Tutoriel n°3:

Exploration du corpus Traductions alignées du discours d'investiture de B. Obama

Corpus alignés, méthodes textométriques pour l'alignement

[Obama1]

Sommaire

Objectifs

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.

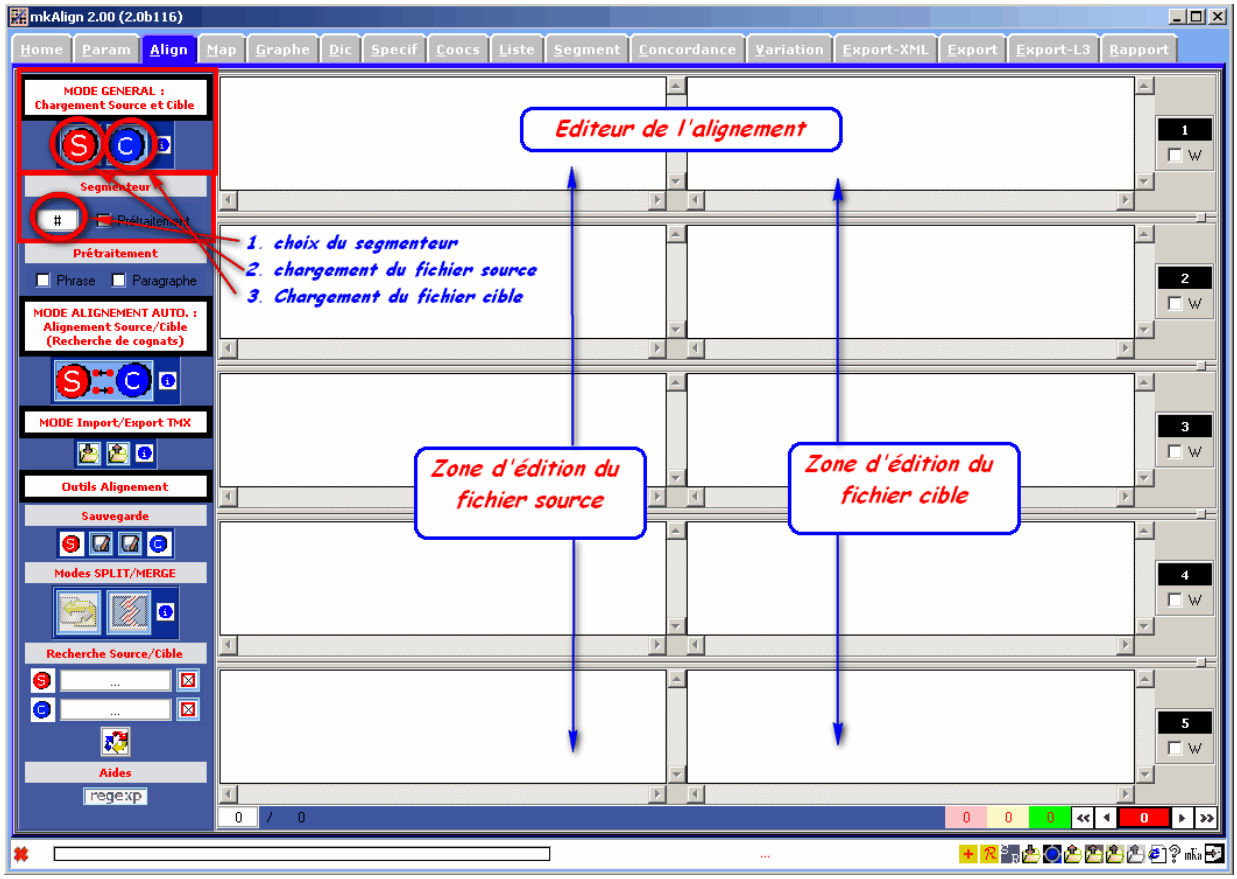


Figure 1 : Interface de l'alignement avec mkAlign

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,

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

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

pour amorcer grossièrement les choses, à charger les 2 volets initiaux en choisissant comme *segmenteur d'alignement* le caractère retour à la ligne.

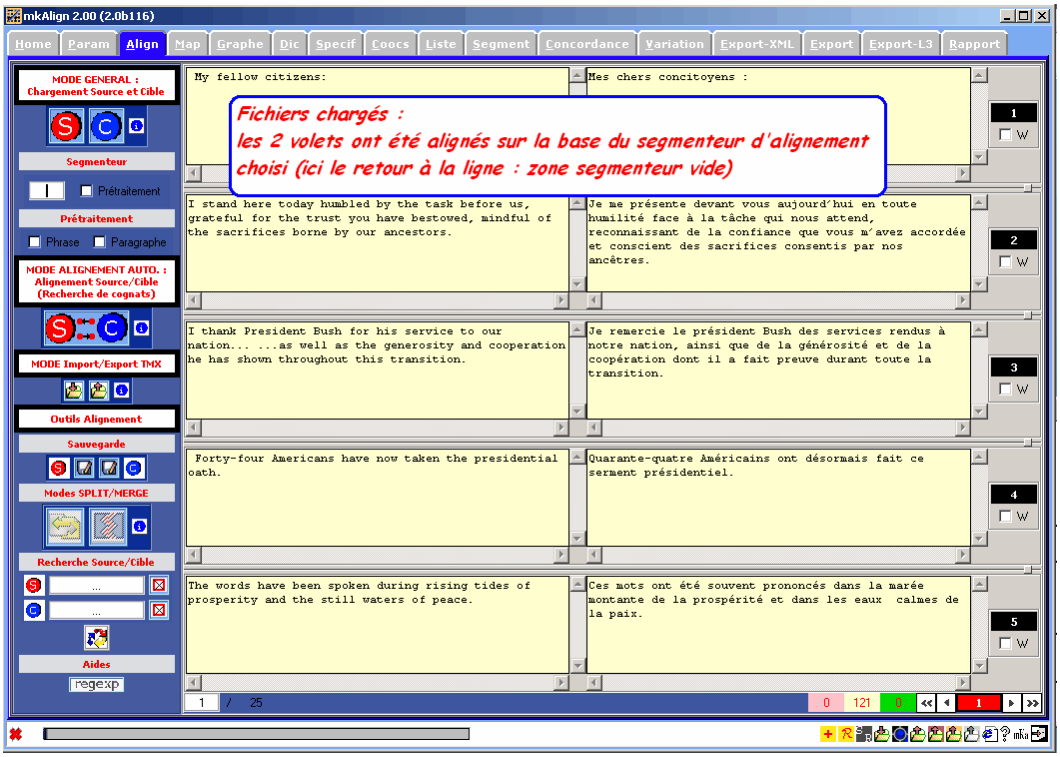


Figure 2 : Chargement des fichiers à aligner

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.

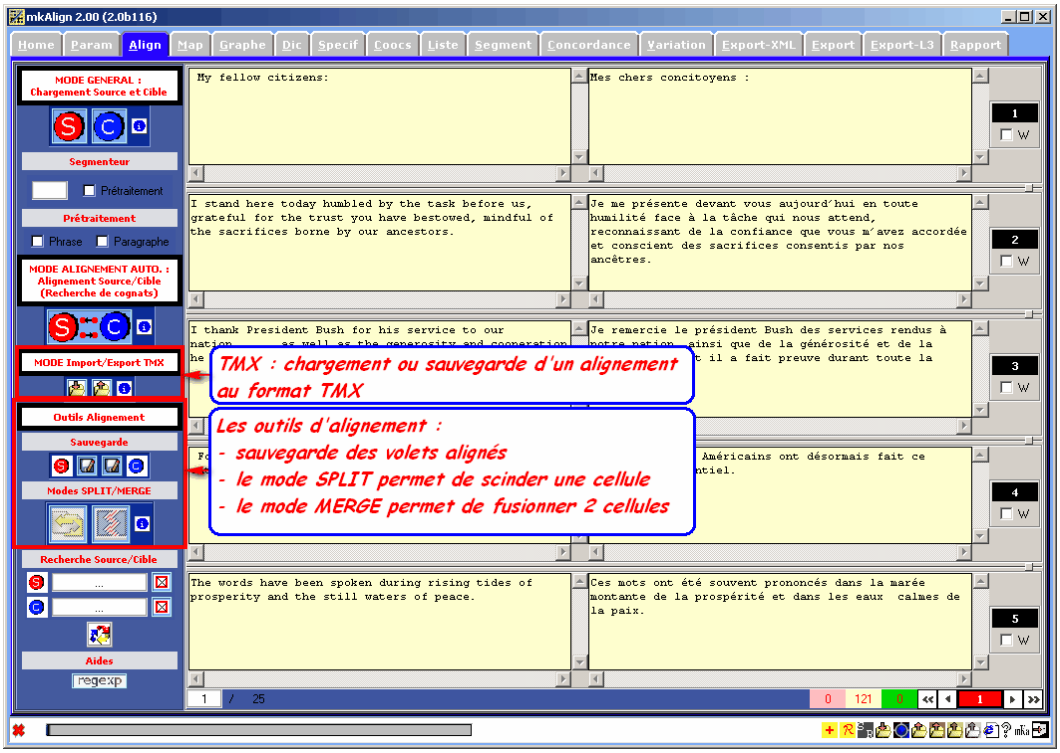
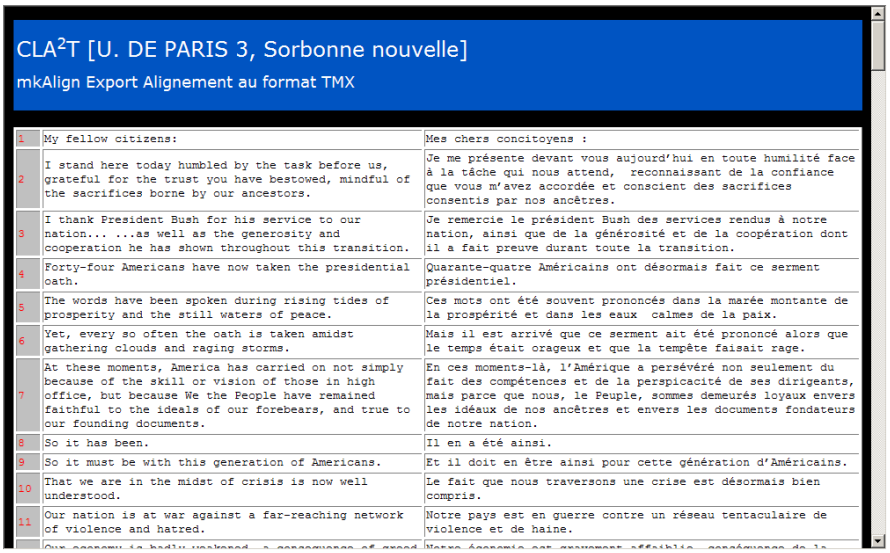


Figure 3 : Les outils de l'alignement

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 :



CLA²T [U. DE PARIS 3, Sorbonne nouvelle]

mkAlign Export Alignement au format TMX

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

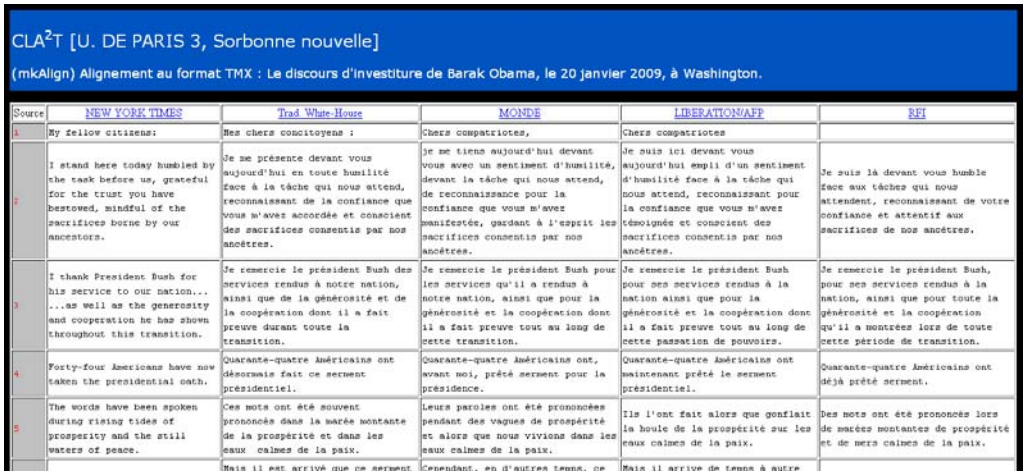
La seconde montre un extrait du code source de ce fichier au format TMX :

```
<?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.

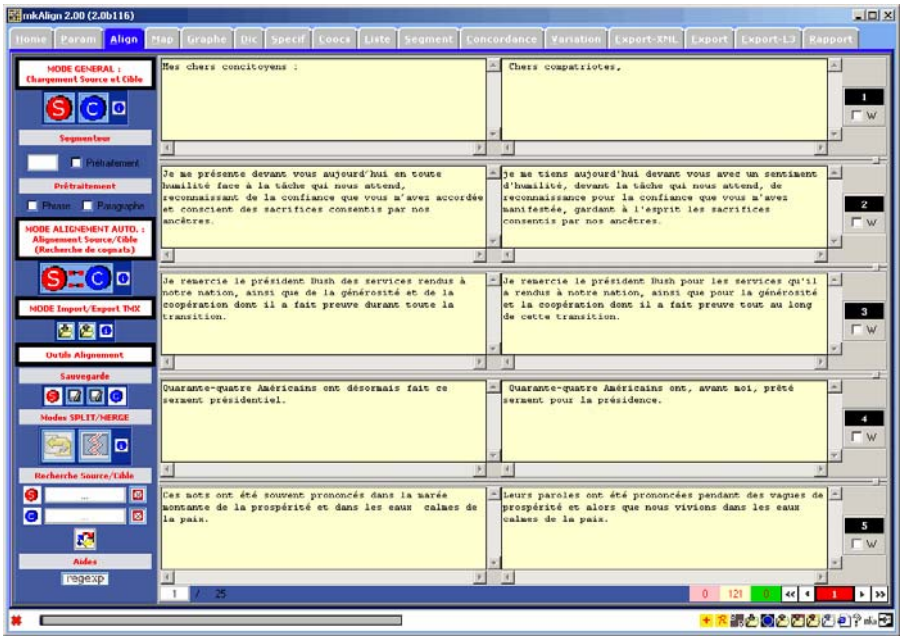


Figure 7: Alignement Volets FR-1 et FR-2

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*) :

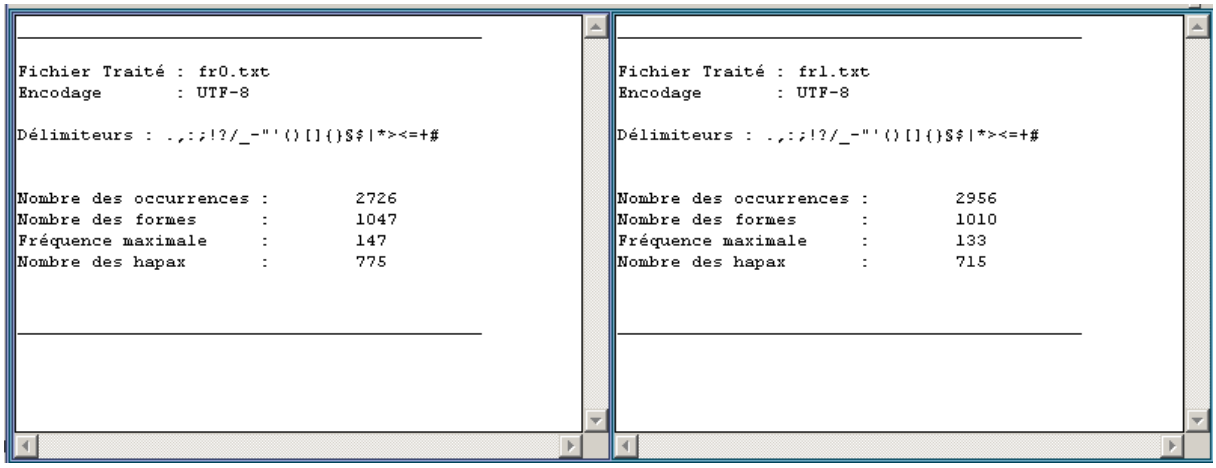


Figure 8: Paramètres lexicométriques des deux volets alignés

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

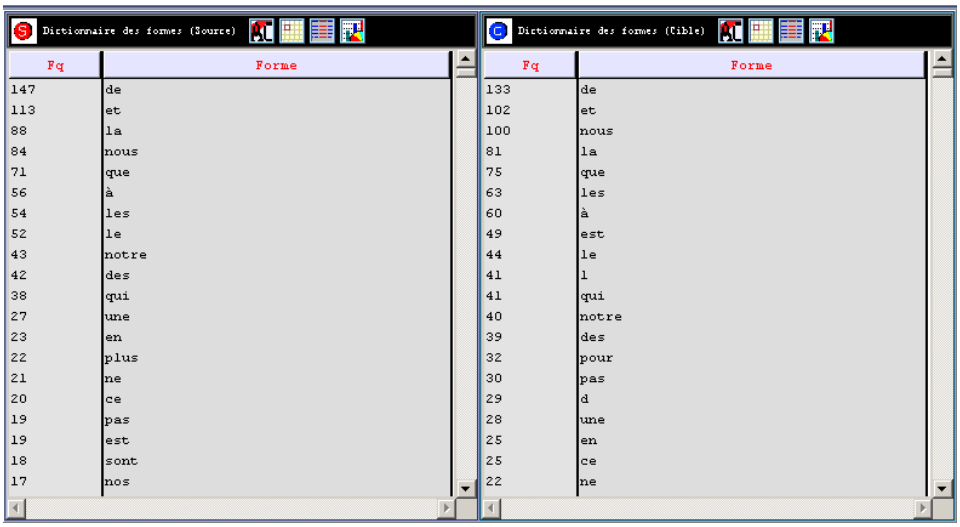


Figure 9: Les dictionnaires de formes issus de l'alignement

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

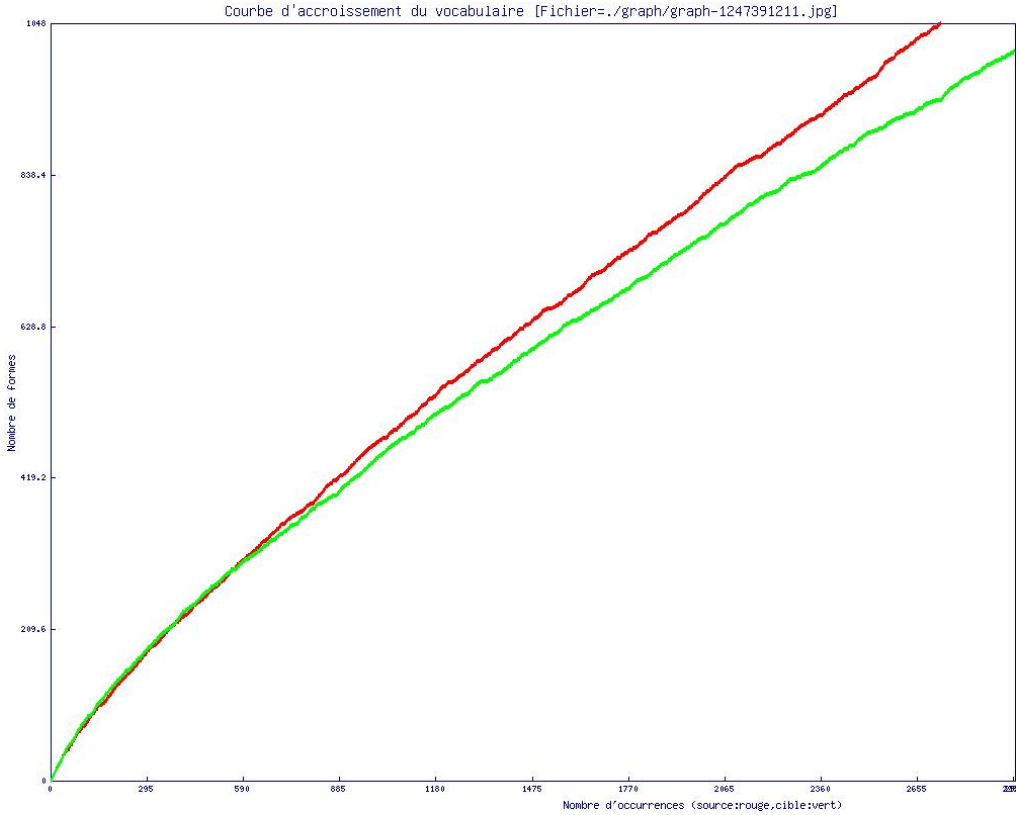


Figure 10 : Accroissement de vocabulaire sur les 2 volets de l'alignement

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.

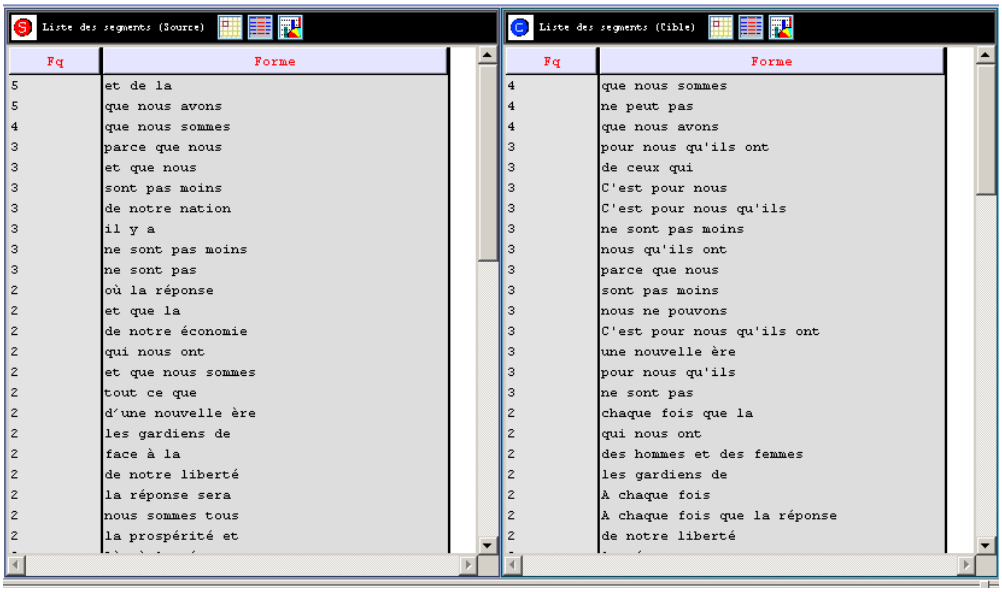


Figure 11: Liste des segments répétés sur les 2 volets du corpus

==== ***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.

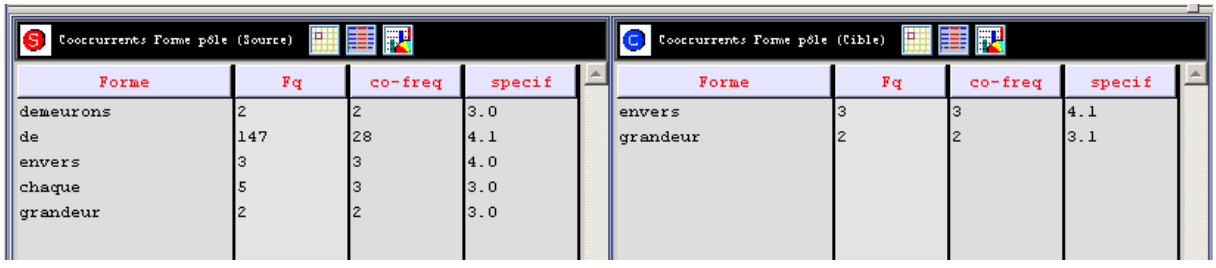


Figure 12 : Les cooccurrents de "nation"

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*

[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.

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

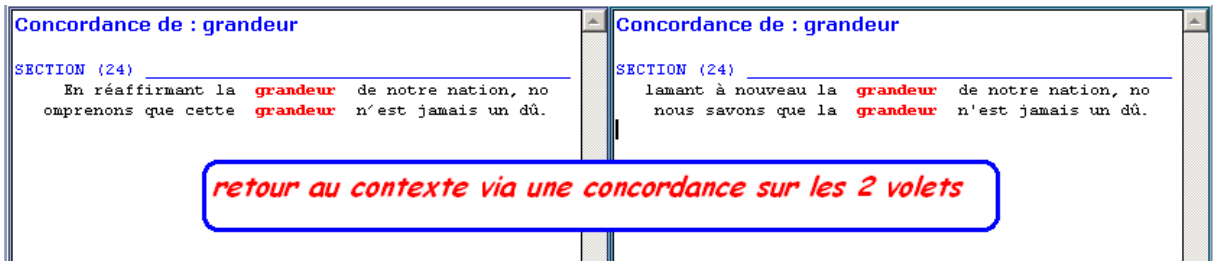


Figure 13 : Retours aux contextes

==== *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).

- 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 :

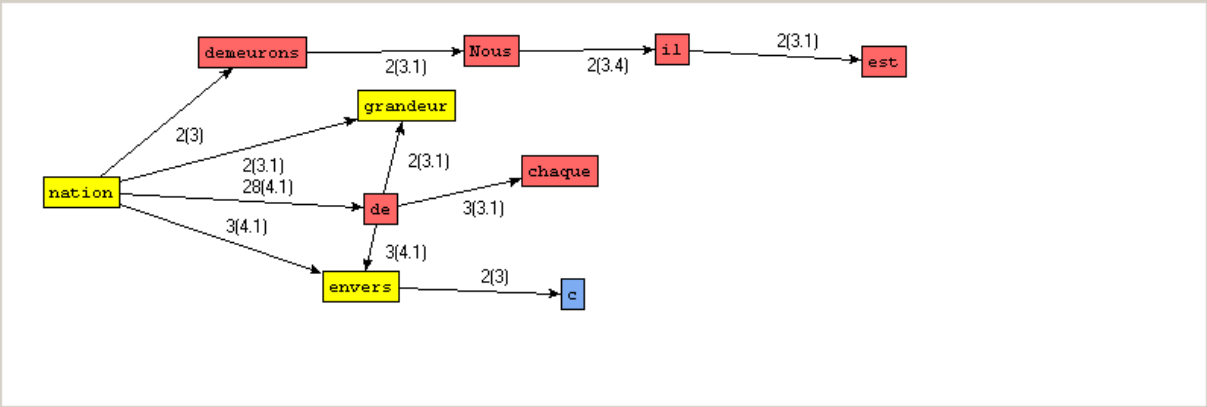


Figure 14 : Les polycooccurrents de la forme "nation"

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**

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

[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).

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.

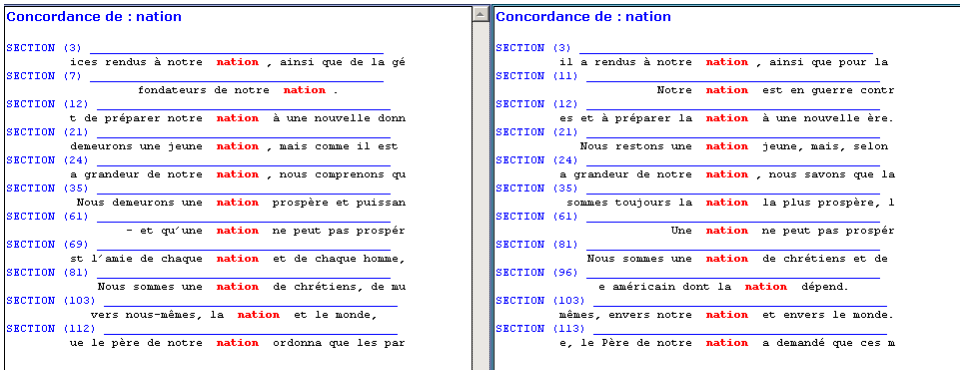


Figure 15 : Concordance de la forme nation sur les 2 volets du corpus

==== ***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 :

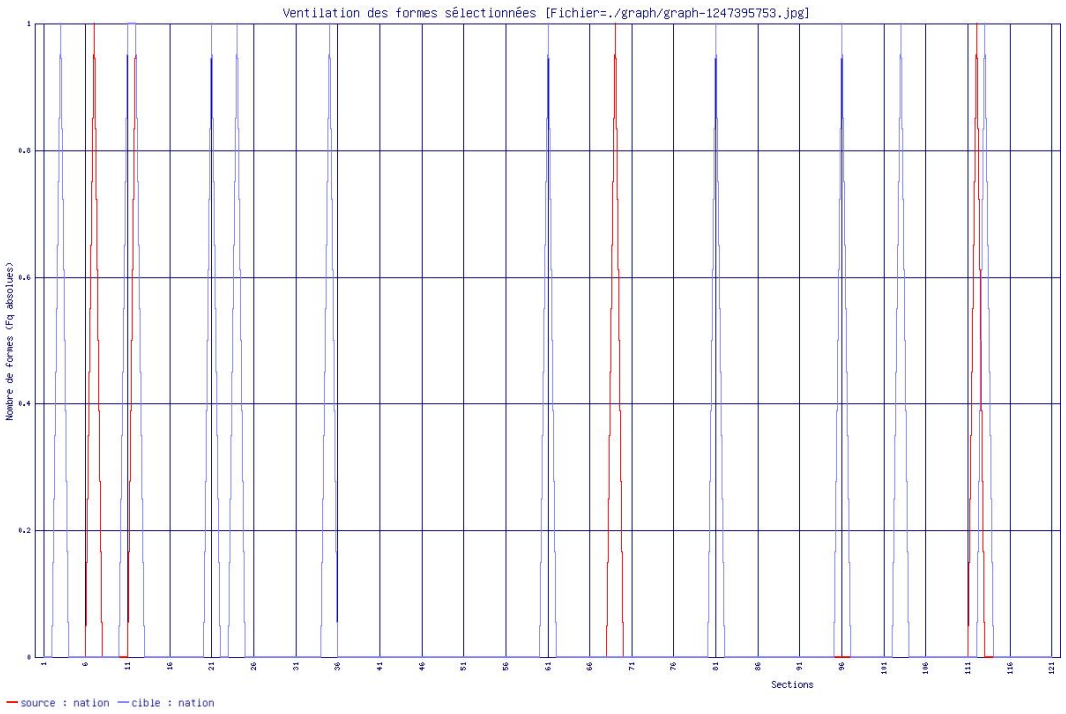


Figure 16 : Ventilation de la forme « nation » sur les 2 volets de l'alignement

==== *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

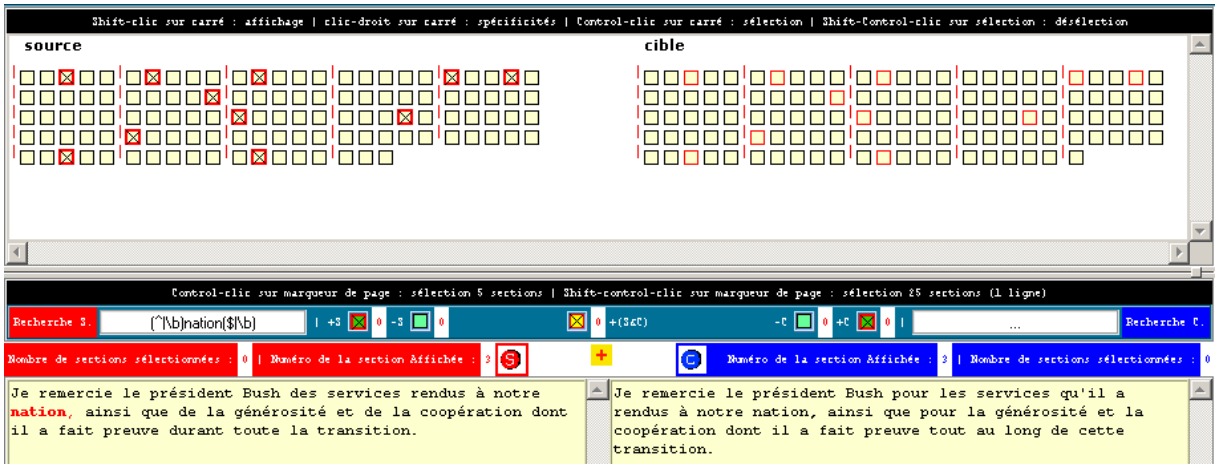


Figure 17 : Carte des sections ; projection de la forme "nation" sur le volet source

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.

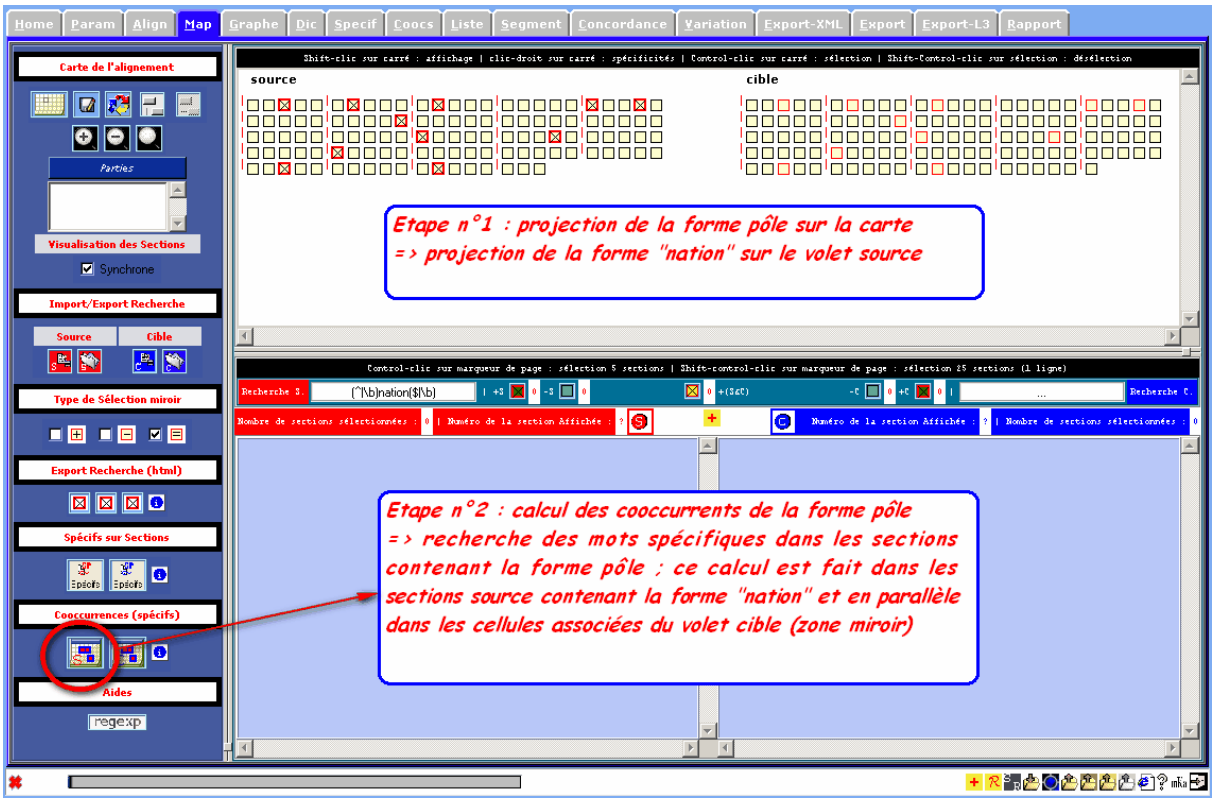


Figure 18 : Calcul des cooccurrents d'une forme par la carte des sections

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*)

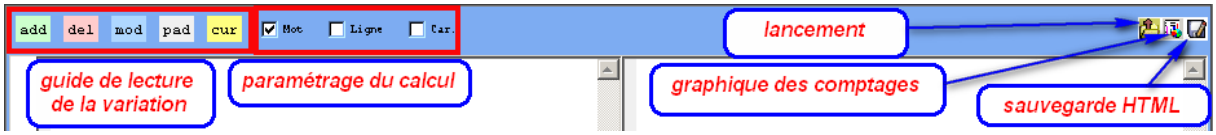


Figure 20 : paramétrage du calcul de la variation

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

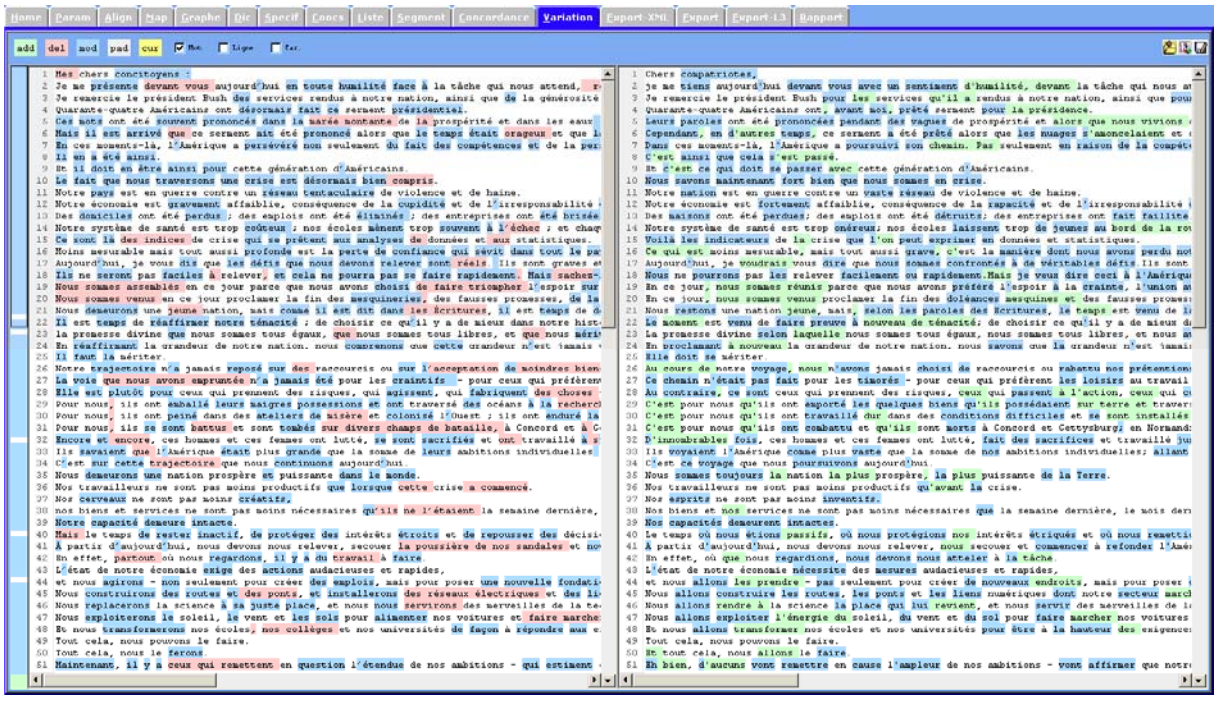


Figure 21 : Mise au jour de la variation (comparaison au niveau du mot)

[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

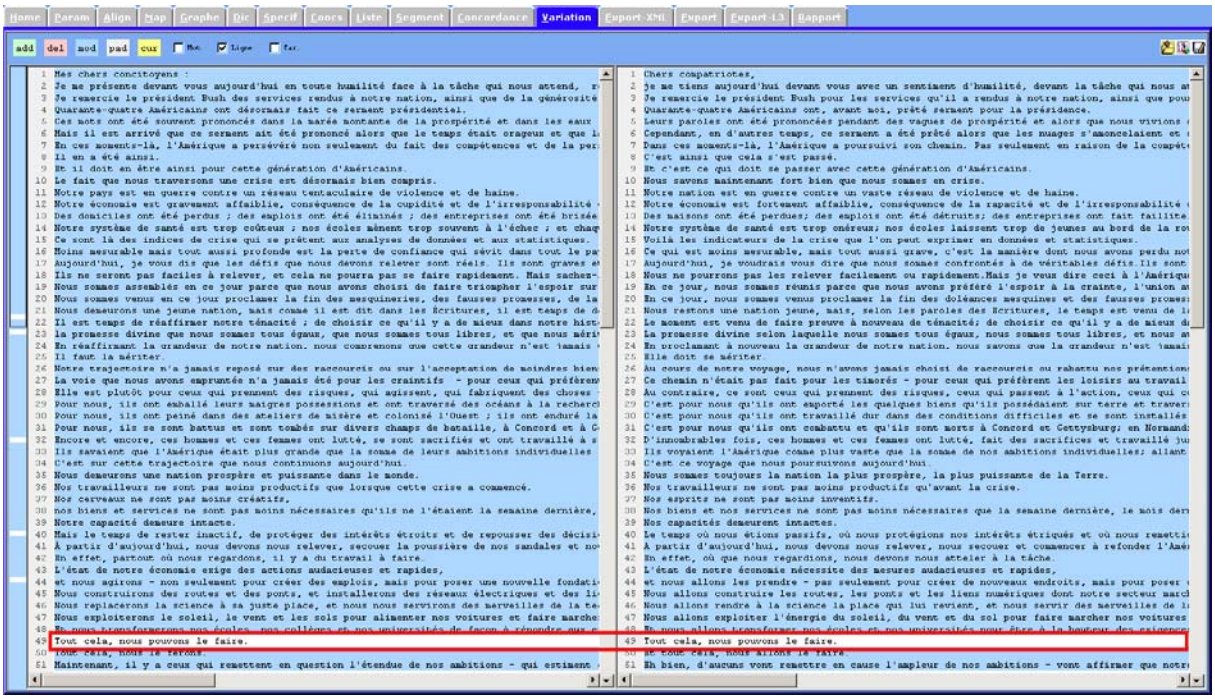


Figure 22 : Mise au jour de la variation (comparaison au niveau des lignes)

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.

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.