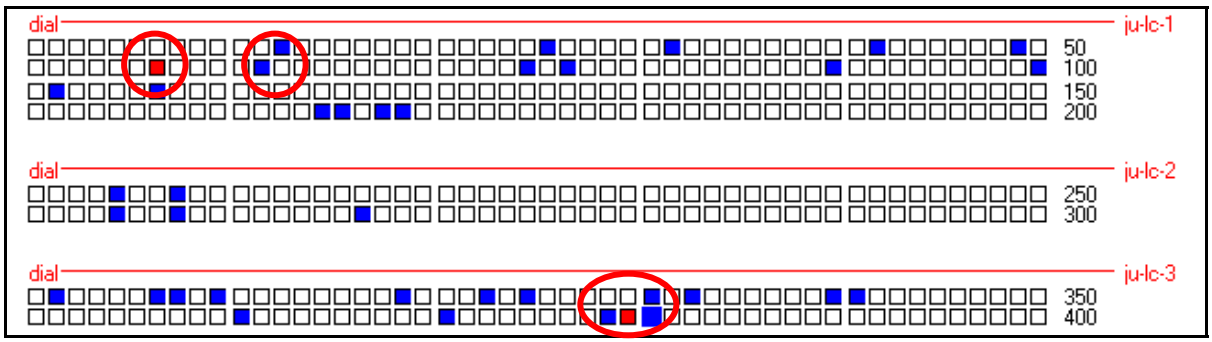


Figure 21
Localisation des faisa *en rouge et des formes canoniques du verbe* faire *en bleu dans le corpus* **JuMa-LC**

Dans l'exemple 9, l'enfant tente une construction au passé simple du verbe *faire*, mais l'adulte ne reprend ni ne reformule cette tentative. Quelques énoncés plus loin, l'enfant réutilise le verbe *faire* à l'imparfait, et l'adulte reprend immédiatement cette production (exemple 10). Dans l'exemple 11, l'enfant énonce un passé composé après quelques hésitations. Après une interrogation de l'adulte, il reprend son énoncé en faisant une tentative de construction au passé simple. L'adulte reformule l'énonciation de l'enfant en utilisant un passé composé, que l'enfant reprend immédiatement. L'adulte n'a donc pas proposé le passé simple canonique mais a repris le passé composé énoncé par l'enfant. L'enfant Mathilde ne produit pas de construction non canonique du verbe *faire*.

Exemple 9 :

J7 mm elle lui **faisa** un petit gilet,, elle décida de / de l'em/mener en classe il apprena,, l'alphabet

A8 Et euh, en fait elle s'occu, madame Bodot / elle s'occupe du serpent comme un / un petit enfant, en fait ?

Exemple 10 :

J11 à / à ça,, toboggan il était très serviable et il montrait comment on **faisait** les nœuds

A13 Il montrait euh à qui,, comment on **faisait** les nœuds ?

Exemple 11 :

J27 mm il a eut une médaille et il a / il a été / on l'a / on l'**a fait** en statue

A29 Le serpent ?

J28 mm on lui **faisa** un jardin

A30 Le serpent **a fait** un jardin ?

J29 nan on lui **a fait** un jardin pour le serpent

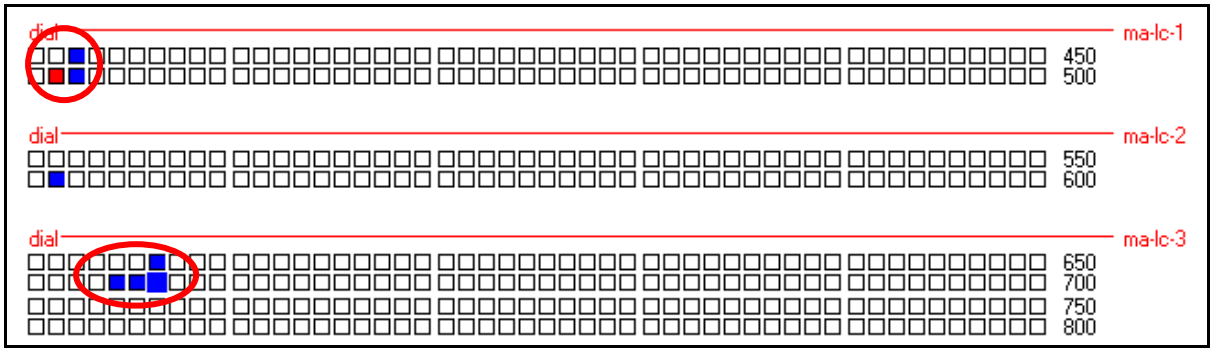


Figure 22
Localisation du ouvra *en rouge et des formes canoniques du verbe* ouvrir *en bleu dans le corpus* **JuMa-LC**

La seule tentative de passé simple du verbe *ouvrir*, énoncée par Mathilde au début du premier dialogue, est immédiatement reprise par l'adulte au même temps (exemple 12). L'enfant valide alors cette construction de passé simple en reprenant immédiatement l'énoncé de l'adulte. Dans le dernier dialogue, Mathilde produit un énoncé spontané avec la forme canonique (exemple 13). Toutes les autres énonciations de ce verbe sont au gérondif.

Exemple 12 :

M2 Bodot et un monsieur lui donna des (ind.) il lui acheta un cadeau elle l'**ouvra** et elle avait (ind.) euh euh

A3 Madame Bodot avait peur quand elle **ouvrit** la boîte.

M3 quand / madame Bodot avait peur quand elle **ouvrit** la boîte

Exemple 13 :

M5 après madame Bodot poussa un cri en l'**ouvrant** c'était un serpent Cric/

A6 Madame / madame Bodot pousse un cri.

M6 en l'**ouvrant**

A7 Car / en **ouvrant** la boîte car dans la boîte il y avait un serpent.

M7 madame Bodot **ouvrit** la boîte car dans la boîte il y avait un serpent aujourd'hui je l'appelle Serpounet

6. Conclusion

La linguistique de l'acquisition du langage s'intéresse à la mise en place du système cognitivo-langagier chez l'enfant en situation d'interactions verbales avec un adulte. L'observation porte, entre autres choses, sur l'organisation syntaxique des éléments de la phrase, et les phénomènes de feed-back correctif. La localisation d'une forme ou d'une structure syntaxique dans un corpus, constitué de plusieurs dialogues, permet de repérer de façon précise les cas de reprise et de reformulation. Il est alors aisé d'analyser, en retournant au texte, les interactions entre les locuteurs.

La cartographie des énoncés permet de localiser la forme ou la structure syntaxique recherchée, dans les énoncés de l'enfant et de l'adulte. Avec le cas des constructions de type *Prep+VInf*, on constate que dans le corpus ***Julien-LC***, l'adulte ne propose pas ce patron syntaxique à l'enfant, mais il reprend ce que l'enfant énonce (Figure 15). En revanche, avec le corpus ***Mathilde-LC***, on s'aperçoit que c'est l'adulte qui propose le plus souvent ce patron syntaxique à l'enfant.

Lorsque nous recherchons les créations enfantines dans le corpus ***JuMa-LC*** pour localiser les feed-back correctif de l'adulte, nous remarquons qu'il y a peu de correction apportée par l'adulte. Quand il reprend la tentative de construction du passé simple de Julien, il utilise un passé composé (Figure 20). En revanche, avec Mathilde, l'adulte reformule un passé simple canonique, que l'enfant réutilisera immédiatement et plus tard dans ses énoncés (Figure 22).

A l'aide de ces localisations précises, nous constatons que l'adulte a changé sa manière d'interagir avec l'enfant. Avec Julien, nous avons remarqué le questionnement incessant (Figure 8), alors qu'avec Mathilde, ce même adulte pose moins de question (Figure 9). D'autre par, l'adulte fait plus attention aux créations enfantines de Mathilde. Il propose également plus de patron syntaxique de type *Prep+VInf* à Mathilde qu'à Julien (Figures 23 et 24 ci-dessous). L'adulte se serait *adapté* à l'enfant au fur et à mesure de ces interactions.

Figure 23
Graphique de répartition des constructions Prep+VInf *entre l'adulte et Julien dans* **JuMaLC**

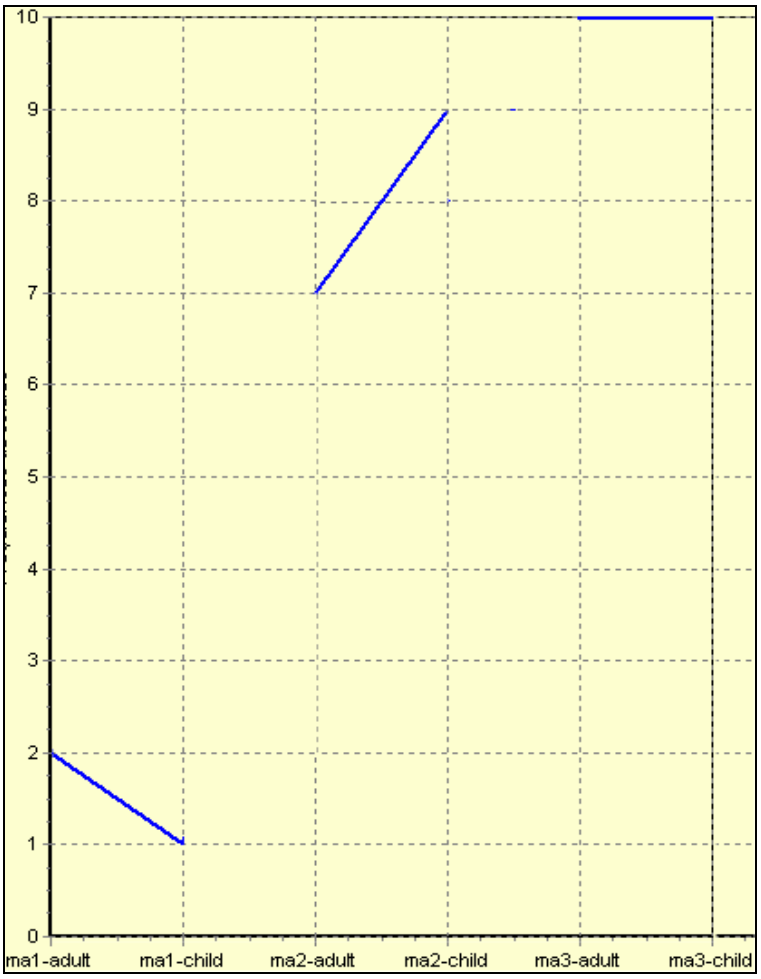


Figure 24
Graphique de répartition des constructions Prep+VInf *entre l'adulte et Mathilde dans* **JuMaLC**

Les fréquences dans les figures 23 et 24 sont absolues, c'est-à-dire que le nombre correspond au nombre de fois que la forme apparaît dans la partie. Les parties sélectionnées pour ces graphiques représentent l'ensemble des énoncés d'un locuteur dans un dialogue. Cette représentation permet de comparer, dans chaque dialogue, la fréquence d'utilisation de la forme recherchée par locuteur. En effet, nous constatons que l'enfant Julien a énoncé plus de patron syntaxique de type *Prep+VInf* que l'adulte, mais que les fréquences d'occurrences restent faibles. En revanche, avec l'enfant Mathilde, il y a beaucoup plus d'occurrences de ce patron de la part des deux locuteurs, et qu'ils sont assez proches quant à la fréquence d'utilisation.

La figure 25, représentant l'analyse factorielle des correspondances entre les locuteurs dans les différents dialogues du corpus ***JuMa-LC***, nous montre l'effort produit par l'adulte pour se rapprocher de l'enfant Julien au fil des dialogues, ainsi qu'un rapprochement davantage marqué de Mathilde. Cette interprétation est confirmée par le fait que l'adulte a d'abord travaillé avec l'enfant Julien, puis a analysé ses dialogues. C'est ensuite que ce même adulte a fait attention de procéder autrement avec un autre enfant en utilisant les mêmes supports pour guider les interactions (Tissier, 2001).

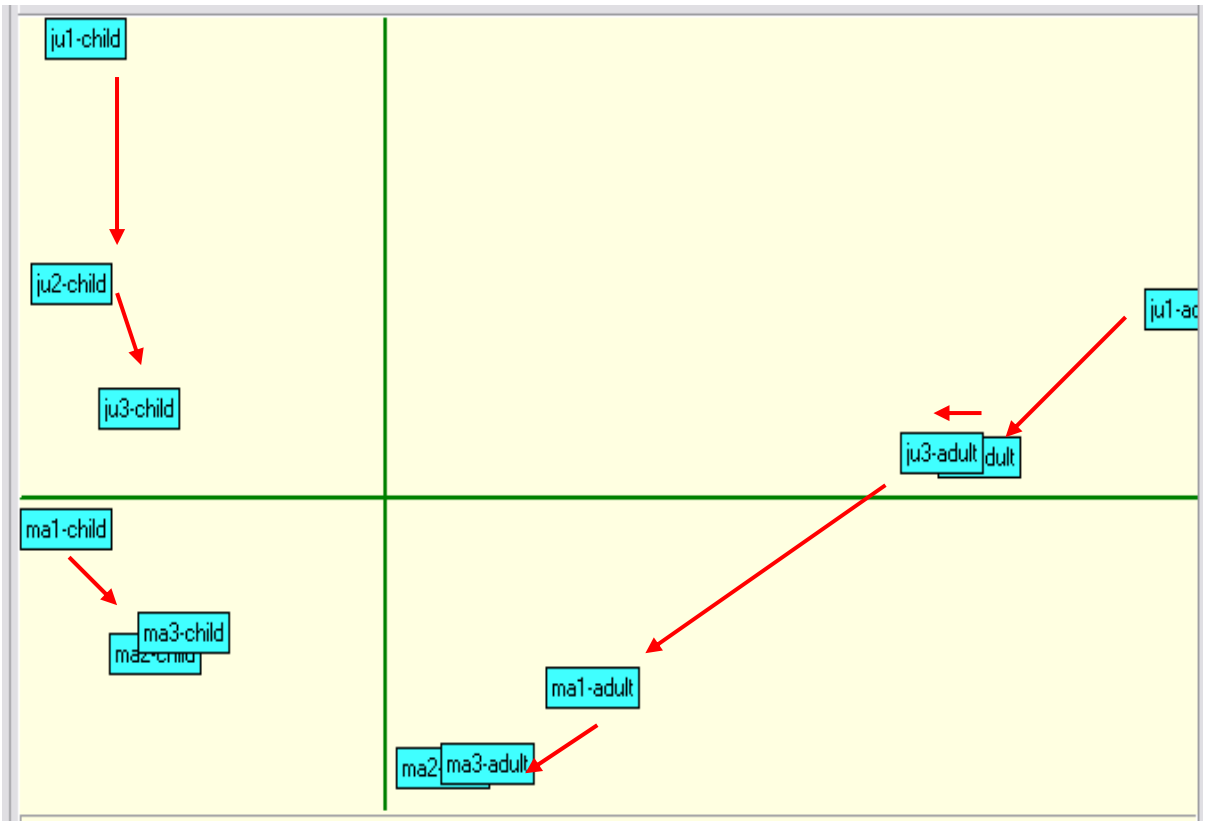


Figure 25
AFC des parties locuteur par dialogue dans le corpus **JuMaLC**

7. Indications bibliographiques

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*, Paris, Ophrys, Collection l'essentiel français.

Lentin L. et al. (1984-1988). *Recherches sur l'Acquisition du Langage*, tome 1&2. Presses de la Sorbonne Nouvelle.

Lentin, L. (1998). *Apprendre à penser, parler, lire, écrire*. Paris, ESF.

Ochs, E. & Schieffelin, B. (1995). « Language socialization and its consequences for language development » in Fletcher P and MacWhinney B. éd., *The Handbook of Child Language*, Blackwell Publishers.

Tissier C. (2001). *Rôle de l'adulte dans l'interaction langagière adulte-enfant (entre 4 ans 9 mois et 6 ans 4 mois) en situation de narrations dans deux corpus longitudinaux*. Paris, Mémoire de Maîtrise, ILPGA, Paris 3 (non publié).

Valli A. & Véronis J., 1999, « Etiquetage grammatical des corpus de parole : problèmes et perspectives », in *Revue Française de Linguistique Appliquée*, Volume IV n°2, décembre 1999, p.113-133.

Véronis J., 2000, « Annotation automatique de corpus : panorama et état de la technique », in Pierrel J.-M. éd., 2000, *Ingénierie des langues*, Hermes Sciences Publications, p.151-171.

Wyatt, G. (1969). *La relation mère-enfant et l'acquisition du langage*, Mardaga, Bruxelles.

8. Fonctionnalités Lexico3 utilisées dans cette navigation

N°	*Fonctionnalité*	Résultat
6	**Partition (clé dial,** pour dialogue**, clé part,** pour locuteur**)**	
7	**Carte des sections (énoncés, 1 locuteur par ligne)**	***Figures 5, 6, 7, 8, 9, 15, 18, 20, 21, 22***
8	**Groupe de Formes (Préposition suivie de Verbe Infinitif)**	***Tableaux 13, 16***
5	**PCLC (corpus, dialogue, locuteur)**	***Tableau 2***
6	**Graphique de Ventilation**	***Figures 23, 24***
5.3	**AFC (locuteur par dialogue)**	***Figure 25***

Interactions homme-machine

Ajustements à l'interlocuteur dans l'échange

Marguerite Leenhardt
EA2290 Syled-Cla2T, Le Sémiopôle
mleenhardt@le-semiopole.fr

Résumé : On utilise la textométrie pour comparer les productions d'intervenants en situation de demande de renseignement dans un contexte industriel. Des traitements adaptés à ces comparaisons mettent en évidence différentes conduites interactionnelles dans les échanges entre humains, d'une part, entre humains confrontés à une machine, d'autre part. Après une présentation du corpus et des codages indispensables à sa prise en charge (§2), une série d'analyses quantitatives permettent de dégager des caractéristiques propres aux différents types d'intervenants (§3), puis de proposer, sur cette base, une typologie conversationnelle des interactions (§4). Ces analyses nous amènent (§5) à discuter la question de l'ajustement conversationnel chez l'humain en situation de demande d'information.

Mots-clés : conversation humain-machine, analyse conversationnelle, textométrie, conduite interactionnelle

Abstract : Textometry is used to compare a verbal inquiry by a human speaker in reference to an industrial context or subject matter. Specific data processing is used to compare and correlate behavioural interaction between a human to human exchange and human to machine conversations. First a presentation is given of the corpus and formatted processing codifications (§2), next a series of quantitative analyses are used to extract speaker specific characteristics and main features (§3). As a result, a conversational topology is proposed for the interaction processes (§4), and an analysis is put forth to reveal questions pertaining to variations in human behaviour in situations of information inquiry (§5).

Key-words : man-machine conversation, conversational analysis, textometry, interactional behaviour

Pour faire face aux demandes de renseignement, toujours plus nombreuses, formulées par les usagers, les grandes entreprises qui interviennent dans le domaine des *services* mettent en place des traitements informatisés de prise en charge téléphonique. Après une période d'essai en contexte industriel, la phase actuelle est consacrée à l'évaluation des systèmes d'automates vocaux, une partie des appels étant désormais traitée par ces *systèmes intelligents*[58], les opérateurs humains restant en charge de l'essentiel du travail de réponse aux usagers[59].

[58] Dans le domaine des interactions humain-machine, on appelle ainsi des systèmes informatiques qui couplent un module de synthèse vocale et un module de traitement de questions/réponses.

[59] Il convient de signaler que ces systèmes ont acquis une certaine qualité de réponse et que de nombreux usagers ne se rendent pas compte, au terme de l'échange, qu'ils ont été confrontés à une machine.

L'étude du comportement de l'usager avec une machine mobilise un effort de recherche conséquent, notamment développé au sein des équipes de recherche en télécommunications[60]. Notre travail va consister à comparer la *conduite interactionnelle* des correspondants humains, en interaction avec un opérateur humain ou avec un opérateur machine. Nous proposerons des *procédures de traitement textométriques*[61] adaptées à ces comparaisons.

1 Contexte et motivations de la recherche

Certaines notions mobilisées pour la description des données du corpus sont empruntées au cadre méthodologique de l'*analyse conversationnelle*[62]. Plusieurs phénomènes relevant de divers niveaux de description linguistique sont analysés, à l'aide des mêmes outils de statistique textuelle. Après une description détaillée des données de travail, les analyses quantitatives nous permettront d'aborder les typologies conversationnelles globales du corpus, pour enfin discuter la question de l'ajustement conversationnel chez l'usager.

==== *Glossaire minimal pour l'analyse des conversations* ====

Situation :	contexte dans lequel sont situées les interactions
Interactants :	locuteurs en relation d'interaction pour mener à bien une activité sociale (parfois *participants* dans la littérature)
Interaction :	échange entre au moins deux interactants, qui peut être verbal ou non verbal
Conversation :	cas particulier de l'interaction, caractérise les échanges verbaux entre les interactants
Séquence :	suite de tours de parole formant un ensemble fonctionnel distinct dans la conversation
Tour de parole :	temps durant lequel l'un des interactants garde la parole (parfois abrégé en *tour* dans la littérature)

La *situation* renvoie au contexte où sont situées les interactions ; en l'occurrence, l'appel téléphonique d'un usager vers la plateforme de renseignements d'une société de services.
L'*interaction* correspond à un échange entre au moins deux *interactants*, qui sont en présence pour mener à bien une activité sociale. Cet échange peut être verbal ou non verbal ; étant donnée la situation d'interaction, le corpus étudié ne contient que des échanges verbaux. La *conversation* est un cas particulier de l'*interaction*. Cette pratique sociale caractérise les échanges verbaux[63] entre des *interactants*.
Ces échanges se structurent en *séquences*, unités qui décrivent des étapes distinctes de la *conversation*. Certaines étapes sont attendues, telles que les séquences d'*ouverture* ou de *fermeture*, qui consistent par exemple en des échanges de salutations en début et en fin de

---

[60] En particulier, la conférence IHM'07 a été consacrée à l'étude de l'engagement de l'utilisateur dans les interactions verbales homme/machine. On peut, par exemple, consulter (Ech Chafai et al., 2007).

[61] Pour davantage de précisions, consulter, par exemple, (Lebart et Salem, 1994).

[62] On renvoie en particulier à (Sacks et al., 1974) pour davantage de précisions sur le domaine de l'*analyse conversationnelle*. Pour une présentation détaillée des unités minimales de l'infrastructure conversationnelle, voir par exemple (Portes et Bertrand, 2005).

[63] Les échanges *signés* – exprimés en langue des signes – sont inclus dans les *conversations*. De plus, avec l'émergence des nouvelles technologies de communication médiatisée par ordinateur (CMO), il est aujourd'hui admis que certains échanges écrits – SMS, messagerie instantanée, par exemple – appartiennent au paradigme conversationnel.

conversation. Chaque *séquence* est constituée de *tours de parole*, un *tour de parole* correspondant au temps durant lequel l'un des *interactants* garde la parole. Les *tours de parole* peuvent entrer en relation de *pertinence conditionnelle* pour former une *paire adjacente*. Une séquence d'ouverture peut par exemple être composée de la paire suivante :

```
O1 (Interactant A): sncf bonjour
C1 (Interactant B): allô bonjour madame est-c'que je pourrais/ pourrais savoir
                    e : le prix d'un billet e :, paris rouen s'il vous plaît
```

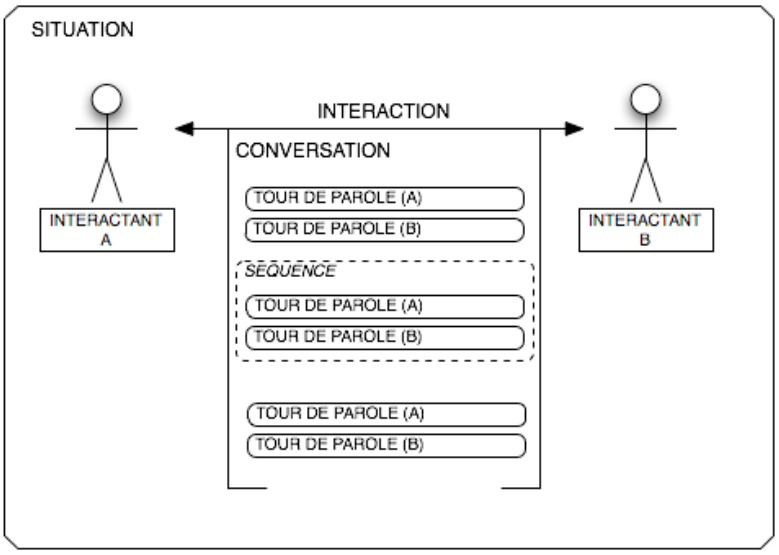


Figure 1
Les différentes unités de description

Objectifs de cette étude

Deux pistes possibles émergent pour l'analyse du corpus *Interactions* selon des procédures textométriques :

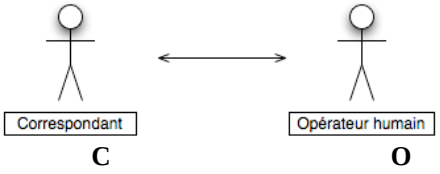
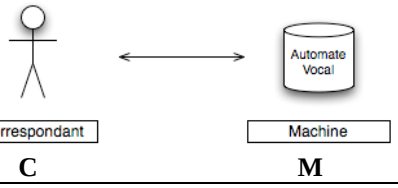
- la comparaison des réponses de la machine (M) à celles de l'opératrice (O),
- la comparaison de l'adaptation du correspondant (C) à un interlocuteur machine (M) d'une part, à un interlocuteur humain (O) d'autre part.

La première piste constitue, selon nous, une question mineure. En effet, les *disfluences* transcrites dans les tours de parole, les *phatiques* en particulier, ne résisteraient pas à l'épreuve des analyses textométriques. Il faudrait, pour ne pas tomber sur des résultats évidents, procéder à une *standardisation* du corpus, et partant, le purger *intégralement* des disfluences présentes dans les tours de parole. Cela s'avère une opération fastidieuse, les conventions de transcription utilisées rendant la normalisation du corpus quasiment impossible à réaliser par des procédures automatiques. Standardiser les disfluences du corpus à la main serait pertinent afin, par exemple, d'évaluer la variété linguistique des routines de la machine, ce qui restreindrait par contre l'étendue des analyses à la dimension lexicale.

La seconde piste est un axe d'étude qui nous paraît plus riche, car elle permet de ne pas dissocier dans l'analyse les dimensions locales et globales du corpus. On peut se demander si le correspondant, en situation de demande d'information, présente, dans les différentes dimensions de son discours, des indices spécifiques d'un *ajustement discursif* à l'interlocuteur avec lequel il converse. En somme, ce qui aurait pu être un obstacle en soi – purger manuellement l'intégralité des disfluences – s'avère un atout important pour la piste d'analyse

que nous privilégions ici. Nous posons en effet l'hypothèse que la présence des disfluences est l'un des paramètres qui permettra d'apprécier les traits distinctifs de l'ajustement du correspondant humain (C), selon qu'il interagit avec une machine (M) ou un humain (O).

Tableau 1 :
Types d'interactions dans le corpus analysé

Interactions entre humains	*Interactions humain-machine*
	
– 117 interactions – 2 713 tours de parole produits par C à destination de O – 2 769 tours de parole produits par O à destination de C	– 143 interactions – 1 198 tours de parole produits par C à destination de M – 1 607 tours de parole produits par M à destination de C
Exemple d'interaction PHASE 1 – COMMUNICATION 11 O1 :-- sncf bonjour C1 :-- allô bonjour madame est-c'que je pourrais/,pourrais savoir e: le prix d'un billet e:, paris rouen, s'il vous plaît O2 :-- un prix paris/pour e:/paris rouen (h) en:: quelle classe s'il vous plaît C2 :-- en deuxième O3 :-- en deuxième vous n'bénéficiez pas de:: réduction C3 :-- non non O4 :-- et vous voulez: un: aller-simple C4 :-- oui O5 :-- d'accord ne quittez pas s'il vous plait - O6 :-- allô? C5 :-- oui O7 :-- (h) oui alors le prix en: plein tarif deuxième classe aller simple est de soixante deux francs C6 :-- et le retour c'est la même chose alors? O8 :-- oui ça vous fait cent vingt-quat(re) francs aller retour C7 :-- d'accord j'vous r'mercie au] r'voir]]	**Exemple d'interaction** PHASE 3 - COMMUNICATION 39 M1 :-- sncf bonjour, énoncez votre demande s'il vous plaît C1 :-- voilà/ je voudrais, e savoir, pour aujourd'hui sam'di, quels sont les trains: qui vont à bonnière-sur-seine dans l'après-midi ? M2 :-- vers quelle heure précise désirez-vous partir? C2 :-- e: entre:: trois heures et d'mi/ entre[trois heures et/ e: et six heures M3 :-- ne quittez pas- M4 :-- le seul train que vous pouvez prendre est un autorail première et deuxième classes partant de la gare de paris saint-lazare à dix-sept heures vingt-sept minutes arrivant à la gare de bonnière à dix-huit heures quinze minutes, ce train vous convient-il? C3 :-- écoutez c'est parfait mad'moiselle M5 :-- désirez-vous d'autres renseignements? C4 :-- oui qu'est-c'que vous faites ce soir ? M6 :-- ceci n'est pas de mon domaine, ne quittez pas je vous passe la personne compétente

2 Le corpus *Interactions*

Ce corpus est constitué par un ensemble d'entretiens[64], centrés sur la réservation de billets de trains et les informations sur leurs horaires de circulation. Il a été enregistré en 1985 par le groupe de recherche PRC Communication Homme-machine du LIMSI[65], pôle de recherche national pour l'élaboration et l'évaluation des systèmes de communication humain-machine.

[64] Ces entretiens ont été enregistrés par la SNCF, *Société Nationale des Chemins de fer France*, qui a possédé le monopole de l'exploitation du réseau ferroviaire français de sa création, en 1938, à l'ouverture du réseau à la concurrence, en 2005.

[65] De plus amples informations sur ce groupe de recherche sont accessibles via le lien suivant : http://www.limsi.fr/RS96FF/CHM/CAM.html

2.1 Les données recueillies

Le corpus comprend 260 interactions. Le tableau 1 récapitule les types d'interaction observés et le nombre de tours de parole produits par chacun des interactants. Dans le cas de l'interaction *humain-machine*, le *système intelligent* avec lequel interagit l'usager possède un module de synthèse vocale synthétisant une voix humaine (féminine en l'occurrence). L'opérateur humain est toujours de sexe féminin et aucune information externe aux tours de parole du corpus ne permet de déterminer le sexe du correspondant humain[66].

Tableau 2 :

Rôles conversationnels et nature des différents interactants du corpus

	Interactants		
	Correspondant humain (C)	*Opératrice humaine (O)*	*Machine (M)*
Rôle conversationnel	Demandeur	Fournisseur	Fournisseur
Nature	Humain	Humain	Machine

Le corpus étudié est formé de textes recueillis dans deux situations d'interaction de type *requête/réponse*[67], qu'on distingue selon la nature des interactants impliqués :

- un *correspondant humain* dialoguant avec *un opérateur humain* formé au renseignement téléphonique de l'entreprise ;
- un *correspondant humain* dialoguant avec un *automate vocal* programmé pour fournir une réponse aux demandes d'information.

Trois *interactants* peuvent être distingués :

- un *correspondant humain (C)* appelant pour obtenir des informations ;
- une *opératrice humaine (O)* prenant en charge une partie des appels de C ;
- une *machine (M)* prenant en charge l'autre partie des appels de C.

Deux *rôles conversationnels* sont donc tenus par les *interactants* :

- le rôle de *demandeur d'information* ;
- le rôle de *fournisseur d'information*.

2.2 Mise en forme des données

L'étape de normalisation a pour principal objectif une exploitation du corpus fondée sur des données comparables par des procédures d'analyse textométriques. Le corpus original, au

---

[66] Le genre du correspondant humain est un paramètre qu'il serait intéressant de prendre en compte dans le cadre d'analyses sur la conduite interactionnelle en situation de demande d'information.

[67] Nous utilisons les critères structurels introduits par (Sacks et al., 1974), dans le cadre du modèle du *Turn Taking System*, pour caractériser les interactions du corpus. (Sacks et al., 1974) formalisent la structuration des échanges conversationnels en *paires adjacentes*, unités de description de la dynamique conversationnelle, fondées sur l'alternance de *tours de parole* entre les interlocuteurs. Les *paires adjacentes* rendent possible l'accomplissement d'activités sociales, la *demande de renseignements* en l'occurrence. Une *paire adjacente* est une suite connexe de deux tours de parole, entretenant une relation de *pertinence conditionnelle* et produits par deux interlocuteurs différents. La notion de *pertinence conditionnelle* renvoie au fait qu'une activité sociale donnée induit la présence de certains types de *paires*. En principe, dans le cadre des interactions du corpus, la réussite de l'activité sociale – l'échange téléphonique pour une demande de renseignements – est satisfaite si chaque question posée par le *correspondant* (C) trouve des éléments de réponse dans les tours de parole de ses interlocuteurs M ou O.

format texte brut, est une transcription orthographique d'interactions téléphoniques où se déroulent des échanges conversationnels. Cette transcription est enrichie de deux niveaux d'annotation, le premier décrivant des phénomènes audibles, le second donnant à voir la structure des échanges. Etant donnés les objectifs de la recherche, nous choisissons de normaliser le corpus pour aplanir un certain nombre de différences évidentes entre les interactants *humain* et *machine*. Nous sommes partie des transcriptions recueillies sur le site de la FreeBank [68], dont on peut voir des exemples en tableau 1. Différents types de *disfluences*[69] sont signalés : les *phatiques*, tels que *euh*, transcrits *e* ; les *recouvrements de parole*, marqués par des combinaisons variables du caractère *]* pour en indiquer le début et la fin. Les *reprises* et *répétitions* sont indiquées par */*. Le marqueur *,* indique les *pauses* et peut être doublé ou triplé pour fournir une information sur la durée de la pause. Les conventions de transcription utilisées (tableau 3), qui faisaient partie intégrante de la ressource téléchargée, présentent comme un *énoncé* ce que nous considérons comme des *tours de parole*[70].

Tableau 3 :
Conventions de transcription utilisées dans le codage du corpus *Interactions*

:--	précédé de l'initiale identifiant le locuteur, ce symbole marque le début d'un énoncé "normal" c'est-à-dire commençant pendant un silence et non simultanément avec d'autres interventions
]	à l'intérieur d'un énoncé, indique qu'à ce moment une autre voix intervient, pouvant provoquer un recouvrement
]]	note la fin du recouvrement des voix
]-	indique que l'énoncé qui suit se présente comme une intervention située à l'intérieur même de l'énoncé du précédent locuteur, provoquant par là un recouvrement de paroles ou une interruption ayant été indiquée par la marque] au moment où elle s'est produite
,	note une pause, même brève (,/,,/,,, selon la durée de la pause)
/	note une reformulation ou une répétition d'un mot dans l'énoncé
-	note une pause finale

Une première phase de normalisation a consisté à purger certains phénomènes de l'oral spontané appartenant au paradigme des *disfluences*. Les marqueurs des *recouvrements de parole* et des *pauses* internes aux tours de parole ont été éliminés, les caractères qui les signalent étant interprétés comme des segmenteurs du fil textuel par les outils de textométrie utilisés par la suite. Le but de cette opération est d'isoler correctement les unités dans la chaine textuelle, afin d'obtenir des décomptes pertinents sur les occurrences de formes. Dans le même temps, les marqueurs de *reprises* et de *répétitions*, ont également été purgés. Seuls les *phatiques*, qui se présentent sous des formes faiblement distinctives dans le corpus

[68] La FreeBank (http://freebank.loria.fr/corpus.php) est la banque de corpus ouverte du LORIA.

[69] Pour davantage de précisions typologiques autour de la notion de *disfluence*, voir par exemple (Schriberg, 1994).

[70] Dans le cadre méthodologique de l'analyse conversationnelle, l'*énoncé* et le *tour de parole* ne recouvrent pas la même réalité linguistique : un *tour de parole* peut être non verbal, consister en un phatique ou comprendre plusieurs énoncés, alors qu'un *énoncé* est soumis des conditions de complétude ou de vérité, selon les approches. Les approches traditionnelles de l'énoncé en *linguistique* ne considèrent par exemple pas qu'un mot isolé ou un phatique peut constituer un énoncé. De la même façon, les approches issues de la *philosophie du langage* considèrent l'énoncé comme une unité à laquelle on peut attribuer une valeur de vérité.

original, ont été conservés, leur transcription rendant délicate l'application d'une procédure automatique pour les normaliser.[71]
La transcription originale distingue les *interactions* (marqueurs *COMMUNICATION*) et les *tours de parole* qui les composent (marqueurs *:*--). Une seconde phase de normalisation consiste à adapter ce découpage du texte en *parties*, pour en rendre comparables ces deux types de *contenants* du texte. On normalise donc d'une part les paramètres de *segmentation* de la chaine textuelle, d'autre part les paramètres de *partition,* pour les rendre adéquats à l'analyse textométrique.

Tableau 4 :
Adaptation de la structuration du corpus pour la normalisation des interactions

Extrait d'une interaction *humain-machine* avant normalisation (extrait)

```
PHASE 3 - COMMUNICATION 39
M1 :-- sncf bonjour, énoncez votre demande s'il vous plaît
C1 :-- voilà/ je voudrais, e savoir, pour aujourd'hui sam'di, quels sont les trains: qui
   vont à bonnière-sur-seine dans l'après-midi ?
M2 :-- vers quelle heure précise désirez-vous partir?
C2 :-- e: entre:: trois heures et d'mi/ entre[ trois heures et/ e: et six heures
M3 :-- ne quittez pas-
```

Extrait d'une interaction *humain-machine* après normalisation (extrait)

```
<COMMUNICATION=238>
<TdP=MC1284> sncf bonjour, énoncez votre demande s'il vous plaît
<TdP=CM0944> voilà je voudrais, savoir, pour aujourd'hui samedi, quels sont les trains qui
               vont à bonnière-sur-seine dans l'après-midi?
<TdP=MC1285> vers quelle heure précise désirez-vous partir?
<TdP=CM0945> entre trois heures et demi entre trois heures et e et six heures
<TdP=MC1286> ne quittez pas
```

3 Analyses quantitatives sur le corpus *Interactions*

Pour mettre en évidence les éléments de typologie globale du corpus, on s'appuie sur le *découpage* du corpus en *contenants*, les *tours de parole*, dont on rend transparente la trajectoire « locuteur courant → interlocuteur ». La procédure de *découpage* appliquée permet de distinguer *quatre types de contenants* :

- ceux adressés par le correspondant humain (C) à la machine (M),
- ceux adressés par le correspondant humain (C) à l'opératrice (O),
- ceux adressés par la machine (M) au correspondant humain (C),
- ceux adressés par l'opératrice (O) au correspondant humain (C).

Le typage des tours de parole permet donc de caractériser chacune des situations d'interaction du corpus selon qu'elles impliquent deux humains ou un humain dialoguant avec une machine : les parties MC et CM caractérisent le premier type, les parties OC et CO le second.

[71] Traiter intégralement un grand corpus tel que celui-ci, pour en purger les *phatiques*, mobiliserait une équipe de plusieurs personnes pour plusieurs semaines de travail. Une telle opération est très coûteuse et ne se justifie que si l'on souhaite, par exemple, comparer les productions de M et de O pour rendre les productions de la machine plus proches des tours de parole humains. Hors, cet aspect de l'amélioration des interfaces humain-machine est déjà fort bien documenté, notre apport serait donc peu utile de ce point de vue.

Tableau 5 :
Tours de parole typés en fonction du rôle et de la nature des interactants

Rôle interactionnel / *Nature des interactants*	*Demandeur d'information*	*Fournisseur d'information*
Humain-humain	CO	OC
Humain-machine	CM	MC

3.1 Premiers décomptes

Nous commençons par quantifier les différentes unités de description des interactions présentées plus haut (figure 1). On s'intéresse en particulier aux différents types d'interactants, ainsi qu'aux deux types d'interaction distingués.

Décomptes par type d'interactant

Tableau 6 :
Principales caractéristiques quantitatives du corpus *Interactions*

	Occurrences	*Formes*	*Hapax*	*Forme la plus fréquente*	
Corpus	79 043	1 971	749	2 486	vous
C	30 812	1 978	803	1 331	oui
O	26 740	1 193	477	871	vous
M	21 491	547	158	1145	vous

Le tableau 6 présente les principales caractéristiques quantitatives du corpus. Le déictique personnel vocatif *vous*, forme la plus fréquente, représente près de 32% des occurrences du corpus. Cela s'explique en partie par un contexte où les situations d'interaction sont de nature *formelle* et où le *vous* de politesse est obligatoire.

Le demandeur d'information (C) produit l'adverbe *oui* plus que toute autre forme, acquiescement qui met en avant la fonction conversationnelle de récepteur d'information. La machine, quant à elle, use d'un nombre de formes relativement restreint, en particulier comparé à l'opératrice. C'est là l'indice d'une redondance dans les productions de la machine.

Décomptes par type d'interaction

Nous commençons par comparer la répartition des tours de parole, entre les conversations entre humains ou humain-machine.

D'un point de vue quantitatif, si le nombre de conversations *humain-humain* et *humain-machine* est équilibré – respectivement 117 dans le premier cas, contre 143 dans le second – on relève un écart plus conséquent au niveau des tours de parole produits – respectivement 5 482 dans les conversations entre humains, contre 2 805 dans les conversations humain-machine : seuls 34% des tours de parole sont produits dans ces dernières. D'après les premiers décomptes sur le corpus brut, on peut déjà dire que la négociation pour *l'allocation des tours*

de parole explique en partie ces différences quantitatives. En effet, les *autosélections*[72] du correspondant (C) sont plus fréquentes lors des échanges avec l'opératrice humaine (O) qu'avec la machine (M). Les *fins de recouvrements de parole*, indiqués par des marqueurs spécifiques, constituent un critère pertinent pour comparer la répartition des phénomènes d'*autosélection* dans les interactions du corpus. En effet, lorsque les tours de parole de deux interactants se recouvrent, c'est que l'un des deux s'est arrogé un *tour de parole* avant que l'interlocuteur n'ait achevé le sien ou n'y ait donné à voir de *point de transition possible*. La *négociation* pour l'attribution d'un tour de parole à l'un ou l'autre des interactants est donc plus longue, puisqu'elle s'étend sur plusieurs tours de parole. On relève un ratio de 6% de *fins de recouvrement de parole* dans les interactions humain-machine, contre 94% dans celles impliquant deux humains[73]. Ces premiers éléments d'observation des disfluences dans le corpus brut montrent que les *autosélections* sont plus fréquentes dans les interactions humain-humain que dans celles impliquant un humain et une machine. La négociation pour *l'allocation des tours de parole*, plus difficile entre les humains, contribue donc à expliquer cette différence quantitative.

Dans le cadre des conversations entre humain et machine, la négociation pour l'allocation des tours de parole est moins longue en moyenne, du fait de l'existence de deux facteurs distincts :

- la machine ne coupe jamais la parole à son interlocuteur, les cas de recouvrement de parole étant déclenchés par l'humain ;
- la structuration du message produit par la machine semble décourager toute interruption intempestive des la part des humains.

3.2 Quelques entrées textométriques

Nous avons appliqué à ce corpus d'interactions les différentes procédures textométriques que l'on emploie pour analyser les ensembles de textes numérisés.

L'accroissement du vocabulaire

La figure 2 montre les courbes d'accroissement du vocabulaire calculées pour chacun des types de situation d'interaction : on l'a vu, les parties MC et CM caractérisent les interactions humain-machine, les parties OC et CO celles entre humains. Le fait que, dans le corpus que nous avons construit, le volume des transcriptions retenues pour chacun des types de communication soit inégal, explique que certaines de ces courbes s'interrompent plus tôt que les autres sur l'axe horizontal. Il est par contre possible de comparer les différentes courbes sur la partie gauche du graphique en ne considérant que des volumes comparables.

On constate tout d'abord que le vocabulaire de la machine (courbe rose – MC, machine/correspondant) croît de façon beaucoup moins importante que les trois autres courbes, qui correspondent à des productions humaines. Le décrochement important que l'on observe à partir de l'abscisse 3000 correspond au démarrage de routines spécifiant l'offre sur les trajets, déclenchées par des questions de confort et de tarifs qui n'avaient pas été introduites dans la partie précédente du corpus. Les thématiques introduites portent par exemple sur la classe du train choisie pour le voyage, caractérisée par le segment répété *première et deuxième classes*.

---

[72] On parle d'*autosélection* lorsque l'un des interactants prend la parole sans que l'interlocuteur la lui ait accordée, ce qui s'appelle, dans le langage courant, *couper la parole*.

[73] Le décompte opéré sur les segments « *]]* », marqueurs des fins de recouvrement de parole entre deux tours de parole, donne les fréquences absolues suivantes : 67 occurrences dans les interactions entre C et M; 1036 occurrences dans les interactions entre C et O. Cette analyse s'est déroulée sur la version du corpus brut.

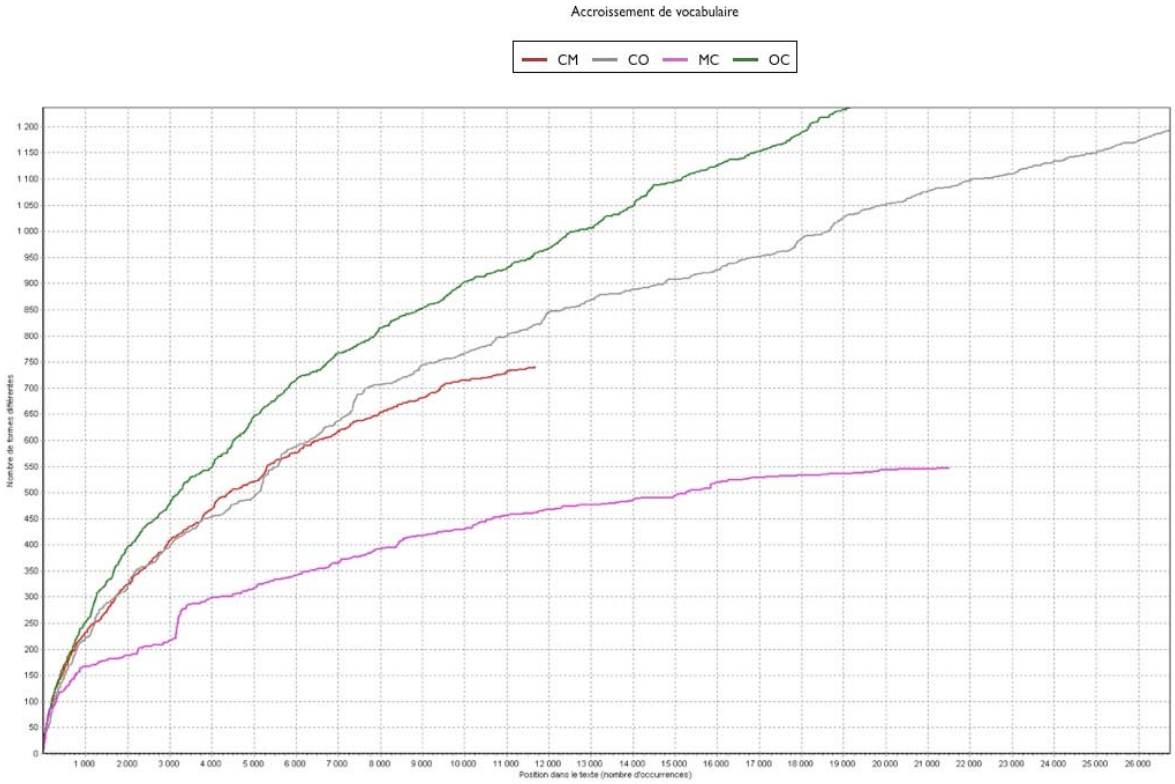


Figure 2

Accroissement du vocabulaire par type de tour de parole
***CM** : correspondant-machine, **CO** : correspondant-opératrice*
***MC** : machine-correspondant, **OC** : opératrice-correspondant*

Les deux courbes qui correspondent à des échanges entre humains (courbe verte – OC, opératrice/correspondant et courbe grise – CO, correspondant/opératrice) sont situées sur le haut du graphique, ce qui indique une variété du vocabulaire plus importante. De ces deux types d'interaction, ce sont les productions de l'opératrice qui possèdent la plus grande variété de vocabulaire, comparé à celles des demandeurs d'information.

Les productions des correspondants confrontés à une machine (courbe rouge – CM, correspondant/machine) occupent une position intermédiaire dans ce classement. On en déduit que, dans cette situation, le demandeur, même si rien ne permet de noter qu'il a conscience de s'adresser à une machine, est amené à réduire l'étendue de son vocabulaire. Ainsi, on peut dire que les productions du correspondant, portent la trace d'un *ajustement conversationnel à* la variété du vocabulaire de chacun de ses interlocuteurs.

Déictiques et clôtures

Il nous a paru intéressant de comparer, de manière similaire, les ancrages et clôtures conversationnels réalisés au cours de chacun des types d'interactions. Nous montrons que les déictiques, comme les usages de fin de conversation, sont des indicateurs importants de l'ajustement conversationnel du correspondant.

Lorsqu'il converse avec l'opératrice, le correspondant produit une plus grande variété de déictiques personnels. Nous avons donc choisi de projeter sur un même graphique (figure 3) les spécificités des principaux déictiques du corpus : *je* (1 176 occurrences), *vous* (2 486

occurrences), *il* (1 382 occurrences), *on* (126 occurrences) et *nous* (16 occurrences)[74]. La présence très spécifique du pronom *je*, dans ses productions confirme l'existence d'une dimension interpersonnelle plus forte dans la conduite interactionnelle du correspondant, lorsqu'il interagit avec un humain.

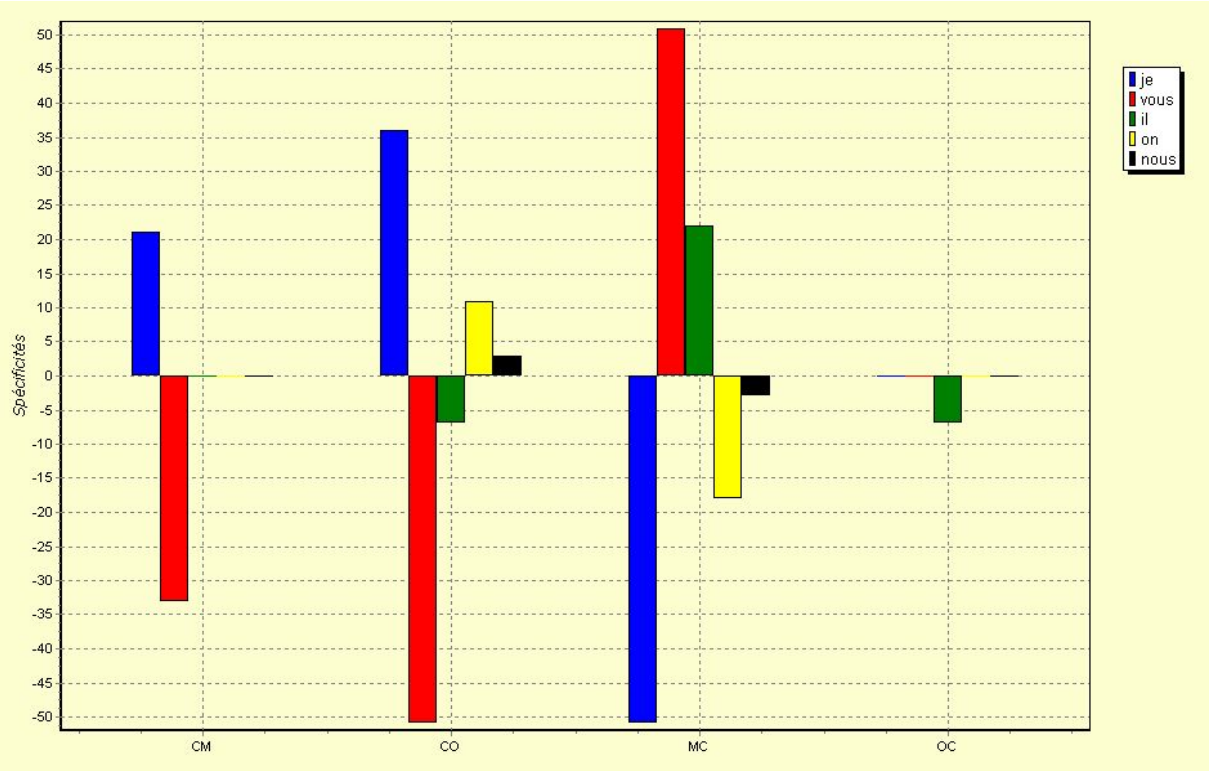


Figure 3

Ventilation des spécificités des formes *je*, *vous*, *il*, *on*, *nous* par type de tour de parole
***CM** : correspondant-machine, **CO** : correspondant-opératrice*
***MC** : machine-correspondant, **OC** : opératrice-correspondant*

Contrairement aux interactions avec l'opératrice, lorsque le correspondant s'adresse à la machine, le pronom *je* figure dans des tours de parole sans disfluence et sans indice explicite de la dimension interpersonnelle.

<TdP=CO0124> donc on peut y aller comme ça d'autre part **je** vais vous demander un renseignement, est-ce que la réduction est valable par exemple sur un billet paris brussell?

<TdP=CO0149> si **je** veux réserver je je je demande le train cinq mille neuf cent quarante-cinq

En particulier, les tours de parole adressés à la machine présentent une régularité structurelle importante : *je voudrais + [informations sur la circulation des trains]*

<TdP=CM1177> **je** voudrais les horaires des trains pour rouen au départ de paris saint-lazare pour ce soir

<TdP=CM1082> pour le lundi vingt-huit janvier, **je** voudrais l'heure d'un train partant de bâle, pour paris

On note le vouvoiement adressé de manière préférentielle à la machine, plutôt qu'à l'opératrice, ce qui indique une différentiation opérée de manière plus ou moins consciente par l'appelant.

<TdP=CM0016> **vous** n'avez pas d'autres trains ?

[74] Ces décomptes sont opérés sur les formes originales, non lemmatisées au préalable.

<TdP=CM0166> très bien, **vous** pouvez me donner le prix du billet

Deux emplois principaux sont observés pour l'utilisation du pronom *il* : d'une part, un emploi impersonnel, en particulier dans les figements de type *il faut* ou *il faudrait* ; d'autre part, un emploi anaphorique, où le pronom renvoie au moyen de transport.

<TdP=CO0198> oui oui je sais bien, **il** faut passer par vous et j'ai appelé déjà et c'était à neuf heures moins le quart je me suis dit peut-être ils font la journée continue

<TdP=CO0517> autrement **il** part de paris à quelle heure le deuxième?

La fréquence de la forme *on* est remarquable dans les interactions entre humains, où ce pronom figure essentiellement dans les confirmations de renseignements donnés par l'opératrice. Cette forme est absente des tours de parole produits par la machine, elle est nettement moins employée par les correspondants qui échangent avec une machine, ce qui constitue un autre indice de l'ajustement conversationnel.

<TdP=CO0097> allô oui bonjour madame, je voudrais avoir des horaires je sais pas **on** m'a donné des horaires e suivants pour paris le creusot le quatorze décembre à seize heures quarante-neuf, et moi je les trouve pas

<TdP=CO0121> **on** n'a pas besoin de photo

La forme *nous* est très peu produite et apparaît de façon privilégiée dans des contextes locaux de disfluence, au sein de tours de parole adressés à l'opératrice. Comme le pronom *on*, sa fréquence est remarquable dans les interactions entre humains.

<TdP=CO1303> non non non, non non non non **nous** partirions mardi je ne sais pas je crois qu'il doit y avoir un train dans l'après-midi fin de l'après-midi

<TdP=CO0474> - oui c'est ça c'est que **nous** on va e je c'est pour une maison d'retraite et je dispose pas de beaucoup de temps si vous voulez

En second lieu, les analyses sur la spécificité des *segments répétés* (SR) ventilés (figure 4) permettent d'identifier des tendances complémentaires dans les stratégies mises en œuvre pour clore les conversations.

Si le SR *je vous remercie* (en vert) n'est pas plus spécifique des tours de parole de type CM que de ceux de type CO, la forme *merci* (en rouge) est par contre caractéristique des échanges du correspondant avec la machine. En corrélant ce constat avec les observations sur l'accroissement du vocabulaire (figure 2) on en déduit que les remerciements adressés à la machine par le correspondant ont des formulations moins variées.

Par ailleurs, la machine ne produit jamais de marque d'agrément du remerciement, ce qu'indiquent les résultats pour le SR *de rien* (en jaune). Cela ajoute au caractère non régulier de la conduite interactionnelle de la machine. Enfin, le parallélisme d'emploi de la clôture conversationnelle *au revoir* (en bleu) entre les tours de parole CO et OC est un indice supplémentaire de l'ajustement conversationnel du correspondant en fonction de son interlocuteur.

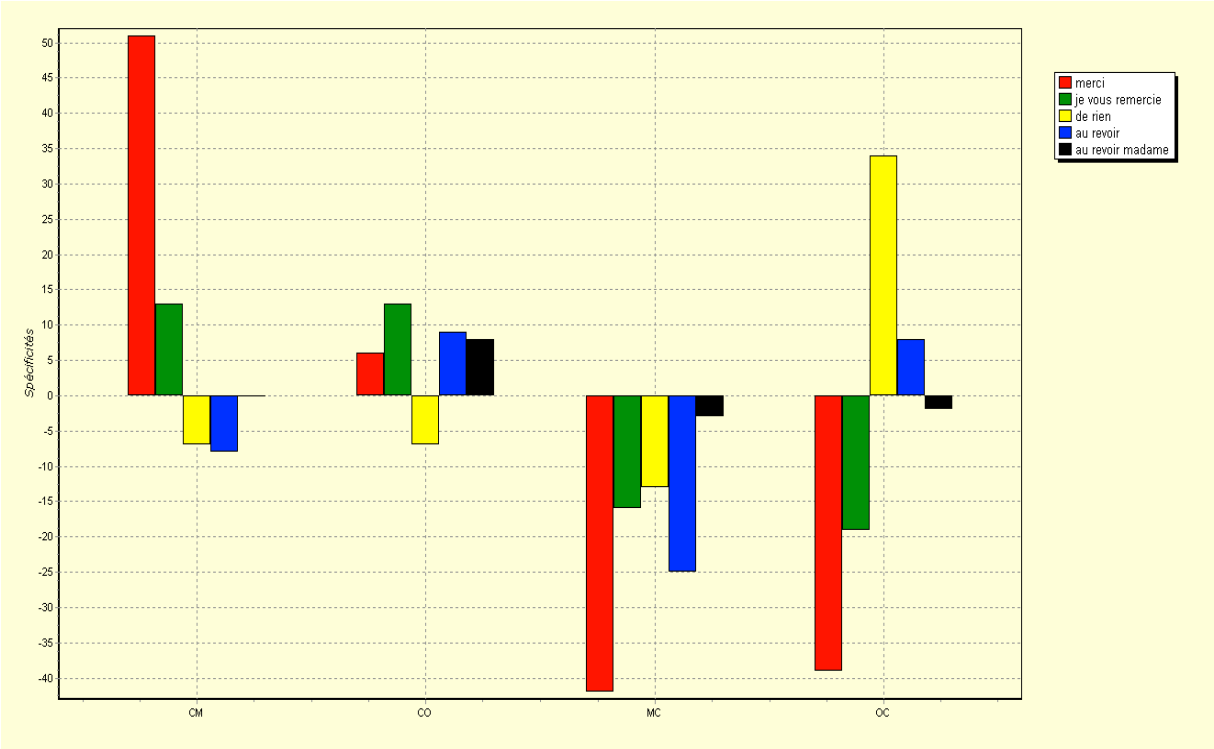


Figure 4

Ventilation des spécificités des segments *merci, je vous remercie, de rien, au revoir, au revoir madame*, par type de tour de parole

***CM** : correspondant-machine, **CO** : correspondant-opératrice*

***MC** : machine-correspondant, **OC** : opératrice-correspondant*

La spécificité du SR *au revoir madame* (en noir) dans les tours de parole de type CO, qui rajoute une dimension de politesse à la clôture conversationnelle avec le substantif de posture sociale *madame*, corrobore cette idée d'ajustement.

Comparaison des types de tours de parole

L'analyse factorielle des correspondances (AFC) donne une visualisation synthétique des proximités entre les différentes parties confrontées, en fonction de leur vocabulaire (figure 5).

L'analyse a été réalisée à partir du tableau croisant les 723 formes de fréquence supérieure à 5 dans le corpus et les quatre types de tours de parole.
Un premier axe, horizontal sur le graphique, se détache nettement (64% de l'inertie totale) qui oppose les tours de parole produits par la machine aux productions des trois autres intervenants humains. On trouvera au tableau 7 les spécificités des productions de chacun de ces groupes d'actants.
Le deuxième facteur (24% de l'inertie totale), oppose les tours de parole adressés à la machine à ceux qui sont échangés entre humains. Nous analyserons plus loin cette opposition comme une adaptation du demandeur à son interlocuteur.

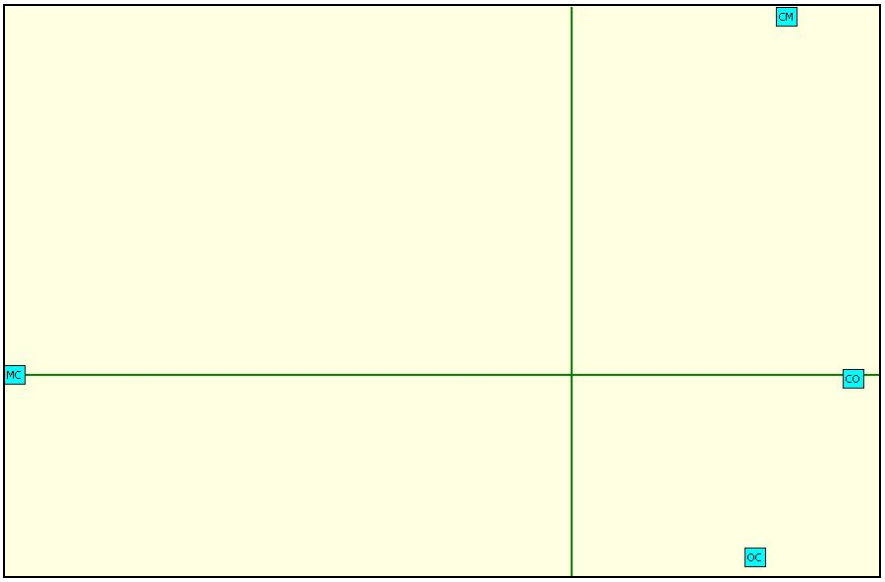


Figure 5

Représentation factorielle des productions par type de tour de parole
CM *: correspondant-machine,* ***CO*** *: correspondant-opératrice*
MC *: machine-correspondant,* ***OC*** *: opératrice-correspondant*

4 Typologies conversationnelles

La situation d'interaction impose aux interactants des rôles conversationnels : on distingue les *demandeurs d'information* (C), des *fournisseurs d'information* (M et O).

4.1 Rôles conversationnels

On utilise la métrique du *calcul des spécificités* pour contraster les différents types de tours de parole, en fonction du rôle des interactants.

Les demandeurs d'information

La première catégorie d'interactants présente dans le corpus est celle des *demandeurs d'information*, représentés par les correspondants. Ce sont des usagers de la SNCF qui soumettent des requêtes sur la circulation et la réservation des trains aux services de renseignement de la société. Pour donner des éléments de typologie de leurs productions, nous présentons en tableau 7 les 30 formes les plus *spécifiques* de leurs tours de parole, selon qu'ils sont en interaction avec une opératrice humaine ou une machine.

Ces résultats font apparaître un premier trait distinctif de la conduite interactionnelle du *demandeur d'information* en fonction de son interlocuteur, ce qui répond à l'un des objectifs de cette recherche (section 1), qui vise à identifier des indices de son *ajustement conversationnel* en fonction de la nature de l'interlocuteur, humain ou machine.

On observe en premier lieu que les phatiques spécifiques des tours de parole de type CO sont plus nombreux. Il y a donc une plus grande variété de phatiques adressée par le correspondant lorsqu'il interagit avec un interlocuteur humain.

Tableau 7 :

Productions des *demandeurs d'information*,
en fonction de l'interlocuteur (les 30 formes les plus spécifiques)

Adressées à la machine (M)				*Adressées à l'opératrice (O)*			
Forme	*Frq. Tot.*	*Fréquence*	*Coeff.*	*Forme*	*Frq. Tot.*	*Fréquence*	*Coeff.*
train	273	196	+32	oui	1188	252	-34
paris	307	193	+21	accord	222	21	-21
merci	297	187	+20	madame	118	3	-20
le	753	395	+19	ah	197	19	-19
horaires	151	104	+16	bonjour	118	6	-17
voudrais	202	129	+15	parce	77	1	-15
non	360	204	+15	bon	228	34	-14
après	192	122	+14	ça	289	53	-13
pour	528	273	+13	ben	193	30	-12
vers	128	85	+12	hein	123	12	-12
matin	137	90	+12	revoir	83	5	-11
horaire	46	38	+11	que	260	48	-11
aimerais	55	44	+11	ouais	55	2	-9
janvier	56	44	+11	pas	299	70	-8
e	360	191	+11	est	700	193	-8
suivant	31	28	+10	alors	240	50	-8
midi	119	77	+10	vais	47	2	-8
trains	134	84	+10	là	107	17	-7
de	732	350	+10	quarante	91	12	-7
décembre	63	45	+9	moi	65	8	-6
très	74	51	+9	par	54	7	-5
un	502	247	+9	mon	26	1	-5
les	326	172	+9	donc	83	15	-5
parfait	22	20	+8	oh	56	8	-5
connaître	58	42	+8	au	159	35	-5
samedi	82	55	+8	on	79	16	-4
départ	85	56	+8	cinquante	99	22	-4
lundi	47	34	+7	tgv	52	9	-4
prix	82	52	+7	peut	57	11	-4
brussel	13	13	+7	sinon	19	1	-4

La présence de plusieurs déictiques dans ces listes nous amène à faire les remarques suivantes :

- les déictiques temporels – *après, matin* – sont spécifiques des tours de parole que le demandeur d'information adresse à la machine ;

- les déictiques personnels caractérisent les tours de parole adressés à l'opératrice humaine – *moi, mon, on* – ce qui indique que le correspondant en situation de demande d'information se met davantage au premier plan avec un interlocuteur humain ;
- les déictiques de lieu produits par le correspondant ont des spécificités différentes selon la situation d'interaction : des noms propres – *paris, brussel* – sont adressés à la machine, alors que les noms communs ou adverbes – *là, tgv* – sont spécifiques des tours de parole destinés à l'opératrice humaine.

Les modes verbaux spécifiques des tours de parole adressés à la machine sont exclusivement au conditionnel – *aimerais, voudrais* – alors que ceux destinés à l'opératrice sont au mode indicatif – *peut, vais*.

Les marques de l'accord simple telles que *oui* ou *ouais* ne sont pas spécifiques des tours de parole adressés à la machine.

Les fournisseurs d'information

La machine et l'opératrice partagent le rôle discursif de *fournisseur d'information*. Les 30 formes les plus *spécifiques* de leurs tours de parole sont présentées en tableau 8.

Tableau 8 :

Productions des *fournisseurs d'information* (les 30 formes les plus spécifiques)

Produites par la machine (M)				*Produites par l'opératrice (O)*			
Forme	*Frq. Tot.*	*Fréquence*	*Coeff.*	*Forme*	*Frq. Tot.*	*Fréquence*	*Coeff.*
renseignements	283	270	***	alors	544	8	***
première	315	291	***	en	474	44	***
autres	194	185	***	ça	250	2	***
convient	259	251	***	hein	399	1	***
gare	759	649	***	non	320	12	***
classes	275	265	***	est	875	164	***
la	771	609	***	oui	714	40	***
ce	490	413	***	mais	220	2	***
minutes	601	593	***	je	407	34	***
de	1574	1032	***	ai	259	15	-44
désirez	429	421	***	E	180	1	-44
deuxième	382	325	***	ben	156	1	-38
quittez	648	489	***	bon	144	1	-35
part	346	282	+47	qui	247	22	-34
corail	172	155	+38	A	191	10	-34
ne	747	499	+37	allô	119	1	-29
train	609	416	+35	au	154	8	-28
phrase	93	93	+34	donc	109	1	-26
vous	2016	1145	+33	Y	163	12	-26

obtenir	85	85	+31	avez	139	7	-25
quels	84	84	+31	les	324	57	-24
votre	207	170	+30	pour	289	49	-23
paris	517	351	+29	là	203	28	-21
a	1882	1042	+25	que	297	59	-19
arrive	431	292	+24	des	137	15	-17
ouvez	182	145	+23	tard	92	6	-16
énoncez	60	60	+22	plus	173	26	-16
formuler	57	57	+21	si	130	16	-15
autrement	27	27	+21	le	623	180	-15
plaît	237	172	+20	voulez	60	1	-14

Les déictiques personnels sont spécifiques des tours de parole des fournisseurs d'information. La machine emploie de façon spécifique le vocatif *vous*, alors que les tours de parole de la seconde sont caractérisés par l'emploi du pronom *je*. Le *vous* de politesse est particulièrement spécifique des productions de la machine.

Le temps verbal dominant dans les deux cas est le *présent*, toujours au mode indicatif chez l'opératrice et parfois à l'impératif pour la machine. Les verbes d'action sont spécifiques des tours de parole de la machine - *obtenir, énoncez, formuler* –, tandis que l'opératrice privilégie l'utilisation des adverbes *oui* et *non*, ainsi que les articulateurs du discours *alors* et *donc*.

4.2 Routines conversationnelles

Au-delà des formes spécifiques employées par chacun des fournisseurs, on remarque que ces derniers utilisent de manière préférentielle un grand nombre de *routines conversationnelles*. Etant donnée la situation d'interaction, les fournisseurs d'information, véritable interface entre l'entreprise de services et les usagers, sont fortement soumis à la norme sociale : ils représentent l'entreprise et la qualité de leur travail est évaluée à partir du respect de ces normes, lorsqu'ils fournissent des informations aux usagers. Cependant, leurs routines sont différentes : leur mise en œuvre de pratiques socialement normées diverge.

Routines machine : la densité d'information

Le tableau 9 comporte des exemples[75] de tours de parole produits par la machine, qui correspondent à de telles routines conversationnelles. On propose un type pour chaque routine. La colonne gauche contient les différents exemples, tandis que la colonne droite comporte nos propositions de typage pour chaque cas de routine.

Tableau 9 :

Exemples de tours de parole produits par la machine M, proposition de typologie des routines

Exemple de routine	*Type de la routine*
`<TdP=MC0174> sncf **bonjour**, quels renseignements désirez-vous obtenir ?`	Routine d'ouverture

[75] Nous nous attachons à montrer des tours de parole caractéristiques des routines de l'un et de l'autre, pour affiner la comparaison des interactions. Nous utilisons à cette fin la carte des sections comme trame d'exploration. Nous y projetons des formes et des segments répétés, caractéristiques des routines de l'opératrice et de la machine.

<TdP=MC0175> quel **jour** désirez-vous partir ?	Routine pour l'obtention du jour
<TdP=MC0176> vers quelle **heure** désirez-vous partir ?	Routine pour l'obtention de l'heure
<TdP=MC0179> désirez-vous d'autres **renseignements** ?	Routine de pré-clôture
<TdP=MC1203> est-ce qu'il s'agit d'une **question** si oui est-ce que vous pouvez exprimer cette question de manière plus précise s'il vous plaît ?	Routine de reformulation
*<TdP=MC1487> pouvez-vous **formuler** votre phrase autrement il vous plaît ?*	Routine de reformulation
<TdP=MC1403> le premier train après douze heures zéro minute est un autorail première et deuxième classes qui part de la gare d'auxerre-saint-gervais à quinze heures quarante-deux minutes, arrive à la gare de laroche-migenne à seize heures zéro quatre minutes là vous devez changer et prendre un express **première et deuxième classes** qui part de la gare de laroche-migenne à seize heures vingt minutes arrive à paris gare de lyon à dix-huit heures trente-trois minutes, ce train vous convient-il ?	Message à caractère informatif
<TdP=MC1244> le dernier train que vous pouvez prendre est un corail **première et deuxième classes**, partant de la gare d'amboise à dix-neuf heures quinze minutes arrivant à la gare de paris-austerlitz à vingt et une heures quarante-trois minutes, ce renseignement vous satisfait-il ?	Message à caractère informatif
<TdP=MC0177> **ne quittez pas**	Routine de clôture

On remarque que la machine utilise le segment *ne quittez pas* en guise de clôture conversationnelle, au lieu d'employer *au revoir*, comme le fait l'opératrice (tableau 10).

Le *système intelligent* sous-jacent, qui gère la production des routines de la machine, est construit pour :

- reproduire des séquences d'ouverture et de clôture de la conversation[76] ;
- amener le correspondant à préciser sa demande ;
- délivrer la réponse à la demande du correspondant sous la forme d'un *message à caractère informatif*.

Routines opératrice : respect des normes conversationnelles

Le tableau 10 donne des exemples de routines produits par l'opératrice, dont certains correspondent à des routines conversationnelles. On propose pour chaque cas le type de la routine.

Tableau 10 :

Exemples de tours de parole produits par l'opératrice O, proposition de typologie des routines

Exemple de routine	*Type de la routine*
<TdP=OC0061> **ne quittez pas**	Routine de mise en attente
<TdP=OC2067> sncf **bonjour**	Routine d'ouverture

[76] Les routines de clôture de la conversation sont non pertinentes dans la situation d'interaction, la machine produisant systématiquement le tour de parole *ne quittez pas*. Un tel procédé pour clôturer une conversation n'est pas *régulier*, c'est-à-dire qu'il n'y a pas, dans les routines de la machine, de formule de politesse telle qu'*au revoir* par exemple. C'est surtout la densité d'information, notamment liée à la longueur des tours de parole, qui caractérise les énoncés de la machine.

<TdP=OC1158> non celui-ci est spécial, ah la la **je** peux c'est c'est un peu c'est un peu oui c'est difficile parce que moi j'ai des, j'ai des mois j'ai des de tel mois à tel mois ça circule mais e, je peux pas prendre en compte	Message à caractère digressif
<TdP=OC1157> oui oui c'est le tgv, mais apparemment **j'en** ai pas j'ai pas autre chose que des tgv	Message à caractère digressif
<TdP=CO2675> **je** l'ai fait une fois oui c'est un vrai pèlerinage	Message à caractère digressif
<TdP=OC1190> m, m bon ben je vais regarder hein **ne quittez pas**	Routine de mise en attente
<TdP=OC0071> - c'est bien pour un vendredi **hein**	Demande de confirmation
<TdP=CO0823> **au revoir** merci	Routine de clôture

Trois types de routines conversationnelles sont systématiquement présents dans les productions de l'opératrice :

- la routine d'ouverture de la conversation ;
- la routine de mise en attente de l'usager ;
- la routine de clôture de la conversation.

Les messages à caractère digressif constituent une grande part des tours de parole de l'opératrice. L'information délivrée au correspondant est diluée dans ses productions. On identifie par ailleurs une conduite interactionnelle centrée sur une application des normes conversationnelles dans les routines de l'opératrice.

5 Ajustements conversationnels de l'usager

La comparaison de l'adaptation du correspondant à un interlocuteur machine d'une part, à un interlocuteur humain d'autre part, est l'axe d'étude majeur de ce travail. Nous avons utilisé différentes procédures d'analyse textométrique pour étudier ce phénomène. En particulier, nous avons mobilisé :

- l'analyse de l'accroissement du vocabulaire comparée pour les quatre *types de tours de parole* du corpus, selon le double critère *foyer énonciatif/cible de l'énonciation* ;
- la ventilation des *pronoms personnels* et des *segments répétés* spécifiques des *tours de parole typés* ;
- la projection des *tours de parole typés* sur la carte des sections pour illustrer les phénomènes d'ajustement du correspondant, dont nous donnons des exemples.

Plusieurs indices corroborent l'idée d'un ajustement conversationnel du correspondant selon qu'il interagit avec un interlocuteur humain ou machine. Ces indices s'observent aussi bien à un niveau local (le tour de parole – spécificités / SR), qu'à un niveau global (proximités linguistiques et de vocabulaire – AFC / Accroissement du Vocabulaire) et font écho aux observations sur la dynamique conversationnelle (négociation des tours de parole).

Nous avons montré, dans un premier temps, que la négociation de l'allocation des tours de parole est (quasi) absente des interactions de l'usager avec la machine. Les différentes analyses sur les *tours de parole typés* et les *segments répétés* produits par les interactants confirment ces différences dans l'*ajustement conversationnel* du correspondant. En effet, confronté à une machine, le *demandeur d'information* humain manifeste une tendance à la réduction de son propre vocabulaire, minimise la complexité de ses productions et la longueur de l'échange. Il va même jusqu'à moduler sa production de clôtures conversationnelles sur celles de la machine.

Nous avons vu plus haut que les productions de l'opératrice sont caractérisées par ce qui constitue, du strict point de vue de l'échange d'information, des échanges digressifs. Les

routines conversationnelles produites sont régulières, étant donnée la formalisation des séquences conversationnelles et la relation de *pertinence conditionnelle* entre les tours de parole systématisées dans le *Turn-Taking System*, notamment concernant les clôtures.

La *négociation des tours de parole* est plus longue dans les interactions humain-humain, qui comportent de nombreuses phases de recouvrement de parole dues à des autosélections plus fréquentes. La dynamique conversationnelle semble plus fluide avec la machine ; en tous cas, la dynamique de l'allocation des tours de parole est plus régulière, presque mécanique. On a par ailleurs noté la faible part de *phatiques* produits par le correspondant en interaction avec la machine. Les phatiques étant caractéristiques de l'oral spontané, nous avons donc des premiers éléments tangibles pour soutenir l'idée d'un *ajustement conversationnel* du *demandeur d'information*.

6 Conclusions – Perspectives

Ce travail nous a permis de montrer l'utilisation des procédures d'analyses textométriques et du cadre méthodologique de l'analyse conversationnelle, pour la description de corpus d'interactions entre différents intervenants. Des stratégies de partition du corpus nous ont permis de gagner en puissance d'analyse, notamment en introduisant des types homogènes de tours de parole produits par les différents interactant en présence. L'approche pluridisciplinaire mobilisée dans ce travail a permis d'identifier et d'analyser des indices de l'ajustement conversationnel de l'appelant humain, aussi bien à des niveaux de description linguistiques locaux - emploi du vocabulaire, spécificités, segments répétés - que globaux - typologies et routines conversationnelles, régularités/irrégularités conversationnelles dans la production de séquences de clôture.

Nous avons pu vérifier sur notre corpus que la situation d'interaction induit un certain nombre de conduites socialement normées et contraintes par une polarité plus ou moins formelle. Ces conduites sont liées aux différentes fonctions assurées par chaque intervenant au fil de la conversation et fondent les rôles conversationnels.

L'analyse textométrique peut-être utilisée pour effectuer des comparaisons à des niveaux de granularité variables, permettant de ne pas dissocier dans l'analyse les dimensions *locale* et *globale* du corpus.

7 Références

Ech Chafaï, N., Ochs, M., Peters, C., Mancini, M., Bevacqua, E., Pelachaud C., (2007) *Des agents virtuels sociaux et émotionnels pour l 'interaction humain-machine*, in *Actes de la 19ème conférence francophone sur l'interaction humain-machine (IHM'07)*, pp. 207-214

Lebart, L., Salem, A., (1994) *Statistique Textuelle*, 342 p., Paris : Dunod, 1994

Portes, C., Bertrand, R., (2005) *De la valeur interactionnelle du « contour intonatif » en français. Résultats préliminaires*, Travaux interdisciplinaires du Laboratoire Parole et Langage, vol. 24, pp. 139-157

Sacks, H., Schegloff, E. A., Jefferson G., (1974) A simplest systematics for the organisation of turn-taking for conversation, in *Language, 50*, pp. 696-735

Schriberg, E., (1994) *Preliminaries to a theory of speech disfluencies,* Ph.D. thesis, University of Berkeley, California

Textométrie hiéroglyphique

[Conte du naufragé]

André Salem, Romuald Schummer
salem@msh-paris.fr, schummer2001@yahoo.fr

They did not know it was impossible, so they did it !
Mark Twain[77]

Résumé : A partir d'un texte hiéroglyphique et de ses translittérations sur un support informatisé, les méthodes textométriques permettent d'explorer directement des récurrences textuelles contenues dans le corpus. Le repérage de séquences répétées dans le texte original ouvre une voie textométrique à l'étude des procédés narratifs à l'œuvre dans le récit. La constitution d'un bitexte constitué du texte original et de sa traduction française alignée au niveau du verset permet d'étudier l'activité de traduction réalisée à partir des textes originaux.

Mots clés : textométrie, hiéroglyphes

L'activité d'*exploration* recèle bien des dangers pour ceux qui s'aventurent sans préparation dans des contrées qu'ils n'ont pas pris le temps de connaître, au moins par les récits de gens qui en sont revenus sains et saufs. En abordant l'exploration textométrique de textes fixés sur parchemin il y a plusieurs millénaires, après avoir connu une existence que l'on peut supposer aussi longue sous forme de poèmes transmis oralement de générations en générations, nous avons pleinement conscience de ne pas avoir préparé notre voyage avec autant de soin qu'il aurait été utile de le faire.

D'un autre côté, nous disposons aujourd'hui d'un corps de méthodes et d'outils textométriques éprouvés sur de très nombreux textes, écrits dans des langues extrêmement diverses. Ces méthodes ont montré qu'en s'appuyant sur la forme matérielle du texte et en y projetant un éclairage quantitatif, il était possible d'y repérer de *faits textuels* de répartition ou de répétition que les spécialistes formés aux sciences humaines, plus naturellement enclins lors de leurs lectures cursives à en extraire ce qui fait sens pour eux, en s'appuyant sur l'érudition acquise à leur contact, risquaient de négliger.

L'intuition textométrique souffle que cet éclairage devrait également prouver son efficacité sur les séquences de caractères hiéroglyphiques[78] que les systèmes informatiques modernes permettent désormais de gérer.

1 Le contexte de la recherche

Dans ce qui suit, notre projet sera double. Nous aimerions, en premier lieu, attirer l'attention des différents spécialistes de l'étude des textes hiéroglyphiques sur l'efficacité des méthodes

[77] Citation placée en exergue sur le site du *Projet Rosette* (http://projetrosette.info/) sur lequel nous avons recueilli l'essentiel des ressources informatisées qui nous ont permis de réaliser cette étude.

[78] Du grec ἱερογλύφος / *hieroglúphos*, composé de ἱερός / *hierós* sacré et γλύφειν / *glúphein* graver.

textométriques et sur les possibilités d'investigation nouvelles qu'elles ouvrent aux chercheurs dans le domaine des études égyptologiques. Par ailleurs, il nous semble que cette première application de méthodes textométriques, souvent éprouvées sur des corpus de textes rédigés dans des langues modernes, à des textes qui relèvent d'un système d'écriture très différent peut permettre du même coup à la communauté des études textométriques de prendre un recul utile par rapport au corps de méthodes qu'elle met régulièrement en œuvre sur les corpus de texte qui retienne son atention.

2 Le système d'écriture hiéroglyphique[79]

Les textes hiéroglyphiques sont en fait composés de phrases regroupant des mots écrits à l'aide de signes-images. Il n'y a pas de ponctuation et, comme c'est le cas pour la plupart des systèmes d'écriture de l'Antiquité, les mots ne sont pas séparés par des espaces. L'ordre dans lequel le texte doit être lu varie d'une inscription à l'autre (gauche-droite, droite-gauche, haut-bas, parcours boustrophédon, etc.).

Le système d'écriture hiéroglyphique permet et encourage même, à des fins esthétiques, des modifications de la séquence linéaire du texte. Les signes sont dessinés à l'intérieur d'un carré imaginaire qu'on appelle *cadrat*. Il sont parfois regroupés en un empilement méthodique, certains signes pouvant être associés ou superposés par rapport à d'autres.

2.1 Classification des hiéroglyphes par leur fonction

On peut classer les signes en trois classes principales :

- *idéogrammes :* certains signes sont utilisés pour coder le nom de l'être, de l'objet ou de l'action qu'ils représentent. L'image d'un taureau **Ä** permet la référence à cet animal, celle d'un plan de maison **O** est utilisée pour signifier *maison*. L'image d'une voile gonflée par le vent ¶ est utilisée pour faire référence au *vent*.
- phonogrammes : *d'autres signes sont principalement utilisés pour représenter un son. L'image d'un serpent* **œ**, *correspond plus ou moins au groupe phonique « dj », celle d'une bouche* **'** *que l'on prononce* « *er* » *sert à représenter la lettre « r », etc.*
- déterminatifs : *pour réduire le nombre des ambiguïtés dues à l'homonymie, on utilise des déterminatifs placés en fin de mot qui ne se prononcent pas. Ainsi, dans cette fonction, l'homme assis* **!** *détermine la séquence qui précède comme : occupations masculines, noms propres, etc.*

Notons qu'un même signe peut avoir des fonctions différentes en fonction du contexte dans lequel il est utilisé.

2.2 Translittérations modernes

En 1927, un siècle après la classification de Champollion, Gardiner propose une classification portant sur les quelques 740 hiéroglyphes, les plus courants. Chacune des 26 catégories de cette classification est symbolisée par une lettre. A l'intérieur de chaque catégorie les hiéroglyphes sont numérotés à partir de 1. Le code A1 correspond, par exemple, au signe **!** (homme assis), le code A2, au signe **#** (homme assis portant la main à la bouche), etc.

Pour les translittérations modernes, on utilise de plus en plus les prescriptions du *Manuel de codage* (dorénavant MdC) adoptées en 1988 par une grande partie de la communauté des

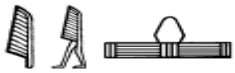
[79] Pour cette présentation des grandes lignes du système d'écriture hiéroglyphique, nous avons utilisé l'ouvrage publié par le ministère français de la culture à l'occasion de l'exposition *Naissance de l'écriture, cunéiforme et hiéroglyphes* - Galeries nationales du Grand Palais, Éditions de la réunion des musées nationaux, Paris, 1982.

égyptologues, qui permettent de transcrire les textes hiéroglyphiques en utilisant à la fois les codes de Gardiner et les translittérations de certains phonogrammes les plus courants.[80]

2.3 Codage informatique des écrits hiéroglyphiques

Le codage informatique moderne s'appuie notamment sur ces dernières méthodes de translittération pour stocker les textes initialement composés sous forme hiéroglyphique. A cette translittération vient souvent d'ajouter un découpage en *mots*. Chaque séquence reconnue comme un mot est précédée par un blanc et/ou caractère informatique particulier, les différents morphèmes grammaticaux étant systématiquement isolés par d'autres caractères[81].

Ainsi, la séquence de signes :



dont le codage dans la liste Gardiner est : **M17 M18 R4** sera notée, dans ce système de codage, à partir de ses valeurs phonétiques : **i ii Htp**.

Dans les transcriptions que nous avons utilisées, les codes « : » et « * » permettent respectivement de transcrire la superposition et la juxtaposition de deux signes. Le groupe de signes :

sera codé : **p*t:pt**, d'après ses valeurs phonétiques ou : **Q3*X1:N1** d'après les codes de la liste de Gardiner (association des signes Q3 et X1 dessinée au-dessus du signe N1).

2.4 Transcriptions, translittérations, traductions

Partant d'un texte hiéroglyphique, ont peut *générer*, en utilisant dans chaque cas des règles dont le degré de formalisation varie selon l'objectif fixé, d'autres *textes* qui permettront à des individus moins versés dans la lecture hiéroglyphique de mieux saisir tel ou tel aspect de la signification ou de la prononciation du texte :

- *une translittération :* substitue à chaque **graphème** d'un système d'écriture un **graphème** ou un groupe de graphèmes d'un autre système, indépendamment de la prononciation. Si les règles de translittération sont explicites et réversibles, il est possible de reconstituer le texte original à partir du résultat de la translittération.
- *une transcription :* substitue à chaque **phonème** d'une langue un **graphème** ou un groupe de graphèmes d'un système d'écriture.
- *une traduction :* tente de restituer dans une autre langue le **sens** contenu dans le texte original. Dans la pratique, les traducteurs choisissent entre plusieurs options dont certaines visent à rester au plus près du texte original pour le *trahir* le moins possible, alors que d'autres prennent, au contraire, le parti de placer la traduction dans un cadre socio culturel familier au lecteur, afin de faciliter au maximum sa perception du texte original.

Comme on le comprend, les *translittérations* et les *transcriptions* peuvent posséder, sous certaines conditions, la propriété de **réversibilité**. Tel est le cas, par exemple, si chaque état du texte est accompagné des règles de translittération qui, associées à ce texte, permettent de

---

[80] Cf. *Manuel de codage des données pour textes hiéroglyphiques sur ordinateur*, *consultable par exemple sur le site : http://projetrosette.info/page.php?Id=205.*

[81] On trouvera, au tableau 2, l'exemple d'un texte hiéroglyphique muni de sa codification dans un codage de ce type.

reconstituer l'état original. Dans ce cas, on peut grosso modo considérer, au plan textométrique, chacune des translittérations obtenues comme des ressources équivalentes au texte original. Comme on le conçoit aisément, cette propriété est rarement associée aux traductions effectuées d'une langue à une autre. Les traductions ne suffisent pas, dans le cas général, à reconstituer de manière univoque le texte original.

2.5 Segmentation en mots

Comme nous l'avons signalé plus haut, la tradition d'écriture hiéroglyphique ne sépare pas systématiquement par des blancs les différents mots qu'un lecteur égyptologue peut identifier dans le texte. Pour venir à bout de cette tâche, il est possible de s'appuyer sur le repérage de certains signes (ex : les déterminatifs) qui apparaissent prioritairement en fin de mot. Cependant, les spécialistes s'accordent sur le fait qu'une solide connaissance de la langue est nécessaire pour découper un texte hiéroglyphique en mots[82].

2.6 Ressources hiéroglyphiques en ligne

Un certain nombre de translittérations, et tout particulièrement celles qui permettent de redessiner les signes hiéroglyphiques originaux à partir des translittérations de type Gardiner, peuvent être confiées à des procédures informatiques. L'utilisation de telles procédures permet du même coup de vérifier le bon encodage du texte translitteré et de garantir l'homogénéité de la translittération elle-même.

Plusieurs sites web proposent des procédures capables d'effectuer automatiquement cette opération[83]. A partir du texte translitteré, ces procédures restituent des images qui permettent de vérifier visuellement la conformité de la translittération réinterprétée au texte d'origine. Les procédures réunies sur le site du ***Projet Rosette*** permettent, de plus, de faire le lien, pour chaque signe hiéroglyphique, avec toute une série de renseignements de type dictionnairique qui concernent : ses variantes scripturales, sa prononciation, sa signification globale, ses différentes significations en contexte, etc.

Ces possibilités de transcriptions automatiques fiables permettent de considérer les corpus de textes hiéroglyphiques translittérés comme des bases de données textométriques susceptibles de servir de point de départ à des traitements textométriques dont les résultats pourront également être translittérés sous leur forme hiéroglyphique originale.

3 Le corpus *Naufragé*

Le ***Conte du Naufragé*** est l'un des textes importants de la littérature de l'Égypte ancienne parvenus jusqu'à nous. Des versions électroniques du texte hiéroglyphique original, composé de 190 versets, ainsi que des traductions, des transcriptions et des translittérations destinées à permettre la conservation de ce texte sur des supports informatisés peuvent être aisément localisés sur différents sites consacrés à l'égyptologie.[84] Le *Conte du naufragé* à donné lieu à

---

[82] De cette certitude partagée par les égyptologues, on peut inférer sans risque de se tromper qu'à l'instar de ce qui se passe pour les textes écrits en d'autres langues, tout découpage d'un texte hiéroglyphique en mots et a fortiori toute tentative de rattacher systématiquement chacun des mots découpés dans la chaîne textuelle à des unités dictionnairiques plus génériques (lemmatisation) sera susceptible de prêter le flan à des critiques qui feront valoir des interprétations du texte ou des arguments de grammairiens conduisant à des découpages et ou à des regroupements différents.

[83] Pour cette étude, nous avons eu recours à l'ensemble des procédures réunies sur le site du Projet Rosette : http://www.projetrosette.info.

[84] La version électronique du texte hiéroglyphique du *Conte du naufragé* que nous avons utilisée pour cette étude a été téléchargée à partir du site du Projet Rosette.

de nombreuses études de caractère littéraire portant essentiellement sur la structure extrêmement remarquable du récit[85].

==== Le conte du naufragé ====

Le papyrus : La seule version de ce conte qui nous soit parvenue est consignée sur un papyrus hiératique[86]. Le document a été découvert dans les réserves du Musée de l'Ermitage, à Saint-Pétersbourg à la fin du 19ème siècle de notre ère. Les historiens qui ont pu faire des rapprochements avec d'autres textes fixés sur papyrus à la même époque pensent que le document a été établi il y a environ 4 000 ans.

Il n'est pas possible d'estimer avec précision la date de la création du récit lui-même. Bien avant sa fixation sous forme écrite, ce texte z pu circuler sous forme d'un récit poétique transmis oralement, sans altération majeure, de générations en générations pendant une très longue période. Le texte peut avoir été traduit ou fortement inspiré par un texte préexistant transmis oralement ou fixé sur un document rédigé dans une autre langue.

L'histoire : Pour rassurer un jeune supérieur, inquiet d'avoir à rencontrer prochainement son suzerain, un vieux serviteur lui raconte qu'embarqué sur un navire il a été victime d'un naufrage qui l'a fait échouer sur une île habitée par un serpent géant. Sa frayeur dissipée, il a raconté son histoire au serpent. Puis le naufragé a écouté l'histoire du serpent, lui même victime de malheurs qui ont abouti à la destruction de sa propre famille, lors d'une période précédente. A l'issue de cette rencontre, le serpent a couvert le naufragé de présents et lui a prédit qu'il vivrait heureux parmi les siens. Le jeune supérieur écoute avec attention ce récit qui ne dissipe cependant pas ses propres craintes.

La critique : Plusieurs critiques modernes ont souligné la composition originale de ce récit. Plusieurs conteurs y enchâssent à tour de rôle des récits personnels ainsi que des commentaires sur les faits qu'ils relatent. On note des symétries dans la manière dont sont agencées les différentes parties du conte. A la description du voyage d'aller correspond celle d'un retour, aux frayeurs initiales, des surprises agréables, etc.

---

[85] Cf., par exemple, D. Benoît, Le conte du naufragé dans le cycle : Les grands textes de l'Égypte ancienne. http://www.thotscribe.net/docs/2004_2005/conte_naufrage.pdf.

[86] L'écriture *hiératique* constitue une forme simplifiée de l'écriture hiéroglyphique permettant d'écrire plus rapidement.

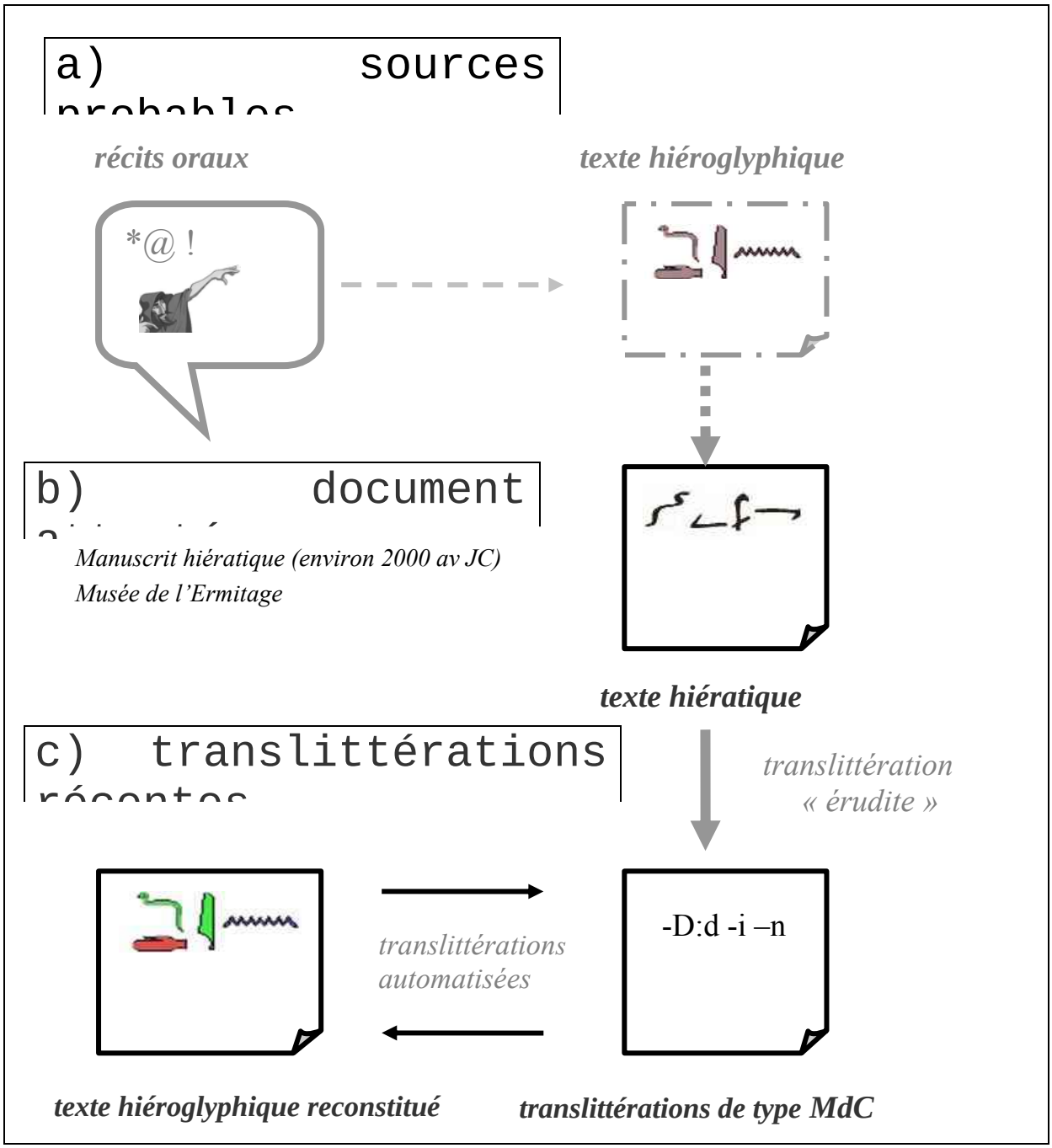


Figure : 1

Le conte du naufragé *:*

*s*ources probables, documents attestés et translittérations modernes

On a rassemblé, sur la figure 1, différents états du récit qui ne nous est parvenu que sous forme d'un papyrus hiératique (section b). Les états antérieurs de ce récit, dont l'existence est probable, sont mentionnés en gris (section a). La dernière section (section c) regroupe les versions informatisées du texte sur lesquelles nous avons pu travailler effectivement.

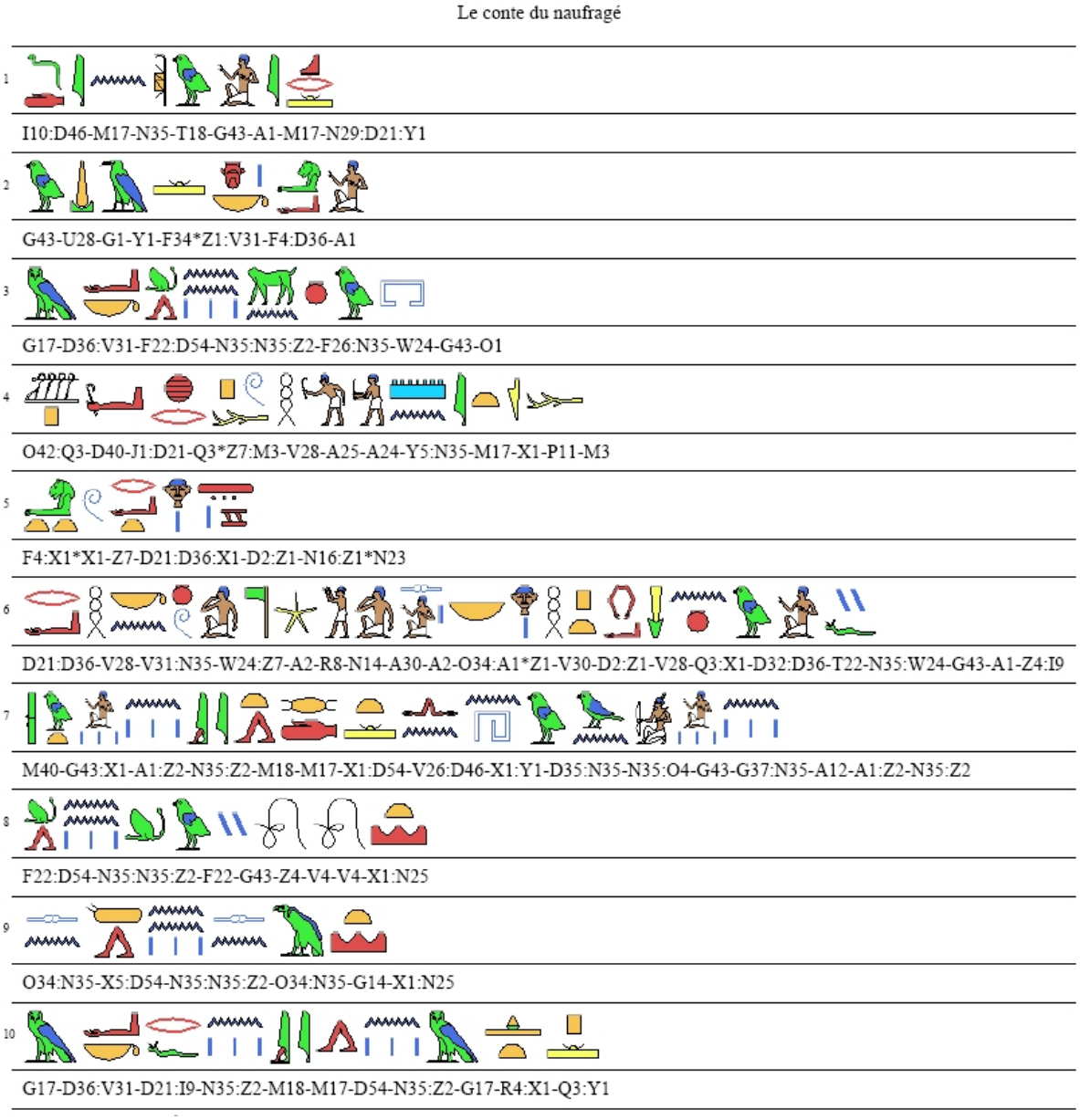


Figure 2

Le début du *Conte du naufragé*
suivi de sa transcription en codes Gardiner

Dans l'état actuel des logiciels de traitement textométrique, il n'est pas envisageable d'appliquer directement les procédures textométriques destinées à permettre des mesures sur le texte aux séquences de signes hiéroglyphiques. Cependant, comme nous l'avons signalé plus haut, les translittérations obtenues à partir des codages de type Gardiner constituent des équivalents extrêmement fidèles de leur source hiéroglyphique.

Dans ce qui suit, nous tenterons de montrer comment des mesures textométriques portant en fait sur les séquences originales du texte hiéroglyphique peuvent être réalisées à partir des translittérations de type Gardiner archivées sur support informatique.

4 Approches textométriques du corpus *Naufragé*

Pour soumettre un texte à des traitements textométriques, il est nécessaire de déterminer deux systèmes complémentaires : un système de *contenants*, parties du texte qui vont être soumises à des comparaisons textométriques et un système de *contenus,* unités textuelles (habituellement : *mots, graphèmes*, etc.) dont on s'attachera ensuite à recenser les occurrences au sein de chacune des parties du texte.

A partir du décompte des occurrences des unités-contenus à l'intérieur des contenants, les méthodes textométriques produisent des jugements quantitatifs qui peuvent ensuite être interprétés en terme de variations dans l'usage du vocabulaire.

Nous avons jugé utile, dans ce qui suit, de faire figurer, en regard des calculs effectués à partir du texte hiéroglyphique, des calculs similaires réalisés à partir de la traduction française du *Conte du naufragé*. On peut voir, sur le tableau 2, un extrait de chacune des deux versions du texte qui constituent ensemble ce que l'on appelle un *corpus aligné multilingue*. L'alignement a été réalisé, ici, au niveau du verset. A côté des calculs que l'on peut effectuer à partir de chacun des volets pris isolément, les investigations multilingues permettent d'effectuer des rapprochements entre fragments du corpus aligné et de mieux analyser l'activité de traduction effectuée entre les deux versions du texte.

4.1 Découpages du corpus

Le document original se présente sous forme d'un texte découpé en 190 lignes, que nous appellerons ici des ***versets***. Une même phrase du texte (un même mot ?) peut se retrouver transcrite à cheval sur la fin d'un verset et sur le début du verset suivant. Nous avons numéroté les versets de 1 à 190 en faisant précéder le verset x de la balise <v=x>.

Tableau : 1
Partition en douze fragments du corpus ***Naufragé***

Partie	occurrences	formes	hapax	fmax	forme
01Intro	313	78	35	31	n
02VoyageEtNaufrage	277	81	37	21	n
03IleDuKa	251	73	40	28	n
04LeSerpent	434	90	39	39	n
05RecitNaufAuSerp	374	91	40	29	n
06DiscduSerpent1	224	61	28	23	n
07RecitduSerpent	270	73	37	31	n
08NaufetSerpent	354	91	41	30	n
09DiscDuSerpent2	597	115	60	60	n
10Retour	214	59	21	31	n
11Epilogue	153	56	29	13	n

Dans le document original, certains groupes de versets sont précédés d'une courte séquence de signes mise en valeur par une coloration rouge qui semble marquer le début d'une nouvelle

partie du récit et suggérer un découpage du texte en parties[87]. Ce découpage provisoire, dont il faut noter que nous ignorons l'origine exacte, ne constituera pas pour nous une donnée indépassable. Nous nous appuierons cependant sur ce découpage pour effectuer une première comparaison à partir des différents fragments du texte.

Nous avons transcrit cette division qui aboutit à une partition du texte en douze fragments par des balises de type <D=y> où *y* varie de 1 à 12. Le tableau 3 fournit les principales caractéristiques lexicométriques calculées pour chacun des fragments.

On trouve au tableau 3 un état qui présente le début de chacun des deux volets du corpus munis des balises qui permettent de distinguer les versets et les regroupements thématiques.

4.2 Les unités de décompte

La question de la détermination des unités les plus aptes à servir de base aux décomptes textométriques a longtemps agité les communautés de chercheurs confrontées aux corpus textométriques[88]. Nous avons signalé que, dans le cas des corpus hiéroglyphiques, la détermination des frontières de mots constituait une tâche hors de portée pour les traitements automatisés. Nous consacrerons l'essentiel de cette première étude au repérage automatique des répétitions contenues dans le texte. Pour effectuer cette tâche, nous allons commencer par considérer le système des unités de décompte constitué par les différents signes hiéroglyphiques.

[87] Pour effectuer ce découpage, nous nous sommes efforcés de suivre les indications du manuscrit original qui ont donné lieu à l'insertion d'intertitres (rédigés par les éditeurs français du manuscrit) sur le site sur lequel nous avons récupéré le texte orignal.

[88] Sur ces questions on consultera, par exemple, [Muller 1963] et [Brunet 2000].

```
<D=01Intro>
<v=001> -D:d -i -n -Sms -w -A1 -i -q:r:Y1 -w -DA -A -Y1 §
<v=002> -ib*Z1:V31A -HAt:a -A1 -m -a:V31A -pH:D54 -n:n:Z2 §
<v=003> -Xn:n -nw -w -pr -Ssp:p -a -x:r -p*W:xt §
<v=004> -H -A25 -A24 -mn:n -i -t -P11 -xt -HAt:t*t -W -r:a:t §
<v=005> -Hr:Z1 -tA:Z1*N23 -r:a -H -V31A:n -nw:W -A2 -nTr -dwA §
<v=006> -A30 -A2 -z:A1*Z1 -nb -Hr:Z1 -H -p:t -D32:a -sn -n:nw -w -A1 -y:f §
<v=007> -iz -w:t -A1 -Z2 -t:n:Z2 -ii -i -t:D54 -aD:d -t:Y1 -D35:n §
<v=008> -n:h -w -wr:n -mSa -A1:Z2 -n:Z2 -pH:D54 -n:n:Z2 §
<v=009> -pH -w -y -wA -wA -t:xAst -z:n -X5:D54 -n:n:Z2 §
<v=010> -z:n -mwt -t:xAst -m -a:V31A -r:f -n:Z2 -ii -i -D54 -n:Z2 §
<v=011> -m -Htp:t -p:Y1 -tA:N23*Z1 -n:Z2 -pH:D54 -n:Z2 -sw -W §
```

```
<D=01Intro>
<v=001> un excellent suivant dit alors : apaise §
<v=002> ton coeur, prince ! vois, nous avons atteint §
<v=003> la résidence. le maillet est saisi et §
<v=004> le poteau d'amarrage est frappé, l'amarre de proue ayant été portée §
<v=005> à terre ; les prières sont dites, le dieu a été remercié §
<v=006> et chaque homme embrasse son semblable, §
<v=007> car notre équipage est revenu sain et sauf, sans §
<v=008> perte pour notre troupe. nous avons atteint §
<v=009> les confins de ouaouat, après avoir doublé §
<v=010> senmout. vois donc, nous revenons  §
<v=011> en paix, notre pays, nous l'avons atteint. §
```

Tableau 2

Le corpus multilingue aligné ***Naufragé***

a) le début du poème codé selon les normes MdC

b) la traduction française de cet extrait

Pour mettre en œuvre ce choix, il nous suffira de considérer, dans le cadre de cette première expérience, les signes d'association (*) et de superposition (:) comme des caractères isolant les différents signes réunis dans un même cadrat. Cette option s'appuie sur l'affirmation trouvée dans les travaux que nous avons pu consulter, que l'habitude de superposer et d'associer différents signes hiéroglyphiques dans un même cadrat prend souvent sa source dans des considérations d'ordre esthétique. Si cette hypothèse est vraie, on peut s'attendre à ce que les séquences de signes ayant donné lieu au regroupement graphique en un même cadrat composite soient traitées de la même manière aux différents endroits du texte dans lesquels elles apparaissent. Notons que la prise en compte du texte sur support informatisé nous permet de vérifier systématiquement cette hypothèse par l'utilisation de la méthode textométrique de base que constitue l'établissement de *concordances*.

4.3 Principales caractéristiques textométriques

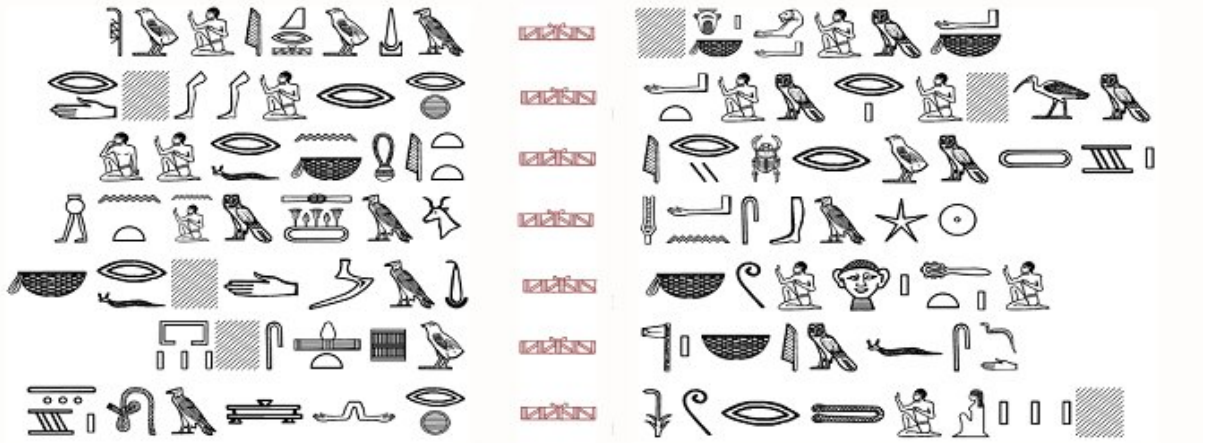
Le dépouillement des deux volets du corpus parallèle amène les caractéristiques lexicométriques que l'on trouve au tableau 3. Ces caractéristiques ne sont pas directement comparables car elles signalent avant tout des différences notables dans les systèmes d'écriture, compte tenu des normes de dépouillement que nous avons utilisées. Dans le cas du volet français du texte, la segmentation s'est faite sur des unités lexicales qui correspondent plus ou moins aux mots de la langue. Dans le cas du corpus hiéroglyphique, la segmentation a abouti à isoler des unités plus ténues qui entrent dans la composition des mots (lettres, phonèmes, morphèmes, déterminants). Les caractéristiques lexicométriques calculées sur chacun des volets du corpus portent la trace de cette importante différence. Les différents modes de segmentation retenus expliquent à eux seuls : d'une part le plus grand nombre d'occurrences et la fréquence maximale nettement plus élevée dans le volet hiéroglyphique, de l'autre, le plus grand nombre de formes et d'hapax dans la traduction française du texte.

Tableau : 3
Principales caractéristiques textométriques
pour les deux volets du corpus ***Naufragé***

	Hiéroglyphes	Français
Nombre d'occurrences	3 741	1 745
Nombre de formes	248	541
Nombre d'hapax	89	316
Fréquence maximale	336	77
forme	n	de

4.5 Concordance d'un signe

Lorsqu'on désire étudier la signification d'une unité textuelle dans l'ensemble d'un corpus ou examiner chacun de ses contextes particuliers d'utilisation, la possibilité de rassembler sur un même document toutes les occurrences d'une forme donnée, accompagnée d'un contexte minimal, constitue l'un des avantages les plus appréciables offerts par la prise en compte d'un corpus informatisé.



Signe

Code Gardiner : **Y1**
EGPZ : 58328 (e3d8)
GlyphBasic : 4-242
Transliteration : mDA.t / dmD / dmd

Signification
écriture, abstraction

Description
rouleau de papyrus scellé (var.Y2)

Commentaire :
- idéogramme dans mDAt 'rouleau de papyrus'
- déterminatif dans les termes liés à l'écriture ou aux notions abstraites

Figure : 3
Extrait d'une concordance réalisée à partir de la forme Y1 *écriture*
(les carrés gris signalent un changement de verset)

Comme on l'a souligné plus haut, dans le cas d'une translittération chacune des occurrences d'une même unité textuelle reçoit un codage identique. Dans notre cas, chacun des signes

hiéroglyphiques reçoit un code identique. Pour réaliser la concordance du signe  que l'on peut voir sur la figure 3, nous avons commencé par réaliser une concordance portant sur les occurrences de la forme Y1 dans le fichier translittéré. Les lignes de contexte générées par le module de concordance ont ensuite été soumises à l'éditeur Rosette[89] qui a rétabli leur forme hiéroglyphique originale.

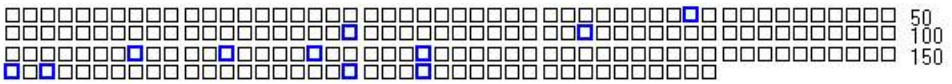
Les états ainsi obtenus permettent d'examiner sous forme *visuelle* l'ensemble des emplois d'une même unité de segmentation dans un corpus de textes hiéroglyphiques.

4.4 Explorations multilingues

Le fait de disposer d'une traduction alignée du texte que l'on étudie se révèle d'une grande utilité pour explorer un texte rédigé dans une langue que l'on ne domine pas. Les méthodes textométriques permettent d'établir des liens entre certaines des unités textuelles qui sont en rapport de traduction au sein d'un bitexte aligné.

Ainsi, par exemple, on peut constater que le terme *île* apparaît onze fois dans le volet français du corpus. Pour tenter de trouver des termes qui correspondent à ce terme dans le volet hiéroglyphique du corpus, on commence par sélectionner les versets qui contiennent la forme île dans le volet français (figure 4a).

4a : versets du volets français contenant la forme *île*



4b : sélection des versets du volet hiéroglyphique correspondant à des versets du volet français contenant la forme *île*

Figure : 4
Extraction de termes en rapport de traduction à partir d'un bitexte

On commence par repérer les sections du volet français dans lesquelles apparaît le terme *île*. Pour chacune de ces sections, on peut localiser, dans le volet hiéroglyphique, une section correspondante laquelle est susceptible de contenir un terme en rapport de traduction avec cette forme lexicale. Le calcul des spécificités (formes surreprésentées) dans la zone du volet hiéroglyphique ainsi mise en évidence nous indique que la séquence de signes -iw:N23*Z1 apparaît 11 fois dans le corpus. L'éditeur du site Rosette nous fournit la forme hiéroglyphique originale de cette translittération et nous informe que ce signe complexe se traduit bien en français par le nom commun *île*[90].

	N18:N23*Z1	iw	**nc** : île

[89] Le site Projet Rosette offre un éditeur *en ligne* qui traduit sous forme hiéroglyphique les séquences de signes translittérés qui lui sont fournies par le biais d'un interface web.

[90] Ce type de procédure a été analysé par Maria Zimina dans sa thèse, cf. [Zimina 2004]. Les versions actuelles de Lexico3 (à partir de la version 3.5.9) permettent d'interroger chacun des volets d'un corpus parallèle à partir d'un sélection effectuée sur l'autre volet.

4.6 L'accroissement du vocabulaire

La figure 5 montre la courbe du vocabulaire réalisée pour le volet hiéroglyphique du corpus ***Naufragé***. La partition du corpus en fragments a été matérialisée sur ce graphique par des lignes verticales qui marquent chacune le début d'un des douze fragments du corpus.

Certains fragments sont caractérisés par des portions presque horizontales de la courbe d'accroissement. Cette circonstance peut s'expliquer par le fait que ces fragments sont le siège de répétitions de signes hiéroglyphiques déjà utilisés dans des fragments précédents.

La seconde courbe rend compte de l'apparition des *hapax* (formes qui ne trouvent qu'une seule occurrence dans le corpus).[91] Dans les dépouillements textométriques pratiqués à partir du découpage du texte en *mots*, on a pu remarquer que, loin de constituer une exception, la propriété d'hapaxie est partagée par un très grand nombre de formes du texte. De ce fait, l'ensemble du texte se trouve parsemé de formes de fréquence 1 et tout fragment du texte en contient un certain nombre plus ou moins proportionnel à sa longueur. La surabondance de formes de fréquence 1 dans un fragment particulier constitue un souvent le signe que le fragment est le lieu de descriptions et d'énumérations de termes qui ne seront plus employés par la suite. A l'inverse, l'absence relative de ces formes est souvent le signe que le fragment contient des répétitions de segments de textes dupliqués dans le corpus.

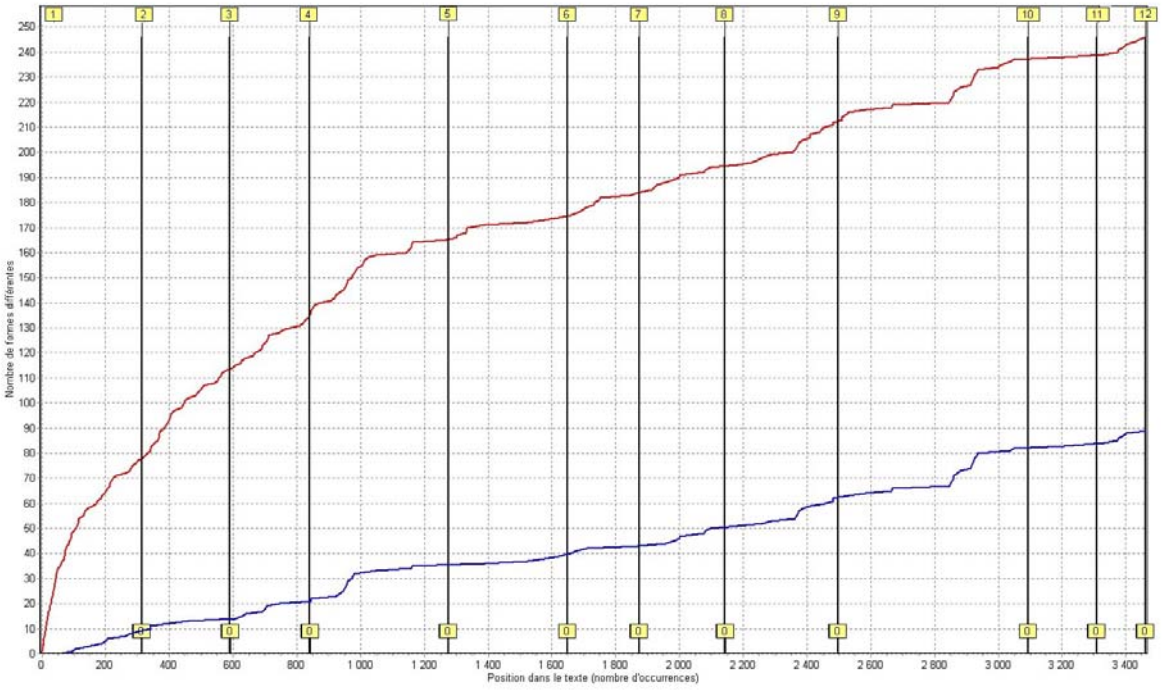


Figure : 5

[91] Dans la longue tradition des études critiques à propos des textes, le concept d'*hapax legomena* (chose dite une fois) a été élaboré pour signaler la propriété attachée à une unité textuelle de constituer un exemple unique d'utilisation dans un corpus donné. Dans la pratique, les copistes et les commentateurs ont souvent noté cette propriété, jugée exceptionnelle, à propos d'unités textuelles remarquables du point de vue de leur forme. Dès le début des études quantitatives appliquées aux textes et avant que les dépouillements textométriques ne soient systématiquement confiés à des ordinateurs, les textométriciens ont noté que le phénomène de l'hapaxie, loin de constituer une propriété exceptionnelle pour certaines formes rares, constituait au contraire un phénomène massif pour tout texte écrit dans une langue naturelle. Depuis la description de la structure quantitative du vocabulaire opérée par G. K. Zipf (cf. [Zipf, 1936]) on sait au contraire que dans la plupart des corpus de textes écrits en langue naturelle, la propriété de n'apparaître qu'une seule fois dans un corpus est partagée par un très grand nombre de formes du texte.

Courbe d'accroissement du vocabulaire et courbe d'accroissement du nombre des hapax calculées pour le volet hiéroglyphique du corpus ***Naufragé***

==== Guide de lecture pour la figure 5 ====

- Le nombre des occurrences du texte se développe le long de l'axe horizontal entre le début et la fin du texte pour lequel la courbe a été établie.
- La *Courbe d'accroissement du vocabulaire* (en rouge, dans la partie supérieure du graphique) s'accroît d'une unité chaque fois que l'on rencontre une forme qui n'a pas encore été rencontrée précédemment. C'est une courbe croissante qui varie de 0 (au début du texte) à NbForm (nombre de formes différentes du texte, valeur atteinte lorsque le texte a été entièrement parcouru).
- La *Courbe d'accroissement du nombre des hapax* (en bleu, dans la partie inférieure du graphique) résulte d'un calcul similaire pour lequel ne sont prises en compte que les formes hapax du texte considéré (i.e. les formes qui ne possèdent qu'une seule occurrence dans l'ensemble du corpus). Cette seconde courbe varie de 0 à NbHap (nombre total des hapax du texte).

Dans le cas du dépouillement en signes hiéroglyphiques que nous avons adopté pour cette étude, l'unité de décompte concerne des unités dont les combinaisons permettent ensuite de former les unités plus étendues que sont les mots. Ces unités peuvent parfois coïncider avec des mots, dans d'autres cas elles n'en constituent qu'un élément. Compromis entre un système basé sur un alphabet extrêmement réduit et un système dans lequel tous les signes auraient valeur d'idéogramme, le système d'écriture hiéroglyphique ne peut donc être totalement assimilé à un système lexical du point de vue de ses caractéristiques textométriques, ce dont témoignent d'ailleurs les décomptes produits au tableau 3.

Malgré ces différences, nous allons montrer que, la raréfaction des hapax constitue bien un signe de redondance du texte contenu dans le fragment par rapport à l'ensemble des fragments qui précèdent. Sur la figure 5, on peut vérifier que certaines portions du texte connaissent un accroissement faible du nombre des hapax (courbe d'apparition des hapax presque horizontale pour le fragment). La suite de notre étude nous permettra de vérifier que ces fragments constituent bien des reprises textuelles par rapport aux fragments précédemment rencontrés dans les parties précédentes du texte.

4.6 Étude des segments répétés du corpus

Les procédure de calcul des *segments répétés* permettent de localiser des suite de signes hiéroglyphiques apparaissant à l'identique à plusieurs endroits du corpus ***Naufragé***. Ainsi par exemple, la séquence de signes translittérés :

```
-A1 -r:f -n:V31A -mi -i -t*t:Y1 -i -r:y -xpr:r
```

peut être localisée à l'identique dans deux versets du texte (versets 21 et 125). L'éditeur du site *Projet Rosette* permet de rétablir la forme originale de cette séquence :

et de vérifier sa présence dans le texte original aux deux endroits indiqués[92]. On trouvera, figure 8, les traductions associées à cette séquence aux endroits du corpus qui la contiennent.

Classification et localisation des répétitions du corpus

Différents travaux consacrés à l'utilisation des recensements de segments répétés dans un corpus de textes montrent que les résultats fournis par ce type de formalisation renvoient la plupart du temps à des phénomènes textuels de niveaux très différents. Dans le cas des dépouillements en mots, les segments courts (i.e. composés de 2-3 formes) renvoient souvent à la présence d'unités lexicales complexes (mots composés, locutions, etc.) alors que la répétition de segments composés d'un plus grand nombre de formes trahit en général la présence de citations ou de reprises textuelles plus systématiques.[93]

L'analyse des segments répétés contenus dans chacun des volets du bitexte ***Naufragé*** fait apparaître toute une série de segments répétés particulièrement longs. L'établissement d'une concordance portant sur les segments les plus longs, tableau 5, nous permet de vérifier que plusieurs de ces segments trouvent une de leurs occurrences dans le fragment n°2 du conte à laquelle correspond une seconde occurrence qui peut être localisée dans le fragment n°5.

L'établissement d'une carte de sections sur laquelle on a signalé la présence des segments appartenant à ce seul groupe nous conduit au constat que la duplication de ces longues séquences résulte de la répétition d'un même récit, repris avec des variations à deux endroits différents du corpus (figure 4).

Tableau : 4

Extrait des concordances réalisées à partir des occurrences des segments répétés les plus longs dans le volet hiéroglyphique du corpus ***Naufragé***

```
Partie : 01Intro, Nombre de contextes : 1
- p : W - D : d - n : V31A - s - D : d - A1 - r : f ! - n : V31A - mi - i - t * t :
Partie : 02VoyageEtNaufrage, Nombre de contextes : 8
: V31A - w - A1 - r - M14 - wr : r - S - m - d : p * t - P1 ! - n : t - mH : a - V1
: mD - mD : mD - m - s - x ! - w - iab - s - s - qd - d - A30 - A1 - V1 - V20 : V20
V1 - V20 : V20 - i - m - s ! - m - stp : Y1 - n - km - m - t : niwt - mA : ir - A -
tA : N23 * Z1 - m - a : V31A - A - a ! - ib : Z1 - s - n : Z2 - r - mA : ir - A - w
 r - S - tp : Z1 - a : Z1 ! - sAH - Y1 - n : 3 - tA : N23 * Z1 - f - A - t - A9 - a
 m - i - i - t - A2 - n : U19 - nw - W - i - i - t - mw ! - i - m - f - n : t - mH
 n - xt : t * Z1 - H - H ! - A25 - A24 - n - A1 - s - aHa - a : n - d : p * t - P1
 a : n - d : p * t - P1 ! - m - t : Z6 - n : t - tyw - Z2 - i - m - s - D35 - z : p
Partie : 03IleDuKa, Nombre de contextes : 1
 - H - n : a - A - p : d - w - zA : Z2 - D35 : n - n : t * t ! - D35 : n - s - t - m
Partie : 04LeSerpent, Nombre de contextes : 6
 V12 : Y1 - sw - w - r - xnt - n : t ! - i - w - wp : p - Z9 : n : f - r * Z1 : f -
: n - A1 - n - m : a - ini - n : t * w - zp : Z1 * Z1 - n : D - z : wr - A1 ! - n -
: n - iTi : t * t - A24 - i - m - A1 ! - i - w - wp : p - Z9 : n : f - r * Z1 : f -
- w - A1 ! - Hr - Z1 - X : t * Z1 - A1 - m - b - bA - A - H - D53 : Y1 - f ! - aHa
- n : A1 - n - m : a - ini - n : t * W - zp - Z1 * Z1 ! - n : D : z - wr - A1 - n -
 - N36 - n : t * y - Aa13 : Z1 - f : y - m - n : U19 - nw - W ! - i - i - mw - aHa
Partie : 05RecitNaufAuSerp, Nombre de contextes : 9
 - i - i - A1 - x - xA - A - m - D41 ! - m - b - bA - A - H - D53 : Y1 - f - D : d
p : p - w - t : D54 ! - sAq : sAq - G7 - m - d : p * t - P1 - n : t ! - mH : a - V1
- mD : mD - m - s - x : w ! - iab : Y1 - s - s - qd - d - A30 - A1 - V1 - V20 : V20
 i - m - s ! - m - s - t : p - w - U21 : Y1 - n : km - m - t : niwt ! - mA : ir - A
 : N23 * Z1 ! - m - a : V31A - A - A24 - ib - Z1 - s - n : Z2 - r - mA : ir - A ! -
: N36 ! - tp - Z1 - a - Z1 - D61 - D54 - n : 3 - tA : N23 * Z1 - f - A - t - A9 - a
```

[92] Rappelons que l'identité que nous avons recherchée porte sur la *séquence* des signes élémentaires qui constituent la séquence hiéroglyphique. En l'occurrence, les deux versions de la séquence repérée présentent quelques écarts minimes qui peuvent concerner la disposition des signes sur la ligne.

[93] Sur la méthode des segments répétés, cf. par exemple [Salem 1994].

```
m - i - i - t - A2 - n : U19 - nw - w - i - i - t - mw ! - i - m - f - n : t - mH
- n - xt : t * Z1 - H - H - A19 - a ! - n : A1 - s - aHa - a : n - d : p * t - P1
n - d : p * t - P1 - m - t : Z6 : t ! - n : t - tyw - Z2 - i - m - s - D35 : z - p
```

Dans le cas de reprise textuelle d'un récit relativement long que nous venons d'explorer, on peut penser que l'existence d'une répétition n'aurait pas échappé à un lecteur attentif, pour peu que celui-ci soit suffisamment à l'aise avec la langue dans laquelle le texte a été rédigé. Une fois identifiées les zones de répétition, le repérage des unités textuelles qui n'apparaissent que dans l'un des deux fragments qui entrent en rapport de duplication peut alors permettre de localiser des variations entre les différentes versions du récit.

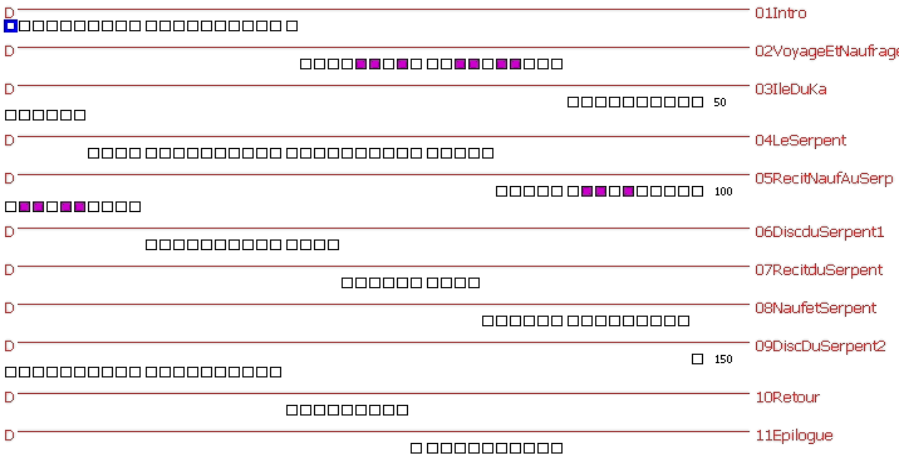
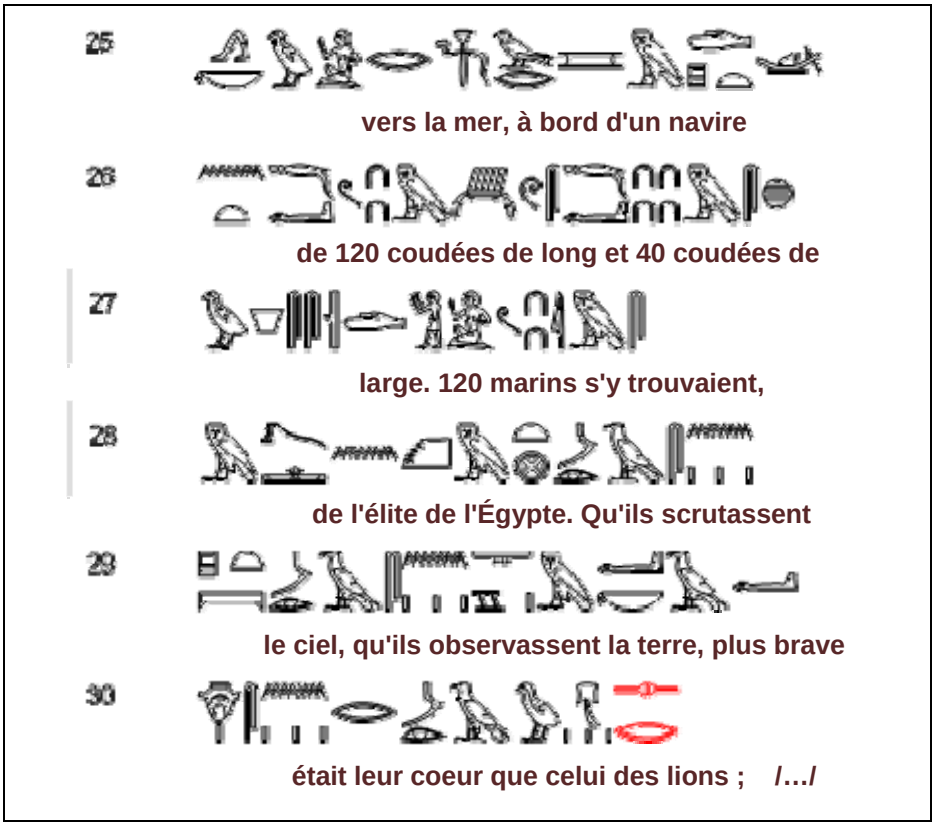


Figure : 6

Ventilation des occurrences des segments répétés longs trouvant dans les fragments 2 et 5 du volet hiéroglyphique du corpus ***Naufragé***



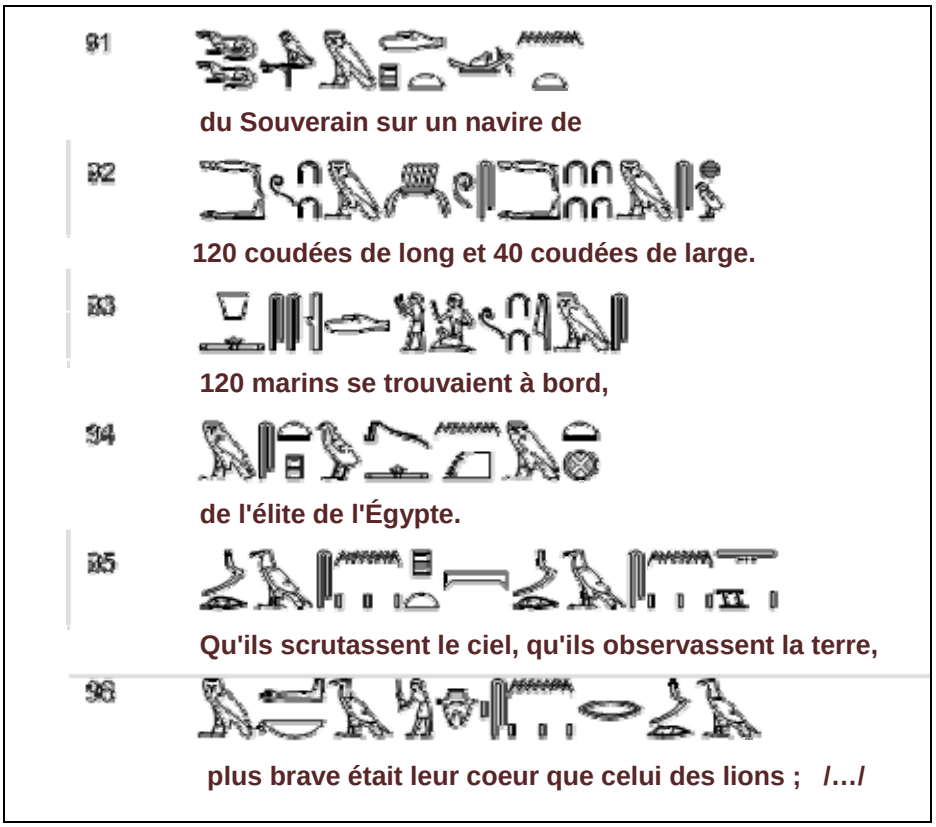


Figure : 7
Deux passages du corpus ***Naufragé*** rapprochés sur la base de leur utilisation de segments répétés communs**.**

La comparaison systématique entre les résultats fournis par la même méthode sur les deux volets du corpus multilingue peut permettre d'interroger utilement le travail du traducteur : a-t-il rendu par des formulations différentes des segments de texte absolument identiques dans le texte original ? a-t-il, au contraire traduit par les mêmes expressions des formulations qui différaient quelque peu dans ce même texte ?

récurrences isolées

La méthode des segments répétés permet également de repérer des récurrences moins systématiques dues à la reprise d'une formule particulière dont l'origine peut être trouvée soit dans l'existence d'un figement linguistique particulier soit au contraire dans la mise en pratique de procédés narratifs utilisés de manière récurrente. On voit par exemple sur la figure 8 le rapprochement que l'on peut opérer en suivant la même méthode entre les propos tenus par le vieux serviteur pour commencer le récit qu'il adresse à son supérieur et ceux prononcés par le Serpent pour commencer le sien.

125 

Laisse-moi donc te raconter quelque chose de semblable qui est arrivé sur cette île

Figure : 8
Fragments du corpus ***Naufragé*** rapprochés sur la base de leur utilisation de segments répétés communs**.**

Dans ce second cas, la méthode textométrique apporte incontestablement un éclairage qui permet seul de localiser des répétitions segmentales importantes pour l'étude de la construction du récit, dans le cas du corpus que nous avons considéré et, a fortiori, dans le cas d'un corpus qui réunirait un plus grand nombre de textes.

5 Reproductibilité des explorations dans le bitexte

Dans ce qui précède, nous avons utilisé la traduction française du conte pour permettre au lecteur francophone de mieux s'approprier les résultats que nous obtenions à partir du volet hiéroglyphique du texte. Dans cette dernière section nous avons regroupé quelques résultats obtenus par la mise en œuvre des mêmes méthodes appliquées cette fois au volet français du bitexte. Ces résultats montrent que les phénomènes constatés sur le texte hiéroglyphique trouvent en quelque sorte un écho mesurable dans les résultats du même type que l'on obtient à partir de la traduction française.

Sur la courbe d'accroissement du vocabulaire établie à partir du volet français du corpus la stagnation est encore plus perceptible que sur la courbe réalisée à partir du volet hiéroglyphique correspondant. Cette stagnation est encore plus marquée sur la courbe, située dans le bas du graphique, qui rend compte de l'apparition des hapax au fil du texte.

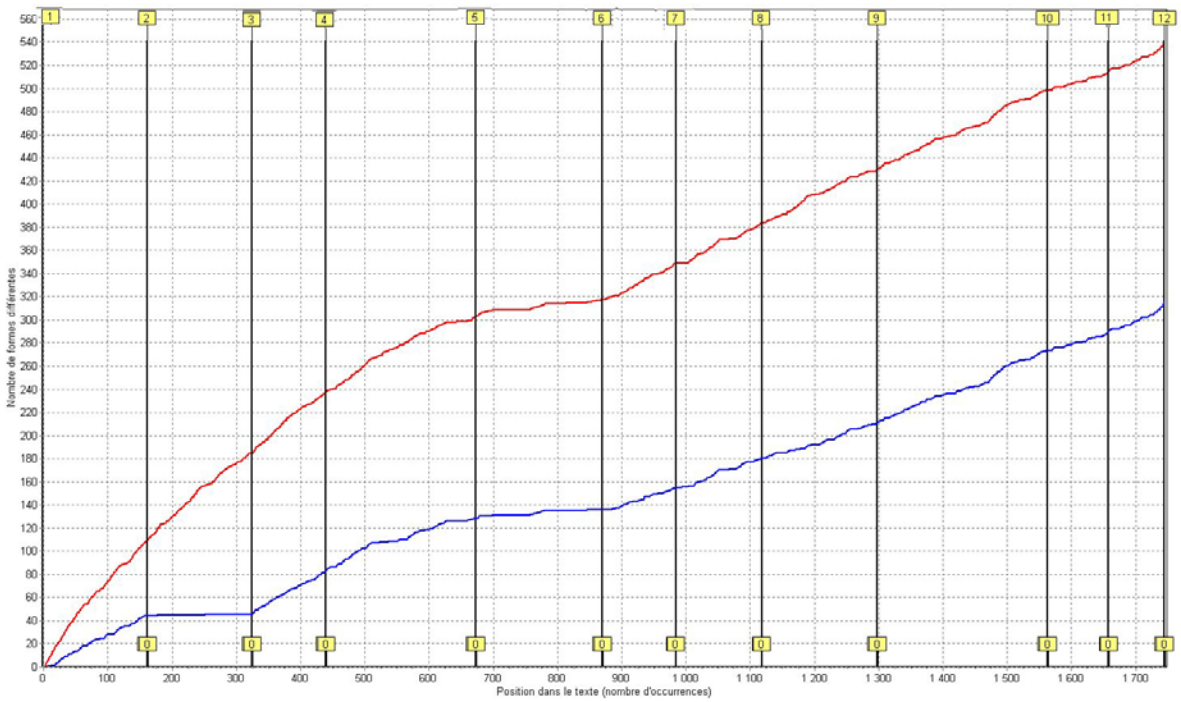


Figure : 9

Courbe d'accroissement du vocabulaire et courbe d'accroissement du nombre des hapax calculées pour le volet français du corpus ***Naufragé***

Comme c'était le cas pour le volet hiéroglyphique du corpus, la ventilation des segments répétés les plus longs montre une répartition privilégiée de certains segments entre les fragments 2 et 5 de la traduction française du conte.

On vérifiera sans surprise que les traductions françaises des deux parties constituées par la répétition d'un même récit dans le corpus original ont amené la création de textes qui sont très proches entre eux.

Tableau : 5

Extrait des concordances réalisées à partir des occurrences des segments répétés les plus longs dans le volet français du corpus ***Naufragé***

```
Partie : 01Intro, Nombre de contextes : 1,
§ car c ' est fatiguant de te parler . laisse - moi donc te raconter § quelque chose de
Partie : 02VoyageEtNaufrage, Nombre de contextes : 5
tais descendu § vers la mer , à bord d ' un navire § de 120 coudées de long et 40 coudées
 large . 120 marins s ' y trouvaient , § de l ' élite de l ' égypte . qu ' ils scrutassent
a venue , un orage § avant son arrivée . une tempête était survenue § alors que nous
rvenue § alors que nous étions en mer et avant § que nous eussions touché terre . le vent
ta pas § un . et je fus déposé § sur une île par une vague de la mer . § je passai trois
Partie : 04LeSerpent, Nombre de contextes : 2
 ' il ouvrit la bouche vers moi , tandis que § j ' étais à plat ventre devant lui , §
. § il ouvrit sa bouche vers moi , alors que § j ' étais à plat ventre devant lui § "
Partie : 05RecitNaufAuSerp, Nombre de contextes : 5
 les mines en mission § du souverain sur un navire de § 120 coudées de long et 40 coudées
 . § 120 marins se trouvaient à bord , § de l ' élite de l ' égypte . § qu ' ils scrutassent
' y avait pas § de maladroit parmi eux . une tempête § était survenue alors que nous
urvenue alors que nous étions en mer , § avant que nous eussions touché terre . § " le
 § voici que j ' ai été déposé sur cette île par § une vague de la mer . § il me dit
```

6 Conclusion

Dans cette étude exploratoire portant sur un corpus de textes hiéroglyphiques, nous avons montré comment des méthodes textométriques pouvaient être requises pour explorer les répétitions segmentales à l'oeuvre dans un corpus de textes. L'étude de ces répétitions permet de mettre en évidence différents types de reprises textuelles : reprises de fragments étendus lorsqu'il s'agit de la répétition d'une portion de récit, reprises de fragments plus courts dans le cas de la répétition de formules, de locutions, d'expressions plus ou moins figées en langue.

L'étude d'un corpus de texte hiéroglyphique pratiquée en liaison avec celle de sa traduction alignée dans une langue plus accessible aux chercheurs contemporains (bitexte aligné) permet d'éclairer les résultats textométriques obtenus sur le corpus hiéroglyphique à l'aide de résultats du même type obtenus à partir de leur traduction. Cette possibilité permet d'envisager l'études systématique des traductions obtenues à partir de corpus hiéroglyphiques nettement plus vastes que le corpus réduit que nous avons considéré pour cette première étude.

 (S34 U28 S29) *Vie, prospérité, santé !*[94]

[94] Formule d'eulogie, (i.e.) courte proposition exclamative appelant toutes sortes de bénédictions sur la personne qui fait l'objet du texte, souvent placée à la fin des textes hiéroglyphiques égyptiens.

7 Références

Brunet, E., (2000). « Qui lemmatise, dilemme attise », in *Lexicometrica*, no 2.

Lamalle, C, Salem, A., (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in Actes *des 6èmes Journées d'analyse des données textuelles,* St Malo.

Mayaffre, D. (2005). De la lexicométrie à la logométrie, *L'Astrolabe.*

Muller, Ch., (1963). « Le Mot, unité de texte et unité de lexique en statistique lexicologique », in *Travaux de linguistique et de littérature,* 1.

Salem, A. (1987). *Pratique des segments répétés*, Publications de l'INaLF, collection "St.Cloud", Klincksieck, Paris.

Zimina, M., (2004). *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles,* Thèse de doctorat , Université de la Sorbonne nouvelle – Paris 3, Paris.

Zipf, G. K., (1935). *The Psychobiology of Language, an Introduction to Dynamic Philology*, Houghton-Mifflin, Boston.

Webographie

Site du *Projet Rosette* : http://projetrosette.info/page.php?Id=1

Présentation et texte intégral du conte du naufragé :

http://pagesperso-orange.fr/sylvie.griffon/textes/naufrage/naufrage.htm