Nouvelles perspectives en sciences sociales
Revue internationale de systémique complexe et d'études relationnelles

Perspectives de l'architecture *Trame/Cadre* pour les alignements multilingues
Perspectives on the *Frame/Thread* Architecture for Multilingual Alignments

Maria Zimina and Serge Fleury

Volume 11, Number 1, November 2015

Sur le thème de l'analyse de données textuelles informatisée

URI: https://id.erudit.org/iderudit/1035940ar
DOI: https://doi.org/10.7202/1035940ar

See table of contents

Publisher(s)

Prise de parole

ISSN

1712-8307 (print)
1918-7475 (digital)

Explore this journal

Cite this article

Zimina, M. & Fleury, S. (2015). Perspectives de l'architecture *Trame/Cadre* pour les alignements multilingues. *Nouvelles perspectives en sciences sociales*, *11*(1), 325–353. https://doi.org/10.7202/1035940ar

Article abstract

Multilingual text alignment is challenging due to the complexity of text and discourse organisation. Multilingual textual space can be explored using a textometric data model (*Thread/Frame*). A *Thread* is a textual flow represented as a system of items with position identifiers. A *Frame* is used to locate different textual objects (*containers* and *contents*) and their contexts. Following these principles, all text parts and annotations (including alignments) are stored and exchanged through different computerised procedures. *Incremental textual resources* trace all processing steps (from the initial segmentation to subsequent explorations and quantitative analyses). The software implementation of this model in *Le Trameur* allows exploring richly annotated multilingual text corpora (*treebanks*).

Perspectives de l'architecture *Trame/Cadre* pour les alignements multilingues

Maria Zimina
Université Paris Diderot – Paris 7
Serge Fleury
Université de la Sorbonne nouvelle – Paris 3

Alignements de corpus : enjeux actuels

L'analyse automatique de corpus multilingues parallèles et comparables[1] a beaucoup progressé ces dernières années[2]. Les avancées concernent aussi bien les couples de langues analysées, que les types de relations que l'on parvient à repérer entre les unités de bi-textes. Malgré le volume important de ressources textuelles multilingues et la diversité des outils informatiques utilisés pour le traitement de ces données, la formalisation du concept d'alignement textuel n'est pas encore aboutie, ni sur le plan linguistique, ni sur le plan de sa modélisation informatique. Il s'agit pourtant d'une direction de recherche importante en sciences humaines. En effet, l'alignement textuel demeure au cœur de plusieurs problématiques en analyse de discours, traduction, terminologie, linguistique contrastive, etc.

1 Les corpus multilingues comparables autorisent des comparaisons au plan traductionnel sans être des traductions.

2 Jörg Tiedemann, *Bitext Alignment. Synthesis Lectures on Human Language Technologies*, San Rafael (CA), Morgan and Claypool Publishers, 2011.

Le concept d'alignement suggère une forme d'organisation des ensembles textuels « équivalents ». Souvent perçu comme un appariement des contextes, la modélisation de ces procédés reflète pleinement la complexité de l'organisation textuelle et discursive. Au lieu de se référer à l'*alignement* textuel, il serait d'ailleurs plus juste d'évoquer des *alignements* multiples liés aux objectifs de l'exploration textuelle. Vus sous cet angle, ces alignements sont des mises en relation d'objets textuels qui se correspondent. La nature des objets visés par ces relations est extrêmement variable et dépend des objectifs de l'analyse. Par ailleurs, ces relations ne suivent pas nécessairement la structure hiérarchique des textes. Par exemple, on peut établir une relation entre un segment textuel qui apparaît dans l'introduction du texte original avec un ou plusieurs segments faisant partie de la section qui suit l'introduction dans la version traduite.

Sur papier, pour représenter les alignements, on est le plus souvent contraint à établir une hiérarchie des liens sous la forme de relations partie-tout ou, à défaut, à envisager un système d'index[3]. Les représentations des textes parallèles que l'on trouve dans des manuscrits anciens illustrent ces principes : les textes sont justifiés, c'est-à-dire adaptés le plus régulièrement possible au cadre de bi-feuille/bi-texte[4]. Pour y parvenir, les copistes adaptaient la forme du texte, en intégrant des abréviations, en jouant sur les espaces entre les mots ou en faisant varier la régularité du tracé des lettres par compression ou dilatation. Ces principes sont encore utilisés de nos jours pour adapter dynamiquement la représentation des passages textuels à l'écran. Ils permettent de gérer notamment les paramètres d'affichage des

3 Hatem Ghorbel et Giovanni Coray, « L'alignement multicritères des documents médiévaux », *Lexicometrica*, nº spécial *Corpus Alignés*, 2002, http://lexicometrica.univ-paris3.fr/thema/thema6.htm, site consulté le 10 octobre 2015.

4 On peut citer ici l'exemple de la version bilingue de l'*Iliade* disponible à la BAV (Biblioteca Apostolica Vaticana), https://www.actualitte.com/education-international/des-centaines-de-manuscrits-du-vatican-deja-numerises-par-ntt-53331.htm, site consulté le 10 octobre 2015. Les deux versions de l'*Iliade* sont en regard l'une de l'autre et synchronisées : la page de gauche est en grec, et celle de droite en latin, afin de permettre au lecteur de comparer.

éléments textuels lorsqu'une fenêtre graphique a une taille limitée.

La numérisation des corpus de textes a permis d'élargir considérablement les possibilités de représentation informatisée des liens textuels[5]. On assiste au développement de champs de recherches riches de perspectives pour la représentation des alignements, réduisant considérablement des contraintes habituellement liées aux limitations physiques du support papier. Pour exploiter pleinement ces opportunités inédites, il appartient aux chercheurs en sciences humaines de repenser le concept d'alignement textuel en faisant abstraction des contraintes liées au support papier. Cette tâche n'est pas aisée car elle nécessite des efforts multiples : une réflexion conceptuelle sur la modélisation du bi-texte informatisé, débutée par Brian Harris[6] dans les années 1980, une formalisation d'un corps de méthodes qui permettent d'interagir avec le bi-texte, un changement de paradigme dans le protocole d'expérimentation, s'appuyant sur des résultats d'analyses quantifiables et reproductibles, indexés sur les résultats d'alignements textuels.

Fort heureusement, on commence à distinguer des signes d'une transition progressive dans ce sens, comme en témoignent des projets pluridisciplinaires récents, tels que *TransRead*[7] qui s'intéresse à la visualisation de textes bilingues et des alignements qui les lient. L'un des objectifs du projet est le développement des stratégies d'exploration de bi-textes dans les nouvelles applications multilingues, rendues possible grâce aux avancées des techniques de visualisation de l'information et à l'arrivée d'équipements mobiles (tablettes tactiles, lecteurs électroniques). Parallèlement,

---

5 Sur ces questions, on consultera, par exemple, Clément Pillias, « Reading Bilingual Texts with Digital Tools : A State of the Art », *Projet ANR 2012 CORD 015 TransRead, Lecture et interaction bilingues enrichies par les données d'alignement*, Deliverable 2.1, 2014, http://transread.limsi.fr/Deliverable2.1.pdf, site consulté le 10 octobre 2015.

6 Brian Harris, « Bi-text : A New Concept in Translation Theory », *Language Monthly,* nº 54, 1988, p. 8-10.

7 ANR 2012 CORD 015, *TransRead, Lecture et interaction bilingues enrichies par les données d'alignement*, 2015, http://transread.limsi.fr/, site consulté le 10 octobre.

une autre réflexion est en cours au sein du projet *Rhapsodie*[8] sur les unités linguistiques et les annotations textuelles multiples (prosodiques, syntaxiques, etc.) avec une volonté de tester les apports croisés de plusieurs cadres théoriques, et sur la façon dont ils éclairent le même jeu de données échantillonné en différents genres discursifs. Ces initiatives témoignent de la richesse des problématiques qui alimentent la réflexion liée à la représentation des alignements et la modélisation des structures textuelles sous-jacentes.

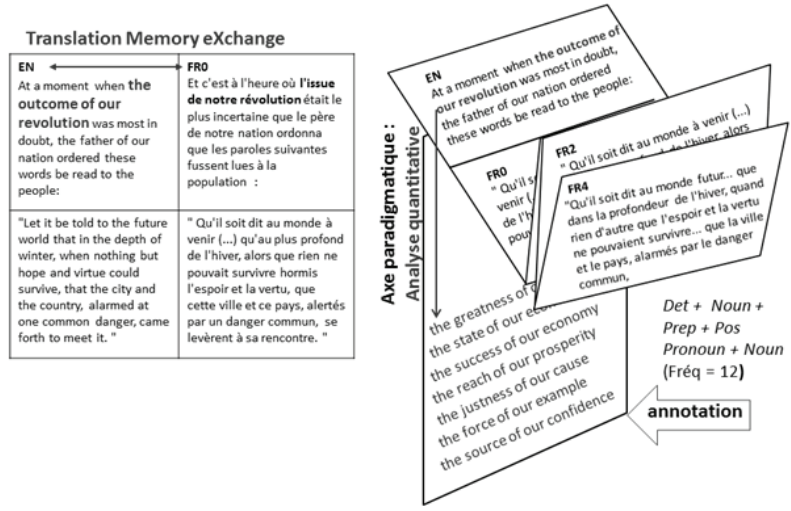
En même temps, la disponibilité des corpus alignés amène les questions d'exploitation et de réutilisation de ces ressources traductionnelles. En TAO (Traduction Assistée par Ordinateur), les MT (Mémoires de Traduction) constituées de segments textuels alignés sont largement utilisées pour faciliter la traduction de nouveaux documents connexes. Dans ce contexte, l'avancée principale est sans doute le développement des standards d'échanges. Par exemple, les formats TMX (*Translation Memory eXchange*) et TBX (*TermBase eXchange*) ont facilité les pratiques de partage de ressources traductionnelles et terminologiques existantes[9]. Toutefois, ce type de standards ne couvre qu'une partie des besoins liés aux échanges d'objets textuels que l'on parvient à identifier en explorant les contextes multilingues. Notamment, il n'est pas toujours aisé d'interagir autour des correspondances traductionnelles multiples, tout en préservant des liens dynamiques qui existent entre elles à plusieurs niveaux d'analyse (mots, syntagmes, phrases, cadres discursifs, etc.), comme l'illustre la figure 1.

8 ANR 07 CORP 030 01, *Rhapsodie, Corpus prosodique de référence en français parlé*, 2015, http://www.projet-rhapsodie.fr/, site consulté le 10 octobre 2015.

9 Pour plus d'information, Harold Somers, *Computers and Translation : A translator's guide*, Amsterdam, John Benjamins B.V., 2003.

Figure 1

Prise en compte de l'axe paradigmatique et des correspondances multiples dans les Mémoires de Traduction



Pour faciliter l'accès aux ressources, l'exploration de corpus textuels peut se faire au sein des plateformes unifiées qui s'orientent vers la standardisation de l'annotation des données textuelles et leur exploitation à l'aide de langages de requêtes[10]. Cependant, les objets construits à l'aide de requêtes sont de nature variable et ne peuvent pas être confondus dans les mêmes décomptes au cours de l'analyse quantitative. Afin de préciser davantage les apports croisés entre qualitatif et quantitatif, nous menons une réflexion sur le stockage de la structure du matériau

10 On peut citer dans ce contexte des systèmes, tels que :
PDT (The Prague Dependency Treebank 2.0), 2015, ufal.mff.cuni.cz/pdt2.0, site consulté le 10 octobre 2015.
GATE (General Architecture for Text Engineering), 2015, http://gate.ac.uk/gate, site consulté le 10 octobre 2015.
ANNIS (ANNotation of Information Structure), 2015, http://annis-tools.org/, site consulté le 10 octobre 2015.
MACAON, chaîne de traitement, 2015, http://macaon.lif.univ-mrs.fr/, site consulté le 10 octobre 2015.

bi-textuel informatisé. Cette réflexion s'appuie sur le développement d'un modèle de données issu des recherches menées en *textométrie*[11].

Dans un premier temps, nous présenterons les principes opérationnels de la textométrie dans le contexte particulier de l'analyse de corpus alignés. Ensuite, la mise en œuvre concrète de ces principes sera illustrée à l'aide des fonctionnalités du logiciel de textométrie *Le Trameur*[12]. Enfin, nous montrerons les avancées récentes de la textométrie multilingue dans le cas de l'analyse des corpus richement annotés, avec l'utilisation des relations syntaxiques pour guider le processus d'alignement au niveau des mots.

Principes opérationnels de la *textométrie* multilingue

Système d'unités et alignements

La *textométrie* propose des méthodes quantitatives permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur l'ensemble des unités textuelles d'un corpus. La démarche textométrique repose principalement sur l'observation des variations de fréquence d'unités textuelles (formes, lemmes, etc.) appelées *contenus* textuels, dans les différentes parties d'un ensemble de textes, considérées comme des *contenants* textuels (parties, sections, zones, chapitres, paragraphes, phrases, séquences, etc.)[13]. Une description formelle de ces deux systèmes d'unités (*contenants* et *contenus*) permet d'obtenir, à l'aide de procédures informatisées, des décomptes sous forme de vastes tableaux statistiques. La *textométrie* mobilise des méthodes d'ADT (Analyse de Données Textuelles) afin

11 Ces recherches sont diffusées, par exemple, par la revue électronique *Lexicometrica*, 2014, http://lexicometrica.univ-paris3.fr/, site consulté le 10 octobre 2015.

12 Serge Fleury, *Le Trameur*, 2015, http://www.tal.univ-paris3.fr/trameur/, site consulté le 10 octobre 2015.

13 Pour en savoir plus sur l'approche textométrique, on consultera l'ouvrage de Ludovic Lebart et André Salem, *Statistique textuelle*, Paris, Dunod, 1994. Sur la formalisation des principes adoptés dans ce travail, on se reportera à : Keyser Söze-Duval, *Pour une textométrie opérationnelle*, 2008, http://www.tal.univ-paris3.fr/trameur/RTI6provisoire.doc, site consulté le 10 octobre 2015.

d'étudier la répartition statistique des *contenus* au sein des *contenants* des corpus textuels. Les synthèses statistiques qui en résultent sont des points de départ pour la mise en évidence des principales dimensions de variation des corpus analysés.

Dans un contexte multilingue, cette analyse des variations de fréquences des unités textuelles fournit un éclairage précieux sur les ressemblances et les oppositions entre les volets multilingues et crée une base d'analyse contrastive. Appuyée sur des restitutions du contexte autour des points saillants des textes, l'analyse textométrique met en évidence les spécificités discursives de chacun des volets du corpus et fait apparaître leurs correspondances et dissonances[14].

Pour y parvenir, on opère sur chaque volet du corpus multilingue un double repérage : le découpage en *contenants* et la segmentation en *contenus*. Les alignements d'empans textuels s'appuient sur le repérage des *contenants* qui se correspondent (sections, paragraphes, phrases). Les alignements de *contenus* (formes, lemmes, segments, etc.) s'appuient sur la segmentation de la chaîne textuelle, le typage des unités génériques et l'analyse de la répartition statistique dans chaque volet du corpus.

Trame et *Cadre*

La segmentation de la chaîne textuelle peut commencer par la définition d'une liste de caractères délimiteurs qui permet un premier découpage en unités disjointes appelées *items*. À partir d'un texte découpé en items, on constitue un système de coordonnées (*Trame*) où chaque item est repéré par son numéro d'ordre dans la chaîne textuelle.

Chaque *item* isolé peut recevoir une étiquette (lemme, catégorie grammaticale, syntaxique, sémantique, stylistique, etc.) au cours d'une ou plusieurs annotations. Les items semblables sont rattachés à un même *type* (à partir des propriétés de leur forme intrinsèque

14 Sur l'utilisation des méthodes textométriques dans le contexte multilingue : Maria Zimina, « Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles », thèse pour le Doctorat en Sciences du langage, Paris, Université de la Sorbonne nouvelle – Paris 3, 2004.

ou sur la base de certaines annotations). Les *types* deviennent ainsi des unités génériques dont on peut recenser les occurrences.

Les empans textuels (parties) sont indexés sur la *Trame* comme suites d'*items* consécutifs, entre la position x_1 et la position x_2. Les systèmes de *contenants* du corpus sont regroupés dans une structure de données appelée *Cadre*. Une ressource textuelle constituée sous la forme de *Trame/Cadre* est une *base textométrique*[15]. Un extrait d'une *base textométrique* bilingue anglais/français est présenté sur la figure 2. Le stockage de données textuelles multilingues à l'aide de l'architecture *Trame/Cadre* peut servir de base à toute exploration textométrique.

Objets génériques *Sélections*

La création d'une *base textométrique* permet de repérer les *contenants* et *contenus* nécessaires à l'analyse textométrique : ce sont des *Sélections*, objets génériques de la textométrie. Les *Sélections* de *contenus* sont des *items* correspondant aux occurrences d'un *type* (forme, lemme, patron morphosyntaxique, expression régulière croisant plusieurs annotations). Les *Sélections* de *contenants* sont constituées d'*items* connexes (zones, parties, sections, paragraphes). Indexées sur une *Trame* commune, les *Sélections* sont analysées au sein des tableaux croisant les décomptes de chacun des *types* (*contenus*) dans chacune des parties (*contenants*). Elles sont transmises entre procédures de traitement : à partir de *Sélections* initiales, les méthodes textométriques produisent de nouvelles *Sélections* en fonction des objectifs visés.

Résonance textuelle et alignements

Dans le contexte multilingue, des *contenants* issus de chaque volet du corpus se correspondent en fonction des alignements réalisés (chapitres, sections, paragraphes, phrases, etc.). Les logiciels de *textométrie* gérant des corpus alignés peuvent visualiser ces correspondances par exemple à l'aide d'une *carte des sections*

15 Serge Fleury, *Le Trameur. Propositions de description et d'implémentation des objets textométriques*, 2013, http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf, site consulté le 10 octobre 2015.

parallèles[16]. Cette approche topographique des alignements est fondée sur les principes de la *résonance textuelle*[17]. On passe d'une liste d'unités (*types*) appartenant à l'un des volets du corpus à des *types* qui lui correspondent dans un autre volet à travers une *Sélection* par seuillage, opérée par un calcul statistique (méthode des *spécificités*[18]). Les résultats permettent d'apparier les *types* les plus caractéristiques de ces zones en correspondance. Parallèlement, les dissonances constatées amènent de nouvelles explorations et de nouveaux parcours interprétatifs.

Dans la section qui suit, nous montrons comment exploiter une *base textométrique* bilingue, gérer des *Sélections* et alignements des types en correspondance à l'aide des fonctionnalités du logiciel *Le Trameur*[19] qui fonctionne selon les principes décrits plus haut.

Le Trameur

Base textométrique anglais/français *Investiture Obama*

Pour illustrer la construction d'une base textométrique, nous utilisons ici l'exemple du corpus bilingue anglais/français *Investiture Obama*. Ce corpus compte approximativement 30 000 occurrences et comprend 5 volets : le discours original en anglais prononcé par Barack Obama le 20 janvier 2009 à Washington, publié sur le site de *The New York Times* (EN), et 4 traductions françaises de ce discours (*FR0-1-2-3*). Les traductions ont été récupérées sur le site officiel de la Maison blanche (*FR0*), sur les sites des journaux français *Le Monde* (*FR1*) et *Libération* (*FR2*),

16 L'approche topographique a été implémentée dans le logiciel *Lexico3*, puis développée et élargie aux corpus multilingues au sein des outils *mkAlign* et *Le Trameur* développés par l'équipe SYLED-CLA2T, Université Sorbonne nouvelle-Paris 3.

17 Maria Zimina et Serge Fleury, « Approche systémique de la résonance textuelle multilingue », dans Émilie Née *et al.* (dir.), *Actes JADT 2014* « *Journées Internationales d'Analyse Statistiques des Données Textuelles* », JADT.org, 2014, p. 717-728.

18 La méthode des *spécificités* est présentée, par exemple, dans l'ouvrage de Ludovic Lebart et André Salem, *op. cit.*

19 Serge Fleury, *Le Trameur*, 2015, *op. cit.*

ainsi que sur le site de *Radio France International* (*FR3*)[20]. Le corpus a été aligné au niveau de la phrase à l'aide du programme *mkAlign*[21]. Chaque volet a été étiqueté avec *Le Trameur* à l'aide du programme *treetagger*[22] puis converti en une *base textométrique*. Les 5 bases ont été agrégées en gardant les mêmes niveaux d'annotations pour l'anglais et le français (forme, lemme, catégorie morphosyntaxique)[23]. Le stockage de données bi-textuelles suit le modèle *Trame/Cadre*. La figure 2 montre un extrait de la base : les 5 volets du corpus sont matérialisés sur une *Trame* commune. L'alignement des phrases est indiqué par le délimiteur « § ». Sur la figure 3, on distingue un système de 5 *contenants*, un pour chaque volet du corpus; ces *contenants* structurent le *Cadre*.

---

20 Serge Fleury, « Exploration du corpus Traductions alignées du discours d'investiture de B. Obama (Tutoriel nº 3, Explorations textométriques avec mkAlign) », *Lexicometrica*, nº spécial *Explorations textométriques*, 2009, http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm, site consulté le 10 octobre 2015.

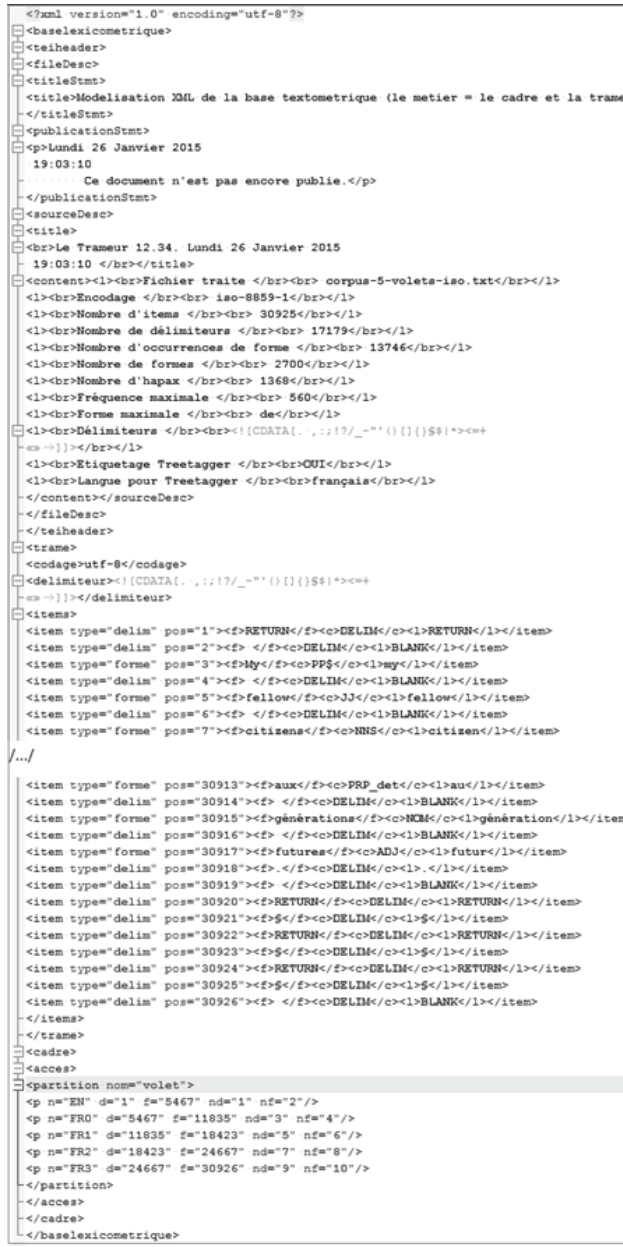
21 *mkAlign* est conçu pour aider l'utilisateur dans l'alignement, la correction et la validation de textes traduits. L'utilisateur garde la maîtrise sur l'ensemble de ces processus, depuis la mise en correspondance initiale des segments équivalents jusqu'à l'export final du bi-texte produit, comme cela est illustré dans Serge Fleury et Maria Zimina, « Exploring translation corpora with MkAlign », dans Gabe Bokor (dir.), *Translation Journal*, 2007, http://translationjournal.net/journal/39mk.htm, site consulté le 10 octobre 2015.

22 *Le Trameur* intègre par défaut le programme *treetagger* : un système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation développé par Helmut Schmid, « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference « New Methods in Language Processing »*, Manchester, UMIST, p. 44-49, 1994.

23 Toutes les étapes de la construction d'une base textométrique de textes bilingues alignés sont détaillées dans Serge Fleury, *Base textométrique de textes alignés*, 2014, http://www.tal.univ-paris3.fr/trameur/MAJ-12.02.pdf, site consulté le 10 octobre 2015.

Figure 2

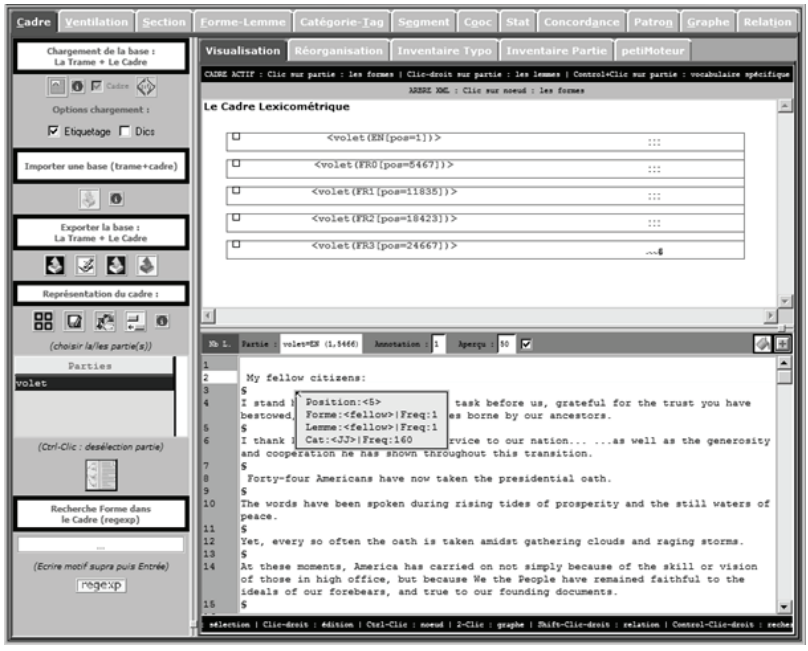
Extrait de la *base textométrique* anglais/français *Investiture Obama*



```
<?xml version="1.0" encoding="utf-8"?>
<baselexicometrique>
<teiheader>
<fileDesc>
<titleStmt>
<title>Modelisation XML de la base textometrique (le metier = le cadre et la trame)
</titleStmt>
<publicationStmt>
<p>Lundi 26 Janvier 2015
 19:03:10
        Ce document n'est pas encore publie.</p>
</publicationStmt>
<sourceDesc>
<title>
<br>Le Trameur 12.34. Lundi 26 Janvier 2015
 19:03:10 </br></title>
<content><l><br>Fichier traite </br><br> corpus-5-volets-iso.txt</br></l>
<l><br>Encodage </br><br> iso-8859-1</br></l>
<l><br>Nombre d'items </br><br> 30925</br></l>
<l><br>Nombre de délimiteurs </br><br> 17179</br></l>
<l><br>Nombre d'occurrences de forme </br><br> 13746</br></l>
<l><br>Nombre de formes </br><br> 2700</br></l>
<l><br>Nombre d'hapax </br><br> 1368</br></l>
<l><br>Fréquence maximale </br><br> 560</br></l>
<l><br>Forme maximale </br><br> de</br></l>
<l><br>Délimiteurs </br><br><![CDATA[.`,;;!?/_-"'()[]{}$$|*><=+
«»→]]></br></l>
<l><br>Etiquetage Treetagger </br><br>OUI</br></l>
<l><br>Langue pour Treetagger </br><br>français</br></l>
</content></sourceDesc>
</fileDesc>
</teiheader>
<trame>
<codage>utf-8</codage>
<delimiteur><![CDATA[.`,;;!?/_-"'()[]{}$$|*><=+
«»→]]></delimiteur>
<items>
<item type="delim" pos="1"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="2"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="3"><f>My</f><c>PP$</c><l>my</l></item>
<item type="delim" pos="4"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="5"><f>fellow</f><c>JJ</c><l>fellow</l></item>
<item type="delim" pos="6"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="7"><f>citizens</f><c>NNS</c><l>citizen</l></item>
/.../
<item type="forme" pos="30913"><f>aux</f><c>PRP_det</c><l>au</l></item>
<item type="delim" pos="30914"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="30915"><f>générations</f><c>NOM</c><l>génération</l></item>
<item type="delim" pos="30916"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="forme" pos="30917"><f>futures</f><c>ADJ</c><l>futur</l></item>
<item type="delim" pos="30918"><f>.</f><c>DELIM</c><l>.</l></item>
<item type="delim" pos="30919"><f> </f><c>DELIM</c><l>BLANK</l></item>
<item type="delim" pos="30920"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="30921"><f>$</f><c>DELIM</c><l>$</l></item>
<item type="delim" pos="30922"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="30923"><f>$</f><c>DELIM</c><l>$</l></item>
<item type="delim" pos="30924"><f>RETURN</f><c>DELIM</c><l>RETURN</l></item>
<item type="delim" pos="30925"><f>$</f><c>DELIM</c><l>$</l></item>
<item type="delim" pos="30926"><f> </f><c>DELIM</c><l>BLANK</l></item>
</items>
</trame>
<cadre>
<acces>
<partition nom="volet">
<p n="EN" d="1" f="5467" nd="1" nf="2"/>
<p n="FR0" d="5467" f="11835" nd="3" nf="4"/>
<p n="FR1" d="11835" f="18423" nd="5" nf="6"/>
<p n="FR2" d="18423" f="24667" nd="7" nf="8"/>
<p n="FR3" d="24667" f="30926" nd="9" nf="10"/>
</partition>
</acces>
</cadre>
</baselexicometrique>
```

Figure 3

Cadre de la *base textométrique Investiture Obama* affiché par *Le Trameur*



Alignements de schémas discursifs avec expressions composées et motifs linguistiques

Dans la version originale du corpus *Investiture Obama* (*EN*), nous constatons une présence caractéristique d'un schéma discursif matérialisé par la répétition du figement morphosyntaxique suivant : *Determiner* + *Noun* + *Preposition* + *Possessive Pronoun* + *Noun* (c'est la plus longue reprise dont la fréquence est supérieure à 10)[24]. On retrouve ce schéma de façon récurrente dans des tournures « chargées » sur le plan expressif, telles que : « *the outcome of our revolution* », « *the father of our nation* », « *the source*

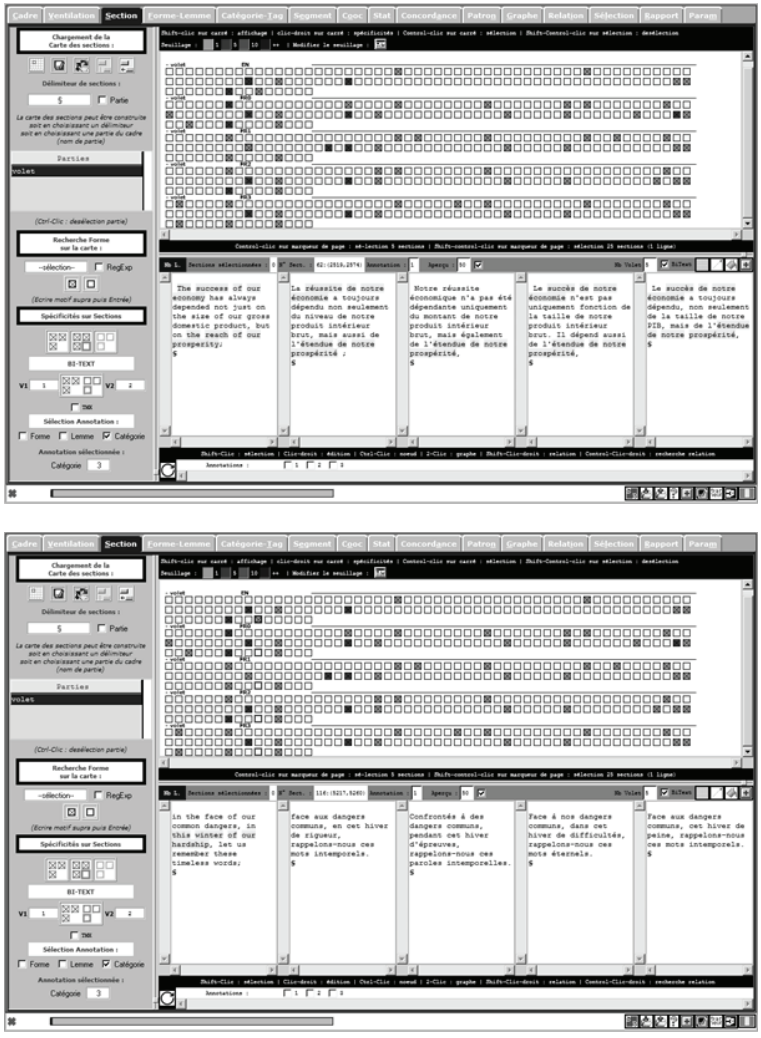
24 Pour découvrir les reprises, le calcul des *segments répétés* a été mené sur la base de l'annotation morphosyntaxique des *items* (en utilisant les jeux d'étiquettes développés pour *treetagger*). La méthode des *segments répétés* utilisée ici est présentée, par exemple, dans Ludovic Lebart et André Salem, *op. cit.*

of our confidence », « *the reach of our prosperity* », « *this winter of our hardship* », « *the justness of our cause* », etc. *La Sélection* par seuillage permet de localiser les phrases du discours original dans lesquelles ce *type* est présent (sur la figure 4, l'intensité de la couleur utilisée sur la *carte des sections* précise si la fréquence est plus ou moins grande). *La Sélection* topographique induite sur les phrases alignées des traductions françaises fait émerger la reprise suivante : *Nom + Préposition + Déterminant possessif + Nom*. C'est le segment le plus *spécifique*[25] de l'échantillon. On trouve ces répétitions dans les constructions, telles que : « issue de notre révolution », « père de notre nation », « réussite de notre économie », « justesse de notre cause », « force de notre exemple », « grandeur de notre nation », « promesse de notre citoyenneté », « source de notre confiance ». Notons que l'absence du *Déterminant* en première place dans le segment français résulte de la variation *Déterminant / Pronom démonstratif* qui caractérise ces traductions.

25 Indice de *spécificité* : 24.1, fréquence totale : 83, fréquence dans la *Sélection* : 40. Sur la méthode des *spécificités*, on peut se référer à l'ouvrage de Ludovic Lebart et André Salem, *ibid*.

Figure 4

Navigation à l'aide de la *carte des sections* parallèles dans la base *Investiture Obama* (5 volets alignés au niveau de la phrase)



Légende : Chaque carré de la *carte des sections* correspond à une phrase. L'alignement vertical des carrés matérialise les alignements. Les ruptures de parallélisme dans le coloriage des sections miroirs signalent les divergences dans la distribution des unités étudiées. La carte permet de naviguer entre les volets textuels afin de comparer le texte original et les trais distinctifs de chaque traduction.

La *carte des sections* parallèles affichée sur la figure 4 montre la distribution des schémas discursifs au fil du texte dans le corpus segmenté en phrases. La présence/absence est indiquée à l'aide de carrés de couleur dont l'intensité varie en fonction des seuils de fréquence (de 1 à 10 et plus). La cartographie des *types* bilingues *Determiner* + *Noun* + *Preposition* + *Possessive Pronoun* + *Noun* en anglais et *Nom* + *Préposition* + *Déterminant possessif* + *Nom* en français permet de repérer une zone d'intersection. Dans cette zone, les distributions parallèles des *types* affichées sur la carte par des carrés de couleur donnent des propriétés graphiques similaires aux sections (phrases) qui se correspondent dans le texte original et les traductions. Ce constat amène des alignements multiples (type source-cible) qui tiennent compte de plusieurs traductions, telles que « *the success of our economy* » mis en correspondance avec « réussite de notre économie » et « succès de notre économie »; « *the justness of our cause* » apparié avec « justesse de notre cause » dans toutes les traductions; ou encore « *the meaning of our liberty* » en correspondance avec « sens de notre liberté » et aussi avec « signification de notre liberté ». Ces exemples illustrent les principes de la construction des axes d'alignements au-delà des « appariements de mots » isolés de la chaîne discursive. Les alignements sont élargis à l'axe paradigmatique et tiennent compte du parallélisme de schémas discursifs.

Principes d'incrémentations successives de la *Trame*

Les alignements de *contenus* sont matérialisés sur la *Trame* par le biais d'une nouvelle annotation que l'on peut ajouter à la *base textométrique* sous *Le Trameur*. Pour chaque *item* du corpus *Investiture Obama*, on peut désormais repérer sa forme, son lemme, son étiquette morphosyntaxique, mais aussi sa catégorie d'alignement, connue lorsque l'*item* a été sélectionné à l'issue du calcul de *spécificités* et de l'alignement dans la *carte des sections parallèles*. Vus sous cet angle, les alignements sous-phrastiques sont des *Sélections* d'ensembles d'*item*s de la *Trame*. Ces objets génériques peuvent être transmis entre procédures de traitement que l'on applique successivement à la même *base textométrique*.

Les traces des traitements apportés à la ressource initiale sont systématiquement conservées sous forme d'incrémentations successives. Elles sont ajoutées au *Cadre* textométrique défini à partir d'une *Trame* commune. Ce système permet d'implémenter les mécanismes d'exploration interactive des corpus multilingues multi-annotés à des fins d'analyse contrastive. Par exemple, on peut réaliser de nouvelles *Sélections* topographiques[26] qui ciblent les divergences dans les distributions des *types* bilingues *Determiner + Noun + Preposition + Possessive Pronoun + Noun* et *Nom + Préposition + Déterminant possessif + Nom*, compte tenu des alignements déjà réalisés dans la zone d'intersection[27]. Ainsi, on parvient à repérer les phrases où l'équivalence des constructions n'est pas observée dans toutes les traductions (« *the success of our economy* » traduit par « notre réussite économique »), ou même absente (« *this winter of our hardship* » traduit différemment dans chaque traduction : « cet hiver de rigueur », « cet hiver d'épreuves », « cet hiver de difficultés » et « cet hiver de peine »), comme le montre la figure 4.

Ce découpage progressif d'une même *Trame* au cours des traitements successifs correspond à l'application du schéma de *ressources textuelles incrémentales*[28]. Pour recourir à ce schéma, il faut préciser les apports de chaque traitement successif et identifier ses résultats sur le couple *Trame/Cadre* d'une même *base textométrique*.

---

26 Maria Zimina, « Equivalences traductionnelles », *Lexicometrica*, nº spécial *Explorations textométriques*, 2013, http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm, site consulté le 10 octobre 2015.

27 Les principes de découverte itérative des appariements en cas de correspondances traductionnelles multiples sont introduits par Maria Zimina (« Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles », dans Gérald Purnelle (dir.), *Les poids des mots*, vol. 2, Louvain-La-Neuve, Presses Universitaires de Louvain, 2004, p. 1195-1202.

28 Keyser Söze-Duval, *op. cit.*

Alignement incrémental de ressources textuelles multilingues richement annotées

Le modèle *Trame/Cadre* mis en œuvre dans *Le Trameur* permet de conduire des explorations sur des ressources textuelles multilingues richement annotées. On peut utiliser ce modèle de stockage de données pour l'analyse textométrique de *treebanks*[29], appelés aussi corpus arborés, par exemple *Rhapsodie*[30], ou encore pour des alignements de treebanks couvrant plusieurs langues.

Treebanks parallèles : *base textométrique ParTUT2Trameur*

Nous illustrons les principes d'exploration textométrique et alignements des *treebanks* multilingues à l'aide des corpus diffusés dans le cadre du projet *ParTUT*[31] (*Multilingual Turin University TreeBank*). Le projet a permis la création et l'alignement phrastique des *treebanks* parallèles en italien, anglais et français à l'aide d'un seul schéma unifié[32]. Les données *ParTUT* utilisent le format CoNLL[33] (*Conference on Computational Natural Language Learning*). Seules les ressources en français et en anglais ont été utilisées dans notre exploration. Elles ont été transcodées dans un format permettant de définir une *base textométrique*

29 Un *treebank* est un corpus enrichi par une analyse syntaxique et/ou sémantique, voire prosodique dans le cas d'un corpus oral.

30 Les principes d'exploration de la *base textométrique Rhapsodie* sont présentés dans Serge Fleury et Maria Zimina, « Trameur : A Framework for Annotated Text Corpora Exploration », dans Lamia Tounsi *et al.* (dir.) *Proceedings of COLING 2014, 25th International Conference on Computational Linguistics : System Demonstrations*, Dublin, 2014, p. 57-61, http://www.aclweb.org/anthology/C14-2013.pdf, site consulté le 10 octobre 2015.

31 Le schéma utilisé pourrait potentiellement être étendu à d'autres langues : ParTUT Project, 2015, http://www.di.unito.it/~tutreeb/partut.html, site consulté le 10 octobre 2015.

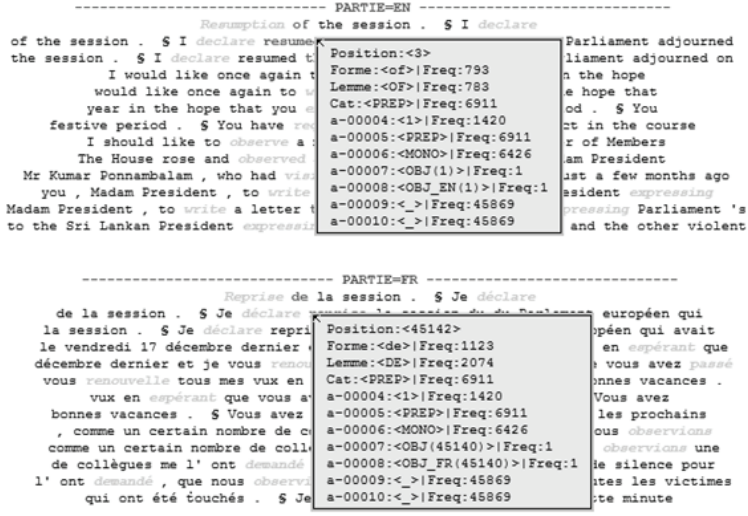
32 Manuela Sanguinetti et Cristina Bosco, « Building the Multilingual TUT Parallel Treebank », dans *Proceedings of the 2nd Workshop on Annotation and Exploitation of Parallel Corpora (AEPC 2)*, 2011, http://www.mt-archive.info/AEPC-2011-Sanguinetti.pdf, site consulté le 10 octobre 2015.

33 Sabine Buchholz and Erwin Marsi, « CoNLL-X Shared Task on Multilingual Dependency Parsing », dans *CoNLL-X'06 Proceedings of the Tenth Conference on Computational Natural Language Learning*, Stroudsburg (PA), 2006, p. 149-164.

intégrant une segmentation annotée. La *Trame* de cette base *ParTUT2Trameur* intègre 10 niveaux d'annotation et met en contraste les volets anglais et français au sein du *Cadre* organisé en deux *contenants* (*EN* et *FR*)[34]. L'alignement des phrases est matérialisé par le délimiteur « § ». Chaque *item* de la *Trame* est décrit par l'ensemble des annotations, comme indiqué sur la figure 5.

Figure 5

Concordances sur les relations de dépendance OBJ (objet) dans la base *ParTUT2Trameur* (anglais/français)



Légende : L'extrait de la première concordance du volet anglais de *ParTUT2Trameur* met au jour l'*item* « *of* » en position 3 et ses 10 annotations. On constate que «of» porte la relation de dépendance OBJ(1) qui traduit sa relation avec l'*item* cible « *Resumption* » en position 1 sur la *Trame*. Le deuxième extrait de concordance affiche la zone miroir en français sur l'*item* « de » en position 45142 et ses 10 annotations. L'*item* « de » porte la relation de dépendance OBJ(45140), traduisant sa relation avec l'*item* « Reprise » en position 45140. La coloration des items source et cible de la relation est produite automatiquement au moment de l'affichage de la concordance dans *Le Trameur*.

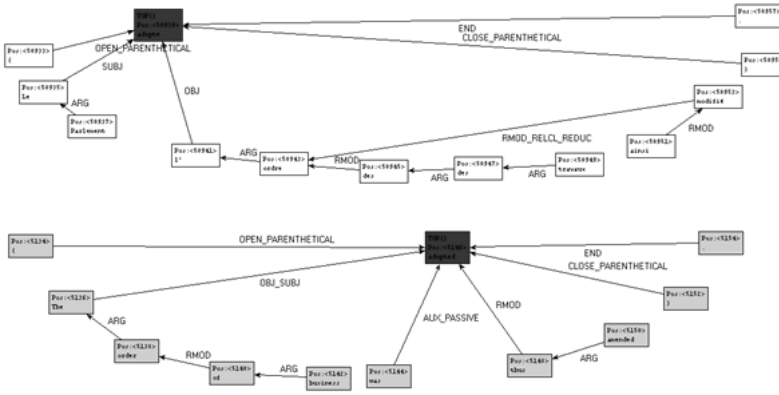
34 Pour l'accès à la base *ParTUT2Trameur* et la documentation sur le processus de construction : Serge Fleury, *Le Trameur*, 2015, *op. cit.*

Dans *ParTUT2Trameur*, chaque *item* peut porter une annotation traduisant une dépendance syntaxique sous la forme : RELATION(CIBLE), où RELATION est une chaîne portant le nom de la relation visée et CIBLE est une valeur numérique pointant vers une position de la *Trame*. La figure 5 donne à voir un extrait de la concordance mettant au jour toutes les relations de dépendance de type OBJ (objet). Seuls les 13 premiers contextes intégrant cette relation sont représentés pour chaque volet.

Sur la figure 6, les graphes de relations générés sous *Le Trameur* affichent une analyse syntaxique complète d'un couple de phrases alignées. Lors de l'exploration du corpus à l'aide de la *carte des sections* parallèles, cette fonction permet de mieux interpréter l'annotation des relations syntaxiques correspondant à chaque alignement phrastique de la *Sélection* étudiée.

Figure 6

Les relations de dépendance dans les phrases alignées affichées à l'aide des graphes sous *Le Trameur*



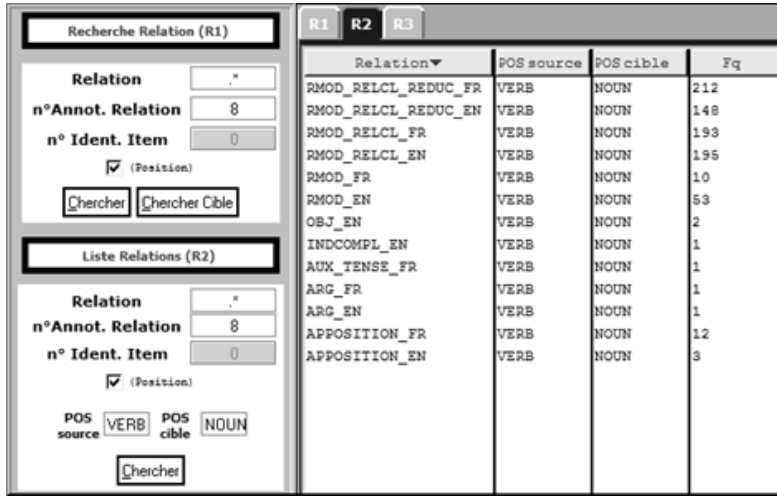
Alignements de relations : croisement d'annotations et résonance textuelle

Dans cette expérimentation, nous montrons que les parallélismes syntaxiques entre les langues peuvent être utilisés pour guider les alignements au niveau des mots et de leurs constructions, en articulation à un calcul statistique[35]. L'exploration parallèle de dépendances syntaxiques peut débuter à l'aide d'un tableau de fréquences croisant les relations (annotation nº 8, avec l'appartenance *EN*/*FR*) et les catégories morphosyntaxiques (annotation nº 3). Sous *Le Trameur*, ce tableau est généré automatiquement, comme le montre la figure 7. On remarque que les fréquences des relations RMOD_RELCL (proposition relative), avec le verbe en position source et le nom en position cible, se ressemblent fortement dans les deux langues (193 en français et 195 en anglais).

35 Notons que l'annotation syntaxique mise en place dans le cadre du projet *ParTUT* suit un schéma unifié et facilite les appariements entre les deux langues (anglais/français). De façon générale, la parenté des langues, le type de traduction (plus ou moins libre) et le schéma d'annotation mis en place dans chaque volet du corpus ont des répercussions sur les alignements.

Figure 7

Fréquences des relations de dépendance entre les verbes (en position source) et les noms (en position cible) dans la base *ParTUT2Trameur*



Légende : On lit sur le tableau ci-dessus que les relations de dépendance RMOD_ RELCL (proposition relative) avec les verbes en position source et les noms en position cible ont des fréquences très proches dans les deux volets (193 en français et 195 en anglais).

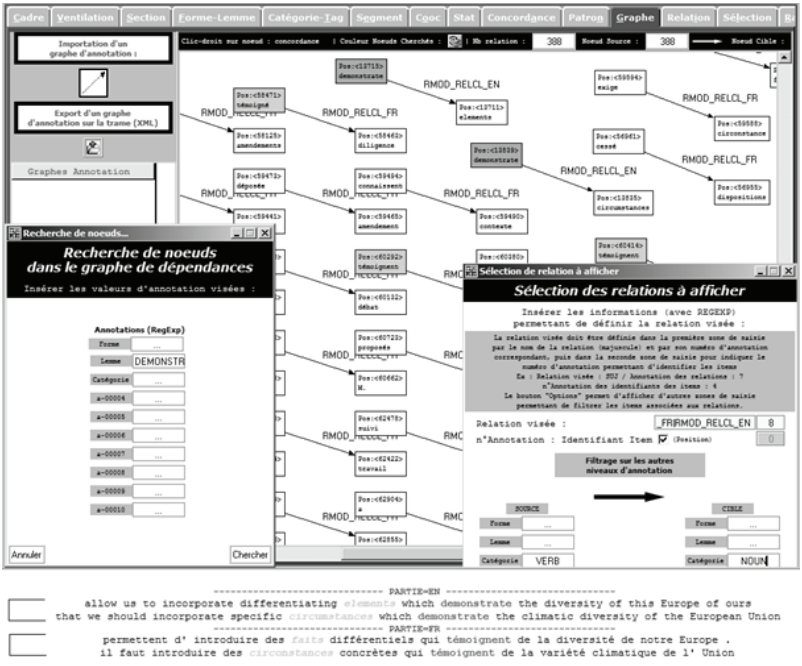
Sur la figure 8, les relations sont visualisées sous forme de graphes, avec croisement d'annotations. Pour aligner les relations qui se correspondent, on mobilise la *résonance textuelle*. Comme nous l'avons déjà montré pour la base *Investiture Obama*, l'exploration se fait à l'aide de la *carte des section*s parallèles. Sous *Le Trameur*, les correspondances sont induites par seuillage à l'aide du calcul des *spécificités*[36], déclenché automatiquement pour des zones miroirs de la carte, voir la figure 9. Dans le cas de la base *ParTUT2Trameur*, les correspondances sont issues des relations que l'on cherche à comparer. Dans notre exemple, après avoir mis en correspondance le lemme (**to**) « *demonstrate* » en anglais et le lemme « témoigner » en français, on opère une *Sélection* des

36 Maria Zimina « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles », *op. cit.*

nœuds correspondants dans les relations RMOD_RELCL, comme cela est indiqué sur les figures 8 et 9. La présence conjointe des relations dans les sections miroirs (phrases) permet d'incrémenter sur la *Trame* les alignements suivants : « *elements* … *demonstrate* » et « faits … témoignent », « *circumstances* … *demonstrate* » et « circonstances … témoignent ». Ces alignements sont issus des propositions relatives.

Figure 8

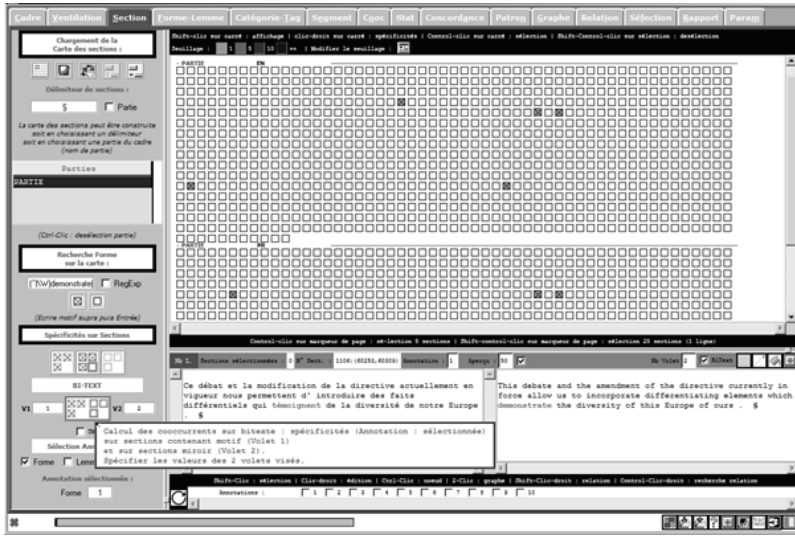
Alignements de relations de dépendance dans *ParTUT2Trameur*



Légende : L'exploration débute par la recherche des relations RMOD_RELCL (proposition relative), avec le verbe en position source et le nom en position cible dans les deux volets. La *résonance textuelle* (représentée sur la figure 8) met au jour la correspondance (*to*) « *demonstrate* » en anglais et « témoigner » en français. Ces nœuds sont signalés à l'aide des couleurs. Les alignements sont effectifs si les relations sélectionnées font partie des sections miroirs (phrases alignées). Ils sont matérialisés sur la figure 9 par des extraits de concordances

Figure 9

Analyse de la *résonance textuelle* à l'aide de la *carte des sections* parallèles dans *ParTUT2Trameur*



Légende : La carte des sections parallèles affiche la distribution du lemme anglais (to) « *demonstrate* » et les sections miroirs en français. Comme l'indique la position du curseur sur la capture d'écran, le calcul des *spécificités* est déclenché pour les zones miroirs afin d'induire des correspondances par seuillage.

Alignements de relations dans les réseaux de cooccurrences

Au lieu de partir d'un ensemble de relations et d'examiner de plus près les réalisations d'une relation pour un mot, comme nous l'avons effectué dans l'expérimentation précédente, l'exploration peut débuter à partir d'un mot et faire apparaître des relations syntaxiques qu'il privilégie. Notons que cette approche permettrait davantage de procéder à des alignements de relations non identiques d'une langue à l'autre.

Comme *Le Trameur* permet d'articuler les différents niveaux d'annotation disponibles dans la base traitée pendant les calculs statistiques, les requêtes sur les relations de dépendance peuvent être combinées avec des méthodes traditionnelles de la statique textuelle, telles que le calcul des cooccurrences. Dans l'exploration

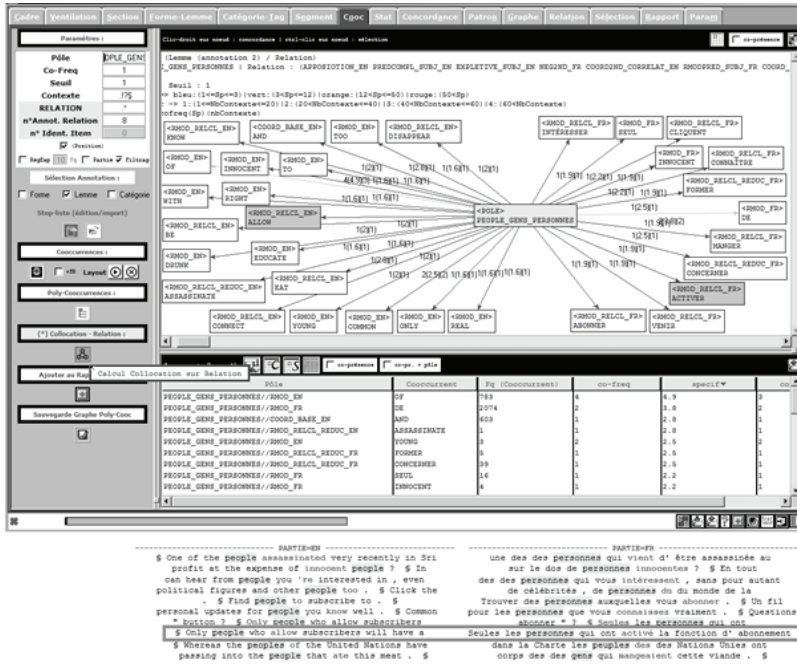
qui suit, nous utilisons conjointement les deux approches afin d'explorer les branches de réseaux de cooccurrences calculées sur les relations de dépendance.

Dans *ParTUT2Trameur*, l'équivalence de « *people* » en anglais et « personnes » ou « gens » en français ressort par *résonance textuelle*. Ce constat oriente l'exploration vers l'analyse des voisinages contextuels correspondants. Les occurrences de « *people* », « personnes » et « gens » sont d'abord regroupées au sein d'un nouveau *type* (59 occurrences au total). Ce *type* devient le pôle du calcul de cooccurrences sur relations dans les phrases *EN*/*FR*.

Sur la figure 10, le réseau de cooccurrences reflète les particularités à la fois lexicales et syntaxiques des voisinages étudiés dans les deux langues. On remarque les mêmes catégories de relations en anglais et en français. Chaque branche du réseau est ensuite localisée sur la *Trame*. On parvient à aligner les cooccurrents anglais et français compte tenu des similitudes de leurs relations de dépendance dans les phrases qui se correspondent.

Figure 10

Alignements de cooccurrences sur relations dans *ParTUT2Trameur*



Légende : L'exploration débute par la recherche des cooccurrences sur relations de dépendance à partir du pôle bilingue composé de « *peuple* » en anglais et « gens » ou «personnes» en français. Pour chaque cooccurrent, on déchiffre la relation de dépendance syntaxique au sein du réseau construit. Les cooccurrents bilingues portés par les mêmes relations de dépendance sont appariés dans les phrases qui se correspondent. Ces alignements sont matérialisés par le code couleur dans les concordances.

Les extraits de contextes affichés sur la figure 10 permettent de vérifier que les cooccurrents alignés font bien partie des équivalences de traduction. On y trouve, par exemple, « *Only* ***people*** *who* ***allow*** *subscribers…* » traduit par « Seules les **personnes** qui ont **activé** la fonction d'abonnement… ». Notons qu'il s'agit des alignements singuliers, qui reflètent les particularités contextuelles.

À tout moment, les alignements peuvent être incrémentés sur la *Trame* ou défaits, compte tenu de la progression de l'exploration et de l'intuition linguistique de l'utilisateur qui garde la main sur tout le processus d'interaction avec la base.

Conclusion et perspectives

Nous avons présenté une approche quantitative des corpus multilingues annotés qui permet d'étudier des correspondances linguistiques d'une langue à l'autre. Les méthodes convoquées dans le cadre de cette approche aident à réaliser des alignements textuels à plusieurs niveaux d'analyse linguistique en faisant émerger des ensembles de traits caractéristiques (structures lexico-grammaticales, syntaxiques, etc.). Elles rendent possible l'exploration interactive des corpus multilingues annotés avec une prise en compte systématique des emplois parallèles à plusieurs niveaux d'analyse linguistique.

Héritées de la *textométrie*, ces méthodes articulées les unes aux autres permettent *in fine* de conduire des parcours interprétatifs sur des données complexes. Elles sont implémentées dans un logiciel appelé *Le Trameur*[37], qui mobilise un modèle de stockage de données particulier *Trame/Cadre*. Ce modèle décrit à la fois les résultats de la segmentation de la chaîne textuelle en unités de décomptes et un système d'empans textuels (découpages) que l'on peut soumettre aux comparaisons statistiques[38].

Actuellement, *Le Trameur* permet de traiter des ressources textuelles richement annotées (*treebanks*) qui intègrent, notamment, des relations de dépendance entre les unités représentées. En particulier, il offre la possibilité d'éditer et de visualiser les ressources annotées disponibles via une interface graphique. Étant au départ un logiciel de textométrie, *Le Trameur* intègre également des processus de calcul qui combinent les approches statistiques et les requêtes sur relations en croisant différents niveaux d'annotation présents dans les données traitées. Son autre spécificité est liée au fait qu'il donne à voir des alignements

37 Serge Fleury, *Le Trameur*, 2015, *op. cit.*
38 Keyser Söze-Duval, *op. cit.*

de *treebanks* en contrastant les vues sur les ressources textuelles dans des langues différentes.

La réflexion mise en place dans le cadre de ce développement logiciel vise notamment à mettre en œuvre des processus de calcul articulant les différents niveaux de représentation dans la *base textométrique* traitée. Ces étapes de développement passent inévitablement par des collaborations entre spécialistes des domaines concernés, par exemple, prosodiste et syntacticien pour le corpus oral *Rhapsodie*[39], syntacticien et traductologue pour le *treebank* multilingue *ParTUT*[40]. L'approche collaborative permet de s'assurer de la fiabilité des représentations de données expérimentales, de la pertinence de l'outillage méthodologique mobilisé dans le développement logiciel, de la validité de l'interprétation des résultats, etc.

L'utilisation de l'architecture de stockage de données *Trame/Cadre* pour les ressources multilingues richement annotées offre une grande souplesse dans la construction des parcours interprétatifs en fonction des objectifs d'analyse particuliers. Les séquences de traitements apportés à la ressource textuelle initiale sont organisées sous forme d'incrémentations successives, ajoutées au *Cadre* textométrique défini à partir d'une *Trame* commune. Particulièrement flexibles pour l'indexation d'objets textuels complexes, les ressources respectant l'architecture *Trame/Cadre* peuvent faciliter plusieurs pratiques linguistiques dans l'espace textuel multilingue informatisé (traduction, échanges de procédés, mises à jour, homogénéisation terminologique, etc.). Ces nouvelles perspectives amènent à développer le concept de bi-texte numérisé vers des représentations plus dynamiques que les Mémoires de Traduction traditionnelles.

39 ANR 07 CORP 030 01 *Rhapsodie, Corpus prosodique de référence en français parlé*, *op. cit.*

40 ParTUT Project, *op. cit.*

Bibliographie

ANNIS (ANNotation of Information Structure), 2015, http://annis-tools.org/, site consulté le 10 octobre 2015.

ANR 07 CORP 030 01, *Rhapsodie, Corpus prosodique de référence en français parlé*, 2015, http://www.projet-rhapsodie.fr/, site consulté le 10 octobre 2015.

ANR 2012 CORD 015, *TransRead, Lecture et interaction bilingues enrichies par les données d'alignement*, 2015, http://transread.limsi.fr/, site consulté le 10 octobre 2015.

BAV (Biblioteca Apostolica Vaticana), 2015, https://www.vatlib.it/, site consulté le 10 octobre 2015.

Buchholz Sabine and Erwin Marsi, « CoNLL-X Shared Task on Multilingual Dependency Parsing », dans *CoNLL-X'06 Proceedings of the Tenth Conference on Computational Natural Language Learning*, Stroudsburg (PA), 2006, p. 149-164.

Fleury, Serge, *Base textométrique de textes alignés*, 2014, http://www.tal.univ-paris3.fr/trameur/MAJ-12.02.pdf, site consulté le 10 octobre 2015.

Fleury, Serge, « Exploration du corpus Traductions alignées du discours d'investiture de B. Obama (Tutoriel nº 3, Explorations textométriques avec mkAlign) », *Lexicometrica*, nº spécial *Explorations textométriques*, 2009, http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm, site consulté le 10 octobre 2015.

Fleury, Serge, *Le Trameur*, 2015, http://www.tal.univ-paris3.fr/trameur/, site consulté le 10 octobre 2015.

Fleury, Serge, *Le Trameur. Propositions de description et d'implémentation des objets textométriques*, 2013, http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf, site consulté le 10 octobre 2015.

Fleury, Serge et Maria Zimina, « Exploring translation corpora with MkAlign », dans Gabe Bokor (dir.), *Translation Journal*, 2007, http://translationjournal.net/journal/39mk.htm, site consulté le 10 octobre 2015.

Fleury, Serge et Maria Zimina, « Trameur : A Framework for Annotated Text Corpora Exploration », dans Lamia Tounsi *et al.* (dir.) *Proceedings of COLING 2014, 25th International Conference on Computational Linguistics : System Demonstrations*, Dublin, 2014, p. 57-61, http://www.aclweb.org/anthology/C14-2013.pdf, consulté le 10 octobre 2015.

GATE (General Architecture for Text Engineering), 2015, http://gate.ac.uk/gate, site consulté le 10 octobre 2015.

Ghorbel, Hatem et Giovanni Coray, « L'alignement multicritères des documents médiévaux », *Lexicometrica*, nº special *Corpus aligné*, 2002, http://lexicometrica.univ-paris3.fr/thema/thema6.htm, site consulté le 10 octobre 2015.

Harris, Brian, « Bi-text : A New Concept in Translation Theory », *Language Monthly*, nº 54, 1988, p. 8-10.

Lebart, Ludovic et André Salem, *Statistique textuelle*, Paris, Dunod, 1994.

MACAON, chaîne de traitement, 2015, http://macaon.lif.univ-mrs.fr/, site consulté le 10 octobre 2015.

ParTUT Project, 2015, http://www.di.unito.it/~tutreeb/partut.html, site consulté le 10 octobre 2015.

PDT (*The Prague Dependency Treebank* 2.0), 2015), ufal.mff.cuni.cz/pdt2.0, site consulté le 10 octobre 2015.

Pillias, Clément, « Reading Bilingual Texts with Digital Tools : A State of the Art », *Projet ANR 2012 CORD 015 TransRead, Lecture et interaction bilingues enrichies par les données d'alignement*, Deliverable 2.1, 2014, http://transread.limsi.fr/Deliverable2.1.pdf, site consulté le 10 octobre 2015.

Sanguinetti, Manuela et Cristina Bosco, « Building the Multilingual TUT Parallel Treebank », dans *Proceedings of the 2nd Workshop on Annotation and Exploitation of Parallel Corpora (AEPC 2)*, 2011, http://www.mt-archive.info/AEPC-2011-Sanguinetti.pdf, site consulté le 10 octobre 2015.

Schmid, Helmut, « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of International Conference « New Methods in Language Processing »*, Manchester, UMIST, 1994, p. 44-49.

Somers, Harold, *Computers and Translation : A translator's guide*, Amsterdam, John Benjamins B.V., 2003.

Söze-Duval, Keyser, *Pour une textométrie opérationnelle,* 2008, http://www.tal.univ-paris3.fr/trameur/RTI6provisoire.doc, site consulté le 10 octobre 2015.

Tiedemann, Jörg, *Bitext Alignment, Synthesis Lectures on Human Language Technologies*, San Rafael, Morgan & Claypool Publishers, 2011.

Zimina, Maria, « Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles », thèse pour le Doctorat en Sciences du langage, Paris, Université de la Sorbonne nouvelle – Paris 3, 2004.

Zimina, Maria, « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles », dans Gérald Purnelle (dir.), *Les poids des mots*, vol. 2, Louvain-La-Neuve, Presses Universitaires de Louvain, 2004, p. 1195-1202.

Zimina, Maria, « Equivalences traductionnelles ». *Lexicometrica*, nº spécial *Explorations textométriques*, 2013, http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm, site consulté le 10 octobre 2015.