Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues

Maria Zimina

CLA2T

ILPGA, Université de la Sorbonne nouvelle - Paris 3

19, rue des Bernardins 75005 Paris

zimina@msh-paris.fr

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*

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*

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.[1] 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 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.[2] 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.[3] En revanche, la recherche automatique des correspondances plus fines demeure une tache difficile compte tenu de la diversité et de la complexité

[1] 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).

[2] 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).

[3] 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).

des liens de traduction au niveau des mots et des syntagmes (cf. Debili, 2000 ; Véronis, 2000a).[4]

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***[5] permet de décrire des 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

[4] 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).

[5] 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 + 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*).

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.[6] 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*[7] 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.

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***.[8]

---

[6] 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).

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

[8] 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.

Etude du vocabulaire bilingue du corpus *Convention* sur des bases quantitatives

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.[9]

1.1 Préparation du corpus pour une étude quantitative

Pour préparer une étude quantitative, nous avons entrepris une série de transformations[10] à partir du corpus aligné au niveau de la phrase qui nous a été fourni.[11] La segmentation parallèle des deux volets du corpus a permis la mise en œuvre de calculs statistiques en allégeant la manipulation de la chaîne textuelle.[12]

1.2 Analyse des gammes des fréquences

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

---

[9] 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.

[10] Les deux volets anglais et français ont été séparés. Les textes ont été ensuite transformés de manière à supprimer les majuscules. A ces exceptions près, le corpus n'a subi aucune annotation ni lemmatisation.

[11] 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 12 131 phrases. Chaque couple de phrases équivalentes a reçu le même identifiant. A la fin de cette opération, le découpage en phrases au sein des paragraphes était correct à environ 90 %, cf. Bourigault *et al.* (1999).

[12] Le graphisme de chacune des formes a été remplacé par son numéro d'ordre lexicométrique. Un dictionnaire de formes graphiques a été généré afin de permettre la reconstitution du graphisme de chacune des formes du texte, cf. Lebart et Salem (1994, pp. 42-45).

Tableau 1 : Résultat de la segmentation du corpus Convention[13]

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

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

Nous constatons sur le *tableau 1* que le nombre total d'occurrences est plus important pour le volet *français* (296396 contre 284958 pour le volet *anglais*). Le volet français du corpus est également plus "riche" en formes diverses (12913 > 9530) et compte beaucoup plus d'hapax[14] (4959 > 3407). La fréquence maximale *the* en anglais est largement supérieure à celle du texte français *de* (29622 > 17572).

Pour une description plus précise de la gamme des fréquences, on utilisera le diagramme de Pareto.[15] La *figure 2* nous permet de constater la plus grande diversité de formes employées dans le volet français. Les écarts entre les deux courbes concernent notamment les deux extrémités des gammes des fréquences : les formes de faible fréquence et celles de fréquence maximale. Dans l'intervalle des fréquences moyennes (20 à 1000 occ.) les courbes correspondant aux deux volets du corpus sont très proches.

Le plus grand nombre des formes dans le volet français s'explique par l'emploi de noms et d'adjectifs soumis à un accord morphologique [ex. *contraignant* (F=2), *contraignante* (F=2), *contraignants* (F=2), *contraignantes* (F=2)]. La confrontation de contextes équivalents où sont attestées les occurrences de formes de haute fréquence révèle que le volet anglais du corpus est particulièrement riche en formes polysémiques lesquelles reçoivent plusieurs traductions en français. 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).

---

[13] La segmentation a été réalisée à l'aide d'outils de statistique textuelle regroupés au sein du logiciel *Lexico* développé par le *Centre de Lexicométrie et d'Analyse Automatique des Textes* (CLA2T), Paris 3 : www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW

[14] Les hapax sont des formes qui ne sont attestées qu'une seule fois dans le texte.

[15] Le diagramme de Pareto permet de donner une représentation graphique de la gamme des fréquences. Sur l'axe vertical, gradué selon une échelle logarithmique, on porte la fréquence de répétition *F* (de 1 à F_{max}). Sur l'axe horizontal, gradué selon la même échelle, on indique pour chacune des valeurs de la fréquence *F*, le nombre *N(F)* des formes répétées au moins *F* fois dans le corpus. Les points ainsi tracés s'alignent approximativement le long d'une ligne droite. Pour une explication plus détaillée de ce phénomène, cf. Lebart et Salem (1994) ; Labbé *et al.* (1988).

Guide de lecture du Diagramme de Pareto :

Sur l'axe vertical (*Y*), gradué selon une échelle logarithmique, on lit la fréquence de répétition *F*.

Sur l'axe horizontal (*X*), gradué selon la même échelle, on porte le nombre des formes répétées au moins *F* fois dans le corpus pour chaque fréquence *F* (entre 1 et F_{max}).

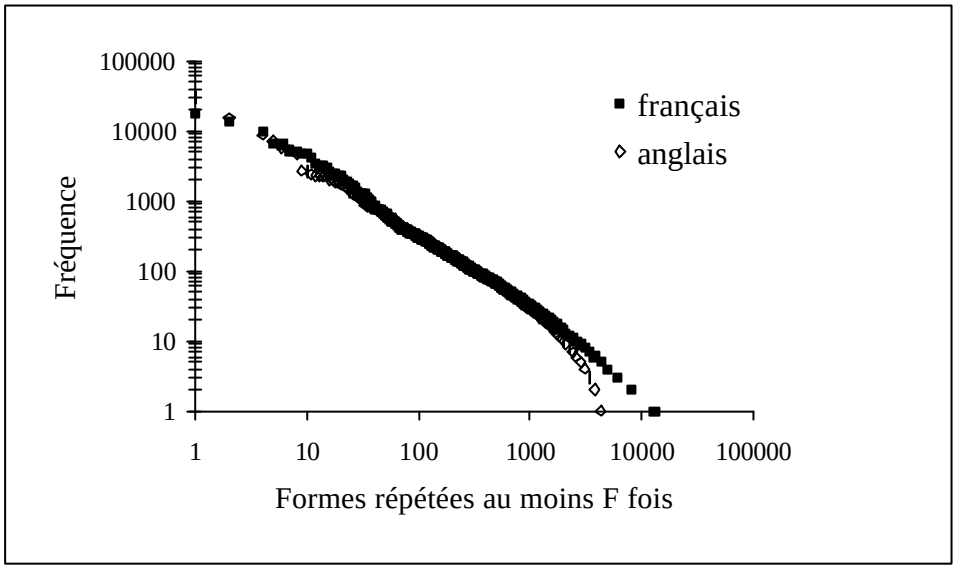

Figure 2 : Diagramme de Pareto pour les volets français et anglais du corpus *Convention*

L'analyse des dictionnaires des segments répétés constitués à partir des deux volets du corpus permet de poursuivre la comparaison des caractéristiques fréquentielles (cf. *Tableau 3*).

Malgré le statut intermédiaire de l'unité type *segment répété* du point de vue de la segmentation du corpus en unités de signification pertinentes, les inventaires de segments répétés issus des deux volets bilingues du corpus donnent des indices précieux pour découvrir les moyens formels d'expression qui servent à maintenir l'équivalence au niveau des mots et des syntagmes. Sur le *tableau 3*, les segments recensés autour des formes équivalentes *cour/court* sont triés par ordre décroissant des fréquences. La confrontation des deux listes laisse apparaître des proximités dans les fréquences générales des équivalences. Ainsi, le segment anglais *the court* (F=1652) est traduit en français par *la cour* (F=1853) ; *the administrative court* (F=126) par *la cour administrative* (F=133) ; *the supreme court* (F=112) par les segments *la cour de cassation* (F=67) et *la cour suprême* (F=58), etc. On remarque que ces similitudes sont d'autant plus mises en relief par les dictionnaires que le rang des fréquences des segments concernés est élevé.

Tableau 3 : Extraits des inventaires des segments répétés autour des formes bilingues *cour/court*

Freq	Lg	Segment FRA	Segment ANG	Lg	Freq
1853	2	la **cour**	the **court**	2	1652
221	2	**cour** d	**court** of	2	315
203	3	la **cour** d	the **court** of	3	213
159	3	de la **cour**	**court** of appeal	3	211
150	3	de la **cour**	administrative **court**	2	209
145	3	à la **cour**	the **court** of appeal	4	147
135	3	devant la **cour**	the administrative **court**	3	126
135	2	**cour** administrative	supreme **court**	2	117
133	3	la **cour** administrative	the supreme **court**	3	112
125	3	de la **cour**	to the **court**	3	108
107	2	**cour** de	before the **court**	3	107
103	3	la **cour** de	of the **court**	3	101
70	3	**cour** de cassation	of **court**	2	77
67	4	la **cour** de cassation	constitutional **court**	2	75
67	3	par la **cour**	rules of **court**	3	72
65	3	la **cour** a	of rules of **court**	4	70
64	3	la **cour** constitutionnelle	the **court** shall	3	67
58	3	la **cour** suprême	of rules of **court** a	5	64
55	4	décision de la **cour**	**court** shall	2	63
53	3	que la **cour**	to the **court**	3	61
51	5	la décision de la **cour**	the constitutional **court**	3	58
50	3	la **cour** n	the **court** notes	3	56
44	4	de la **cour** d	**court** a	2	56
44	3	la **cour** à	insurance **court**	2	56
43	4	déférée à la **cour**	**court** by	2	55
41	6	la juridiction obligatoire de la **cour**	**court** had	2	55
41	4	président de la **cour**	jurisdiction of the **court**	4	51
40	3	la **cour** estime	**court** is	2	51
38	5	été déférée à la **cour**	**court** by the	3	50
37	7	a été déférée à la **cour** par	regional **court**	2	49
36	8	affaire a été déférée à la **cour** par	the **court** by	3	48
34	7	reconnaissant la juridiction obligatoire de la **cour**	a **court**	2	47
33	5	déférée à la **cour** par	to the **court** by the	5	46
33	3	la **cour** avait	the **court** notes that	4	46
32	4	de la **cour** administrative	referred to the **court**	4	46
32	3	la **cour** rappelle	president of the **court**	4	45
32	3	la sheriff **court**	**court** has	2	45
30	11	affaire a été déférée à la **cour** par la commission européenne	the president of the **court**	5	44

Lorsque les effectifs sont faibles, il faut faire appel à des méthodes qui vont au-delà de la simple comparaison des fréquences. Les approches hybrides qui allient les méthodes de *topographie textuelle* et l'*analyse des spécificités* peuvent contribuer considérablement à l'appariement.

Recherche d'équivalences faibles sur le plan quantitatif

Dans le corpus pré-aligné au niveau de la phrase, l'analyse des distributions de *TGen(s)*[16] 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.

1.3 Spécificités

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.[17] 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.

---

[16] 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).

[17] 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 co-occurrent. Le diagnostic est fourni sous la forme *±Exx* 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).

Application au corpus Convention : *ETE*

La recherche de correspondances par la méthode des spécificités s'appuie sur les résultats de l'alignement du corpus au niveau de la phrase. 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 4a-b* 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***

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 5* 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).

Tableau 4a: Spécificités majeures

texte français

terme	F	f	spec.	orig.
démocratie	9	9	+*E27*	*
de la démocratie	5	5	+*E15*	*
démocratie apte à se défendre	4	4	+*E12*	*
apte	4	4	+*E12*	*
défendre	22	5	+*E11*	
se défendre	11	4	+*E10*	
à se	29	4	+*E08*	
instaurer une	3	2	+*E06*	
cauchemar	2	2	+*E06*	*
de la démocratie et	2	2	+*E06*	*
la volonté d'	3	2	+*E06*	
le cauchemar du nazisme	2	2	+*E06*	*
nazisme	2	2	+*E06*	*
volonté d'	5	2	+*E05*	
justifiant	11	2	+*E05*	
la volonté	9	2	+*E05*	
instaurer	6	2	+*E05*	
allemands	11	2	+*E05*	
valeurs	15	2	+*E04*	
allemagne	38	2	+*E04*	
idée	27	2	+*E04*	
mieux	20	2	+*E04*	
volonté	15	2	+*E04*	
particulière	33	2	+*E04*	

Guide de lecture du tableau :

Un emploi caractéristique d'un type (forme, segment répété, co-occurrence etc.) dans le fragment de texte français correspondant à l'échantillon des phrases où est attestée la forme *démocratie*, 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. Le caractère * (astérisque) indique que le type n'est présent que dans le fragment de texte courant.

Tableau 4b: Spécificités majeures

texte anglais

terme	F	f	spec.	orig.
democracy	10	9	+*E26*	
of democracy	4	4	+*E12*	*
democracy capable of defending itself	4	4	+*E12*	*
of defending	5	4	+*E11*	
defending	7	4	+*E11*	
capable of	34	4	+*E08*	
capable	34	4	+*E08*	
nightmare	2	2	+*E06*	*
democracy and	3	2	+*E06*	
the nightmare of nazism	2	2	+*E06*	*
its constitution	3	2	+*E06*	
values of democracy	2	2	+*E06*	*
being based	3	2	+*E06*	
based on the principle	3	2	+*E06*	
of democracy and	2	2	+*E06*	*
on the principle	4	2	+*E06*	
nazism	2	2	+*E06*	*
led to its constitution being based/ /on the principle of	2	2	+*E06*	*
justifying	7	2	+*E05*	
founded	11	2	+*E05*	
values of	7	2	+*E05*	
imposed on	25	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*	
she	246	4	+*E04*	
constitution	118	3	+*E04*	
notion	20	2	+*E04*	

Guide de lecture du tableau :

Un emploi caractéristique d'un type (forme, segment répété, co-occurrence etc.) dans le fragment de texte anglais équivalent à l'échantillon français (cf. *Tableau 4a*) 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. Le caractère * (astérisque) indique que le type n'est présent que dans le fragment de texte courant.

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 co-occurrentes *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. *Tableaux 4a-b*).

1.3.1 Prise en compte des co-occurrences multiples

Lorsque l'équivalence des types est posée, on peut utiliser ce pôle bilingue, correspondant à une nouvelle *ETE*, pour la détection de l'ensemble d'unités surreprésentées dans les phrases du corpus où il est attesté. Cette recherche de correspondances basée sur l'itération du calcul des spécificités permet de saisir les attractions lexicales multiples entre les types. Par exemple, l'exploration de l'environnement lexical des *Tgen(s)* équivalents correspondant à la co-occurrence des formes *démocratie + apte + défendre* (F=4)[18] en français et à celle des formes *democracy + defending + capable* (F=4) en anglais, met en évidence les éléments divers qui y sont associées dans les phrases (cf. *Tableau 6*).

Une recherche approfondie des correspondances parmi les unités bilingues qui illustrent la spécificité d'un fragment bi-textuel peut être appuyée sur l'exploration de l'ensemble de phrases du corpus qui contiennent les occurrences de ces unités.

Comme le montre le *tableau 7*, lorsque la prise en comte des ventilations des types dans les phrases du fragment ne permet pas de départager les unités de même fréquence locale (*f*) [ex. *souhaitait* 8[+3](1), *fondant* 25[+3](1), *avoid* 13[+3](1), *founding* 2[+3](1)], l'ambiguï té peut être levée si l'on prend soin de comparer les ventilations de ces mêmes unités dans tout le corpus. Appuyée sur le marquage cartographique en *présence/absence* (cf. *Figure 10*), cette approche fait ressortir systématiquement des profils de distribution similaires. Nous constatons ainsi que les formes verbales *fondant* (F=25) et *fonder* (F=13) apparaissent dans les phrases françaises équivalentes à celles qui contiennent la forme anglaise *founding* (F=2) (cf. *Tableau 7*). Lorsqu'il s'agit des hapax du corpus, la mise en correspondance peut être appuyée sur des ressources dictionnairiques.

[18] La fréquence d'une co-occurrence correspond au nombre d'*unités de contextes* où est attestée la rencontre de deux ou plusieurs formes. L'unité de contexte retenue pour notre exploration correspond à la longueur de la phrase.

Tableau 5 : 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/											
/based on the principle of	2	*+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 :

La partie droite du tableau indique la ventilation des unités dans les phrases alignées du fragment bi-textuel. Chaque fragment correspond à un 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*.

Tableau 6 : Co-occurrences multiples

Co-occurrents de ***démocratie + apte + défendre*** (NbUC=4)*

terme	F	f	spec.	NbUC
cauchemar	2	2	+E07	2
nazisme	2	2	+E07	2
instaurer	6	2	+E06	2
volonté	15	2	+E05	2
république	116	2	+E04	1
allemagne	38	2	+E04	2
pilier	**1**	**1**	**+E04**	**1**
wehrhafte	1	1	+E04	1
a	3003	5	+E03	2
se	1292	4	+E03	4
éviter	25	1	+E03	1
fondant	**25**	**1**	**+E03**	**1**
expérience	23	1	+E03	1
revêt	18	1	+E03	1
conduit	14	1	+E03	1
relevé	12	1	+E03	1
nouvel	10	1	+E03	1
constituée	8	1	+E03	1
souhaitait	8	1	+E03	1
weimar	5	1	+E03	1
connue	4	1	+E03	1
expériences	4	1	+E03	1
répétition	4	1	+E03	1

Co-occurrents de ***democracy + defending + capable*** (NbUC=4)

terme	F	f	spec.	NbUC
nazism	2	2	+E07	2
nightmare	2	2	+E07	2
itself	103	3	+E05	3
led	30	2	+E05	2
principle	103	2	+E04	2
republic	102	2	+E04	1
germany	37	2	+E04	2
founding	**2**	**1**	**+E04**	**1**
cornerstone	**1**	**1**	**+E04**	**1**
itself	1	1	+E04	1
wehrhafte	1	1	+E04	1
its	739	3	+E03	3
being	283	2	+E03	2
based	130	2	+E03	2
constitution	118	2	+E03	2
notion	20	1	+E03	1
avoid	13	1	+E03	1
founded	30	1	+E03	1
idea	9	1	+E03	1
weimar	5	1	+E03	1
experiences	4	1	+E03	1
repetition	4	1	+E03	1

**NbUC* / Nombre d'unités de contextes / = nombre de phrases où est attestée la co-occurrence.

1.3.2 Réseaux de co-occurrences

Pour exhiber de manière automatique les attractions simultanées entre les unités qui maintiennent l'équivalence traductionnelle entre les deux volets du corpus, il est possible de faire appel au calcul des réseaux de co-occurrences. La méthode élaborée par W. Martinez[19] explore les contextes spécifiques à partir d'un pôle et révèle à chaque étape du calcul un élément supplémentaire du réseau de co-occurrences qui s'élabore à partir de celui-ci. L'application de cette méthode parallèlement aux deux volets du corpus révèle des similitudes entre les systèmes co-occurrentiels des pôles *démocratie* et *democracy* (cf. *Figures 8a-b ; Figure 9*).

1.4 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*. La réitération systématique du calcul appuyée sur le repérage des profils de ventilation homogènes a fait apparaître de nouveaux éléments faisant partie de l'univers lexical du pôle.

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

---

[19] Elaboré autour du logiciel *Lexico*, le module des réseaux des co-occurrences est fondé sur la réitération du calcul :

Etape 1 : On calcule pour le pôle A les co-occurrents spécifiques B, C et D
Etape 2 : Dans leurs contextes communs, on calcule pour les pôles A+B les co-occurrents spécifiques E et F
Etape 3 : Les pôles A+B+E ont pour co-occurrent spécifique H
Etape 4 : Les pôles A+B+E+H n'ont pas de co-occurrent spécifique et l'exploration s'interrompt pour ce chemin
Etape 5 : Les pôles A+B+F ont pour co-occurrents spécifiques I, etc.

Durant l'exploration, différents filtrages conditionnent l'épuisement des explorations contextuelles et réduisent le bruit dans les résultats pour privilégier l'information la plus spécifique : seuils maximaux de fréquence et de spécificité du co-occurrent, nombre minimal de contextes où se produit la co-occurrence et exclusion des mots-outils. A l'issue du calcul, on sélectionne les chemins originaux en écartant les chemins qui se contiennent (AB et ABC contenus dans ABCD) ou qui se répètent (ACB, BAC, BCA, CAB et CBA contenus dans ABC). Sur la description de la méthode de calcul des réseaux de co-occurrences dans l'appariement des mots, cf. Martinez et Zimina (2002).

à l'aide d'outils permettant l'accès au langage des *expressions régulières*.[20] 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*[21] du corpus permet de localiser les zones textuelles correspondant à une forte concentration d'occurrences correspondant à ce type.

Sur la *figure 10*, la *description cartographique*[22] des deux volets du corpus divisé en phrases, transcrit simultanément la ventilation des types équivalents français/anglais *démocrat+* et *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 10*).

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 11*).

[20] 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 questions de la syntaxe des expressions régulières, Desgraupes (2001). Sur les utilisations spécifiques dans l'analyse textuelle Habert *et al.* (1998).

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

[22] 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).

Tableau 7 : Retours au contexte

	fragment
$_ceux-ci seraient en effet le **pilier** 1[+4](1) d'une "*démocratie apte* à se *défendre*" :	$_the civil service was the **cornerstone** 1[+4](1) of a "*democracy capable* of *defending* itself " :
$_l'allemagne **souhaitait** 8[+3](1) éviter la **répétition** 4[+3](1) de ces **expériences** 4[+3](1) en **fondant** 25[+3](1) son nouvel etat sur l'idée de "*démocratie apte* à se *défendre* " .	$_germany wished to **avoid** 13[+3](1) a **repetition** 4[+3](1) of those **experiences** 4[+3](1) by **founding** 2[+3](1) its new state on the idea that it should be a "*democracy capable* of *defending* itself " .

corpus

*$ _ ils étaient rentrés en turquie de leur propre volonté et avec un but bien précis , **fonder** le parti communiste unifié turc (paragraphes 7 et 13 ci - dessus) ; ils ne pouvaient ignorer qu ' ils seraient poursuivis pour cela .*
*$_ they had returned to turkey of their own accord and with the specific aim of **founding** the turkish united communist party (see paragraphs 7 and 13 above) and they could not be unaware that they would be prosecuted for this .*

Guide de lecture du tableau :

Le tableau permet de localiser dans le contexte quelques co-occurrents des pôles bilingues *démocratie* + *apte* + *défendre* et *democracy* + *defending* + *capable*. Un diagnostic est fourni pour chaque co-occurrent sous la forme *F*[s]*f* où *F* = fréquence générale, [*s*] = indice de spécificité, (*f*) = fréquence locale. On note que l'analyse de l'ensemble du corpus permet d'affiner la description des équivalences lorsque les fréquences et la localisation dans les phrases du fragment courant ne fournissent pas d'indices suffisamment précis pour l'appariement.

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.

Remerciements

Je tiens à remercier Jean VÉRONIS, Serge FLEURY et André SALEM pour leurs avis et leurs conseils qui m'ont beaucoup aidée au cours de ce travail.

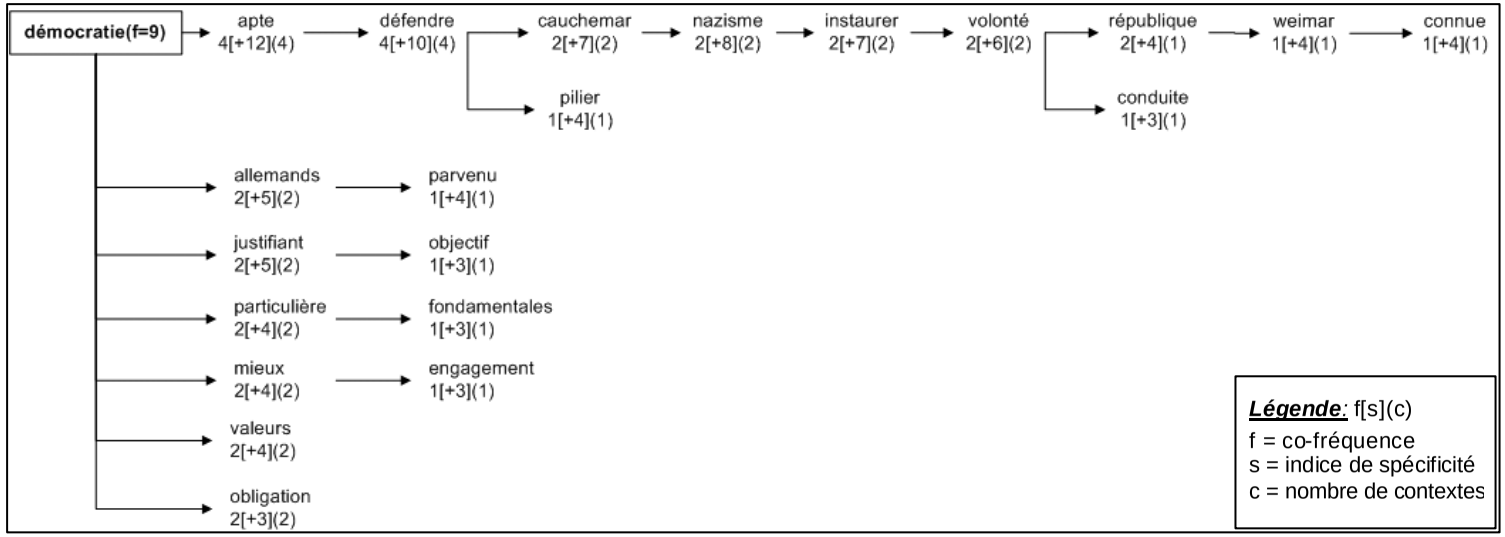

Figure 8a: Vue partielle du réseau de co-occurrences élaboré à partir du pôle *démocratie*

Guide de lecture de la figure :

Basée sur la réitération du calcul des co-occurrences, la recherche de réseaux met en évidence des associations multiples au sein de l'unité contextuelle de la phrase. La figure présente sous la forme d'une arborescence les résultats du calcul exploratoire pour la forme pôle démocratie. Sur la première ligne qui correspond à la branche la plus spécifique du réseau, est rapportée une forte co-occurrence du pôle avec les formes défendre et apte : 4 rencontres dans 4 phrases pour une spécificité >=+10. A partir de ce premier résultat on précise l'exploration contextuelle en disséquant les contextes où apparaissent ensemble ces deux formes. A l'étape suivante du calcul on mesure une co-occurrence avec les formes cauchemar et nazisme dans 2 phrases. En répétant ainsi le processus de comptage jusqu'à épuisement on détermine des chemins de co-occurrence correspondant à des "squelettes" de phrases avérées dans le corpus.

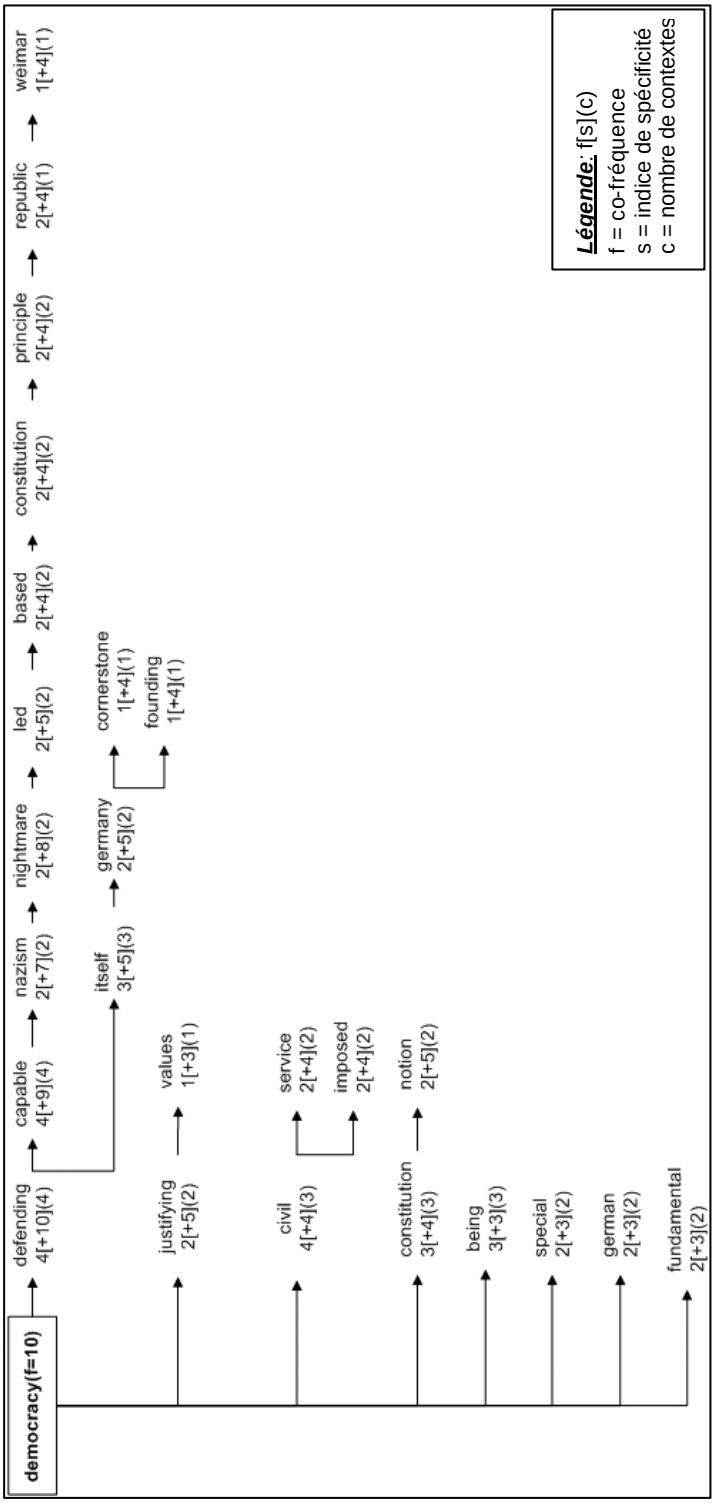

Figure 8b: Vue partielle du réseau de co-occurrences élaboré à partir du pôle *democracy*

$_en l'espèce, l'**obligation** faite aux fonctionnaires **allemands** de professer et de **défendre** activement et constamment le régime fondamental libéral et démocratique au sens de la loi fondamentale repose sur l'idée que la fonction publique est le garant de la constitution et de la ***démocratie*** .	$_in this case the obligation **imposed** on **german civil** servants to bear witness to and actively uphold at all times the free democratic constitutional system within the meaning of the basic law is founded on the **notion** that the **civil service** is the guarantor of the **constitution** and ***democracy*** .
$_elle revêt une importance **particulière** en allemagne en raison de l'expérience que celle-ci a **connue** sous la **république** de **weimar** et qui, lorsque la **république** fédérale a été constituée après le **cauchemar** du **nazisme**, a conduit à la **volonté** d'**instaurer** une " ***démocratie*** **apte** à se **défendre** ".	$_this **notion** has a special importance in **germany** because of that country's experience under the **weimar republic**, which, when the federal **republic** was founded after the **nightmare** of **nazism**, **led** to its **constitution being based** on the **principle** of a "***democracy*** **capable** of **defending** itself " .
$_ceux-ci seraient en effet le **pilier** d'une " ***démocratie*** **apte** à se **défendre** " :	$_the **civil service** was the **cornerstone** of a "***democracy*** **capable** of **defending itself** " .
$_même si aucun reproche ne lui avait été fait dans l'exercice de ses fonctions en soi, elle avait néanmoins, en tant qu'enseignante, une responsabilité **particulière** dans la transmission des **valeurs fondamentales** de la ***démocratie***.	$_even though no criticism had been levelled at the way she actually performed her duties, she had had, nevertheless, as a teacher, a **special** responsibility in the transmission of the **fundamental values** of ***democracy*** .
$_elle aurait la ferme conviction de pouvoir servir au **mieux** la cause de la ***démocratie*** et des droits de l'homme par son **engagement** au sein du dkp ;	$_she was firmly convinced that she could best serve the cause of ***democracy*** and human rights by her political activities on behalf of the dkp ;
$_l'allemagne souhaitait éviter la répétition de ces expériences en **fondant** son nouvel etat sur l'idée de " ***démocratie*** **apte** à se **défendre** " .	$_**germany** wished to avoid a repetition of those experiences by **founding** its new state on the idea that it should be a "***democracy*** **capable** of **defending itself** " .
$_elle a aussi relevé que " le **cauchemar** du **nazisme** " l'a **conduite** à " la **volonté** d'**instaurer** une " ***démocratie*** **apte** à se **défendre** " .	$_it also noted that 'the **nightmare** of **nazism**" **led** to its **constitution being based** on the **principle** of “a ***democracy*** **capable** of **defending itself**".
$_j'ajoute que ce principe constitutionnel représentait aussi à l'époque à considérer pour la présente affaire un **objectif** légitime **justifiant** l'**obligation**, imposée à tous les fonctionnaires, de loyauté envers les **valeurs** de la ***démocratie*** et la prééminence du droit .	$_may i add that this constitutional **principle** also represented at the time material for the present case a legitimate aim **justifying** the duty **imposed** on **civil servants** of loyalty to the **values** of ***democracy*** and the rule of law .
$_je suis donc parvenu à la conclusion que, sur cet aspect de l'affaire, les autorités et juges **allemands** étaient **mieux** placés pour apprécier si l'ingérence était nécessaire à la défense de la ***démocratie***, l'une des principales raisons **justifiant** les restrictions dans l'intérêt de la sécurité nationale, et qu'il faut donc leur laisser dans le cadre de leur marge d'appréciation un pouvoir discrétionnaire plus large que celui reconnu par la majorité .	$_therefore, i came to the conclusion that the **german** authorities and judges in this respect of the case were in a better position to assess whether the interference was necessary in defence of ***democracy***, that **being** one of the main reasons **justifying** restrictions in the interests of national security, and should therefore be given a wider discretion within their margin of appreciation than that recognised by the majority .

Figure 9 : Les contextes spécifiques où se réalisent les réseaux de co-occurrences calculés à partir des pôles *démocratie* - *democracy*

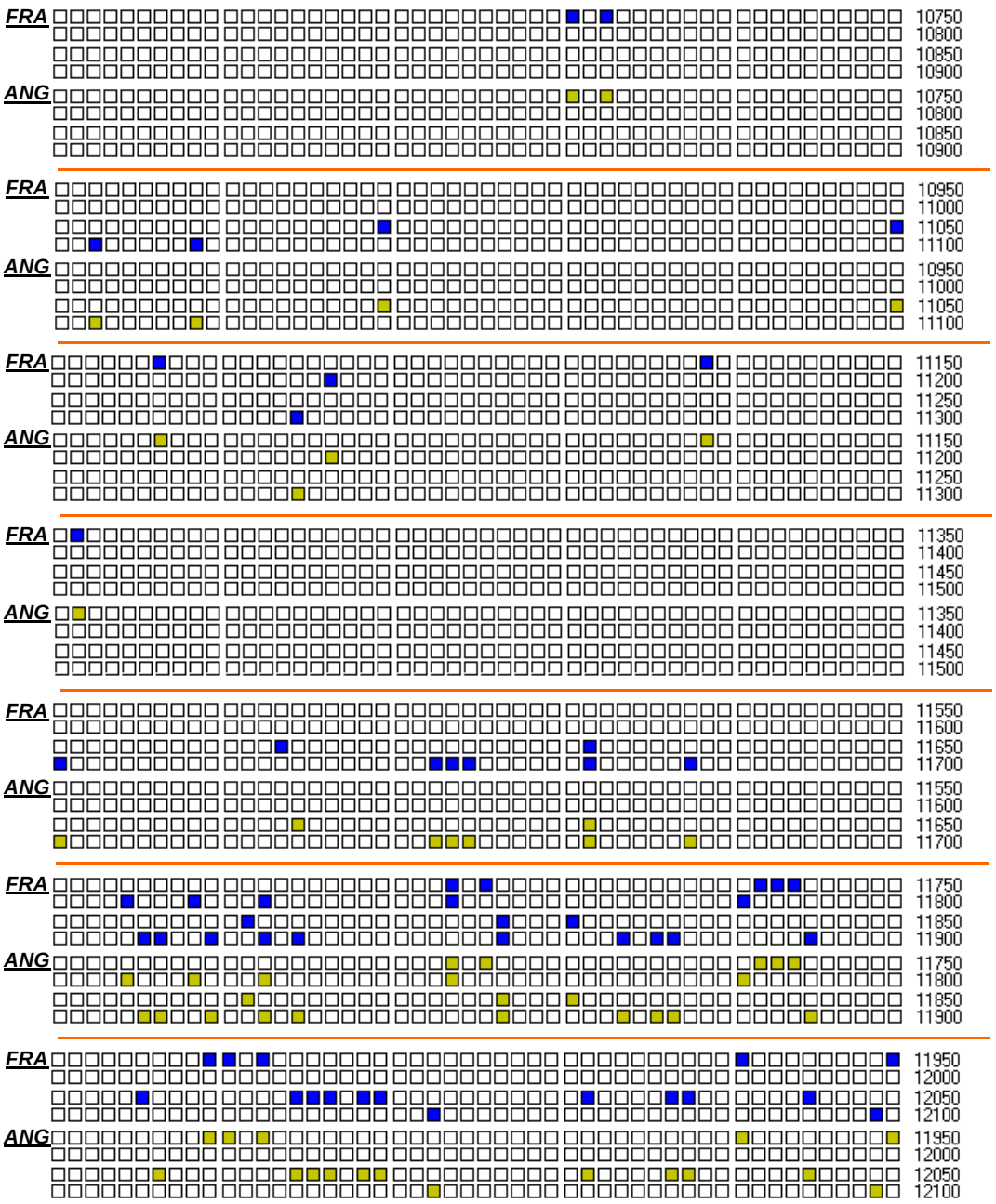

Figure 10 : Les occurrences des types bilingues démocrat+ et democra+ dans un extrait du corpus *Convention*

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

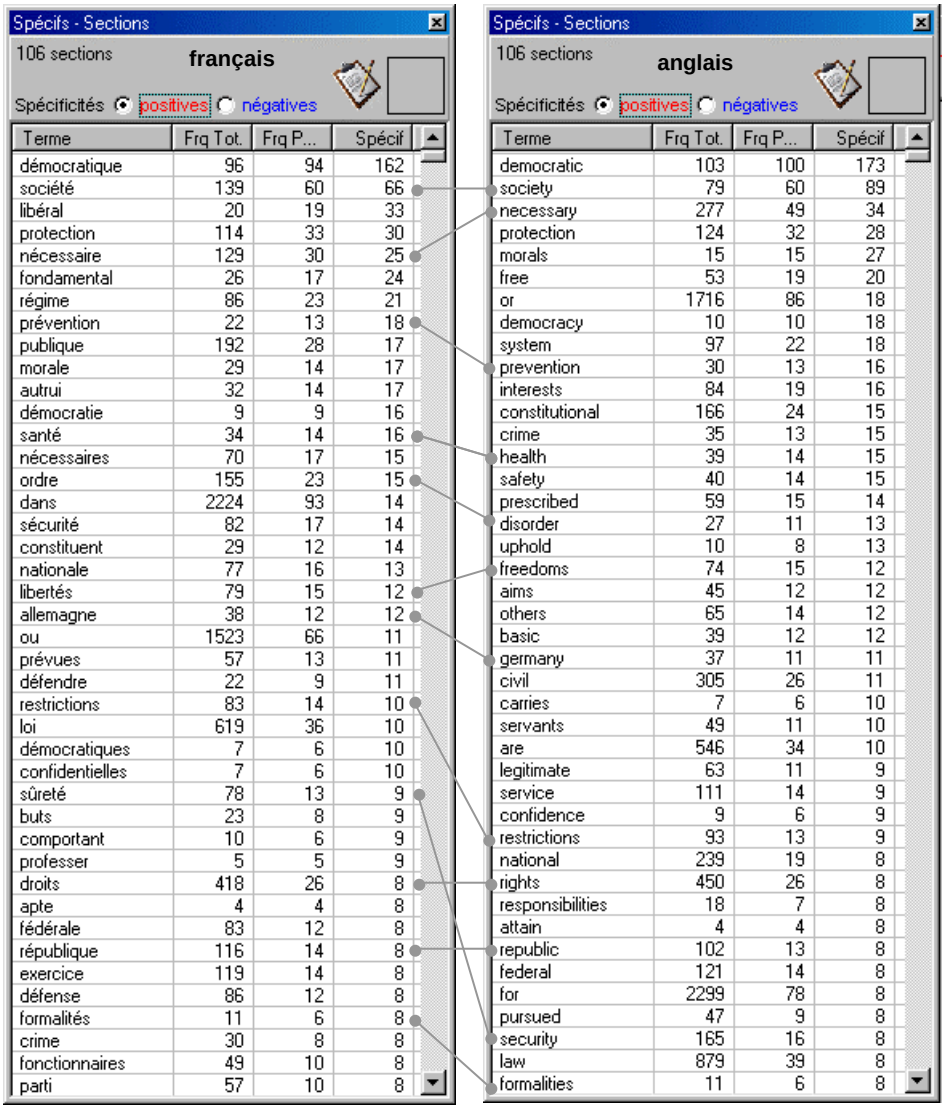

Figure 11 : Vocabulaire caractéristique des phrases contenant le pôle *démocrat+/democra+*

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.

DESGRAUPES Bernard (2001). *Introduction aux expressions régulières*. Paris : Vuibert Informatique, 272 p.

GALE William, CHURCH Kenneth (1991). "A Program for Aligning Sentences in Bilingual Corpora." In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkley, 1991, pp. 177-184.

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.