Alignement lexical dans les corpus bilingues : la méthode des spécificités

Maria Zimina

Université de la Sorbonne nouvelle - Paris 3

ED 268

LANGAGE & LANGUES

DESCRIPTION, THEORISATION, TRANSMISSION

RJC'2001

5ème rencontre *jeunes chercheurs*, 19 octobre 2001

LANGAGE & LANGUES, ***ILPGA,*** 19 rue des Bernardins, 75005 Paris

OCTOBRE 2001

ACTES DE LA 5EME JOURNEE *Jeunes chercheurs* de l'ED268 DU 19 OCTOBRE 2001

articles réunis et mis en page par :
Angélique AMELOT, Lucile CADET,
Olivier CORBIN, Cédric GENDROT,
Vincent GUIGUE, Charlotte JACQUEMOT,
Hana SKRABALOVA, Julie LEFEBVRE,
Maria ZIMINA.

Université de la Sorbonne nouvelle - Paris 3

ED 268 - *LANGAGE & LANGUES*

SOMMAIRE

RJC'2001 : Actes de la 5ème journée *Jeunes chercheurs* de l'ED 268, Université de la Sorbonne nouvelle - Paris 3

ALIGNEMENT LEXICAL DANS LES CORPUS BILINGUES : LA MÉTHODE DES SPÉCIFICITÉS

Maria Zimina
CLA2T – ILPGA, Université de la Sorbonne nouvelle - Paris 3
19, rue des Bernardins 75005 Paris
zimina@msh-paris.fr

Résumé
Dans les corpus bilingues alignés au niveau de la phrase, le repérage des équivalences lexicales à faible fréquence peut être effectué sur des bases quantitatives. Basée sur la pratique du calcul des spécificités, notre méthode explore parallèlement les contextes équivalents pour repérer des correspondances dans les emplois caractéristiques des différents types d'unités textuelles. L'intégration du calcul des co-occurrences multiples permet d'affiner la description des unités lexicales complexes. La réitération systématique de ce processus dans le corpus, éventuellement appuyée sur l'utilisation d'un dictionnaire ou d'un lexique bilingue, offre de nouveaux moyens d'appariement des mots et des syntagmes.

MOTS-CLES : corpus bilingues, lexicométrie, équivalence de traduction

Abstract
The approach suggested in this article enables statistic identification of low-frequency word correspondences of bilingual texts aligned on phrase level. Corresponding lexical units are discovered through characteristic element computation in parallel contexts. An extensive description of translation equivalence is obtained through the study of multiple co-occurrences. The calculation undergoes systematic reiteration in order to embrace the entire corpus. The exploratory results show that the use of quantitative methods in combination with a bilingual lexicon or a dictionary offers new prospects for improving automatic word alignment.

KEYWORDS: bilingual corpora, lexicometrics, translation equivalence

1. INTRODUCTION

Les corpus parallèles sont constitués de plusieurs volets qui correspondent chacun à une version d'un même texte dans deux langues différentes ou plus. Des ressources linguistiques obtenues à partir de corpus de textes bilingues (ou multilingues) peuvent être réutilisées efficacement dans des domaines tels que la lexicographie, la terminologie, la traduction, la recherche d'information, l'enseignement des langues.[91]

Pour rendre exploitable le potentiel des corpus parallèles et pour faciliter l'analyse et l'extraction d'équivalences à partir de ce type de textes, il faut d'abord procéder à la mise en correspondance automatique des unités liées sur le plan d'analyse de la traduction. Les différentes approches permettant l'appariement de segments de traduction constituent un secteur du *traitement*

[91] Les applications principales de textes parallèles alignés sont présentées dans Isabelle et Warwick-Armstrong (1993, pp. 301-303) ; Langlois (1996) ; Véronis (2000a, pp. 152-159) ; Véronis, (2000b, pp. 9-14).

automatique du langage (TAL) que l'on appelle l'*alignement automatique*. Un des principaux objectifs de recherche dans le domaine de l'alignement est le stockage des correspondances dans une sorte de *mémoire de traduction* gérée par ordinateur.[92] L'existence de bases de données de textes alignés permet de concevoir toutes sortes d'outils de recherche et d'extraction de ressources bilingues à base de corpus.

Les connaissances acquises dans l'alignement automatique des textes permettent actuellement d'identifier des correspondances de traduction au niveau de la phrase.[93] En revanche, la recherche automatique des correspondances plus fines demeure une tache difficile compte tenu de la diversité et de la complexité des liens de traduction au niveau des mots et des syntagmes (cf. Debili, 2000 ; Véronis, 2000a).[94]

2. Le statut des méthodes lexicométriques dans l'alignement

Les méthodes quantitatives trouvent des applications nouvelles dans le domaine de la mise en correspondance de textes bilingues. Le recours à ces méthodes pour l'alignement est motivé par le fait que les correspondances qu'elles produisent ne résultent pas de connaissances *a priori* sur les textes mais de similitudes qu'ils présentent au plan quantitatif.

Les procédures de segmentation automatique de la séquence textuelle servent de base aux comparaisons statistiques destinées à mettre en évidence des segments de traduction. Dans les études lexicométriques, les textes sont d'abord segmentés en occurrences de formes graphiques (chaînes de caractères bornées par deux caractères délimiteurs). Ces formes sont ensuite regroupées pour recenser dans la chaîne textuelle les différents *types* d'unités sur la base de leur identité ou de leur ressemblance. Le concept de *type généralisé* ***TGen***[95] permet de décrire des

---

[92] Dans la pratique, un ensemble de normes uniques d'encodage bi-textuel se met progressivement en place pour concevoir ce que l'on appelle la mémoire de traduction. Ces normes sont axées sur le standard TMX (Translation Memory Exchange Standard) proche de SGML/XML. Le standard TMX a été développé par LISA (Localization Industry Standards Association). Le TMX devrait permettre à moyen terme de converger vers un système commun d'archivage électronique des traductions existantes alignées, cf. Melby (2000).

[93] Les systèmes actuels d'alignement des phrases de textes parallèles multilingues ont fait récemment l'objet d'une étude d'évaluation menée au sein du projet ARCADE financé par l'AUPELF-UREF dans le cadre des Actions de Recherches Concertées "Ingénierie de la langue". Les résultats de l'étude témoignent de l'existence d'avancées méthodologiques importantes dans les techniques d'alignement des phrases. Lorsque les textes ne présentent pas de divergences importantes au niveau structurel (pas d'omissions, etc.), le taux de précision des systèmes évalués est estimé, en moyenne, à 98,5%, cf. Langlais *et al.*, (1998) ; Véronis et Langlais (2000).

[94] Malgré de nombreuses difficultés dans l'automatisation totale de l'alignement au niveau des mots et des syntagmes, il y a eu des avancées importantes dans la réflexion sur l'utilisation conjointe de plusieurs méthodes pour réaliser ce type de tâche, cf. Debili et Zribi (1996) ; Gaussier (1998) ; Choueka *et al.* (2000) ; Wu (2000).

[95] Le concept de *type généralisé* ***TGen*** est une définition très générique de type d'unité à recenser, cf. Lamalle et Salem (2002, p. 404-405). On peut recenser au-delà des occurrences des formes graphiques : les occurrences d'un segment répété (exemple : *démocratie apte à se défendre*) ; la rencontre de deux formes (ou co-occurrence) à l'intérieur d'une fenêtre de x-formes graphiques ou d'une phrase (*démocratie*

ensembles d'occurrences sélectionnés systématiquement dans le texte pour recenser un *segment répété*, une *co-occurrence* de deux ou plusieurs formes, ou un autre type d'unité lexicale défini en fonction de critères formels de l'étude (cf. Lamalle et Salem, 2002).

Privilégiant le point de vue lexicométrique, on peut procéder à la mise en correspondance automatique des segments de traduction issus de textes bilingues par localisation de *TGen(s)* avec des ventilations similaires. Pour effectuer cette comparaison, il faut d'abord identifier les zones de textes bilingues dans lesquelles seront recherchées des similitudes. Lorsque la fréquence des unités que l'on souhaite étudier est suffisamment élevée, il est possible d'effectuer des comparaisons dans l'ensemble du corpus. La démarche consiste à décomposer parallèlement les textes bilingues en parties de longueur fixe et à comparer les profils de répartition des unités. Des méthodes de classification automatique, telles que la *classification ascendante hiérarchique*, rapprochent ensuite des unités ayant des distributions similaires dans les deux volets du corpus (cf. Zimina, 2000). L'implication de ce type de méthodes dans l'alignement permet également le rapprochement automatique des groupes de mots équivalents et fournit des indices pour l'alignement des phrases.

La plupart du temps, une étude statistique portant sur la répartition des unités au sein de parties consécutives découpées dans le corpus se révèle insuffisante pour localiser avec précision les équivalences parmi les unités de basse fréquence. La prise en compte des résultats de l'alignement préalable des phrases du corpus permet d'affiner la description de segments de traduction.[96] Pour un repérage exhaustif des correspondances lexicales, il est utile d'établir, à côté des mesures statistiques portant sur la ventilation des unités dans l'ensemble du corpus, des diagnostics portant sur la répartition des mêmes unités dans les portions de texte équivalentes choisies en raison de l'abondance relative des occurrences d'un type donné. Ce type d'analyse peut être envisagé si l'on a recours à une *représentation topographique*[97] du texte.

Une cartographie de la *présence-absence* des unités bilingues au sein des phrases en correspondance donne de nouveaux moyens pour le recensement des équivalences parmi les unités souvent présentes dans les mêmes phrases ou dans les mêmes paragraphes que les occurrences du type considéré. On pourra ainsi mettre en évidence et localiser de manière automatique les unités de deux langues dont la présence significative au sein des phrases (ou paragraphes) équivalentes permet de faire une hypothèse sur l'existence d'une relation de correspondance, y compris lorsque leurs effectifs sont faibles.

+ *république*) ; le type constitué par les occurrences d'un ensemble de formes graphiques défini en raison de la parenté lexicale des ces dernières (exemple *démocratique, démocratie, démocratiques, démocrate*).

[96] L'existence d'un large spectre de méthodes d'alignement des phrases donne accès à de nombreux moyens pour automatiser ce type de tâche, cf. Brown *et al.* (1991) ; Gale et Church (1991) ; Kay et Röscheisen (1993) ; Simard *et al.* (1992) ; Véronis (2000b).

[97] La topographie textuelle a pour objectif une localisation graphique des phénomènes mis en évidence par l'étude statistique, cf. Lamalle et Salem (2002).

Pour illustrer l'utilisation de méthodes lexicométriques dans l'alignement, nous emprunterons des exemples à un corpus de textes juridiques anglais-français de la *Convention de sauvegarde des droits de l'homme et des libertés fondamentales*, désormais ***Convention***.[98]

3. DESCRIPTION DU CORPUS BILINGUE *CONVENTION*

Le corpus bilingue ***Convention*** est constitué des textes officiels de la *Convention de sauvegarde des droits de l'homme et des libertés fondamentales*, ainsi que des protocoles et des arrêts rendus par la Cour européenne des droits de l'homme de Strasbourg en 1995.[99]

Les premiers comptages réalisés sur le corpus donnent un aperçu grossier des principales caractéristiques lexicométriques des deux volets bilingues (cf. *Tableau 1*).

	occurrences	**formes**	**hapax**	**fmax**	
français	296396	12913	4959	*de*	17572
anglais	284958	9530	3407	*the*	29622

Liste de caractères-délimiteurs : .,::!?/_\"'()[]{}$$

Tableau 1 : Résultats de la segmentation du corpus *Convention*

4. RECHERCHE D'ÉQUIVALENCES FAIBLES SUR LE PLAN QUANTITATIF : SPÉCIFICITÉS

Dans le corpus pré-aligné au niveau de la phrase, l'analyse des distributions de *TGen(s)*[100] au sein des couples de phrases appariées est susceptible de mettre en relief des équivalences de traduction au niveau des mots et des syntagmes. Dans l'expérimentation qui suit, la version numérisée du corpus *Convention* est soumise à une série de traitements statistiques qui prennent en compte l'appariement des phrases.[101]

---

[98] Nous remercions Didier Bourigault (Equipe de Recherche en Syntaxe et Sémantique, CNRS – Université Toulouse II), qui a accepté de mettre en notre disposition le corpus *Convention*. Chaque volet du corpus compte approximativement 300 000 mots.

[99] Elaborée au sein du Conseil de l'Europe, la *Convention* définit un certain nombre de droits fondamentaux et institue un mécanisme de contrôle et de sanction propre à assurer le respect de ces droits par les Etats signataires. Il existe parallèlement deux versions officielles des documents mentionnés ci-dessus : l'une en français, l'autre en anglais, et il est impossible de distinguer une langue source et une langue cible.

[100] Nous appellerons *occurrence* chacun des éléments découpés par un algorithme de segmentation automatique au fil d'un corpus de texte et *TGen* (ou *type*) les divers regroupements de ces occurrences que l'on peut opérer sur la base de leur identité ou de leurs ressemblances, cf. Lamalle et Salem (2002, p. 404).

[101] Le vocabulaire bilingue du corpus *Convention* a fait l'objet de l'étude terminologique menée dans le cadre du projet "Lexique des Droits de l'Homme" financé par le Ministère français de la Recherche et de l'Enseignement Supérieur. Il visait la construction d'un *lexique bilingue des Droits de l'Homme* à partir de l'analyse d'un corpus textuel composé de la *Convention*. Au cours du projet, l'agencement des textes juridiques du corpus (la numérotation des parties, alinéas, paragraphes) a été transformée en une structuration logique qui peut être manipulée par l'ordinateur. Chaque volet du corpus a été découpé en 12

Le repérage des *spécificités* ou *vocabulaires caractéristiques* met en évidence, pour un groupe de phrases donné, les unités dont la fréquence connaît une variation importante dans ce fragment de texte.[102] La méthode permet de sélectionner pour un sous-ensemble quelconque de phrases du corpus une série de types surreprésentés dans ce fragment par rapport à l'ensemble du corpus. Il est possible d'établir ce genre de diagnostic parallèlement pour deux volets du corpus bilingue. Une fois repérées les unités dont les occurrences connaissent une abondance relative dans les fragments de texte équivalents, on peut mettre en évidence des liens de correspondance par une série de comparaisons statistiques entre elles. Dans l'exemple qui suit, nous montrerons que la faible fréquence des unités dans le corpus ne constitue pas un obstacle à leur rapprochement automatique par la méthode proposée.

L'exploration débute par le marquage au fil du texte dans un volet du corpus d'un sous-ensemble d'occurrences correspondant à un type quelconque. Le repérage des phrases équivalentes dans l'autre volet permet de construire deux fragments de texte équivalents qui seront confrontés au reste du corpus à des fins de comparaison. Pour illustrer notre propos, nous considérerons les phrases qui contiennent la forme *démocratie* (F=9) et le sous-ensemble de phrases équivalentes en anglais. Le calcul des spécificités permet de sélectionner parallèlement pour chacun de ces fragments une série de types particulièrement caractéristiques de ces parties du corpus. Le *tableau 2* présente quelques-uns de ces types mis en évidence statistiquement par la méthode. On constate sur ce tableau que les unités bilingues issues de l'exploration se correspondent au plan traductionnel. Ainsi, la forme française *démocratie* (spec.+*E27*), ayant servi de point d'entrée pour la construction de l'échantillon de phrases, peut être directement appariée avec la forme *democracy* (spec.+*E27*), la plus caractéristique du fragment anglais.

Nous appellerons ***Equivalence Traductionnelle Elémentaire (ETE)*** la liste des adresses d'un sous-ensemble d'occurrences des types bilingues appariés attestées dans les phrases équivalentes. Nous observons que la valeur de l'indice de spécificité est proportionnelle au nombre de rencontres des types avec l'*ETE démocratie/democracy* qui est la plus caractéristique du fragment :

ETE

démocratie	**+E27**	democracy	**+E26**
démocratie apte à se défendre	**+E12**	democracy capable of defending itself	**+E12**
apte	**+E12**	capable	**+E08**
défendre	**+E11**	defending	**+E11**
le cauchemar du nazisme	**+E06**	the nightmare of nazism	**+E06**
valeurs	**+E04**	values	**+E04**

131 phrases. Chaque couple de phrases équivalentes a reçu le même identifiant, cf. Bourigault *et al.* (1999).

[102] Fondée sur le *model hypergéométrique* [Lafon, 1984, pp. 54-68], la méthode des spécificités permet d'effectuer une comparaison entre l'ensemble du corpus (*T*) et l'échantillon des contextes contenant la forme pôle (*t*). En fonction de la fréquence totale d'une forme (*F*) et de sa fréquence locale (*f*), on affecte un indice de spécificité au cooccurrent. Le diagnostic est fourni sous la forme ±E*xx* où le signe indique un sur-emploi ou un sous-emploi de la forme et la valeur indique son degré de spécificité. Pour une analyse détaillée des principes fondamentaux de cette méthode, cf. Lebart et Salem (1994).

Le repérage des *TGen(s)* caractéristiques donne ainsi un premier aperçu de la structure des équivalences qui se forment au niveau des mots et des syntagmes dans cette partie du corpus. La comparaison des fréquences locales (*f*) et la localisation des unités dans les phrases alignées du fragment font apparaître des indices supplémentaires pour l'appariement. Le *tableau 3* permet de vérifier que les unités spécifiques sont liées sur le plan de la traduction lorsque leurs ventilations dans les phrases alignées du fragment sont similaires.

Par exemple, le segment français *démocratie apte à se défendre* (f=4, spec.+*E12*) est le segment anglais *democracy capable of defending itself* (f=4, spec.+*E12*) se correspondent dans le corpus (cf. *Tableau 5*). Pour obtenir une description plus précise des *TGen(s)* équivalents, il faut tenir compte des relations d'inclusion entre les différents types. Ainsi, les formes cooccurrentes *apte* (F=4, f=4) et *défendre* (F=22, f=5) font partie d'un segment plus large en français : *démocratie apte à se défendre* (F=4, f=4). De même pour les formes anglaises *capable* (F=34, f=4) et *defending* (F=7, f=4) attestées dans le segment *democracy capable of defending itself* (F=4, f=4) (cf. *Tableau 2*).

5. TOPOGRAPHIE TEXTUELLE : "L'ORGANISATION SPATIALE" D'ÉQUIVALENCES DE TRADUCTION

Comme nous l'avons montré dans les sections précédentes, des comparaisons statistiques à base de corpus ont permis d'apparier les *Tgen(s)* caractéristiques issus des fragments de texte équivalents élaborés autour du pôle bilingue *démocratie/democracy*. Les expériences sur le corpus montrent que le repérage des équivalences peut être complété si l'on parvient à une description plus large du pôle bilingue sur lequel s'appuie la sélection des phrases soumises au calcul des spécificités. Au delà des types évoqués plus haut, il est possible de définir d'autres séries d'unités à l'aide d'outils permettant l'accès au langage des *expressions régulières*.[103] Ce langage fournit des moyens formels pour mettre en évidence des ensembles de formes graphiques liées au plan lexical, telles que *démocratie*, *démocratique*, *démocrates*, *démocratiquement*, *démocratiser*, etc. Du point de vue de l'analyse sémantique, ces unités représentent un thème qui est matérialisé dans le texte du corpus au travers d'un vaste ensemble d'occurrences que l'on peut considérer comme un nouveau type. La *représentation topographique*[104] du corpus permet de localiser les zones textuelles correspondant à une forte concentration d'occurrences correspondant à ce type.

Sur la *figure 4*, la *description cartographique*[105] des deux volets du corpus divisé en phrases, transcrit simultanément la ventilation des types équivalents français/anglais *démocrat+* et

[103] Le langage des expressions régulières offre la possibilité de représenter des portions de texte à l'aide d'un ensemble riche d'opérateurs. Il est accessible sur la plupart des systèmes et plateformes informatiques. Sur les utilisations spécifiques des expressions régulières dans l'analyse textuelle, cf. Habert *et al.* (1998).

[104] On trouvera dans Lamalle et Salem (2002) des exemples d'études réalisées par le recours systématique à une représentation topographique du texte.

[105] La représentation topographique du corpus a été générée grâce aux fonctionnalités développées récemment dans le cadre du logiciel ***Lexico 3***, cf. Lamalle *et al.* (2001).

democra+. Chacun de ces types est constitué par l'ensemble d'occurrences des formes graphiques qui révèlent de la même famille morphologique :

> ***démocrat+*** [démocratique (96 occ.), démocratie (9 occ.), démocratiques (7 occ.), démocrate (1 occ.)]
> ***democra+*** [democratic (103 occ.), democracy (10 occ.), democrat (1 occ.)]

Pour chaque volet du corpus, les carrés de couleur sombre indiquent la présence, au sein de la phrase concernée, d'une occurrence au moins du type cartographié. La confrontation des deux graphiques révèle une correspondance presque totale dans la répartition des types à l'intérieur du corpus (cf. *Figure 4*). La localisation des phrases équivalentes où sont présents parallèlement les types *démocrat+* et *democra+*, permet d'envisager une analyse plus approfondie de la hiérarchie de correspondances qui se forment autour du thème de la démocratie dans les deux volets du corpus. Ainsi, les premiers résultats de l'étude des co-occurrences dans les mêmes phrases que les occurrences des types bilingues *démocrat+* et *democra+* permettent de compléter les diagnostics obtenus à partir de la seule correspondance *démocratie – democracy* (cf. *Figure 5*).

6. Conclusions

Au terme de cette étude, nous avons défini une série de méthodes qui permettent un rapprochement automatique des unités équivalentes dans les corpus de textes bilingues pré-alignés au niveau de la phrase. Privilégiant le point de vue lexicométrique, notre approche repose entièrement sur les ressources fabriquées à partir du corpus. Les fréquences et les distributions des formes servent de points de repère pour l'identification et l'extraction des correspondances.
Les résultats de nos expérimentations montrent que des équivalences peu fréquentes peuvent être mises en évidence par des méthodes lexicométriques lorsque l'exploration du corpus tient compte de l'alignement des phrases. La localisation des différents types d'unités textuelles dans les phrases en correspondance apporte une plus grande précision dans l'analyse des distributions et des attractions simultanées des unités. Par conséquent, la description des équivalences de traduction obtenue avec ce type de méthodes est beaucoup plus fine.

L'utilisation d'outils lexicométriques dans le cadre de l'alignement automatique des corpus constitue une piste de recherche prometteuse. Appuyée sur l'utilisation des ressources dictionnairiques, cette approche permet d'envisager la construction de nouvelles procédures informatiques susceptibles de dévoiler la complexité de la structure des équivalences qui se forment au niveau des mots et des syntagmes dans les textes originaux et leurs traductions.

Tableau 2 : Spécificités majeures

terme	F	f	spec.	orig	terme	F	f	spec.	orig.
démocratie	9	9	+E27	*	democracy	10	9	+E26	*
de la démocratie	5	5	+E15	*	of democracy	4	4	+E12	*
démocratie apte à se défendre	4	4	+E12	*	democracy capable of defending itself	4	4	+E12	*
apte	4	4	+E12	*	of defending	5	4	+E11	
défendre	22	5	+E11		defending	7	4	+E11	
se défendre	11	4	+E10		capable of	34	4	+E08	
à se	29	4	+E08		capable	34	4	+E08	
instaurer une	3	2	+E06		nightmare	2	2	+E06	*
cauchemar	2	2	+E06	*	democracy and	3	2	+E06	
de la démocratie et	2	2	+E06	*	the nightmare of nazism	2	2	+E06	*
la volonté d'	3	2	+E06		its constitution	3	2	+E06	
le cauchemar du nazisme	2	2	+E06	*	values of democracy	2	2	+E06	*
nazisme	2	2	+E06	*	being based	3	2	+E06	
volonté d'	5	2	+E05		based on the principle	3	2	+E06	
justifiant	11	2	+E05		of democracy and	2	2	+E06	*
la volonté	9	2	+E05		on the principle	4	2	+E06	
instaurer	6	2	+E05		nazism	2	2	+E06	*
allemands	11	2	+E05		led to its constitution being based/				
valeurs	15	2	+E04		/on the principle of	2	2	+E06	*
allemagne	38	2	+E04		justifying	7	2	+E05	
idée	27	2	+E04		founded	11	2	+E05	
mieux	20	2	+E04		values of	7	2	+E05	
volonté	15	2	+E04		imposed on	25	2	+E04	
particulière	33	2	+E04		values	15	2	+E04	
					civil	302	4	+E04	
					led to	23	2	+E04	
					principle	103	3	+E04	
					based on the	27	2	+E04	
					itself	103	3	+E04	
					germany	37	2	+E04	
					led	30	2	+E04	
					a special	14	2	+E04	
					the principle of	35	2	+E04	

Guide de lecture du tableau : Un emploi caractéristique d'un type (forme, segments répété, co-occurrence etc.) dans le fragment de texte français correspondant à l'échantillon des phrases où est attestée la forme *démocratie* et le fragment de texte anglais équivalent, est indiqué par un *indice de spécificité*. Le diagnostic est fourni sous la forme *±Exx* où le signe indique un *sur-emploi* ou un *sous-emploi* du type et la valeur indique son degré de spécificité. Seul les spécificités positives majeures sont représentées. L'astérisque indique que le type n'est présent que dans le fragment de texte courant.

Tableau 3 : Localisation des unités spécifiques dans les phrases alignées

			Phrases alignées du fragment								
terme	f	spec.	01	02	03	04	05	06	07	08	09
démocratie	9	+E27	1	1	1	1	1	1	1	1	1
democracy	9	+E26	1	1	1	1	1	1	1	1	1
de la démocratie	5	+E15	1			1	1			1	1
of democracy	4	+E12				1	1			1	1
démocratie apte à se défendre	4	+E12		1	1			1	1		
democracy capable of defending/											
/itself	4	+E12		1	1			1	1		
la volonté d'	2	+E06		1					1		
instaurer une	2	+E06		1					1		
led to its constitution being/	2										
/based on the principle of		+E06		1					1		
valeurs	2	+E06				1				1	
values of democracy	2	+E06				1				1	
allemagne	2	+E04		1				1			
germany	2	+E04		1				1			
particulière	33	+E04		1		1					
a special	14	+E04		1		1					

Guide de lecture du tableau : Le tableau indique la ventilation des unités dans les phrases alignées du fragment bi-textuel. Le fragment correspond à l'échantillon des phrases alignées du corpus où est attestée l'équivalence *démocratie/democracy*. Les deux premières colonnes permettent de confronter les fréquences locales (*f*) des unités bilingues équivalentes, ainsi que leurs indices de *spécificité*.

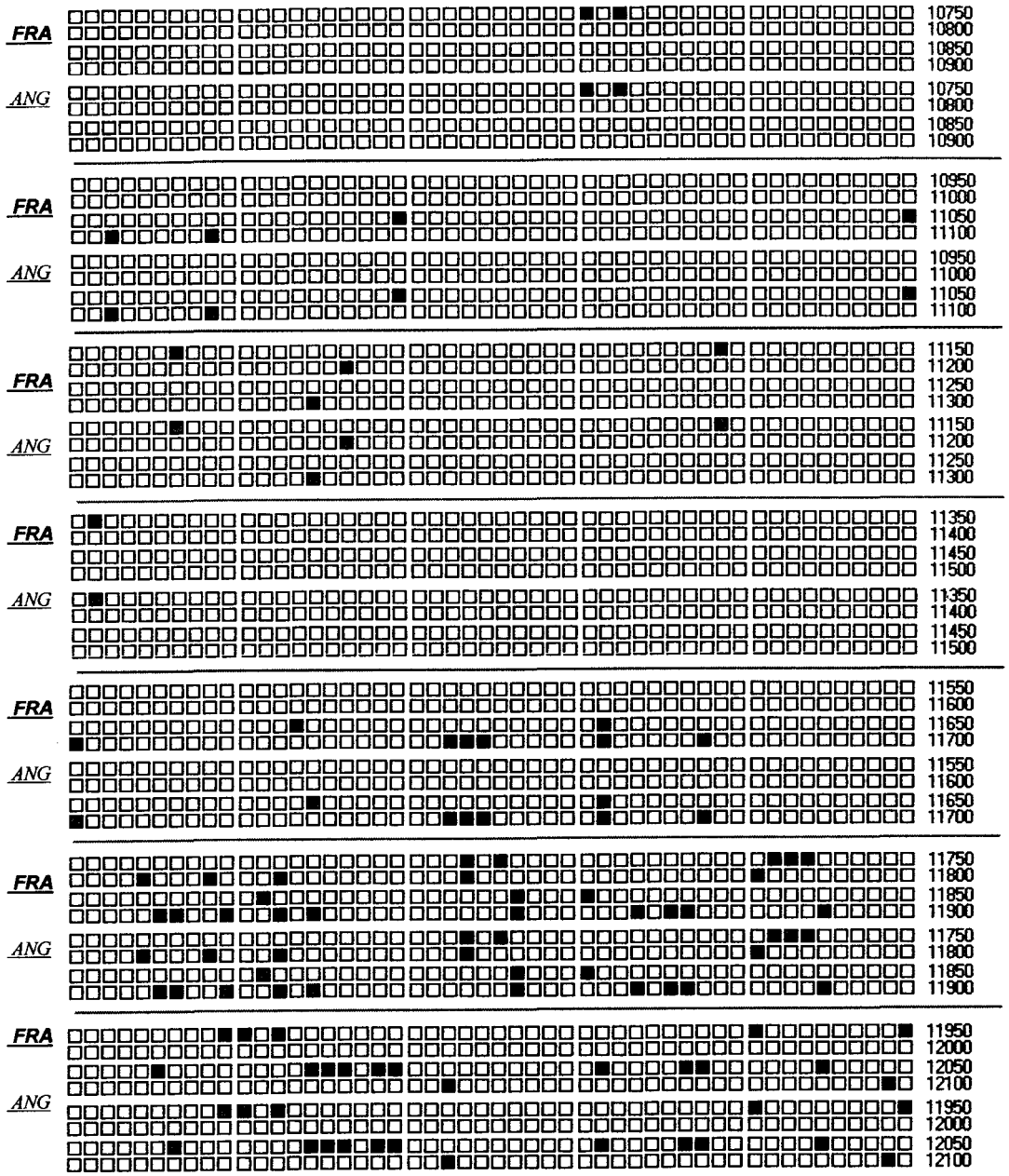
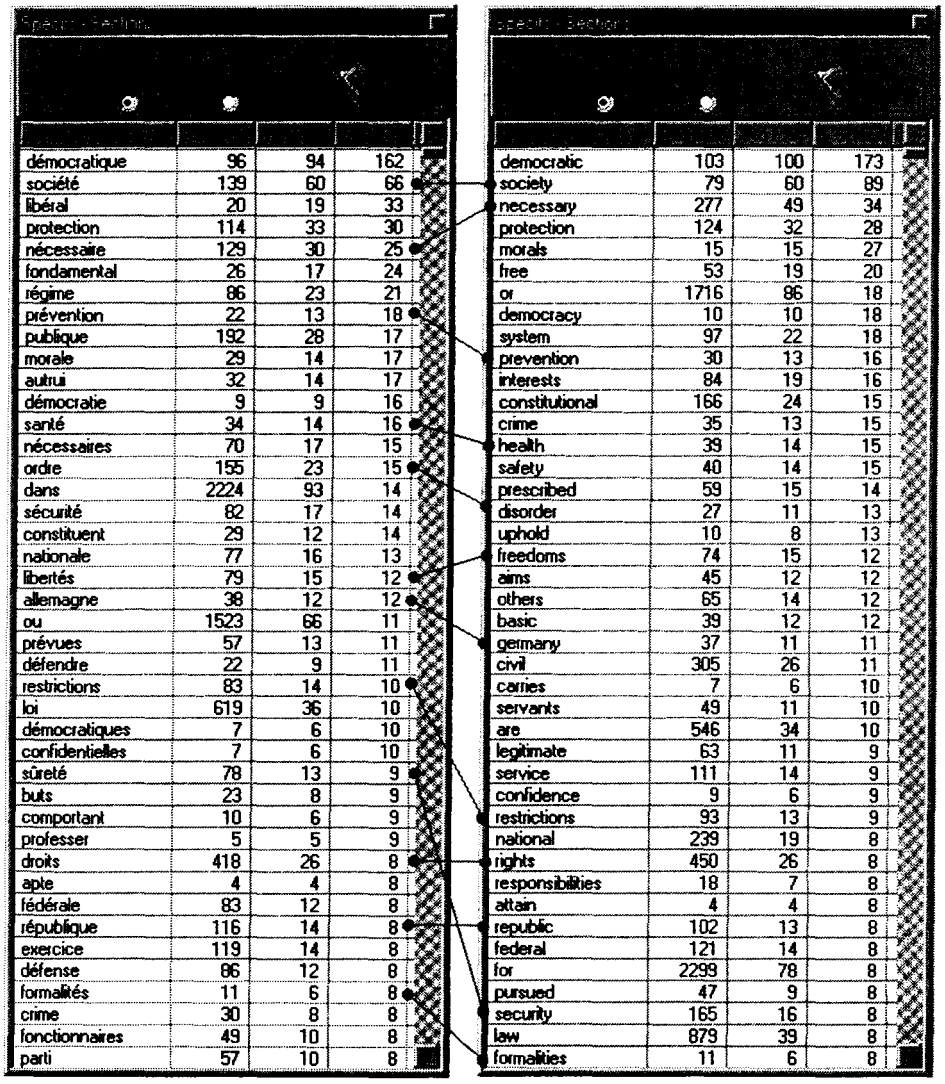


Figure 4 : Les occurrences des types bilingues ■*démocrat*+ et ■*democra*+ dans un extrait du corpus ***Convention***

Guide de lecture : La division du corpus en phrases est matérialisée par des carrées. Les carrées de couleur sombre indiquent la présence du type concerné dans la phrase.



démocratique	96	94	162
société	139	60	66
libéral	20	19	33
protection	114	33	30
nécessaire	129	30	25
fondamental	26	17	24
régime	86	23	21
prévention	22	13	18
publique	192	28	17
morale	29	14	17
autrui	32	14	17
démocratie	9	9	16
santé	34	14	16
nécessaires	70	17	15
ordre	155	23	15
dans	2224	93	14
sécurité	82	17	14
constituent	29	12	14
nationale	77	16	13
libertés	79	15	12
allemagne	38	12	12
ou	1523	66	11
prévues	57	13	11
défendre	22	9	11
restrictions	83	14	10
loi	619	36	10
démocratiques	7	6	10
confidentielles	7	6	10
sûreté	78	13	9
buts	23	8	9
comportant	10	6	9
professer	5	5	9
droits	418	26	8
apte	4	4	8
fédérale	83	12	8
république	116	14	8
exercice	119	14	8
défense	86	12	8
formalités	11	6	8
crime	30	8	8
fonctionnaires	49	10	8
parti	57	10	8

democratic	103	100	173
society	79	60	89
necessary	277	49	34
protection	124	32	28
morals	15	15	27
free	53	19	20
or	1716	86	18
democracy	10	10	18
system	97	22	18
prevention	30	13	16
interests	84	19	16
constitutional	166	24	15
crime	35	13	15
health	39	14	15
safety	40	14	15
prescribed	59	15	14
disorder	27	11	13
uphold	10	8	13
freedoms	74	15	12
aims	45	12	12
others	65	14	12
basic	39	12	12
germany	37	11	11
civil	305	26	11
carries	7	6	10
servants	49	11	10
are	546	34	10
legitimate	63	11	9
service	111	14	9
confidence	9	6	9
restrictions	93	13	9
national	239	19	8
rights	450	26	8
responsibilities	18	7	8
attain	4	4	8
republic	102	13	8
federal	121	14	8
for	2299	78	8
pursued	47	9	8
security	165	16	8
law	879	39	8
formalities	11	6	8

Figure 5 : Vocabulaire caractéristique des phrases contenant le pôle *démocrat+/democra+*

Guide de lecture : La confrontation des diagnostics obtenus pour chacun des volets du corpus met en évidence des équivalences spécifiques de l'univers lexical du pôle bilingue.

RÉFÉRENCES

BOURIGAULT Didier, CHODKIEWICZ Christine, HUMBLEY John (1999). "Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné". In *Actes de la troisième conférence 'Terminologie et Intelligence Artificielle'*, Nantes, 1999, pp. 70-77.

BROWN Peter, LAI Jennifer, MERCER Robert (1991). "Aligning Sentences in Parallel Corpora." In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkley, 1991, pp. 169-176.

CHUEKA Yaacov, CONLEY Ehud, DAGAN Ido (2000). "A comprehensive bilingual word alignment system. Application to disparate languages: Hebrew and English." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers, pp. 69-96.

DEBILI Fathi, ZRIBI Adnane (1996). "Les dépendances syntaxiques au service de l'appariement des mots." In *Actes du 10ème Congrès 'Reconnaissance des Formes et Intelligence Artificielle'*, Rennes, 1996.

DEBILI Fathi (2000). "L'appariement : quels problèmes ?" In Chibout K., Mariani J., Masson N. et al. (Eds.), *Ressources et évaluation en ingénierie des langues.* Bruxelles : De Boeck & Larcier s.a., pp. 101-125.

GAUSSIER Eric (1998). "Flow Network Models for Word Alignment and Terminology extraction from Bilingual Corpora." In *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, 1998, pp. 444-450.

HABERT Benoît, FABRE Cécile, ISSAC Fabrice (1998). *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques.* Paris : Masson, 320 p.

HABERT Benoît, NAZARENKO Adeline, SALEM André (1997). *Les linguistiques de corpus.* Paris: Armand Colin/Masson, 240 p.

KAY Martin, RÖCHEISEN Martin. (1993). "Text-Translation Alignment." Computational Linguistics, 19(1), pp. 121-142.

LABBÉ Dominique, THOIRON Philippe, SERANT Daniel (Eds.) (1988). *Etudes sur la richesse et la structure lexicales.* Paris-Genève : Slatkine-Champion, 172 p.

LAFON Pierre (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris: Slatkine-Champion., 217 p.

LAMALLE Cédric, SALEM André (2002). "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels." In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, 2002, pp. 403-412.

LAMALLE Cédric, MARTINEZ William, FLEURY Serge, SALEM André *et al.* (2001). *Lexico3 – Outils de statistique textuelle.* http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW

LANGE Jean-Marc, GAUSSIER Eric (1995) "Alignement de corpus multilingues au niveau des phrases." *TAL*, 36(1-2), pp. 67-80.

LANGLAIS Philippe, SIMARD Michel, VÉRONIS Jean et al. (1998). "ARCADE: A co-operative research project on bilingual text alignment." In *Proceedings of First International Conference on Language Resources and Evaluation*, Granada, 1998, pp. 289-292.

LANGLOIS Lucie (1996). "Bilingual Concordancers: A New Tool for Bilingual Lexicographers." In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montréal, 1996, pp. 34-42.

LEBART Ludovic, SALEM André (1994). *Statistique Textuelle*. Paris : Dunod, 342 p.

MARTINEZ William, ZIMINA Maria (2002). "Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues." In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, 2002, pp. 495-506.

MELBY Alan (2000). "Sharing of translation memory databases derived from aligned parallel text." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers, pp. 347-368.

SIMARD Michel, FOSTER George, ISABELLE Pierre (1992). "Using Cognates to Align Sentences in Bilingual Corpora." In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, 1992, pp. 67-81.

SOMERS Harold (1998). "Further Experiments in Bilingual Text Alignment." *International Journal of Corpus Linguistics*, 3(1), pp. 115-150.

VÉRONIS Jean (2000a). "Alignement de corpus multilingues." In Pierrel J.-M. (Ed.), *Ingénierie des langues.* Paris, Editions Hermès, pp. 151-171.

VÉRONIS Jean (Ed.) (2000b). *Parallel Text Processing: Alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers, 402 p.

VÉRONIS Jean, LANGLAIS Philippe (2000). "Evaluation of parallel text alignment systems. The ARCADE project." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers, pp. 369-388.

WU Dekai (2000). "Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers, pp. 139-167.

ZIMINA Maria (2000). "Alignement de textes bilingues par classification ascendante hiérarchique." In *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, Lausanne, 2000, pp. 171-178.