SYLED - CLA2T
Université de la Sorbonne Nouvelle - Paris 3

Explorations textométriques

Volume 3 : *corpus multilingues*

Sous la direction de
André Salem et Serge Fleury

S. Fleury, M. Zimina, J. Miao,
A. Salem, J-H. Cho, Christian Jean

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

Tutoriel n°3 : Exploration du corpus « Traductions alignées du discours d'investiture » de B. Obama

Corpus alignés, méthodes textométriques pour l'alignement

[Obama1]

Serge Fleury

Apprendre à :

- Construire une ressource textométrique alignée
- Utiliser les outils textométriques de base sur un alignement de textes
- Conduire une exploration textométrique sur un corpus aligné

1. Le corpus « Traductions alignées du discours d'investiture de B.Obama » (Investiture Obama)

Le corpus ***Investiture Obama*** est constitué de 5 volets : le discours original en anglais prononcé par B. Obama le 20 janvier 2009 à Washington et 4 traductions en français de ce discours.

Ces différents volets ont été récupérés sur différents site web :

Volet EN : le discours en anglais disponible sur le site du New York Times. Cette page n'est plus accessible à ce jour. On peut accéder à une version de cette page sauvegardée pour cette étude à cette adresse :

http://tal.univ-paris3.fr/mkAlign/corpus/obama-tmx-v5/PDF/nyt.pdf

Volet FR-1 : traduction en français fournie par les services de la Maison Blanche. On peut accéder à une version de cette page sauvegardée pour cette étude à cette adresse :

http://tal.univ-paris3.fr/mkAlign/corpus/obama-tmx-v5/PDF/traduction-maison-blanche.pdf

Volet FR-2 : traduction fournie sur le site du Monde. Cette page n'est plus accessible à ce jour. On peut accéder à une version de cette page sauvegardée pour cette étude à cette adresse :

http://tal.univ-paris3.fr/mkAlign/corpus/obama-tmx-v5/PDF/LeMonde.pdf

Volet FR-3 : traduction fournie sur le site de Libération (via l'AFP). Cette page n'est plus accessible à ce jour. On peut accéder à une version de cette page sauvegardée pour cette étude à cette adresse :

http://tal.univ-paris3.fr/mkAlign/corpus/obama-tmx-v5/PDF/libe.pdf

Volet FR-4 : traduction fournie sur le site de RFI. Cette page n'est plus accessible à ce jour. On peut accéder à une version de cette page sauvegardée pour cette étude à cette adresse :

http://tal.univ-paris3.fr/mkAlign/corpus/obama-tmx-v5/PDF/RFI.pdf

2. Construction du corpus aligné

Les contenus textuels des différentes pages web contenant le discours ou sa traduction ont été sauvegardés dans 5 fichiers différents au format texte brut : en.txt (volet EN), fr-0.txt (volet FR-1), fr-1.txt (volet FR-2), fr-2.txt (volet FR-3), fr-3.txt (volet FR-4). Les volets EN et FR-1 ont servi d'amorce pour construire l'alignement global. Ces deux volets étant alignés, on a ensuite aligné FR-1 avec FR-2, FR-2 avec FR-3 et enfin FR-3 avec FR-4.

Cet alignement a été construit avec *mkAlign*[1] qui fournit des outils d'aide à l'alignement dans un éditeur à 2 volets ; il permet aussi de sauvegarder l'alignement dans un format normalisé (le format TMX[2]) permettant de stocker pour une ressource textuelle donnée différents volets associés (comme ses différentes traductions par exemple).

2.1 Etape n°1 : alignement de 2 volets initiaux

- En entrée : `en.txt`, `fr-0.txt` (les 2 volets initiaux)
- En sortie : `en_mkAlign.txt`, `fr-0_mkAlign.txt`, `obama-alignement-en-fr1.tmx` (les 2 fichiers sauvegardés à l'issue de l'alignement et la version TMX de l'alignement)

La figure suivante donne à voir l'interface de *mkAlign* permettant de construire un alignement.

[1] http://tal.univ-paris3.fr/mkAlign/

[2] http://en.wikipedia.org/wiki/Translation_Memory_eXchange

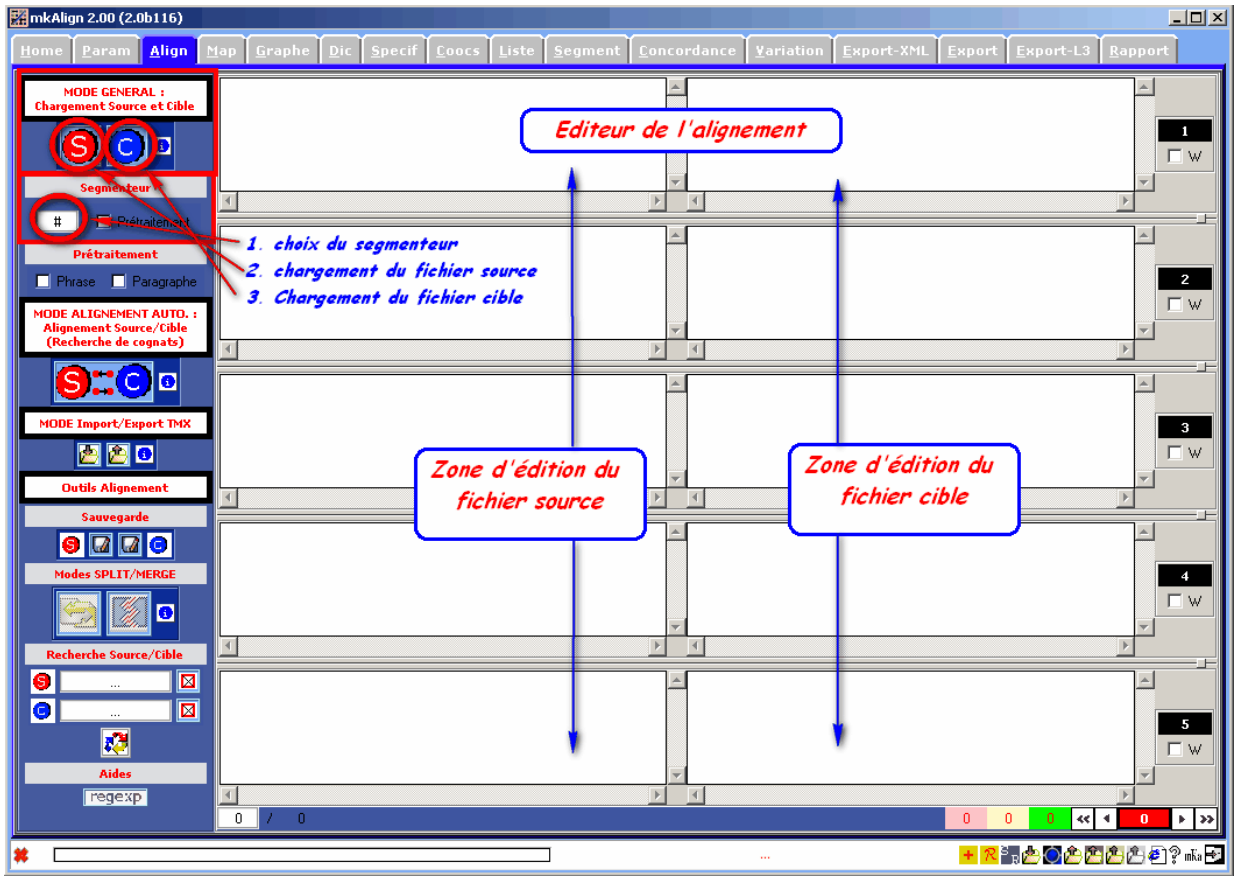


Figure 1 : Interface de l'alignement avec mkAlign

Pour cette étude, nous avons choisi d'aligner au niveau de la phrase. *mkAlign* permet de sélectionner un caractère (le *segmenteur d'alignement*) permettant de découper les textes à aligner pour ensuite charger les différentes sections résultantes dans les zones d'édition disponibles : chaque page contient 5 zones d'édition alignées permettant de visualiser chaque couple de sections textuelles alignées. Notre objectif d'alignement phrastique nous a conduit, pour amorcer grossièrement les choses, à charger les 2 volets initiaux en choisissant comme *segmenteur d'alignement* le caractère retour à la ligne.

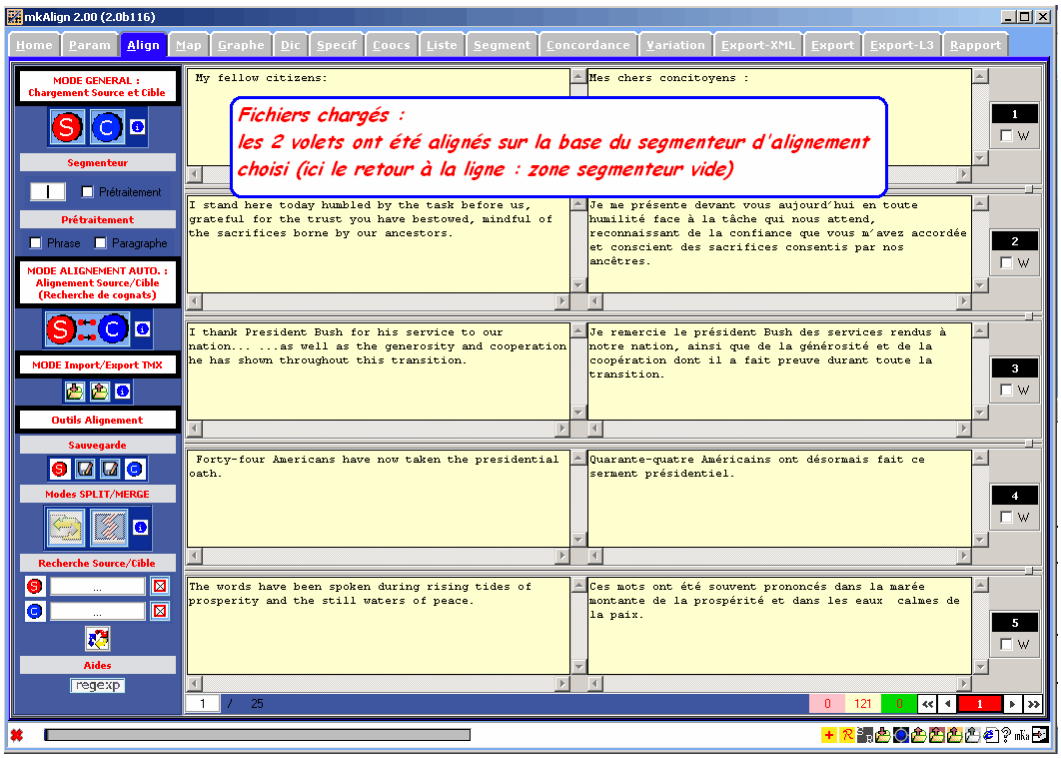


Figure 2 : Chargement des fichiers à aligner

Les 2 volets étant chargés, on peut ensuite affiner l'alignement en utilisant les outils idoines pour scinder certaines sections ou en fusionner d'autres.

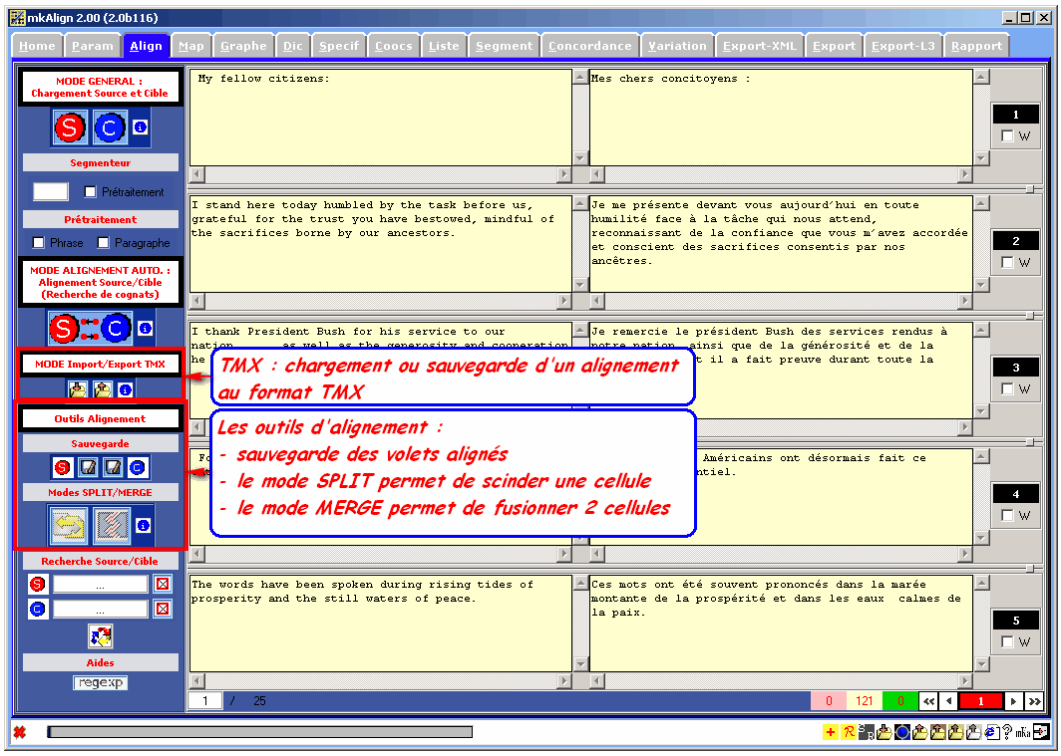
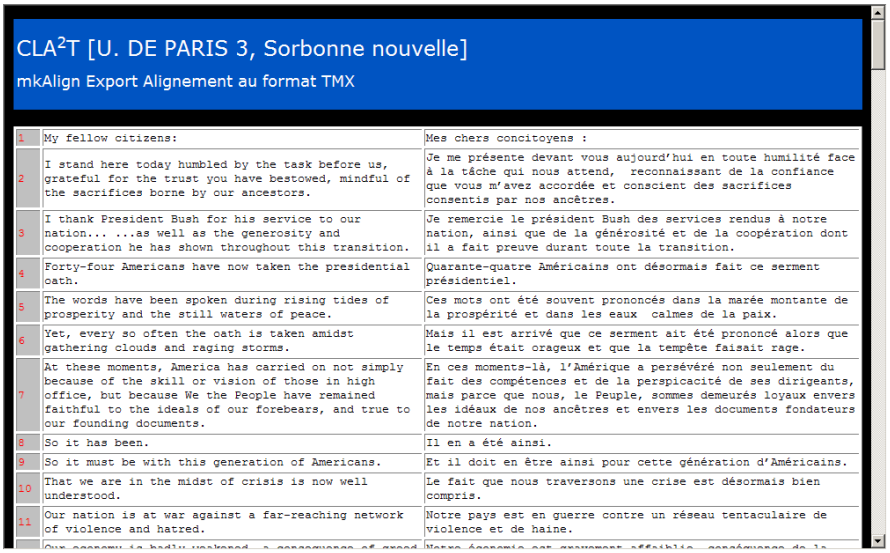


Figure 3 : Les outils de l'alignement

Au final, on dispose dans l'éditeur de l'alignement d'un corpus aligné avec lequel on peut mener des explorations textométriques (*cf infra*). On peut aussi sauvegarder chacun des volets ainsi remodelés (dans 2 fichiers) ou exporter les 2 volets dans un fichier au format TMX, ce type de fichier permettant de stocker de manière séquentielle les différentes sections alignées. La première figure qui suit montre l'état de l'alignement exporté au format TMX tel qu'il est affiché dans un navigateur avec une feuille de styles fournie :



1	My fellow citizens:	Mes chers concitoyens :
2	I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors.	Je me présente devant vous aujourd'hui en toute humilité face à la tâche qui nous attend, reconnaissant de la confiance que vous m'avez accordée et conscient des sacrifices consentis par nos ancêtres.
3	I thank President Bush for his service to our nation... ...as well as the generosity and cooperation he has shown throughout this transition.	Je remercie le président Bush des services rendus à notre nation, ainsi que de la générosité et de la coopération dont il a fait preuve durant toute la transition.
4	Forty-four Americans have now taken the presidential oath.	Quarante-quatre Américains ont désormais fait ce serment présidentiel.
5	The words have been spoken during rising tides of prosperity and the still waters of peace.	Ces mots ont été souvent prononcés dans la marée montante de la prospérité et dans les eaux calmes de la paix.
6	Yet, every so often the oath is taken amidst gathering clouds and raging storms.	Mais il est arrivé que ce serment ait été prononcé alors que le temps était orageux et que la tempête faisait rage.
7	At these moments, America has carried on not simply because of the skill or vision of those in high office, but because We the People have remained faithful to the ideals of our forebears, and true to our founding documents.	En ces moments-là, l'Amérique a persévéré non seulement du fait des compétences et de la perspicacité de ses dirigeants, mais parce que nous, le Peuple, sommes demeurés loyaux envers les idéaux de nos ancêtres et envers les documents fondateurs de notre nation.
8	So it has been.	Il en a été ainsi.
9	So it must be with this generation of Americans.	Et il doit en être ainsi pour cette génération d'Américains.
10	That we are in the midst of crisis is now well understood.	Le fait que nous traversons une crise est désormais bien compris.
11	Our nation is at war against a far-reaching network of violence and hatred.	Notre pays est en guerre contre un réseau tentaculaire de violence et de haine.

Figure 4: Alignement au format TMX, affichage dans le navigateur

La seconde montre un extrait du code source de ce fichier au format TMX :

```
<?xml version="1.0" encoding="UTF-8" ?>
<tmx version="1.4">
<header adminlang="en" creationdate="20090712T110800Z" creationtool="mkAlign" creationtoolversion="2.00 (2.0b116)"
datatype="xml" o-tmf="unknown" segtype="block" srclang="en"/>
<body>
<tu>
<tuv xml:lang="en">
<seg> My fellow citizens:

</seg>
</tuv>
<tuv xml:lang="fr">
<seg>Mes chers concitoyens :

</seg>
</tuv>
</tu>
<tu>
<tuv xml:lang="en">
<seg>I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices
borne by our ancestors.

</seg>
</tuv>
<tuv xml:lang="fr">
<seg>Je me présente devant vous aujourd'hui en toute humilité face à la tâche qui nous attend,  reconnaissant de la confiance
que vous m'avez accordée et conscient des sacrifices consentis par nos ancêtres.

</seg>
</tuv>
</tu>
<tu>
<tuv xml:lang="en">
<seg>I thank President Bush for his service to our nation... ...as well as the generosity and cooperation he has shown
throughout this transition.

</seg>
</tuv>
<tuv xml:lang="fr">
<seg>Je remercie le président Bush des services rendus à notre nation, ainsi que de la générosité et de la coopération dont il
a fait preuve durant toute la transition.

</seg>
</tuv>
```

Figure 5: Code source du fichier d'alignement au format TMX

2.2 Etape n°2 : Généralisation de l'alignement

L'opération décrite dans l'étape précédente a été répétée sur les différents couples de textes disponibles. Les fichiers TMX construits à chaque étape ont ensuite été « fusionnés » pour fournir au final un fichier regroupant les différents volets alignés : l'alignement construit ici est composé pour chaque section d'alignement de 5 volets, le volet anglais et ses 4 traductions.

CLA²T [U. DE PARIS 3, Sorbonne nouvelle]

(mkAlign) Alignement au format TMX : Le discours d'investiture de Barak Obama, le 20 janvier 2009, à Washington.

Source	NEW YORK TIMES	Trad. White-House	MONDE	LIBERATION/AFP	RFI
1	My fellow citizens:	Mes chers concitoyens :	Chers compatriotes,	Chers compatriotes	
2	I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors.	Je me présente devant vous aujourd'hui en toute humilité face à la tâche qui nous attend, reconnaissant de la confiance que vous m'avez accordée et conscient des sacrifices consentis par nos ancêtres.	Je me tiens aujourd'hui devant vous avec un sentiment d'humilité, devant la tâche qui nous attend, de reconnaissance pour la confiance que vous m'avez manifestée, gardant à l'esprit les sacrifices consentis par nos ancêtres.	Je suis ici devant vous aujourd'hui empli d'un sentiment d'humilité face à la tâche qui nous attend, reconnaissant pour la confiance que vous m'avez témoignée et conscient des sacrifices consentis par nos ancêtres.	Je suis là devant vous humble face aux tâches qui nous attendent, reconnaissant de votre confiance et attentif aux sacrifices de nos ancêtres.
3	I thank President Bush for his service to our nation... ...as well as the generosity and cooperation he has shown throughout this transition.	Je remercie le président Bush des services rendus à notre nation, ainsi que de la générosité et de la coopération dont il a fait preuve durant toute la transition.	Je remercie le président Bush pour les services qu'il a rendus à notre nation, ainsi que pour la générosité et la coopération dont il a fait preuve tout au long de cette transition.	Je remercie le président Bush pour ses services rendus à la nation ainsi que pour la générosité et la coopération dont il a fait preuve tout au long de cette passation de pouvoirs.	Je remercie le président Bush, pour ses services rendus à la nation, ainsi que pour toute la générosité et la coopération qu'il a montrées lors de toute cette période de transition.
4	Forty-four Americans have now taken the presidential oath.	Quarante-quatre Américains ont désormais fait ce serment présidentiel.	Quarante-quatre Américains ont, avant moi, prêté serment pour la présidence.	Quarante-quatre Américains ont maintenant prêté le serment présidentiel.	Quarante-quatre Américains ont déjà prêté serment.
5	The words have been spoken during rising tides of prosperity and the still waters of peace.	Ces mots ont été souvent prononcés dans la marée montante de la prospérité et dans les eaux calmes de la paix.	Leurs paroles ont été prononcées pendant des vagues de prospérité et alors que nous vivions dans les eaux calmes de la paix.	Ils l'ont fait alors que gonflait la houle de la prospérité sur les eaux calmes de la paix.	Des mots ont été prononcés lors de marées montantes de prospérité et de mers calmes de la paix.
		Mais il est arrivé que ce serment	Cependant, en d'autres temps, ce	Mais il arrive de temps à autre	

Figure 6: Alignement du corpus « Obama Investiture ». Affichage dans un navigateur

2.3 Etape n°3 : Exploration textométrique de l'alignement

mkAlign permet de mener des explorations textométriques sur des couples de textes alignés. Dans notre cas, le fichier TMX étant composé de 5 volets, il est nécessaire de sélectionner au préalable 2 volets avec de démarrer cette exploration. Dans les exemples qui suivent nous travaillerons avec les 2 volets FR-1 et FR-2. La figure qui suit montre l'état de l'alignement de ces 2 volets.

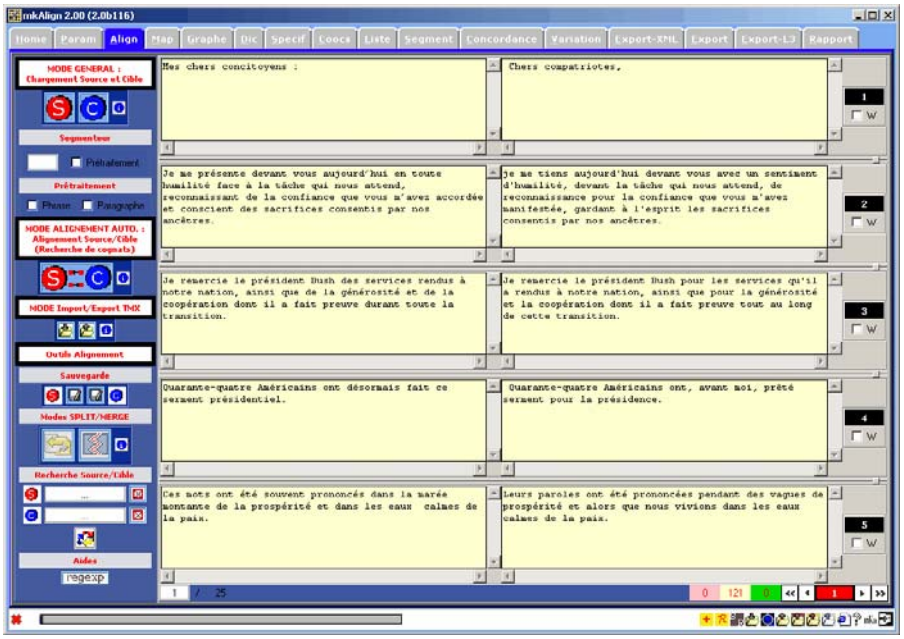
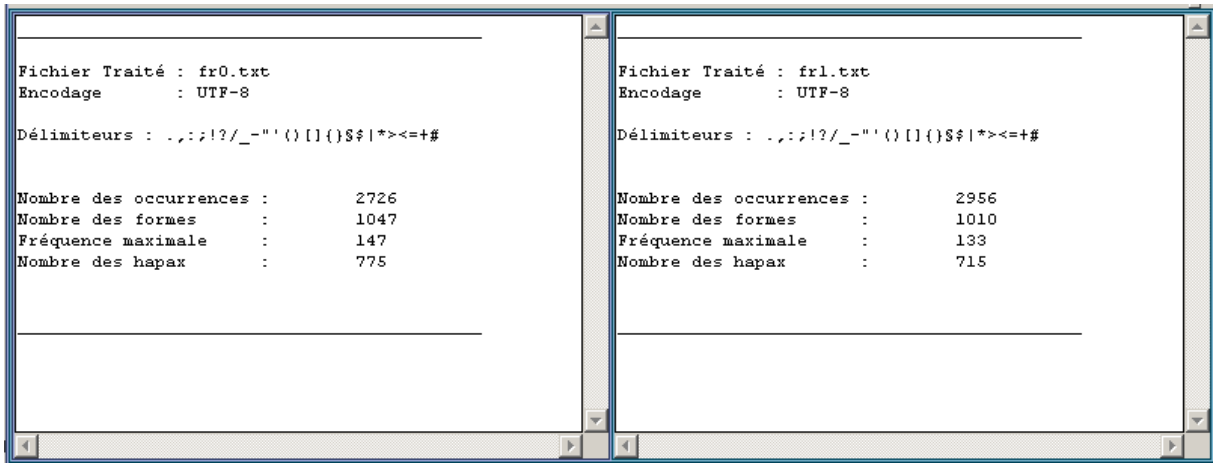


Figure 7: Alignement Volets FR-1 et FR-2

2.3.1 Le dépouillement en formes graphiques

Comme pour toute exploration textométrique, *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)*. Dans le cas de *mkAlign*, cette segmentation des 2 volets en unités est réalisée au chargement des fichiers. Le dépouillement des 2 volets en formes graphiques délimitées par les délimiteurs proposés par défaut conduit aux résultats suivants (visibles dans l'onglet Rapport de *mkAlign*) :



```
Fichier Traité : fr0.txt
Encodage       : UTF-8

Délimiteurs : .,:;!?/_-"'()[]{}$$|*><=+#

Nombre des occurrences :        2726
Nombre des formes      :        1047
Fréquence maximale     :        147
Nombre des hapax       :        775
```

```
Fichier Traité : fr1.txt
Encodage       : UTF-8

Délimiteurs : .,:;!?/_-"'()[]{}$$|*><=+#

Nombre des occurrences :        2956
Nombre des formes      :        1010
Fréquence maximale     :        133
Nombre des hapax       :        715
```

Figure 8: Paramètres lexicométriques des deux volets alignés

Cette segmentation conduit à la génération des 2 dictionnaires de formes, chacun étant associé à un des volets du corpus aligné :

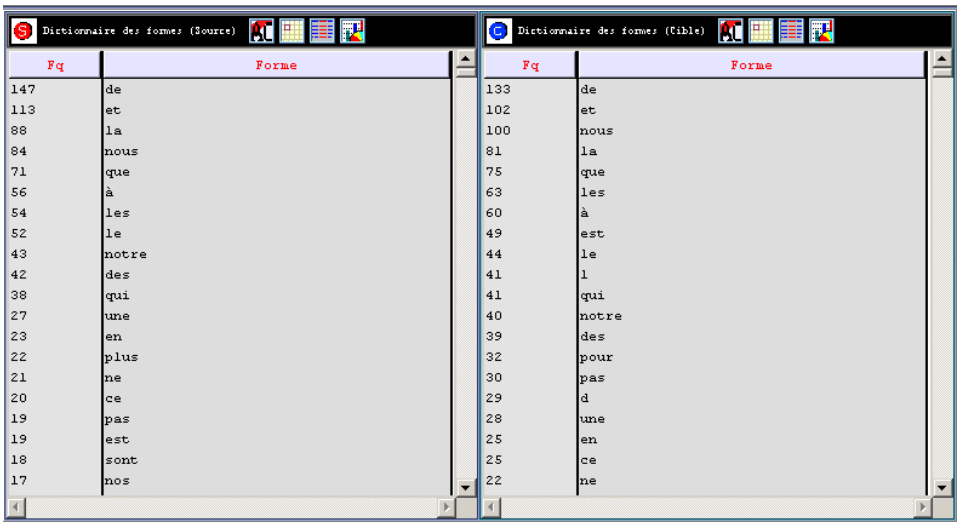


Figure 9: Les dictionnaires de formes issus de l'alignement

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. Dans la figure précédente, certains de ces outils sont visibles dans la partie supérieure sous la forme d'icône. Après avoir sélectionné des items dans la liste, on active l'opération visée pour ces items.

2.3.2 Etude globale des types simples

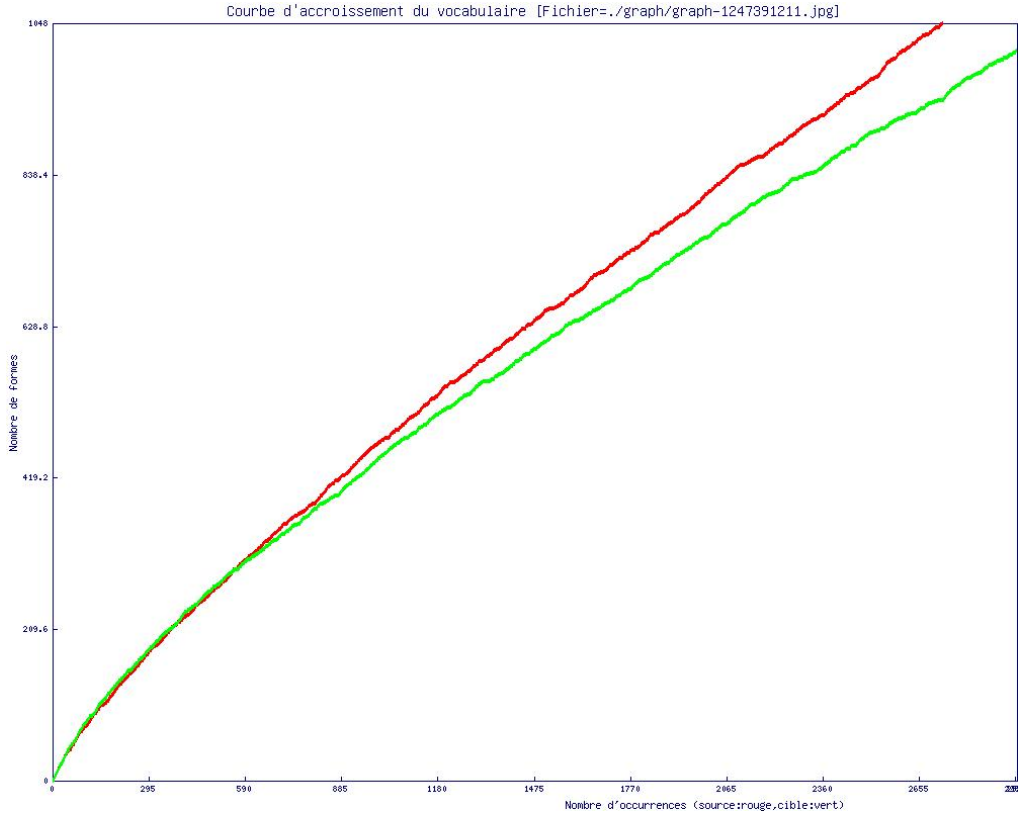


Figure 10 : Accroissement de vocabulaire sur les 2 volets de l'alignement

Le ***Diagramme d'accroissement du vocabulaire*** 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. Dans le cas de *mkAlign*, on peut observer cette ventilation sur les 2 volets chargés.

==== ***mkAlign*** ==== ***Accroissement du vocabulaire***

- ✓ Dans l'onglet Graphe, activez le bouton AC
- ✓ Le diagramme apparaît dans la zone d'édition de l'onglet Graphe.

2.3.3 Les types complexes

Les segments répétés

La fonctionnalité ***Segments répétés*** permet d'établir la liste de toutes les séquences de formes répétées (pour les 2 volets alignés) 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.

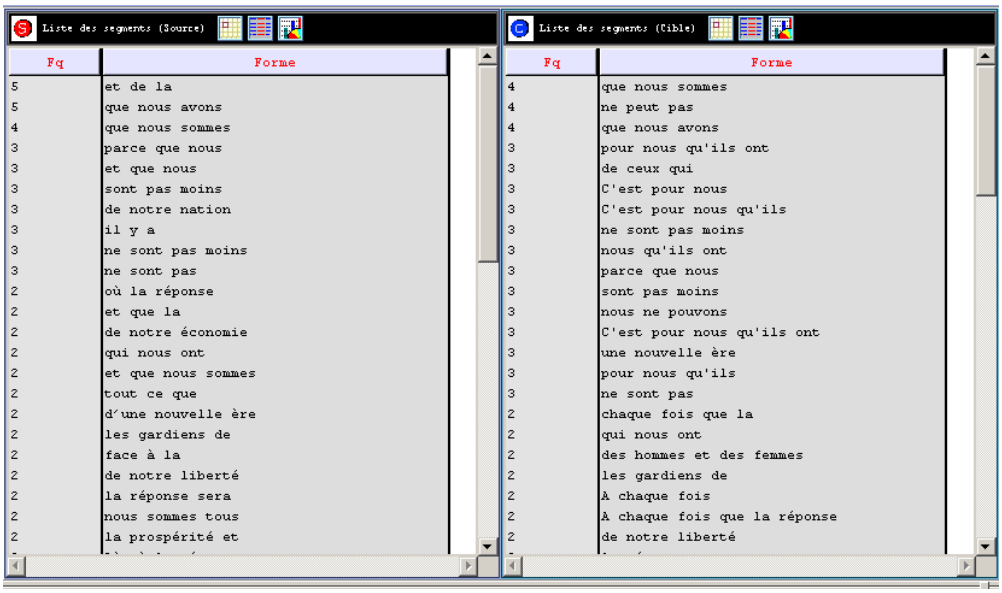


Figure 11: Liste des segments répétés sur les 2 volets du corpus

==== ***mkAlign*** ==== ***Segments répétés***

- ✓ Dans l'onglet Param, sélectionner un seuil de fréquence minimal pour les segments
- ✓ Dans l'onglet Segments, activez le calcul
- ✓ Les segments apparaissent dans la zone d'édition de l'onglet Segments sous la forme de 2 listes. 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 à l'aide des différents outils disponibles (concordance, histogramme, carte des sections, etc.) au dessus de chaque liste.

Cooccurrences et polycooccurrences pour un type donné

Un alignement induit un découpage du corpus en sections (les différentes cellules alignées). Pour une forme-pôle (nous prendrons comme ci-dessus l'exemple de la forme : ***nation***) il est possible de constituer la liste des formes qui trouvent, d'après un calcul statistique particulier[3], un nombre élevé d'occurrence dans les mêmes sections que la forme-pôle sur chacun des volets.

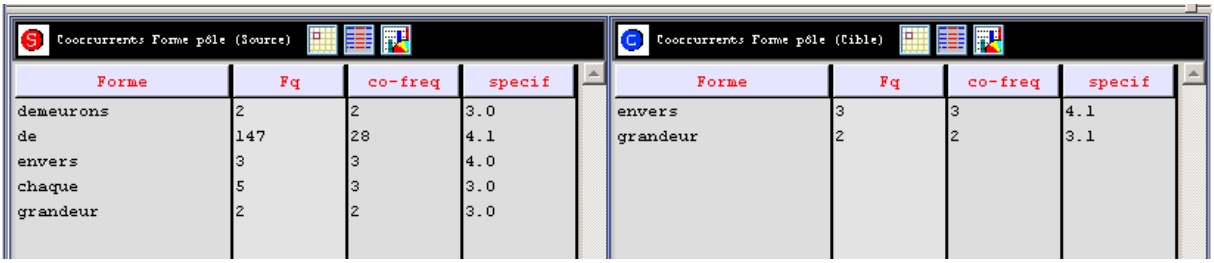


Figure 12 : Les cooccurrents de "nation"

Nous trouvons ici pour la forme-pôle sur le volet FR-1 : *demeurons, de, envers, chaque, grandeur* et pour cette même forme-pôle sur le volet FR-2 : *envers, grandeur*

Le retour aux contextes confirmera que ces formes entrent avec le pôle choisi dans des associations récurrentes :

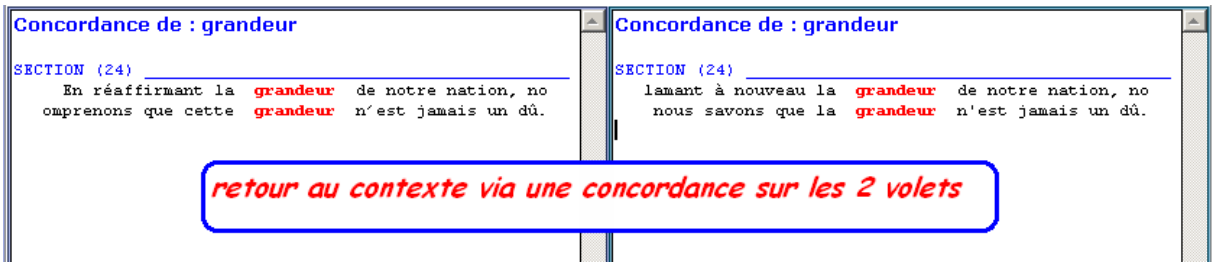


Figure 13 : Retours aux contextes

==== *mkAlign* ==== *Cooccurrences*

- ✓ Dans l'onglet Coocs, sélectionner la forme pôle (volet source et volet cible)
- ✓ Charger éventuellement une liste de forme à exclure du calcul (stop-liste) sur chacun des 2 volets
- ✓ Choisir une fréquence minimale et un seuil de probabilité pour les cooccurrents
- ✓ Appuyer sur l'icône des cooccurrences,

On verra *infra* qu'il est possible de déterminer cette liste de cooccurrents en utilisant dans *mkAlign* une autre méthode basée sur la représentation graphique de l'alignement.

A partir de la liste de cooccurrents, on peut ensuite activer le calcul des polycooccurrents. Ce calcul reprend la démarche mise en œuvre dans le travail de William Martinez (2002, 2003, 2006).

---

[3] Un calcul hypergéométrique est utilisé ici 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.

- Une *cooccurrence* désigne l'apparition de deux mots en même temps et dans le même contexte.

Le module de cooccurrences mis en œuvre prend appui sur l'alignement en cours, les contextes dans lesquels on examine la co-présence sont donc ceux qui coïncident aux différentes cellules dans l'éditeur d'alignement (ou aux sections dans la carte des sections)

- Le terme *poly-cooccurrence* désigne les attractions lexicales au-delà de la cooccurrence binaire.

Le module de poly-cooccurrences intégré reprend l'algorithme décrit dans [Martinez, 2006] :

- On calcule pour le pôle A les cooccurrents spécifiques B, C et D
- Dans leurs contextes communs, on calcule pour les pôles A+B les cooccurrents spécifiques E et F
- Les pôles A+B+E ont pour cooccurrent spécifique H
- Les pôles A+B+E+H n'ont pas de cooccurrent spécifique et l'exploration s'interrompt pour ce chemin
- Les pôles A+B+F ont pour cooccurrents spécifiques I, etc.
- Durant l'exploration, différents filtrages conditionnent l'épuisement des explorations contextuelles et réduisent le bruit dans les résultats pour privilégier l'information la plus spécifique : seuils maximaux de fréquence et de spécificité du cooccurrent.

Le calcul des cooccurrents étant terminé, l'activation du module de polycooccurrence construit les chemins de polycooccurrence ; le graphique suivant construit par *mkAlign* synthétise l'ensemble de ces chemins que nous insérons[4] plus bas :

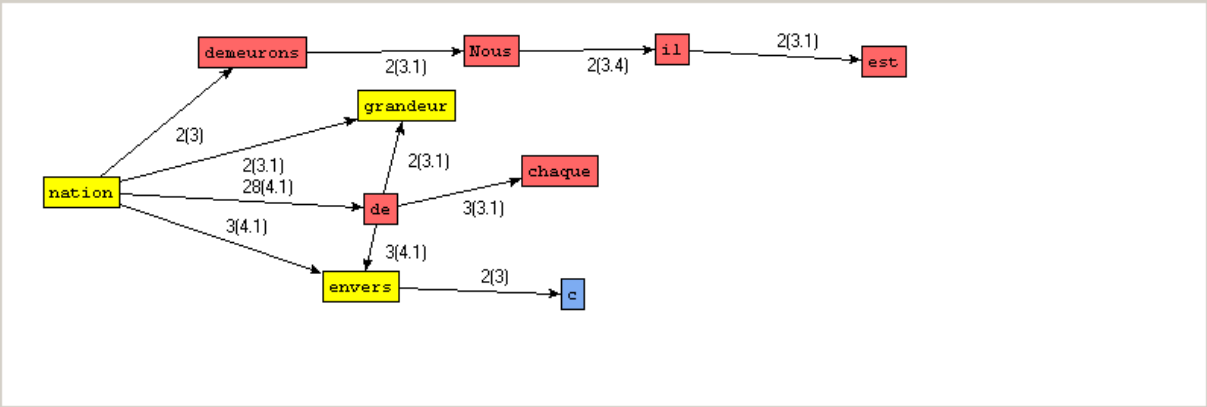


Figure 14 : Les polycooccurrents de la forme "nation"

Dans ce graphique, les formes en jaune sont présentes dans les 2 volets, les formes en rouge sont présentes dans le volet source (FR-1) et les formes en bleu sont présentes dans le volet cible (FR-2).

Polycooccurrents : (FR-1) nation (co-freq : 2, seuil : 3)
nation-2(3)->**demeurons**-2(3.1)->**Nous**-2(3.4)->**il**-2(3.1)->**est**
nation-28(4.1)->**de**-2(3.1)->**grandeur**
nation-28(4.1)->**de**-3(4.1)->**envers**
nation-28(4.1)->**de**-3(3.1)->**chaque**

Polycooccurrents : (FR-2) nation (co-freq : 2, seuil : 3)
nation-3(4.1)->**envers**-2(3)->**c**
nation-2(3.1)->**grandeur**

[4] Les chemins de polycooccurrence sont accessibles après sauvegarde des résultats du calcul dans le rapport d'exploration (*cf* « sauvegarder un rapport » dans le manuel d'utilisation).

Le graphique des chemins de polycooccurrences permet aussi de réaliser des retours au contexte en sélectionnant des nœuds « forme » (Control-Clic sur un nœud) que l'on peut ensuite projeter sur la carte des sections de l'alignement (icône carte des sections dans la partie haute de la zone d'édition du graphe dans l'onglet Cooc). Cette projection permet de mettre au jour les sections contenant l'ensemble des formes sélectionnées (Option « Global » cochée) ou celles contenant au moins l'une des d'entre elles. On peut ainsi visualiser rapidement les sections contenant des chemins complets de polycooccurrences.

==== ***mkAlign*** ==== ***Polycooccurrences***

- ✓ Dans l'onglet Coocs, sélectionner la forme pôle (volet source et volet cible)
- ✓ Charger éventuellement une liste de forme à exclure du calcul (stop-liste) sur chacun des 2 volets
- ✓ Choisir une fréquence minimale et un seuil de probabilité pour les cooccurrents
- ✓ Appuyer sur l'icône des cooccurrences
- ✓ Appuyer sur l'icône des polycooccurrents
- ✓ Le graphe des polycooccurrents apparaît dans la zone supérieur de la zone d'édition de l'onglet Coocs. Les chemins de cooccurrence seront accessibles dans le rapport si les résultats produits y sont ajoutés

3. Etude la distribution d'un type

3.1 Les outils de base

3.1.1 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. 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. La concordance est ici disponible pour chacun des volets du corpus aligné.

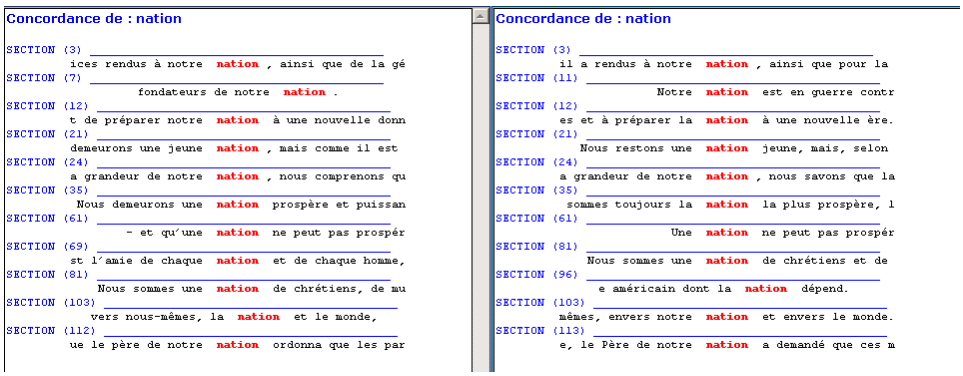


Figure 15 : Concordance de la forme nation sur les 2 volets du corpus

==== ***mkAlign*** ==== ***Concordances***

- ✓ Dans l'onglet ***Concordances***
- ✓ Entrer une forme dans la zone de saisie ***(ex : nation)***
- ✓ Choisir [éventuellement] un regroupement par parties (si une partition a été sélectionnée)

3.1.2 L'outil ventilation par sections d'alignement

Cet outil permet de juger de la répartition des occurrences relevant d'un même type dans les différentes sections de l'alignement :

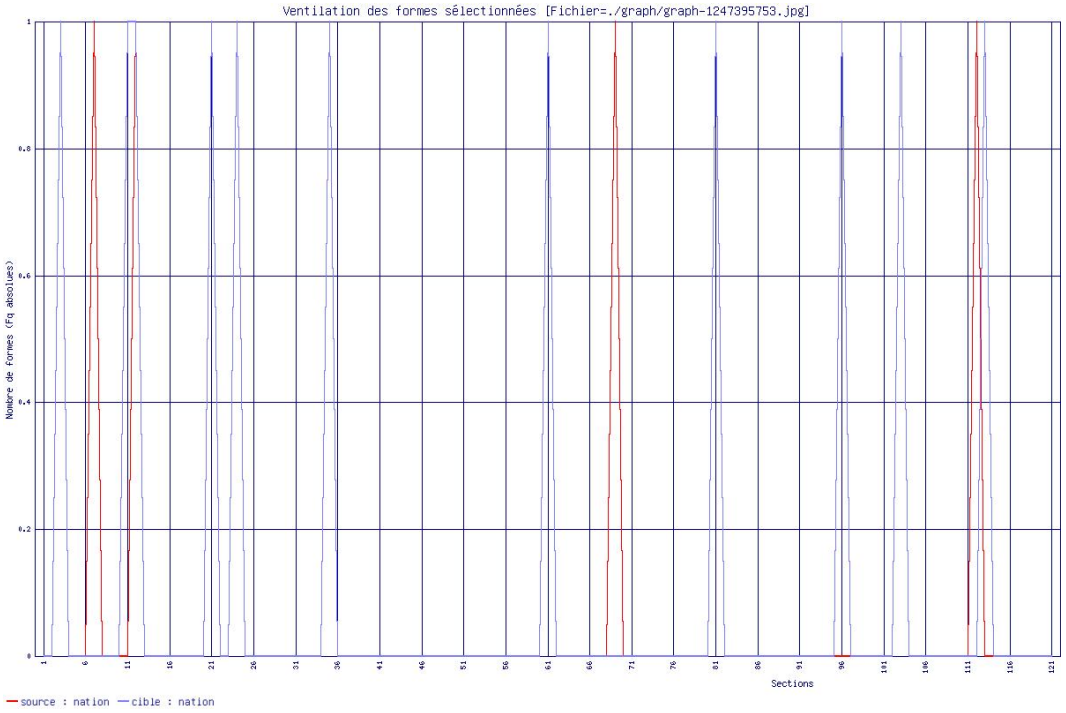


Figure 16 : Ventilation de la forme « nation » sur les 2 volets de l'alignement

==== *mkAlign* ==== *Ventilation par section d'alignement*

- ✓ Dans l'onglet Dic (et dans chaque onglet donnant à voir des listes de formes)
- ✓ Sélectionner une (ou plusieurs) forme(s)
- ✓ Activez le bouton Ventilation, la ventilation concernera l'ensemble des formes sélectionnées dans le volet source et dans le volet cible

3.1.3 L'outil carte des sections

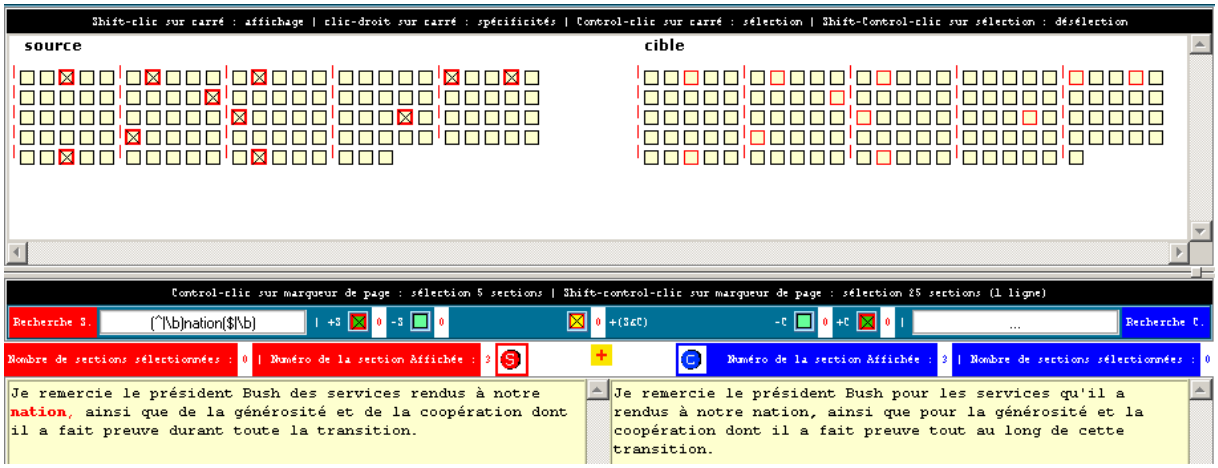


Figure 17 : Carte des sections ; projection de la forme "nation" sur le volet source

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 (constitué ici de 2 volets et donc de 2

cartes). Chacun des carrés représente un élément particulier du texte découpé en sections : les sections correspondent ici aux sections de l'alignement construit (les cellules alignées dans l'éditeur de l'alignement).

Chacun des carrés de la séquence du haut représente une des sections du texte original (volet source à gauche et volet cible à droite). La forme *nation* a été projetée sur la carte des sections à partir du dictionnaire (source) provoquant ainsi le marquage par une croix et le coloriage du contour des sections ou elle est attestée. Le texte d'une des sections sélectionnée par l'utilisateur est affiché en bas de la figure. Les occurrences de la forme sélectionnée y sont mises en évidence.

==== ***mkAlign*** ==== ***Carte des sections***

- ✓ Dans l'onglet ***Map***
- ✓ Activez la construction de la carte
- ✓ Projet une forme sur la carte à partir du dictionnaire par exemple ***(nation)***
- ✓ Choisir [éventuellement] un regroupement par parties, si une partition a été sélectionnée

4. 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.

4.1 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[5] (on se reportera au Tutorial n°1 pour des informations complémentaires sur la méthode des spécificités).

Exemple n°1 : Calcul des cooccurrents d'une forme à partir de la carte des sections de l'alignement

La carte des sections construit par définition un découpage du corpus en sections correspondant à l'état de l'alignement. Une forme-pôle étant choisi (sur le volet source ou le volet cible), la projection de la forme sur la carte des sections donne à voir la localisation de la forme dans la carte des sections. Nous reprenons ci-dessous l'exemple de la forme : ***nation*** et la projection construite dans la figure précédente. A partir de cette carte, il est possible de constituer la liste des formes et des segments répétés qui trouvent, d'après un calcul statistique particulier[6], un nombre élevé d'occurrence dans les mêmes sections que la forme-pôle (les cooccurrents de cette forme).

[5] 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].

[6] 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.

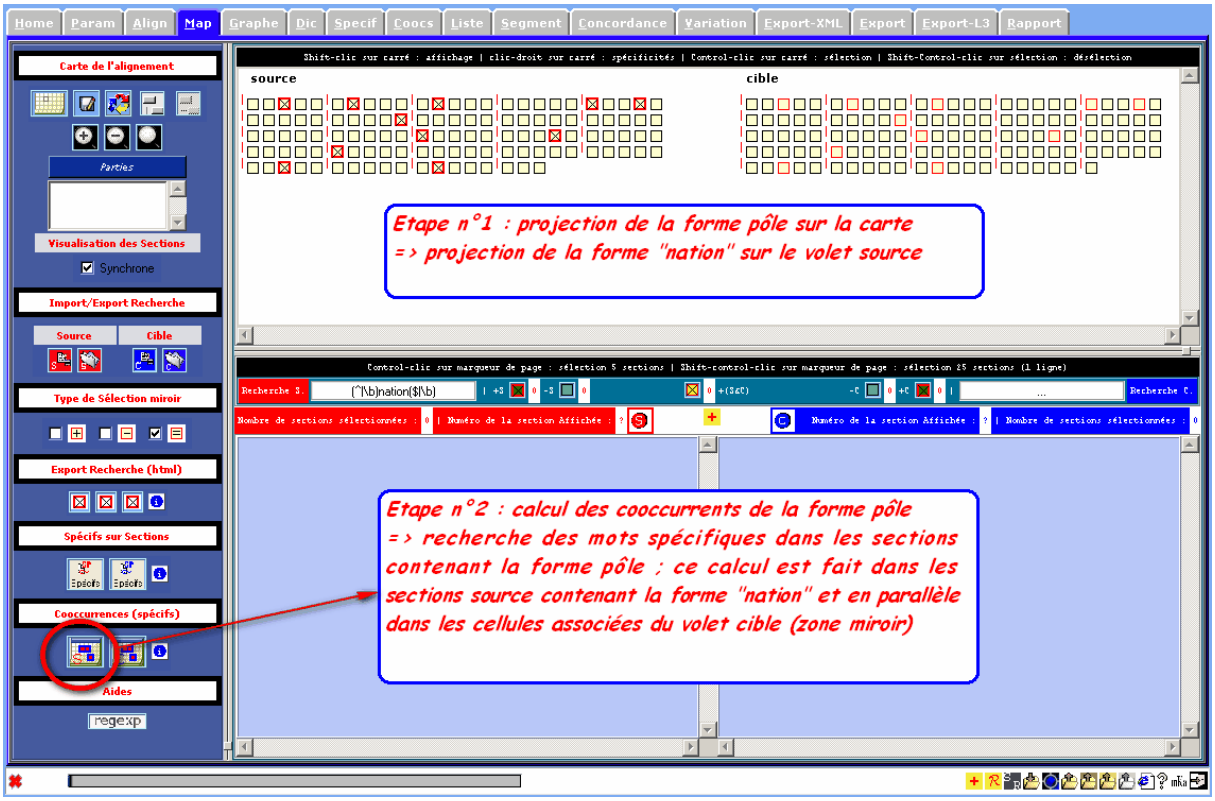


Figure 18 : Calcul des cooccurrents d'une forme par la carte des sections

Le résultat est constitué par deux listes donnant à voir d'une part les mots spécifiques de la forme-pôle (pour le volet source) et les mots spécifiques dans les sections associées du volet cible :

Specificites du vocabulaire sur les sections SOURCE contenant le motif : <nation>

Nombre d'occurrences du texte global = 2726
Nombre d'occurrences dans la partie visee = 284
Seuil : 5
(Specificites positives en haut de liste, negatives en bas)
Le fichier construit : EXPORT/TXT/resultspecif-source-1247416002.txt

Forme	Ind-Specif	Fq-Totale
nation	11.9	11
de	4.1	147
envers	4.0	3
chaque	3.0	5
demeurons	3.0	2
grandeur	3.0	2
En	2.2	4
ses	2.2	4
il	2.2	9
notre	2.1	43
ainsi	2.1	5
sommes	2.0	11
donne	2.0	1
part	2.0	1
membres	2.0	1
prospères	2.0	1
reconnaissance	2.0	1
Écritures	2.0	1
favorise	2.0	1
affaiblie	2.0	1
loyaux	2.0	1
remercie	2.0	1
persévéré	2.0	1
population	2.0	1
assumer	2.0	1
femme	2.0	1
transition	2.0	1
incapacité	2.0	1
rendus	2.0	1
cupidité	2.0	1
président	2.0	1
prospère	2.0	1
préparer	2.0	1
compétences	2.0	1
documents	2.0	1
Peuple	2.0	1

Specificites du vocabulaire sur les sections CIBLE associees aux sections SOURCE

Nombre d'occurrences du texte global = 2956
Nombre d'occurrences dans la partie visee = 297
Seuil : 5
(Specificites positives en haut de liste, negatives en bas)
Le fichier construit : EXPORT/TXT/resultspecif-cible-1247416003.txt

Forme	Ind-Specif	Fq-Totale
nation	6.9	11
envers	4.0	3
grandeur	3.0	2
la	2.9	81
de	2.7	133
preuve	2.3	4
ère	2.3	4
sachez	2.3	4
à	2.1	60
sommes	2.1	17
fortement	2.0	1
Pas	2.0	1
chrétiens	2.0	1
aspirent	2.0	1
Ecritures	2.0	1
neige	2.0	1
proclamant	2.0	1
mêmes	2.0	1
puissante	2.0	1
pairs	2.0	1
musulmans	2.0	1
favorise	2.0	1
nantis	2.0	1
affaiblie	2.0	1
remercie	2.0	1
vision	2.0	1
collective	2.0	1
restons	2.0	1
transition	2.0	1
moments	2.0	1
générosité	2.0	1
rapacité	2.0	1
incapacité	2.0	1
rendus	2.0	1
toujours	2.0	1
athées	2.0	1

Figure 19 : Liste des cooccurrents de la forme pôle et liste des mots spécifiques de la zone miroir

Nous retrouvons normalement ici les résultats déjà vus plus haut. Le corpus étant aligné, la forme en tête de liste est sans surprise la forme « nation » : les deux traductions convergent

sur cette forme localisée dans les mêmes sections dans les 2 volets, par contre les divergences entre les traductions se traduisent par des comportements lexicaux spécifiques propres à chaque volet.

4.2 Mise au jour de la variation entre les 2 volets du corpus aligné

Dans l'exemple traité dans ce tutorial, les volets français sont issus par une dérivation de traduction du même texte original. Dans ce cas précis, si on choisit 2 volets français particuliers, ces deux textes sont théoriquement proches (mais différents : les traductions n'étant pas complètement similaires 2 à 2). On peut donc vouloir essayer de mettre au jour les différences entre ces volets traduits du même texte de départ. Cette mise au jour de la variation est possible dans *mkAlign* : une fois les textes alignés, le module de variation donne à voir globalement les différences entre les 2 volets chargés. Ce processus s'appuie sur l'implémentation de la commande `diff`[7] dans la bibliothèque `Tk::DiffText`[8] (*composite widget for colorized diffs*)

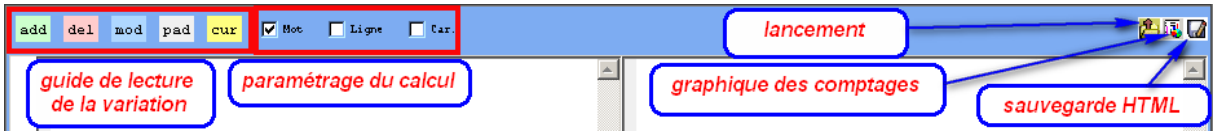


Figure 20 : paramétrage du calcul de la variation

Après avoir choisi le grain de la variation (mot, ligne, caractère), on lance la visualisation de la variation en activant le bouton idoine :

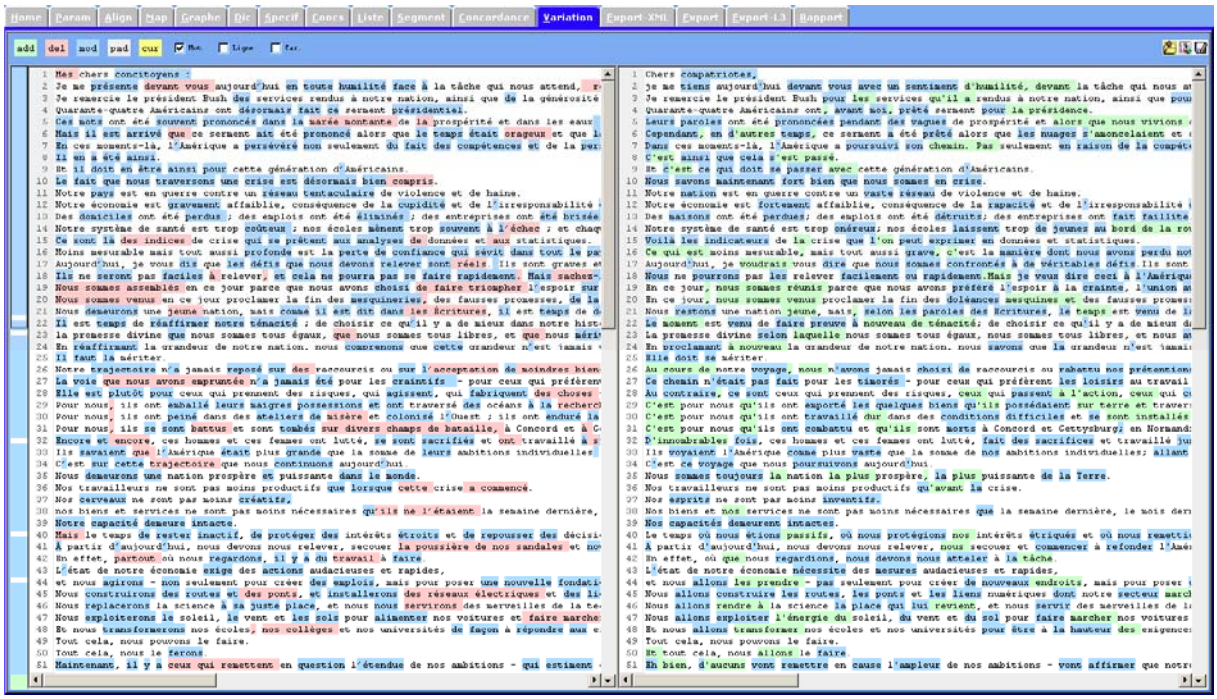


Figure 21 : Mise au jour de la variation (comparaison au niveau du mot)

[7] Compare deux fichiers et affiche les différences (*cf* http://fr.wikipedia.org/wiki/Diff)

[8] http://search.cpan.org/~mjcarman/Tk-DiffText-0.17/lib/Tk/DiffText.pm

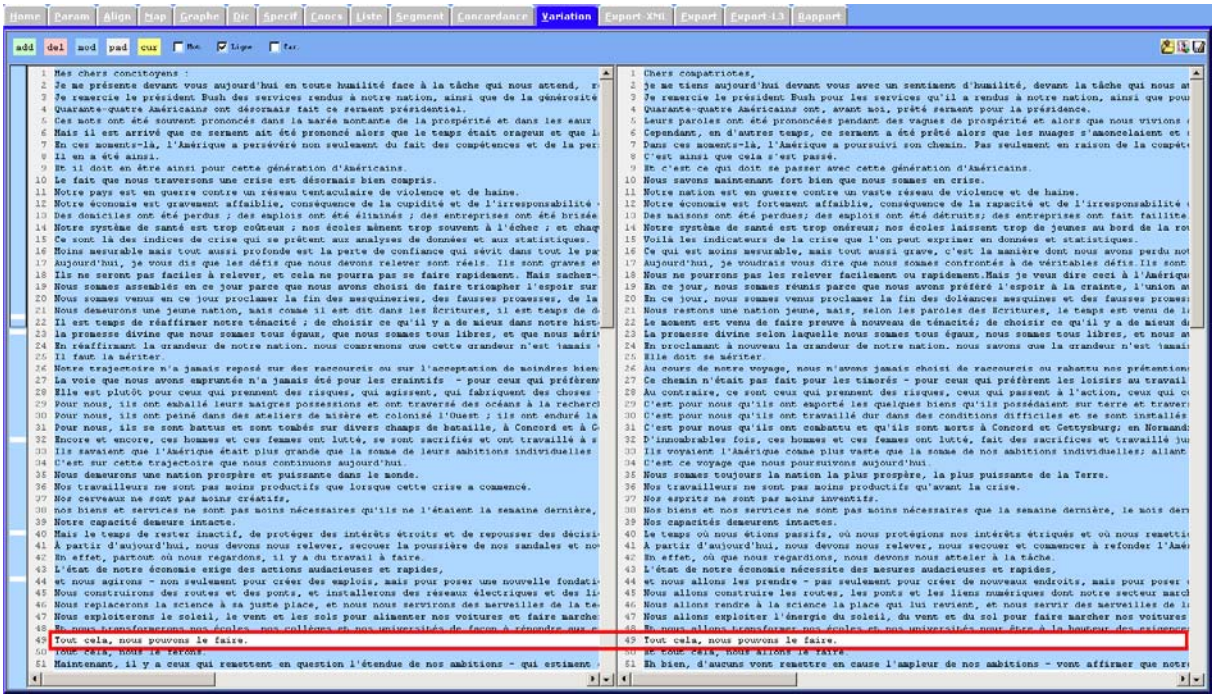


Figure 22 : Mise au jour de la variation (comparaison au niveau des lignes)

Pour ce module le texte source (à gauche) est considéré comme le texte de référence à partir duquel on mesure les différences. La coloration permet de mettre au jour :

- Les éléments supprimés dans le texte source (zones rouges dans le volet à gauche)
- Les éléments ajoutés dans le texte cible (zones vertes dans le volet à droite)
- Les éléments modifiés dans les 2 volets (zones bleues dans les 2 volets)
- Les éléments inchangés d'un volet à l'autre restant non colorés : dans la seconde comparaison, la seule ligne inchangée dans la partie visible à l'écran est cerclée de rouge.

Ce résultat est exportable au format HTML ; on trouve en ligne plusieurs illustrations de ces exports :

- Deux traductions du discours d'investiture de B. Obama :
  - export comparaison : http://tal.univ-paris3.fr/mkAlign/mkalign-variation/variation-obama-export.html
  - graphique de comptage de la variation http://tal.univ-paris3.fr/mkAlign/mkalign-variation/graph-variation-obama.jpg
- Deux discours de Ségolène Royal (campagne 2007) :
  - export comparaison (après alignement automatique) http://tal.univ-paris3.fr/mkAlign/mkalign-variation/variation-royal-export.html
- Deux discours de Nicolas Sarkozy (conférence de presse 2008) :
  - export comparaison (après alignement automatique) http://tal.univ-paris3.fr/mkAlign/mkalign-variation/variation-sarko-export.html

On peut aussi calculer des indicateurs de la variation (fond commun, mots ajoutés, supprimés, modifiés…) : le graphique produit donne à voir pour chaque section d'alignement un décompte des variations sur chaque section. On trouvera en ligne (*supra*) des exemples de telles sorties.

5 Bibliographie

Fleury Serge, **Zimina Maria**, "*Exploring Translation Corpora with* ***mkAlign***", in *Translation Journal*, Volume 11, n°1 January 2007.
http://accurapid.com/journal/39mk.htm

Fleury Serge, **Zimina Maria**, "*Utilisations de* ***mkAlign*** *pour la traduction philologique*" (PDF), in Actes JADT 2008, Journées Internationales d'Analyse Statistiques des Données Textuelles, Lyon, 2008.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/fleury-zimina.pdf
http://tal.univ-paris3.fr/mkAlign/Slides%20-%20JADT2008/
http://tal.univ-paris3.fr/mkAlign/Demo_mkAlign%20-%20JADT2008/

Leblanc Jean-Marc, **Martinez William**, "*L'analyse contrastive des réseaux de cooccurrence Le monde dans les discours des présidents de la Cinquième République*", in Actes JADT 2006, Journées Internationales d'Analyse Statistiques des Données Textuelles, Besançon, 2006.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/II-054.pdf

Martinez William, **Zimina Maria**, "*Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues*", in Actes JADT 2002, Journées Internationales d'Analyse Statistiques des Données Textuelles, St Malo, 2002.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/martinez_zimina.pdf

Véronis Jean, *Alignement de corpus multilingues*, in Pierrel, J.-M., éditeur, *Ingénierie des langues*, Informatique et systèmes d'information, chapitre 6, pages 151–172. Hermès Sciences, 2000.
http://www.up.univ-mrs.fr/~veronis/pdf/2000hermes6.pdf

Zimina Maria, *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles.* Présentation à la soutenance de thèse, Université de la Sorbonne nouvelle - Paris 3, le 26 novembre 2004.
http://www.cavi.univ-paris3.fr/ilpga/ED/student/stmz/ED268-PagePersoMZ_fichiers/stmz/page6_fichiers/26novembre_MZ.zip

Zimina Maria, *L'alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles.* Conférence aux 7es Journées internationales d'Analyse statistique des Données Textuelles *JADT'2004*, Louvain-la-Neuve (Belgique), 2004.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_118.pdf

Zimina Maria, *Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*, in *Actes des 7es Journées scientifiques du Réseau de chercheurs "Lexicologie, Terminologie, Traduction"*, Institut supérieur de traducteurs et interprètes (ISTI), Bruxelles, 8-10 septembre 2005.
http://perso.univ-lyon2.fr/~thoiron/JS%20LTT%202005/pdf/Zimina.pdf

Zimina Maria, *Corpus multilingues : exploration textométrique dans l'espace intertextuel*, in Ballard M., Pineira-Tresmontant C. (éd) *Les corpus en linguistique et en traductologie"* (p. 107-121), Artois Presses Université, 2007.

Equivalences traductionnelles

[Equivalences]

Maria Zimina

zimina@msh-paris.fr

Résumé : Les *Types* bilingues français/anglais ***administr+/administ+*** sont appariés en raison de leur parenté sémantique dans le corpus parallèle. Dans le bi-texte découpé en sections, leurs distributions respectives présentent des divergences. Une suite d'opérations textométriques permet de cerner les causes de ces discordances. On découvre deux phénomènes sensiblement différents : 1) Les asymétries sont dues au décalage dans l'alignement des sections ; 2) Il existe des contextes originaux où les mots français commençant par la chaîne ***administr+*** (*administration*, *administrer* etc.) ne sont pas traduits par des mots anglais commençant par la chaîne ***administ+*** (*administration*, *administering* etc.) et réciproquement. On en déduit deux méthodes de travail sur corpus parallèles : 1) Une méthode de synchronisation d'alignement phrastique à l'aide de la carte des sections bi-textuelle ; 2) Une méthode d'exploration bi-textuelle permettant le repérage de passages originaux où sont attestées des équivalences lexicales peu communes.

1 Contexte de la recherche

Le corpus ***Convention*** est constitué de textes juridiques français/anglais de la *Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales*, de ses protocoles intégraux, et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995. Deux versions de chaque document existent parallèlement ; il est difficile de distinguer une langue source et une langue cible. Ce corpus a été réuni dans le cadre d'une étude plus large qui avait pour objectif la construction d'un lexique bilingue des droits de l'homme à base de corpus parallèles alignés au niveau de la phrase (Bourigault *et al.*, 1999). Au cours du projet, le corpus ***Convention*** a été aligné semi automatiquement jusqu'au niveau du paragraphe. On estime le taux de précision du découpage en phrases à 90 % environ.

Le corpus compte 12 913 formes pour 296 396 occurrences dans le volet français et 9 530 formes pour 284 958 occurrences dans le volet anglais. La partition naturelle du corpus en 3 parties dont chacune correspond à un ensemble de documents juridiques d'un certain type amène les résultats que l'on peut voir au tableau 1.

Tableau 1

Structure du corpus ***Convention***

Corpus ***Convention***	volet français 296 396 occ.	volet anglais 284 958 occ.
Convention européenne des Droits de l'Homme	5 953 occ.	5 710 occ.
Protocoles intégraux de la Convention	8 984 occ.	8 773 occ.
Arrêts de la Cour Européenne des Droits de l'Homme	281 459 occ.	274 475 occ.

Les arrêts de la Cour européenne constituent la principale partie du corpus ***Convention***. On trouve un extrait du texte des arrêts en français et en anglais au tableau 2 ci-dessous.

Tableau 2

Convention : Arrêts de la Cour européenne des Droits de l'Homme (extraits)

volet français	volet anglais
<texte="fr"> § du côté gibraltarien de la frontière, les fonctionnaires des douanes et de la police en service normal ne furent ni informés ni associés à la surveillance, au motif que cela impliquerait que l'information soit communiquée à un trop grand nombre de personnes.	<texte="en"> § on the *gibraltar side of the border, the customs officers and police normally on duty were not informed or involved in the surveillance on the basis that this would involve information being provided to an excessive number of people.
<texte="fr"> § aucune mesure ne fut prise pour ralentir la file de voitures lors de leur entrée, ou pour examiner tous les passeports, car on craignait que cela puisse alerter les suspects.	<texte="en"> § no steps were taken to slow down the line of cars as they entered or to scrutinise all passports since it was felt that this might put the suspects on guard.
<texte="fr"> § une équipe de surveillance distincte se trouvait cependant à la frontière et un groupe préposé à l'arrestation était posté dans le secteur de l'aéroport voisin.	<texte="en"> § there was, however, a separate surveillance team at the border and, in the area of the airfield nearby, an arrest group.
<texte="fr"> § le témoin *m, qui dirigeait une équipe de surveillance postée à la frontière, exprima sa déception au vu du manque apparent de coopération entre les divers groupes impliqués à *gibraltar, mais il comprit que les choses étaient ainsi organisées pour des questions de sécurité.	<texte="en"> § witness *m who led a surveillance team at the frontier expressed disappointment at the apparent lack of co-operation between the various groups involved in *gibraltar but he understood that matters were arranged that way as a matter of security.

Guide de lecture du tableau 2 :

Dans cet extrait du corpus parallèle ***Convention***, plusieurs types de codage sont mis en évidence :

- *la clé <texte> texte qui distingue deux langues (français : "fr" , anglais : "en") ;*
- *le caractère § qui matérialise l'alignement des phrases ;*
- *le caractère * qui permet d'identifier des lettres (à l'origine) en majuscules.*

2 Asymétries distributionnelles des *Types* bilingues appariés

La confrontation des dictionnaires de formes graphiques constitués à partir de chacun des volets du corpus nous amène à nous interroger sur les particularités d'un ensemble de

vocabulaire associé dans les deux langues à la notion d'*administration* (en anglais : *administration*).

Nous allons constituer un type particulier, que nous appellerons ***administr+*** à partir de toutes les formes graphiques commençant par cette chaîne de caractères dans le volet français du corpus.[9] Puis, de la même façon, nous allons construire un deuxième type à partir de toutes les formes graphiques commençant par la chaîne ***administ+*** dans le volet anglais du corpus. *A priori*, on peut s'attendre à ce que ces entités soient liées sur le plan de la traduction.

Tableau 3

Convention : transformation pour une exploration parallèle sous *Lexico3*

```
§
<texte="fr"> aucune mesure ne fut prise pour ralentir la file de voitures
lors de leur entrée, ou pour examiner tous les passeports, car on craignait
que cela puisse alerter les suspects.

<texte="en"> _no _steps _were _taken _to _slow _down _the _line _of _cars
_as _they _entered _or _to _scrutinise _all _passports _since _it _was _felt
_that _this _might _put _the _suspects _on _guard.
§
```

Sur la figure 4, chacun des types ***administr+*** [478 occ.] et ***administ+*** [482 occ.] (français/anglais) est constitué par l'ensemble d'occurrences des formes graphiques regroupées en raison de leur parenté sémantique dans le corpus transformé pour une exploration parallèle sous ***Lexico3*** (voir l'extrait présenté au tableau 3) : [10]

[9] Sous ***Lexico3***, le langage des « expressions régulières » permet à l'utilisateur de constituer des groupes de mots correspondant au type de son choix et d'enregistrer la liste de ces unités pour une exploration ultérieure.

[10] Dans l'état actuel, les fonctionnalités de ***Lexico3*** ne permettent pas encore de charger séparément les dictionnaires de formes correspondant à chaque volet d'un corpus bi-textuel. Pour contourner cette difficulté, nous avons différencié les deux langues en introduisant le caractère « _ » (*underscore*) devant chaque forme graphique du volet anglais. Automatisée par une opération Rechercher/Remplacer, l'insertion de cette marque a permis d'éviter toute confusion entre les vocabulaires correspondant à chaque volet du corpus.

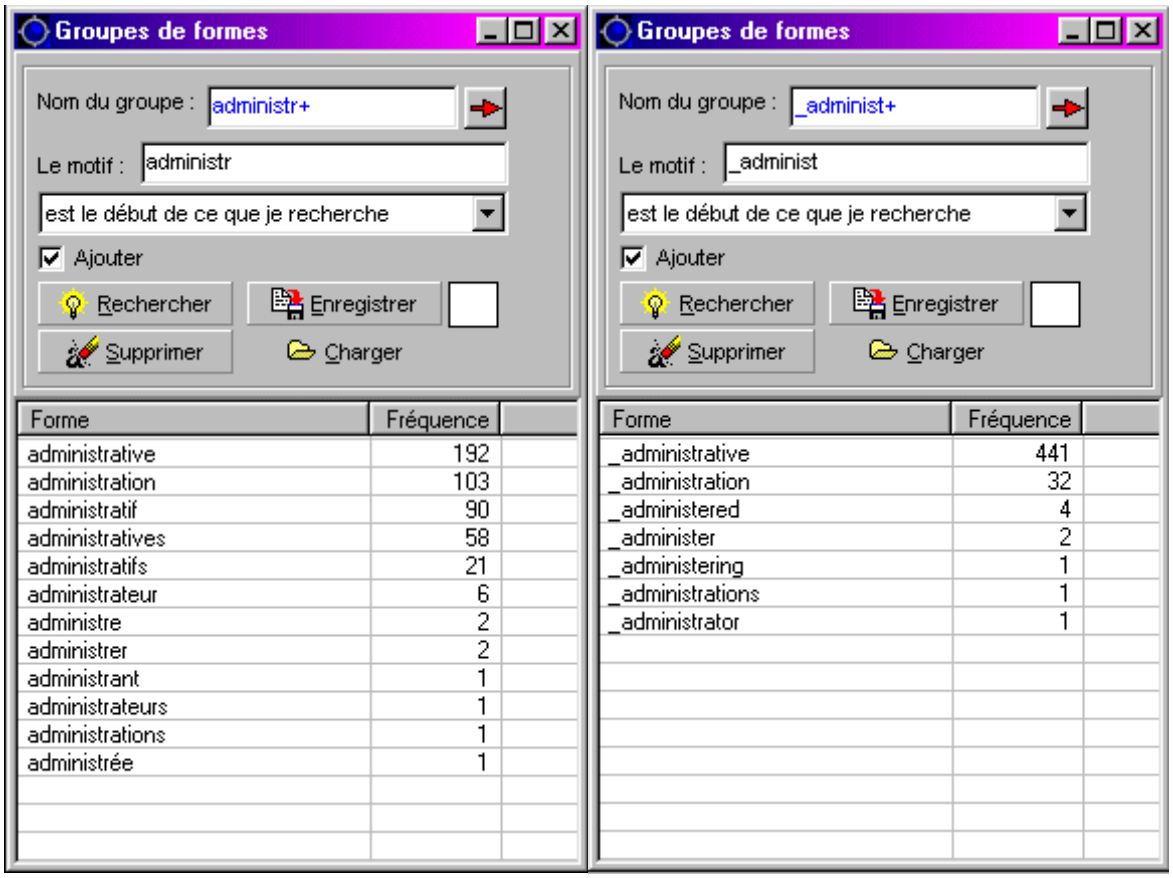


Figure 4

Sélection des *Types* bilingues pour une exploration parallèle

Afin de poursuivre notre exploration, nous allons créer une carte bi-textuelle en s'appuyant sur l'alignement des sections parallèles.[11]

[11] La mise en correspondance des parties équivalentes du corpus parallèle a été réalisée l'aide du logiciel ***mkAlign*** qui permet de construire ou de corriger un alignement de deux textes. L'outil permet de visualiser l'alignement en cours et de le modifier via un éditeur à double entrée (dans notre exemple, le caractère **§** sert de délimiteur de sections appariées). ***mkAlign*** donne la possibilité d'exporter l'alignement au format ***Lexico3***. Pour plus d'informations sur les fonctionnalités de cet outil, on consultera la documentation à l'adresse suivante :

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

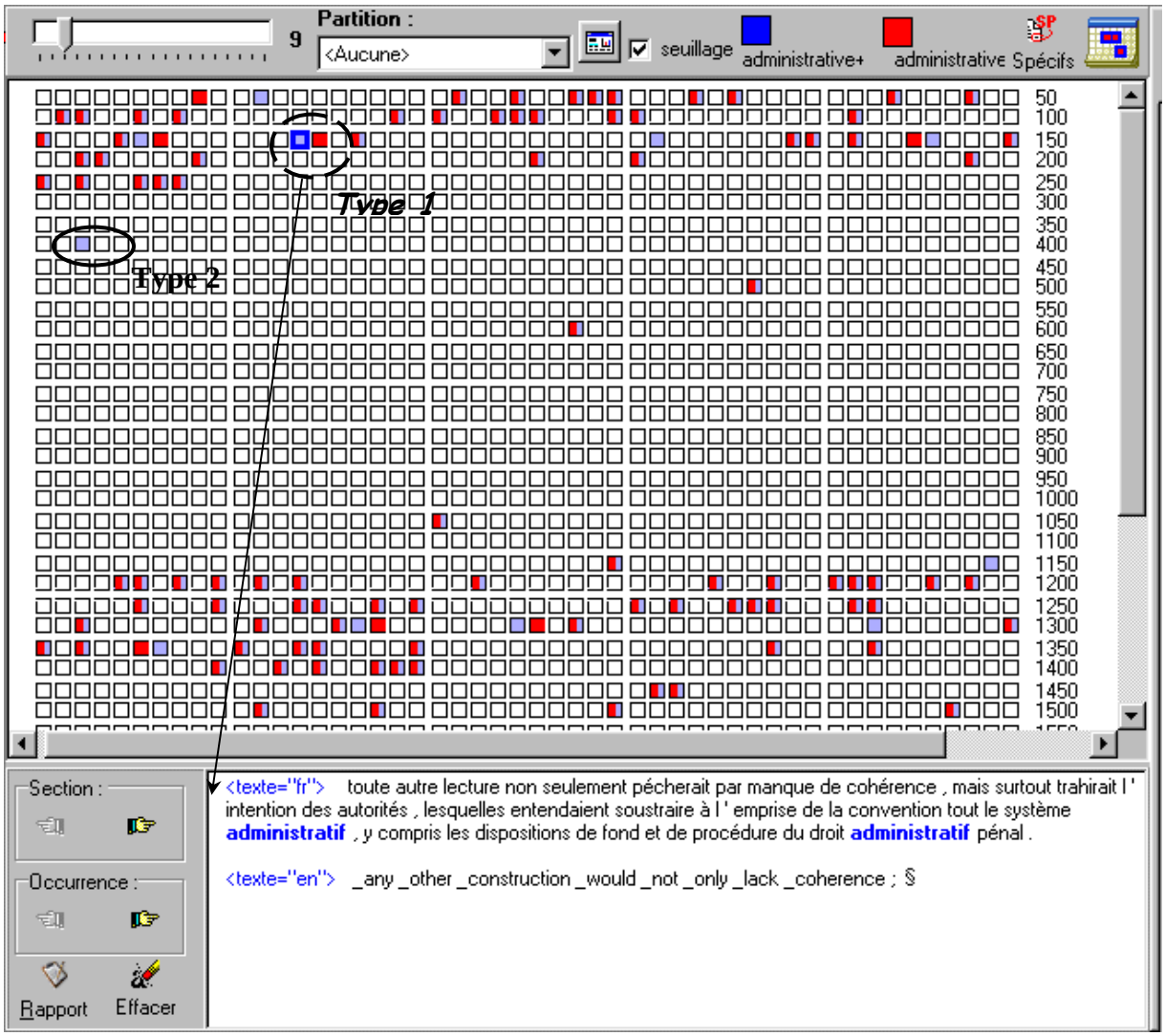


Figure 5

Ventilations des *Types* français/anglais ***administr+ / administ+*** dans le corpus aligné au niveau de la phrase : recherche d'asymétries distributionnelles

Guide de lecture de la figure 5 :

L'alignement des *sections* (phrases) du bi-texte est matérialisé par des carrés. Le coloriage des carrés indique la présence des *types* étudiés dans les sections concernées :

– les *carrés bicolores* de la carte signalent les sections bi-textuelles où les mots français commençant par la chaîne ***administr+*** (*administration*, *administrer* etc.) sont traduits par des mots anglais commençant par la chaîne ***administ+*** (*administration*, *administering* etc.).

– les *carrés monochromes* correspondent aux sections du bi-texte où le type français ***administr+*** et le type anglais ***administ+*** ne se correspondent pas dans le corpus. En cliquant sur un *carré monochrome* (bleu ou rouge), il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types ne sont pas liés. On peut ensuite étudier les particularités de ces contextes et trier entre les cas qui correspondent aux *décalages dans*

l'alignement des sections parallèles du corpus (***Type 1***) et les autres (susceptibles de révéler des équivalences lexicales peu communes – ***Type 2***).

Rappel sur les fonctionnalités de la carte des sections bi-textuelle

Pour étudier la ventilation des *types* sur la carte des sections, on procède de la façon suivante :

On sélectionne le *Tgen* (à partir du dictionnaire, du *Garde-mots*, de la liste des segments répétés, etc.) et on le fait glisser sur la carte (bouton gauche maintenu enfoncé).

On sélectionne la section à visualiser dans la fenêtre du bas en cliquant sur le carré qui la représente dans la carte des sections.

La case *seuillage* permet de régler deux seuils en probabilités qui entraîneront un coloriage (plus ou moins sombre) des sections.

Pour une représentation simultanée de deux *Tgen(s)*, ce processus doit être réitéré (en prenant soin de changer la couleur dans la boîte correspondante). Il faut maintenir la touche Control en position basse lors du second glisser/déposer.

La figure 5 montre la ventilation des types ***administr+*** / ***administ+*** dans les sections appariées du corpus. Une conclusion s'impose : dans le corpus ***Convention***, même si l'on peut constater des similitudes importantes qui concernent des parties équivalentes, les distributions des ces types présentent des divergences.

Ce constat amène une question : *Quelles sont les particularités des contextes où les mots français commençant par la chaîne* ***administr+*** *ne sont pas en correspondance avec des mots anglais commençant par la chaîne* ***administ+*** *?*

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

Type 1 : il existe des *décalages dans l'alignement* des sections parallèles du corpus, ce qui expliquerait la présence de sections bi-textuelles où les deux types ne sont pas en correspondance.

Type 2 : le type ***administr+*** n'est pas toujours traduit par le type ***administ+*** et il existe des contextes originaux, où sont attestées des équivalences lexicales peu communes, susceptibles d'intéresser le chercheur.

La figure 5 permet de trier entre les cas qui correspondent à la première hypothèse et les autres.

3 Résolution du problème

Les fonctionnalités de la carte des sections rendent possible une visualisation simultanée de la présence/absence des types bilingues. Comme indiqué sur la figure 5, la couleur bleu est utilisée pour matérialiser le type français ***administr+*** et le rouge pour le type anglais ***administ+***. En cliquant sur un *carré bicolore*, il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types sont liés. L'analyse de ces sections signale l'équivalence lexicale des types appariés :

volet français	volet anglais

<texte="fr"> les extraits du dossier **administr**atif que cite l'appelant à l'appui de sa thèse ne confortent toutefois pas cette affirmation.	**<texte="en">** the passages from the **administr**ative file which the appellant cites in evidence in this connection do not, however, support that assertion.

La présence de sections monochromes sur la carte montre qu'il existe des cas de non-correspondance entre les types. En cliquant sur un *carré monochrome* (bleu ou rouge), il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où les deux types ne sont pas liés :

volet français	volet anglais
<texte="fr"> toute autre lecture non seulement pécherait par manque de cohérence, mais surtout trahirait l'intention des autorités, lesquelles entendaient soustraire à l'emprise de la *convention tout le système **administr**atif, y compris les dispositions de fond et de procédure du droit administratif pénal.	**<texte="en">** any other construction would not only lack coherence;

Type 1

Lorsque deux sections monochromes coloriées en bleue et rouge se succèdent sur la carte, on peut généralement constater les décalages dans l'appariement des sections. Par exemple :

volet français	volet anglais
<texte="fr"> toute autre lecture non seulement pécherait par manque de cohérence, mais surtout trahirait l'intention des autorités, lesquelles entendaient soustraire à l'emprise de la *convention tout le système **administr**atif, y compris les dispositions de fond et de procédure du droit **administr**atif pénal.	**<texte="en">** any other construction would not only lack coherence;
<texte="fr"> cela vaudrait même dans le cas où, comme en l'espèce, l'accusé ne se voit infliger qu'une amende, dès lors qu'à défaut de paiement une peine d'emprisonnement s'y substitue.	**<texte="en">** it would also run counter to the authorities' intention, which had been to remove from the scope of the *convention the whole **administr**ative system, including the substantive and procedural provisions of **administr**ative criminal law. that would be so even in a case where, as in this instance, the accused was merely fined, in so far as default on payment of that fine would entail committal to prison.

Les erreurs de l'alignement initial peuvent être corrigées si l'on prend soin de sauvegarder les sections concernées dans un rapport. Pour ajouter une section au rapport, il suffit de cliquer sur l'icône *Rapport*  située en bas de la fenêtre de la carte des sections (cf. *Figure 5*).[12]

Type 2

La *présence isolée de sections monochromes* coloriées en bleu ou en rouge révèle des contextes originaux où les mots français commençant par la chaîne ***administr+*** (*administration*, *administratif*, etc.) ne sont pas traduits par des mots anglais commençant par la chaîne ***administ+*** (*administration*, *administrative*, etc.) et réciproquement.

La matérialisation de ces sections sur une carte représentant le corpus parallèle permet de dresser une véritable topographie bi-textuelle. Il devient possible d'isoler des contextes singuliers où sont attestées des équivalences lexicales originales, susceptibles d'intéresser l'expert humain pour la construction de ressources textuelle (cf. *Tableau 6*) :

- le recours **administr**atif ~ the non-contentious application
- l'**administr**ation des douanes ~ the customs
- bonne **administr**ation ~ good governance
- dépositions **administr**atives ~ provisions
- l'**administr**ation du district ~ district authority
- l'**administr**ation des eaux ~ water-rights authority
- procédures antérieures ~ earlier **administr**ative proceedings

Tableau 6

Convention : Contextes originaux repérés à l'aide de la topographie bi-textuelle

volet français	volet anglais
<texte="fr"> 1. le recours **administr**atif /.../	**<texte="en">** 1. the non-contentious application /.../
<texte="fr"> il prononça la confiscation des marchandises saisies et infligea aux prévenus une amende, assortie de la contrainte par corps, à payer à l'**administr**ation des douanes, partie poursuivante jointe et qui s'était constituée partie civile à l'audience.	**<texte="en">** the court also ordered confiscation of the goods seized and sentenced the defendants to pay a fine, with imprisonment in default, to the customs, which was a co-prosecutor and had also joined the proceedings as a civil party.
<texte="fr"> en pareil cas, le tiers peut aussi chercher à démontrer que le directeur a agi en violation d'un principe général de bonne **administr**ation (algemeen beginsel van behoorlijk bestuur).	**<texte="en">** in so doing, the third party may also base his claim of unlawfulness on the allegation that the *commissioner has acted in breach of a general principle of good governance (algemeen beginsel van behoorlijk bestuur).

[12] Les erreurs recensées dans l'alignement des sections bi-textuelles peuvent être corrigées à l'aide du programme ***mkAlign*** (Fleury, 2005).

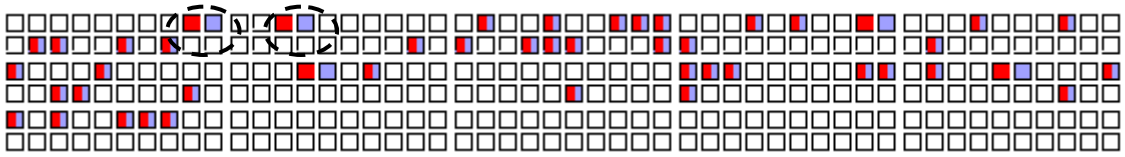
Tableau 6 (suite)

Convention : Contextes originaux repérés à l'aide de la topographie bi-textuelle

volet français	volet anglais
<texte="fr"> en outre, la réserve n'entre en jeu que lorsqu'ont été appliquées des dispositions **administr**atives de fond et de procédure d'une ou plusieurs des quatre lois qu'elle spécifie.	**<texte="en">** moreover, the reservation only comes into play where both substantive and procedural provisions of one or more of the four specific laws indicated in it have been applied.
<texte="fr"> il ressort des mémoires soumis par les parties à la procédure devant elle et des dossiers des procédures antérieures qu'une audience ne contribuera sans doute pas à éclaircir l'affaire.	**<texte="en">** it is apparent to the *court from the pleadings of the parties to the proceedings before it and from the files relating to the earlier **administr**ative proceedings that an oral hearing is not likely to clarify the case further.

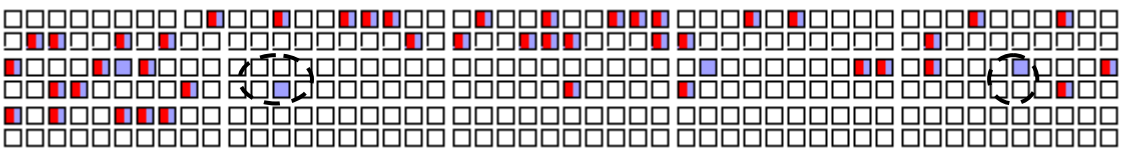
4 Une méthode de synchronisation de l'alignement

On pose l'équivalence de types bilingues issus de chaque volet du corpus parallèle aligné au niveau du paragraphe ou de la phrase. Le rapprochement des types peut être effectué en prenant en considération leur proximité sémantique ou thématique dans le corpus. On matérialise les distributions des types sur une carte des sections bi-textuelle. Si les distributions sont toujours parallèles mais très légèrement décalées dans certaines parties du corpus, les ruptures du parallélisme signalent le décalage dans l'alignement des sections. Les *paires de sections monochromes* voisines signalent généralement les passages où il existe des erreurs. Voici un diagramme sommaire réalisé à partir d'une telle ventilation :



5 Une méthode de repérage de passages originaux dans la traduction

On matérialise les distributions des types bilingues appariés sur une carte des sections bi-textuelle. Si les distributions se ressemblent, à quelques asymétries près, la *présence isolée de sections monochromes* montre le plus souvent des passages originaux dans la traduction où sont attestées des équivalences lexicales susceptibles d'intéresser le chercheur. Le diagramme d'une telle ventilation se présente de la façon suivante :



6 Conclusion

La démarche proposée permet de comprendre les raisons d'asymétries dans les distributions parallèles du vocabulaire bilingue correspondant aux *Types* appariés. La suite des opérations textométriques convoquées pour localiser les ruptures de parallélisme sur un diagramme représentant le bi-texte aligné constitue une méthode largement applicable à d'autres corpus pluritextuels.

A la phase de repérage direct, appuyée sur la topographie bi-textuelle, succède une phase de remise en contexte des particularités distributionnelles constatées. Cette dernière phase débouche sur une édition contrastée des erreurs d'alignement phrastique et de contextes originaux, où sont attestées des équivalences lexicales peu communes, difficiles à postuler *a priori*.

7 Références

Bourigault D., Chodkiewicz Ch., Humbley J. « Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné. », in *actes de la troisième conférence 'Terminologie et Intelligence Artificielle'*, Nantes, 1999.

Fleury S. « MKAlign », *documentation*. Paris : Université de la Sorbonne nouvelle – Paris 3, (Travaux du SYLED-CLA^{2}T, 2005), http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

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.

Zimina M. « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. », in *actes des 7emes journées d'analyse statistique des données textuelles*, Presses universitaires de Louvain, Louvain-la-neuve, 2004

Zimina M. *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de Doctorat en Sciences du langage. Université de la Sorbonne nouvelle – Paris 3, 2004.

Zimina M. « Exploration textométrique de corpus de traduction », in *actes du colloque « Pour une traductologie proactive » – META'50*, Presses de l'Université de Montréal, Montréal, 2005 (à paraître).

8 Fonctionnalités *Lexico3* utilisées dans cette navigation

N°	*Fonctionnalité*	*Résultat*
8.2	**Sélection d'un Type (occurrences de formes graphiques commençant par une chaîne de caractères définie)**	Figure 4
7	**Carte des sections (sections bi-textuelles, présence/absence des *Types* bilingues français/anglais *administr+* / *administ+*)**	***Figure 5***

Comparaisons textométriques de traductions franco-chinoises

[Traductions franco-chinoises]

Jun MIAO, André SALEM[13]

silaomiaomiao@yahoo.fr, salem@msh-paris.fr

Résumé : Après un bref rappel sur le système d'écriture chinoise et ses prises en charge par différents systèmes de codage informatique (§3), on compare les dépouillements textométriques d'un texte français et d'une de ses traductions chinoises. Après le dépouillement du texte chinois en caractères isolés (§4), on compare un dépouillement automatisé en mots de ce même texte avec le texte français original (§5). La dernière section est consacrée à l'étude des perspectives ouvertes par la démarche textométrique pour l'analyse des différentes traductions chinoises utilisées pour rendre un même mot français (§6).

Mots-clés : textométrie; caractères chinois (hanzi); littérature; traductologie.

Abstract : After a short recall of the Chinese writing system and on its various encoding systems (§3), the authors apply textometric methods to compare a French text with its Chinese translation. After an examination of the Chinese text with isolated characters (§4), the same text cut into words with a Chinese word separation program is compared with the French original (§5). The last section aims at studying the perspectives of textometric approaches in the analysis of different Chinese translations of French words (§6).

Key-words: Textometry, Chinese characters (Hanzi), literature, translation studies.

摘要:中国文字（书写）系统是基于汉字的一种古老语言誊写方式。在对此文字系统以及其现代各种信息编码作一简单描述后(§3) 作者运用词量法对法语著作及其中文翻译进行了比较与分析。首先运用中间加空隔的方式将中文翻译当中的汉字相互独立开来(§4),其次运用分词软件对这同一翻译进行了单词的自动切分并加与原文做对比(§5)。文章的最后一部分侧重于运用词量法对法语单词的各种汉语翻译的考察(§6)。

关键词：词量法；中文（汉字）；文学；翻译学

1 Contexte de la recherche

Parmi les nombreuses raisons qui peuvent expliquer le fait que les méthodes d'analyse des textes sur ordinateur, de plus en plus largement répandues dans le monde occidental, ne se sont développées que plus tardivement dans la sphère culturelle chinoise, on doit considérer les facteurs liés à l'existence d'un système d'écriture très ancien, dont certaines qualités sont indiscutables, mais dont l'informatisation s'est révélée beaucoup plus complexe que celle des systèmes basés sur l'utilisation d'un alphabet réduit.

Dans la période récente, parallèlement aux efforts entrepris par les linguistes chinois pour simplifier la représentation des caractères *hanzi*, les problèmes liés à l'informatisation des systèmes d'écritures complexes ont été dépassés par la mise en place de normes internationales (telle la norme *Unicode*) et de technologies permettant la saisie et l'affichage

[13] Les auteurs remercient Kim Gerdes, Serge Fleury et Cédric Lamalle pour leur aide et leurs conseils précieux dans la réalisation de ce travail.

de textes écrits dans des langues jusqu'alors difficilement accessibles au traitement sur ordinateur.

Ces avancées technologiques ouvrent la voie à un formidable développement des études textuelles appliquées à des gisements textuels que les codages traditionnels étaient incapables de prendre en charge. Au delà de l'exploration des corpus électroniques à des fins de recherches linguistiques ou sociolinguistiques, la fouille de données textuelles concerne dorénavant un très vaste ensemble de textes saisis dans le cadre d'activités entreprises dans tous les secteurs de la vie socio-économique d'un pays en plein développement.

L'étude de bitextes, dont l'un des volets est constitué par la traduction de l'autre, constitue une entrée privilégiée dans le domaine des études comparatives entre textes rédigés dans des langues différentes. Dans ce cas, en effet, les caractéristiques quantitatives calculées à partir de chacun des volets du corpus peuvent être directement utilisées pour cerner les différences entre les langues mises en présence. C'est ce que nous allons tenter de faire dans l'étude qui suit afin de poser les premiers jalons d'études traductologiques que nous nous proposons d'entreprendre par la suite.

2 Le système d'écriture chinois

Les écritures chinoise, japonaise et coréenne utilisent, toutes trois, les caractères ***Han***, caractères d'origine chinoise dits 汉字 *(hanzi)* en chinois, ainsi que des caractères nationaux propres à chacune des langues[14]. Le chinois possède, pour sa part, un système d'écriture qui n'est ni alphabétique, ni phonétique. On peut dire que chaque caractère correspond plus ou moins à un morphème et à une syllabe de l'oral.

Le nombre de *hanzis* différents utilisés par ces systèmes d'écriture se compte en milliers (parfois en dizaines de milliers) dépassant de très loin le nombres des lettres qui permettent de transcrire les écritures alphabétiques. On dit que pour lire un journal, un lecteur chinois doit pouvoir identifier sans mal 5 000 *hanzis* environ.

2.1 Les caractères chinois

Chaque caractère chinois est composé d'un certain nombre de *traits* que l'on peut retrouver dans une série d'autres caractères. Les caractères correspondent à la fois à un segment sonore, la syllabe, et à une unité de sens[15].

人 – rén /# *homme* #
大 (一+人) – dà /# *grand*; 天 (二 +人) – tiān /# *ciel* 1 ##

木 – mù /# *bois*; 林 (木+木) – lín, *fôret*; 森 (木+林) – sēng, *grande forêt*.

Chaque caractère véhicule une signification, mais ne constitue pas nécessairement à lui seul un mot. Certains caractères changent de sens dans la combinaison avec d'autres.

---

[14] Le Consortium Unicode et l'ISO considèrent que les caractères chinois, coréens et japonais sont les mêmes, que seuls les *glyphes* diffèrent. On peut rapprocher cette différence d'aspect des traditions différentes qui ont longtemps prévalu en allemand (police de caractères gothique), en français (police à sérifs) et en anglais (police sans sérifs). Les caractères sont codés de la même façon. Chaque tradition utilise une police appropriée pour afficher les caractères dans le style qui convient le mieux aux habitudes locales.
Après l'établissement de la République Populaire de Chine en 1949, les autorités ont entrepris des efforts pour simplifier les caractères chinois. En 1955, le Comité pour la Réforme de l'Écriture (*Wenzi gaige wei yuanhui*) a publié une proposition de caractères simplifiés. En 1964, il a publié une deuxième liste de simplifications. Cette dernière liste règle actuellement l'emploi des caractères chinois.

[15] Cf. , par exemple, [ALLETON 1997],. p.11-18.

东 – dōng /#Est/# #西 – xī /#Ouest># 东西 –#dōngxī /#chose1##

2.2 Les mots chinois

C'est la combinaison de deux caractères ou parfois de trois caractères qui constitue le mot.

你 *nǐ, tu, toi*

好 *hǎo, bon, bien*

你好！ *nǐhǎo! Bonjour! Comment ça va?*

Dans la langue moderne, il existe beaucoup de mots bi-syllabiques, voire tri-syllabiques. Par suite de l'évolution de la langue et de l'adoption de mots empruntés à d'autres langues. Par exemple :

(1)	(2)	(3)	(4)
手,	手机	邂逅	巧克力
shǒu	shǒu jī	xiě hòu	qiǎo kè lì
main	portable	rencontre par hasard	chocolat

Dans le premier exemple, le caractère 手 (shǒu) signifie *main,* il constitue une syllabe et correspond en même temps à un sens indépendant. Dans ce cas, il peut être considéré comme un mot.

Dans le deuxième exemple, 手机, le même caractère est associé au caractère 机 (jī, *machine, appareil*) Il garde dans ce cas le sens *main*, mais la combinaison des deux caractères prend un nouveau sens : *téléphone mobile, portable*.

Dans le troisième exemple, la combinaison des deux caractères 邂逅 (xiè hòu) signifie *se rencontrer par hasard*, mais ces caractères perdent leur sens lorsqu'il sont isolés.

Dans le mot 巧克力 (anglais *chocolate*), chacun des caractères 巧, 克, 力 possède un sens propre sans rapport immédiat avec le mot (巧: *adroite*, *habile*; 克: *convaincre*; 力: *force*).

Produit courant, 茉莉花茶 （mò lì huā chā, *le thé au jasmin*) est un mot, dont les composants identifiables sont difficiles à segmenter. On peut considérer 茉莉 (mò lì, *jasmin*) comme un mot bi-syllabique composé de deux caractères dépourvus de sens propre. Mais en combinaison avec le caractère 花 (huā, *fleur*), le mot qui désigne toujours le jasmin, renvoie à la fleur de l'arbuste. On peut considérer le caractère 茶 (chā, *thé*) comme un mot monosyllabique. Mais précédé par le caractère 花 (huā, *fleur*), on peut également considérer que les caractères combinés 花茶 (huā chā, *thé aux fleurs*) qui sont différents de 绿茶 (lǜ chá, *thé vert*) ou de 红茶 (hóng chá, *thé noir*) forment un nouveau mot.

2.3 Les phrases et la ponctuation

Comme dans le cas des mots, il est difficile de définir clairement les limites de la phrase chinoise. Les définitions et les classifications de la phrase que l'on trouve dans les grammaires chinoises (phrases énonciatives, interrogatives, impératives, exclamatives, etc.) permettent difficilement de segmenter un texte en phrases de manière automatisée.

La ponctuation est d'usage récent en chinois. En 1919, on a commencé à utiliser la ponctuation moderne en se référant au système de ponctuation occidental. Le système utilisé

actuellement conserve la trace des réformes successives de l'écriture chinoise. C'est pourquoi la ponctuation chinoise moderne, malgré ses similarités avec celle utilisée en occident, reste distincte de cette dernière.

L'utilité des repères liés à la notation de la ponctuation chinoise est d'autant plus importante que, comme on s'en souvient, les mots (ou plutôt les caractères) chinois sont écrits l'un après l'autre sans être séparés par des espaces. [16]

3 Le codage informatique des caractères chinois

En raison de leur nombre élevé et contrairement à ce qui se passe pour les systèmes d'écriture des langues qui utilisent un alphabet restreint, les caractères chinois ne peuvent être représentés à l'aide d'un codage sur un seul octet. La norme *Unicode* qui permet de représenter chaque caractère sur plusieurs octets fournit une bonne solution pour représenter les caractères chinois.[17].

3.1 Logiciels supportant le traitement de textes chinois.

Dans leurs versions récentes, les logiciels de traitement de textes permettent de manipuler, en plus des textes codés en unicode qui vont rapidement constituer la norme, des polices multioctets qui permettent d'afficher correctement les textes chinois (entre autres écritures non latines). Avec le logiciel Word[18], par exemple, lorsqu'on tente d'enregistrer un texte chinois, avec l'option texte seulement une boite de dialogue permet de sélectionner le codage `Chinois simplifié (GB2312)` comme on peut le voir sur la figure 1.

3.2 Lexico3 et les textes chinois

Dans ses versions actuelles (3.5.0.2), ***Lexico3*** manipule des chaînes de caractères codés sur un seul octet. Cette limite, qui est en voie d'être dépassée[19], n'entraîne cependant pas l'impossibilité de traiter des chaînes de caractères codées sur plusieurs octets. Comme on comprend, en les comparant octet par octet, il est possible de conclure que deux chaînes de caractères multioctets sont identiques ou qu'elles sont différentes. De plus, les systèmes informatiques modernes permettent d'afficher correctement certaines représentation multioctets qui ne sont pas des représentations unicode .

Pour le présent travail, nous avons utilisé un codage **`Chinois simplifié . Mainland China`** proposé par le logiciel ***Word***. On prend en charge ce codage sous Lexico3 en activant l'article `Chinois simplifié.Mainland China` proposé par le menu `Options` (couteau suisse) de ***Lexico3***.

Les composants utilisés dans ***Lexico3*** (Edition du texte, Concordances, Carte des sections, etc.) affichent ce codage correctement lorsqu'on choisit de le visualiser avec le codage `Chinois GB2313` des navigateurs :

---

[16] A l'instar de très nombreux systèmes d'écriture parmi lesquels ceux de l'antiquité (latin, grec, hébreu, sumérien, etc.).

[17] Un grand nombre de systèmes d'écriture occidentaux, dont le système du français ont utilisé jusqu'à une date récente le code ASCII (127 caractères), puis le code ASCII étendu (255 caractères) qui permettait de coder en outre les voyelles accentuées du français.

[18] Nous avons utilisé, pour cette étude, la version 2003 du logiciel ***Word*** distribué par Microsoft.

[19] Plusieurs versions de la série *Lexico*, en cours d'achèvement, permettent déjà de traiter les chaînes de caractères unicodes. Le logiciel *MKAlign*, développé par S. Fleury dans l'équipe Syled-Cla2t permet également de traiter les textes encodés sous ces formats.

Bouton droit -> Codage -> Plus ->. Chinois simplifié (GB2312)

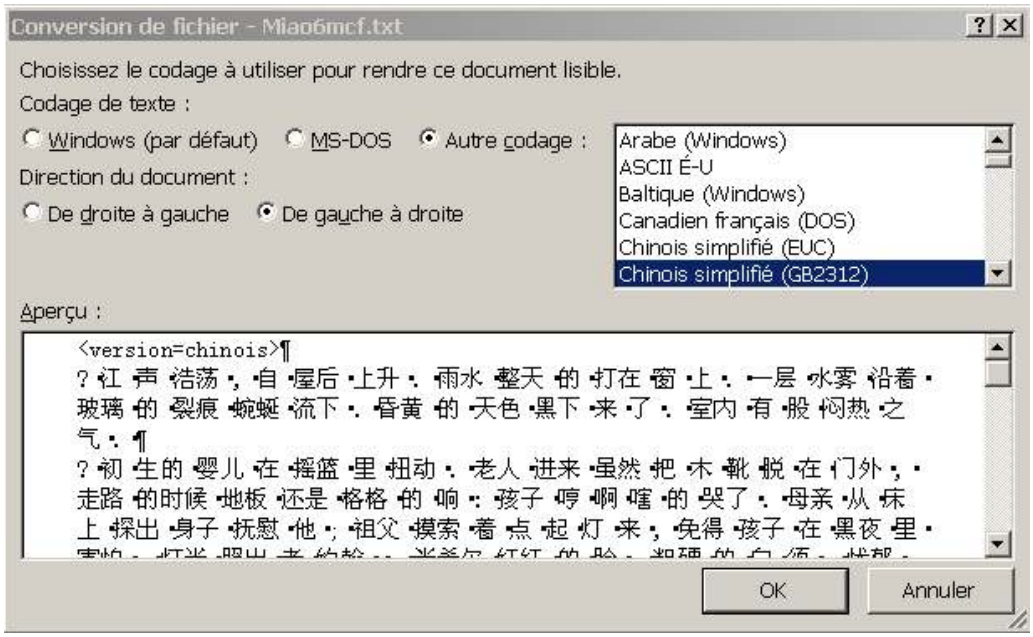


Figure 1 :

Word 2003 : Paramétrage de l'enregistrement du texte

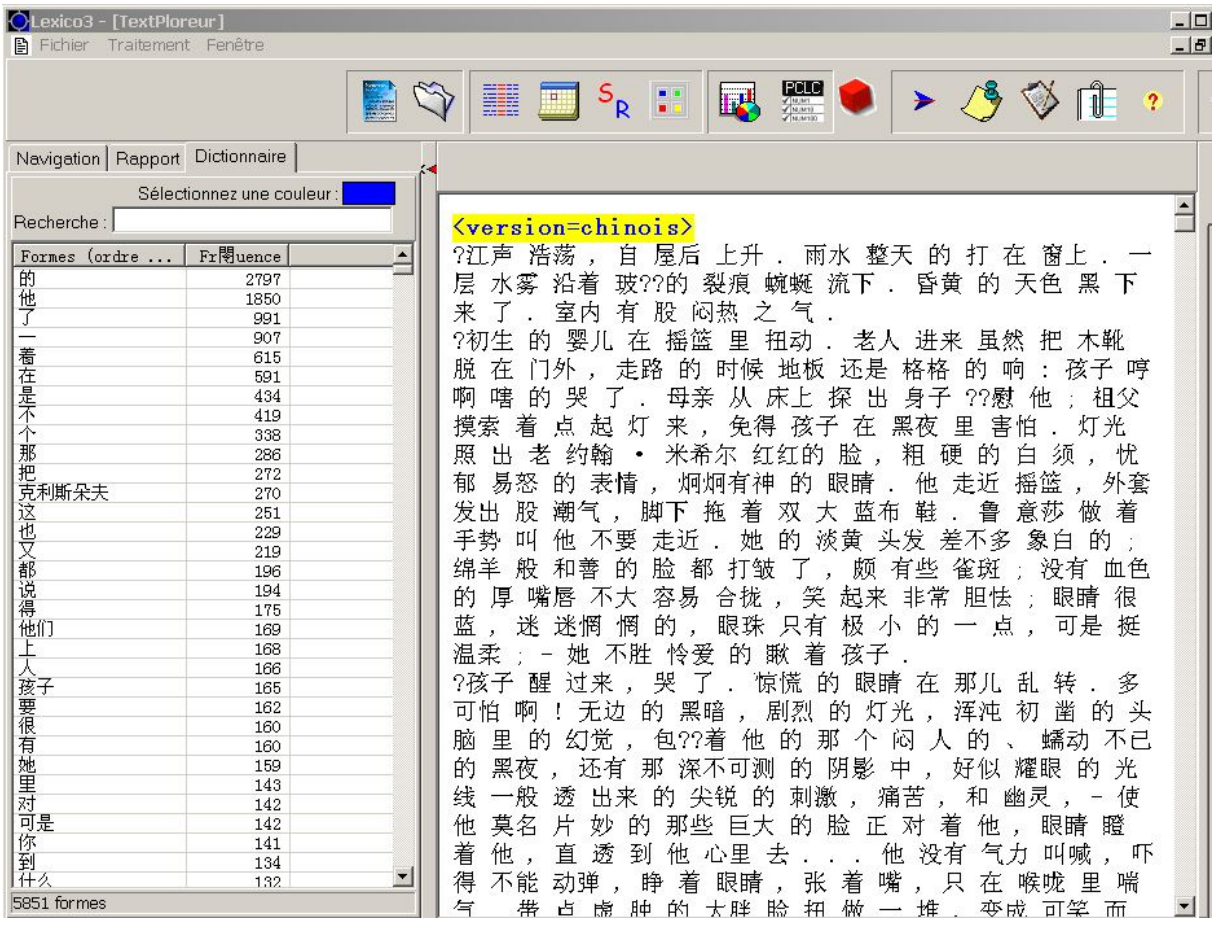


Figure 2 :
Lexico 3 : Affichage du texte avec le codage « Chinois simplifié (GB2312) »

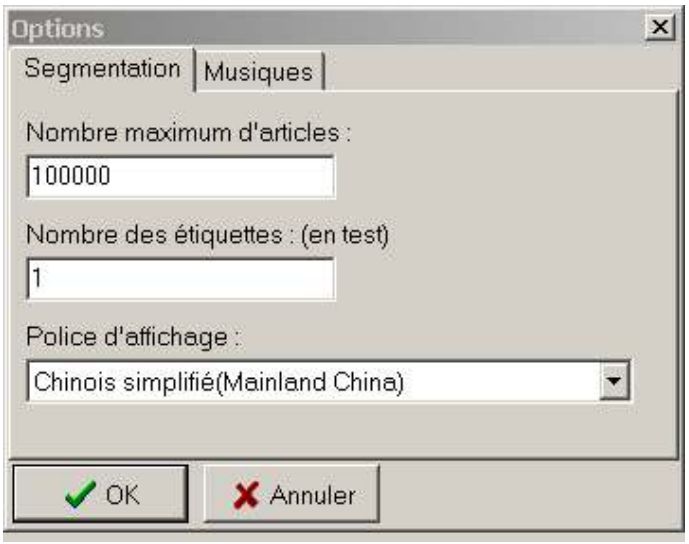


Figure 3 :
Lexico 3 : Fenêtre de réglage du paramètre « encodage des caractères »

4 Un corpus d'application

Pour illustrer ces possibilités, nous utiliserons un corpus bilingue dont le volet français est constitué par le premier chapitre du roman *Jean-Christophe* publié en 1904 par Romain Rolland (1866-1944). On trouve, au tableau 1, ci-dessous un extrait du texte original de Romain Rolland. Le second volet du corpus est constitué par la traduction de ce texte en chinois par Fu Lei (1908-1966). Nous appellerons respectivement ces deux corpus ***JC1-Français*** et ***JC1-Chinois.***

4.1 Segmentation du texte en caractères

Comme on l'a vu plus haut, sans que cela constitue une gêne pour le lecteur expérimenté, le système d'écriture chinois n'utilise pas d'espace entre les unités lexicales placées côte à côte. Cette circonstance constitue une difficulté spécifique pour l'exploitation textométrique des textes chinois.

Sur quels critères peut-on s'appuyer pour découper des unités statistiques au fil du texte afin de réaliser des comparaisons entre textes ? Pour cette première analyse, nous nous appuierons sur une segmentation automatique, relativement facile formaliser et à mettre en œuvre sur un ordinateur, qui isole chaque caractère *hanzi*.[20]

[20] Pour réaliser cette segmentation en caractères, nous avons remplacé, en utilisant pour cela une expression régulière, chaque caractère du texte de départ par ce même caractère précédé d'un espace (code ASCII 32). Le fichier ainsi modifié réalise l'isolation de tous les caractères du corpus. Une procédure de ce type est disponible à l'adresse : http://www.cavi.univ-paris3.fr/

Tableau 1

*Extrait des corpus **JC1-Fr** et **JC1-Chin**,*

Le grondement du fleuve monte derrière la maison. La pluie bat les carreaux depuis le commencement du jour. Une buée d'eau ruisselle sur la vitre au coin fêlé. Le jour jaunâtre s'éteint. Il fait tiède et fade dans la chambre.

Le nouveau-né s'agite dans son berceau. Bien que le vieux ait laissé, pour entrer, ses sabots à la porte, son pas a fait craquer le plancher : l'enfant commence à geindre. La mère se penche hors de son lit, afin de le rassurer ; et le grand-père allume la lampe en tâtonnant, pour que le petit n'ait pas peur de la nuit. La flamme éclaire la figure rouge du vieux Jean-Michel, sa barbe blanche et rude, son air bourru et ses yeux vifs. Il vient près du berceau. Son manteau sent le mouillé ; il traîne en marchant ses gros chaussons bleus. Louisa lui fait signe de ne pas s'approcher. Elle est d'un blond presque blanc ; ses traits sont tirés ; sa douce figure mouton est marquée de taches de rousseur ; elle a des lèvres pâles et grosses, qui ne parviennent pas à se rejoindre et qui sourient avec timidité ; elle couve l'enfant des yeux – des yeux très bleus, très vagues, où la prunelle est un point tout petit, mais infiniment tendre.

§ l'enfant s'éveille et pleure. son regard trouble s'agite. quelle épouvante ! les ténèbres, l'éclat brutal de la lampe, les hallucinations d'un cerveau à peine dégagé du chaos, la nuit étouffante et grouillante qui l'entoure, l'ombre sans fond d'où se détachent, comme des jets aveuglants de lumière, des sensations aiguës, des douleurs, des fantômes : ces figures énormes qui se penchent sur lui, ces yeux qui le pénètrent, qui s'enfoncent en lui, et qu'il ne comprend pas - il n'a pas la force de crier ; la terreur le cloue immobile, les yeux, la bouche ouverts, soufflant du fond de la gorge. sa grosse tête boursouflée se plisse de grimaces lamentables et grotesques ; la peau de sa figure et de ses mains est brune, violacée, avec des taches jaunâtres.

Romain Rolland, *Jean-Christophe,* 1904

第一部

江声浩荡, 自屋后上升. 雨水整天的打在窗上. 一层水雾沿着玻**璃**的裂痕蜿蜒流下. 昏黄的天色黑下来了. 室内有股闷热之气.

初生的婴儿在摇篮里扭动. 老人进来虽然把木靴脱在门外, 走路的时候地板还是格格的响: 孩子哼啊嗐的哭了. 母亲从床上探出身子**抚**慰他;祖父摸索着点起灯来, 免得孩子在黑夜里害怕. 灯光照出老约翰·米希尔红红的脸, 粗硬的白须, 忧郁易怒的表情, 炯炯有神的眼睛. 他走近摇篮, 外套发出股潮气, 脚下拖着双大蓝布鞋. 鲁意莎做着手势叫他不要走近. 她的淡黄头发差不多象白的;绵羊般和善的脸都打皱了, 颇有些雀斑;没有血色的厚嘴唇不大容易合拢, **笑**起来非常胆怯;眼睛很蓝, 迷迷惘惘的, 眼珠只有极小的一点, 可是挺温柔;-她不胜怜爱的瞅着孩子.

孩子醒过来, 哭了. 惊慌的眼睛在那儿乱转. 多可怕啊!无边的黑暗, 剧烈的灯光, 浑沌初凿的头脑里的幻觉, 包围着他的那个闷人的*蠕动不已的黑夜, 还有那深不可测的阴影中, 好似耀眼的光线一般透出来的尖锐的刺激, 痛苦, 和幽灵, 使他莫名片妙的那些巨大的脸正对着他, 眼睛瞪着他, 直透到他心里去. . . 他没有气力叫喊, 吓得不能动弹, 睁着眼睛, 张着嘴, 只在喉咙里喘气. 带点虚肿的大胖脸扭做一堆, 变成可**笑**而又可怜的怪样子;脸上与手上的皮肤是棕色的, 暗红的, 还有些黄黄的斑点.

Traduction chinoise par Fu Lei, 1957[21]

Le tableau 1 montre un extrait du texte original suivi de sa traduction chinoise.

[21] Nous avons utilisé la version complète, réunie en 1957 par les Éditions Littéraires Populaires (人民文艺出版社), à partir d'une révision par Fu Lei de la première version de 1953.

La figure 4 montre, dans la fenêtre de droite, l'affichage par ***Lexico 3*** du texte chinois dans lequel les caractères ont été isolés par insertion d'un caractère espace entre chaque caractère. Dans la fenêtre de gauche on peut lire le résultat du dépouillement statistique réalisé sur la base du décompte des caractères isolés. Les caractères sont triés par ordre de fréquence décroissante dans le corpus analysé.

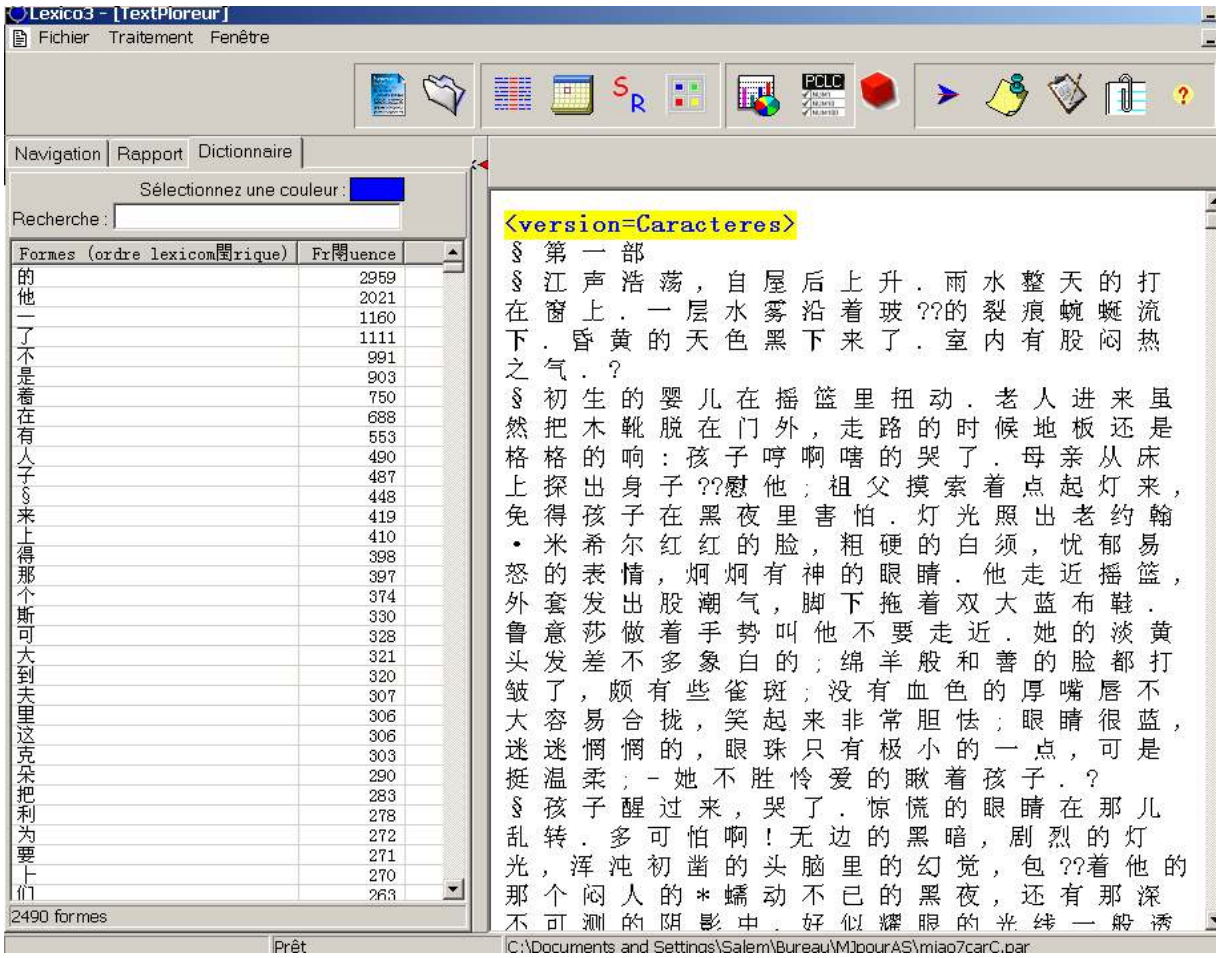


Figure 4

Exploitation avec ***Lexico3*** du texte chinois découpé en caractères

Le texte ainsi modifié va nous permettre d'obtenir un premier dépouillement en caractères (*hanzi*) du volet chinois du corpus. On peut voir les principales caractéristiques quantitatives de ce dépouillement au tableau 2.

Tableau 2

Principales caractéristiques quantitatives résultant du dépouillement en caractères (*hanzi*) du volet chinois du corpus

Partie	NB de caract.	Caract. différents	hapax	FMax	
Caractères	56 797	2 478	579	2 959	的

Le tableau 2 montre que les 56 797 caractères que compte le corpus ***JC1-Chin*** sont des occurrences de 2 478 hanzis différents. Un quart environ de ces caractères, soit 579, ne trouvent qu'une seule occurrence dans le corpus. Le caractère le plus fréquent est le caractère 的 (qui correspond plus ou moins à la préposition *de* en français).

La figure 5, qui rend compte de l'apparition de nouveaux caractères au fur et à mesure que l'on parcourt le texte, permet de préciser ces observations. La courbe d'accroissement réalisée à partir des caractères *hanzis* montre qu'on atteint, dès les 5 000 premiers caractères du texte le seuil de 1 000 caractères différents. Les 5 000 caractères suivants n'apportent que 500 nouveaux *hanzis*. Comme dans le cas des courbes d'accroissement de vocabulaire constituées à partir des mots, les tranches successives apportent de moins en moins d'unités nouvelles. Dans le cas des *hanzis* cependant on peut remarquer que l'accroissement initial est plus fort que dans le cas des courbe d'accroissement réalisées à partir d'unités lexicales (cf. § 5, infra).

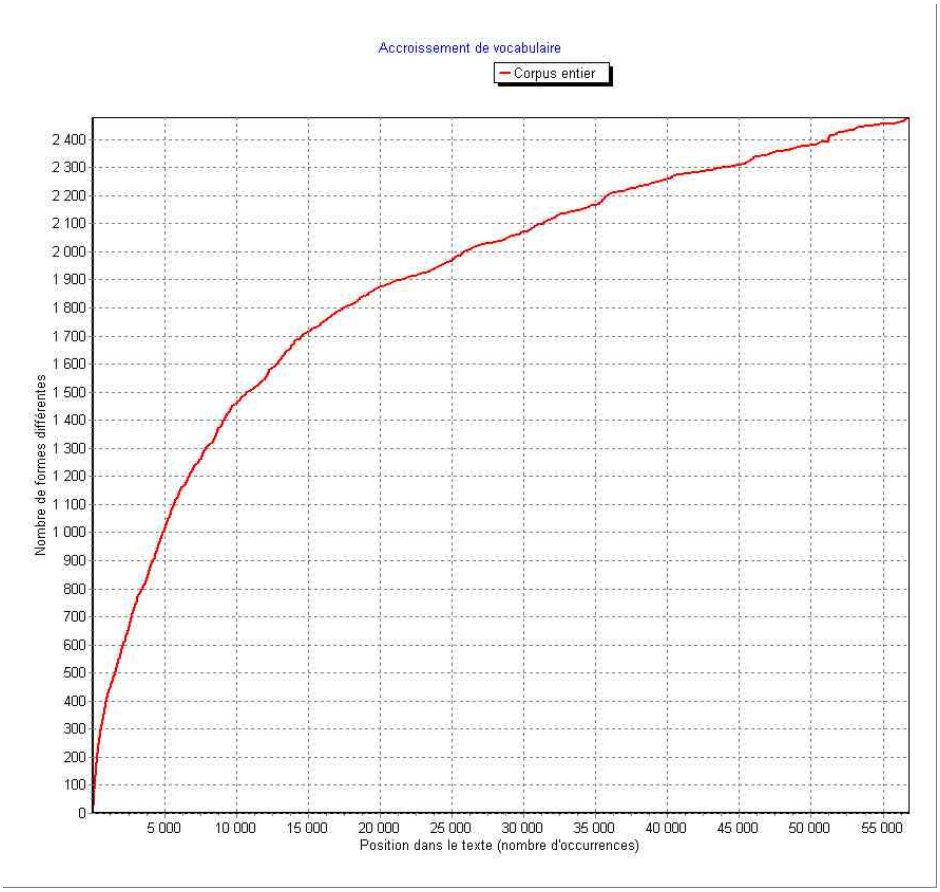


Figure 5

Apparition progressive des caractères dans le volet chinois.

4.2 Segmentation automatique en « mots »

Certains professionnels du Traitement Automatique des Langues proposent sur le web des procédures qui permettent de découper un texte chinois en « mots ». Dans cette section, nous utiliserons un découpage automatique en mots réalisé par un logiciel de segmentation spécialement conçu pour les textes chinois[22]. On peut voir au tableau 3 le résultat de cette segmentation en mots réalisée à partir de l'extrait de texte présenté au tableau 1.

[22] Pour cette première étude, nous avons utilisé le logiciel 海量智能分词研究版 (*Hailanda Segmentation intelligente* - version d'essai) réalisé par le Centre d'intelligence artificielle *Hailanda*, disponible à l'adresse suivante : http://www.mydown.com/code/234/234301.html . En plus de la segmentation, ce logiciel réalise une catégorisation des mots du texte orientée vers la recherche d'information technico-commerciale. Nous n'avons pas utilisé cette catégorisation pour notre étude. Il existe d'autres logiciels de segmentation du chinois, que l'on peut trouver sur l'Internet : ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System), et

Tableau 3

Extrait du volet chinois ***JC1-Chin*** segmenté en mots

(Chaque mot isolé par le logiciel *Hailanda* est suivi d'un blanc)

第一 部

江 声 浩荡 ， 自 屋后 上升 ． 雨水 整天 的 打在 窗 上 ． 一层 水雾 沿着 玻璃 的 裂痕 蜿蜒 流下 ． 昏黄 的 天色 黑下 来 了 ． 室内 有 股 闷热 之气 ． 初 生的 婴儿 在 摇篮 里 扭动 ． 老人 进来 虽然 把 木 靴 脱 在 门外 ， 走路 的时候 地板 还是 格格 的 响 ： 孩子 哼 啊 嗐 的 哭了 ． 母亲 从 床上 探出 身子 抚慰 他 ； 祖父 摸索 着 点 起 灯 来 ， 免得 孩子 在 黑夜 里 害怕 ． 灯光 照出 老 约翰 • 米希尔 红红 的 脸 ， 粗硬 的 白 须 ， 忧郁 易怒 的 表情 ， 炯炯有神 的 眼睛 ． 他 走近 摇篮 ， 外套 发出 股 潮气 ， 脚下 拖着 双 大 蓝布 鞋 ． 鲁 意 莎 做着 手势 叫 他 不要 走近 ． 她的 淡黄 头发 差不多 象 白的 ； 绵羊 般 和善 的 脸 都 打 皱 了 ， 颇有 些 雀斑 ； 没有 血色 的 厚 嘴唇 不大 容易 合拢 ， 笑 起来 非常 胆怯 ； 眼睛 很 蓝 ， 迷 迷惘 惘 的 ， 眼珠 只有 极小 的 一点 ， 可是 挺 温柔 ； － 她 不胜 怜爱 的 瞅着 孩子 ．

孩子 醒过来 ， 哭了 ． 惊慌 的 眼睛 在 那儿 乱转 ． 多可 怕 啊 ！ 无边 的 黑暗 ， 剧烈 的 灯光 ， 浑沌 初 凿 的 头脑 里 的 幻觉 ， 包围着 他的 那个 闷人 的 * 蠕动 不已 的 黑夜 ， 还有 那 深不可测 的 阴影 中 ， 好似 耀眼 的 光线 一般 透出 来的 尖锐 的 刺激 ， 痛苦 ， 和 幽灵 ， － 使 他 莫名 片 妙 的 那些 巨大 的 脸 正对着 他 ， 眼睛 瞪着 他 ， 直 透 到 他 心里 去 ． ． ． 他 没有 气力 叫喊 ， 吓得 不能 动弹 ， 睁着 眼睛 ， 张 着 嘴 ， 只 在 喉咙 里 喘气 ． 带 点 虚肿 的 大 胖 脸 扭 做 一堆 ， 变成 可笑 而 又 可怜 的 怪样子 ； 脸上 与 手上 的 皮肤 是 棕色 的 ， 暗红 的 ， 还有 些 黄黄 的 斑点

5 Comparaisons quantitatives à partir des *mots*

Les comptages réalisés à partir des mots ainsi découpés par l'algorithme de segmentation permettent de comparer les résultats obtenus sur le texte chinois à ceux que l'on obtient de la même manière sur la version française du texte.

Tableau 4

Principales caractéristiques quantitatives du dépouillement en *mots*
réalisé sur les volets français ***JC1-Fr*** et chinois ***JC1-Chin*** du corpus.

Partie	Occurrences	Formes	Hapax	F. Max	
JC1-Chin	34 743	7 196	3 781	2313	的
JC1-Fr	39 666	6 673	3 970	1578	de

Comme on le voit au tableau 4, la traduction chinoise compte nettement moins de mots graphiques que le texte français. On notera qu'elle compte cependant nettement plus de formes différentes. La proportion des formes qui n'apparaissent qu'une seule fois dans chacun

---

Java Lucene segmentation du chinois, etc.. Signalons qu'en raison de l'absence d'un consensus sur la définition de ce qu'est un mot chinois, aucun logiciel ne peut prétendre fournir de résultats parfaits.

des textes est moindre dans le texte chinois alors que la forme la plus fréquente y trouve nettement plus d'occurrences que dans le texte français[23].

La comparaison entre le système des mots chinois et celui des caractères chinois, pour lequel nous avons présenté plus haut des décomptes comparables montre que les mots chinois sont composés en moyenne de 1,6 caractères et que le mot le plus fréquent rassemble presque toutes les occurrences du caractère le plus fréquent (dans les deux cas le caractère : 的, *de*).

Tableau 5

Les formes les plus fréquentes pour chacun des volets du corpus

	Français		Chinois	
1	de	1 578	2313	的
2	il	1 044	1581	他
3	et	1 034	638	了
4	le	908	373	在
5	la	841	368	是
6	les	575	276	夫
7	Il	515	275	朵
8	se	463	274	克利斯
9	lui	448	235	把
10	des	447	208	着
11	ne	439	204	也
12	un	407	184	他的
13	en	399	158	又
14	que	394	156	孩子
15	pas	376	147	他们
16	qui	375	143	都
17	son	362	142	可是
18	dans	329	139	来
19	une	314	139	个
			136	她

La comparaison entre les formes les plus fréquentes dans chacun des volets du corpus montre que les fréquences décroissent plus rapidement dans le volet chinois. L'étude comparée des

[23] Il nous a semblé intéressant de publier ces premiers comptages sur la comparaison textométrique entre textes chinois et textes français. Cependant, ces résultats présentés dans le but de fournir une comparaison sur deux systèmes d'écriture très différents doivent être pris avec de grandes précautions. Nous étudierons par la suite l'influence que peut avoir la lemmatisation de chacune des listes de formes sur les résultats produits de la sorte (ainsi par exemple, la fréquence de la forme chinoise la plus fréquente 的 2313 occ. renvoie à la forme française *de* 1578 occ . mais aussi aux formes *du* 243 occ., *des* 447 occ., etc.).

courbes d'accroissement du vocabulaire, figure 6, précise les résultats obtenus par la comparaison des principales caractéristiques lexicométriques des volets français et chinois du corpus. La courbe située dans le haut du graphique correspond à l'enrichissement du vocabulaire chinois au fil du texte. Le fait que ce texte comporte moins d'occurrences est responsable de l'interruption de la courbe correspondante (abscisse 34 743) avant la courbe qui correspond au texte français (abscisse 39 666). La courbe correspondant à l'apparition de nouveaux mots chinois est située, dès que l'on atteint le premier tiers du corpus, largement au-dessus de celle qui correspond à l'apparition des mots français, ce qui confirme l'existence d'un plus grand nombre de formes en chinois.

On peut remarquer que des paliers créés par le ralentissement de l'accroissement du vocabulaire au cours du récit peuvent être mis en rapport d'une courbe à l'autre. Au ralentissement qui survient sur la courbe correspondant au texte français (abscisse 20 000) correspond un ralentissent dans la traduction chinoise (abscisse 17 000). A celui qui survient pour le texte français (abscisse 32 500) correspond également un ralentissent dans la traduction chinoise (abscisse 28 000).

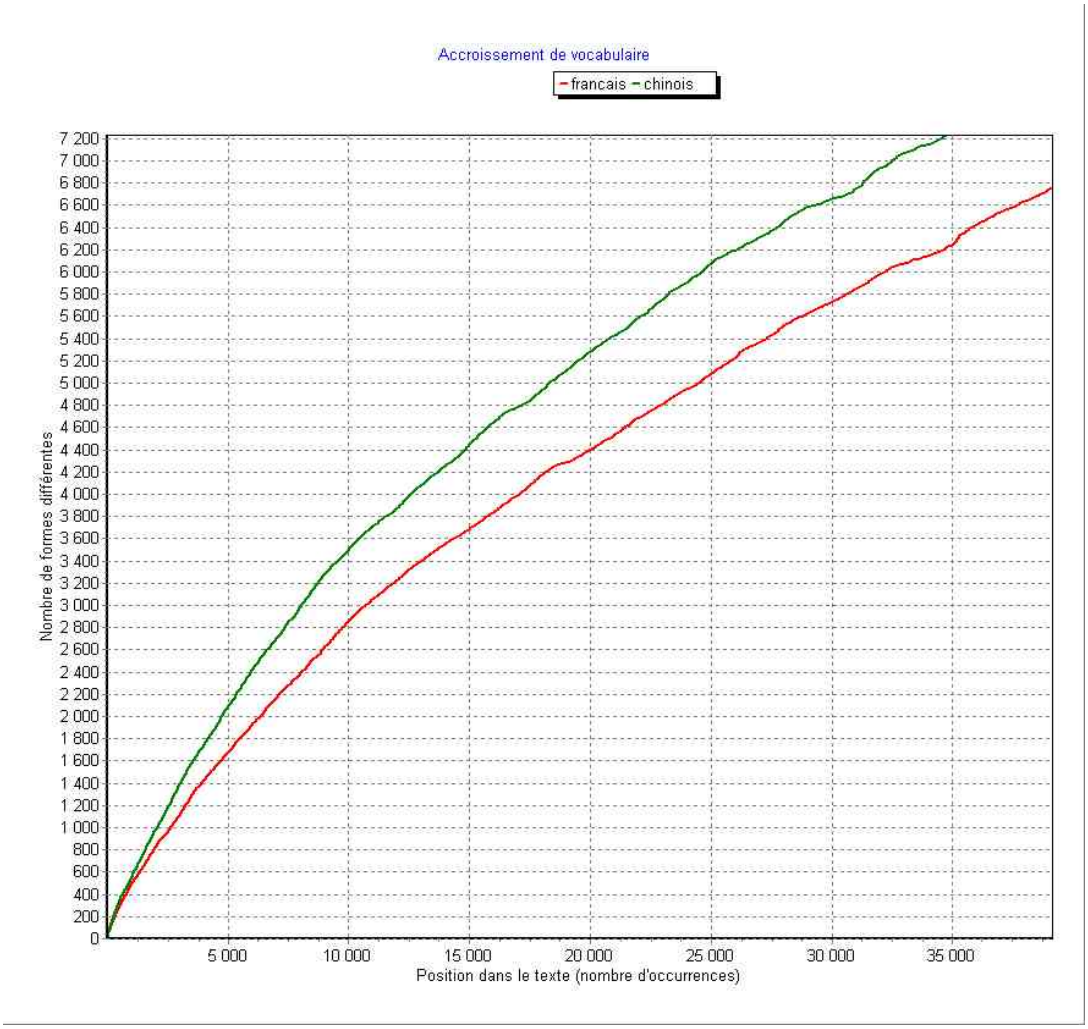


Figure 6

Courbes d'accroissement du vocabulaire réalisée
sur les volets français ***JC1-Fr*** et chinois ***JC1-Chin*** du corpus.

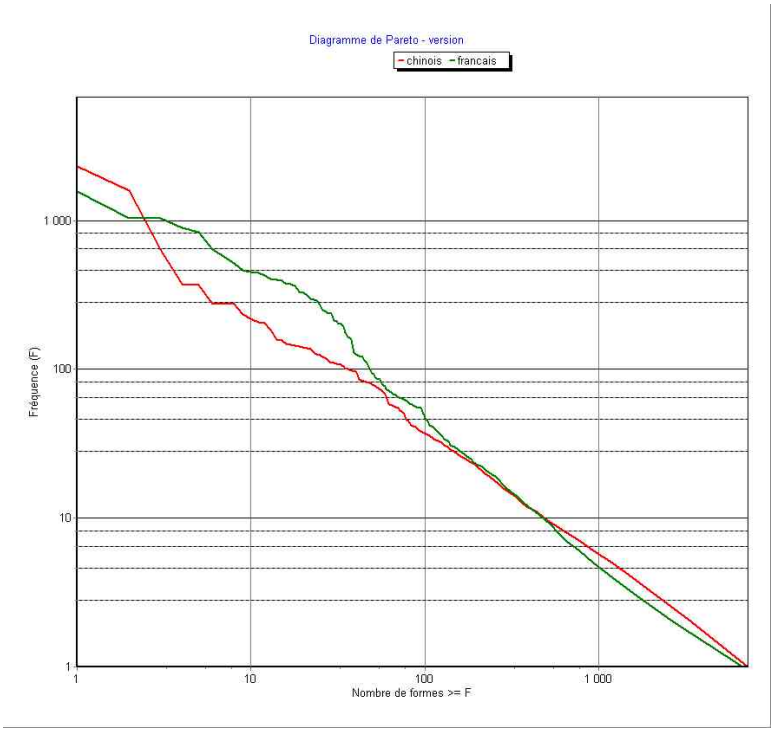


Figure 7 :
Diagramme de Pareto pour les deux volets du corpus

==== Guide de lecture pour la figure 7 ====

Pour un texte ***T*** dépouillé en unités statistiques appelées *formes*, le ***Diagramme de Pareto*** 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*

Pour en savoir plus :

Zipf, GK (1935), *The Psychobiology of Language, an introduction to Dynamic Philology*, Boston, Houghton-Mifflin.

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

La comparaison des deux courbes fait apparaître des différences assez nettes dans la structure des gammes de fréquences des deux textes. Le texte français possède nettement plus de

formes dans la zone de fréquences qui s'étend de 50 occurrences à 1000 occurrences environ. De son côté, le chinois crée plus de formes différentes dans la zone des très basses fréquences.

6 Un exemple d'étude parallèle

Aligner un bitexte, c'est construire une représentation qui met en correspondance des unités textuelles en rapport de traduction mutuelle. Le tableau 6 montre un alignement des deux volets du bitexte réalisé à partir du corpus ***JC1*** au niveau du paragraphe[24].

A partir d'un tel alignement on peut s'intéresser aux traductions de ce qui constitue une unité dans la langue source dans l'autre volet du corpus. Cette comparaison peut être menée simultanément du point de vue distributionnel , à l'aide de l'outil concordance (cf. tableau 7) et d'un point de vue *spatial* (cf. figure 8).

6.2 Le groupe vieux/vieillard *et son correspondant* 老人 (lao ren)

A titre d'exemple, nous examinerons les traductions chinoises d'un ensemble de mots qui rendent en français le concept de ***vieillesse*** : *vieux, vieillard,* etc.[25] Pour cette famille de mots, nous obtenons une fréquence globale de 95 occurrences qui se répartissent comme suit :

vieux 77, vieille 7, vieil 3, vieillard 3, vieilles 2, vieillards 1, vieillissait 1, vieillots 1.

On trouve au tableau 7 un extrait de concordance réalisée autour du pôle 老(lao, *vieux*), dont les lignes sont triées par ordre d'apparition dans le texte chinois. La localisation des occurrences de chacun de ces termes dans la carte des sections établie pour le texte français (figure 8) permet de repérer des sections correspondantes du texte chinois dans lesquelles on peut s'attendre à ce que soit rendue, en chinois, l'idée de *vieux*. La liste des mots les plus spécifiques dans le texte chinois qui correspond à ces dernières sections, nous laisse penser que le concept *vieux, vieillard, etc.,* est souvent rendu en chinois par les termes 老人 (lao ren, *vieil homme*) et 老(lao, *vieux*) qui constituent par ailleurs les équivalences traductionnelles les plus adaptées pour traduire le concept de *vieux*.

Dans une seconde étape, nous introduisons les mots 老人 et 老 sur la carte des sections découpées à partir du texte chinois. La comparaison des deux volets montre que la correspondance est loin d'être parfaite. On a rassemblé dans le tableau 8 des paires, sélectionnées à partir du concept français *vieux*, qui se trouvent être en rapport de traduction avec des expressions chinoises. L'analyse des discordances dans la localisation de ces formes révèle avant tout un écart entre le champ sémantique du mot français *vieux* et celui du *hanzi* chinois 老 (lao, *vieux, ancien, etc.*). En français, le mot *vieux* possède un lien étroit avec l'âge et le temps, mais il véhicule aussi une valeur parfois péjorative lorsqu'il s'applique à des objets ou à des personnes dans certains contextes (*vieux vêtements, vieille caisse*). En chinois, tout au contraire, le mot 老, dont le champ sémantique est un peu plus large, est employé pour désigner des personnes anciennes, respectables, honorables (老师 *professeur,* 老师傅 *vieux maître*).

---

[24] Cet alignement a été réalisé en utilisant le logiciel MKAlign proposé par Serge Fleury.. ce logiciel peut être téléchargé sur le site : http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm.

[25] Maria Zimina-Poirot a étudié dans sa thèse [Zimina 2004] des correspondances traductionnelles de ce type. Les logiciels de textométrie permettent désormais l'étude systématique de ce genre de correspondances traductionnelles. Les termes de la correspondance peuvent être étendus par l'utilisation du système des expressions rationnelles. Dans notre cas, le motif : *vie[iu]* permet de localiser toutes les occurrences des formes détaillées plus haut.

Tableau 6 :

Alignement en paragraphes sur les deux volets du corpus

§ le grondement du fleuve monte derrière la maison. la pluie bat les carreaux depuis le commencement du jour. une buée d'eau ruisselle sur la vitre au coin fêlé. le jour jaunâtre s'éteint. il fait tiède et fade dans la chambre.	§ 江声 浩荡， 自 屋后 上升. 雨水 整天 的 打 在 窗上. 一 层 水雾 沿着 玻璃 的 裂痕 蜿蜒 流下. 昏黄 的 天色 黑 下来 了. 室内 有 股 闷热 之 气.
§ le nouveau-né s'agite dans son berceau. bien que le vieux ait laissé, pour entrer, ses sabots à la porte, son pas a fait craquer le plancher : l'enfant commence à geindre. la mère se penche hors de son lit, afin de le rassurer ; et le grand-père allume la lampe en tâtonnant, pour que le petit n'ait pas peur de la nuit. la flamme éclaire la figure rouge du vieux jean-michel, sa barbe blanche et rude, son air bourru et ses yeux vifs. il vient près du berceau. son manteau sent le mouillé ; il traîne en marchant ses gros chaussons bleus. louisa lui fait signe de ne pas s'approcher. elle est d'un blond presque blanc ; ses traits sont tirés ; sa douce figure mouton est marquée de taches de rousseur ; elle a des lèvres pâles et grosses, qui ne parviennent pas à se rejoindre et qui sourient avec timidité ; elle couve l'enfant des yeux – des yeux très bleus, très vagues, où la prunelle est un point tout petit, mais infiniment tendre.	§ 初生 的 婴儿 在 摇篮 里 扭动. 老人 进来 虽然 把 木靴 脱 在 门外， 走路 的 时候 地板 还是 格格 的 响 : 孩子 哼 啊 嗐 的 哭 了. 母亲 从 床上 探 出 身子 抚慰 他 ; 祖父 摸索 着 点 起 灯 来， 免 得 孩子 在 黑夜 里 害怕. 灯光 照 出 老 约翰 • 米希尔 红红的 脸， 粗 硬 的 白 须， 忧郁 易怒 的 表情， 炯炯有神 的 眼 睛. 他 走近 摇篮， 外套 发出 股 潮气， 脚下 拖 着 双 大 蓝布 鞋. 鲁 意莎 做 着 手势 叫 他 不要 走近. 她 的 淡黄 头 发 差不多 象白 的 ; 绵羊 般 和善 的 脸 都 打皱 了， 颇 有些 雀斑 ; 没有 血色 的 厚 嘴唇 不大 容易 合拢， 笑 起来 非 常 胆怯 ; 眼睛 很 蓝， 迷 迷惘 惘 的， 眼珠 只有 极 小 的 一 点， 可是 挺 温 柔 ; -她 不胜 怜爱 的 瞅 着 孩子.
§ l'enfant s'éveille et pleure. son regard trouble s'agite. quelle épouvante ! les ténèbres, l'éclat brutal de la lampe, les hallucinations d'un cerveau à peine dégagé du chaos, la nuit étouffante et grouillante qui l'entoure, l'ombre sans fond d'où se détachent, comme des jets aveuglants de lumière, des sensations aiguës, des douleurs, des fantômes : ces figures énormes qui se penchent sur lui, ces yeux qui le pénètrent, qui s'enfoncent en lui, et qu'il ne comprend pas - il n'a pas la force de crier ; la terreur le cloue immobile, les yeux, la bouche ouverts, soufflant du fond de la gorge. sa grosse tête boursouflée se plisse de grimaces lamentables et grotesques ; la peau de sa figure et de ses mains est brune, violacée, avec des taches jaunâtres.	§ 孩子 醒 过来， 哭 了. 惊慌 的 眼睛 在 那儿 乱 转. 多 可怕 啊 ! 无边 的 黑 暗， 剧烈 的 灯光， 浑沌 初 凿 的 头脑 里 的 幻觉， 包围 着 他 的 那 个 闷 人 的 、 蠕动 不已 的 黑夜， 还有 那 深不 可测 的 阴影 中， 好似 耀眼 的 光线 一 般 透 出来 的 尖锐 的 刺激， 痛苦， 和 幽灵， -使 他 莫名 片 妙 的 那些 巨大 的 脸 正 对 着 他， 眼睛 瞪 着 他， 直 透 到 他 心里 去 . . . 他 没有 气力 叫喊， 吓 得 不能 动弹， 睁 着 眼睛， 张 着 嘴， 只 在 喉咙 里 喘气. 带 点 虚 肿 的 大胖 脸 扭 做 一 堆， 变成 可笑 而 又 可怜 的 怪 样子 ; 脸上 与 手上 的 皮 肤 是 棕色 的， 暗红 的， 还 有些 黄 黄 的 斑点.

Pour rendre le sens vaguement péjoratif associé en français à *vieux vêtement*, il faut, en chinois, avoir recours à d'autres mots. La traduction mot à mot en chinois de : *vieux rideau* et *vieille caisse* ne signifierait pas forcément, que les objets considérés sont en mauvais état mais soulignerait simplement leur ancienneté, sans liaison explicite avec leur état au moment

du récit. Fu Lei emploie 破 (po, *abîmé, déchiré*) et 破旧 (po jiu, *abîmé, usé, déchiré, etc.*) pour rendre accessible aux lecteurs chinois le sens original.

Tableau 7

Extrait de la concordance autour du pôle **老** (lao, *vieux*)

着 他的 要求 哼 一??歌词 没有 意义 的 **老调** 。 父亲 觉得 那种 音乐 是 胡闹 ； 可是 克利斯 那儿 摇晃 。 瘦削 的 树 好似 奇形怪状 的 **老人** 。 路旁 界石 上 的 反光 ， 象 青灰色 的 ?， 尤其是 把 人家 的 敬意 看得 很重 的 **老人** 。 他们 常常 跟 他 说 些 过火 的 笑话 ， 而 一想 到 就 觉得 心灰意冷 。 # 可怜 的 **老人** ！ 在 无论 哪 方面 ， 他 都不能 完全 表露 党?他所 有的 小计 划 ， 仿佛 他们 俩 是 **老朋友** ； 他 说 他 怎样 想做 一个 象 哈斯莱 那样 ?不会 说的 吧 ？ … … — — （ 他 指着 **老人** ） — — 瞧 ， 祖父 就在 那边 。 我 真爱 它们 象 牛 ， 象 巨人 ， 象 帽子 ， 象 **老婆婆** ， 象 广漠 无垠 的 风景 。 他 和 它们 低声 忱?， 快活 得 脸红 了 。 比他 更快 活的 **老人** ， 装着 若无其事 的 声音 和 他 说 （ 因为 器具 和 动物 的 尸身 ， 裹着 大氅 ， 象 **老太太** 般 ， 一边 庄严 的 前进 ， 一边 行 着 礼 低 的 吼 着 。 孩子 一个又一个 的 听 上 **老** 半天 ， 听 它们 低下 去 ， 没有 了 ； 它们 的时候 。 往往 你得 不 声 不 响的 等 个 **老** 半天 ， 正当 克利斯 朵 夫 想着 “ 他 今晚 但 就 因为 厌恶 ， 反而 常常 要看 。 他 **老** 半天 的 瞪着 它们 ， 不时 向 四下里 溜 一眼 贵族 ??生 的 家长 出来 散步 。 那时 他 **会老** 半天 的 停下来 ， 深深 的 鞠躬 ， 说着 一大 ： # “ 噢 ！ 祖父 ！ 祖父 ！ … … ” # **老人** 把 他 拉到 身边 。 他 扑在 老人 膝上 ， 峡?罢 。 ” # “ 那 也 该 回来 啦 ， “ **老人** 不高兴 的 说 。 # 他 踌躇 了 一会 ， 很不 命运 。 他 尤其 为 一个 美人 儿 颠倒 ， **不老** 不少 的 年纪 ， 金黄 的 长发 ， 大得 有点 所 教 的 东西 了 。 给 骂 了 一顿 ， 他 **老大** 不愿意 的 继续 下去 。 这样 当然 招来 了 ， 他 没有 ， ” 鲁 意 莎 抢着 回答 。 # **老人** 瞅着 她 ， 她 把 眼睛 躲开 了 。 # “ 哼 发愁 ， 时时刻刻 从 窗 里 张望 。 终于 **老人** 出现 了 ， 他们 俩 动身 了 。 他的 心 在 似的 告诉 他 ， 说 有些 东西 给 他 看 。 **老人** 打开 书桌 ， 检出 一本 乐器 放在 钢琴 上 岁 的 姑娘 ， 腮 帮 通红 ， 非常 壮健 ， **老** 带着 笑容 。 奥蒂 丽 的 长处 正好 和 克拉拉 十八世纪 的 雕 有 人像 的 柜子 ； 那 是 **老人** 从来不 肯 割爱 的 ， 虽然 古董 商 华姆塞 问道 ： # “ 那末 您 呢 ， 祖父 ？ ” # **老人** 打了 个 寒噤 。 # “ 什么 ？ “ 他 问 。 有心 装做 对 故事 的 下文 满不在乎 ， 使 **老人** 大为 难过 。 — — 但 眼前 他 是 完全 给 ” # 孩子 迷 迷 忽忽 的 ， 对着 灯光 和 **老人** 的 目光 愣住 了 ， 这 时 才 醒过来 ， 哭了 也 做 过 这些 东西 ？ ” # “ 当然 ， ” **老人** 的 声音 有点儿 不高兴 。 # 说完 他 不做 时候 常常 带着 他 一块儿 去 。 孩子 拉着 **老人** 的 手 在旁边 急急忙忙 的 搬 着 小步 。 他们 夜里 ， 还能 看出 他 憔悴 的 脸 ， 好似 **老人** 的 一样 。 她 开始 ??睡 了 ， 乱哄哄 的 吃了一惊 。 大家 一起 笑了 ； 大公 爵 向 **老人** 道贺 ， 他却 慌 做 一团 ， 想 解释 又 解释 ， 影子 的 头 会 爬上 去 ， 过后 又 回到 **老** 地方 ； 口 环 变得 很大 ， 象个 破 气球 ， 他 茫然若??， 发觉 自己 还是 在 **老** 地方 ， 在 黑 ？？的 楼梯 上 。 在 几步 的 某一个 人 ， 但 英勇 的 事迹 使 他 和 **老人** 都 骄傲 得 心花怒放 ， 仿佛 那些 事 就是 朵 夫 立刻 凑 上去 。 他们 俩 很 投机 。 **老人** 非常 喜欢 孙子 ； 有个 愿意 听 他 说话 的 嚷 起来 。 母亲 嘲笑 他 。 曼希沃 说是 **老人家** 疯了 ， 与其 把 孩子 弄得 神魂颠倒 ， 还不

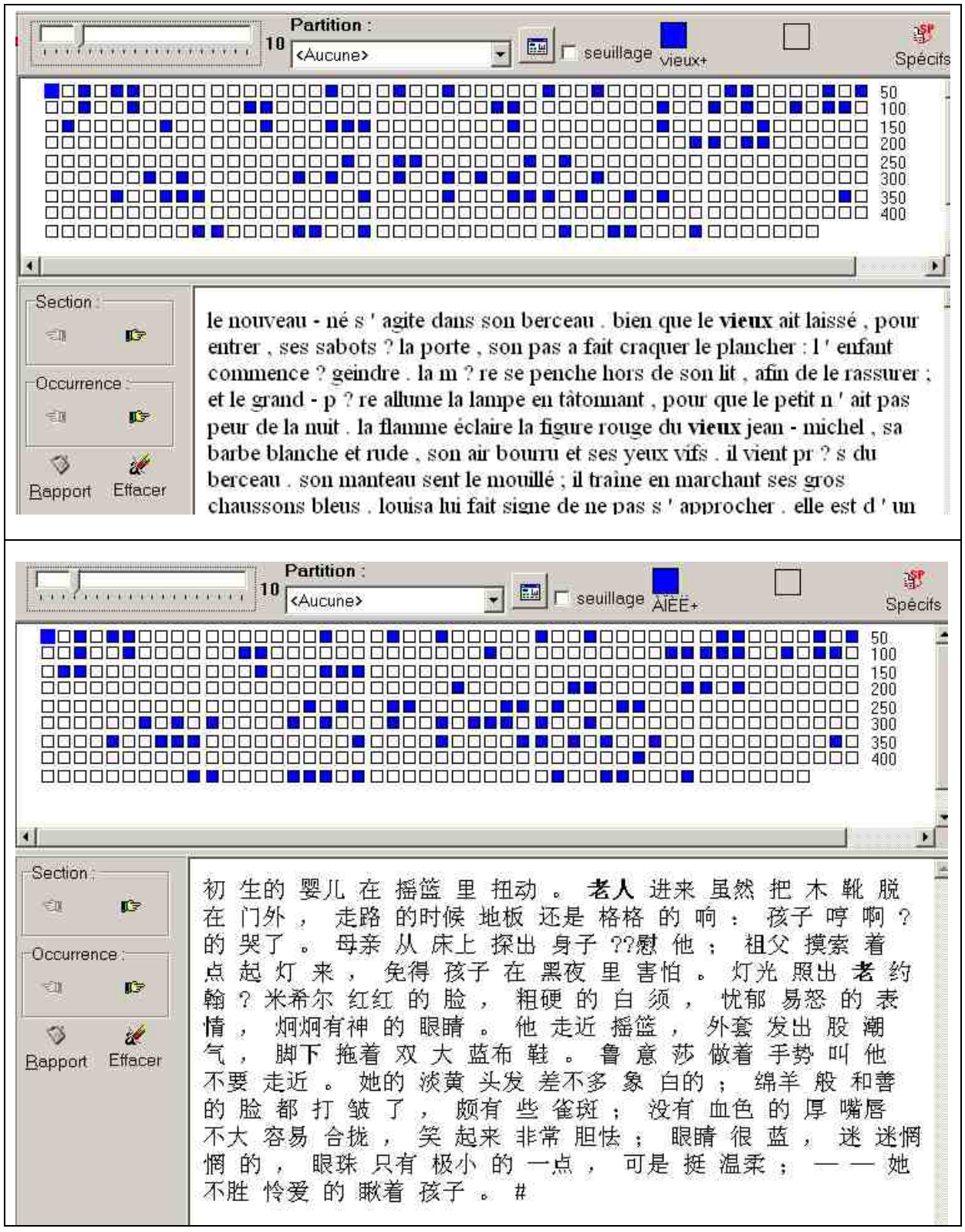


Figure 8 :

Localisation des correspondances de *vieux* et 老 dans le bitexte

à l'aide du logiciel *Lexico3*.

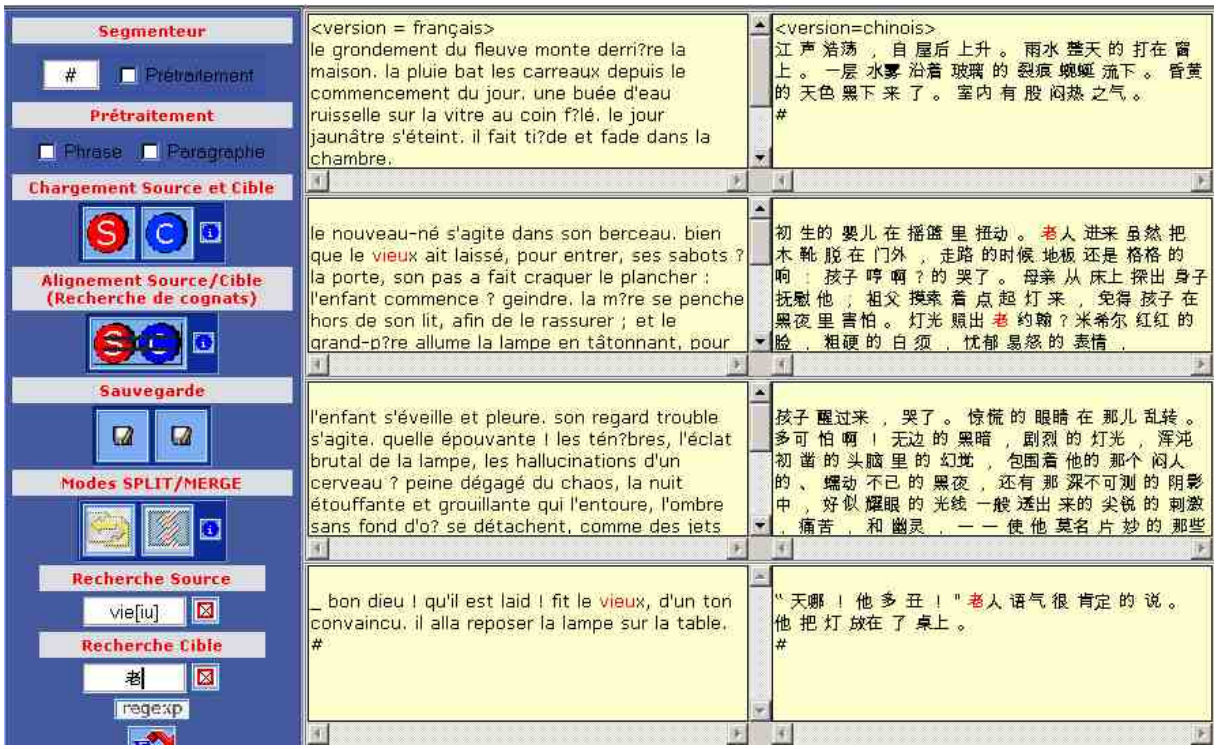


Figure 9 :

Visualisation des correspondances de *vieux* et 老 dans le bitexte à l'aide du logiciel mkAlign.

La localisation des concordances et des discordances dans la localisation des termes qui qui sont réputés constituer des équivalences traductionnelles permet d'approfondir l'étude traductologique et de mieux cerner les techniques propres à chaque traducteur pour rendre compte du sens véhiculé par le texte source.

Tableau 8 :

Traductions attestées dans le volet chinois pour le terme *vieux*

français	traduction chinoise	français	traduction chinoise
vieille maison	旧屋子 (maison ancienne)	de vieux amis	老朋友 (vieux amis)
vieille ficelle	旧绳子 (ficelle usagée)	vieux grand père	祖父 (grand-père)
de vieux habits	旧衣衫 (vêtements usagés)	vieux grand père	老祖父(vieux grand-père)
vieux veston bleu	旧蓝上装 (veston usagé)	le vieux jean-michel	老祖父(vieux grand-père)
vieille chanson	老调(mélodie ancienne)	le vieux	老人家 (un vieil homme)
vieille chanson	老歌 (chanson ancienne)	pauvre vieux	老人家 (vieil homme)
vieil escalier	黑魆魆的楼梯(escalier noir)	vieilles dames	老太太 (vieilles dames)
vieux rideau	破帘子 (rideau usagé)	vieilles dames	老婆婆 (vieilles dames)
vieille caisse	破旧匣子 (caisse abîmée)	il vieillissait	年纪越大(il prenait de l'âge)

7 Conclusion

La complexité apparente, le système d'écriture chinois ne constitue pas un obstacle incontournable à l'exploration textométrique des textes. Les traitements informatisés élaborés pour les textes codés à l'aide d'écritures alphabétiques peuvent être adaptés, moyennant des modifications mineures à l'étude des textes chinois.

Malgré des difficultés importantes dans la définition de l'entité *mot* en chinois, l'introduction de cette notion et sa prise en charge par des logiciels de segmentation automatique permet d'augmenter l'efficacité de l'exploration textométrique du bitexte franco-chinois et de dépasser l'exploration fondée sur les caractères *hanzis* considérés comme des entités isolées.

Les résultats, obtenus sur la base de la comparaison textométrique du bitexte aligné découpé en mots ouvrent, au plan traductologique, des pistes de comparaison qui semblent extrêmement prometteuses. Elles permettent d'envisager la comparaison simultanée des moyens lexicaux utilisés dans les corpus de traduction mis en confrontation et des procédés employés par les traducteurs pour faire saisir à leurs lecteurs les différents sens, nuances et connotations véhiculés par le texte d'origine.

8 Références

ALLETON V. 1997. *L'écriture chinoise*, « Que sais-je ?», 5è édition corrigée, 1re édition : 1970, Paris, Presses universitaires de France.

FU LEI (傅 雷). 1998 .*La grande série de la traduction de Fu Lei 傅雷译文全集* , He fei, Éditions de l'art d'An Hui, 安徽文艺出版社.

FLEURY S., MKAlign : *Manuel d'utilisation*, http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

GRANGER S., LEROT J., PETCH-TYSON S. (eds.). 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Amsterdam – New York, Editions Rodopi.

HABERT B., NAZARENKO A., et SALEM A. 1997. *Les linguistiques de corpus.* Paris, Armand Colin/Masson.

HOA M. 2005. *C'est du chinois*! I, volume "Lire et écrire", 3è édition. Paris, Édition You-Feng.

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

OLOHAN M. 2004. *Introducing Corpora in Translation Studies*. London and New York, Routledge.

SALEM A., "Introduction à la résonance textuelle", Actes *des 7èmes Journées d'analyse des données textuelles,* Louvain la neuve, 2004

WEI N. et alii. 2005. *Corpora in use 语料库应用研究.* Shanghai, Éditions de l'enseignement des langues étrangères de Shanghai 上海外语教育出版社.

ZIMINA, M. 2004. *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de doctorat, Université de la Sorbonne nouvelle – Paris3.

ZIMINA, M. 2005. *Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Actes des 7es Journées scientifiques du Réseau de chercheurs "Lexicologie, Terminologie, Traduction", Institut supérieur des traducteurs et interprètes (ISTI), Bruxelles.

ZIPF, G.., K. 1935. *The Psychobiology of Language, an introduction to Dynamic Philology.* Boston, Houghton-Mifflin.

ZHOU Q., DUAN H., 周 强, 段慧明. 2007. *Traitement de segmentation et de marquage des mots dans les corpus chinois modernes 现代汉语语料库加工中的切词与词性标注处理*, disponible sur http://hi.baidu.com/jagard/blog/item/dcdb653844fd842097ddd8ec.html

9 Fonctionnalités *Lexico3* utilisées dans cette exploration

N°	*Fonctionnalité*	*Résultat*
5.5	Courbe d'accroissement des caractères (hanzis)	*Figure 5*
5	Principales caractériques lexicométriques (PCLC)	*Tableau 4*
5.5	Accroissement du vocabulaire (chinois et français)	*Figure 6*
5.4	Diagramme de Pareto (chinois et français)	*Figure 7*
7	Carte des sections (volets français et chinois)	*Figure 8*

Traductions franco-coréennes

[franco-coréen]

Cho Joon-Hyung
chojh4net@gmail.com

Résumé : L'approche quantitative nous permet d'explorer la ventilation des mots en correspondance de traduction à partir d'une segmentation des séquences textuelles dans le corpus. Avec cette méthode, nous pouvons directement comparer des mots contenus dans le corpus parallèle en langues sans parenté, bien que celles-ci n'aient aucune structure syntaxique en commun. Dans le présent article, nous présenterons comment cette méthode est applicable aux corpus parallèle en langues hétérogènes à travers l'analyse textométrique d'un couple de mots traductionnel français/coréen dans un corpus parallèle coréen-français.

Mots clés : corpus bilingues, coréen, traductologie, textométrie

Abstract : A quantitative approach enables us to explore the distribution of words in translational correspondence obtained from the segmentation of the textual sequences in a corpus. With this method, we can directly compare the words from the parallel corpus in languages without cognates, although they do not have any syntactic structure in common. In this article, we will present how this method is applicable to parallel corpora in heterogeneous languages through the textometric analysis of a couple of French/Korean translational words in a parallel French-Korean corpus.

Keywords : bilingal corpora, korean, traductology, textometrics

1 Contexte de la recherche

Les corpus parallèles bilingues, sont des corpus composés de deux textes en langues différentes dont l'un constitue la traduction de l'autre. Chacun des textes est découpé en un système d'unités de traductions qui peuvent être mises en correspondance deux à deux. Ce type de corpus est actuellement utilisé dans diverses études comparatives : stylistique comparée, lexicographie bilingue, traductologie, traitement automatique des langues, désormais TAL (cf. Véronis, 2000).

La méthode textométrique nous permet, à partir de la segmentation des séquences textuelles, d'explorer, dans chacun des volets du corpus, la ventilation des formes graphiques ainsi que les réseaux de cooccurrences autour d'une forme-pôle. Cette méthode permet, dans certains cas, d'entreprendre des analyse directes basée sur la forme graphique des unités lexicales qui entrent en rapport de traduction, écartant dans un premier temps, l'obstacle que constitue les caractéristiques syntaxiques différentes de chaque langue. Cependant, les comparaisons fructueuses entreprises à partir de textes écrits dans des langues proches deviennent plus compliquées à mettre en œuvre lorsque les bitextes associent des langues qui ne présente aucune parenté.

Dans cette étude, nous commencerons par présenter les principales caractéristiques morphosyntaxiques du coréen que nous comparerons très brièvement à celle du français (§ 2). Nous analyserons ensuite les différences quantitatives induites par ces caractéristiques pour les dépouillements de bitextes franco-coréens(§ 3). Nous envisagerons enfin l'approche textométrique des équivalences traductionnelles dans le cadre de l'étude d'un corpus parallèle coréen-français (§ 4).

2 Le coréen *et son système d'écriture*

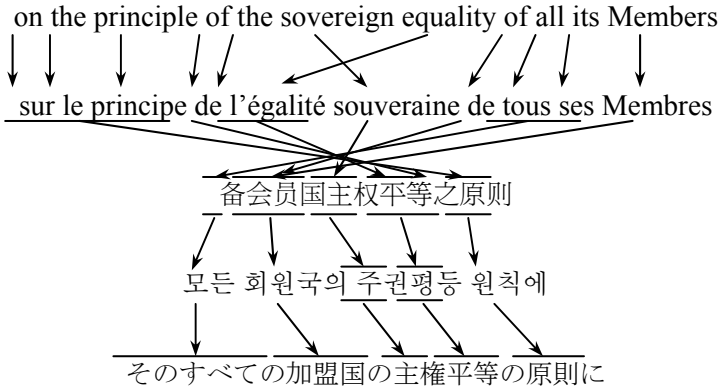
Le coréen est langue parlée en Corée par environ 72 millions de personnes. L'alphabet coréen, appelé *Hangul*, se compose fondamentalement de 24 lettres de base (14 consonnes et 10 voyelles). Mais on utilise en fait 40 lettres, si on inclut les consonnes et les voyelles doubles.

2.1 Caractéristiques linguistiques

Le coréen se distingue du chinois et du japonais, qui appartiennent à la même région culturelle et géographique par quelques caractéristiques typographiques et grammaticales.
On trouve ci-dessous, à titre d'exemple, cinq traductions, commentées grammaticalement, d'un même article de la *Charte des Nations Unies* (chapitre I, article 2.1): anglais, français, chinois, coréen et japonais.[26]

- The Organization *(sujet)* [[is based] *(verbe)* [on the principle of the sovereign equality of all its Members] *(complement)*] *(prédicat)*. (anglais)
- L'Organisation *(sujet)* [[est fondée] *(verbe)* [sur le principe de l'égalité souveraine de tous ses Membres] *(complément)*] *(prédicat)*. (français)
- 本组织 *(sujet)* [[系基于] *(verbe)* [备会员国主权平等之原则] *(complément)*] *(prédicat)*. (chinois)
- 기구는 *(sujet)* [[모든 회원국의 주권평등 원칙에] *(complément)* [기초한다] *(verbe)*] *(prédicat)*. (coréen)
- この機構は *(sujet)*, [[そのすべての加盟国の主権平等の原則に] *(complément)* [基礎をおいている] *(verbe)*] *(prédicat)*. (japonais)

Bien qu'il s'agisse de langues différentes, l'anglais et le français partagent, en plus de l'alphabet latin, des structures syntaxiques proches. En revanche, les trois dernières langues orientales possèdent des caractéristiques qui diffèrent fortement des premières et qui ne sont pas les mêmes à l'intérieur du second groupe. En premier lieu, les trois langues orientales utilisent depuis longtemps certains caractères chinois à des fins de communication. Mais ceux-ci se diffèrent dans chaque cas par la prononciation et la forme.
En coréen moderne, les caractères chinois (*hanja* caractères phonétiques, idéogrammes indispensables à l'écriture du chinois et du japonais) ont pour rôle principal d'aider à lever de nombreuses ambiguïtés sémantiques qui résultent de la transcription en *Hangul* des mots chinois.



[26] Les versions anglais/français/chinois de la Charte de l'ONU se trouvent sur le site officiel de l'ONU (http://www.un.org). Les versions coréenne et japonaise, peuvent être consultées respectivement sur les sites du Ministère des affaires étrangères et du commerce en Corée (http://www.mofat.go.kr) et sur celui du Centre d'information des Nations Unies au Japon (http://www.unic.or.jp/know/kensyo.htm).

Par ailleurs, le chinois possède fondamentalement une structure phrastique qui n'est pas sans rapport avec les deux premières langues occidentales (sujet-verbe-complément), alors que le coréen et le japonais recourent à une structure phrastique inverse (sujet-complément-verbe). Par contre, le chinois suit, pour la position des attributs, un ordre identique à celui des deux langues orientales.

Le coréen fait partie, avec le japonais et le turc, des langues agglutinantes caractérisées par la combinaison des radicaux avec des particules auxiliaires qui déterminent les propriétés grammaticales des radicaux. Comme nous le verrons plus loin, ces particularités entraînent des conséquences importantes au plan quantitatif. Le grand nombre des formes différentes dans les textes coréens dépouillés en formes graphiques résulte avant tout de cette agglutination des particules auxiliaires aux radicaux qui complique singulièrement l'analyse morphologique.

2.2 Les caractéristiques typographique

Le coréen moderne utilise généralement les signes de ponctuation occidentaux pour marquer les limites de la phrase et celles de la proposition. Il utilise de surcroît quelques ponctuations coréennes comme 「 」, 『 』 pour noter les titres d'œuvres. On note aussi quelques différences entre la ponctuation du coréen et celle du français : par exemple, le coréen utilise pour les citations des guillemets anglais (“ ”) au lieu des guillemets français (« »).
Comme en français et en anglais, les mots coréens sont séparés par des espaces. Les corpus de textes coréens se prêtent donc sans grande difficulté à la segmentation automatique en mot par la sélection d'un ensemble de *délimiteurs* (signes de ponctuation et espace).

La structure syllabique originale du coréen est caractérisé par la combinaison de 2 à 3 lettres par syllabe, disposées en *carré virtuel*, on recense effectivement 11 172 combinaisons de ce type qui peuvent être identifiées à des caractères. La version actuelle de *Lexico3* n'accepte pas encore la table *Unicode*. Elle rencontre, de ce fait, des problèmes pour afficher simultanément le coréen et le français.

2.3 Encodage des textes coréens pour Lexico3

Le *couteau suisse* de *Lexico3* permet d'afficher les caractères coréens lorsqu'ils sont encodés avec la table de caractères *win-949*, basée sur l'*ASCII*, qui correspond au codage « Coréen Wansung ». Mais, dans le cas du traitement informatique d'un corpus multilingue constitué par des couples *langues occidentales /langues orientales*, les outils informatiques ont du mal à afficher simultanément les caractères correspondant aux deux systèmes d'écriture.

3 Le corpus

Pour illustrer notre propos, nous avons sélectionné un corpus de textes juridiques constitué par une série de conventions, protocoles, chartes, etc., publiés à propos du thème des droits de l'homme, par le Haut-Commissariat des Nations Unies aux droits de l'homme, le Conseil de l'Europe, la Commission Interaméricaine des Droits de l'Homme et le Bureau International du Travail.[27]

---

[27] On peut consulter les textes originaux du corpus *Droit* sur les sites suivants :
Haut-Commissariat des Nations Unies aux droits de l'homme (http://www.ohchr.org/french);
Conseil de l'Europe (http://conventions.coe.int/Treaty/FR/v3DefaultFRE.asp);

Le corpus ***Droits*** se compose de deux volets : le premier est constitué par le texte orignal en français, le second par sa traduction en coréen. Les traductions coréennes ont été officiellement publiées par la représentation de l'UNESCO en Corée et par la Commission nationale des Droits de l'Homme de Corée.[28] Signalons que les traductions coréennes n'ont pas été réalisées directement à partir des textes français mais à partir de leurs équivalents anglais. Cependant, dans la mesure où l'anglais et le français sont les deux langues officielles de ces organisations qui effectuent pour leur compte des traductions de qualité, nous avons considéré, pour cette expérience, que le bitexte franco-coréen pouvait être considéré comme un corpus parallèle de bonne qualité.

Le corpus ***Droits*** a déjà été aligné au niveau des phrases. Il ne contient aucune balise véhiculant des informations linguistiques à l'exception de quelques caractères spéciaux portant sur la structure des textes et sur leur alignement en phrases: le paragraphe (§), la phrase (#), l'indice des phrases alignées ($) et les lettres en majuscules contenues dans l'original (*)[29]. Une relecture attentive du corpus nous a permis de corriger certaines erreurs de traduction. Certains fragments absents dans l'un des volets ont été supprimés dans le volet correspondant pour constituer un corpus d'expérimentation acceptable.

français	coréen
§$# *article 1 §$# *tous les êtres humains naissent libres et égaux en dignité et en droits. $# *ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité. §$# *article 2 §$# *chacun peut se prévaloir de tous les droits et de toutes les libertés proclamés dans la présente *déclaration, sans distinction aucune, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique ou de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation.	§$# 제1조 §$# 모든 사람은 태어날 때부터 자유롭고, 존엄성과 권리에 있어서 평등하다. $# 사람은 이성과 양심을 부여받았으며 서로에게 형제의 정신으로 대하여야 한다. §$# 제2조 §$# 모든 사람은 인종, 피부색, 성, 언어, 종교, 정치적 또는 그 밖의 견해, 민족적 또는 사회적 출신, 재산, 출생, 기타의 지위 등에 따른 어떠한 종류의 구별도 없이, 이 선언에 제시된 모든 권리와 자유를 누릴 자격이 있다.

Tableau 1 :

Extrait du corpus *Droits*

Le corpus ***Droits*** se compose de quarante parties qui correspondent chacune à une convention ou à un protocole. Les deux volets du corpus comptent respectivement 7 867 phrases françaises et 7 947 phrases coréennes. Le nombre de couples des phrases alignées est de 7 721, en raison des divers types de correspondances entre les phrases alignées. Pour cette

---

Commission interaméricaine des Droits de l'Homme (http://www.cidh.org/docdebase.htm); et
Bureau international du Travail (http://www.logos-net.net/ilo/150_base/fra/instr/afri_2.htm).

[28] Il est possible d'obtenir les textes traduits en coréen que nous avons utilisés sur les sites suivants:
UNESCO en Corée (http://www.unesco.or.kr/hrtreaty), Commission nationale des Droits de l'Homme de Corée (http://humanrights.go.kr/eng/index.jsp).

[29] Les caractères identiques contenus dans les textes originaux ont été remplacés par d'autres signes de ponctuation.

étude lexicométrique, les deux textes ont été segmentés en occurrences de formes graphiques afin d'obtenir une première comparaison des caractéristiques lexicales des deux langues, sur la base de ce type de segmentation.[30]

Partie	Occurrences	Formes	Hapax	Fréq. Max	Forme Max
français	214 313	7 821	2 548	12 576	de
coréen	114 006	21 068	11 732	1 642	또는

Tableau 2 :

Principales caractéristiques lexicométriques du corpus *Droits*

Le Tableau 2 montre que la taille du volet français, mesurée en occurrence de formes graphiques, est près deux fois supérieure à celle du volet coréen. A l'inverse, le nombre des formes du volet coréen est 3 fois plus élevé que celui qui a été calculé pour le volet français. Le volet coréen compte beaucoup plus d'hapax[31] que le volet français, conséquence des particularités morphologiques propres à la langue coréenne que nous avons mentionnées plus haut. Dans le volet coréen, plus de la moitié des formes, soit 55,7 % des formes graphiques, apparaissent en tant qu'hapax, ce qui contraste avec le taux de 32,6 % calculé pour le volet français.

3.1 Accroissement du vocabulaire

L'étude de l'apparition de nouvelles formes graphiques au fil du corpus confirme les différences quantitatives entrevues plus haut entre le coréen et le français. La courbe d'accroissement de vocabulaire calculée simultanément pour les deux volets du corpus (Figure 1) montre que la croissance du vocabulaire français s'épuise plus rapidement que celle du vocabulaire coréen[32]. De plus, l'accroissement du vocabulaire français devient de plus en plus faible au fur et à mesure que l'on avance dans le texte, alors que la courbe qui correspond au texte coréen maintient une pente relativement stable. Plus que le texte français, le texte coréen voit sans cesse apparaître de nouvelles formes graphiques.

[30] Les présents travaux, y compris la segmentation du corpus, ont été effectués à l'aide du logiciel *Lexico 3*, développé par le *CLA2T* (Centre de Lexicométrie et d'Analyse Automatique des Textes), Université de la Sorbonne Nouvelle - Paris 3. (http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/).

[31] Les hapax sont les formes dont la fréquence est égale à un dans le corpus..

[32] Signalons que ce corpus particulier montre un accroissement du vocabulaire relativement constant pour un texte français. Cela est sans doute,à mettre sur le compte d'une certaine hétérogénéité des documents rassemblés dans le corpus à partir de sources diverses, bien que concernant le thème des *droits de l'homme*.

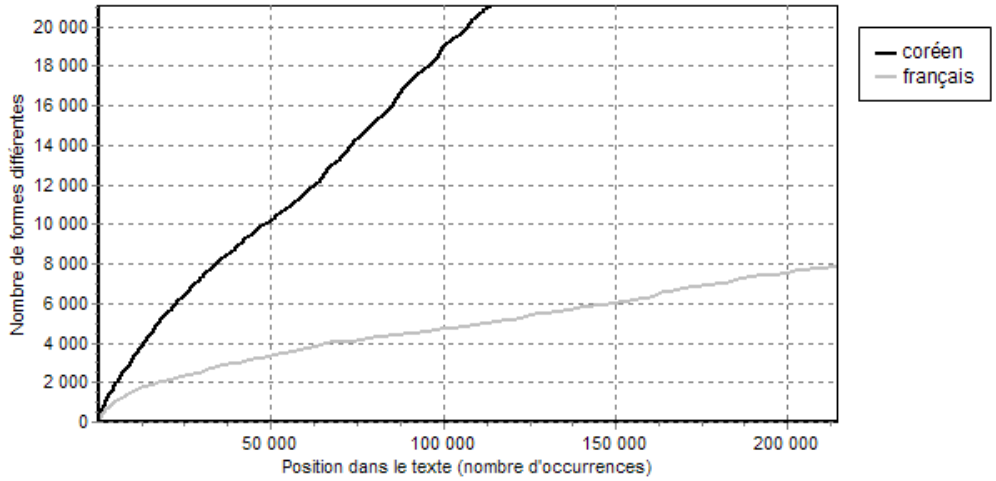


Figure 23:

Accroissement de vocabulaire dans les deux volets du corpus *Droits*

3.2 Diagramme de Pareto

Le diagramme de Pareto, figure 2, permet de visualiser la gammes des fréquence du vocabulaire pour chacune des deux langues rassemblées dans le corpus ***Droits*** [33].

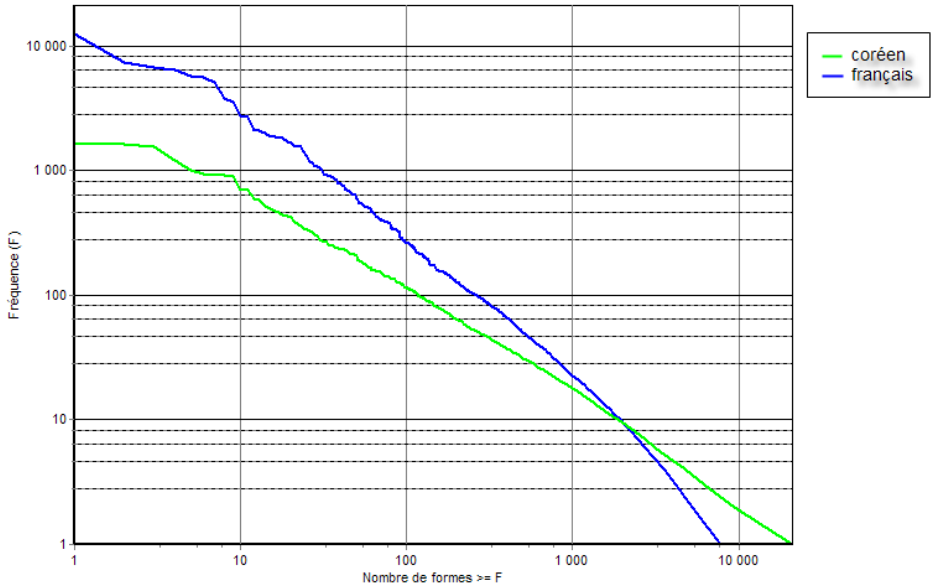


Figure 24 :

Diagramme de Pareto pour les deux volets du corpus *Droits*

[33] « Le diagramme de Pareto fournit une représentation très synthétique des renseignements contenus dans la gamme des fréquences. […] Sur l'axe vertical, gradué selon une échelle logarithmique, on porte la fréquence de répétition *F*, qui varie donc de 1 à *Fmax*, la fréquence maximale du corpus. Sur l'axe horizontal, gradué selon la même échelle logarithmique, on porte, pour chacune des valeurs de la fréquence *F* comprises entre 1 et *Fmax*, le nombre *N(F)* des formes répétées au moins *F* fois dans le corpus. La courbe obtenue est donc une courbe cumulée. » (Lebart et Salem, 1994 : 48)

Les différences que l'on peut constater aux deux extrémités du Diagramme confirment que le français utilise plus de formes de haute fréquence et moins d'hapax que le coréen. Ainsi, le taux de formes ayant plus de 10 occurrences atteint 24,6 % pour le français, tandis qu'il est environ de 8,8 % pour le coréen. Près de 91,2 % des formes graphiques du coréen ont une fréquence inférieure à 9 occurrences.

Les résultats statistiques présentés ci-dessus conduiraient à penser que le coréen utilise un grand nombre de mots monosémiques. Comme nous l'avons déjà signalé, il s'agit sans doute d'un artefact lié à la segmentation en formes graphiques que nous avons opérée sur la base de la distinction entre caractères délimiteurs et caractères non-délimiteurs. Nous reporterons à une autre étude l'analyse de l'incidence des propriétés agglutinantes que avons mentionné plus haut sur les calculs de fréquence.

Cet obstacle lié a la segmentation en formes graphiques peut cependant être contourné, pour les analyses qui suivent, par un repérage systématique, utilisant notamment le langage des expressions régulières qui offre une possibilité de repérer les différentes compositions réalisées à partir d'un même radical.

4 Analyse des équivalences traductionnelles français/coréen

Pour l'analyse textométrique, les textes sont d'abord segmentés en occurrences de formes graphiques qui sont ensuite regroupées par type. Les corpus textuels ainsi découpés permettent d'observer directement des formes ou des séquences textuelles sans référence particulières aux structures syntaxiques particulières des langues considérées.

Les résultats obtenus à l'aide du calcul statistique à partir de textes qui entrent en correspondance de traduction, constituent des données parallèles particulièrement précieuses pour les études contrastives[34]. Les travaux lexicométriques de M. Zimina (Zimina 2000) portant sur des corpus parallèles français-anglais constitués de documents concernant la *Convention de sauvegarde des Droits de l'Homme et des libertés fondamentales*, ont illustré les possibilités de cette méthode pour contribuer à l'alignement des unités correspondantes dans les deux volets du corpus. En comparant les fréquences globales et locales des termes français et de leurs traductions anglaises, ils ont mis en évidence des similarités distributionnelles entre les répartitions des termes des deux volets. D'autre part, l'analyse multidimensionnelle des formes qui entrent en rapport de cooccurrence avec un terme-pôle a permis de mettre en lumière des similarités distributionnelles qui concernent les réseaux de cooccurrences.

Le français et le coréen sont deux langues qui n'ont aucune parenté structurelle et qui, de plus, utilisent des caractères différents. Ces différences interdisent de s'appuyer sur la ressemblance des formes graphiques pour comparer la ventilation de termes qui entrent en rapport de traduction dans les deux langues. L'approche lexicométrique est-elle susceptible d'apporter un éclairage intéressant pour l'étude des corpus parallèles coréen-français ?

Dans ce qui suit, nous montrerons l'utilité de la méthode textométrique, sur l'exemple de l'analyse d'un ensemble de formes qui entrent en rapport de traduction dans le corpus français/coréen ***Droits***.

---

[34] Des analyses lexicométriques de ce type ont été réalisées à propos de corpus parallèles, parmi lesquelles : Martinez et Zimina (2002), Salem (2004) et Zimina (2000, 2002, 2004a, 2004b), etc.

4.1 Etude de l'équivalence traductionnelle homme/'사람 saram'

La forme *homme* est, en français, une forme polysémique capable de désigner plusieurs concepts du générique au particulier. Dans des contextes ordinaires, cette forme est fréquemment traduite en coréen par les quatre formes : '사람saram', '인간ingan' ; '인류illyu' (en fr., *humanité*) ; '남자namja' (en fr., l'antonyme de *femme*).

A l'inverse de ce qui se passe pour les confrontations entre langues proches comme le français et l'anglais pour lesquelles les comparaisons peuvent s'appuyer sur des ressemblances typographiques (*homme/human*, *administration/administration, etc.*), les confrontations entre textes français et coréens ne peuvent s'appuyer sur des ressemblance de ce type. Pour recenser l'ensemble des équivalences traductionnelles d'un terme particulier appartenant à un des volet du corpus, il est nécessaire d'examiner, autant que possible, l'ensemble du vocabulaire de l'autre volet. On peut optimiser ce genre de recherche en s'appuyant sur la fréquence et la répartition des formes attestées dans chacun des volets du corpus.

Le nom commun français connaît deux variations grammaticales, le singulier et le pluriel. Dans le volet français du corpus ***Droits***, la forme singulière *homme* compte 1 046 occurrences et son pluriel *hommes* 41 occurrences. En coréen, le nom commun est susceptible de prendre un assez grand nombre de variations au plan de la forme graphique.

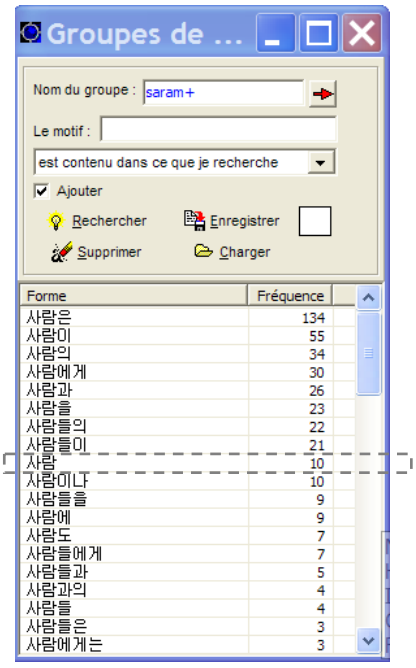


Figure 25 :

Groupe de formes *saram* dans le volet coréen du corpus *Droits*

Le mot coréen *saram* connaît deux principales variations grammaticales *saram* (singulier) et *saram**deul*** (pluriel). Dans le volet coréen, nous nous trouvons du fait de la structure agglutinante de la langue coréenne, de nombreuses occurrences qui concernent également la forme *saram* : '사람은sarame*un*' (F=134), '사람이sarami' (F=55), '사람의saram*ui*' (F=34), '사람에게sarame*ge*' (F=30), '사람과saram*gwa*' (F=26), '사람을sarame*ul*' (F=23), '사람들의saram*deurui*' (F=22), etc. Dans notre corpus (cf. figure 3), ces formes trouvent, pour la plupart, une fréquence supérieure à celle de la forme *saram* laquelle ne compte que 10 occurrences.

Dans le cadre du dépouillement en formes graphiques à partir de la sélection de caractères délimiteurs, la variation graphique associée à un nom commun français provient de la marque éventuelle du pluriel par rapport au singulier. Dans le cas d'un texte coréen cette variation est augmentée par la combinaison possible avec différents mots fonctionnels ou particules auxiliaires[35]. C'est la raison pour laquelle le dépouillement des textes coréens génère, comme nous l'avons déjà signalé au § 2, beaucoup plus de formes et d'hapax[36] que celui des textes équivalents français.
Faute de posséder une procédure de segmentation adaptée à la morphologie de la langue coréenne, il est nécessaire, pour repérer des traductions possibles du terme *homme,* d'examiner, au delà de la chaîne de caractères isolée *saram*, les occurrences de toutes les formes contenant la séquence de caractères *saram*.
Pour venir à bout de cette tâche, le concept de *Type généralisé* (***TGen)*** va se révéler d'une grande utilité[37]. Le ***TGen*** *homme*+ (désormais *homme_fr*) nous permet de rassembler les variations de la forme *homme* attestées dans le volet français du corpus (*hommes* et *hommes*). De la même façon, on constitue le ***TGen*** *saram*+ en rassemblant toutes les occurrences contenant *saram*. Nous pouvons faire de même pour chacune des formes traductionnelles coréennes mentionnées ci-dessus et rassembler l'ensemble de ces occurrences du corpus coréen dans un ***TGen*** *homme_co* que nous allons comparer au ***TGen*** français *homme_fr*.

TGen	Fréquence
saram+	428
ingan+	135
illyu+	18
namja+	0
Total	581

Tableau 3 :

Fréquence des mots traductionnels coréens correspondants au type *homme_fr* dans le volet coréen du corpus *Droits*

La comparaison des fréquences de chaque sous-groupe de formes du ***TGen*** *homme_co* révèle que, dans le corpus ***Droits,*** les types *saram*+ et *ingan*+ sont nettement plus fréquents pour traduire le terme français *homme* (Tableau 3), Au contraire, la fréquence du ***TGen*** *namja*+ est nulle dans la présente enquête. Ce résultat peut laisser penser que la forme *homme* n'est jamais utilisée comme antonyme de *femme* dans le corpus ***Droits.***
La question qui reste posée est celle de comprendre les raisons qui peuvent être à l'origine de l'écart fréquentiel entre les deux ***TGen*** *homme_fr* (F=1 087) et *homme_co* (F=581). Dans ce qui suit, nous allons chercher ces raisons à partir de l'exploration des fréquences locales de ces deux ***TGen*** dans les parties du corpus.

---

[35] Dans nos exemples, '-은eun' (nominatif), '-이i' (nominatif), '-의eui' (génitif), '-에게ege' (datif) '-을eul' (accusatif) ; '-과gwa' (conjonction) appartiennent aux particules auxiliaires. Elles ne définissent que la position du nom dans une phrase et n'entraînent aucun changement au plan sémantique. Ce phénomène est un des traits particuliers des langues agglutinantes telles que le coréen et le japonais.

[36] Dans l'état actuel, bien que la forme coréenne ait une seule occurrence, il serait difficile d'affirmer que cette forme est un hapax. Par exemple, les formes coréennes '사람들도saramdeuldo' et '사람으로sarameuro' ont une seule occurrence dans le corpus ***Droits***. En pratique, nous recensé 12 hapax contenant '사람' dans le volet coréen du corpus ***Droits***.

[37] Le TGen (Type généralisé) est un ensemble d'occurrences sélectionnées parmi les occurrences du texte. (cf. Lamalle et Salem, 2002).

4.2. Comparaison des fréquences locales dans les parties du corpus

Le type *homme_fr* compte 1 087 occurrences dans le volet français du corpus. Comme nous l'avons vu, la fréquence du *TGen* correspondant dans le volet coréen, *homme_co*, est beaucoup moins élevée (581 occurrences). Pour expliquer cet écart important, il est nécessaire d'explorer les fréquences locales du couple *homme_fr*/*homme_co* dans les parties du corpus. L'exploration de la variation des fréquences locales nous permettra de comprendre les raisons de cette disparité globale.

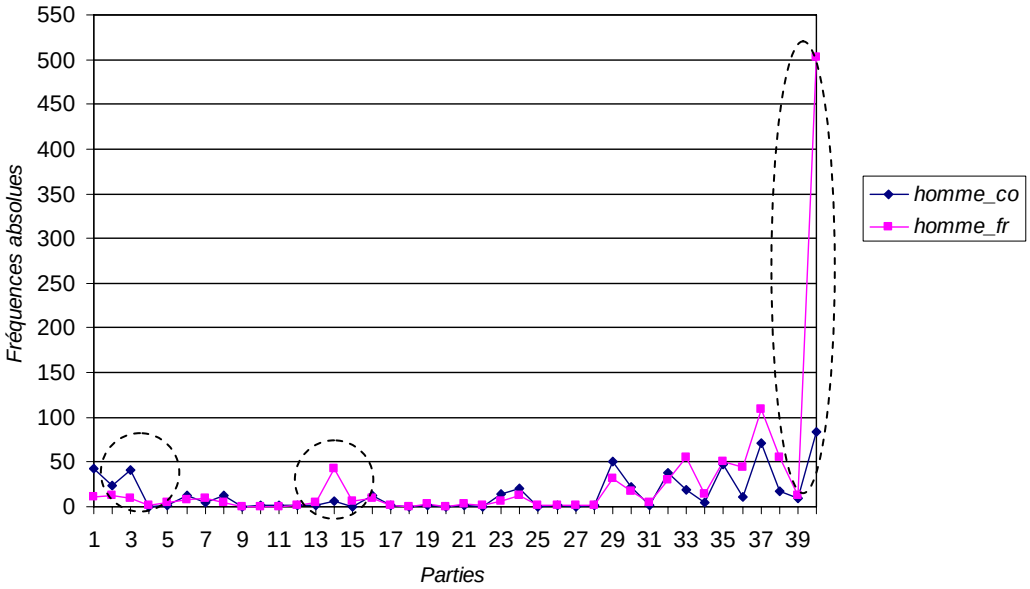


Figure 26 :

Fréquences locales des deux types *homme_fr* et *homme_co* dans les quarante parties du corpus *Droits*

Comme nous l'avons signalé plus haut, le corpus ***Droits*** est constitué de quarante parties. On voit, sur la Figure 26, que les deux courbes présentent un profil distributionnel similaire à quelques exceptions près. Le TGen *homme_fr* ne dépasse la cinquantaine d'occurrences que dans quelques parties. Dans les parties 37 et 40, *homme_fr* compte respectivement 109 occ. et 502 occ. Les parties 04, 09-12, 17-18, 20, 22 contiennent au maximum occurrence.
Dans le volet coréen, la fréquence locale du TGen *homme_co* dans chaque partie reste inférieurs à 50 occurrences, à l'exception des parties 37 et 40, dans lesquelles leur fréquence atteint respectivement 71 et 83 occurrences.
Les parties 4-5, 9-10, 12, 15, 17-22, 25, 27 comptent chacune une occurrence au plus. On a répertorié au tableau 4 des parties du corpus pour lesquelles la différence fréquentielle entre les deux volets est particulièrement importante.

Parties	01	03	14	33	36	37	38	40
homme_fr	11	10	42	55	44	109	56	502
homme_co	42	41	7	19	11	71	18	83

Tableau 4 :

Extrait des fréquences locales de *homme_fr* et *homme_co* dans les parties du corpus *Droits*

Une cartographie textuelle permet de visualiser, au niveau de chaque section, la présence ou l'absence des occurrences de chacun des *TGen*s. La carte des sections (Figure 27) montre des écarts dans la répartition des *TGens homme_fr* et *homme_co* entre les deux volets du corpus ***Droits***. Dans chacun des volets de la carte des sections, un carré représente une séquence (en général une phrase) alignée avec une sélection appartenant à l'autre volet du corpus[38]. Dans le volet français, la ventilation du *TGen homme_fr* est représentée par des carrés noirs ; celle du *TGen homme_co* est représentée par des carrés vert foncé dans le volet coréen.

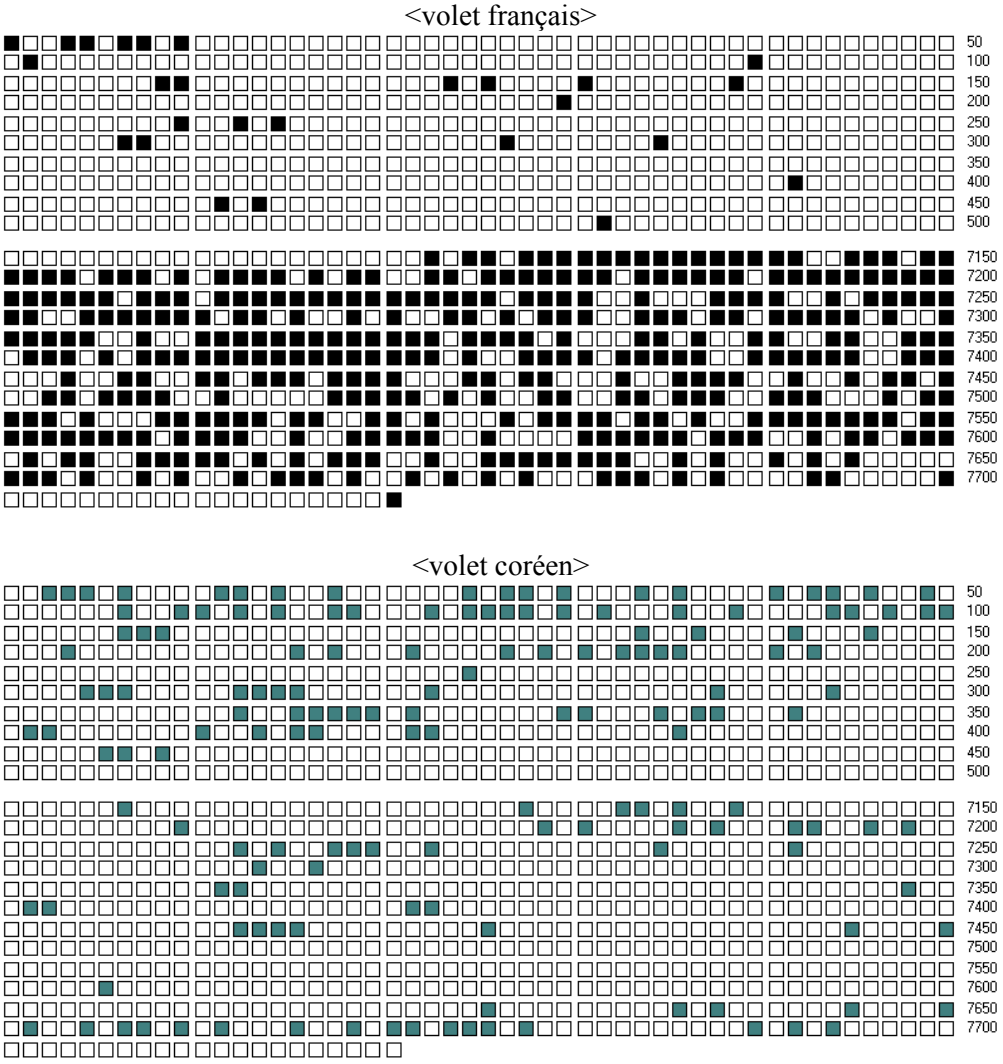


Figure 27 :
Extrait de la carte des sections ■ *homme_fr* et ■ *homme_co*
dans le corpus *Droits*

La distribution du type *homme_co* ne s'accorde que très partiellement avec celle du type *homme_fr* (Figure 5). Une fréquence supérieure du TGen *homme_co* dans certaines parties nous amènera au constat que différentes expressions françaises : *êtres humains*, *individu*, *personne humaine* ; ainsi que des formes qui constituent des reprises anaphoriques de ces

[38] Dans certains cas, un même carré peut contenir plus de deux phrases en fonction de la relation de correspondance avec l'autre volet.

dernières, le pronom personnel *ils* ; et le pronom *chacun*, *tous* (Tableau 5) sont rendues en coréen par des formes relevant du TGen *homme_co*. L'écart des fréquences locales dans les parties 01 et 03 s'explique par la présence de ces équivalences traductionnelles.

coréen	français
§$# 모든 **사람**은 태어날 때부터 자유롭고, 존엄성과 권리에 있어서 평등하다.	§$# *tous les **êtres humains** naissent libres et égaux en dignité et en droits.
$# **사람**은 이성과 양심을 부여받았으며 서로에게 형제의 정신으로 대하여야 한다.	$# ***ils** sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.
§$# **모든 사람**은 인종, 피부색, 성, 언어, 종교, 정치적 /…/.	§$# ***chacun** peut se prévaloir de tous les droits et de toutes les libertés proclamés /…/.
§$# 모든 **사람**은 생명권과 신체의 자유와 안전을 누릴 권리가 있다.	§$# *tout **individu** a droit à la vie, à la liberté et à la sûreté de sa personne.
§$# 이러한 권리는 **인간**의 고유한 존엄성으로부터 유래함을 인정하며,	§$# *reconnaissant que ces droits découlent de la dignité inhérente à la **personne humaine**,
§$# 1. 성년에 이른 **남**녀는 인종, 국적 또는 종교에 따른 어떠한 제한도 받지 않고 혼인하여 가정을 이룰 권리를 가진다.	§$# 1. *à partir de l'âge nubile, l'**homme** et la femme, sans aucune restriction quant à la race, la nationalité ou la religion, ont le droit de se marier et de fonder une famille.
$# /…/, 고등교육도 능력에 따라 모든 **사람**에게 평등하게 개방되어야 한다.	$# /…/ # l'accès aux études supérieures doit être ouvert en pleine égalité à **tous** en fonction de leur mérite.
§$# 2. 교육은 인격의 완전한 발전과 **인권** 및 기본적 자유에 대한 존중의 강화를 목표로 하여야 한다.	§$# 2. *l'éducation doit viser au plein épanouissement de la personnalité humaine et au renforcement du respect des **droits de l'homme** et des libertés fondamentales.

Tableau 5 :

Exemple des expressions françaises correspondantes au type *homme_co* dans le corpus *Droits*

4.3. droits de l'homme/'인권ingwon'

Plusieurs méthodes (sélection des termes cooccurrents, calcul des segments répétés) permettent de constater que, dans notre corpus, la forme *homme* est en cooccurrence étroite avec la forme *droits*. Le segment *droits de l'homme* compte 986 occurrences dans le corpus. Cependant, on ne trouve aucune occurrence de la traduction littérale du segment français qui serait constituée par l'expression '인간의 권리inganui gwolli' . Le segment *droits de l'homme* est souvent traduit par la seule forme '인권ingwon' qui compte 1 244 occurrences. Si nous tentons de localiser ces occurrences à partir des pôles de recherche *saram* et/ou *ingan*, nous ne localiserons pas les occurrences de la forme *ingwon*. L'écart important des fréquences que l'on a constaté entre les types *homme_fr* et *homme_co* dans les parties 36-38, 40 tient bien fait que la majorité des occurrences qui relèvent de la forme *homme* apparaissent dans le corpus ***Droits*** en cooccurrence avec la forme *droits*, la plupart du temps sous la forme *droits de l'homme*. On localise les occurrences correspondantes du type *ingwon+* dans les dernières parties du corpus (cf. Figure 28).

Dans les cas ou le segment subit une inclusion, il est à nouveau rendu par *ingan*. Par exemple, les *droits fondamentaux de l'homme* est traduit par '**인간**의 (de l'homme) 기본 (fondamentaux) **권리** (droits)' et non plus '기본 **인권**gibonjeok *ingwon*'.

On vérifie, sur la figure 6, que les distributions dans les parties du corpus du couple *droits de l'homme/ingwon+* sont assez similaires, à quelques expressions dues à la présence de segments comme *droits fondamentaux de l'homme, etc.*.

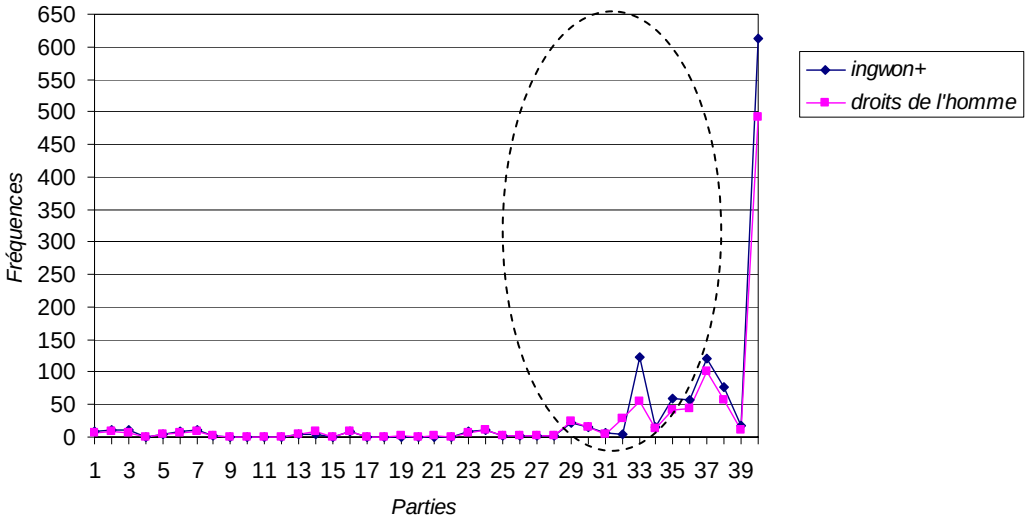


Figure 28 :

Les fréquences locales du couple *ingwon+/droits de l'homme* dans les quarante parties du corpus *Droit*

L'écart constaté à propos de la partie 40 tient essentiellement au phénomène que nous venons de décrire plus haut. Cependant, après la prise en compte de ces variantes traductionnelles, les parties 33 et 40 montrent encore des écarts importants au plan fréquentiel dans la répartition des occurrences des deux *TGen* dont nous avons entrepris le rapprochement. Dans la partie 33, les types *ingwon+* et *droits de l'homme* comptent respectivement 123 occurrences et 55 occurrences Dans le volet français, la fréquence du type *homme_fr* s'élève également à 55 occurrences, ce qui signifie que la forme *homme* n'apparaît dans cette partie que dans le contexte plus large *droits de l'homme*. Dans la partie correspondante du volet coréen, la fréquence locale du type *ingwon+* dépasse largement celle de *droits de l'homme*. Cette différence provient du fait que le nom des organisations internationales contenant ce segment et leurs sigles respectifs sont fréquemment traduits en coréen par le même segment coréen.

Commission des ***droits de l'homme*** : **인권**위원회
Commission
Haute Commissariat des Nation Unies aux ***droits de l'homme*** : 유엔**인권**고등판무관실
HCDH

4.4. homme/'남자namja'

On peut fournir une explication du même type pour rendre compte de la fréquence nulle du ***TGen*** coréen *namja+* (Voir Tableau 3). Le retour au texte permet de vérifier néanmoins la présence d'une opposition *homme/femme*. Dans les contextes où *homme* apparaît en cooccurrence avec *femme*, la plupart des occurrences coréennes apparaissent sous la forme : '남녀namnyeo' (F=31), '남성namseong' (F=40). *Namnyeo* est un mot composé indiquant « homme (남nam) et femme (녀nyeo) » et *namseong*, synonyme de *namja*, signifie, entre autres choses, un homme adulte.

La cartographie textuelle permet de représenter simultanément la localisation des occurrences du type *homme+* et celle du type *femme+* (*femme* 120 occurrences et *femmes* 55 occurrences). On compare ces résultats à la ventilation des occurrences du type coréen *namja+* à partir du dépouillement de *namnyeo* et *namseong*.

Dans le volet français de la carte des sections (Figures 7 et 8), les carrés noir indiquent la présence d'une occurrence du type *homme+* ; un carré gris celles des occurrences du type *femme+*. Les carré bicolores (noir et gris) signalent la cooccurrence au sein d'une même section des types *homme+*/*femme+*. De manière symétrique, les carrés noirs de la carte des sections réalisée pour le volet coréen indiquent la présence des occurrences du type *namja+*. La cartographie révèle que le type ◨ *homme+femme* génère une représentation qui ressemble considérablement à celle établie à partir du type ■ *namja+* pour le volet coréen. Le tableau 6 rassemble quelques cas qui font exception à cette règle et qui intéresseront le traducteur

coréen	français
/…/ 매춘행위를 목적으로 하는 **남녀**의 인신매매를 방지하기 위하여 본 협약에 의하여 그들의 의무로서 요구되는 조치를 채택하거나 /…/	/…/ les mesures destinées à combattre la traite des personnes de **l'un ou de l'autre sexe** aux fins de prostitution.
/…/ 유엔**인권**위원회와 **여성**의 지위에 관한 위원회의 진정 절차에 제출된 /…/	/…/ selon des procédures spéciales devant la * commission des **droits de l'homme** et la * commission de la condition de la **femme**.

Tableau 6 :

Exemples de cooccurrences *homme & femme* ne correspondant pas au *TGen namja+* dans le corpus *Droits*

La différence de fréquence constatée dans la partie 14 (Tableau 4) s'explique bien par la relation de cooccurrence du couple *homme/femme*. Le retour au contexte nous montre quelques segments comme *droits de l'homme et de la femme*, *entre l'homme et la femme*, *égalité de l'homme et de la femme*. La fréquence locale du type *namja+* dans la partie 14 est effectivement beaucoup plus élevée que dans les autres parties (F=38).

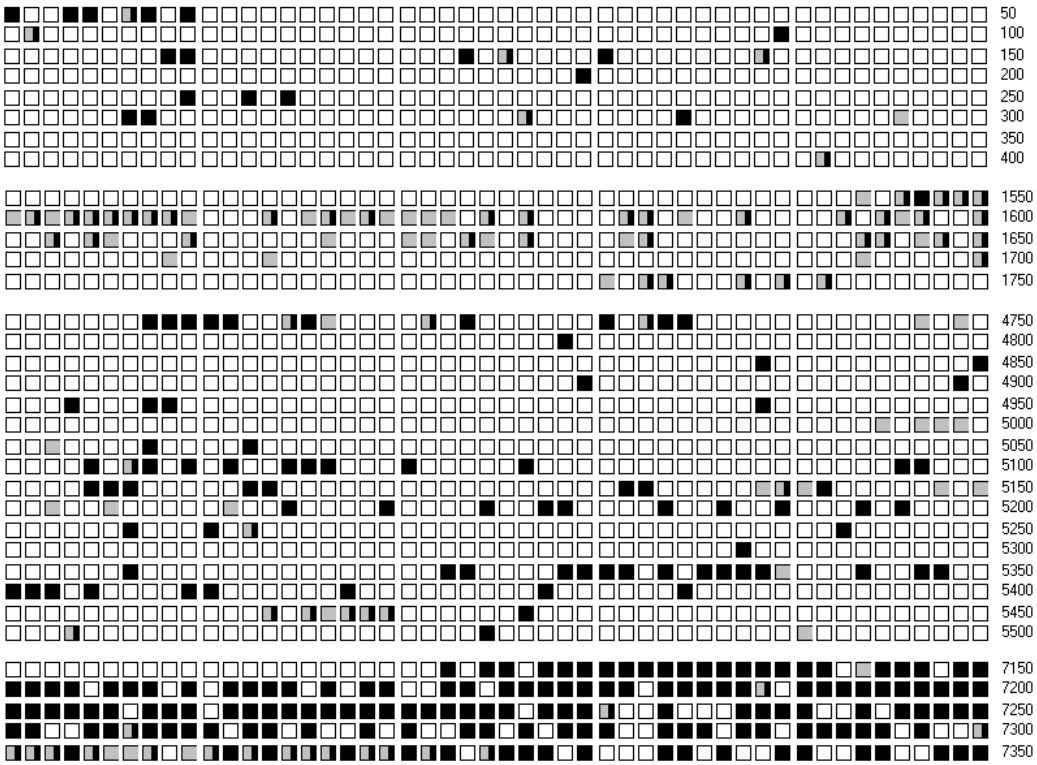


Figure 29 :

Extrait de la carte des sections ■ *homme_fr*, ■ *femme_fr* et ◨ *homme+femme* dans le volet français

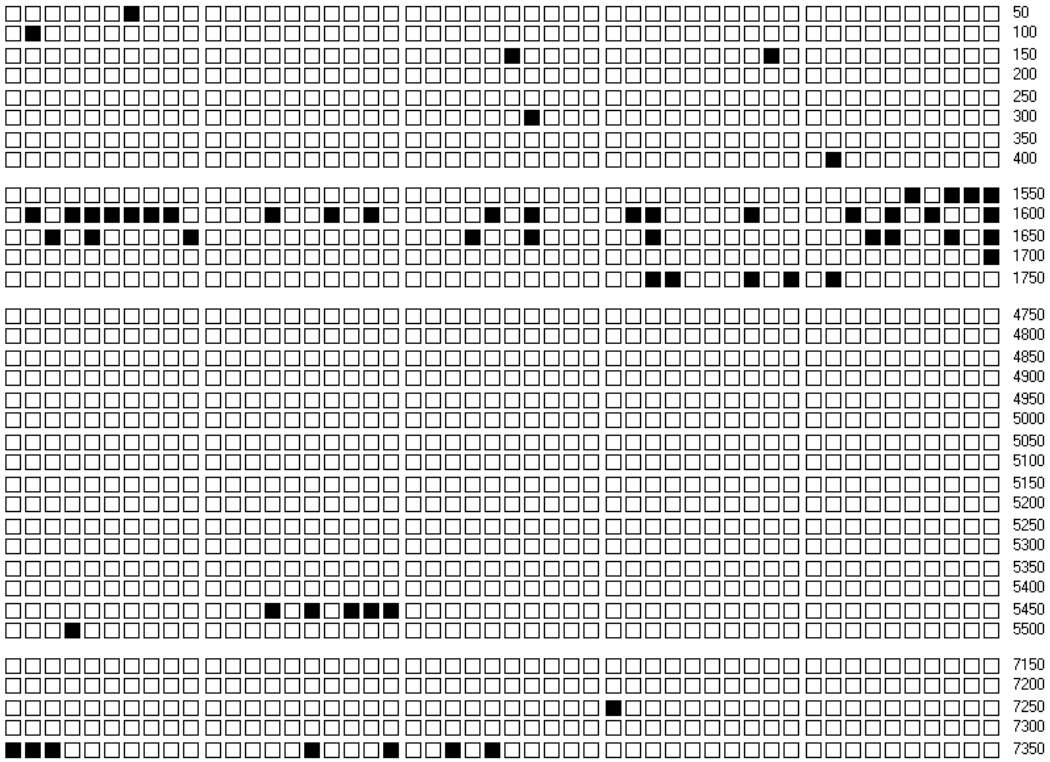


Figure 30 :

Extrait de la carte des sections ■ *namja+* dans le volet coréen

5. Conclusion

La traduction qui se donne pour objectif de transférer le sens d'un texte d'une langue à une autre mobilise des processus très complexes dans le cerveau humain. Lorsqu'il s'agit de langues n'ayant aucune parenté, la traduction des unités de la langue source vers des unités équivalentes dans la langue cible demande un travail encore plus complexe.
A partir de l'analyse lexicométrique du corpus ***Droits*** nous avons établi un certains nombre de rapports de correspondance pour le couple traductionnel *homme_fr/homme_co*. La complexité de ces rapports de traduction trouve sa source dans les différences profondes qui existent au plan linguistique et au plan culturel entre le français et le coréen. Cependant, l'observation des différences distributionnelles locales nous a permis d'établir un *schéma de traduction* du couple *homme_fr/homme_co* valable, pour le moins, à l'intérieur du corpus ***Droits***.

- homme → *saram*, *ingan*
  - si *homme* accompagne le mot *femme* → *namnyeo* ou *namseong*
- droits de l'homme → *ingwon*
  - si inclusion d'autres expressions lexicales → *ingan*
    *ex : droits **fondamentaux** de l'homme*
  - si il est suivi par le mot *femme* → *namja* ou *namseong*
    *ex : droits de l'homme et de la **femme***
- Autres expressions : *êtres humains*, *individu*, *personne humaine*, *chacun*, *tous* → *saram*, *ingan*

Dans cette étude, nous nous sommes attachés à la seule entité traductionnelle *homme_fr*/*homme_co* sans épuiser l'exploration des réseaux de cooccurrence autour de ces notions. Malgré ces limites, nous pensons avoir montré que l'analyse lexicométrique constitue désormais un outil extrêmement utile pour l'analyse des corpus parallèles qui concernent des langues sans parenté.

6 Références

Isabelle, P. et Warwick-Armstrong, S. (1993). « Les corpus bilingues : une nouvelle ressource pour le traducteur ». In P. Bouillon et A. Clas (Dir.), *La Traductique : études et recherches de traduction par ordinateur*, Les Presses de l'Université de Montréal, pp. 288-306.

Isahara, H. et Haruno, M. (2000). « Japanese-English aligned bilingual corpora ». In J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht / Boston / London : Kluwer Academic Publishers, pp. 313-334.

Lamalle, C. et Salem, A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels ». In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, 2002, pp. 403-412.

Lebart, L. et Salem, A. (1994). *Statistique textuelle*. Paris : Dunod.

Martinez, W. et Zimina, M. (2002). « Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues ». In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 495-506.

Rastier, F. (2005). « Enjeux épistémologiques de la linguistique de corpus ». In G. Williams (Dir.), *La linguistique de corpus*, Rennes : Presses Universitaires de Rennes, pp. 31-45.
Salem, A. (1987). *Pratique des segments répétés, Essai de statistique textuelle*. Paris : Klincksieck.
Salem, A. (2004). « Introduction à la résonance textuelle ». In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, pp. 986-992.
Salkie, R. (2000). « Quelques questions méthodologiques dans l'exploitation des corpus multilingues », in M. Bilger (Ed.), Corpus : *Méthodologie et applications linguistiques*, Paris : Honoré Champion, pp. 180-195.
Shin, J. H., Han, Y. S. et Choi, K.-S. (1996). « Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method (Korean-English Alignment at Word and Phrase Level) ». In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 230-235.
Simard, M., Foster, G. et Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Montreal, Canada, pp. 67-81.
Véronis, J. (2000). « From the Rosetta stone to the information society ». In J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht / Boston / London : Kluwer Academic Publishers, pp. 1-24.
Zimina, M. (2000). « Alignement de textes bilingues par classification ascendante hiérarchique ». In *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, Lausanne, pp. 171-178.
Zimina, M. (2002). « Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues ». In J. Véronis (Ed.), Revue électronique *Lexicometrica*, n. spécial « Corpus alignés ».
Zimina, M. (2004a). « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles ». In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, pp. 1195-1202.
Zimina, M. (2004b). *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de doctorat, Université Paris III.

7 Fonctionnalités *Lexico3* utilisées dans cette exploration

N°	Fonctionnalité	Résultat
5.5	Courbe d'accroissement du vocabulaire	*Figure 5*
5	Principales caractéristiques lexicométriques (PCLC)	*Tableau 2*
5.5	Courbe d'accroissement du vocabulaire	*Figure 1*
5.4	Diagramme de Pareto	*Figure 2*
6	Ventilation dans les parties	*Figure 4, 6*
8	Groupe de formes	*Figure 3*
7	Carte des sections	*Figure 5, 7, 8*

Le thaï. De la segmentation aux maux

[français-thaï]

Christian Jean
chr_jean2000@yahoo.fr

Résumé : Le thaï ou siamois[39] est une des langues d'Asie-du-Sud-Est à écriture non segmentée dérivée de la dévanagari indienne.

Pour le chercheur qui tente de pénétrer le domaine des études thaïes, la mise à disposition, sur des sites webs thaïlandais, de traductions de textes français réalisées par des traducteurs dont le thaï est la langue maternelle, constitue une occasion précieuse d'avancer dans la compréhension de la langue et de la culture thaïes.

La présente étude est consacrée à l'exploration en corpus à l'aide des outils fournis par ***Lexico3*** des problèmes de segmentation du thaï dans l'optique d'une étude textométrique comparative ultérieure. En effet, des études portant sur le thaï dans le domaine du traitement automatique des langues sont de plus en plus présentées en France. Toutes introduisent une spécificité du thaï à savoir l'utilisation d'une écriture non segmentée mais rares sont celles montrant les intrications entre les notions de syllabe, de morphème et d'unité lexicale dans le système de la langue thaïe.

Pour réaliser cette étude nous disposons d'un segmenteur automatique permettant de segmenter les textes thaïs en trois niveaux : la syllabe, le morphème lexical et l'unité lexicale. Les méthodes de segmentation de cet outil ont fait l'objet d'une publication en thaï [Asa2003]. Nous nous appuierons sur cette étude pour définir les notions de syllabes, de morphèmes lexicaux et d'unités lexicales. Acquiesçons que sans cet outil et sans cette publication, la présente étude aurait été impossible à réaliser.

Nous disposons par ailleurs d'un corpus parallèle de nouvelles françaises traduites en thaï. Ce corpus initialement préparé dans le but de faire une étude textométrique comparative entre le français et le thaï, permettra d'apprécier le sens des mots thaïs en fournissant le référentiel sémantique d'origine en plus de fournir des mots inconnus au segmenteur.

La section §1 présente les particularités du système d'écriture thaï ainsi que les trois niveaux de segmentation utilisés. La section §2 présente le corpus sélectionné. La navigation dans les syllabes, les morphèmes et les unités lexicales débute véritablement dans la section §3. La dernière section §4 est consacrée à un approfondissement des problèmes de segmentation en unités lexicales.

---

[39] Le terme thaï (ไทย) est la manière dont les Thaïs nomment leur langue, leur pays et eux-mêmes depuis 1939. Le siamois est le dialecte du centre de la Thaïlande (ancien royaume du siam) promu au rang de langue officielle, on l'appelle aussi thaï standard.

1 Présentation du thaï

Nous commencerons par décrire quelques propriétés du thaï sur lesquelles les chercheurs s'accordent en général et qui nous seront utiles pour notre étude.

La langue et son système d'écriture

Le thaï est une langue isolante c'est-à-dire que tous les mots sont invariables : le masculin, féminin, singulier et pluriel ne sont pas morphologiquement marqués. Les verbes ne se conjuguent pas. C'est une langue à tendance monosyllabique dont les nombreux emprunts au sanskrit, au pâli et plus récemment à l'anglais ont introduit de nombreux mots constitués de plusieurs syllabes.

Comme on le voit sur l'extrait de traduction présenté ci-dessous, le thaï possède une écriture non segmentée. Les mots ne sont pas séparés les uns des autres par des espaces. Il n'y a pas de délimiteur de phrase comparable aux signes de ponctuation de l'alphabet latin bien que l'espace[Tha1978] puisse sembler jouer parfois ce rôle.

L'écriture thaïe[40] utilise 44 signes consonnes et 19 signes supplémentaires qui en se combinant permettent de représenter 32 voyelles. À cela il faut ajouter 4 marques tonales, 2 diacritiques, 10 chiffres traditionnels, 3 marques additionnelles pour les mots pâli/sanskrit et 6 signes typographiques utilisés principalement dans les œuvres versifiées. Dans le corpus que nous avons réuni, on remarque aussi la présence de guillemets.

Segmentations préalables des textes thaïs

Afin de rendre le texte thaï analysable par ***Lexico3*** nous l'avons préalablement segmenté en utilisant l'outil Kucut[41] développé par l'unité de recherche NaiST[42] de l'université Kasetsart spécialisée dans le traitement automatique des textes écrits en thaï.

La méthode de segmentation utilisée par ce segmenteur est décrite dans [Asa2003]. Le taux de reconnaissance des mots déclaré est d'environ 80% pour la segmentation des mots inconnus et de 65% pour la fixation des frontières de l'unité lexicale. Cet outil permet de réaliser la segmentation sur trois niveaux différents.

Le premier niveau est la syllabe. Cette segmentation consiste à regrouper des caractères afin de former une syllabe prononçable. Par exemple :

- Le mot ทรราช /thorara:t/[43] sera découpé en 2 syllabes ทร/thon/-[44]ราช/ra:t/ (mot d'origine sanskrite)

---

40 Tous les caractères thaïs sont répertoriés dans le seul standard existant : le TIS 620-25335 défini en 1990 par l'Institut des Standards Industriels Thaïlandais. Il est encodé principalement par deux tables d'encodage 8 bits très similaires : la tis620, table officielle et la Windows-874 très utilisée dans le monde Microsoft. Ce jeu de caractères est aussi représenté dans Unicode.

41 Kucut est un programme écrit en Python et téléchargeable gratuitement : http://naist.cpe.ku.ac.th/wordcut/static/kucut-1.2.2.tar.gz

42 Natural Language Processing and Intelligent Information SystemTechnology Research http://naist.cpe.ku.ac.th/

43 Notes sur la translittération : la translittération utilisée ici est une solution ad hoc ayant pour but l'identification des mots par le lecteur. Elle renseigne peu sur la façon de lire car ni les tons, ni les valeurs et ni les longueurs de voyelles ne sont vraiment représentés.

- Le mot เขลา /khlaw/ sera découpé en une seule syllabe bien qu'on aurait pu le découper en deux syllabes เข/khe:/-ลา/la:/ mais dans ce cas, on aurait eu soit deux mots thaïs, soit un mot d'origine étrangère. Le ข/kh/ et le ล/l/ forment un groupe consonantique.

Le second niveau de segmentation est celui du morphème lexicale[45]. Il est défini comme la plus petite unité ayant un sens et apparaissant dans le dictionnaire de mots du segmenteur. Par exemple :

- พ่อ /phau:/, père; แม่ /mè:/, mère; หุง /hung/, cuire; สะพาน /sapha:n/, pont.

Le troisième niveau est celui de l'unité lexicale. L'unité lexicale est soit un morphème lexical, soit un mot composé[46]. Un mot composé est la fusion de plusieurs morphèmes dont le sens est changeant par rapport à ces morphèmes. Par exemple :

- Simple : พ่อ : père; น้ำ /nam/ : eau;
- Composée **พ่อ-แม่** : parents; แม่-น้ำ /mè- nam/ : rivière, fleuve;.

Le but de notre étude est de pouvoir observer en corpus les formes les plus et les moins spécifiques de chacun de ces niveaux, d'initier le lecteur à la complexité de différencier un mot composé d'un syntagme nominale et de déterminer à quoi correspond réellement ce niveau d'unité lexicale.

2 Le corpus

Nous présentons dans cette partie le corpus de travail, sa structure logique ainsi que les problèmes d'encodage.

Constitution

Ce corpus est constitué d'un ensemble de nouvelles françaises[47] ainsi que de leurs traductions en thaï. Elles sont disponibles sur le site http://www.wanakam.com. Un travail de normalisation ainsi qu'un alignement manuelle en unités de traduction a été effectué. Celle-ci varie d'une phrase à plusieurs paragraphes selon les nouvelles.

Nous disposons de deux fichiers de travail qui ont chacun une finalité et donc une structure différente.

Le premier fichier de travail *sylmorwor-corpus-th-cp874* a pour but l'étude des différents niveaux de segmentation du thaï. Il contient les textes thaïs en trois exemplaires divisés en parties selon leur niveau de segmentation. Elles sont identifiées par la clé <langue> dont les

---

44 On utilisera tout au long de l'article le – pour marquer la segmentation des syllabes.

45 Nous employons le terme morphème lexical bien qu'il puisse s'agir de mots outils pour indiquer qu'il n'est pas question de morphèmes comme dans les langues flexionnelles ou agglutinantes.

46 Pour le lecteur curieux, ouvrir un dictionnaire thai-anglais à l'entrée □□ /kham/ que l'on donne comme traduction du mot « mot » peu impressionner tant la liste de mots composés à partir de ce morphème lexical est longue. Par exemple dans le SE-ED's thai-english dictionnary la liste débute à la fin de la page 133 et s'achève à la fin de la page 136.

47 Auteurs de ces nouvelles : Alphonse Allais, Apollinaire, Aragon, Baudelaire, Bertot, Bloy, Daudet, Didier Daeninckx, Flaubert, Jean Hourgon, JMG Leclezio, Marcel Schwob, Maupassant, Perrault, Raymond Radiguet, Renard, Roegiers, Sagan, Sartre, Zola.

valeurs sont thsyl pour la partie segmentée en syllabes, thmor pour celle segmentée en morphèmes et thlex pour celle segmentée en unités lexicales. Chacune de ces parties est divisée en nouvelles identifiées par la clé <nouvelle> dont la valeur est composée d'un numéro et préfixée de la lettre A, B ou C pour les nouvelles segmentées respectivement en syllabes, morphèmes et unités lexicales. Par exemple la nouvelle 001 est identifiée par les valeurs <nouvelle=A001>, <nouvelle=B001> et <nouvelle=C001>.

Illustration 1: Extrait de la traduction de la nouvelle Arfled d'Alphonse Allais

```
<langue="frth">
<nouvelle="001"><auteur="1">
<par="00001">
|
```

Le Dr Joris-Abraham-W. Snowdrop, de Pigtown (U.S.A.), était arrivé à l'âge de cinquante-cinq ans, sans que personne de ses parents ou amis eût pu l'amener à prendre femme.

หมอ จอรีส __ อับราฮัม __ ดับเบิลยู __ สโนว์ดรอป __ เมือง พิกทาวน์ __ (สหรัฐอเมริกา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โกโหติกา หรือ เพื่อน สนิท ผู้ใด สามารถ โน้มน้าว ให้ เขา แต่ง ภรรยา ได้

```
<par="00002">
```

Texte 1: Extrait du fichier en relation de traduction.

Le deuxième fichier de travail dont nous disposons, corpus-frth-al-win contient la version française en relation de traduction avec la version thaïe segmentée en unités lexicales. La partition langue unique est identifiée par la clé <langue=frth>. La valeur de la clé nouvelle est seulement composée du numéro de la nouvelle sans être préfixée d'une lettre.

La clé <par> ainsi que le symbole délimiteur de section sont utilisés de manière à maintenir la relation de traduction. L'intérêt de cette structure est de pouvoir retrouver facilement les unités lexicales en relation de traduction à l'aide de la carte des sections.

Encodage des textes thaïs pour Lexico3

Le couteau suisse de ***Lexico3*** permet d'afficher les caractères thaïs lorsqu'ils sont encodés avec win874. Cependant, on doit prendre quelques précautions car les caractères § et ฯ partagent le même code 8 bits. Il faut donc exclure § de la liste des séparateurs et y ajouter le caractère | qui sert de délimiteur de sections dans notre corpus. Comme on veut garder la trace des espaces originaux, on exclu aussi le caractère _ de la liste des délimiteurs.

La table win874, idéale pour des textes bilingues anglais-thaï, permet de travailler simultanément avec les caractères ASCII et les caractères thaï mais pas avec les caractères français accentués. Ainsi il faudra faire un choix d'affichage lorsqu'on travaillera avec les fichiers contenant à la fois les versions françaises et les versions thaïes des nouvelles.

<langue="fr"><nouvelle="001"><auteur="1"> Le Dr Joris-Abraham-W. Snowdrop, de Pigtown (U.S.A.), était arrivé à l'âge de cinquante-cinq ans, sans que personne de ses parents ou amis eût pu l'amener à prendre femme. L'année dernière, quelques jours avant Noël, il entra dans le grand magasin du 37th Square (Objets artistiques en Banaloïd), pour y acheter ses cadeaux de Christmas.
<langue="th"> <nouvelle="001"><auteur="1"> หมอจอรีส อับราฮัม ดับเบิลยู สโนว์ครอป เมืองพิกทาวน์ (สหรัฐอเมริกา) ย่างเข้าสู่วัยห้าสิบห้าโดยไม่มีญาติโกโหติกาหรือเพื่อนสนิทผู้ใดสามารถโน้มน้าวให้เขาแต่งภรรยาได้ ปีที่แล้ว สามสี่วันก่อนวันคริสต์มาส หมอจอรีสเข้าไปซื้อของขวัญคริสต์มาสในห้างสรรพสินค้าย่านจัตุรัสสามสิบเจ็ด (ชิ้นงานศิลปะที่ทำด้วยพลาสถ่อย)
<langue="thsyl"><nouvelle="A001"><auteur="1"> หมอ จอ รีส __ อับ รา ฮัม __ ดับ เบิล ยู __ สโนว์ ครอป __ เมือง พิก ทาวน์ __ (สหรัฐ อเม ริ กา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โก โห ติ กา หรือ เพื่อน สนิท ผู้ ใด สา มารถ โน้ม น้าว ให้ เขา แต่ง ภรร ยา ได้ ปี ที่ แล้ว __ สาม สี่ วัน ก่อน วัน คริสต์ มาส หมอ จอ รีส เข้า ไป ซื้อ ของ ขวัญ คริสต์ มาส ใน ห้าง สรรพ สิน ค้า ย่าน จัตุ รัส สาม สิบ เจ็ด __ (ชิ้น งาน ศิล ปะ ที่ ทำ ด้วย พลาส ถ่อย)
<langue="thmor"><nouvelle="B001"><auteur="1"> หมอ จอ รีส __ อับ รา ฮัม __ ดับ เบิล ยู __ สโนว์ ครอป __ เมือง พิก ทาวน์ __ (สหรัฐอเมริกา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โก โห ติ กา หรือ เพื่อน สนิท ผู้ใด สามารถ โน้มน้าว ให้ เขา แต่ง ภรรยา ได้ ปี ที่ แล้ว __ สาม สี่ วัน ก่อน วัน คริสต์ มาส หมอ จอ รีส เข้า ไป ซื้อ ของขวัญ คริสต์ มาส ใน ห้างสรรพสินค้า ย่าน จัตุรัส สาม สิบ เจ็ด __ (ชิ้น งาน ศิลปะ ที่ ทำ ด้วย พลาส ถ่อย)
<langue="thlex"><nouvelle="C001"><auteur="1"> หมอ จอรีส __ อับราฮัม __ ดับเบิลยู __ สโนว์ครอป __ เมือง พิกทาวน์ __ (สหรัฐอเมริกา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โกโหติกา หรือ เพื่อน สนิท ผู้ใด สามารถ โน้มน้าว ให้ เขา แต่ง ภรรยา ได้ ปี ที่แล้ว __ สาม สี่ วัน ก่อน วันคริสต์มาส หมอ จอรีส เข้า ไป ซื้อ ของขวัญ คริสต์มาส ใน ห้างสรรพสินค้า ย่าน จัตุรัส สาม สิบ เจ็ด __ (ชิ้น งาน ศิลปะ ที่ ทำ ด้วย พลาส ถ่อย)

Tableau 7: Les différentes versions d'une nouvelle.

Guide de lecture du tableau 1

La première partie du Tableau 1 correspond à la version originale de la nouvelle Collage d' *Alphonse Allais*. La deuxième partie du tableau correspond à la version traduite en thaï. On remarque que le texte n'est globalement pas segmenté hormis quelques espaces ici ou là.

Chacune des parties suivantes a été segmentée par l'outil Kucut. Il a remplacé les espaces originels par la suite de caractères __ puis il a ajouté des espaces afin de délimiter les segments. La troisième, quatrième et cinquième partie du tableau correspondent aux versions thaïes segmentées respectivement en syllabes, morphèmes et unités lexicales.

3 Navigation dans les segmentations du thaï

Nous essayons de caractériser dans cette partie les différents niveaux de segmentation en observant leurs formes avec les outils statistiques de ***Lexico3***.

Principales caractéristiques

Partie	occurrences	formes	hapax	Fréq.Max	Forme
thsyl	110235	3991	1083	4125	__
thmor	98199	5978	2276	4125	__
thlex	89178	6493	2656	4125	__
Corpus	297612	8050	1353	12375	__

Tableau 8: Principales Caractéristiques Lexicographiques.

On observe[48] dans le Tableau 2 conformément à ce que l'on pouvait supposer que plus l'unité est petite telle la syllabe, plus la forme est en moyenne répétée et moins elle est susceptible d'être hapax. Inversement, plus l'unité est grande comme l'unité lexicale, moins la forme est répétée et plus il y a d'hapax. Le nombre élevé de syllabes différentes peut frapper mais sachant que le système d'écriture thaï peut théoriquement produire plus de 1.400.000 syllabes différentes[Ber2004], le nombre attesté est relativement faible.

Les sommations sur l'ensemble du corpus montrent que les parties ne sont pas au sens strict des partitions. En effet, il existe des formes et des hapax communs aux différentes parties.

Une dernière remarque concerne la forme la plus fréquente, le symbole __ qui représente les espaces présents initialement dans le corpus. Son utilisation reste fréquente bien que l'espace ne sert pas à séparer les mots[49].

Accroissement de vocabulaire

L'Illustration 2 montre les courbes d'accroissement de vocabulaire pour chacune des parties. On observe une forte corrélation entre les courbes des morphèmes et des unités lexicales. L'écart entre ces deux courbes tend à se stabiliser plus on avance dans le corpus alors que la courbe des syllabes a un comportement différent, elle se tasse beaucoup plus rapidement. On observe cependant dans deux secteurs du corpus, entourés en gris, une accélération de l'accroissement du vocabulaire pour chacune des parties. Ceci indique que l'apport de nouveaux mots et de nouveaux morphèmes est en partie réalisé par l'apport de nouvelles syllabes. Peut-être s'agit-il de mots empruntés transcrits comme des noms propres ?

---

[48] Nous rappelons que les partitions thsyl, thmor, thlex correspondent au corpus segmenté respectivement en syllabes, morphèmes et unités lexicales.

[49] Une étude textométrique de son usage à travers par exemple des concordances serait intéressante à mener ultérieurement.

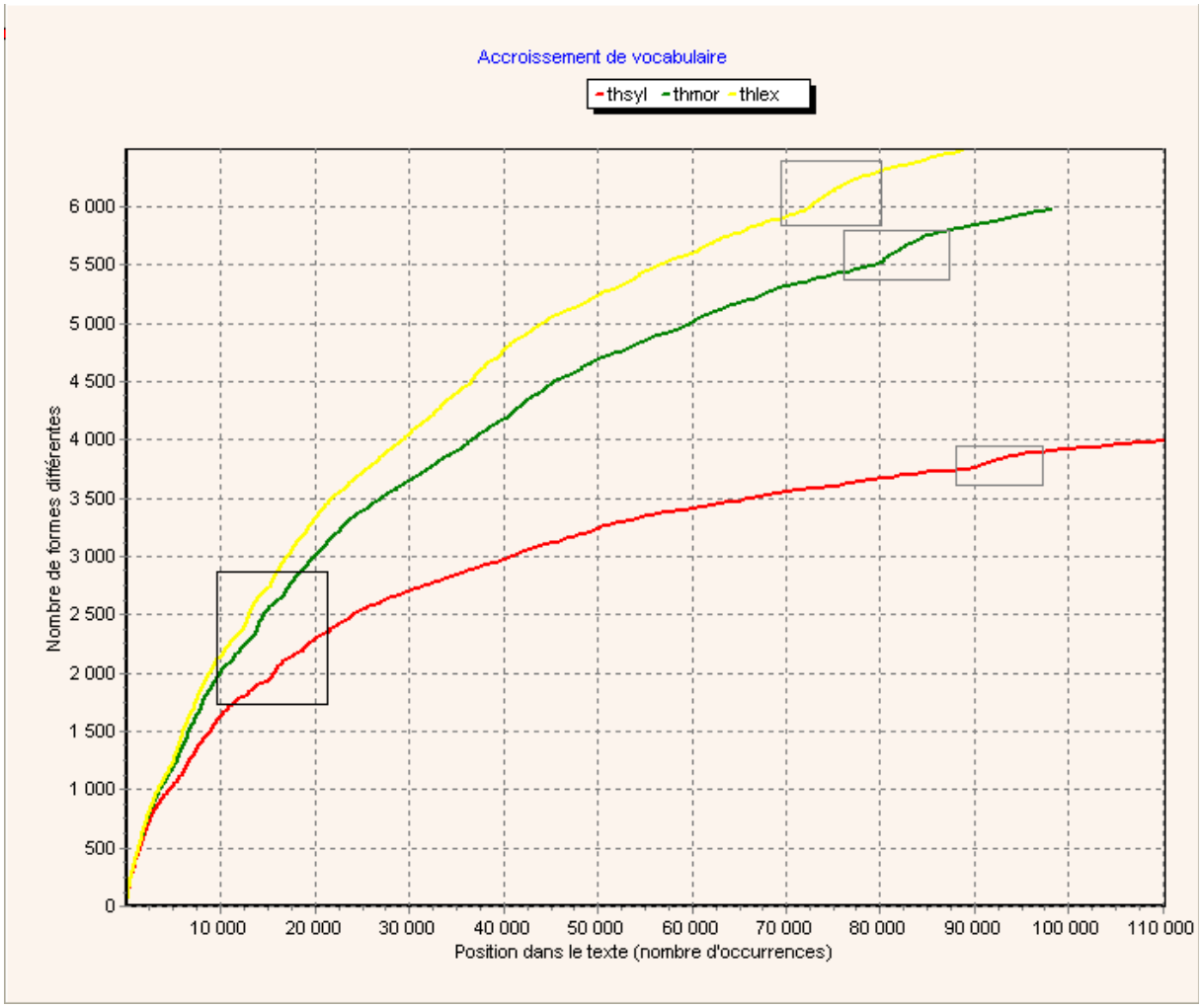


Illustration 2: Courbes d'accroissement de vocabulaire

Diagramme de Pareto

Le diagramme de Pareto, Illustration 3, montre que les syllabes, les morphèmes et les unités lexicales suivent à peu près la loi de Zipf. Il confirme que les syllabes sont plus utilisées que les morphèmes, ceux-ci plus utilisés que les unités lexicales. Cependant on observe que les courbes des morphèmes lexicaux et des unités lexicales sont très proches alors que celle des syllabes est un peu plus éloignée.

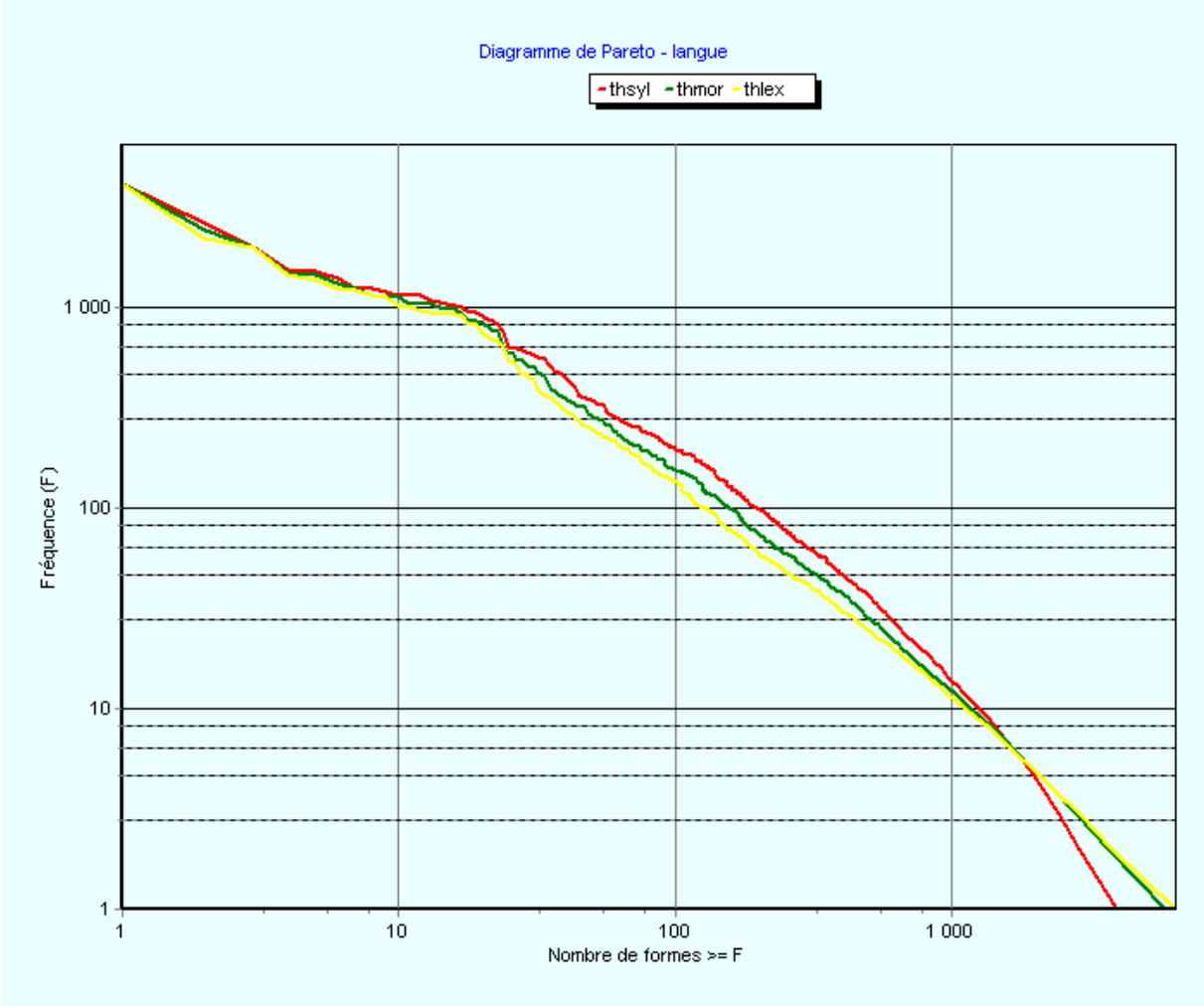


Illustration 3: Diagramme de Pareto

Les segmentations vues par les spécificités

Nous avons observé dans les parties précédentes que les syllabes et plus encore les morphèmes et les unités lexicales ont des comportements textométriques assez similaires. Par conséquent, nous allons utiliser les spécificités[50] de chacune des parties par rapport au corpus entier pour faire ressortir les formes spécifiques à chaque partie.

Les syllabes

Le Tableau 3 montre dans le volet gauche les cinq formes les plus spécifiques de la partie thsyl qui n'est autre que les traductions thaïes segmentées en syllabes. Il montre dans le volet droit les cinq formes les moins spécifiques de cette partie. On observe que toutes ces formes sont monosyllabiques. De plus, on remarque que les occurrences des formes du volet gauche sont presque exclusivement dans cette partie alors que les occurrences des formes du volet droit ne compte que pour un tiers des occurrences totales.

---

50 Nous avons retenu un seuil de probabilité de 5 et une fréquence minimale de 10.

Syllabes	*Spécificités positives*					*Spécificités négatives*		
Formes	Fréq.	Fréq.Tot.	Coef.		Formes	Fréq.	Fréq.Tot.	Coef.
ระ	256	301	***		ผม	1980	5938	-10
กระ	437	556	***		และ	1248	3735	-7
อะ	211	216	***		จะ	1144	3400	-6
ประ	423	536	***		เธอ	1151	3441	-6
สา	174	181	***		เขา	1013	3028	-6

Tableau 9: Spécificités pos/nég thsyl sur thsyl+thmor+thlex

On peut corroborer ce constat numérique par des connaissances linguistiques. En effet, il semble difficile d'attribuer un sens aux formes de gauche alors qu'on sait par connaissance du thaï qu'elles sont présentes dans de nombreuses unités lexicales. On peut confirmer cette affirmation par une recherche à l'aide de l'outil groupe de formes. Quelques exemples sont donnés dans le Tableau 4. Quant à la présence d'occurrences de ces formes dans les parties thmor ou thlex, il peut s'agir d'erreur de segmentation.

Motif **:** ^ระ 40 formes.	Fréq.	Motif : ระ**$** 12 formes.	Fréq.	Motif : **.+ระ.+** 198 formes.	
ระหว่าง	60	พระ	39	กระ.+	85
ระฆัง	26	ธุระ	7	ประ.+	66
ระดับ	13	กระ	6	Motif **:** ^**[**^ค**|**ป**]**.+ระ.+ 33 formes.	
ระบม	1	วาระ	2	จนกระทั่ง	8
ระบาด	1	ขรุขระ	1	ถึงกระนั้น	8
ระบายสี	1	ตระกรุมตระ	1	หลักประกัน	1

Tableau 10: Extraits de résultats de recherche de groupe de formes dans la partie thlex

En revanche, les formes du volet droit sont bien connues comme unité lexicale. Par exemple les formes ผม /phom/, เธอ /theu/ et เขา /khaw/ peuvent être utilisées comme des substituts du nom (je, tu/elle/il, il/elle/ils/elles) ou avoir une valeur lexicale (cheveux, ,montagne) quant à และ /lè/ c'est une conjonction de coordination étant presque équivalant à notre « et ». Le จะ /ja/ est une particule marquant l'inaccompli.

On peut confirmer cette connaissance linguistique par l'utilisation du concordancier pour décompter ces formes par partie. Par exemple, cela donne pour la forme ผม les résultats suivants : thsyl, 1980 ; thmor, 1979 ; thlex, 1979, confirmant ainsi le statut de syllabe, morphème lexicale et unité lexicale de cette forme.

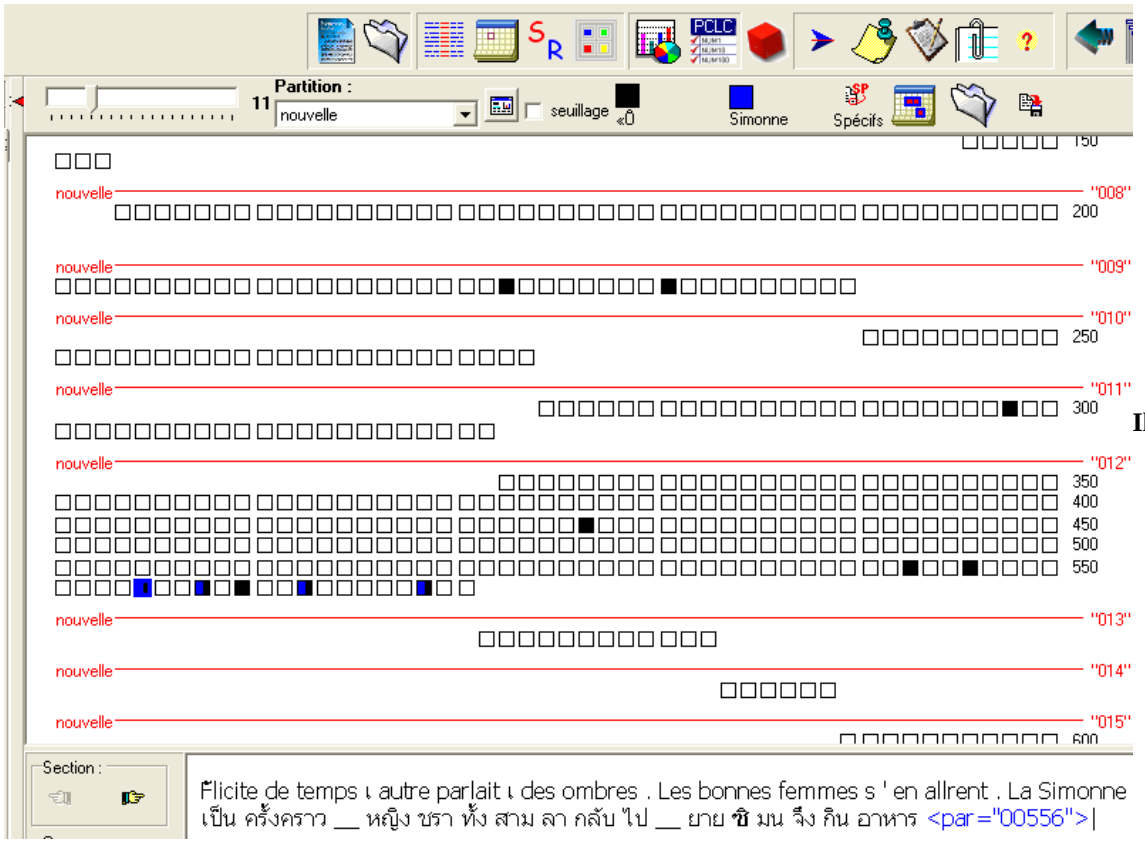
Les morphèmes lexicaux

Le Tableau 5 montre dans le volet gauche les cinq formes les plus spécifiques de la partie thmor qui n'est autre que les traductions thaïes segmentées en morphèmes lexicaux. Ceux-ci étant défini par le segmenteur comme la plus petite unité ayant un sens selon son dictionnaire. Il montre dans le volet droit les cinq formes les moins spécifiques.

Morphèmes	*Spécificités positives*						*Spécificités négatives*		
Forme	thmor	thlex	thsyl	Fréq.Tot.	Coef.	Forme	thmor	Fréq.Tot.	Coef.
อะไร	205	205	0	410	13	อะ	4	216	-32
ซิ	184	12	184	381	10	สา	6	181	-24
เวลา	140	143	0	283	9	เว	8	172	-20
ลิ	108	0	108	216	8	ตะ	1	117	-20
มาดาม	147	167	0	314	8	วิด	2	119	-18

Tableau 11: Spécificités pos/neg thmor sur thsyl+thmor+thlex

On observe dans le volet gauche trois formes composées de deux syllabes อะ-ไร/a-raj/, เว-ลา/we-la/ et มา-ดาม /ma-dam/ et deux formes composées d'une seule syllabe : ซิ /si/ et ลิ /li/. Le nombre de syllabes est aussi déductible par l'observation de la distribution des fréquences selon les parties. Les morphèmes dissyllabiques sont clairement des morphèmes lexicaux, en effet on a อะไร (pronom interrogatif), เวลา (Le temps) et มาดาม qui est une translittération de madame.



ation 4: Carte des sections pour ชิ et Simon-ne

Que sont les 12 occurrences de ชิ dans la partie thlex ? Une concordance groupée par nouvelle montre que sur les douze occurrences de ชิ, cinq, répartie dans quatre nouvelles, ont une autonomie réelle en tant que particule d'insistance. Comme le montre la carte des sections , Illustration 4, les sept autres occurrences sont localisées dans une seule nouvelle et n'ont qu'une valeur syllabique en tant que constituant d'un nom propre nom reconnu par le segmenteur ชิมน / Simonne. Les deux formes ชิ /si/ et ลิ /li/ sont apportées majoritairement par le prénom Félicité/เฟลิชิเต. Le segmenteur essaie de reconstituer les mots inconnus uniquement lors de la segmentation en unités lexicales, il est donc normal de retrouver les formes ลิ et ชิ dans la partie thmor lors de la segmentation en morphème lexical. En revanche, laisser telles quelles les occurrences de ชิ dans la partie thlex lorsqu'elles font parties du prénom ชิมน est clairement une erreur du segmenteur. Le problème spécifique de la reconnaissance des noms propres sera abordé ultérieurement.

Les formes du volet droit sont toutes monosyllabiques et ont une fréquence faible dans la partie thmor. On observe que deux des formes อะ /a/, เว /we/ sont des syllabes des formes อะไร /araj/ et เวลา /wela/ respectivement. Les quatre occurrences de อะ dans la partie thmor proviennent essentiellement d'emprunt dont certaines des syllabes sont connues comme des mots thaïes ainsi อะกู (aku, pronom malais signifiant je) où la syllabe กู signifie aussi je (familier) en thaï. La plupart de ces formes, à l'exception des instances de noms propres, seront reconstituées correctement dans la partie thlex.

Les unités lexicales

Le Tableau 6 montre dans le volet gauche les cinq formes les plus spécifiquement employées comme unité lexicale et dans le volet droit, les cinq formes les moins spécifiques.

Unité lexicale	Principaux sens	*Spécificités positives*					*Spécificités négatives*		
Forme		Fréq	Fréq.Tot.	Coef.		Forme	Fréq	Fréq.Tot.	Coef.
รู้สึก	Ressentir, sentir, sentiment,*	233	335	***		สึก	1	369	***
เฟลิซิเต	Félicité	76	76	41		อา	8	472	***
ทำให้	Causer, faire en sorte de ...	225	355	39		ประ	3	536	***
กำลัง	Modificateur d'aspect temporel/ pouvoir:N	159	255	27		กระ	6	556	***
ตัวเอง	Pronom personnel réflexif.	111	157	26		ใจ	82	988	***

Tableau 12: Spécificités pos/neg thlex sur thsyl+thmor+thlex

On observe que toutes les formes de gauche sont polysyllabiques alors que celles de droite sont monosyllabiques.

Les formes de gauche sont variées quant à leur nature. En effet, nous avons un verbe, un nom commun, un nom propre ainsi que des mots outils[51]. On remarque que le mot outils ทำ-ให้ /tham-haj/ est composé de deux syllabes dont l'une est principalement un verbe ทำ(faire) rentrant dans la composition d'un nombre assez important d'unités lexicales et l'autre est aussi un mot outil dérivé du verbe ให้ /haj/ (donner). Ils sont très fréquents. Par exemple ทำ apparaît dans les parties thsyl, thmor, thlex respectivement 581, 360, 210 fois et ให้ respectivement 1002, 854 et 748 fois.

Les fréquences des formes de droite, hormis celle de la forme ใจ[52], sont faibles dans cette partie. Ainsi il n'y a qu'une seule occurrence de สึก /seuk/ contre 132 dans la partie thmor et 236 dans la partie thsyl. On ne manquera pas de remarquer qu'elle rentre en composition dans la forme รู้สึก /ruuseuk/, celle-ci apparaît 102 fois dans la partie thlex, ce qui nous permet de déduire par calcul que la séquence รู้ สึก apparaît 131 fois dans la partie thmor[53]. Il existe donc

---

[51] Conformément à l'expression utilisée dans la méthode de langue intitulée « Pratique du Thaï » de Wanee Pooput et Michèle Conjeaud.

[52] La formeใจ /cai/ est une des plus belles formes du thaï dont le sens est associé à celui de cœur au sens propre comme figuré. Je laisse son étude aux doctorants ou au romancier amoureux du thaï (cf. http://www.learningthai.com/hearttalk.htm).

[53] On notera au passage que la segmentation en morphèmes lexicaux n'est pas stable puisqu'il n'y a pas de raison de découper la séquence รู้สึก tantôt en รู้สึก tantôt en รู้ สึก. Cela n'est pas très grave car l'étape morphème

une occurrence de สึก dans la partie thlex et une dans la partie thmor. L'utilisation des concordances groupées montre que c'est la même.

On a remarqué précédemment que les formes ประ /pra/ et กระ /kra/ participaient en tant que syllabe à la formation de nombreux mots (cf. Tableau 4) mais il s'avère que ces deux formes ont aussi une signification autonome attestée par l'existence d'entrée dans différents dictionnaires. Cependant, il reste à confirmer le statut de leurs occurrences dans nos textes.

Formes	Occurrences dans le texte		Entrées de dictionnaire
ประ	หลังคา **ประ** ทุน	capote	หลังคา N:toit **ประ**ทุน N:couverture
ประ	**ประ** เหมาะ	convenir	**ประ** V: ? เหมาะ Adj:être adapté **ประ**เหมาะ : Abs. dico.
ประ	ผู้ **ประ** เหมาะ เคราะห์ ซวย	La malheureuse	ผู้ : N:personne **ประ** : V:? เหมาะ : V: être fait pour เคราะห์ : N chance ou malchance ซวย : Adj malchanceux
สึก	เซาะ เสีย จน **สึก**	usés	เซาะ V:éroder เสีย : V: être abîmé จน : prép. Jusqu'à **สึก** : V: être éroder

Tableau 13: Occurrences en contexte d'unités lexicales les moins spécifiques

Le Tableau 7 montre les occurrences des formes ประ /pra/ et สึก /seuk/ de la partie thlex, c'est-à-dire considérées comme une unité lexicale après segmentation du texte original. On voit que leur statut respectif n'est pas simple puisqu'à chacune des séquences où apparaissent ces formes correspond un seul mot source français. La première ligne du tableau montre que la séquence est mal segmentée puisque les formes **ประ** /pra/ et ทุน /thun/ auraient dû être fusionnées en **ประ**ทุน/prathun/ conformément à l'entrée des dictionnaires. Quant à savoir si les formes หลังคา /langka/ et **ประ**ทุน doivent être fusionnées, il s'agit d'un autre problème.

La deuxième occurrence de ประ laisse à penser que ce sont bien deux unités séparées car la séquence **ประ** /pra/ เหมาะ /maw/ n'est attestée dans aucun de nos dictionnaires[54]. Cependant le sens de ประ est légèrement modifié par rapport aux différents sens donnés par ces dictionnaires.

La troisième occurrence de ประ ajoute encore au doute. En effet, on retrouve de nouveau la séquence **ประ** เหมาะ. En outre, on observe la séquence เคราะห์ /khraw/ et ซวย /suey/ qui est une accumulation de deux formes au sens proche ce qui légitimerait la composition en เคราะห์ซวย. Quant à la forme ผู้ /phou/ elle est souvent décrite dans les méthodes de langues comme un préfixe permettant la création de nombreux mots relatifs à une personne. Ainsi, si la forme

lexical pour le segmenteur est une sorte de pré-traitement pour constituer les unités lexicales. Ce n'est pas une analyse d'une unité lexicale en morphèmes.

54 Voir la liste des dictionnaires utilisés dans les références.

เขียน /khien/ signifiant écrire est précédée de ผู้ pour former ผู้เขียน, le tout signifie auteur, à ne pas confondre avec écrivain qui s'écrit นักเขียน. Ceci laisse à penser que la séquence complète de la troisième ligne constitue une seule unité lexicale construite à des fins littéraires mais dont le sens est parfaitement décomposable.

Nous voyons donc que la notion d'unité lexicale n'est pas simple et que les spécificités donc le segmenteur, ne se sont pas trompées en nous présentant la forme ประ comme peu représentative d'une unité lexicale et en nous présentant les noms propres et les mots outils comme des unités lexicales. Toutefois, on peut s'interroger sur la pertinence de la segmentation des séquences plus longues comme celles du Tableau 7.

Bilan de la navigation

Les observations faîtes sur les courbes d'accroissement de vocabulaire à savoir que les accroissements de syllabes, de morphèmes et d'unités lexicales sont corrélés, ont été confirmées par l'analyse des spécificités par partie. Ainsi on a vu que les syllabes les plus spécifiques rentrent dans la composition de nombreuses formes polysyllabiques ayant autant le statut de morphème lexicale que d'unité lexicale. On a aussi observé que certaines syllabes très fréquentes sont aussi des morphèmes et des unités lexicales notamment des mots à usage grammatical comme les substituts du noms.

On a aussi montré qu'il ne fallait pas trop se fier à la partie morphème lexicale lorsqu'il s'agissait d'analyser la composition d'une unité lexicale car bien souvent la segmentation était instable : soit l'unité lexicale apparaissait telle quelle, soit elle apparaissait segmentée.

Conformément à la description de cette méthode employée par le segmenteur[Asa2003], la segmentation en morphèmes lexicaux doit être vue comme une étape intermédiaire vers la construction des unités lexicales à partir des syllabes.

Enfin, l'observation des spécificités sur la partie unité lexicale a montré que si les mots outils, les noms propres semblent constituer le gros des unités lexicales c'est que les frontières des unités composées ne semblent pas très nette.

4 Les maux de l'unité lexicale

On vient d'observer que la nature des formes les plus spécifiques de la partie thlex est variée (noms propres, mots outils, verbe). Cependant, si on sélectionne les quinze premières formes au lieu de cinq, on remarque une large prédominance des noms propres. Ces formes complémentaires sont consignées dans le Tableau 8.

L'identification des noms propres et notamment des personnages est intéressante puisque notre corpus est constitué de nouvelles françaises traduites en thaï. L'enjeu est donc la restitution des noms de personnes, mots vraisemblablement inconnus des dictionnaires du segmenteur mais dont la limite signifiant/signifié est claire.

Par conséquent, nous utiliserons dans un premier temps les outils de ***Lexico3*** pour vérifier si les occurrences de noms propres ont été correctement identifiées et analyser, le cas échéant, les problèmes de non reconnaissance. Dans un deuxième temps nous essayerons de saisir la complexité de la notion d'unité lexicale en l'illustrant par un exemple tiré des formes les plus spécifiques, à priori simple, la forme อาหาร /ahan/(aliment, nourriture).

Forme	Principaux sens	Fréq/Fréq.Tot.		Forme	Principaux sens	Fréq/Fréq.Tot.
เฟลิซิเต	Félicité	76/76		รู้สึก	Ressentir, sentir, sentiment,*	233/335
โอแบง	Aubain	40/40				
มาร์เกอริต	Marguerite	39/39		อาหาร	Repas, diner, aliment,...	88/131
ปารีส	Paris	35/35		ประตู	Porte:V, *	75/110
จิเยร์	Gier	31/31				

Tableau 14: Formes extraites parmi les 15 unités lexicales les plus spécifiques.

Problèmes de segmentation des noms propres

Le Tableau 8 montre que les formes référençant des noms propres situées dans le volet gauche n'apparaissent que dans la partie thlex. Ceci indique qu'elles ont été découpées différemment dans la partie thmor. C'est à première vue surprenant puisque ce sont des mots empruntés donc impossible à analyser morphologiquement mais il faut garder à l'esprit que le segmenteur n'analyse pas en morphèmes les unités lexicales. En effet, il découpe d'abord le texte en syllabes, puis en morphèmes lexicaux enfin recompose les unités lexicales à partir de ces morphèmes.

Ceci étant dit, on peut avoir affaire à deux problèmes. Le premier est un problème de sous-segmentation c'est-à-dire que des parties de noms propres sont rattachées à d'autres unités lexicales. Le second est un problème de sur-segmentation c'est-à-dire que des bouts de morphèmes de noms propres n'ont pas été rattachés ensemble.

La méthode pour retrouver des occurrences de formes mal segmentées avec ***Lexico3*** consiste à calculer les segments répétés sur le corpus segmenté en trois parties puis à utiliser conjointement l'outil de recherche de groupe de formes et les expressions rationnelles.

Expressions	Exemples de formes		Expressions	Exemples de formes
.*เฟลิซิเต.*	**เฟลิซิเต** **เฟลิซิเตก็อต** **เฟลิซิเตพามันออกไป** **เฟลิซิเตไป** **เฟลิซิเตไอ** **เฟลิซิเตรักการ** . . . **เฟลิซิเตออกท่า**		.*จิเยร์.*	**จิเยร์** มาดาม กรอง **จิเยร์** เจ้า หนู กรอง **จิเยร์** กรอง **จิเยร์** มอง เมอซิเออร์ กรอง **จิเยร์**
14 formes trouvées, la plupart sont des hapax.			14 segments répétés, distribution variée.	

Tableau 15: Groupe de formes avec segments répétés

Les résultats de la recherche consignées dans le Tableau 9 montrent un problème de sous-segmentation pour la forme เฟลิซิเต /félicité/ puisque nous avons trouvé un certain nombre de formes contenant เฟลิซิเต. On voit tout l'intérêt d'utiliser les segments répétés puisqu'on remarque que la forme จิเยร์ /jier/ n'est pas un nom propre. Le vrai nom propre est กรองจิเยร์ (Grangier) puisqu'en contexte, la séquence est précédée de มาดาม (Madame), เมอซิเออร์ (Monsieur), ou เจ้า หนู (le petit [Grangier]). On vient donc d'identifier un problème de sur-segmentation.

Le problème de sur-segmentation de la forme กรองจิเยร์ s'explique partiellement par le fait que la forme กรอง /krong/ est un mot thaï. On utilise la carte des sections pour trouver des occurrences de กรอง n'apparaissant pas en coocurrence avec la forme จิเยร์.

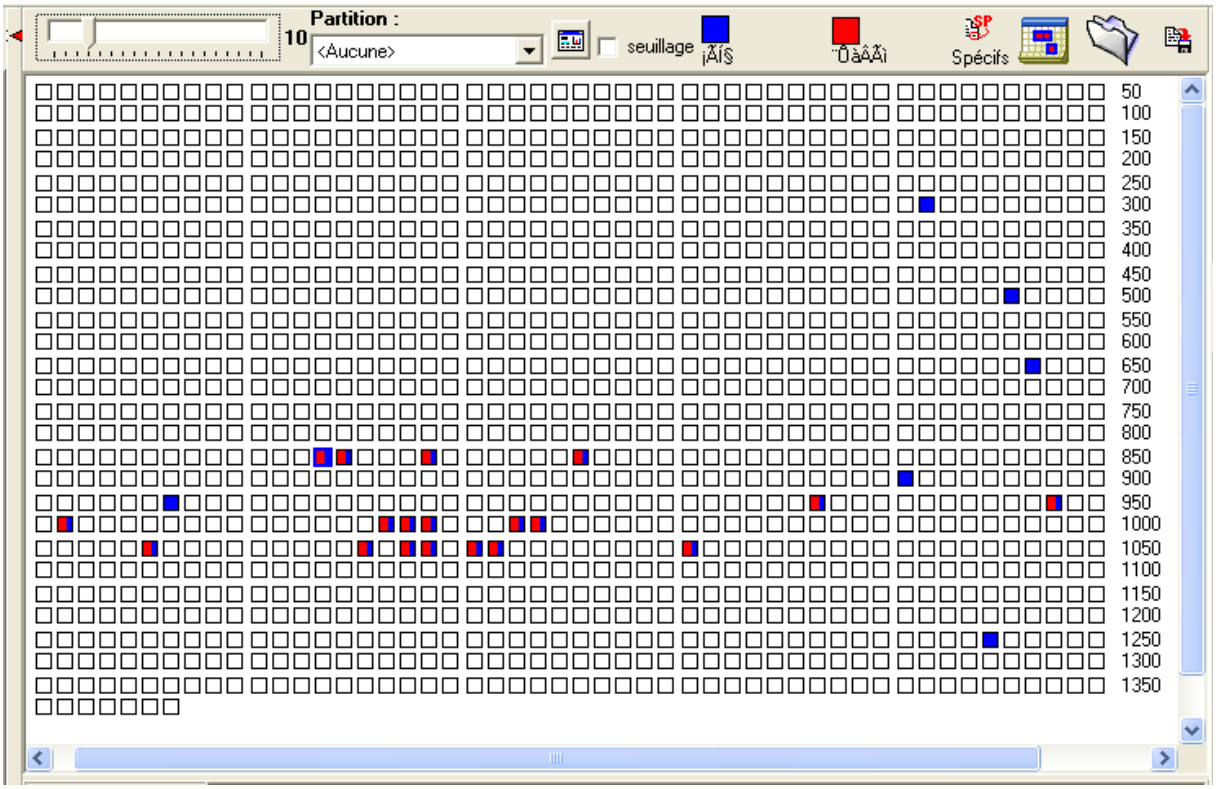


Illustration 5: Carte des sections : en bleu กรอง, en rouge จิเยร์

On remarque que la forme กรอง /krong/ apparaît sans la forme จิเยร์ /jier/ dans six sections. On a pu répartir ces occurrences de กรอง en trois groupes après analyse.

Deux occurrences réfèrent à l'unité lexicale thaïe dont le sens attesté par nos dictionnaires est le verbe filtrer. Dans nos textes, elles sont en relation de traduction avec le nom commun : filtre.

Trois autres occurrences sont des erreurs de segmentation concernant des noms de lieu non reconnus : ถนน **กรอง** ปง | rue Grand – Pont, ที่ **กรอง** วิลล์ | à Granville, ที่ **กรอง** วิลล์ | à Granville. La forme กรอง /krong/ est ici une transcription approximative du son gran qui n'existe pas en thaï.

Enfin, la dernière occurrence n'aurait pas dû exister. En effet nous avons le segment suivant เภสัช **กรอง** ฟร้ว pour Onfroy l'apothicaire, qui est une erreur de segmentation de niveau caractère. En effet la séquence aurait dû être segmentée de la façon suivante : เภสัช**กร** **อง**ฟร้ว | apothicaire Onfroy.

Nous tenons à faire remarquer que nous n'avons pas utilisé le segmenteur dans ses conditions optimales puisque pour résoudre les problèmes de mots inconnus, il utilise une méthode de segmentation basée sur des statistiques globales et locales. On aurait probablement gagné en précision si on avait segmenté nouvelle par nouvelle au lieu du corpus

dans sa globalité. Ainsi les occurrences de กรอง /krong/ dans les autres nouvelles n'auraient peut-être pas interféré avec celles liées à จิเยร์ /gier/.

Cette exploration des noms propres a permis d'expliciter quelques problèmes de segmentation provoqué par le fait que les formes empruntées sont composées de syllabes correspondant à des mots thaï. Ces problèmes ne concernent pas uniquement des textes traduits mais aussi les textes proprement thaïs puisque bien souvent les noms et prénoms thaïs sont des noms venant du sanskrit et du pâli ayant leur propre sens notamment dans le domaine religieux et royal.

Globalement le segmenteur basé sur des méthodes statistiques a réussi à correctement segmenter de nombreuses occurrences de noms propres. Peut-être qu'un post-traitement symbolique de reconnaissance d'entités nommés permettrait d'améliorer cette segmentation.

Problèmes de composition lexicale

Le Tableau 8 montre que les formes du volet droit ont une distribution différente des formes nominales du volet gauche. Par exemple, la forme /ahan/ apparaît 88 fois dans la partie thlex et seulement 43 fois dans la partie thmor. On a déjà expliqué ce phénomène précédemment. De plus, ce qui nous intéresse pour la suite de cette étude est de trouver des formes ou des segments répétés dans la partie thlex contenant la forme /ahan/ afin de déterminer la limite de l'unité lexicale.

La méthode pour retrouver ces formes avec ***Lexico3*** consiste à calculer les segments répétés sur le corpus aligné puis à utiliser conjointement l'outil de recherche de groupe de formes et les expressions rationnelles comme dans l'Illustration 6.

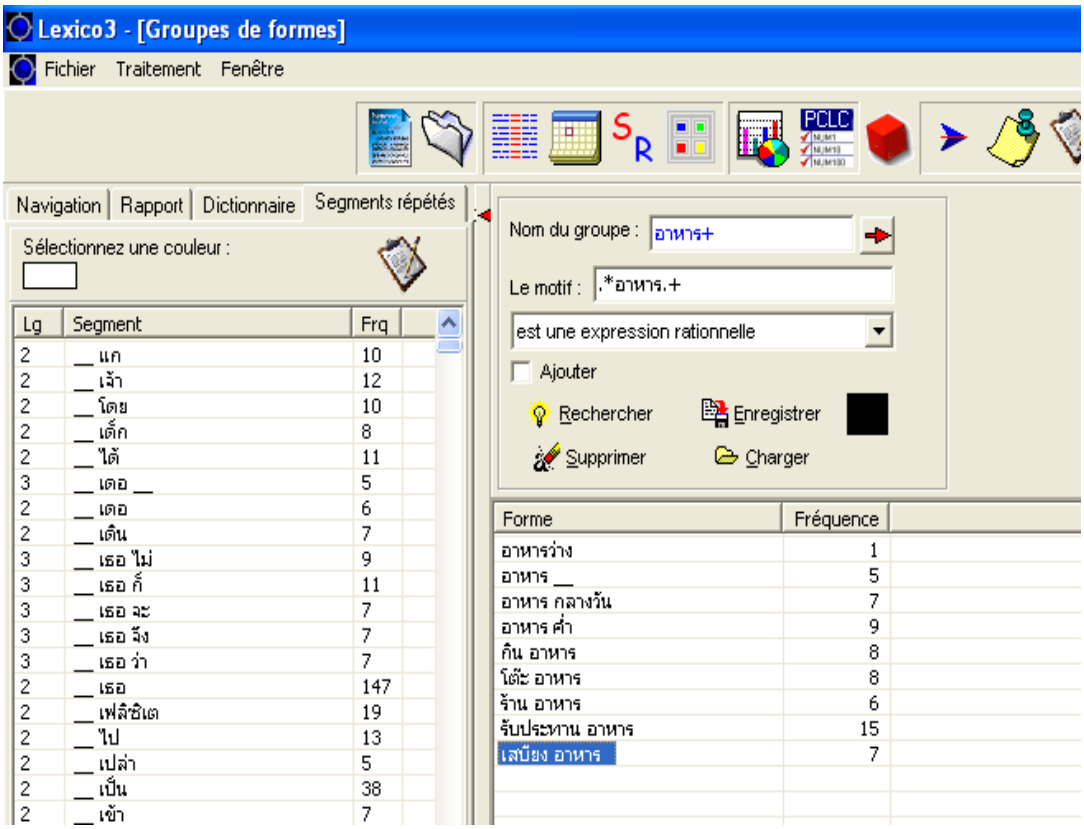


Illustration 6: Recherche groupe de formes, segments répétés, อาหาร

On voit déjà apparaître quelques segments intéressants mais pour compléter la recherche on réalise un inventaire distributionnel sur l'ensemble du groupe. Une fois que nous disposons de ces formes composées, on recherche l'expression correspondante source dans les textes français afin de déterminer le sens en contexte. Pour analyser les résultats, on construit une matrice dite de composition lexicale (cf Tableau 10) où les formes de la première colonne se combinent avec certaines formes de la première ligne pour traduire un mot source.

+	อาหาร	อาหาร กลางวัน	อาหาร เย็น	อาหาร ค่ำ
_	nourriture aliment (หุง หา+) faire la cuisine (เหล้า ย่อย+) eau de vie (เสิร์ฟ+) อาหาร servir	(ยก||)déjeuner:N		Souper:N, (เวลา+)(heure [du]) dîner:N, dîner:N
รับประทาน	(ใน ระหว่าง ที่+) [pendant le] repas:N, (ร่วม โต๊ะ+) manger à table, manger,dîner:V	(เวลา+)(heure [du])Déjeuner :N, (นั่ง ร่วม วง+||หลังจาก) Déjeuner :V,	(ร่วม+) Dîner:V (หลัง+)Dîner:N	Dîner:V
กิน	Manger:V,déjeuner:V, (+เลิศ รส) être nourrie de	(นั่ง+) déjeuner:V		Dîner:V
โต๊ะ	Table:N, (นั่ง+) [se mettre à] table			
เสบียง	Réserve:N, provisions:N, (การ รับประทาน+) [manger ses] provisions:N			
ร้าน	Cabaret, Restaurant (+ถูก ๆ) gargote:N,			

Tableau 16: Matrice de composition lexicale

Le Tableau 10 est une matrice de composition lexicale. On la lit en combinant les formes de la première colonne avec les formes de la première ligne. Par exemple _ + อาหาร = nourriture. Cela signifie que le sens de la forme à la place du caractère _ était clairement séparé de celui de la forme อาหาร. De plus une séquence entre parenthèse précise le contexte, par exemple (นั่ง+) + โต๊ะ + อาหาร = [se mettre à] table ou ร้าน + อาหาร + (+ถูก ๆ) = gargote.

D'après cette matrice en regardant la première ligne et la première colonne on peut isoler le sens de อาหาร comme étant nourriture ou aliment.

La séquence หุง หา **อาหาร** (hung ha ahan) existe en entrée de dictionnaire avec le sens de cuisiner dans un niveau de langue littéraire. On peut décomposer cette séquence de la manière suivante หุง /hung/ : cuisiner (attesté dans le même dictionnaire), หา /ha/ : Il existe en tant que verbe à multiple sens (chercher), mais je pense qu'ici il a une valeur euphonique (ha) plus que sémantique. อาหาร/ahan/ **:** ici nourriture. Il sert de complément à หุง de la même manière que pour เสิร์ฟ (servir). Il va de soi que si หา /ha/ a une valeur euphonique alors il faut considérer l'expression entière comme une seule unité lexicale.

La séquence เหล้า (liqueur) ย่อย (digérer) อาหาร (nourriture) n'est pas attestée dans les dictionnaires, mais il semble que ce soit un bon équivalent du mot digestif si on calcule le sens global à partir de chaque unité.

La forme ร้าน /ran/va permettre d'illustrer en corpus la notion de termes génériques bien connues des étudiants de thaï. En effet, à opposer à la séquence ร้าน + อาหาร (nourriture) = restaurant nous avons dans le corpus les séquences :

- ร้าน + กาแฟ (café (boisson))=café (le lieu), ร้าน /ran/+ ขาย /khaj/ (cf Tableau 11)
- ร้านค้า (boutiques, commerces, …) ค้า (V: commercer, marchander, ...)
- ร้าน : On trouve quelques occurrences isolées mais toujours en coocurrence dans un paragraphe avec une autre des formes composées. La seule séquence isolée dans une nouvelle est งาน (fêtes) ออก (sortir) ร้าน qui est utilisé pour traduire fête foraine.

+	_	เหล้า (alcool)	เนื้อ (viande)	ยา (médicament)	ของเก่า (vieille chose)	...
ร้าน + ขาย(vendre)	boutique	estaminet	boucherie	pharmacie	brocante	...

Tableau 17: Composition lexicale ร้าน + ขาย + X

On peut déterminer le sens de ร้าน /ran/ à partir de ces exemples. C'est un terme générique désignant un local dans lequel s'exerce une activité commerciale. Il peut s'utiliser avec une certaine autonomie, ce n'est donc pas un préfixe au sens de l'analyse morphologique mais la plupart du temps il est spécialisé par un ou plusieurs autres morphèmes lexicaux.

Le fait que le segmenteur a traité différemment ร้านค้า/rankha/ des autres formes composées de ร้าน s'explique certainement par le modèle statistique utilisé, basé sur le score d'information mutuelle. Toutes ces formes devraient être traitées de la même manière qu'on les considère comme une seule unité lexicale ou comme plusieurs. Si on se réfère à l'article [Asa2002], il est possible que les auteurs du segmenteur considèrent une séquence débutant par ร้าน comme un syntagme nominal et non comme une unité lexicale puisque la séquence ne fait que préciser le sens de ร้าน sans changer le concept fondamental auquel il réfère.

On a dénombré six occurrences de เสบียง /sabieng/ อาหาร répartie dans deux nouvelles dont cinq comme traductions de provisions au sens de produits alimentaires et une comme traduction de réserves au sens de réserves alimentaires en coocurrence dans le même paragraphe avec provisions. On a aussi dénombré trois occurrences de เสบียง sans อาหาร comme traduction de provisions. De plus, nous avons aussi relevé une occurrence de la séquence suivante การ/kan/ รับประทาน /rappratan/ เสบียง /sabieng/ อาหาร /ahan/ comme traduction de provisions. On a aussi relevé la séquence ตู้/tou/ เสบียง/sabieng/ comme traduction de wagon-restaurant. On en déduit donc que la présence de อาหาร à la suite de เสบียง n'est pas obligatoire à la construction du sens mais servirait plutôt un but littéraire.

Le Tableau 11 laisse clairement apparaître deux autres oppositions que nous ne détaillerons pas. Il s'agit de l'opposition entre รับประทาน /rappratan/ et กิน /kin/ qui est normalement une opposition de registre de langue, l'emploi de รับประทาน étant plus soutenu que กิน. La seconde opposition concerne อาหาร กลางวัน(milieu du jour), อาหาร เย็น (soirée:N/frais:ADJ), อาหาร ค่ำ(N:nuit) où les trois formes viennent préciser อาหาร en ajoutant une information temporelle.

Cette partie a montré quelques problèmes de composition lexicale puisque même si อาหาร /ahan/ est décrit par les spécificités et donc par le segmenteur comme une des formes les plus

représentative de l'unité lexicale, on a vu bien des cas où elle rentre en composition avec d'autres unités lexicales pour être en relation de traduction avec un seul mot français.

Toute cette analyse doit nous permettre de réinterpréter les courbes d'accroissement de vocabulaire Illustration 2 page 7. En effet, les morphèmes lexicaux, sans parler de l'instabilité de cette segmentation, n'est qu'une étape intermédiaire de la syllabe vers l'unité lexicale. L'unité lexicale regroupe les morphèmes lexicaux parmi lesquels certains on été recomposés en noms propres et en mots composés de certains types. Toutefois de nombreuses séquences pouvant être considérées comme unité lexicale vis-à-vis du référentiel sémantique français telle celle commençant par la forme ร้าน /ran/ n'ont pas été recomposées. Sous l'hypothèse que ces séquences s'apparentent à des syntagmes nominaux, ce segmenteur thaï imite les segmenteurs pour les langues à écriture segmentée en ne les recomposant pas, laissant, si besoin est, le soin à un analyseur morpho-syntaxique de les reconstituer. Mais quelle est la différence réelle entre syntagme nominal et unité lexicale dans une langue dite isolante qui n'isole rien à l'écrit ?

5 Conclusion

Cette première étude a illustré en corpus l'intrication entre les syllabes, les morphèmes lexicaux et l'unité lexicale en thaï et par conséquent certains problèmes de segmentation qui en découlent. La méthode originale d'utilisation des outils de ***Lexico3*** tel que le calcul des spécificités par partie, segmentée selon un niveau, pour faire émerger des formes spécifiques ainsi que l'utilisation des segments répétés associée à la recherche par expression rationnelle à permis de trouver des exemples pertinents.

L'analyse des formes ainsi repérées et de leurs contextes à permis de préciser la manière dont travaille le segmenteur. Ainsi les syllabes semblent correctement segmentées. La segmentation en morphèmes lexicaux ne constitue pas véritablement une analyse morphologique mais une étape intermédiaire vers la construction des unités lexicales. Enfin, il semble que la segmentation en unités lexicales ne corresponde pas à la plus grande composition lexicale possible au point de ne plus distinguer l'unité lexicale du syntagme nominal mais à la composition de morphèmes lexicaux en une unité dont le sens n'est pas vraiment calculable à partir de ceux-ci. C'est aussi l'étape de reconstitution des mots inconnus tels les noms de personnes qui sont imparfaitement mais assez bien reconstitués.

Cette étude a donc montré qu'il était possible en utilisant ce segmenteur de réaliser une étude textométrique avec ***Lexico3*** mais qu'il fallait prendre quelques précautions quant à la définition de l'unité lexicale notamment dans le cas d'étude comparative.

6 Références

[Tha1978] Kobkool THAWARANON, 1978.

[Asa2002] Nattakan Pengphon, Asanee Kawtrakul, Mukda Suktarachan, : Word Formation Approach to Noun Phrase Analysis for Thai

[Asa2003] S.P et Kawtrakul Asanee : Thai Word Segmentation based on Global and Local Unsupervised Learning.

[Kos2003] Krit Kosowat : Méthodes de segmentation et d'analyse automatique de textes thaï, thèse de doctorat Université Marne-La-Vallée.

[Ber2004] Vincent Berment : Méthodes Pour Informatiser Des Langues Et Des Groupes De Langues « Peu Dotées », thèse de doctorat Université Joseph Fourier.

Dictionnaires : HAAS, Stanford 1964, Thai-English Students dictionnary. SE-ED'S, Bangkok 2001, Modern Thai-Enlish dictionnary. , พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๔๒ (Dictionnaire en ligne de l'institut royal 2542 : http://rirs3.royin.go.th/dictionary.asp)