Exploration textométrique de corpus de traduction

MARIA ZIMINA
Université de la Sorbonne nouvelle - Paris 3, Paris, France
zimina@msh-paris.fr

RÉSUMÉ
L'article présente les résultats d'une série de recherches consacrées au développement d'une nouvelle famille d'outils d'exploration textométrique intertextuelle. L'utilisation de ces outils dans le contexte multilingue est illustrée par des échantillons de ressources traductionnelles obtenues à partir du corpus parallèle français/anglais de la *Convention de sauvegarde des Droits de l'Homme*. Les perspectives ouvertes par cette approche offrent aux traducteurs, enseignants en langues étrangères, terminologues, lexicographes, etc., des moyens automatisés pour explorer la structure des équivalences lexicales dans les corpus de traduction.

ABSTRACT
The article presents the results of a series of experiments devoted to the development of new tools for intertextual textometric exploration of translation corpora. The use of these tools in a multilingual context is illustrated by sample translation resources obtained on quantitative bases from the parallel French/English corpus of the *Convention for the Protection of Human Rights*. The suggested approach opens up new horizons for automatic exploration of lexical equivalences of translation corpora by a variety of users: translators, foreign language teachers, terminologists, lexicographers, etc.

MOTS-CLÉS / KEYWORDS
alignement, bi-texte, corpus parallèles, correspondances traductionnelles, statistique textuelle, textométrie

1. Les corpus parallèles

La notion de corpus parallèle, qui émerge actuellement dans les travaux de différents chercheurs comme : corpus comportant plusieurs volets qui correspondent chacun à une version d'un même texte dans deux ou plusieurs langues différentes, renvoie à des situations connues de coexistence de textes présentant des liens forts dans leur structuration. Paléographes, historiens, théologiens, juristes, philologues, linguistes, traductologues manipulent depuis fort longtemps des corpus de textes rassemblant plusieurs volets dont chacun est constitué par une version du même texte dans une langue différente. Parmi les documents parallèles les plus connus, on citera, par exemple, les textes inscrits sur *la pierre de Rosette* ainsi que les différentes éditions de la Bible.

On parle de parallélisme non seulement entre les textes en relation de correspondance traductionnelle, mais aussi entre réécritures et réinterprétations de textes monolingues ou multilingues susceptibles de permettre une comparaison.

Dans le contexte récent de l'informatisation des études des corpus de textes, la notion de parallélisme textuel reçoit des formalisations plus strictes qui permettent de manipuler conjointement les différents volets d'un corpus pluritextuel. L'objectif poursuivi est avant tout d'utiliser les données textuelles parallèles et les structures des documents alignés pour extraire, à partir des corpus, des ressources traductionnelles utilisables dans d'autres contextes.

2. Le contexte multilingue

Dans le contexte multilingue, les corpus parallèles sont généralement composés de textes sources et de leurs traductions existantes ou des textes dont chacun est une traduction de l'autre sans qu'il soit possible de déterminer lequel a servi de source. Actuellement, le terme « corpus comparables » est utilisé pour se référer à des corpus composés de textes traitant des mêmes thèmes dans plusieurs langues sans être des traductions.

Avec la croissance du marché de la traduction, les agents économiques et les organisations internationales s'intéressent de plus en plus à l'archivage électronique conjoint de textes et de leurs traductions dans différentes langues. Ces documents représentent le noyau de la communication multilingue et rendent possible l'échange d'information entre communautés. L'information qu'ils contiennent revêt une importance capitale dans plusieurs domaines socio-économiques.

De vastes corpus de textes sont systématiquement archivés dans les banques textuelles et bases de données informatiques. Le Web fournit une source de plus en plus riche de documents parallèles multilingues. Ces banques textuelles sont souvent consultées par les spécialistes pour récupérer des références terminologiques ou pour comparer plusieurs versions d'un même document. Le problème est alors de disposer d'un accès rapide, structuré et efficace à l'information contenue dans ces corpus. L'archivage électronique des données textuelles ainsi que la création de systèmes de recherche documentaire (*information retrieval*) fournissent des solutions partielles à ce problème. Néanmoins, pour rendre facilement consultables les ressources présentes dans ces documents, il est nécessaire d'établir un système de mise en relation entre les segments correspondants dans des couples de textes (lexies, locutions, syntagmes, phrases, etc.).

3. La textométrie multilingue

Dans ce qui suit, on utilisera le terme « textométrie » pour se référer à l'ensemble des méthodes quantitatives permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes (Lebart et Salem 1994 : 314).

La spécificité de la démarche textométrique réside dans le statut privilégié conféré aux données textuelles qui ne prennent que très peu d'appui sur des savoirs *a priori* (linguistiques, pragmatiques, etc.).

L'analyse textométrique s'appuie sur des comptages réalisés à partir du repérage des occurrences d'unités lexicales dans les différentes parties d'un corpus. Les textes sont d'abord segmentés en occurrences de formes graphiques (chaînes de caractères bornées par deux caractères délimiteurs, en anglais - *token*). Une fois cette segmentation réalisée, on peut définir des règles d'identification entre des unités de segmentation ainsi obtenues. La formalisation de ces règles permet de produire des décomptes portant sur les occurrences d'un même type aux différents endroits d'un texte. (Salem 1987), (Lebart et Salem 1994).

Le concept de type généralisé *TGen* permet de décrire des ensembles d'occurrences sélectionnés systématiquement dans le texte (Lamalle et Salem 2002). Ainsi, on peut recenser, au-delà des occurrences des formes graphiques, les occurrences d'un segment répété [Ex. : démocratie apte à se défendre], d'un quasi-segment répété [Ex. : (ang.) : *there had been a breach of the article* et *there had been a violation of the article*], la rencontre de deux formes (cooccurrences) à l'intérieur d'une fenêtre de x-formes graphiques ou d'une phrase [Ex. : démocratie + république], d'un type constitué par les occurrences d'un ensemble de formes graphiques défini en raison de leur parenté sémantique dans le corpus [Ex. : démocratique, démocratie, démocratiques, démocrate], etc.

Dans un contexte multilingue, la textométrie offre des perspectives de recherches prometteuses pour de multiples dimensions d'analyse de corpus dans des langues différentes (l'alignement automatique, l'extraction de ressources traductionnelles, l'exploration intertextuelle, la synthèse de l'information bi-textuelle, etc.).

Pour présenter les développements récents dans ce domaine, nous emprunterons des exemples à un corpus de textes juridiques français/anglais de la *Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales*, désormais CONVENTION. Ce corpus a été

constitué à partir du texte officiel de cette Convention, de ses protocoles intégraux, et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995.

4. L'exploration textométrique du corpus bilingue CONVENTION

L'utilisation de la textométrie pour l'analyse des corpus parallèles implique une tentative de mettre au point des méthodes particulières d'étude de données textuelles multilingues portant sur les distributions conjointes des vocabulaires dans les textes. Notre objectif consiste maintenant à vérifier si les règles formelles du dépouillement automatique peuvent aider à rapprocher des lexèmes bilingues en correspondance de traduction.

4.1. Rapports de correspondances lexicales multilingues

L'analyse des dictionnaires des formes graphiques constitués à partir des deux volets du corpus CONVENTION donne un premier aperçu des caractéristiques lexicométriques. On constate que pour une partie des formes bilingues en correspondance de traduction, la confrontation des dictionnaires classés par ordre lexicométrique[1] fait apparaître des similitudes entre les rangs lexicaux. L'analyse sémantique de ces équivalences lexicales en corpus permet de comprendre les causes de ces correspondances quasi-univoques.

Tableau 1 : Corpus CONVENTION : rangs lexicaux des formes équivalentes

français	rang lexical	fréquence totale	anglais	rang lexical	fréquence totale
convention	31	1223	*convention*	26	1228
commission	38	889	*commission*	35	832
paragraphe	43	740	*paragraph*	40	772
gouvernement	44	721	*government*	43	740
décision	57	08	*decision*	54	557
droits	69	418	*rights*	65	444
mais	78	380	*but*	79	392
article	80	362	*article*	85	351
citation	83	352	*citation*	86	351
détention	96	314	*detention*	95	336

Guide de lecture du tableau 1 : Pour une partie des formes en correspondance, les rangs lexicaux sont très proches. Le rang est attribué en fonction de la place occupée par la forme dans le dictionnaire des formes trié par ordre lexicométrique.

4.1.1. Les correspondances quasi-univoques

Comme le montre le tableau 1, certaines formes en correspondance de traduction ont des fréquences totales très proches. Dans le corpus, seulement une partie des équivalences de traduction présentent de telles similitudes. Les retours au contexte montrent que ce type d'équivalences concerne notamment des unités lexicales dont les champs sémantiques sont particulièrement proches dans les deux volets du corpus. On peut imputer ce phénomène aux facteurs suivants :

a) Les deux mots (ou syntagmes) renvoient à un nom propre ou à un nombre.

Exemple :

corpus français	fréquence totale	corpus anglais	fréquence totale
Petzold	29	*Petzold*	29
Ii	77	*ii*	77
1992	11	*1992*	10

b) L'équivalence est imposée par le respect d'usages terminologiques.

Exemple :

corpus français	fréquence totale	corpus anglais	fréquence totale
les gouvernements signataires	3	*the governments signatory hereto*	3
privilèges et immunités	8	*privileges and immunities*	8
hautes parties contractantes	42	*high contracting parties*	42

c) Le contexte réduit l'espace des sens possibles pour les deux mots.

Exemple :

corpus français	fréquence totale	corpus anglais	fréquence totale
article	362	*article*	351

Le mot *article* est polysémique en anglais comme en français. Cependant, dans le corpus CONVENTION ce terme est employé dans les deux langues pour désigner la structure de découpage des documents juridiques.

4.1.2. Les correspondances multiples

Lorsqu'il s'agit de mots dotés d'un large éventail de sens dans le corpus, les correspondances lexicales entre les deux volets forment un réseau complexe et la comparaison des fréquences globales des formes graphiques ne constitue pas toujours une bonne indication pour l'appariement. Par exemple, la forme anglaise *case* (F=1009) est traduite par « affaire » (F=396), « cause » (F=276), « espèce » (F=271), « procès » (F=116), etc. De même pour la forme *applicant* (F=1244) qui reçoit plusieurs traductions : « requérant » (F=643), « requérante » (F=323), « intéressée » (F=190) et « intéressé » (F=73).

4.1.3. Les équivalences contextuelles

Une partie des équivalences traductionnelles du corpus relèvent de la traduction contextuelle. Ces équivalences sont peu autonomes sur le plan lexical. Dans le corpus CONVENTION, ce phénomène peut être illustré par des équivalences singulières recensées autour du mot français « monde » (F=9). L'équivalence « principale » monde – *world* (F=5) est dominante : elle a été utilisée près d'une fois sur deux dans la traduction. Les autres équivalences, plus rares, concernent notamment des expressions et des syntagmes complexes où le mot « monde » et son voisinage lexical immédiat sont liés au sein de la même unité de traduction, par exemple : l'adhésion à une conception philosophique du monde ~ *adherence to an ideology* ; les hauts responsables du monde politique et

judiciaire ~ *senior politicians and judges*. Il serait impossible d'isoler un élément correspondant au mot « monde » dans ces équivalences.

L'étude des équivalences contextuelles en corpus permet d'analyser les stratégies descriptives employées par les traducteurs pour produire des effets similaires chez les lecteurs de chacune des langues concernées. Cependant, sur le plan sémantique, ce type d'équivalences ne fournit pas de correspondances lexicales « stables », susceptibles d'être reproduites aisément dans d'autres contextes.

4.2. Alignement lexical et résonance textuelle

Lors de l'étude simultanée des traductions d'un même texte, il est utile de considérer les variations conjointes de différentes unités textuelles dans les deux volets du corpus. Le terme de « résonance textuelle » (Salem 2004) permet de décrire les perspectives de recherche d'un nouveau courant de la textométrie lié à l'analyse de textes dont chacun entretient avec l'autre des rapports étroits.
Le schéma de la résonance textuelle permet de décrire des rapports de correspondance entre des ensembles textuels de nature différente : traductions mutuelles, tours de parole polémiques, données d'acquisition, etc.

Dans le cas de corpus de traduction, la notion de résonance textuelle est fondée sur l'alignement préalable des fragments en correspondance (phrases, paragraphes, sections etc.).
A partir de cette mise en correspondance, toute sélection d'un sous-ensemble d'unités dans un des volets du corpus induit une sélection topographique correspondante dans l'autre volet.

On peut envisager plusieurs types de sélection des unités textuelles pour amorcer le processus de résonance. Par exemple, une « sélection par seuillage » (Salem 2004), (Zimina 2004a) permet de décrire la sélection bi-textuelle opérée en deux étapes :

> 1) On repère, dans l'un des volets du corpus bilingue, des paragraphes (ou des phrases) dans lesquels la fréquence locale (f_{locale}) d'une forme, d'un segment répété ou de toute autre unité textuelle dépasse un seuil fixé ;
>
> 2) Le fragment sélectionné dans le premier volet provoque la sélection par résonance des paragraphes (ou des phrases) correspondants dans l'autre volet du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence, dans l'immense majorité des cas, des expressions qui sont liées sur le plan de la traduction à la première expression. Le recensement automatique d'ensembles d'unités caractéristiques des fragments bi-textuels sélectionnés s'appuie sur le calcul des « spécificités » (Lafon 1984).[2]

Une « sélection topographique » se réalise à partir de considérations portant sur la localisation des unités bilingues dans les fragments textuels appariés (phrases, paragraphes, sections).[3]
L'ensemble de ces fragments est alors présenté sous forme d'une « carte des sections parallèles ».
La carte des sections permet une visualisation du corpus découpé en sections par la promotion d'un (ou de plusieurs) caractères particuliers au statut de délimiteurs de section (Lamalle *et al.* 2003).

Dans le cas des corpus de traduction, le découpage en sections peut être effectué parallèlement, en s'appuyant sur des codes attribués aux phrases (ou paragraphes) en correspondance. La cartographie de la présence/absence des unités textuelle bilingues dans les sections appariées permet le repérage des liens de correspondances entre elles.

Appuyé sur les outils de navigation textométrique du logiciel *Lexico3*, le schéma de résonance textuelle nous a permis de mettre au point une série de pratiques d'exploration de corpus de traduction. Ces pratiques peuvent aider l'utilisateur des données textuelles multilingues (traducteur, lexicographe, terminologue etc.) dans l'extraction de ressources traductionnelles à partir de corpus de textes dans des langues différentes. Dans ce qui suit, nous allons présenter des applications du concept de résonance textuelle à l'exploration d'équivalences lexicales du corpus parallèle CONVENTION aligné au niveau de la phrase.

4.3. La mise en évidence des correspondances lexicales multiples

4.3.1. Outils de navigation textométrique de *Lexico3*

Les fonctionnalités développées au sein du logiciel *Lexico3* (Lamalle *et al.* 2003) permettent à l'utilisateur de visualiser une carte des sections, puis de constituer une sélection arbitraire de sections dont on étudiera ensuite le vocabulaire spécifique.

L'utilisateur dispose d'un ensemble d'outils permettant de choisir (à partir du dictionnaire, du « garde-mots »[4], de la liste des segments répétés, etc.) un type d'unité sur lequel portera son exploration. Après avoir sélectionné le type, il est possible de le faire glisser sur la carte (glisser/déposer). La ventilation du type étudié devient alors visible. Les sections dans lesquelles il est présent apparaissent en couleur. Ce processus peut être réitéré.

4.3.2. Exemples d'extraction de ressources traductionnelles

Exemple 1 – cour, tribunal / *court*

Objectif : découvrir les principales traductions du mot anglais *court* dans le volet français du corpus CONVENTION.

Pour cette exploration, nous allons faire appel au schéma de résonance textuelle décrit dans la section 4.2. Sur la figure 2, les volets français et anglais du corpus CONVENTION sont représentés sur une seule carte de sections bi-textuelle. Chaque carré représente un couple de phrases appariées. La technique de repérage des équivalences par seuillage (Zimina 2004b : 118-126) appliquée aux sections du volet français équivalentes à celles qui contiennent le mot anglais *court* permet de signaler à l'utilisateur le mot français *cour* comme la principale traduction de ce terme dans le corpus. La localisation de l'équivalence cour/*court* dans les zones correspondantes du bi-texte montre qu'il existe d'autres traductions du mot *court* dans le corpus.

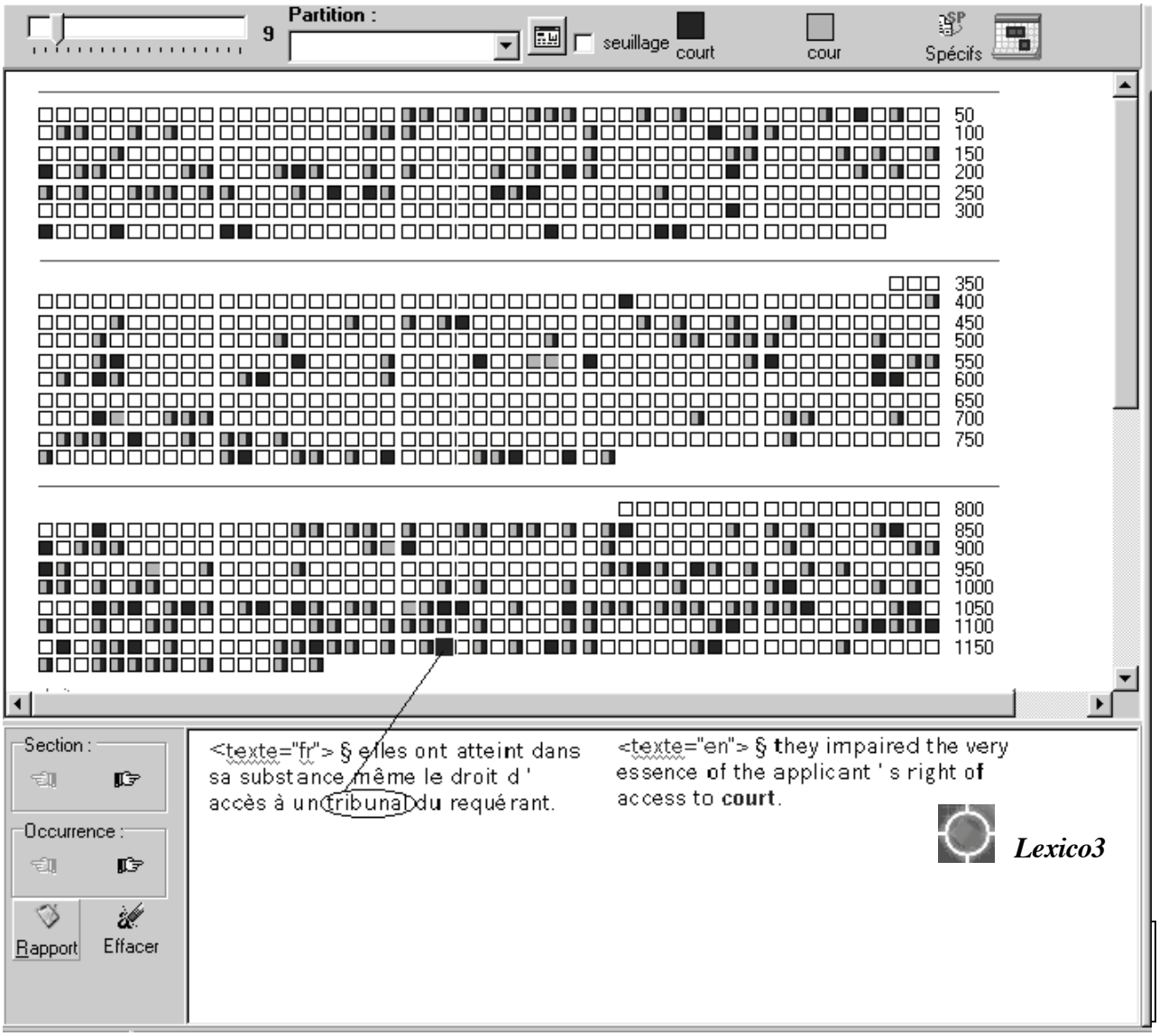
Les fonctionnalités de *Lexico3* rendent possible une visualisation simultanée de la présence des formes cour/*court* dans les sections bi-textuelles. Sur la figure 2, les sections bicolores indiquent les sections où *court* est traduit par « cour ». La présence de sections monochromes sur la carte montre qu'il existe d'autres traductions du mot *court* dans le corpus. En cliquant sur un carré monochrome, il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où *court* n'est pas traduit par « cour ». L'itération du calcul des « spécificités » dans les sections monochromes permet de repérer le terme « tribunal » qui est la deuxième correspondance de *court* dans le corpus.

Exemple 2 – fonctionnaires / *servants, officials, officers*, etc.

Objectif : découvrir l'ensemble de traductions du mot français « fonctionnaires » dans le volet anglais du corpus CONVENTION.

L'exploration commence par le repérage dans le volet français du corpus des phrases dans lesquelles sont présentes les occurrences de la forme « fonctionnaires » (F=49). Le fragment sélectionné dans le volet français provoque la sélection par résonance des phrases correspondantes dans l'autre volet du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence la forme anglaise *servants* (F=50, f_{locale}=31) et le segment répété *civil servants* (F=46, f_{locale}=29) qui sont parmi les traductions du terme français « fonctionnaires ».

Figure 1 : Le repérage topographique des traductions du terme anglais *court*



La fréquence globale de la forme « fonctionnaires » (F=49) est supérieure à celle de la forme anglaise *servants* dans le fragment (f_{locale}=31). Nous pouvons en conclure que la forme « fonctionnaires » reçoit d'autres traductions dans le corpus. Pour découvrir l'ensemble des équivalences lexicales correspondant à la forme-pôle, on soumet au calcul des spécificités les seules phrases du fragment anglais dans lesquelles la forme *servants* est absente. La réitération du calcul des spécificités dans ce sous-ensemble de phrases met en évidence une série d'unités les plus

caractéristiques de ce nouveau fragment : *officers* (F=38, f_{locale}=10), *senior police officers* (F=3, f_{locale}=3), *officials* (F=16, f_{locale}=7).

Le retour au contexte confirme que ces unités constituent bien des traductions de la forme fonctionnaires (au même titre que la forme *servants* et le segment *civil servants* découverts précédemment).

Durant cette exploration, le schéma de résonance textuelle nous a permis de découvrir les principales traductions de la forme-pôle « fonctionnaires » (F=49) : *servants* (f_{locale}=31), *officers* (f_{locale}=10) et *officials* (f_{locale}=7). Un léger écart entre la fréquence totale de cette forme-pôle et le cumul des fréquences locales de ces correspondances en anglais montre qu'il existe au moins un contexte pour lequel la traduction n'a pas été identifiée par notre exploration. Nous pouvons affiner nos constats à travers un retour au texte. Il suffit d'écarter toutes les phrases du fragment anglais où la forme « fonctionnaires » est traduite par *officers*, *officials* ou *servants*. Pour ce faire, on procède à la sélection des phrases du fragment anglais dans lesquelles ces trois formes sont absentes ou contenues en nombre inférieur au total d'occurrences de la forme *fonctionnaires* dans la phrase correspondante en français. Cette recherche aboutit à la localisation sur la carte du corpus du couple de phrases suivantes :

français	anglais
aux termes de /…/ la loi-cadre sur les **fonctionnaires** des länder /…/ seul peut être nommé fonctionnaire celui qui « offre la garantie qu'il prendra constamment fait et cause pour le régime fondamental libéral et démocratique au sens de la loi fondamentale. »	*by virtue of /…/ the **civil service** (general principles) act for the länder, appointments to the civil service are subject to the requirement that the persons concerned "satisfy the authorities that they will at all times uphold the free democratic constitutional system within the meaning of the basic law".*

(Corpus CONVENTION)

Dans ces dernières phrases, c'est l'expression *civil service* qui correspond à la forme « fonctionnaires ». Comme nous l'avons montré dans la section 4.1.3, ce type de correspondance entre les unités lexicales relève de la notion d'équivalence contextuelle. Le segment *civil service* et son voisinage lexical immédiat sont liés au sein de la même locution : *the civil service (general principles) act*. Sur le plan sémantique, il s'agit d'unités traductionnelles singulières qui nécessitent un traitement particulier. Il appartient à l'expert humain de s'appuyer sur les blocs alignés pour examiner dans le détail les parallèles et les divergences entre ce type de séquences : la loi-cadre sur les fonctionnaires ~ *the civil service (general principles) act*.

Exemple 3 – administr+ ≠ *administ*+

Objectif : rechercher les contextes où les mots français commençant par la chaîne administr+ (administration, administrer etc.) ne sont pas traduits par des mots anglais commençant par la chaîne *administ*+ (*administration*, *administering*, etc.).

Pour cette recherche, nous utiliserons les principes de navigation textométrique similaires à ceux utilisés dans les deux exemples précédents. Sous *Lexico3*, le langage des « expressions régulières »[5] permet à l'utilisateur de constituer des groupes de mots correspondant au type de son choix et d'enregistrer la liste de ces unités pour une exploration ultérieure. Le repérage des sections de la carte en fonction de la présence/absence des types bilingues administr+ / *administ*+ (français/anglais) laisse apparaître des sections monochromes où sont attestées des équivalences traductionnelles originales telles que :

français	anglais
l'**administr**ation des douanes	*the customs*
bonne **administr**ation	*good governance*
dépositions **administr**atives	*procedural provisions*
le recours **administr**atif	*the non-contentious application*

5. Conclusion

Au terme de cette étude, nous avons défini une approche qui permet d'accéder à la description automatique de relations de correspondance entre des unités polysémiques qui possèdent plusieurs traductions au sein du corpus bi-textuel.

Les méthodes quantitatives convoquées au cours de nos expérimentations reposent entièrement sur des ressources construites automatiquement à base de corpus. Ces méthodes s'appuient sur des algorithmes qui utilisent les fréquences et les distributions des unités textuelles prises comme points de repère pour l'identification et l'extraction des correspondances. Nous avons montré que les comparaisons des fréquences des unités textuelles recensées dans les deux volets bilingues du corpus sont souvent insuffisantes pour détecter les correspondances traductionnelles au niveau lexical. Les différents sens dans lesquels le lexème est employé dans un contexte donné induisent la plupart du temps autant de traductions différentes. Les mots dotés d'un large éventail de sens dans le corpus forment des réseaux de correspondances souvent complexes. Ces facteurs entraînent des écarts entre les fréquences des unités équivalentes prises dans des contextes particuliers.

Une description automatique des multiples relations d'équivalence entre unités bilingues peut être obtenue par le biais d'appariements statistiques lorsque l'exploration du corpus bi-textuel s'appuie sur le schéma de résonance textuelle. Cette approche peut être utilisée pour le repérage des équivalences lexicales y compris dans le cas où leurs fréquences dans le corpus sont peu élevées.

L'éclairage quantitatif est incontournable pour construire des analyses nuancées de ressources textuelles multilingues. Les possibilités d'exploration intertextuelle ouvertes par cette approche facilitent la mise en évidence de phénomènes traductionnels complexes, relevant de différents niveaux de l'analyse linguistique : la variation des traductions d'un terme en fonction des contextes, le repérage thématique d'équivalences lexicales, la découverte de constellations lexicales parallèles, etc. L'observation de ces phénomènes enrichit la pratique quotidienne de traducteurs, lexicographes, terminologues, enseignants en langues étrangères, spécialistes de l'analyse de discours, etc.

NOTES

1. Ordre lexicométrique (pour les formes graphiques) : ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes ; les formes de même fréquence sont classées par ordre lexicographique.

2. La méthode des spécificités (Lafon, 1984) met en évidence pour chaque unité de décompte les fragments de corpus dans lesquelles l'unité possède de nombreuses occurrences (spécificités positives) ainsi que celles où son effectif est au contraire anormalement faible (spécificités négatives). On calcule le diagnostic de spécificité relatif à l'effectif constaté à base des paramètres suivants : sous-fréquence de l'unité dans le fragment textuel (f_{locale}), fréquence de l'unité dans l'ensemble du corpus (F), nombre des unités dans le fragment, nombre total des unités du corpus. Un

calcul probabiliste permet de porter un jugement sur l'effectif analysé. Si l'effectif se situe dans les limites de ce que le calcul permettait d'espérer, la répartition constatée est considérée « banale ». Si ce n'est pas le cas, on calcule un indice de spécificité de l'unité.
3. La recherche dans le domaine de l'alignement a montré que le repérage automatique des correspondances est relativement simple dans le cas d'unités de texte de taille importante, telles que chapitres, sections, articles, paragraphes, etc. L'utilisation des méthodes probabilistes a donné lieu à des avancées rapides dans l'alignement des phrases. Les comptes rendus d'expériences publiés récemment décrivent des algorithmes permettant d'apparier les phrases d'un corpus parallèle avec un taux de réussite élevé (Véronis 2000).
4. Le *garde-mots* est une fonctionnalité du logiciel de statistique textuelle *Lexico3* (http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/) permettant de mémoriser différents types d'unité textuelle (formes, segments, etc.) pour une utilisation ultérieure. Pour stocker une unité quelconque dans le garde-mots il suffit de la faire glisser sur l'icône de cette fonctionnalité.
5. Les expressions régulières sont des ensembles d'opérateurs (méta-caractères) offrant la possibilité de représenter des portions de texte de manière générique.

RÉFÉRENCES

LAFON, P. (1984) : *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.
LAMALLE, E., C., MARTINEZ, W., FLEURY, S., SALEM, A. *et al*. (2003) : *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation*, SYLED–CLA2T, Université de la Sorbonne nouvelle –Paris 3. Disponible sur : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/manuelsL3/.
LAMALLE, C. et SALEM, A. (2002) : « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. », in *Actes des JADT'02*, p. 403-412.
LEBART, L. et SALEM, A. (1994) : *Statistique Textuelle*, Paris, Dunod.
SALEM, A. (1987) : *Pratique des segments répétés : essai de statistique textuelle*, Paris, Klincksieck.
VÉRONIS, J. (Ed.) (2000) : *Parallel Text Processing: Alignment and use of translation corpora*, Dordrecht, Kluwer Academic Publishers.
ZIMINA, M. (2004a) : « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles », in *Actes des JADT'04*, p. 1195-1202.
ZIMINA, M. (2004b) : *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*, Thèse de Doctorat en Sciences du langage, Université de la Sorbonne nouvelle – Paris 3. Disponible sur : http://www.cavi.univ-paris3.fr/ilpga/ED/student/stmz/ED268-PagePersoMZ_fichiers/stmz/page8.htm.