Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues

William Martinez, Maria Zimina

CLA2T - Université de la Sorbonne nouvelle Paris 3 – France – martinez@msh-paris.fr, zimina@msh-paris.fr

Abstract

The focus of bilingual text alignment is automatic reconstruction of translation correspondences between the segments of source text and the segments of its translation. Progress in alignment technology would allow for new translation support tools based on existing translation resources. Quite a few algorithms have already been proposed for aligning sentences of bilingual texts. The challenge is now to discover the links between finer segments on the word level. Using lexicometrics in bilingual text analysis facilitates the discovery of translation correspondences on the basis of word frequency distributions. The purpose of this article is to demonstrate how the statistical study of the intensity of lexical relations through collocation may allow an identification of translation equivalence within the lexical environments of a couple of words representing mutual translations.

Résumé

L'alignement de corpus bilingues vise la constitution automatique de correspondances de traduction entre le maximum d'éléments de textes source et cible afin de rendre possible la réalisation d'outils d'extraction et la réutilisation des ressources de traductions existantes. La recherche dans l'alignement a déjà abouti au développement de méthodes permettant d'apparier les phrases. L'appariement des mots et des syntagmes demeure une opération complexe. L'approche lexicométrique du matériau bi-textuel fournit des informations fréquentielles susceptibles de laisser apparaître les correspondances de traduction au niveau des mots et des syntagmes. L'article montre comment une application particulière de la méthode des cooccurrences permet de repérer des similitudes traductionnelles parmi les combinaisons lexicales qui s'opèrent dans les deux textes autour d'un couple de formes - traductions mutuelles.

Mots-clés : corpus bilingues, alignement, correspondances de traduction, lexicométrie, cooccurrences, réseaux de cooccurrences, polycooccurrences.

1. Introduction

L'appariement des mots constitue actuellement un problème fondamental dans le domaine de l'alignement des textes bilingues. En effet, l'identification des relations de correspondance entre les mots de phrases représentant des traductions mutuelles est une opération complexe même lorsqu'il s'agit de l'appariement manuel. Dans le domaine lexicométrique, les approches cooccurrentielles permettent de dévoiler la dimension collocative du langage, fondée sur une mesure statistique des attirances et des distributions que les couples de mots manifestent dans un corpus de texte. Les collocations relevées dans les textes bilingues sont ensuite exploitées pour construire des équivalences traductionnelles au niveau des mots et des syntagmes.

2. Le domaine de l'alignement

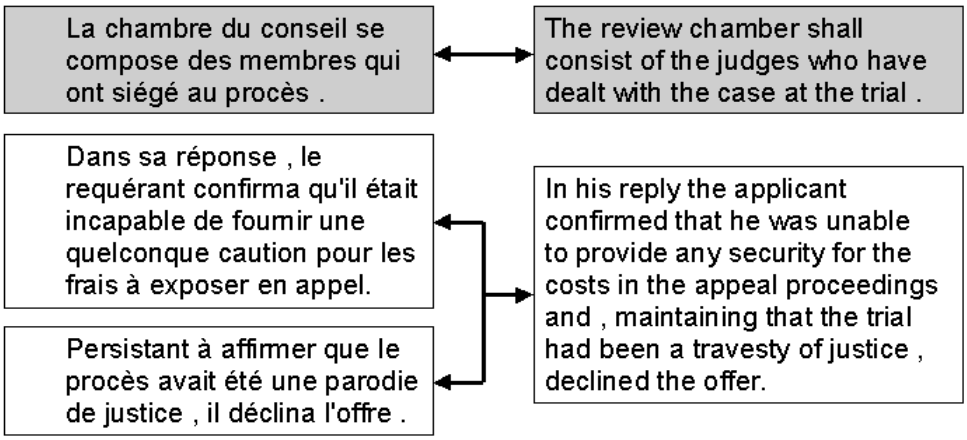


*Figure 1 : Corpus **Convention** : exemple d'alignement des phrases*

Aligner des corpus consiste à mettre en relation des unités textuelles qui se correspondent (cf. figures 1-2). Les unités en correspondance se positionnent à plusieurs niveaux : mots, syntagmes, phrases, paragraphes etc.. Contrairement à ce qui se passe dans le domaine de la traduction automatique (TA), l'alignement ne vise pas à produire des traductions d'une manière automatisée. L'objectif de l'alignement est de mettre en valeur le potentiel des traductions existantes pour en extraire des solutions aux problèmes de traduction.

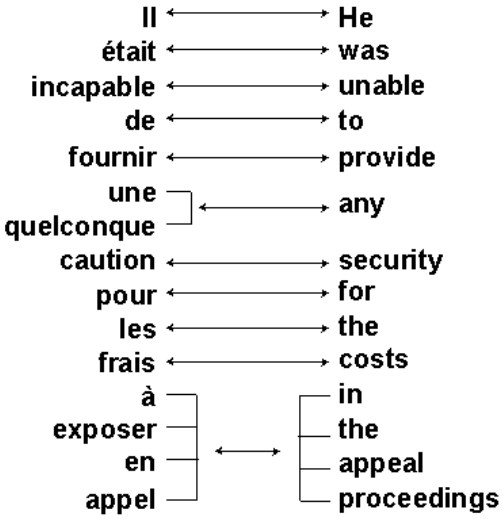


*Figure 2 : Corpus **Convention** : exemple d'alignement des mots et des syntagmes*

L'alignement des données de traduction facilite le travail des terminologues, lexicographes, traducteurs etc… en mettant à leur disposition une sorte de mémoire de traduction, permettant la découverte et l'exploitation d'usages et d'expressions ne figurant pas encore dans les dictionnaires (Isabelle et Warwick-Armstrong, 1993).

Pour rendre la richesse des traductions existantes facilement exploitable, il faut construire des procédures permettant de retrouver automatiquement des correspondances entre le maximum de segments du texte source et leurs équivalents dans le texte traduit. 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..

Les comptes-rendus d'expériences publiés récemment dans le domaine de l'alignement de corpus décrivent un certain nombre d'algorithmes permettant d'apparier les phrases avec un taux de réussite élevé (Brown et al., 1991 ; Gale et Church, 1991 ; Kay et Röscheisen, 1993 ; Simard et al.)[1]. En revanche, l'appariement des mots à partir de phrases bilingues demeure un problème fondamental dans le domaine de l'alignement (Debili et Zribi, 1996).

Le niveau des mots et des syntagmes est le plus apte à servir de base au découpage des corpus bi-textuels en *unités minimales d'équivalence élémentaire*.[2] L'identification de ces unités permettrait d'augmenter considérablement les performances des systèmes d'aide à base de corpus, tels que vérificateurs et extracteurs de terminologie, générateurs semi-automatique des dictionnaires, détecteurs des erreurs de traduction (omission, faux amis etc.), ainsi que celles des nombreuses applications de recherche documentaire. La recherche automatique des unités d'équivalence élémentaire présente une double difficulté car il faut tenir compte à la fois de la structure des unités de chacun des textes et des liens de correspondances qui existent entre eux. En effet, dans les phrases en correspondance de traduction, les mots entretiennent souvent des relations d'équivalence très complexes (cf. figure 3).

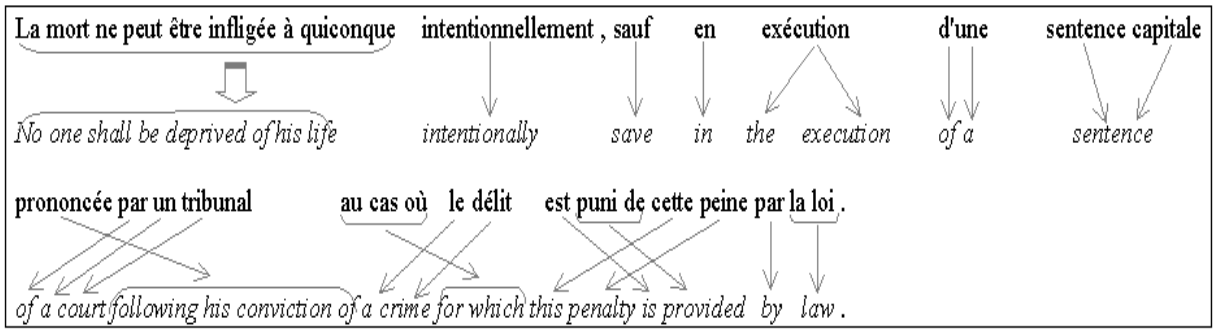


Figure 3 : Relations d'équivalence entre les mots dans deux phrases en relation de traduction

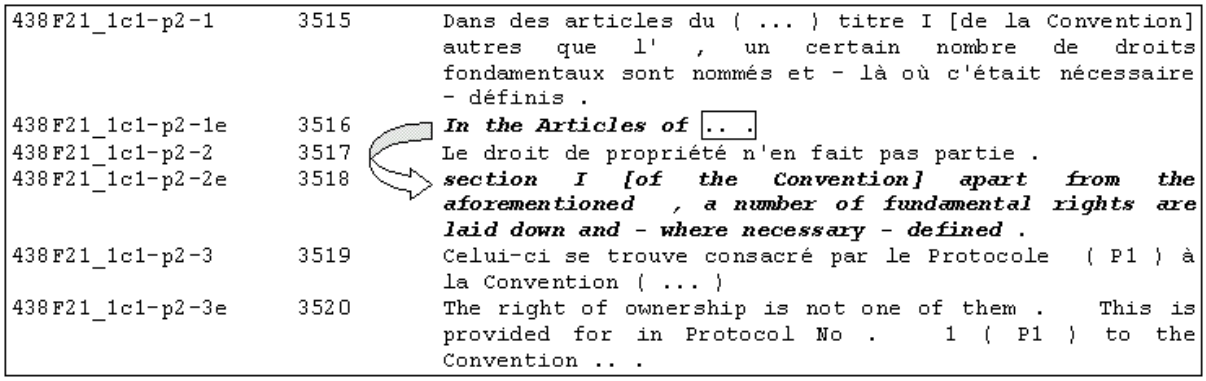
3. Description du Corpus de travail *Convention*

Nous proposons d'aborder le problème du repérage automatique des équivalences traductionnelles au niveau des mots et des syntagmes en nous appuyant sur l'analyse lexicométrique de textes bilingues. Notre corpus de travail est constitué des textes officiels de

[1] L'alignement à base de méthodes probabilistes utilise des critères formels tels que la similitude des longueurs de chaînes à apparier, les mots apparentés (*cognates*) ou l'ordre des mots dans les phrases. Certaines méthodes prévoient également le recours à un dictionnaire bilingue pour identifier les correspondances à base d'acceptions dictionnairiques. Des scores statistiques sont ensuite attribués en fonction de l'ensemble de ces critères pour effectuer l'appariement le plus probable.

[2] Le choix des unités de décompte a été largement étudié dans le cadre de l'analyse statistique de la chaîne textuelle monolingue (cf. Lebart et Salem, 1994). Des principes de segmentation en unités minimales ont été suggérés pour définir une norme permettant d'isoler de la chaîne textuelle les différentes unités que l'on peut étudier du point de vue de la méthode statistique. Dans le domaine de l'alignement, le principe de compositionnalité de traduction est largement utilisé pour la formulation de correspondances hiérarchisées entre deux textes (cf. Isabelle et Warwick-Armstrong, 1993). Selon ce principe, la traduction d'une unité est généralement fonction de la traduction des parties de cette unité. Il nous semble utile de confronter l'expérience du traitement lexicométrique d'ensembles textuels avec les connaissances acquises dans le domaine de l'alignement.

la *Convention de sauvegarde des droits de l'homme et des libertés fondamentales*[3], ainsi que des protocoles et des arrêts rendus par la Cour européenne des droits de l'homme de Strasbourg en 1995 (cf. figure 4).[4]



438F21_1c1-p2-1	3515	Dans des articles du (...) titre I [de la Convention] autres que l' , un certain nombre de droits fondamentaux sont nommés et - là où c'était nécessaire - définis .
438F21_1c1-p2-1e	3516	***In the Articles of .. .***
438F21_1c1-p2-2	3517	Le droit de propriété n'en fait pas partie .
438F21_1c1-p2-2e	3518	***section I [of the Convention] apart from the aforementioned , a number of fundamental rights are laid down and - where necessary - defined .***
438F21_1c1-p2-3	3519	Celui-ci se trouve consacré par le Protocole (P1) à la Convention (...)
438F21_1c1-p2-3e	3520	The right of ownership is not one of them . This is provided for in Protocol No . 1 (P1) to the Convention .. .

Figure 4 : Format initial du corpus de travail ***Convention*** *avec exemples d'erreurs recensées*

4. Contributions lexicométriques à l'alignement de corpus

Lorsque l'on compare les dictionnaires des formes graphiques constitués à partir des deux volets du corpus, certaines proximités se dessinent. Par exemple, l'analyse des 100 formes les plus fréquentes du corpus *Convention*, laisse apparaître certaines correspondances de traduction. La confrontation des dictionnaires numérisés triés par ordre décroissant des fréquences permet de repérer des proximités dans les rangs lexicaux des formes en correspondance (cf. figure 5). Le retour au contexte montre que la plupart des *paires de traductions* repérées à base de comparaison des fréquences générales correspondent à des mots dont le champ sémantique est étroitement défini à la fois dans les deux volets du corpus bilingue. Par exemple, le terme *convention* en anglais et en français réfère toujours à la *Convention de sauvegarde des Droits de l'Homme et des libertés fondamentales*, de même pour les termes *commission/commission*, *gouvernement/government*, *paragraphe/paragraph, droits/rights*.

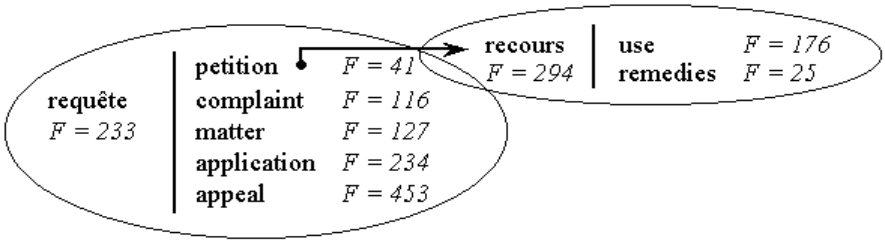
[3] Elaborée au sein du Conseil de l'Europe, la *Convention* définit un certain nombre de droits fondamentaux et institue un mécanisme de contrôle et de sanction propre à assurer le respect de ces droits par les Etats signataires. Il existe parallèlement deux versions officielles des documents mentionnés ci-dessus : l'une en français, l'autre en anglais, et il est impossible distinguer une langue source et une langue cible. La forte structuration des corpus anglais (273 685 occurrences) et français (285 961 occurrences) a permis de ne pas faire appel aux techniques statistiques d'alignement de phrases. En effet, le découpage en sections, très répandu dans les textes de droit, rend possible une approche formelle de l'alignement à base de balisage et segmentation parallèle de deux textes (Bourigault et al., 1999). Au cours du projet "Lexique des Droits de l'Homme" cette structuration a été transformée en une structuration logique manipulable par l'ordinateur : chaque couple de phrases équivalentes a reçu le même identifiant. Cette approche a généré toutefois un certain nombre d'erreurs dans l'alignement de phrases. Il s'agit notamment des phrases incluant une marque de ponctuation forte au milieu (point, deux points etc.). Dans ces cas, les fins de ces phrases n'ont pas été repérées correctement, ce qui a provoqué un décalage dans l'appariement des sections.

[4] Les auteurs aimeraient exprimer leur reconnaissance à Benoît Habert et Didier Bourigault pour la gentillesse avec laquelle ils ont accepté de mettre à notre disposition le corpus *Convention* sur lequel ont porté les expérimentations présentées.

français	rang lexical	fréquence générale	anglais	rang lexical	fréquence générale
convention	32	1 223	convention	26	1 228
commission	39	889	commission	36	832
paragraphe	44	740	paragraph	38	802
gouvernement	45	721	government	44	740
décision	59	508	decision	55	563
droits	66	418	rights	65	444
mais	80	380	but	79	392
article	82	362	article	87	351
citation	85	352	citation	88	351
détention	98	314	detention	96	336

Figure 5 : Extraits des dictionnaires des formes graphiques issus des deux volets du corpus

On observe dans le corpus une correspondance biunivoque entre ces mots qui induit sur le plan quantitatif par une ressemblance de leurs fréquences générales.[5] Naturellement, ce type de correspondance de traduction ne couvre qu'une petite partie des mots du corpus. Dans la plupart des cas, les relations de correspondance observées ne sont pas univoques ; un mot polysémique possède plusieurs correspondants qui dépendent de l'entourage contextuel dans lequel il se retrouve. Le terme français *requête* reçoit ainsi plusieurs traductions :



Le retour au contexte montre que le terme anglais *petition* (l'un des correspondants de *requête*) est polysémique lui-même :

conv_a201.2_p2-1 157 Le membre de la Commission élu au titre de la Haute Partie contractante contre laquelle une requête a été introduite a le droit de faire partie de la chambre saisie de cette □□□□□□□□

435D86_1-p2-2 5347 Avec pour corollaire que l'Etat pourrait limiter l'acceptation du droit de □□□□□□□ individuel à son territoire national , comme il l'a fait en l'occurrence .

conv_a201.2_p2-1e 158 The member of the Commission elected in respect of a High Contracting Party against which a petition has been lodged shall have the right to sit on a Chamber to which the □□□□□□□□□ has been referred .

435D86_1-p2-2e 5348 As a corollary , the State can limit acceptance of the right of individual □□□□□□□□ to its national territory - as has been done in the instant case.

De même pour le terme français *recours* :

[5] Les correspondances biunivoques résultent de situations diverses. Il peut s'agir de noms propres ou de mots dont les acceptions dictionnairiques se recouvrent dans les deux langues (*animal/animal*). Des correspondances biunivoques se forment également entre termes dont l'équivalence est imposée par les normes terminologiques (*privilèges et immunités/privileges and immunities*). Finalement, le contexte est parfois susceptible de restreindre le champ sémantique des mots polysémiques en les mettant en correspondance biunivoque (*convention/convention*).

conv_a2.2_p1-1 21 La mort n'est pas considérée comme infligée en violation de cet article dans les cas où elle résulterait d'un □□□□□□□ à la force rendu absolument nécessaire :	conv_a2.2_p1-1e 22 Deprivation of life shall not be regarded as inflicted in contravention of this article when it results from the □□□ of force which is no more than absolutely necessary :
conv_a26_p1-1 213 La Commission ne peut être saisie qu'après l'épuisement des voies de □□□□□□□ internes ...	conv_a26_p1-1e 214 The Commission may only deal with the matter after all domestic □□□□□□□□ have been exhausted ...

Dans le cas de ce type de correspondances, la comparaison des fréquences globales n'est pas une bonne indication pour l'appariement. Pour une étude complète des régularités dans la structuration des deux discours il est nécessaire d'exploiter des méthodes plus complexes qui vont au-delà de la simple comparaison forme à forme. Les approches hybrides qui allient les méthodes d'analyse de données multidimensionnelles et la détection des polyformes peuvent contribuer considérablement à l'appariement. Nous avons montré que l'utilisation des méthodes de *classification automatique* permet de découvrir des régularités dans les répartitions lexicales des *formes* et des *segments répétés* représentant des traductions mutuelles (Zimina, 2000).

L'utilisation des segments répétés dans l'analyse lexicométrique des textes source et cible nous rapproche encore de la notion d'unité de traduction. Les segments répétés permettent de repérer des collocations connexes ayant des distributions similaires dans les deux volets du corpus. Cependant, ces collocations ne couvrent qu'une partie des associations lexicales que l'on observe dans les deux textes bilingues. Les segments sont obtenus en opérant des coupures en fonction de critères formels qui ne permettent pas de tenir compte des notions de syntagme ou expression. De plus, l'ordre des mots dans les phrases en correspondance de traduction ainsi que leurs statuts syntaxiques influent considérablement sur les propriétés des segments répétés dans les textes sources et cibles. Par conséquent, on ne peut pas poser systématiquement l'équivalence de traduction entre deux unités de ce type. Afin d'élargir le champ d'investigation au-delà des limites de l'ordre syntaxique, nous allons maintenant exploiter la *méthode des cooccurrences*.

5. L'exploitation des cooccurrences pour l'alignement

La comparaison des fréquences locales permet de vérifier l'évolution parallèle des formes *droits* et *rights* dans les deux volets du corpus (cf. figure 6). Cependant, les régularités dans les profils de ventilation de ces formes n'apportent qu'une indication superficielle des correspondances inter-corpus. Les formes *droits* et *rights* évoluent en parallèle mais qu'en est-il de leur environnement lexical ? Pour un repérage exhaustif des correspondances de traduction au niveau des mots et des syntagmes nous allons recourir au calcul des cooccurrences. Leur exploitation dans l'alignement textuel permet d'aller au-delà des associations contiguës des formes et de prendre en considération toutes les formes associées sur l'axe syntagmatique. Fondé sur le modèle hypergéométrique[6], la méthode des

[6] L'exploitation du modèle hypergéométrique pour le calcul des cooccurrences spécifiques s'inspire de la méthode élaborée par Pierre Lafon [1984]. Une comparaison s'effectue 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 *±Exx* où le signe indique un sur-emploi ou un sous-emploi de la forme et la valeur indique son degré de spécificité.

cooccurrences mesure les attractions lexicales les plus intenses autour d'un pôle donné et livre les résultats sous la forme d'une liste hiérarchisée. Le tableau 6 montre une partie des cooccurrences spécifiques calculées pour chacun des pôles bilingues *droits* et *rights*.

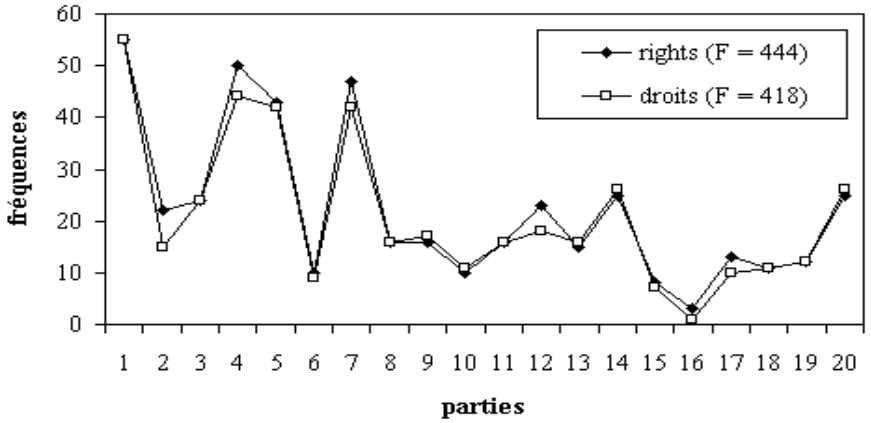


Figure 6 : Profils de répartition similaires des formes DROITS et RIGHTS

Tableau 6 : Principaux cooccurrents binaires des formes DROITS et RIGHTS dans l'unité contextuelle de la phrase jusqu'au seuil +E15

cooccurrent FRA	Fréq	co-fréq	spécificité	cooccurrent ANG	Fréq	co-fréq	spécificité
des	4 122	522	+E51	human	202	192	+E51
homme	222	192	+E51	european	173	86	+E51
européenne	116	75	+E51	protection	124	74	+E51
strasbourg	87	62	+E51	strasbourg	87	62	+E51
libertés	79	69	+E51	freedoms	74	66	+E51
palais	62	61	+E51	building	64	61	+E51
protection	114	54	+E37	fundamental	63	39	+E33
ouvrent	35	32	+E37	convention	1 228	168	+E27
sauvegarde	31	29	+E34	parental	33	26	+E27
fondamentales	33	27	+E29	public	423	87	+E26
déroulés	30	26	+E29	delivered	55	32	+E26
déférée	49	32	+E28	done	79	36	+E24
convention	1 223	168	+E27	month	74	33	+E22
prononcé	54	31	+E25	english	65	31	+E22
publique	192	54	+E23	referred	205	50	+E19
puis	91	37	+E23	laid	122	37	+E18
anglais	61	31	+E23	article	351	65	+E17
décidé	63	31	+E22	morals	15	14	+E17
garantis	19	18	+E22	took	92	31	+E16
parentaux	19	18	+E22	hearing	406	67	+E15
article	362	67	+E18	place	135	36	+E15
civil	90	32	+E18				
français	110	33	+E16				

***Guide de lecture du tableau*:** Le calcul des cooccurrences spécifiques rend compte des attractions lexicales les plus intenses autour d'un pôle dans une fenêtre d'exploration contextuelle donnée (ici, la phrase). A partir des fréquences globales des formes étudiées et de leurs co-fréquences, c'est-à-dire du nombre de rencontres, on calcule la spécificité de la cooccurrence. En comparant la liste de leurs principaux cooccurrents on constate que les univers lexicaux des deux pôles sont très semblables.

Les ressemblances entre les deux listes qui s'étendent jusqu'aux co-fréquences et aux indices de spécificité facilitent le repérage de couples de formes graphiques correspondant aux traductions mutuelles (ex. : *homme*/*human*, 192 occ., +E51). Ces équivalences donnent un premier aperçu des grands traits structuraux qui maintiennent l'équivalence traductionnelle au niveau lexical. Cette information permet d'apparier non plus des pôles isolés mais deux ensembles de cooccurrents, c'est-à-dire deux univers lexicaux. Cependant l'exploitation des résultats reste limitée dans la mesure où le calcul des cooccurrences binaires ne fournit un indice précis que pour l'association individuelle d'un pôle avec un cooccurrent.

6. Les réseaux de cooccurrences

Pour saisir la complexité des attractions lexicales sur l'axe syntagmatique, il est nécessaire de reconsidérer le phénomène de la cooccurrence au-delà des associations forme à forme pour s'intéresser aux liens simultanés qu'entretient un pôle avec l'ensemble de ses cooccurrents. Nous avons donc défini une approche qui mesure la cooccurrence en tant phénomène global en considérant les formes cooccurrentes comme parties intégrantes d'un système dont les différents composants entretiennent une relation de dépendance simultanée.[7] Ainsi, notre méthode[8] explore les contextes spécifiques du pôle et révèle à chaque étape du calcul un élément supplémentaire du réseau de cooccurrences qui s'élabore à partir de celui-ci. L'application de cette méthode parallèlement aux deux volets du corpus bilingue révèle une correspondance presque exacte entre les systèmes cooccurrentiels des pôles *droits* et *rights* (cf. tableau 7).

Tableau 7 : Réseaux de cooccurrences spécifiques (extrait)

[illegible]	[illegible]
[illegible]	[illegible]

[7] L'analyse des cooccurrents révèle la nature sémantico-syntaxique de leur relation avec le pôle dans les deux volets du corpus. On remarque que les attractions lexicales simultanées des formes permettent leur désambiguïsation. Ainsi, le mot *case* en anglais est fortement polysémique mais dans le sous-ensemble des phrases avec des attestations des formes *droits/rights* il est toujours employé dans le sens *affaire (a question to be decided in a court of law,* Longman Dictionary of Contemporary English, 1987, Second Edition). Contrairement aux fréquences générales dans le corpus, les co-fréquences des formes *case/affaire* sont identiques. On peut donc procéder à l'appariement de ces formes :

	Fréq gén (F)	*co-fréq(f)*
case	209	31
affaire	93	31

[8] Elaboré autour du logiciel Lexico d'A. Salem [1994] notre module des réseaux des cooccurrences est fondé sur la réitération du calcul :
Etape 1 : On calcule pour le pôle A les cooccurrents spécifiques B, C et D
Etape 2 : Dans leurs contextes communs, on calcule pour les pôles A+B les cooccurrents spécifiques E et F
Etape 3 : Les pôles A+B+E ont pour cooccurrent spécifique H
Etape 4 : Les pôles A+B+E+H n'ont pas de cooccurrent spécifique et l'exploration s'interrompt pour ce chemin
Etape 5 : Les pôles A+B+F ont pour cooccurrents spécifiques I, etc.
Durant l'exploration, différents filtrages conditionnent l'épuisement des explorations contextuelles et réduisent le bruit dans les résultats pour privilégier l'information la plus spécifique : seuils maximaux de fréquence et de spécificité du cooccurrent, nombre minimal de contextes où se produit la cooccurrence 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).

Pour faciliter la visualisation des réseaux, les résultats du calcul sont présentés sous la forme d'arborescences (cf. figures 8a et 8b). Les indications statistiques fournies par l'exploration contextuelle permettent d'affirmer que, même si l'ordre et le nombre de leurs éléments diffèrent, ces réseaux tendent à former des ensembles équivalents. A partir de ces informations précises il devient possible de retourner au corpus pour extraire dans chaque volet les 9 phrases spécifiques où se réalisent ces réseaux de cooccurrences (cf. tableau 9).

7. Conclusions

A travers différentes expériences menées dans le domaine des cooccurrences, nous avons mis au point des approches qui livrent des informations cooccurrentielles diverses : mesures des attractions lexicales autour d'un pôle donné, indices de la nature sémantico-syntaxique de ces relations et représentations du fonctionnement de ces formes en système. L'exploitation de ces informations dans le cadre de l'alignement de textes permet de régler certains problèmes d'ambiguïté lexicale que l'on rencontre lors de l'appariement de formes en correspondance de traduction. Ainsi, en étudiant les caractéristiques collocationnelles de deux pôles parallèlement dans chacun des volets d'un corpus bilingue, on compare leurs univers lexicaux pour y chercher des similitudes indiquant leur liaison. Mais c'est sous sa forme la plus complexe, celle des réseaux de formes, que l'information cooccurrentielle se révèle la plus utile pour l'alignement. A partir de ces données, livrées sous forme de 'structures de phrases', on peut effectuer un rapprochement automatique des groupes de mots équivalents qui aident à cerner les pôles en correspondance. Nous avons montré enfin comment aligner des contextes à partir de mots isolés. Parmi les perspectives qu'offre cette méthode, on citera la possibilité d'un alignement automatisé basé sur l'application systématique de ce processus, éventuellement appuyée sur l'utilisation d'un dictionnaire ou d'un lexique bilingue, à l'ensemble des pôles d'un même corpus.

Références

Bourigault D., Chodkiewicz C. and Humbley J. (1999). Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné. In *TIA'99*, Nantes.

Brown P., Lai J. C. and Mercer R. L. (1991). Aligning sentences in parallel corpora. In *Proc of the 29th ACL*, Berkeley, 169-174.

Debili F. and Zribi A. (1996). Les dépendances syntaxiques au service de l'appariement des mots. In *RFIA'96*, 81-90.

Gale W. A. and Church K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, vol.19(3), 75-102.

Isabelle P. and Warwick-Armstrong S. (1993). Les corpus bilingues : une nouvelle ressource pour le traducteur. In Bouillon, P. and Clas, A. editors, *La Traductique: Études et Recherches de traduction par ordinateur.* Montréal : Les Presses de l'Université de Montréal.

Kay M. and Röscheisen M. (1993). Text-translation alignment. *Computational Linguistics*, vol.19(3), 121-142.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod.

Simard M., Foster G. and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proc of the Fourth TMI*, Montréal, Canada, 67-81.

Zimina M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique. In *JADT'2000*, Lausanne, 171-178.

Tableau 9 : Mise en évidence de contextes spécifiques

Corpus français	Corpus anglais
1 - Ratio : 0.41 Phrase n° 1452 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # [9] de la **convention**.	1 - Ratio : 0.48 Phrase n° 1464 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
2 - Ratio : 0.41 Phrase n° 3116 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	2 - Ratio : 0.48 Phrase n° 3189 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
3 - Ratio : 0.41 Phrase n° 6300 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	3 - Ratio : 0.48 Phrase n° 6453 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
4 - Ratio : 0.41 Phrase n° 6510 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	4 - Ratio : 0.48 Phrase n° 6672 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
5 - Ratio : 0.41 Phrase n° 6685 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	5 - Ratio : 0.48 Phrase n° 6863 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
6 - Ratio : 0.41 Phrase n° 9349 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	6 - Ratio : 0.48 Phrase n° 9593 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
7 - Ratio : 0.40 Phrase n° 5678 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	7 - Ratio : 0.47 Phrase n° 5808 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**"), within the three-**month period laid down** by # and # of the **convention**.
8 - Ratio : 0.32 Phrase n° 5947 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**") puis par le gouvernement français ("le gouvernement"), les # et # respectivement, dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	8 - Ratio : 0.38 Phrase n° 6087 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**") and by the french government ("the government") # and # respectively, within the three-**month period laid down** by # and # of the **convention**.
9 - Ratio : 0.30 Phrase n° 2856 l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**") puis par le gouvernement de la confédération suisse ("le gouvernement") # et #, dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	9 - Ratio : 0.36 Phrase n° 2927 the **case** was **referred** to the court by the **european commission** of **human rights** ("the **commission**") # and # by the government of the swiss confederation ("the government") # and #, within the three-**month period laid down** by # and # of the **convention**.

Guide de lecture du tableau : A l'issue du calcul des réseaux de cooccurrences on extrait les contextes où se réalisent les structures spécifiques mises en évidence par le comptage. La présentation des deux listes en vis-à-vis laisse apparaître une symétrie complète entre les corpus français et anglais. En comparant le nombre de cooccurrents spécifiques et le nombre de mots dans l'unité contextuelle, le ratio mesure la densité de chaque contexte par rapport au nombre de formes spécifiques qui y figurent.

[9] Le # remplace un mot systématiquement absent dans le corpus qui nous a été fourni.

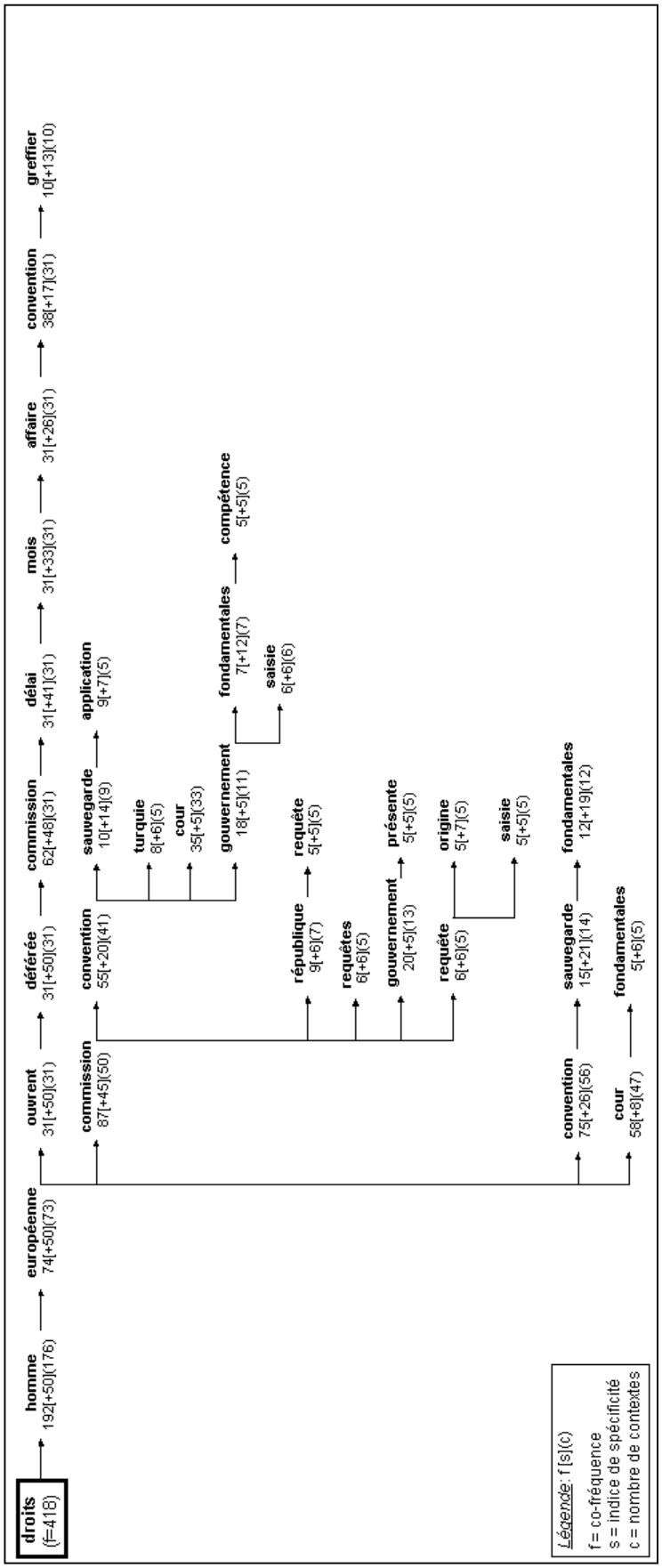


Figure 8a: Vue partielle du réseau de cooccurrences élaboré à partir du pôle DROITS

Guide de lecture de la figure **:** Basée sur la réitération du calcul des cooccurrences, 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 (à lire de la gauche vers la droite et du haut vers le bas) les résultats du calcul exploratoire pour la forme pôle *rights*. Sur la première ligne qui correspond à la branche la plus spécifique du réseau, est rapportée une forte cooccurrence du pôle avec la forme *human* : 192 rencontres dans 176 phrases pour une spécificité de +50. 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 forte cooccurrence avec *european* dans 81 phrases. En itérant ainsi le processus de comptage jusqu'à épuisement on détermine des chemins de cooccurrence correspondant à des 'squelettes' de phrases

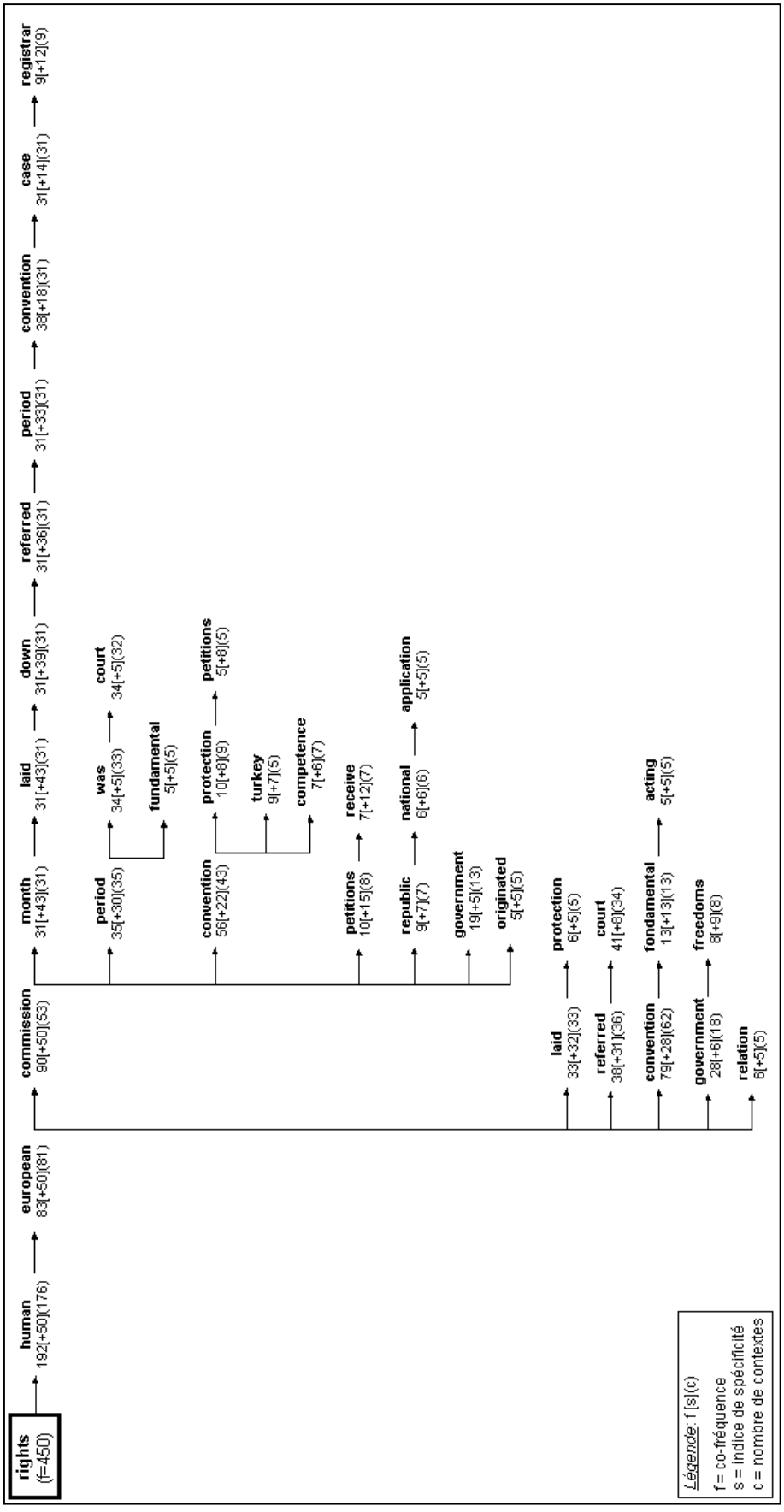


Figure 8b: Vue partielle du réseau de cooccurrences élaboré à partir du pôle RIGHTS