Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles

Maria Zimina-Poirot

UNIVERSITÉ PARIS 3 – SORBONNE NOUVELLE

ÉCOLE DOCTORALE : Langage et langues

N° attribué par la bibliothèque

|_|_|_|_|_|_|_|_|_|_|_|_|

Doctorat

Discipline : Sciences du langage

Maria ZIMINA-POIROT

Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles

Thèse dirigée par

André SALEM

Soutenue le vendredi 26 novembre 2004

Jury :

M. Eric GAUSSIER (Examinateur)

M. Benoît HABERT (Rapporteur)

M. André SALEM (Directeur)

Mme Monique SLODZIAN (Rapporteur)

Remerciements

Ce travail doit beaucoup au **Centre de Lexicométrie et d'Analyse Automatique des Textes** (CLA2T – SYLED) de l'Université de la Sorbonne nouvelle – Paris 3 au sein duquel il s'est développé. Ces remerciements s'adressent à tous les membres de cette équipe qui m'ont aidé dans mes recherches : André Salem, Serge Fleury, Cédric Lamalle, William Martinez.

Je remercie particulièrement André Salem pour l'appui qu'il a donné à ce projet, pour ses précieux conseils, ses critiques ainsi que ses relectures attentives.

Merci également à Benoît Habert et Didier Bourigault pour la gentillesse avec laquelle ils ont accepté à mettre à ma disposition le corpus parallèle de la *Convention des Droits de l'Homme* qui a servi de base pour mes explorations. Merci aussi à Serge Fleury qui m'a aidé à préparer la version étiquetée du corpus.

Je tiens à remercier tous ceux qui m'ont donné des conseils à l'occasion de rencontres informelles, de séminaires, de colloques, d'échanges par courrier électronique : Benoît Habert, Monique Slodzian, Eric Gaussier, Jean Véronis, Thierry Poibeau.

Merci enfin à mes relecteurs Serge Fleury et Thierry Poibeau pour leurs précieuses remarques qui m'ont permis d'améliorer la qualité du manuscrit.

A ces remerciements particuliers, je voudrais associer mon époux Jérôme, la famille et les amis qui m'ont aidé à progresser dans ce travail jour après jour. Leurs encouragements ont été essentiels dans l'aboutissement de cette thèse.

Table des matières

Introduction

Dans le contexte récent de l'informatisation de la société, chercheurs et praticiens sont confrontés à une croissance spectaculaire des corpus de textes multilingues provenant d'archives de textes traduits, de bases documentaires multilingues numérisées, de sites Web internationaux. Pour des raisons variées, diverses communautés s'intéressent dorénavant aux données textuelles multilingues. Les *historiens*, les *juristes*, les *philologues* analysent les corpus multilingues avec des outils d'exploration permettant d'observer plus finement les correspondances entre différents volets de corpus comparés. Les *informaticiens* utilisent les ressources obtenues à partir de ces corpus pour améliorer les performances des outils de traduction automatique ou des moteurs de recherche pour le Web. Enfin, les ressources traductionnelles obtenues à partir des corpus multilingues sont utilisées avec profit pour les études menées dans le cadre de plusieurs disciplines des *sciences du langage* qui s'étendent de la *linguistique contrastive* à la *lexicographie*, de la *traduction assistée par ordinateur* à l'*enseignement des langues étrangères*, de *l'analyse du discours* à la *linguistique computationnelle*.

L'extraction des ressources traductionnelles à partir de corpus de textes multilingues volumineux suppose le recours à des outils informatiques adaptés. Le développement de tels outils fait l'objet des recherches menées dans le domaine du *Traitement Automatique des Langues* (TAL) [1]. Pour analyser des corpus de textes rédigés dans des langues différentes et en extraire des ressources réutilisables, les systèmes actuels de TAL font souvent appel à des savoirs *a priori* (linguistiques, pragmatiques, etc.). Les explorations appuyées sur les

[1] Ce domaine de recherches interdisciplinaire a pour objet la création de programmes informatiques capables de *traiter* automatiquement les *langues naturelles*.

approches de statistique textuelle sont moins courantes. Dans ce contexte, nous avons pu confirmer notre intuition que les connaissances des propriétés fréquentielles du matériau textuel acquises dans le domaine de la *textométrie* [2] donnent accès à des savoir-faire précieux susceptibles d'enrichir les pratiques actuelles du traitement automatique de corpus multilingues.

Conçues pour mesurer des corrélations fréquentielles complexes entre les unités textuelles, les méthodes textométriques offrent une base précieuse pour mettre en évidence des structures lexicales bi-textuelles qui résultent d'équivalences traductionnelles dans les textes en correspondance. Dans le cas de *corpus parallèles* [3], cette approche nous a permis de procéder à l'extraction de ressources traductionnelles à partir de la distribution des unités textuelles au sein de chaque volet [Zimina, 2000 ; 2002 ; 2004].

***

La première partie de notre travail est consacrée à un exposé des enjeux du traitement automatique de corpus multilingues (chapitres 1-2).

Le chapitre 1 tente de cerner le concept de *parallélisme textuel* dans le contexte multilingue. Le lecteur y trouvera des exemples de corpus parallèles composés de textes sources et de leurs traductions (effectuées par des traducteurs humains) ou de textes dont chacun est une traduction de l'autre sans qu'il soit possible de déterminer lequel a servi de source.

Ce chapitre rappelle également les enjeux de l'analyse automatique des corpus parallèles ainsi que les acquis obtenus par les principaux courants de recherche du domaine du traitement automatique des langues. Il s'agit en particulier du secteur du traitement automatique des langues que l'on appelle l'*alignement automatique*. L'objectif de l'alignement est la mise en relation des unités

[2] La spécificité de la démarche textométrique réside dans le statut privilégié conféré aux données textuelles. Elle se caractérise par deux opérations principales qui visent à découper le texte en unités puis à comparer leurs distributions en les opposant les unes aux autres dans les différents contextes où ces unités apparaissent. Le domaine de la textométrie multilingue est présenté au chapitre 3.

[3] La notion de *corpus parallèle* est développée au chapitre 1.

textuelles (*mots*, *syntagmes*, *phrases*, *paragraphes*, etc.) qui se correspondent dans un *bi-texte*[4]. L'automatisation de l'alignement est indissociable d'une réflexion sur la segmentation de corpus de traduction en unités minimales. Les recherches en traductologie ont montré que cette question se pose de manière particulièrement complexe dans le contexte multilingue.

Dans la première partie du chapitre 2, nous recensons les problèmes nés dans le contexte de la segmentation de corpus parallèles en équivalences traductionnelles. Des exemples montrent la difficulté de déterminer des mécanismes formels permettant d'automatiser cette segmentation au niveau lexical. La deuxième partie décrit les principales méthodes d'alignement automatique de corpus. Nous y retraçons brièvement l'histoire du domaine de l'alignement, les acquis, les difficultés rencontrées ainsi que les tendances actuelles. Le lecteur y trouvera la description et la comparaison de quelques grandes familles d'algorithmes d'alignement automatique développés au cours des vingt dernières années.

La deuxième partie de ce travail (chapitres 3-6) présente des applications textométriques que nous avons mises au point pour l'extraction de ressources traductionnelles à partir de corpus parallèles.

Le chapitre 3 introduit les fondements de l'analyse textométrique des corpus multilingues. Les explorations présentées dans ce chapitre montrent que les propriétés textométriques du matériau textuel repérées dans les études quantitatives des corpus de textes monolingues fournissent des indices précieux pour la mise en évidence des correspondances traductionnelles du bi-texte.

Les chapitres 4-6 sont consacrés aux applications des différentes méthodes statistiques couramment utilisées en textométrie (*extraction de segments répétés*, *classification automatique*, *spécificités*, *cooccurrences multiples*, *topographie textuelle*, etc.) pour l'analyse de corpus parallèles. Pour chaque méthode exposée, nous présentons des exemples concrets d'utilisation dans le contexte multilingue

[4] Voir *Chapitre 2*.

accompagnés d'échantillons de ressources traductionnelles obtenues à partir du corpus français/anglais de la *Convention de sauvegarde de Droits de l'Homme*.

Le chapitre 7 aborde des perspectives de recherche peu explorées jusqu'ici et, en premier lieu, les perspectives d'exploration textométrique de *corpus parallèles catégorisés*. L'étiquetage de corpus parallèles offre des points d'appui précieux pour l'extraction de ressources traductionnelles du bi-texte. Cependant, une homogénéisation des jeux d'étiquettes morphosyntaxiques utilisés pour la catégorisation de deux volets bilingues d'un corpus parallèle se révèle nécessaire avant l'exploration bi-textuelle.

La dernière partie du chapitre 7 décrit sommairement les structures de données indispensables au stockage informatique des équivalences traductionnelles au niveau lexical. Des systèmes d'*index bi-textuels* appuyés sur les techniques de *numérisation textométrique* de la séquence textuelle, couramment utilisées en textométrie monolingue pourraient servir au développement des futurs systèmes de *mémoires de traduction* [5].

L'objectif essentiel que nous nous sommes fixé au cours de ce travail est d'attirer l'attention des spécialistes de nombreux domaines des sciences du langage (linguistes, traducteurs, lexicographes, terminologues, enseignants en langues étrangères, etc.) sur les possibilités ouvertes par ces nouvelles approches quantitatives des corpus parallèles. Nous espérons montrer au fil de ce travail que ces méthodes donnent accès à un réservoir de renseignements précieux qui permettent d'enrichir les pratiques actuelles de l'analyse des données de traduction. Nous sommes convaincue que les outils de la statistique textuelle utilisés à partir de postes de travail équipés de ressources traductionnelles informatisées trouverons très rapidement de nombreuses applications en sciences du langage.

[5] La notion de *mémoire de traduction* est présentée au chapitre 1.

« Il est maintenant possible d'enregistrer et de manipuler par ordinateur des masses pratiquement illimitées de textes /.../ Compte tenu de l'évolution récente en informatique, tout indique en effet que les traducteurs pourront bientôt accéder facilement à de très vastes corpus bilingues contenant leurs propres traductions et celles de leurs collègues. »

[Isabelle et Warwick-Armstrong, 1993, p. 288]

Chapitre 1

Les corpus parallèles

La notion de *corpus parallèle*, qui émerge actuellement dans les travaux de différents chercheurs comme : *corpus comportant plusieurs volets qui correspondent chacun à une version d'un même texte dans deux ou plusieurs langues différentes*, renvoie à des situations connues de coexistence de textes présentant des liens forts dans leur structuration. Paléographes, historiens, théologues, juristes, philologues, linguistes, traductologues manipulent depuis fort longtemps des corpus de textes rassemblant plusieurs volets dont chacun est constitué par une version du même texte dans une langue différente.

On rencontre des documents parallèles dans des combinaisons de langues variées (vieux persan/babylonien/élamite, latin/vieil anglais, vieux slavon/russe, français/arabe, etc.). Comme le remarque Jean Véronis [2000a, p. 151], *« Jusqu'à nos jours, l'Histoire est constellée de textes parallèles (contrats, traités, oeuvres sacrées, littérature, etc.), datant de toutes les époques et concernant presque tous les couples de langues en contact /.../ »* . Parmi les

documents parallèles les plus connus, on citera, par exemple, les textes inscrits sur *la pierre de Rosette* [1] ainsi que les différentes éditions de la Bible.

On parle de parallélisme non seulement entre les textes en relation de correspondance traductionnelle, mais aussi entre réécritures et réinterprétations de textes monolingues susceptibles de permettre une comparaison [2].

Dans le contexte récent de l'informatisation des études des corpus de textes, la notion de parallélisme textuel reçoit des formalisations plus strictes qui permettent de manipuler conjointement les différents volets d'un corpus pluritextuel. L'objectif poursuivi est avant tout d'utiliser les données textuelles parallèles et les structures des documents alignés pour extraire, à partir des corpus, des *ressources traductionnelles* utilisables dans d'autres contextes. Le présent chapitre est consacré au rappel des principaux concepts utilisés dans le domaine de l'analyse automatique des corpus parallèles. La première partie définit la notion de parallélisme textuel. Une seconde partie présente la description des enjeux et courants que l'on rencontre dans le domaine de l'étude informatisée des corpus parallèles multilingues.

[1] La *pierre de Rosette* date de 196 av. J.-C. Les inscriptions sur cette pierre relatent les honneurs rendus au roi Ptolémée V par les prêtres des différents temples d'Égypte. Les textes sont présentés sous forme d'un corpus parallèle en deux langues (le grec et l'égyptien) et trois écritures (les textes égyptiens étant écrits à la fois en hiéroglyphes et en démotique). La *pierre de Rosette* est actuellement conservée au *British Museum.*

[2] Actuellement, il existe des outils de comparaison de textes monolingues qui détectent les différences provenant d'éventuelles insertions et omissions dans les fichiers correspondants. Ce type d'outil est souvent utilisé dans la programmation pour détecter les fichiers identiques par comparaison de leurs contenus respectifs octet par octet. Cette information sous format binaire est ensuite associée aux unités textuelles correspondantes par attribution d'*identificateurs de phrases* (sous forme de clés isolées) ou *valeurs-indices* décrivant la position des mots dans le texte [Heather et Rossiter, 1988].

1.1 Définitions

1.1.1 Le parallélisme textuel

Selon Heather et Rossiter [1988], on peut distinguer quatre types de parallélisme textuel en fonction de l'organisation sémantique et structurelle de l'ensemble de données à l'intérieur des documents[3].

Parallélisme explicite : Les deux textes partagent les mêmes identificateurs d'unités textuelles sous forme de clés facilement accessibles par l'ordinateur. *Exemple :* les différentes éditions de la Bible[4].

Parallélisme fonctionnel : Les deux textes ont, essentiellement, la même structure mais possèdent des identificateurs différents. Une correspondance fonctionnelle peut être établie. *Exemple :* deux versions successives d'un document juridique comportant des différences dans le système de numérotation de sections, paragraphes, phrases, etc. (*partial mapping*), ainsi que des différences dans le contenu.

Parallélisme implicite : Les deux textes sont présentés sous un format qui ne permet pas d'établir des correspondances directes. Néanmoins, il y a suffisamment d'information pour mettre en correspondance les différentes parties de ces textes. *Exemple :* deux versions d'un même traité (en latin et anglais) : *Statutum de Marleberge / The Statute of Marlborough*, 1267. [HMSO Technical Services, 1981], (cf. *Figure 1.1*).

Parallélisme latent : Il s'agit de textes qui sont proches dans leurs contenus. Cependant, cette proximité n'est pas manifestée au niveau structurel. Pour mettre

---

[3] Sur les mécanismes permettant de détecter les similitudes linguistiques et structurelles entre les textes comparés, on consultera, par exemple, [Ghorbel *et al.*, 2002].

[4] Pour la plupart des livres de la Bible, les chapitres actuels sont issus d'une standardisation des divisions de la Bible latine effectuée dans le cadre de l'Université de Paris par Stephen Langton avant 1209. La division de Langton a été ensuite légèrement modifiée dans la suite du XIIIe siècle par le dominicain Hugues de Saint-Chef, auteur de la première concordance biblique.

en évidence les liens sémantiques qui réunissent l'ensemble de ces textes, il faut entreprendre une réorganisation sémantique ou insérer des identificateurs supplémentaires. *Exemple :* plusieurs textes traitant des mêmes thèmes. On parle aussi dans ce cas de *corpus comparables* [5].

[5] Sur l'extraction des ressources traductionnelles des corpus comparables, on consultera, par exemple, [Chiao et Zweigenbaum, 2002].

Statutum de Marleberge / The Statute of Marlborough 1267 (extrait)

Latin

" Anno gre M C C Lx septimo, regni autem Dni H. Regis filii Regis Johis quinquagesimo secundo, in Octav Sci Martini, qvidente ipo Dno Rege, ad regni sui Angl melioracoem & exhibicoem justicie qut regalis officii exposcit utilitas meliorem, convocatis discrecorib3 ejusdem regni tam ex majorib3 qam minorib3; qvisum est & statutu ac concordit ordinatu, ut cu regnu Angl mltis tribulaconib3 & dissensionu incomodis nuq deqssum, reformacone legum & Juriu, quib3 pax & tansquillitas incolarum conservetr, ilndigeat, ad quod remediu salubre q ipm Regem & suos fideles oportuit adhiberi, qvisiones, ordinacoes & statuta subscipta, ab omib3 regni ipius incolis tam majorib3 qam minorib3 firmit ac inviolabilit temqib3 qpetuis obsventr. "

Anglais

" In the Year of Grace, One thousand two hundred sixty-seven, the two- and- fiftieth Year of the Reign of King Henry, Son of King John, in the Utas of Saint Martin, the said King our Lord providing for the better Estate of his Realm of England, and for the more speedy Ministration of Justice, as belongeth to the Office of a King, the more discreet Men of the Realm being called together, as well of the higher as of the lower Estate: 'DaIt was Provided, agreed, and ordained, that whereas the Realm of England of late had been disquieted with manifold Troubles and Dissensions; for Reformation whereof Statutes and Laws be right necessary, whereby the Peace and Tranquillity of the People must be observed; wherein the King, intending to devise convenient Remedy, hath made these A'ks, Ordinances, and Statutes underwritten, which he willeth to be observed for ever firmly and inviolably of all his Subje'ks, as well high as low. "

Figure 1.1 : Exemple de parallélisme textuel implicite

Guide de lecture : La figure reprend un extrait de l'acte du Parlement britannique connu sous le nom de *Statut de Marlborough* (The Statute of Marlborough). Les deux versions du statut (l'une en latin et l'autre en anglais) sont présentées côte à côte. L'analyse des deux textes permet de repérer certaines différences contextuelles. La traduction moderne du statut (encore en vigueur aujourd'hui) est publiée dans *Halsbury's statutes of England* (Hardinge Stanley Giffard Halsbury, 1968).

1.1.2 Le contexte multilingue

Dans le contexte multilingue, les corpus parallèles sont généralement composés de textes sources et de leurs traductions existantes ou des textes dont chacun est une traduction de l'autre sans qu'il soit possible de déterminer lequel a servi de source. Actuellement, le terme *corpus comparables* est utilisé pour se référer à des corpus composés de textes traitant des mêmes thèmes dans plusieurs langues sans être des traductions [6].

Avec la croissance du marché de la traduction, les agents économiques, les organisations internationales s'intéressent de plus en plus à l'archivage électronique conjoint de textes et de leurs traductions dans différentes langues. Ces documents représentent le noyau de la *communication multilingue* [7] et rendent possible l'échange d'information entre communautés. L'information qu'ils contiennent revêt une importance capitale dans plusieurs domaines socio-économiques.

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

[6] Sur les critères pouvant servir à calculer le degré de comparabilité/comparaison entre deux collections de textes, on consultera, par exemple, [Déjean et Gaussier, 2002] ; [Fung, 2000].

[7] Sur les enjeux de la communion multilingue, on consultera, par exemple, [Slodzian et Souillot, 1997].

segments correspondants dans des couples de textes (lexies, locutions, syntagmes, phrases, etc.).

1.1.3 Le corpus bilingue Convention

Pour illustrer les principes de mise en correspondance des segments de traduction des documents parallèles multilingues, 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.*** Le corpus ***Convention*** [8] a été constitué à partir du texte officiel de cette Convention, de ses protocoles intégraux, et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995 [9]. Les deux volets (anglais et français) sont présentés sous forme électronique (cf. *Tableau 1.2*).

[8] Nous remercions Didier Bourigault (Equipe de Recherche en Syntaxe et Sémantique, CNRS – Université Toulouse II) et Benoît Habert (Université Paris X – Nanterre) qui ont accepté de mettre à notre disposition les versions électroniques des textes de cette Convention. Chaque volet du corpus compte approximativement 300 000 mots graphiques.

[9] La *Convention de sauvegarde des Droits de l'Homme et des Libertés fondamentales* a été signée à Rome le 4 novembre 1950. Elle est entrée en vigueur le 3 septembre 1953. Élaborée au sein du Conseil de l'Europe, elle définit un certain nombre de droits fondamentaux (moins nombreux que ceux de la *Déclaration universelle des Droits de l'Homme*) et institue un mécanisme de contrôles et de sanctions propres à assurer le respect de ces droits par les États signataires. Depuis la date de sa création le texte de la Convention a été révisé. Une dizaine de protocoles ont été rajoutés. Aujourd'hui, ils font partie intégrante du document. De 1959 à 1997, la Cour Européenne des Droits de l'Homme a rendu plus de huit cents arrêts (rédigés en anglais et en français). Deux versions de chaque document existent parallèlement ; il est difficile de distinguer une langue source et une langue cible. On peut trouver les textes de la Convention en anglais et en français ainsi que les informations générales sur les divers aspects historiques et juridiques concernant sa création sur le site officiel de la Cour Européenne des Droits de l'Homme : http://www.echr.coe.int. On trouvera à l'annexe A la description de la structure du corpus ***Convention***. L'ensemble du corpus ***Convention*** peut être consulté sur le Cd-rom joint à cette thèse : /fichiersCD/stmz/page1_fichiers/Conv.txt.

Tableau 1.2 : Format initial du corpus ***Convention***

conv_a0_p1-1 1 Les gouvernements signataires, membres du Conseil de l'Europe,	**conv_a0_p1-1e 2** The governments signatory hereto, being members of the Council of Europe,
conv_a0_p2-1 3 Considérant la Déclaration universelle des Droits de l'Homme, proclamée par l'Assemblée générale des Nations Unies le 10 décembre 1948 ;	**conv_a0_p2-1e 4** Considering the Universal Declaration of Human Rights proclaimed by the General Assembly of the United Nations on 10th December 1948;
conv_a0_p3-1 5 Considérant que cette déclaration tend à assurer la reconnaissance et l'application universelles et effectives des droits qui y sont énoncés ;	**conv_a0_p3-1e 6** Considering that this Declaration aims at securing the universal and effective recognition and observance of the Rights therein declared;
conv_a0_p4-1 7 Considérant que le but du Conseil de l'Europe est de réaliser une union plus étroite entre ses membres, et que l'un des moyens d'atteindre ce but est la sauvegarde et le développement des droits de l'homme et des libertés fondamentales ;	**conv_a0_p4-1e 8** Considering that the aim of the Council of Europe is the achievement of greater unity between its members and that one of the methods by which that aim is to be pursued is the maintenance and further realisation of human rights and fundamental freedoms;
conv_a0_p5-1 9 Réaffirmant leur profond attachement à ces libertés fondamentales qui constituent les assises mêmes de la justice et de la paix dans le monde et dont le maintien repose essentiellement sur un régime politique véritablement démocratique, d'une part, et, d'autre part, sur une conception commune et un commun respect des droits de l'homme dont ils se réclament ;	**conv_a0_p5-1e 10** Reaffirming their profound belief in those fundamental freedoms which are the foundation of justice and peace in the world and are best maintained on the one hand by an effective political democracy and on the other by a common understanding and observance of the human rights upon which they depend;
conv_a0_p6-1 11 Résolus, en tant que gouvernements d'Etats européens animés d'un même esprit et possédant un patrimoine commun d'idéal et de traditions politiques, de respect de la liberté et de prééminence du droit, à prendre les premières mesures propres à assurer la garantie collective de certains des droits énoncés dans la Déclaration universelle,	**conv_a0_p6-1e 12** Being resolved, as the governments of European countries which are like-minded and have a common heritage of political traditions, ideals, freedom and the rule of law, to take the first steps for the collective enforcement of certain of the rights stated in the Universal Declaration,
conv_a0_p7-1 13 Sont convenus de ce qui suit :	**conv_a0_p7-1e 14** Have agreed as follows:
conv_a1_p1-1 15 Les Hautes Parties contractantes reconnaissent à toute personne relevant de leur juridiction les droits et libertés définis au titre I de la présente Convention :	**conv_a1_p1-1e 16** The High Contracting Parties shall secure to everyone within their jurisdiction the rights and freedoms defined in Section I of this Convention .
conv_a2.1_p1-1 17 Le droit de toute personne à la vie est protégé par la loi.	**conv_a2.1_p1-1e 18** Everyone's right to life shall be protected by law.

Guide de lecture : Chaque couple de phrases équivalentes est introduit par un code. Les numéros indiquent (dans l'ordre) : le type de document (convention, protocole, etc.) ; le numéro de l'arrêt ; la partie de l'arrêt ; le numéro de section et/ou du paragraphe ; le numéro de la phrase dans le corpus, précédé par la lettre « e » pour les phrases en anglais.

1.2 L'alignement des corpus parallèles

Le traitement de corpus parallèles suppose une phase préalable d'*alignement*, c'est-à-dire de mise en correspondance dans chacun des volets de différents types d'unités textuelles. La difficulté majeure de cette phase réside dans la manipulation des données recueillies et dans la lourdeur de la tâche d'extraction manuelle de ressources. L'apport de l'ordinateur consiste en grande partie dans l'automatisation du dépistage des équivalences.

L'alignement « manuel » est particulièrement fastidieux lorsqu'il s'agit de ressources textuelles volumineuses. Ce problème a conduit les chercheurs à mettre en œuvre des pratiques d'alignement automatique de corpus parallèles. Les différentes approches permettant l'appariement de segments textuels en correspondance constituent un secteur du *Traitement automatique des langues* (TAL) que l'on appelle l'*alignement automatique*.

Aligner des corpus de textes originaux et de leurs traductions c'est mettre en relation des unités textuelles qui se correspondent (cf. *Figures 1.3-4*) [10]. On peut établir des correspondances entre des unités de différents niveaux : mots, syntagmes, phrases, paragraphes, etc. Les traductions produites par les experts humains constituent, en général, un point de départ qui sert à apprécier la pertinence des ressources produites automatiquement et à décider de leur réutilisation.

[10] Actuellement, la notion d'alignement est présente dans plusieurs domaines de recherche. Des méthodes analogues à l'alignement bi-textuel sont utilisées dans l'appariement de séquences phonétiques, cf. Kondrak [2000] ou dans l'alignement du texte avec la transcription phonétique et le signal sonore, cf. Malfrère et Dutoit [2000], Somers [1998b]. Dans le domaine de la biologie moléculaire, les spécialistes sont confrontés à des problèmes d'alignement lorsqu'il s'agit, par exemple, d'apparier les séquences d'aminoacides dans les protéines, cf. Jennings *et al.* [2001].

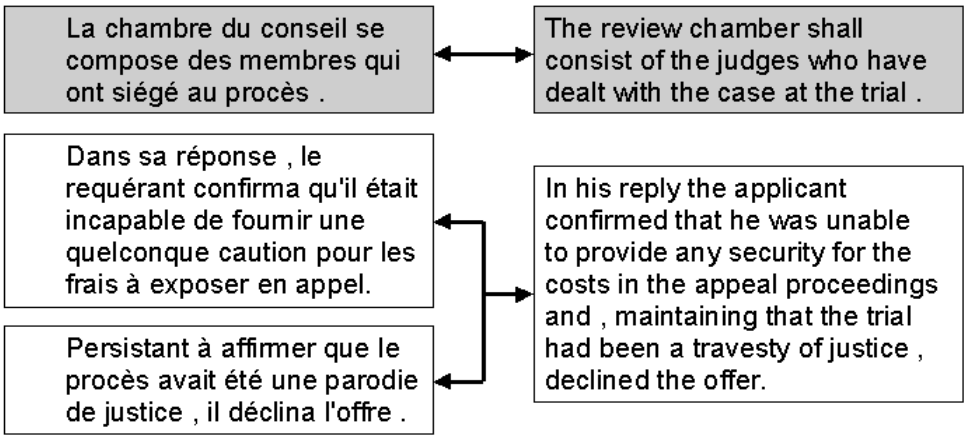


Figure 1.3 : Corpus *Convention* : exemple d'alignement des phrases

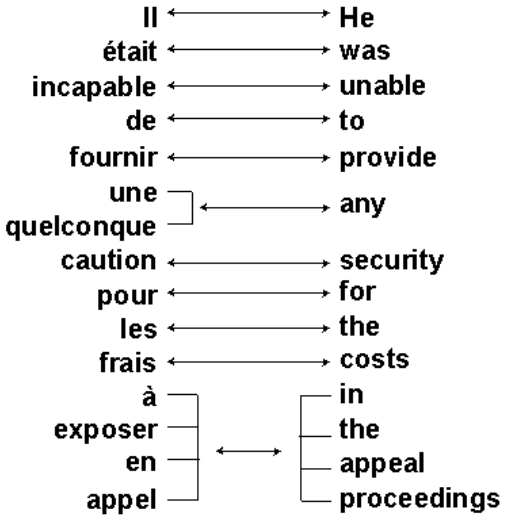


Figure 1.4 : Corpus *Convention* : exemple d'alignement de mots et de syntagmes

1.2.1 L'alignement des phrases

La recherche dans le domaine de l'alignement a montré que le repérage automatique des correspondances est relativement simple dans le cas d'unités de texte de taille importante, telles que chapitres, sections, articles, paragraphes, etc., cf. Isabelle [1992ab], Janicijevic [1997], Véronis [2000ab] [11]. L'utilisation des méthodes probabilistes a donné lieu à des avancées rapides dans l'alignement des phrases [12]. Les comptes-rendus d'expériences publiés récemment décrivent des algorithmes permettant d'apparier les phrases d'un corpus parallèle avec un taux de réussite élevé, cf. Brown P. *et al*. [1991], Gale et Church [1991b], Kay et Röscheisen [1993], Simard *et al*. [1992], Véronis [2000ab] [13]. En revanche, la recherche des correspondances plus fines est particulièrement complexe lorsqu'elle implique le découpage des « unités de sens » qui rentrent dans la construction de l'espace sémantique de la traduction au sein des phrases équivalentes (cf. *Figure 1.5*).

1.2.2 L'alignement des mots et des syntagmes

La recherche automatique de segments de traduction équivalents dans les phrases 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. Malgré de nombreuses difficultés dans l'automatisation de l'alignement au

---

[11] Le découpage en sections ou paragraphes est généralement identique dans le texte original et sa traduction. De plus, le marquage logique et typographique (titres, mise en forme, numérotation, retour chariot, etc.) facilite la création de procédures automatiques pour la mise en correspondance de sections et paragraphes bi-textuels.

[12] Sur l'utilisation de méthodes probabilistes dans l'alignement des phrases, cf. *Chapitre 2*.

[13] Les systèmes actuels d'alignement des phrases de textes parallèles multilingues ont fait récemment l'objet d'une 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 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].

niveau des mots et des syntagmes, on note des avancées importantes dans la réflexion sur l'utilisation coordonnée de plusieurs méthodes pour réaliser ce type de tâche, cf. Ahrenberg *et al*. [2000], Choueka *et al.* [2000], Déjean *et al*. [2002], Fung [2000], Gaussier [1998], Piperidis *et al.* [2000], Wu [2000].

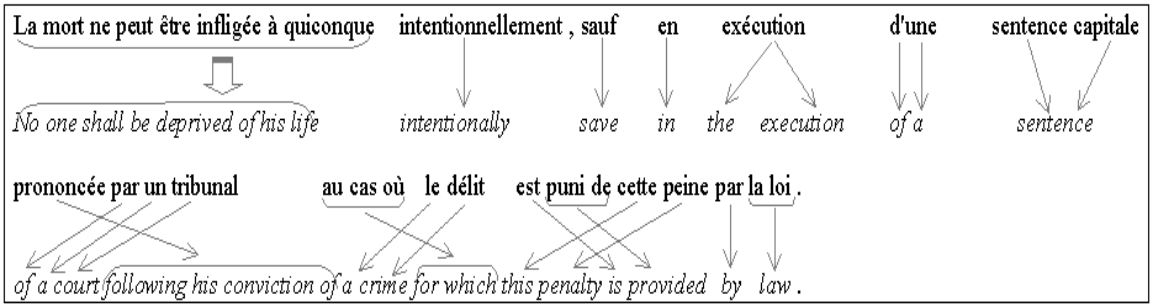


Figure 1.5 : Relations d'équivalence au niveau des mots et des syntagmes

Au cours du processus d'alignement, les données des corpus de textes parallèles sont structurées pour la reconstitution automatique des correspondances de traduction. Cette dimension interactive permet ensuite de concevoir toutes sortes d'outils de recherche et d'extraction de *ressources linguistiques* à base de corpus [14].

[14] Dans ce qui suit, on appellera « outils » les procédures automatiques applicables aux corpus de texte. Ces outils, les produits qu'ils génèrent et les données provenant de la mise en forme sur support numérique de savoirs antérieures (linguistiques, sémantiques, etc.) seront appelés « ressources » (textuelles, traductionnelles, etc.). Sur ces questions, on consultera également le site du *Ministère de la recherche* à l'adresse suivante :
http://www.recherche.gouv.fr/appel/2002/technolangue.htm

1.3 Les outils et les ressources bi-textuels

De nombreux travaux ont montré l'utilité des corpus parallèles alignés pour le développement des applications de traitement automatique des langues, cf. Isabelle et Warwick-Armstrong [1993, pp. 301-303], [Langlois, 1996], [Véronis, 2000a, pp. 152-159], [Véronis, 2000b, pp. 9-14].

Les corpus parallèles alignés permettent de constituer des ressources pour la construction de mémoires de traduction, extraction de dictionnaires et de listes terminologiques bilingues, extraction de connaissances pour la recherche d'information multilingue, construction d'exemples pour l'enseignement assisté par ordinateur (EAO) ou la linguistique constrastive, création de corpus d'entraînement pour les systèmes de désambiguï sation automatique, cf. Véronis [2000a, p. 152-157]. La figure 1.6 montre quelques exemples de ressources bi-textuelles obtenues à partir du corpus ***Convention*** [15].

1.3.1 L'extraction de lexiques bilingues

Comme nous l'avons montré plus haut, l'alignement intégral des unités qui se situent au-dessous du niveau de la phrase pose de nombreux problèmes tant au niveau de la délimitation des segments en correspondance qu'au niveau de l'automatisation de leur appariement. Toutes les correspondances lexicales ne présentent pas le même intérêt pour la constitution de lexiques bilingues, listes terminologiques, dictionnaires, etc. Dans cette optique, l'extraction automatique d'équivalences réutilisables, qui ont acquis une certaine autonomie au niveau du sens et qui pourraient être réutilisées en dehors de leur contexte d'origine, constitue un enjeu particulièrement important pour le « recyclage » des ressources de traductions existantes.

[15] D'autres exemples de ressources traductionnelles obtenues à partir du corpus ***Convention*** sont présentés sur le Cd-rom : /fichiersCD/stmz/page5.htm.

Sur le plan des méthodes, l'*alignement lexical* et l'*extraction de lexiques bilingues* (ou multilingues) représentent deux pôles de recherche connexes. Plusieurs publications récentes sont consacrées à la description des modèles statistiques et linguistiques que l'on peut mobiliser pour l'extraction de ressources traductionnelles des corpus parallèles [16]. Les approches qui sont à l'origine de ces méthodes sont issues, par exemple, des domaines de la *sémantique distributionnelle* en corpus et de la *traduction automatique* probabiliste.

Dans le cadre de l'approche distributionnelle la mise en relation de deux mots ou locutions de langues différentes s'établit sur un plan sémantique, par analyse de leurs distributions sur un ensemble de contextes issus des deux volets d'un corpus parallèle. Cette approche repose sur l'utilisation des ressources syntaxiques et lexicales tirées des corpus eux-mêmes, cf. Blank [2000], Dagan et Church [1997], Déjean et Gaussier [2002], Fung [2000], Gaussier *et al.* [2000], Haruno *et al.* [1996].

A la base de l'approche probabiliste on trouve des modèles statistiques élaborés dans le cadre de recherches sur la traduction automatique. A l'origine, le développement de ces modèles visait, pour des fins de traduction automatique, la génération d'un texte cible à partir d'un texte source, cf. Brown P. *et al*. [1991]. Par la suite, ces calculs probabilistes ont été utilisés pour l'alignement et l'extraction de lexiques bilingues, cf. Brown P. *et al*. [1993], Brown R. [1998], Brown R. *et al*. [2000].

[16] Voir, par exemple, [Véronis, 2000ab], [Somers, 2001].

Corpus *Convention* : exemples de correspondances traductionnelles

Extrait de la concordance autour du terme anglais *recourse*

un raisonnement déductif minimal	minimal *recourse* to logic
choisir plusieurs avocats	to have *recourse* to several lawyers
faire appel à des experts	to have *recourse* to experts

Extrait de la concordance autour du terme anglais *claim*

les prétentions de l'intéressé	applicant's *claim*
les *demandes* de satisfaction équitable	the *claims* for just satisfaction

Extrait de la concordance autour du terme français *administration*

bonne administration de la justice	proper *administration* of justice
bonne administration	good *governance*

Exemples de correspondances traductionnelles multiples

se prévaloir de ***qqch*** (des traités)	to avail ***o.s.*** of ***sthg*** (treaties)
se prévaloir de ***qqch*** en justice	to rely on ***sthg*** in legal proceedings

Exemples d'équivalences contextuelles

la mort ne peut être infligée à quiconque intentionnellement /…/	no one shall be deprived of his life intentionally /…/
les pouvoirs des tribunaux anglais étaient suffisant pour /…/	the scope of the powers of the English courts were sufficient to /…/
Le requérant s'étant soustrait à la justice, il a demandé le rejet de la requête.	Following the applicant's flight, he requested that the application be rejected.

Figure 1.6 :
Exemples de ressources bi-textuelles obtenues à partir du corpus ***Convention***

1.3.2 Les outils bi-textuels

L'automatisation de la mise en correspondance des unités lexicales issues des deux volets d'un corpus parallèle permet d'augmenter considérablement les performances des systèmes d'aide à base de corpus (vérificateurs et extracteurs de terminologie, générateurs semi-automatique de dictionnaires, détecteurs d'erreurs de traduction), ainsi que celles de nombreuses applications de recherche documentaire.

En *Traduction assistée par ordinateur* (TAO), les outils bi-textuels permettent d'acquérir des équivalences traductionnelles sans passer par la traduction automatique, souvent inadaptée. Appuyée sur les pratiques de l'exploration intertextuelle, cette nouvelle famille d'outils « à base de corpus » donne accès à une variété d'usages souvent ignorés par les ouvrages de référence classiques, tels que les dictionnaires [17].

Sur les *figures 1.7-8*, nous avons comparé les ressources sélectionnées par un concordancier parallèle et celles résultant du recours à un traducteur automatique pour traduire le terme juridique français *faits incriminés*. Les résultats de cette comparaison montrent que la qualité de traductions obtenues à l'aide d'une concordance bilingue est bien plus riche que dans le cas de la traduction automatique.

[17] Les outils bi-textuels représentent une ressource précieuse pour un large panel d'utilisations dans plusieurs domaines des sciences du langage. On consultera sur ces questions [Barlow, 1996] ; [King et Woolls, 1996] ; [Klavans et Tzoukermann, 1990] ; [Langlois, 1996] ; [Lixun, 2001] ; [Romary *et al.*, 1995]. Bien évidemment, il appartient à l'expert humain d'évaluer la qualité et la pertinence de ressources traductionnelles mises à sa disposition par les outils bi-textuels. Le statut de ce type de ressources est différent de celui que l'on peut attribuer à des banques terminologiques, dictionnaires, etc. Une systématisation et une étude critique des ressources de corpus de traductions existantes s'imposent lorsque l'on envisage leur intégration à des bases de références classiques.

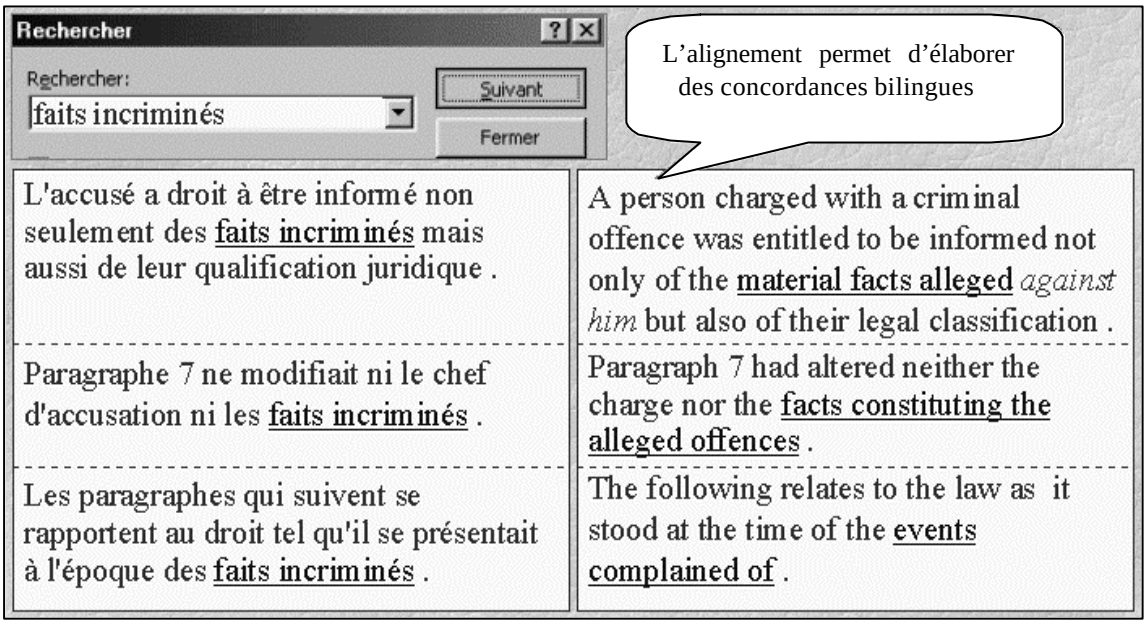


Figure 1.7 :
Exemple d'une concordance bilingue obtenue à partir du corpus ***Convention***

Guide de lecture : Dans un concordancier de ce type dont nous avons simulé ici le fonctionnement, l'utilisateur saisit un terme ou une expression dans une des langues concernées (ici *faits incriminés* en français) puis lance une recherche dans le corpus aligné afin d'accéder à une liste ordonnée des occurrences de ce terme et des traductions de cette expression munies de leur contexte immédiat.

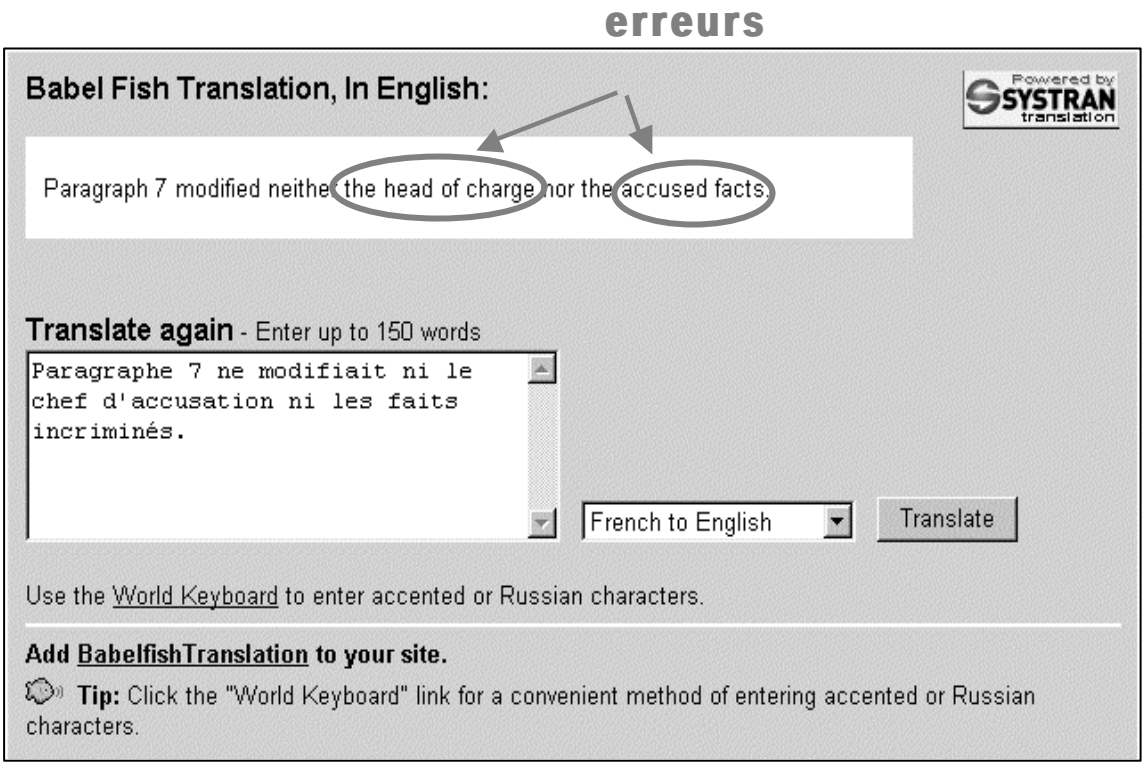


Figure 1.8 : Problèmes de l'acquisition automatique d'équivalences de traduction

Guide de lecture : La figure 1.8 montre les résultats d'une tentative de traduction effectuée à l'aide de l'outil *BabelFish* (sans doute un des plus utilisés sur le Web). L'utilisateur introduit le texte en français puis active le processus de traduction en anglais. La traduction obtenue est syntaxiquement acceptable mais comporte des équivalences inacceptables pour les syntagmes français *le chef d'accusation* et *les faits incriminés* (cf. *Figure 1.7*).

Actuellement, les programmes de ***concordances bilingues*** sont parmi les outils les plus utilisés dans les études informatisées des corpus parallèles, cf. Isabelle [1992ab], Isabelle et Warwick-Armstrong [1993]. Simard, Foster et Perrault [1993, p. 4] décrivent la *concordance bilingue* comme *« /.../ une liste ordonnée de toutes les occurrences de mots d'une paire de textes qui sont des traductions réciproques, à l'intérieur de laquelle chaque occurrence est accompagnée de son contexte, et de la traduction de celui-ci dans l'autre langue »* [18]. On conçoit que la structure d'un programme capable de gérer automatiquement des liens entre les textes dans deux ou plusieurs langues est plus complexe que celle qui est employée dans la construction d'un concordancier monolingue.

Traditionnellement, les programmes de concordance monolingue (ou concordanciers) permettent d'afficher avec chaque occurrence d'un mot la quantité de contexte spécifiée par l'utilisateur. Il est possible souvent de sélectionner les sous-ensembles de la concordance intégrale qui intéressent l'utilisateur à l'aide d'un langage de requêtes spécialisé, qui permet de décrire les mots ou expressions dont on désire examiner les contextes d'utilisation. Une structuration préalable du texte se révèle indispensable pour aboutir à ce genre d'interaction sur un corpus. Elle s'appuie sur une forme d'indexation du texte qui permet de localiser en temps réel toutes les occurrences d'une chaîne de caractères donnée, cf. Salem [1987], Simard *et al*. [1993], Lebart et Salem [1994, pp. 52-55].

Dans le cas d'une concordance bilingue (ou multilingue) il s'agit non seulement d'afficher le contexte qui entoure le mot donné, mais aussi de déterminer sa

[18] Cette définition soulève un certain nombre de points importants lié à la définition du contexte bilingue. Dans une concordance monolingue, le contexte correspond soit à un nombre précis de caractères des deux côtés de l'occurrence du mot faisant objet de la requête, soit à un contexte correspondant au découpage en phrases ou paragraphes dans lequel figure cette occurrence. Dans une concordance bilingue, il faut également pouvoir afficher la traduction de ce contexte. Il faut donc pré-calculer les correspondances entre les textes source et cible et stocker cette information avec le corpus. Le terme de *mémoire de traduction* est parfois employé pour décrire ce type d'architecture bi-textuelle.

traduction[19]. Il faut, de plus, limiter ce contexte à une sous-chaîne du texte de la traduction. Le coût de cette localisation dépend de la taille du corpus bilingue étudié. Elle peut être effectuée en temps réel à condition que l'alignement des unités des deux textes ait été effectué préalablement [20].

Dès leur apparition, les systèmes de concordances parallèles utilisent l'alignement des phrases pour la construction de contextes parallèles. Pour faciliter l'exploitation des données bi-textuelles et pour contourner le problème d'absence d'alignement intégral au niveau des mots, plusieurs solutions sont proposées à l'utilisateur :

a) *Recherche par forme-pôle* : Les concordanciers bilingues permettent d'obtenir la liste des phrases d'un des volets du corpus parallèle où apparaît la forme recherchée ainsi que les phrases de l'autre volet qui y sont liées sur le plan de la traduction. Il appartient à l'utilisateur de reconnaître la traduction de cette forme dans les contextes affichés.

b) *Recherche par couple de formes bilingues* : Ce type de requête permet de rechercher des occurrences de traductions spécifiques. La recherche se révèle utile lorsque l'utilisateur souhaite trouver des phrases où l'une des formes est traduite par l'autre. Extraire, par exemple, tous les contextes d'un corpus de textes parallèles anglais/français où le mot anglais *commitment* est traduit en français par *attachement*. Une telle recherche demande à l'utilisateur de connaître la traduction du terme recherché. Elle se révèle problématique lorsque le mot français *attachement* se trouve accidentellement avec un sens différent

[19] Dans les travaux récents sur l'alignement, les concordanciers bilingues capables de localiser et de signaler automatiquement à l'utilisateur la traduction du terme faisant objet de sa requête sont parfois appelés des systèmes *BiKWIC* (Bilingual Key Word In Context), cf. Barlow [1999].

[20] Sur les programmes de concordances parallèles multilingues, on consultera les sites suivants :
TransSearch : http://www.tsrali.com
Multiconcord (1): http://info.ox.ac.uk/ctitext/resguide/resources/m145.html
Multiconcord (2): http://artsweb.bham.ac.uk/pking/multiconc/l_text.htm
ParaConc : http://www.ruf.rice.edu/~barlow/parac.html
Logiciels de LORIA : http://www.loria.fr/equipes/led/outils.php

[ex. : *attachement* – fichier joint à un courrier électronique] dans la phrase équivalente à celle qui contient le mot anglais *commitment.*

On utilise aussi la recherche par couple de formes bilingues pour afficher les phrases où l'un des termes recherchés n'est pas traduit par la forme indiquée dans la zone de recherche. Par exemple, extraire tous les contextes où le mot anglais *head* n'est pas traduit par *tête* en français.

c) *Recherche formulée à l'aide de méta-caractères* : L'utilisation de méta-caractères permet une plus grande flexibilité dans la reconnaissance des unités lexicales complexes. Le langage des *expressions régulières* [21] est communément utilisé pour constituer des groupes de mots faisant objet de la recherche.

Exemples :

- Dans le système *TransSearch* [Simard *et al.*, 1993], l'expression `f(adresse) e(address) ~f(postale)` donne accès à l'ensemble de contextes français/anglais où *adresse* et *address* apparaissent ensemble, sauf dans les cas où ils sont en cooccurrence avec *postale* dans la partie française.
- Dans le programme de Barlow *ParaConc* [Barlow, 2002], l'expression `sp[eo]a?k[se]?n?` [22] identifie les chaînes de caractères correspondant aux mots anglais *speak*, *speaks*, *spoke*, *spoken* etc.

d) *Recherche appuyée sur les dictionnaires* : L'intégration de dictionnaires (comportant des indications de *lemmes*, de *catégories*, de *lexies*, etc.) aux systèmes de concordances bilingues offre des moyens de reconnaissance de l'expansion morphologique. Les deux dictionnaires (un pour chaque langue)

[21] 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, voir [Desgraupes, 2001]. Sur les utilisations des expressions régulières dans l'analyse textuelle, cf. Habert *et al.* [1998].

[22] Dans cette expression, les caractères entre crochets constituent des alternatives. Par exemple, sp[*eo*] permet de rechercher des mots commençant par une chaîne de caractères sp*e* ou sp*o*.

permettent de réaliser la conversion de chacune des formes de la requête en une disjonction de formes fléchies. Par exemple, lorsqu'une recherche dictionnairique est déclenchée par l'utilisateur, pour l'expression *mordre la poussière*, le système produira toutes les occurrences de *mordre la poussière*, mais aussi *mord la poussière, mordit la poussière,* etc.[23]

1.3.3 Les mémoires de traduction

Comme nous l'avons montré dans la section précédente, la réalisation d'une concordance bilingue implique une certaine forme de stockage des équivalences avec les textes source et cible afin d'assurer des accès rapides à l'information. Le concept de mémoire de traduction (*Translation Memory*) est actuellement utilisé pour décrire les systèmes permettant la structuration et l'exploitation ultérieure des données multilingues alignées[24]. Il s'agit, en quelque sorte, d'un système de « recyclage » des données de traductions existantes qui regroupe à la fois des outils et des ressources bi-textuels.

Au moment de leur intégration dans un système de mémoire de traduction, les corpus de textes parallèles doivent être soumis à alignement pour fixer des liens formels qui maintiennent la relation d'équivalence de traduction. De récentes publications montrent qu'il n'existe pas encore de consensus quant à la définition de la structure et des fonctions d'une mémoire de traduction, cf. Dennett [1995],

---

[23] Notons dans ce contexte que les expérimentations de Simard, Foster et Perrault [1993] avec le système *TransSearch* amènent à croire que l'intégration de connaissances linguistiques est susceptible de résoudre des ambiguï tés lexicales et améliorer le filtrage du « bruit » lors de la réalisation de concordances parallèles. Les auteurs soulignent notamment l'importance d'étiquetages morpho-syntaxiques et sémantiques des mots de corpus parallèles pour réaliser ce type de tâches.

[24] Dans la pratique, un ensemble de normes uniques d'encodage bi-textuel (ou multitextuel) se met progressivement en place pour concevoir ce que l'on appelle ici 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) : http://www.lisa.org/tmx. Le TMX devrait permettre à moyen terme de converger vers un système commun d'archivage électronique des traductions existantes alignées. Sur l'analyse du concept de mémoire de traduction, on consultera [Melby, 2000].

Webb [1999], Gaussier *et al*. [2000], Melby [2000], Simard [2003]. L'utilisation principale envisagée est celle du stockage de textes alignés et indexés et l'extraction de contextes bilingues (ou multilingues) qui se rapprochent, autant que possible, de la requête initiale de l'utilisateur [25]. Le caractère interactif d'un système de mémoire de traduction offre au traducteur des possibilités pour résoudre un grand nombre de problèmes qu'il rencontre dans son travail en prenant en considération une variété de solutions pour la traduction d'une expression donnée.

L'utilisation de ce type d'aide débouche sur un concept plus large de *poste de travail du traducteur* (PTT), englobant les outils d'aide à la lecture, à la rédaction, à la gestion et à la transmission de textes (traitements de textes, vérification d'orthographe, dictionnaires électroniques, bases de données terminologiques, etc.). Plusieurs publications récentes sont consacrées à la description de l'ensemble des outils destinés à être intégrés au poste de travail du traducteur, cf. Melby [1981], Macklovitch [1992, 1993].

[25] EAGLES (Expert Advisory Group on Language Engineering Standards) définit le concept de mémoire de traduction (*translation memory*) comme « */.../ a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions* » [EAGLES, 1995].

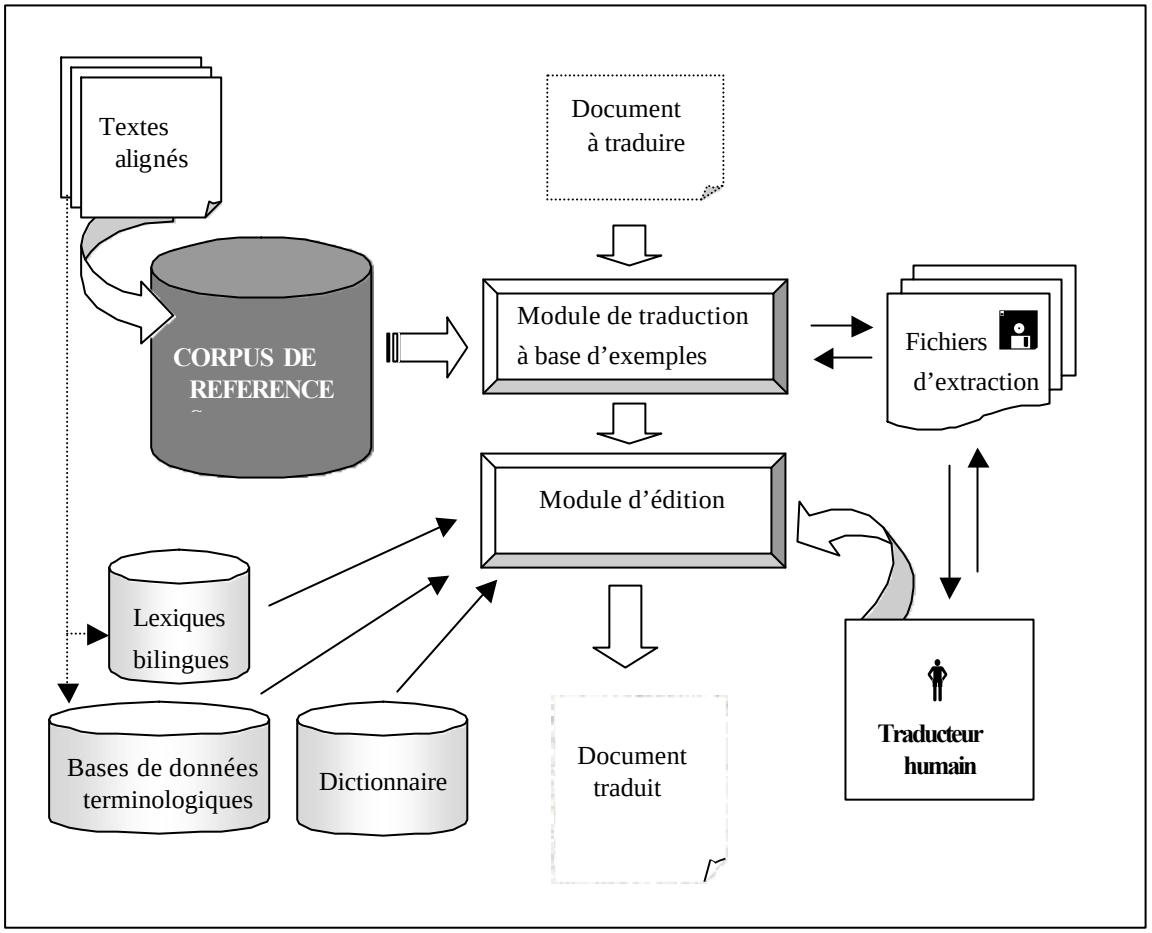


Figure 1.9 : Architecture d'un système de mémoire de traduction

Guide de lecture : Un système de mémoire de traduction comprend les éléments suivants :
- un corpus de référence composé de textes alignés ;
- un module de traduction qui permet de rechercher dans le corpus des exemples de traduction se rapprochant le plus possible du segment de texte à traduire ;
- un module d'édition qui synchronise l'affichage du texte à traduire et de la traduction de référence suggérée par le système ;
- des ressources dictionnairiques et des listes de termes pour assister le traducteur humain dans la phase d'édition.

Nous avons constitué ce schéma à partir de la description de systèmes de mémoire de traduction proposée par Dennett [1995].

1.4 Domaines d'application

Dans les sections précédentes, nous avons montré que les outils bi-textuels à base de corpus parallèles alignés trouvent de nombreuses applications dans les domaines des sciences du langage.

Pour le traducteur, les corpus alignés fournissent un accès au savoir-faire de la communauté lorsqu'il s'agit de traduire une expression pour laquelle les solutions proposées par des ouvrages de référence (dictionnaires, base de données terminologiques, etc.) sont insatisfaisantes. L'alignement de textes source et cible aide à vérifier l'adéquation de la traduction à l'original et à repérer les omissions, faux-amis, etc. Les corpus alignés aident à contourner des limites rencontrées dans l'automatisation de la traduction, cf. Isabelle [1992a].

En lexicographie, l'alignement des données de traduction met en lumière la spécificité et la richesse du vocabulaire bilingue. Il permet de découvrir des usages et des expressions ne figurant pas encore dans les dictionnaires. La création automatique des lexiques bilingues alignés permet également de normaliser la terminologie dans un champ de communication. Elle aide à la maintenance de la documentation technique multilingue.

Les travaux sur les corpus alignés contribuent actuellement au développement des moteurs de recherche multilingues sur le Web, cf. Resnik [1998], Diab [2000]. L'intégration des données textuelles alignées permet d'améliorer les performances des logiciels de traduction automatique.

1.5 Perspectives de recherche

D'une façon générale, les recherches actuelles dans le domaine du traitement automatique de corpus parallèles visent à mettre en place les systèmes nécessaires pour la production, pour la maintenance, l'amélioration et l'évaluation des données textuelles multilingues alignées.

Pour encourager l'accès aux ressources de corpus parallèles multilingues, les systèmes futurs seront fondés sur les notions de ressources libres (*open source*) munis d'interfaces de communication ergonomiques, cf. Casillas *et al*. [2000], Johansson *et al.* [1996], Mangeot-Lerebours [2001], Romary [2000], Romary et Bonhomme [2000][26]. Leur spécificité résidera en une approche *anthropocentrée*, c'est à dire une approche dédiée à un utilisateur en fonction de ses besoins [Beust, 2002]. Axés sur les principes de recyclage de données textuelles multilingues, ces nouveaux systèmes devront intégrer plusieurs fonctionnalités :

- Accès au texte du corpus initial. Visualisation de chaque couple de segments textuels appariés, avec possibilité de mettre en évidence tel ou tel type d'information lexicale (surlignage, coloriage, etc.).
- Identification automatique de thèmes et représentation des équivalences qui y sont associées dans les cadres synthétiques.
- Navigation à plusieurs niveaux de cartes pluritextuelles : thèmes, sous-thèmes, réseaux sémantiques.
- Consultation transversale et requêtes complexes. Prise en compte de mots composés et de leur variation lexicale.
- Intégration d'outils spécialisés de visualisation d'objets complexes (arbres, graphes, etc.).

Différents projets logiciels élaborés dans le cadre de collaborations internationales visent à offrir à l'utilisateur de corpus parallèles des fonctionnalités de ce type, cf. Véronis [2000a, pp. 157-158] [27].

---

[26] Le standard XML (eXtensible Markup Language) permet de bénéficier d'un format générique de représentation des ressources linguistiques multilingues pour de nombreux développements logiciels, par exemple, cf. Romary [2000].

[27] Sur les projets menés actuellement dans les domaines du traitement automatique de corpus parallèles, on consultera les sites référencés dans la *CyberBibliographie* sur le Cd-rom qui accompagne ce volume : /fichiersCD/stmz/page3.htm.

Chapitre 2

L'alignement automatique des corpus

Dans ce chapitre, nous nous intéressons tout particulièrement aux différents courants de recherches issus de la *traduction automatique*, *linguistique computationnelle* et *traductologie* qui ont contribué au développement du domaine que l'on appelle l'*alignement automatique des corpus bilingues ou multilingues*. Nous tenterons de recenser les faits saillants de l'histoire du domaine de l'alignement et d'illustrer les approches les plus caractéristiques par des exemples concrets de méthodes utilisées actuellement [1].

[1] Debili [2000, p. 101] formule les problèmes qui demeurent dans le domaine de l'alignement automatique de la façon suivante : « *Les problèmes que pose l'appariement de paires de textes bilingues ou monolingues se situent à plusieurs niveaux. D'abord* au niveau des unités *que l'on souhaite apparier : les paragraphes, les phrases, les mots, les expressions, ou même les morphèmes. On tombe ici sur le problème classique que posent la définition et la délimitation automatique de ces unités.* Ensuite au niveau de l'appariement *lui-même : il n'y a pas de correspondance biunivoque, loin s'en faut… Enfin* au niveau algorithmique *: quelles connaissances et comment les faire intervenir pour apparier ?* ».

2.1 La traduction automatique et l'alignement des textes

Les premiers travaux dans le domaine de l'alignement de textes originaux et de leurs traductions datent de la fin des années quatre-vingt [2]. Ils sont nés des difficultés rencontrées dans le développement des systèmes de *traduction automatique*. A l'époque, la traduction automatique est marquée par une vague d'innovations conceptuelles importantes qui bouleversent les recherches dans le domaine [3]. Avant de les décrire plus en détail, nous allons analyser brièvement les recherches menées autour de systèmes de traduction automatique traditionnelle qui ont suscité des acquis méthodologiques importants dans ce domaine.

2.2.1 Les enjeux et les limites de la traduction automatique

Les tentatives de traduction automatique ont vu le jour dans les années cinquante. L'objectif essentiel était alors de traduire des articles scientifiques et techniques [4]. Ces systèmes pionniers traduisent pratiquement mot à mot, utilisant de manière annexe quelques stratégies rudimentaires d'analyse linguistique pour traiter des problèmes tels que l'ordre des mots et les flexions [5]. Les méthodes de programmation restent rudimentaires, les capacités de mémoire des ordinateurs sont insuffisantes pour enregistrer les quantités nécessaires d'information linguistique.

---

[2] Sur l'émergence et le développement de méthodes d'alignement de corpus parallèles au fil des vingt dernières années, on consultera Véronis [2000a] ; Somers [2001].

[3] Voir Hutchins [1998] ainsi que les autres publications de ce même auteur sur le site : http://ourworld.compuserve.com/homepages/WJHutchins/

[4] Voir [Hutchins, 1994, 2001] pour un exposé sur l'histoire de la traduction automatique.

[5] Dans le contexte d'après-guerre, la possibilité de traduire automatiquement ce type de textes a un intérêt stratégique. Il intéresse au plus haut niveau l'armée et les services d'espionnage en Union Soviétique et aux États Unis. Les deux pays favorisent le développement des programmes de recherches sur l'automatisation de la traduction du russe vers l'anglais et l'inverse [Bouillon, 1998, p. 9].

Dans un compte-rendu sur les progrès de la traduction automatique, Bar-Hillel [1960][6] estime qu'il est impossible de concevoir une traduction entièrement automatique de haute qualité sans prendre en compte le *sens*. Dominée alors par le *structuralisme*, la théorie linguistique ne s'intéresse que très peu aux problèmes de représentation de sens et ne permet pas la création de modèles conceptuels pour le développement de systèmes informatiques de traduction.

Les recherches dans ce domaine sont alors rapidement stoppées. En 1966, les défauts majeurs des systèmes de traduction automatique de l'époque sont sévèrement critiqués dans le rapport ALPAC (Automatic Language Processing Advisory Committee)[7] qui conclut que les recherches en cours ne sont pas rentables pour l'État américain.

A partir des années soixante-quinze, les recherches en traduction automatique reprennent de l'importance en Europe. La Communauté Européenne déclanche un plan d'action dont le but est de coordonner différents projets qui traitent du multilinguisme et de la traduction automatique [Bouillon, 1998, pp. 11-12]. Le recours au système de traduction automatique *Systran*[8] pour aider les experts humains à faire face à l'accroissement alarmant du besoin de traductions au sein de la Communauté Européenne, stimule les investissements dans ce domaine dans le secteur privé. Les outils de traduction automatique commerciaux se multiplient.

Dans son article *« Vers une nouvelle époque en traduction automatique »*, Hutchins [1994, p. 1] constate que de 1975 à 1988 la plupart des systèmes commerciaux disponibles sur le marché utilisent la méthode « directe » de traduction et quelques éléments de la méthode de « transfert syntaxique ». Ces

---

[6] Voir aussi la présentation du compte-rendu de Bar-Hillel sur le site de John Hutchins : http://ourworld.compuserve.com/homepages/WJHutchins/Miles-3.pdf

[7] ALPAC [1966]. *Language and machines*: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee (ALPAC), Washington, DC, National Academy of Sciences. Sur les principales conclusions de ce rapport, cf. Hutchins [2001].

[8] Voir Hutchins et Somers [1992] : http://ourworld.compuserve.com/homepages/WJHutchins/IntroMT-TOC.htm

systèmes se fondent sur des dictionnaires bilingues relativement riches, assortis d'une analyse linguistique assez superficielle. Dans le milieu de la recherche, domine une approche basée sur des règles linguistiques telles que les règles morphologiques et syntaxiques de génération, les règles de transfert lexical, les règles de désambiguïsation, etc. [9]

L'utilisation des méthodes de « transfert syntaxique » pour la traduction automatique est influencée par le développement des *théories syntaxiques formelles* [10] dans la recherche linguistique. Le trait caractéristique de ces systèmes basés sur des règles linguistiques est le codage des représentations sous forme d'arbres étiquetés. Ces systèmes s'appuient sur des grammaires et des règles de transformation qui déterminent les contraintes et limitent les possibilités de transfert aux différents niveaux linguistiques (d'un arbre morphologique à un arbre syntaxique, d'un arbre syntaxique à un arbre sémantique, d'un arbre d'interface du texte source à un arbre équivalent du texte cible, etc.). Des formalismes complexes fondés sur ces types de contraintes ont été développés pour définir les systèmes de règles abstraites dirigeant ces transformations. [Hutchins, 1994, pp. 4-8] [11].

Pour dépasser les limites mises en évidence par cette deuxième génération de traducteurs automatiques, les recherches successives se sont orientées vers des stratégies de dissociation partielle de l'approche sémantique et de l'approche syntaxique. Avec le temps, l'approche sémantique s'est beaucoup rapprochée des

---

[9] Vers le milieu des années quatre-vingt, on assiste à l'émergence de systèmes basés sur des connaissances directement issues du domaine des textes à traduire, voir, par exemple, [Brachman et Schmolze, 1985].

[10] Sur la présentation de divers formalismes syntaxiques qui servent de base au traitement automatique des langues, y compris en traduction automatique, voir [Miller et Torris, 1990].

[11] Citons, à titre d'exemple, le système *Ariane* (GETA – *Groupe d'étude pour la traduction automatique*) qui fait partie des traducteurs automatiques de la deuxième génération relevant de l'approche de transfert. Voir la présentation de Blanchon à l'adresse suivante : http://www-clips.imag.fr/geta/herve.blanchon/Docs/Intro%20TA.

recherches sur la modélisation des connaissances menées dès cette époque en *intelligence artificielle* [12].

Progressivement, l'orientation syntaxique des premiers systèmes de transfert perd de son importance et on voit croître l'intérêt porté à la construction de lexiques spécifiquement conçus pour la traduction automatique. Le processus de traduction est vu comme une identification et une sélection progressive d'unités lexicales de la langue cible satisfaisant à des contraintes sémantiques liées aux points de départ lexicaux de la langue source [13]. L'acquisition de ressources lexicales devient alors un objectif prioritaire pour un large panel d'applications.

Cependant, l'acquisition de ce type de ressources se révèle rapidement une opération relativement complexe et coûteuse car les sources lexicographiques existantes ne couvrent pas toujours les besoins des domaines particuliers. Ces circonstances font apparaître plus clairement la nécessité d'exploiter les corpus de textes bilingues. L'intérêt pour les ressources traductionnelles existantes est stimulé par le rapide succès des expérimentations à base de corpus de textes [14]. L'efficacité des nouveaux systèmes basés sur des collections « d'exemples de traductions » entraîne dans les années suivantes l'expérimentation de méthodes statistiques [Habert *et al*., 1997].

---

[12] Le domaine de l'Intelligence Artificielle (IA) rassemble des courants de recherches dont l'objectif est de rendre les machines capables d'imiter certains des comportements humains. L'approche IA est souvent utilisée pour le dialogue homme-machine au moyen de la parole, de l'écrit, etc. Par exemple, certaines procédures d'Intelligence Artificielle sont actuellement utilisées dans la recherche documentaire, pour des fonctions d'aide au diagnostic (diagnostic industriel, diagnostic médical, etc.). En traduction automatique, l'approche IA a pour objectif de construire une représentation des univers extralinguistiques dans lesquels on plongera les textes à étudier et de tenter de simuler le raisonnement humain dans le cadre de ces univers, afin d'effectuer des tâches élémentaires liées à leur interprétation dans une autre langue.

[13] Cette conception de la traduction automatique implique l'intégration de différentes connaissances relatives à la langue : *« /.../ quels sont les différents mots, comment ils se prononcent, ce qu'ils signifient, comment ils se combinent pour former une phrase et comment le sens des différents mots contribue au sens de la phrase. »* [Bouillon, 1998, p. 12-13].

[14] L'approche à base de corpus est présentée dans [Leech, 1987] ; [Aijmer et Altenberg, 1991], [Habert *et al.*, 1997] ; [Bourigault et Slodzian, 1999].

L'apparition de méthodes et de stratégies «à base de corpus» vers la fin des années quatre-vingt représente une approche radicalement nouvelle des problèmes de la traduction automatique. Depuis, cette nouvelle tendance basée sur l'exploitation des exemples de traduction n'a fait que se renforcer, elle est actuellement dominante en traduction assistée par ordinateur [15].

2.1.3 La traduction semi-automatique à base de corpus

Les tentatives de traduction entièrement automatisée n'ont connu qu'un succès limité. Actuellement, malgré des investissements importants, les solutions techniques existantes sont encore loin d'être satisfaisantes. Elles constituent souvent un compromis partiellement acceptable entre l'exigence de qualité des résultats, les objectifs d'automatisation totale et ceux de l'extension du domaine d'application.

Certains programmes de recherches appliquées à la traduction automatique spécialisée ont montré la rentabilité de systèmes conçus pour fonctionner dans le cadre de *sous-langages* [16] (domaines dans lesquels le vocabulaire et la syntaxe de textes rencontrés sont considérablement restreints). L'exemple le plus connu de l'application de la traduction automatique à un sous-langage est le système *METEO*® [17] qui sert à traduire les prévisions publiques, agricoles et maritimes émises par Environnement Canada. Les tentatives d'arriver à un résultat similaire avec un sous-langage ayant des règles syntaxiques et sémantiques plus

---

[15] Sur ces questions, voir, par exemple, l'exposé de Chandioux à l'adresse suivante : http://www.univ-tlse2.fr/gril/TAL/TRAD/TRADAUTO1.htm

[16] La notion de *sous-langage*, c'est-à-dire une langue aux règles propres plus restrictives que celles de la langue générale, pour les domaines scientifiques et techniques, a été développée dans les travaux de Zellig Harris. Celle-ci, en limitant le domaine, permet la simplification de la grammaire et la limitation du lexique et des contraintes contextuelles [Harris, 1988]. Sur les langues de spécialité et la traduction, on consultera également Gémard [1999].

[17] Le système *METEO* a été développé dans le cadre du projet TAUM (Traduction Automatique de l'Université de Montréal). Les recherches ont commencé en 1970. *METEO*® est entré en exploitation en 1977 à Montréal. Son exploitation s'est révélé être un vrai succès. Voir [Isabelle et Macklovitch, 1990].

complexes (celui des manuels techniques dans le domaine de l'aviation) ont échoué.

Pour avancer dans la voie de l'acquisition automatisée d'équivalences de traduction, les recherches successives se sont orientées vers la constitution de corpus de traductions. Cette nouvelle approche *à base de corpus* repose sur l'hypothèse qu'une masse importante de données traductionnelles est susceptible de fournir des correspondances pertinentes, précédemment établies par des traducteurs humains, que l'on pourra ensuite réutiliser pour de nouvelles traductions. Ainsi, en mettant systématiquement en évidence l'ensemble de liens qui existent entre les segments respectifs de textes originaux et de leurs traductions, il devient possible de résoudre un grand nombre de problèmes de traduction en effectuant le « recyclage » des données collectées au fil des corpus de textes bilingues.

Cette nouvelle approche change la place réservée à la machine dans l'automatisation des tâches liées à la traduction: *« Les seuls systèmes d'applicabilité générale qui soient présentement réalisables en traduction sont ceux qui laissent au traducteur humain l'entière initiative du processus de traduction, mettant à sa disposition un ensemble intégré d'outils susceptibles de l'aider dans sa tâche »* [Isabelle et Macklovitch, 1990, p. 7][18]. L'orientation vers la traduction semi-automatique supplante désormais les tentatives d'automatisation totale. On peut distinguer deux courants parallèles :

- la traduction humaine assistée par ordinateur ;
- la traduction automatique assistée par l'humain.

Pour le premier courant, l'être humain est au centre du processus de traduction. La machine est considérée comme un appui pour l'accomplissement des tâches périphériques. A l'inverse, la traduction automatique assistée par l'humain place l'intervention humaine dans le cadre de la phase ultime d'édition. Cette intervention indispensable jette parfois le doute quant à l'utilité globale de ce

[18] Notons dans ce contexte que la traduction automatique approximative permet néanmoins de saisir le sens global d'un texte. Elle est souvent utilisée pour la traduction de pages Web (ou des courriers électroniques) rédigées dans des langues étrangères.

type de systèmes. En revanche, le gain économique venant de l'utilisation de systèmes de traduction machine assistés par l'humain est confirmé par de nombreux spécialistes, à propos de plusieurs couples de langues. Dans cette perspective, la création d'archives de textes bilingues appariés sur support numérique stimule les investissements des entreprises dans le domaine, surtout dans les pays où sont utilisées communément plusieurs langues officielles.

Ainsi, par exemple, le système sociopolitique canadien exige des traductions de haute qualité produites dans des très brefs délais. Les traducteurs sont régulièrement confrontés à des problèmes complexes pour lesquels les solutions toutes faites obtenues en utilisant des ouvrages de référence standards (dictionnaires, base de données terminologiques, etc.) ne conviennent pas. Dans le même temps, de vastes corpus de textes déjà traduits sont systématiquement archivés. Cependant, leurs ressources sont inaccessibles sans développement de systèmes à un certain niveau d'automatisation qui permettraient d'en extraire des solutions toutes faites pour de nouvelles traductions. Dans ces conditions, le concept de *bi-texte* [Harris B., 1988] a été introduit pour désigner le noyau d'une nouvelle génération d'aides informatisées à la traduction.

La notion de bi-texte

Les différents schémas visant une description formelle des rapports d'équivalence qui s'établissent entre les segments de textes source et cible s'intègrent dans la description du concept de *bi-texte*. A l'origine, le terme a été introduit pour désigner une mise en correspondance interactive entre les segments de textes bilingues [19].

[19] Sur l'origine du concept de *bi-texte*, [Melby, 1982] ; [Harris B., 1988]. Plusieurs publications récentes sont consacrées aux questions de la représentation d'un bi-texte et l'évaluation de sa qualité. Voir [Isabelle et Simard, 1996] ; [Simard et Plamondon, 1996] ; [Langlais *et al.*, 1998] ; [Véronis et Langlais, 2000].

Actuellement, on rencontre plusieurs définitions techniques de ce terme étroitement liées aux principes retenus pour la segmentation parallèle de textes source et cible en un ensemble fini d'équivalences traductionnelles [20].

D'après la définition de Pierre Isabelle [1992a, pp. 724-725], un ***bi-texte*** peut être représenté par le schéma suivant :

> un quadruplet $<T_1, T_2, Fs, C>$, dans lequel :
>
> - T_1 et T_2 sont deux textes ;
> - ***Fs*** est une fonction qui décompose ces deux textes en des ensembles de segments ;
> - **C** est un ensemble de correspondances entre $Fs(T_1)$ et $Fs(T_2)$.

Cette définition repose sur le principe de *compositionnalité de traduction* [21]. Ce principe pose que l'équivalence sémantique globale est axée sur une hiérarchie structurée de correspondances. Isabelle [1992a] propose d'assimiler la fonction de segmentation *Fs* à une fonction d'analyse syntaxique. La décomposition commence par constituer des unités larges (sections, paragraphes, etc.) pour descendre ensuite vers des unités de plus en plus petites (syntagmes, mots, morphèmes).

---

[20] Par exemple, Melamed [2000b, p. 26] utilise le concept de *bi-texte* pour définir l'espace commun formé par un couple de textes bilingues dans lequel seront recherchées des correspondances. En outre, le terme *bi-text mapping* est utilisé dans la littérature anglo-saxonne pour désigner la génération automatique d'une table de correspondances possibles (même approximatives et ambiguës) dans les zones équivalentes de corpus parallèles. Ces zones peuvent correspondre, par exemple, à des couples de phrases appariées, voir [Brown R., 1998].

[21] Le principe de *compositionnalité* est utilisé depuis longtemps pour la mise en évidence des correspondances de traduction. Dans des domaines tels que l'enseignement des langues, les études des documents anciens, la compilation de dictionnaires, etc., certains modes de mise en page font appel à ce principe : la mise en page en colonnes associe les paragraphes, la mise en page interlinéaire – les phrases. Les manuscrits de textes parallèles, par exemple, sont conçus de manière à faciliter le repérage des liens qui existent entre segments de texte équivalents. Sur le parallélisme des documents anciens [Véronis, 2000a] ; [Ghorbel *et al.*, 2002].

Le résultat de cette décomposition correspond à une hiérarchie de correspondances de traduction : la traduction d'une section est composée de l'ensemble des traductions des paragraphes qui la constituent, la traduction d'un paragraphe se réduit à la traduction des phrases qui le composent, etc. Isabelle [1992a] remarque qu'une segmentation adéquate doit permettre de décrire des correspondances complexes réunissant plusieurs groupes d'éléments de différents niveaux. Il suffit de prendre des exemples de lexèmes anglais qui se traduisent en français par trois ou quatre lexèmes et inversement :

français	**anglais**
base de données *note de bas de page* *pomme de terre*	*database* *footnote* *potato*
brun *embouteillage* *purée*	*dark-haired man* *traffic jam* *mashed potatoes*

Toutefois, dans les syntagmes suivants la décomposition hiérarchique assimilée à la fonction d'analyse syntaxique pose des problèmes sur le plan sémantique :

français	**anglais**
un régime politique véritablement démocratique	*an effective political democracy*
porter atteinte au droit que possèdent les Etats	*to impair the right of a State*

Source : corpus ***Convention***

Les structures lexicales qui révèlent des équivalences *idiomatiques* ou *terminologiques*, les *collocations*, les *clichés* ou *lieux communs*, les *stéréotypes* de situation, etc. se prêtent particulièrement mal au découpage et sont à traiter « en bloc » séparément dans chaque volet du corpus (cf. *Tableau 2.1*) :

anglais	**russe**
Joan repeated the conversation word for word.	*Джоан слово в слово повторила услышанный разговор*[22]*.*

Source : Kunin [1984, p. 836]
(English-Russian Phraseological Dictionnary)

La formalisation de la hiérarchie des correspondances est également difficile dans le cas des composants discontinus :

français	**anglais**
L'avion décolle dans un instant.	*The plane is taking off in a moment.*
Le garçon n'a pas mis son manteau.	*The boy has not put on his coat.*

Dans certains cas, on peut rencontrer des correspondances *un-pour-trois/quatre* au niveau des mots et des correspondances *un mot en anglais ≈ un syntagme en français* au niveau de la hiérarchie des équivalences.

Lorsqu'une phrase est traduite dans une autre langue par un groupe de phrases, l'appariement au niveau des mots et des syntagmes est encore plus aléatoire :

[22] Джоан слово в слово повторила услышанный разговор (russe) ~ litt.: *Joana a répété mot à mot la conversation entendue.*

français	**anglais**
Développer ainsi une jurisprudence au coup par coup aboutit nécessairement à des incertitudes concernant aussi bien la portée exacte de chaque arrêt que le contenu précis de la jurisprudence de la Cour,	*A case-law that is developed on a strict case-by-case basis necessarily leads to uncertainty as to both the exact purport of each judgment and the precise contents of the Court's doctrine.*
d'où la nécessité d'ajouter des commentaires et la possibilité pour des arrêtistes de spéculer, ce qui ajoute encore à l'incertitude.	*Hence the need for comments.*

Source : corpus ***Convention***

Les tableaux ci-dessous complètent la description des difficultés rencontrées dans la formalisation de l'appariement bi-textuel au niveau lexical. Nous y présentons des exemples issus du corpus ***Convention*** qui montrent que l'automatisation intégrale de l'alignement lexical pose de nombreux problèmes dus à la complexité de rapports de correspondances traductionnelles entre les textes source et cible (cf. *Tableaux 2.1-2*).

Tableau 2.1 : Segments de traduction pour lesquels la décomposition formelle parallèle est impossible

Unités de traduction	**Langues**	**Exemples**[23]		
locutions, expressions figées	*anglais / russe*	*цыплят по осени считают*[24]		*don't count your chickens before they are hatched*
	français / anglais	*lapsus*		*slip of the tongue*
stéréotypes de situation	*russe / français*	*Говорит Москва!* [25]		*Ici Moscou*
	russe / français / anglais	*Есть!*	*A vos ordres !*	*Sir, yes, sir!*
termes	*russe / français / anglais*	*продажа в нагрузку*[26]	*vente jumelée*	*tie-in sale*
	russe / français / anglais	*осуществлять сбыт на мировом рынке*[27]	*internationaliser les ventes*	*to sell worldwide*

[23] Les ouvrages suivants ont contribués à l'élaboration de la liste d'exemples : [Miniyar-Belorouchev, 1996] ; [Kunin, 1984] ; [Gavrichina *et al.*, 1993].

[24] цыплят по осени считают (russe) ~ litt. : *on compte les poussins en automne*.

[25] Говорит Москва! (russe) ~ litt. : *Moscou parle*.

[26] продажа в нагрузку ~ litt. : *vente en charge*.

[27] осуществлять сбыт на мировом рынке (russe) ~ litt. : *réaliser l'écoulement sur le marché mondial*.

Tableau 2.2 : Corpus ***Convention*.** Difficultés d'appariement au niveau des mots et des syntagmes

Types de difficulté	Langues	Exemples
l'omission ou l'ajout de segments de traduction	*français*	*Les **paragraphes** qui suivent se rapportent au droit tel qu'il se présentait à l'époque des faits incriminés.*
	anglais	*The following relates to the law as it stood at the time of the events complained of.*
	français	*Vous ne manipulez pas de l'argent de Monopoly, alignant les zéros simplement parce qu'ils font bon effet.*
	anglais	*You do not deal in Mickey Mouse money just reeling off noughts because they sound good, **I know you will not**.*
recouvrement sémantique total	*français*	*Cette affirmation manque totalement de fondement.*
	anglais	*There is no merit in that submission.*
	français	*N'est pas considéré comme « travail forcé ou obligatoire » au sens du présent article /…/*
	anglais	*For the purpose of this article the term "forced or compulsory labour" shall not include /…/*
des correspondences discontinues	*français*	***Il lui faut** cependant **se souvenir** que le revers de la médaille est que les autorités nationales sont obligées de rechercher des directives dans sa jurisprudence.*
	anglais	***It should**, however, **not overlook** that the reverse side of this coin is that national authorities are obliged to seek guidance in its case-law.*
recatégorisation	*français*	*Après **ce remplacement**, ils continuent de connaître des affaires dont ils **sont** déjà **saisis**.*
	anglais	*After **having been replaced**, they shall continue to deal with such cases as they already **have under consideration**.*

Les unités de traduction et l'alignement

L'étude des mécanismes de la traduction humaine et l'analyse des propriétés de la notion d'équivalence traductionnelle sont fondamentales pour la recherche d'éléments formels qui se correspondent dans les textes source et cible. Au cours de son développement, la traductologie a développé un large éventail de méthodes qui permettent de repérer des *unités d'équivalence de traduction* [28]. Les questions liées à la définition et la systématisation des unités de traduction (UT) constituent l'objet central de cette discipline *: « Le concept d'unité est capital en traductologie car il repose sur les notions d'équivalence, d'analyse et de hiérarchisation des processus. Il est à la fois un instrument d'observation et d'analyse dans la pratique du commentaire de traduction et un moyen de rapporter à une grille structurée et évolutive les résultats dégagés grâce à ces investigations »* [Ballard, 1993, p. 260] [29].

Les différentes approches du repérage des unités de traduction révèlent des problèmes liés à la définition de la taille de l'UT et sa situation dans la hiérarchie des unités du texte [30]. Faut-il opérer le découpage en UT au niveau des morphèmes, des mots, des syntagmes, des phrases ? Le niveau d'analyse exige une très grande flexibilité. Il est conditionné par le type de discours, le genre de texte, le niveau et le registre de langue et le style de l'auteur [31]. Dans le domaine de l'alignement automatique, le côté pratique résultant de l'exploitation des

---

[28] Voir, par exemple, [Seleskovitch, 1968], [Mounin, 1976], [Barkhudarov, 1993], [Bennett, 1994], [Minyar-Belorouchev, 1996], [Vegliante, 1996].

[29] Un survol de la littérature consacrée à ce sujet montre qu'il s'agit d'un concept d'une grande complexité et il n'existe pas d'unanimité en ce qui concerne la définition ou la typologie de l'UT. Les principales tendances de la théorie de la traduction et des problèmes associés au concept d'unité de traduction sont présentés, par exemple, dans [Mounin, 1976] ; [Ballard, 1993], [Vegliante, 1996], [Miniyar-Belorouchev, 1996] ; [Janicijevic, 1997]. Les ouvrages montrent la complexité de délimitation des unités de traduction et synthétisent les acquis des recherches précédentes.

[30] On consultera [Ballard, 1993, pp. 223-245] ; [Miniyar-Belorouchev, 1996, pp. 76-90].

[31] Deux approches théoriques co-existent au sein de l'analyse de la traduction : la première concerne notamment les théories *formalistes*, généralement liées à la didactique de la traduction ; la deuxième approche est liée notamment aux études interprétatives et communicatives des procédés de traduction.

données bi-textuelles stockées sur support informatique entraîne une préférence pour les unités de petites tailles qui ne s'imposent pas forcément au plan théorique [32].

La notion d'unité de traduction implique la possibilité de décider d'une équivalence de sens : *« /.../ les deux textes devraient produire des effets similaires chez les lecteurs de chacune des langues concernées. »* [Isabelle et Warwick-Armstrong, 1993, p. 290]. L'équivalence est décomposable *« /.../ en une mise en équation de signifiant de deux codes avec pour objet l'estimation de l'analogie de leurs effets de sens »* [Ballard, 1993, p. 252]. Cette définition implique le recours à des modèles sémantiques permettant un plus grand degré d'abstraction dans l'analyse des unités de traduction. Considérons le couple de phrases suivantes (français/anglais) [33] :

français	**anglais**
/ Tant devant les autorités nationals /	/ The applicant had made formal undertakings not to abscond, /
/ qu'au moment du dépôt de sa requête à la Commission, /	/ both before the national authorities /
/ le requérant avait formellement promis de ne pas se soustraire à la justice. /	/ and at the time of filing his application with the Commission. /

Source : corpus ***Convention***

Le découpage que nous avons opéré établit des « unités de sens » qui rentrent dans la construction de l'espace sémantique de la traduction. Bien que ce découpage soit acceptable du point de vue de l'équivalence du sens, il n'est pas

[32] Comme le note Janicijevic [1997], il faut éviter le morcellement inutile de textes source et cible. Si les textes alignés sont découpés en unités trop petites, la machine ne sera pas capable de repérer correctement les traductions ; par contre, si l'unité de traduction est trop longue, nous devrions nous attendre à un niveau de bruit très élevé.

[33] Des *théories interprétatives* de l'analyse de la traduction (telles que celle de Danica Séleskovitch) mettent en avance le concept d'*unité de compréhension*. Séleskovitch insiste sur une déverbalisation, sur un oubli volontaire du signifiant par le traducteur pour ne retenir que l'image du signifié qui sera « interprétée » librement dans la langue d'arrivée, cf. Seleskovitch, [1968, p. 35].

suffisamment précis pour mettre en valeur les structures lexicales susceptibles d'intéresser le traducteur :

français	**anglais**
requérant	*applicant*
se soustraire à la justice	*to abscond*
déposer la requête	*to file one's application*
promettre formellement	*to make formal undertakings*

Source : corpus ***Convention***

On en conclut que l'identification des unités constituantes du bi-texte implique à la fois une cohérence dans la décomposition de la traduction en unités structurées du point de vue de leur forme linguistique et de leur contenu (compréhension), et une nécessité de mettre en valeur des équivalences originales, susceptibles d'enrichir les nouvelles traductions [34].

Il serait donc difficile de proposer un modèle unique de segmentation parallèle de textes qui rende compte de la majorité de correspondances et fournisse des représentations sémantiques et syntaxiques objectives [35]. Il serait encore plus difficile de trouver des mécanismes formels permettant d'automatiser cette segmentation.

L'une des solutions pour spécifier les correspondances consiste à faire varier la *résolution* [36]. L'alignement à basse résolution montre davantage les correspondances entre les unités du sommet de la hiérarchie (paragraphes, phrases, etc.). A l'inverse, l'alignement à haute résolution mettra en évidence les correspondances entre unités plus petites (syntagmes, mots, etc.). Cependant, la

[34] Les problèmes de délimitation des unités de traduction à des fins d'extraction d'équivalences réutilisables sont soulevés par Kraif [2002, p. 4] qui remarque que « /…/ *d'un côté, les corpus de traductions recèlent un grand nombre d'équivalences, intéressantes à la fois d'un point de vue contrastif et traductionnel, mais diffuses et difficiles à isoler de leur contexte ; de l'autre côté, dans une perspective dictionnairique, on voudrait des unités bien délimitées, transcodables et équivalentes sur le plan de leur valeur (au sens saussurien).* »

[35] Sur l'utilisation des modèles de traduction dans le cadre de la mise en correspondance automatique de textes source et cible, [Santos, 2000].

[36] La *résolution* mesure le degré de précision de l'alignement que l'on peut varier pour mettre en évidence les correspondances entre les unités textuelles des différents niveaux. Pour une résolution donnée, un alignement correct est maximum, s'il est composé du plus petit des couples de segments possible. On consultera sur ces questions [Isabelle et Simard, 1996].

variation du taux de résolution de l'alignement ne permet pas de mettre directement en évidence des correspondances lexicales réutilisables, qui font la richesse du bi-texte du point de vue de son exploitation ultérieure. Dans cette perspective, la recherche sur des méthodes d'extraction automatique de *lexiques bilingues* à base de corpus parallèles représente une voie prometteuse qui permet de contourner les difficultés de l'alignement intégral de segments de traduction au niveau des mots et des syntagmes.

Nous avons essayé de faire un bref relevé des problèmes nés dans le contexte du domaine de l'alignement automatique de corpus parallèles [37]. Dans les sections qui suivent, nous présenterons des solutions techniques pour l'appariement du matériau bi-textuel [38].

2.2 L'alignement automatique des phrases

Les travaux pionniers dans l'alignement ont concerné l'appariement des corpus parallèles au niveau de la phrase. Parmi les premiers algorithmes proposés on citera Gale et Church [1991b] ; Brown, Lai et Mercer [1991] ; Kay et Röscheisen [1993].

2.2.1 L'alignement par longueurs de segments

Les méthodes proposées par Gale et Church ou Brown, Lai et Mercer reposent sur un certain nombre d'hypothèses communes qui permettent de les considérer comme un même type d'alignement que l'on appellera ici *alignement par longueurs des segments* [39]. Ce type d'alignement fait appel à un ensemble de règles opératoires fondées sur l'observation de corrélations entre la longueur

[37] Sur ces questions, on consultera, par exemple, [Kraif, 2001], disponible sur : http://www.u-grenoble3.fr/kraif/publis/these.pdf.

[38] Sur la description de principaux algorithmes d'alignement, voir [Klevbacke, 2001] : http://www.dtek.chalmers.se/~d95ankle/algorithms-ac-project.html.

[39] Voir [Isabelle et Warwick-Armstrong, 1993, pp. 294-296].

d'un segment source et celle de sa traduction. Les auteurs partent de la constatation que la longueur des phrases dans le texte source et celles de leur traduction dans le texte cible sont fortement corrélées et qu'il existe un rapport assez constant entre ces longueurs d'une langue à l'autre [40].

Les principales différences entre les deux familles d'algorithmes résident dans la métrique employée pour calculer la longueur des phrases. Brown, Lai et Mercer fondent leur algorithme sur le nombre des mots dans chaque phrase [41]. Pour chaque mot dans la phrase du texte cible, on choisit un unique mot source. L'appariement est unidirectionnel : les calculs de correspondance ne fonctionnent que dans le sens source-cible. De ce fait, certains mots du texte cible ne sont pas pris en compte et restent sans correspondance directe dans le texte source. Si l'algorithme ne trouve pas de correspondance pour un mot source, ce mot est lié à un mot vide et reçoit une correspondance *zéro*. Gale et Church estiment que les données en nombre de caractères conviennent d'avantage pour l'alignement parce que les caractères sont plus nombreux et constituent une donnée plus stable.

Les deux méthodes ne font aucune hypothèse directe sur le contenu lexical des phrases à apparier et demeurent dans le cadre de l'alignement statistique. Les textes bilingues soumis au traitement sont d'abord alignés au niveau des paragraphes. Le calcul des correspondances de paragraphes est fait soit par un simple appariement un-pour-un, soit par l'utilisation du même algorithme

---

[40] L'étude de corpus de textes bilingues effectuée dans le cadre du projet ARCADE a montré, par exemple, qu'il existe un rapport relativement constant entre la longueur des textes français mesurée en nombre de caractères et leurs équivalents anglais. Les textes français sont généralement plus longs, cf. Langlais *et al.* [1998] ; Véronis et Langlais [2000].

[41] Plusieurs algorithmes d'alignement nécessitent une segmentation préalable des phrases en mots. Cette information est requise pour l'implémentation des mécanismes de recherche axés sur les mots plutôt que sur les chaînes de caractères. La segmentation en mots est faite en utilisant les frontières habituelles (espace, retour chariot, ponctuation), aussi bien que des règles spécifiques à chaque langue pour segmenter des chaînes de caractères qui sont des agrégats de plusieurs mots. Par exemple : *jusqu'alors -> jusqu'* + *alors* ; women's rights -> *women* + s' + *rights* (le caractère « + » marque l'endroit où la segmentation a été faite). Ce repérage est très délicat lorsqu'il s'agit d'un certain nombre de caractères qui fonctionnent soit comme séparateurs, soit comme composants de mots (le trait d'union, par exemple) : *dis-le* (valeur grammaticale) et *casse-tête* (unité polylexicale).

d'alignement par longueur des segments. A l'intérieur des couples de paragraphes ainsi obtenus, on procède à l'alignement des phrases. Une probabilité est ensuite attribuée à chaque paire de phrases proposée pour l'appariement. Cette valeur est utilisée par un algorithme de *programmation dynamique* [42] pour trouver une paire de phrases dont les caractéristiques sont potentiellement appropriées pour l'alignement [43]. Les types d'alignement admis sont limités à des schèmes de traduction suivants décrits ci-dessous :

- une phrase d'un texte source est traduite par une phrase dans un texte cible ;
- deux phrases consécutives se traduisent par une phrase ;
- une phrase se traduit par deux phrases qui se suivent ;
- deux phrases consécutives se traduisent par deux phrases qui se suivent ;
- une phrase d'un texte source reste sans traduction ;
- une phrase sans équivalent dans un texte source est introduite par le traducteur.

Les algorithmes d'alignement par longueur respectent deux limitations :

- conserver l'ordre des segments

a) correspondances admises

```
S1 – – –  C1
S2 – – –  C2
```

b) correspondances rejetées

- conserver les correspondances plusieurs-pour-plusieurs à condition qu'elles ne dépassent pas un petit nombre (généralement de deux) [44].

[42] La *programmation dynamique* est une méthode de résolution qui permet de déterminer une solution optimale d'un problème à partir des solutions de tous les sous-problèmes.

[43] Chaque alignement possible reçoit un score qui reflète la qualité des corrélations de longueur qu'il contient. Un bon score dépend des conditions suivantes : 1) les longueurs des phrases ont une bonne corrélation ; 2) l'alignement donne une bonne résolution (le mécanisme de pointage permet de pénaliser un alignement qui réduit la résolution).

[44] Voir [Gale et Church, 1991b].

Les premiers succès dans l'alignement de phrases de corpus bilingues ont suscité des expérimentations plus ambitieuses sur des corpus de textes parallèles composés de plusieurs volets multilingues [45]. Simard [2000] a développé une approche originale d'alignement qui repose sur l'utilisation de multiples versions d'un même texte pour obtenir une meilleure estimation de « similarité » de séquences textuelles dans des langues différentes [46].

2.2.2 L'alignement par mots apparentés

Les algorithmes d'alignement fondés sur la longueur des segments n'utilisent presque pas de connaissances linguistiques. La création d'un bi-texte s'appuie exclusivement sur le calcul des *octets* [47] d'information. Cette méthode est techniquement simple et économique sur le plan des calculs. Cependant, les résultats sont imparfaits. Simard, Foster et Isabelle [1992] proposent d'éliminer certains cas d'échec en ajoutant des procédures de *resynchronisation* d'un bi-texte. Cet algorithme reprend le mécanisme d'alignement par longueur et rajoute un paramètre supplémentaire à caractère auxiliaire, celui de la ressemblance de chaînes de caractères.

Lorsqu'une erreur survient lors du processus d'appariement, les algorithmes d'alignement par longueur ne sont pas suffisamment robustes pour en tenir compte et trouver des indices de resynchronisation. On risque alors d'avoir, comme résultat, un décalage qui provoque un alignement erroné depuis la source d'erreur jusqu'à un endroit situé beaucoup plus bas dans le texte.

---

[45] Voir, par exemple, le corpus MULTEXT : http://www.lpl.univ-aix.fr/projects/multext.

[46] Cette approche est inspirée par des méthodes d'alignement de séquences de nucléotides utilisées en biologie moléculaire, voir, par exemple, [Jennings *et al.*, 2001]. Les recherches dans ce domaine ont montré que l'augmentation du nombre de séquences similaires comparées permet d'augmenter la qualité des appariements obtenus. Dans le cas d'alignement textuel multilingue, le même principe permet de privilégier un appariement le plus optimal en tenant compte d'un plus grand nombre de probabilités combinatoires au cours du calcul statistique. Sur les détails du calcul, [Simard, 2000, pp. 53-60] ; [Klevbacke, 2001, pp. 7-9].

[47] L'*octet* (en anglais *byte*) est une unité d'information composée de 8 bits. Il permet de stocker un caractère, telle qu'une lettre ou un chiffre. En langage informatique, une unité d'information composée de 16 bits est généralement appelée *mot* (en anglais *word*).

Pour tenir compte de l'erreur et la corriger, il est possible d'utiliser les *mots apparentés* (*cognats*). Le terme *cognat* est une traduction (pas vraiment réussie) du mot anglais *cognate*. Les cognats sont des mots étymologiquement reliés, tels que les groupes de chiffres (1789, 1515), les noms propres préservés dans la traduction, etc. [48]. Selon l'hypothèse de Simard, Foster et Isabelle, un couple de phrases *S* et *C* qui sont en correspondance de traduction, possède plus d'indices apparentés qu'un couple de phrases choisies au hasard.

Notons que la détection automatique des mots apparentés au cours de l'alignement ne fait aucun appel à des connaissances linguistiques *a priori*, ni même aux données construites à partir de corpus. Le peu de connaissances linguistiques utilisées concerne le contenu des unités textuelles à apparier et fonctionne sans aucun appel aux ressources extérieures. Le modèle de Simard, Foster et Isabelle contribue à l'alignement par longueur de segments et propose une modification opératoire permettant d'utiliser le critère de ressemblance des mots [49].

D'autres chercheurs se sont penchés sur les questions de l'utilisation des mots apparentés dans la mise en correspondance de segments de textes source et cible. A la suite du travail de Simard, Foster et Isabelle, Church [1993] a montré l'importance de méthodes à base de mots apparentés dans l'alignement de corpus bilingues présentant certaines divergences au niveau structurel (omissions, insertions d'éléments graphiques, absence de cohérence dans le découpage en

[48] De manière informelle, les mots apparentés sont des mots qui partagent des propriétés phonologiques, orthographiques et sémantiques facilement repérables. Ces mots représentent souvent des traductions mutuelles. Parmi les exemples types pour l'anglais et le français, des mots comme *comprehension/compréhension, text/texte*, etc. Logiquement, les noms propres sont souvent dans cette catégorie (*Paris/Paris*, *London/Londres, Russia/Russie*), ainsi que les expressions numériques, et même parfois des signes de ponctuation (le point d'interrogation, les parenthèses, etc.).

[49] Généralement, l'alignement par mots apparentés est utilisé comme complément d'algorithme d'alignement par longueur des segments. Les calculs de mots apparentés permettent d'améliorer les résultats, à condition que les liens de parenté entre les deux langues soient relativement forts. Employé seul, ce modèle est moins efficace que l'alignement par longueur. Il est aussi plus coûteux en calcul.

paragraphes, etc.) [50]. Globalement, les travaux actuels montrent qu'il est possible de rapprocher les mots non seulement lorsqu'ils commencent par des chaînes de caractères identiques *(« truncation » method*) mais aussi à base de calcul de *n-grammes* [51] de caractères partagés [McEnery et Oakes, 1995], [Brew et McKelvie, 1996]. La similarité des deux mots en termes de caractères partagés peut être calculée en utilisant le coefficient de Dice [1945], correspondant au double du rapport du nombre de lettres communes à la somme des lettres des deux mots [52].

Kondrak [2001] a développé des mesures de similarité permettant la détection de mots apparentés au niveau phonétique. Ces études montrent que la détection de cognats par des méthodes faisant appel à des traits phonologiques des mots est plus efficace que l'approche orthographique**.** Les cognats phonétiques sont particulièrement utiles lorsqu'il s'agit de langues qui emploient deux alphabets différents (le français et le russe, par exemple).

Actuellement, les techniques de détection de correspondances entre cognats phonétiques et orthographiques font partie intégrante de plusieurs algorithmes d'alignement [53]. Ces techniques permettent d'obtenir des points d'ancrage lexicaux dans les textes pour le calcul automatique d'une table de correspondances (*bi-text mapping*) [Melamed, 1999, p. 108]. Dans ce type de systèmes, les mots apparentés sont considérés comme des *îlots de confiance* permettent d'apparier les segments de texte qui gravitent autour d'eux (cf. *Figure 2.3*) [54].

---

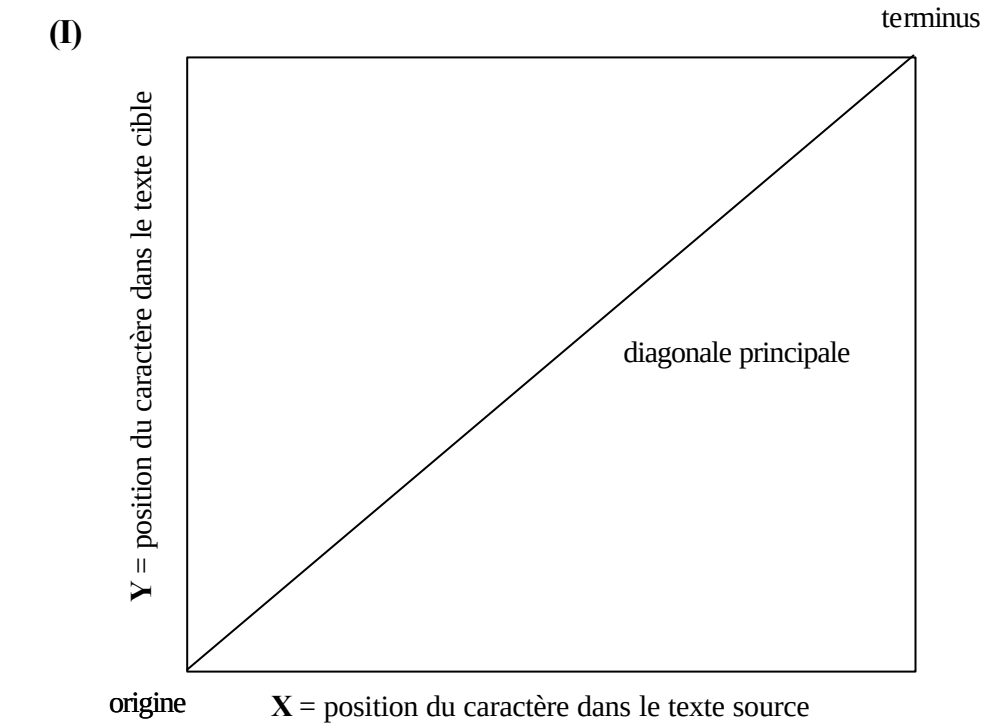
[50] Le programme *Char_align* [Church, 1993] est un exemple d'implémentation pratique de principes d'alignement par mots apparentés qui met en œuvre ces principes généraux.

[51] A la base, dans le contexte multilingue, le calcul de *n-gramme* de caractères a été proposé pour identifier automatiquement la langue dans laquelle un document est écrit. Cette méthode trouve actuellement de nombreuses applications notamment dans la recherche d'information multilingue, voir [Tauritz et Sprinkhuizen-Kuyper, 2000].

[52] Sur l'utilisation du coefficient de Dice dans l'alignement bi-textuel, voir, par exemple, [Hofland et Johansson, 1998].

[53] Voir Johansson et Hofland [1994], Melamed [1997 ; 1998 ; 2000ab], Resnik et Melamed [1997], Ahrenberg *et al.* [2000], Romary et Bonhomme [2000], Kraif [2001].

[54] Sur ces questions, on consultera Melamed [2000ab].



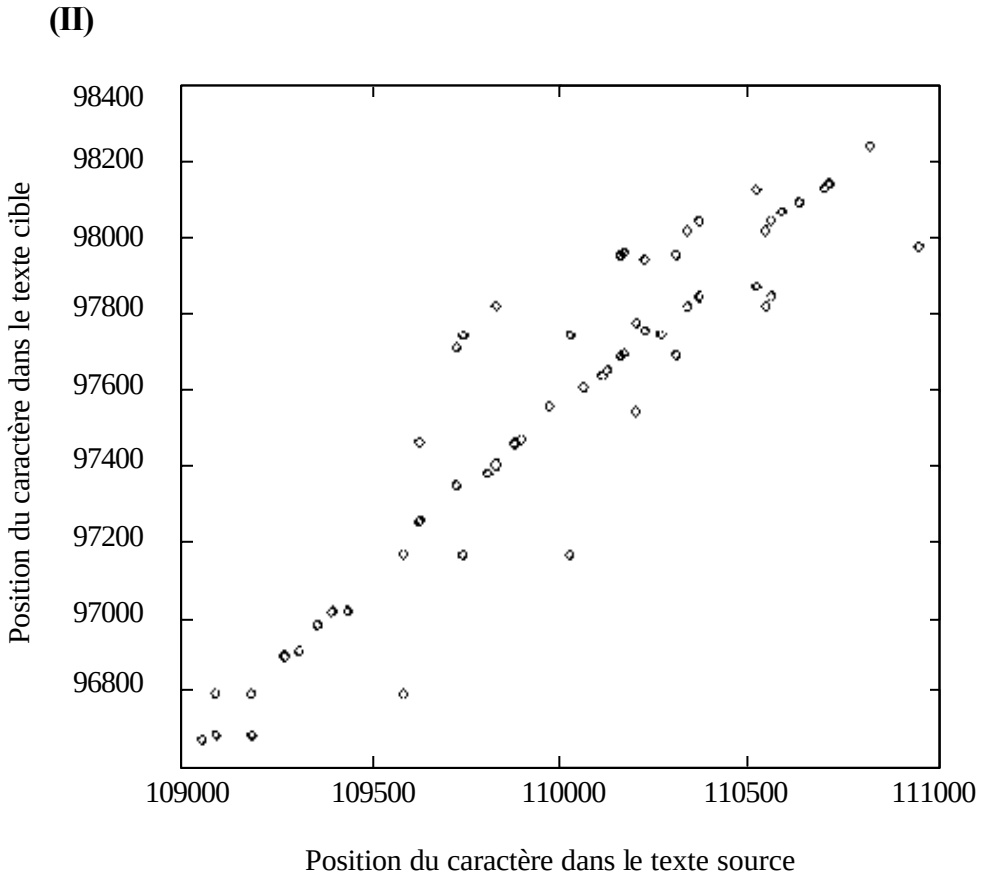


Figure 2.3 :
Modèle de calcul d'une table de correspondances dans l'espace bi-textuel (*bi-text mapping*) développé dans les travaux de Melamed [1999 ; 2000ab]

Guide de lecture de la figure 2.3 :

(I) On considère qu'un bi-texte est défini par un *espace bi-textuel* commun formé par un couple de textes bilingues. Les axes *x*, *y* représentent les source et cible. Les longueurs des axes sont proportionnels aux longueurs respectives des deux textes calculés, généralement, en nombre de caractères. Les points d'ancrage préalables (les mots apparentés, les indices structurels de découpage en phrases, sections, etc.) sont indiqués par un système de coordonnées *x*, *y*.

(II) Les expérimentations menées par Melamed [1999 ; 2000ab] ont permis de constater que la pente générée par le chemin d'alignement dans l'espace bi-textuel est proche de la diagonale *x=y*. Les écarts par rapport à ce schéma concernent notamment les omissions de larges segments de texte dans un des volets du corpus. Compte tenu de cette observation, l'algorithme de Melamed permet de calculer le chemin optimal d'alignement en tenant compte du bruit dû aux omissions et à la discontinuité.

2.2.3 L'alignement par correspondances de mots

Dans la section précédente nous avons montré que les *mots apparentés* (les traductions mutuelles dont la stabilité est relativement élevée) se révèlent utiles pour l'alignement. Cette approche peut être vue comme une étape vers l'utilisation d'un autre critère d'alignement basé sur les *correspondances de mots* dans les phrases liées sur le plan de la traduction.

Historiquement, l'alignement par correspondance de mots est né pratiquement en même temps que l'alignement par longueur de segments. Les deux algorithmes possèdent un certain nombre de points en commun. Ils utilisent des points d'ancrage dans le corpus pour apparier le matériau textuel qui se situe autour, cf. Somers [2001]. La différence principale entre les deux algorithmes réside dans la nature de ces points d'ancrage. L'alignement par longueur privilégie des indices structurels (le découpage en sections, paragraphes, etc.), tandis que l'alignement par correspondance de mots fait appel aux propriétés distributionnelles des unités lexicales.

L'alignement par correspondances de mots pose comme hypothèse temporaire, et pas toujours régulière que lorsque deux phrases source possèdent des mots en

commun, leurs traductions respectives tendent aussi à posséder des mots en commun. Cette propriété est liée directement à la notion de *compositionnalité de traduction* (cf. section 2.1.3). Elle montre que l'on peut exploiter des régularités statistiques dans la distribution des mots en correspondance à l'intérieur des phrases pour déterminer ensuite l'appariement adéquat des ces dernières.

Une première description détaillée de l'algorithme d'alignement par correspondances de mots a été donnée par Kay et Röscheisen [1993]. Les auteurs soulignent la difficulté d'obtenir une mise en correspondance précise entre les mots des deux textes mais considèrent que leur appariement, même imparfait, est susceptible de conduire à un alignement satisfaisant au niveau des phrases. Le calcul est effectué en deux étapes :

(1) On détermine d'abord un ensemble de paires de phrases candidates à l'appariement. On décide que les premières phrases de chaque texte se correspondent ainsi que les dernières. Respectivement, les phrases intermédiaires sont en correspondance dans un couloir diagonal relativement étroit (cf. Figure 2.3).

(2) Pour chaque phrase du texte source on calcule l'ensemble de phrases qui peuvent lui correspondre dans le texte cible, en respectant les contraintes suivantes (les mêmes que pour l'algorithme d'alignement par longueur des segments) :

- les correspondances croisées sont interdites ;
- le nombre maximal de correspondances plusieurs-pour-plusieurs est restreint.

Exemple : Supposons que S_1, S_2, S_3 … S_n et T_1, T_2, T_3 … T_m sont des phrases respectives de textes source et cible. Selon les contraintes retenues, S_1 pourra correspondre soit à T_1, soit à la combinaison de T_1,–T_2, mais ne peut être liée à T_2 toute seule. Lorsque l'on s'approche du milieu du texte, le nombre d'appariements possibles augmente.

(3) On procède à l'analyse des correspondances lexicales compatibles avec l'alignement initial des phrases. L'algorithme compare les distributions des mots à l'intérieur de l'ensemble de phrases appariées. Pour chaque couple de mots, on procède à la vérification à partir d'un seuil préétabli. Si les distributions dans les phrases sont proches, on considère que ces mots sont en rapport de traduction[55].

Exemple : Si le mot X apparaît n fois dans le texte cible, et le mot $Y - n$ fois dans le texte source, on teste l'existence d'un appariement associant chaque phrase qui contient X avec une phrase qui contient Y. On fait l'hypothèse que et Y sont des correspondances de traduction.

Les deux étapes sont répétées de manière itérative, pour assurer un échange de données entre les deux calculs, et la convergence vers une solution optimale. Les mots fournissent alors un ensemble de points d'ancrage qui permettent de réduire le couloir diagonal des alignements de phrases candidates. Un alignement complet est généralement obtenu en 3-4 cycles. Notons que la méthode de Kay et Röscheisen met en valeur un principe de *dépendance réciproque dans l'alignement des phrases et des mots.* On retrouve ce principe à la base d'un grand nombre de systèmes d'alignement automatique.

L'algorithme proposé par Chen [1993] repose sur un certain nombre d'hypothèses proches de celles élaborées par Brown P. *et al.* [1991] et Kay et Roschëisen [1993]. La principale originalité réside dans ce cas dans l'utilisation plus importante des informations lexicales qui permettent d'augmenter la précision, notamment lorsqu'il s'agit de corpus parallèles avec un taux de bruit élevé (omissions, différences structurelles, etc.).

[55] Pour un couple de mots, la mesure de similarité distributionnelle est donnée par le calcul du coefficient de Dice [1945] : $2c / (N_A(u) + N_B(w))$, où c est le nombre de correspondances de mots trouvées dans les phrases appariées, $N_T(x)$ - le nombre d'occurrences du mot X dans le texte T.

L'ancrage lexical est central dans la méthode *DK-vec* (*Dynamic K-vec*[56]) [Fung et McKeown, 1994]. L'algorithme reçoit en entrée deux textes parallèles segmentés en occurrences (*tokens*) de formes graphiques ou lemmes (le choix des unités minimales de décomptes dépend des objectifs de l'expérimentation). Pour chaque mot *w* de ces textes, la méthode permet de calculer un *vecteur de distance* $D^w=<d_1^w,\ldots,d_n^w>$, qui représente les distances relatives exprimées en nombre d'occurrences consécutives du mot *w*. La notation d_i^w correspond à la distance relative entre l'occurrence du mot *w* dans la position *i* et son occurrence précédente dans le texte. Notons que d_1^w marque la distance entre la première occurrence de *w* et le début du fichier [Choueka *et al.*, 2000]. Selon l'hypothèse de départ, les fragments de textes en correspondance de traduction ont des longueurs similaires, d'où les similitudes dans les valeurs de distance associées aux positions des occurrences de mots liés sur le plan de la traduction. En conséquence, les vecteurs des mots en correspondance se ressemblent beaucoup plus que ceux des mots qui ne se correspondent pas [57].

---

[56] Dans sa première version, l'algorithme est connu sous le nom de *K-vec* [Fung et Church, 1994]. Les distributions lexicales sont représentées par des *vecteurs binaires* de présence/absence des occurrences des mots au travers des textes bilingues, divisés parallèlement en *K*-fragments de même longueur. Ainsi, pour le mot *w* présent dans le 3ème, 5ème et le 8ème fragment du texte divisé en 10 fragments (*K*=10), on note : V_w =<0,0,1,0,1,0,0,1,0,0>. Les vecteurs V_w et $V_{w'}$ sont comparés au cours d'un calcul probabiliste de similarité qui utilise l'information sur la présence/absence réciproque des mots bilingues *w* et *w'* dans les fragments de texte en correspondance (*mutual information similarity metrics*). L'algorithme a démontré que les informations fréquentielles et positionnelles obtenues sur les mots du corpus sans appel à des connaissances *a priori*, peuvent être suffisantes pour l'appariement lexical. Dans la version ultérieure de l'algorithme (*Dynamic K-Vec*), les informations positionnelles utilisées sont complétées par un calcul de distance entre les occurrences consécutives [Fung et McKeown, 1994].

[57] Plusieurs mesures statistiques de similarité ont été proposées pour porter une estimation sur les ressemblances des vecteurs, cf. Ahrenberg *et al.* [2000] ; Choueka *et al.* [2000] ; Fung, [2000] ; Fung et McKeown [1994] ; Jones et Somers [1995] ; Somers [2001]. La mesure de distorsion temporelle dynamique (*Dynamic Time Warping Score*) utilisée par Fung et McKeown permet de déterminer la paire de mots les plus proches en utilisant la distance normale à la droite *x=y* avec un point d'abscisse la position dans le texte *S* et pour ordonnée la position dans le texte *C*. Le filtrage sur la fréquence générale est utilisé pour réduire le nombre de mots - candidats à l'appariement.

2.2.4 L'alignement à l'aide de dictionnaires bilingues

Des méthodes d'alignement similaires à celles de Kay et Röscheisen ont été développées par deux autres groupes de chercheurs, celui de Catizone, Russel et Warwick [1989] et celui de Debili et Sammouda [1992]. La différence principale est que ces nouvelles méthodes font appel à un dictionnaire bilingue : *« /.../ la motivation est de simplifier la tâche en utilisant des connaissances linguistiques a priori. Au lieu de chercher les corrélations statistiques à travers toutes les paires formées par les mots contenus dans les phrases candidates à l'alignement, on restreint l'attention aux paires de mots qui figurent dans le dictionnaire bilingue. La combinatoire des possibilités est alors considérablement réduite »* [Isabelle et Warwick-Armstrong, 1993, p. 299].

Malgré quelques différences, la méthode de Debili et Sammouda est fortement influencée par celle de Kay et Röscheisen. Les auteurs proposent d'intégrer des connaissances linguistiques pour pouvoir non seulement aligner deux phrases, mais aussi les comparer. Ils soulignent que l'approche statistique et l'approche linguistique proprement dite possèdent des potentiels différents pour l'alignement : *« L'approche statistique ne pose que le problème de l'appariement des phrases, et ne peut répondre lorsque l'on a à comparer que deux phrases. L'approche linguistique, qui s'inspire de ce que nous ferions nous-mêmes intuitivement, paraît plus puissante, puisqu'elle porte en elle les ingrédients qui permettraient de répondre aussi lorsque, à la comparaison, ne sont soumises que deux phrases »* [Debili *et al*., 1994, p. 6]. L'algorithme appuyé sur l'emploi d'un dictionnaire fait intervenir des relations de dépendance qui s'établissent respectivement entre les différents mots de deux phrases, pour examiner tous les appariements possibles et choisir le meilleur parmi eux. Globalement, la densité d'appariement des mots détermine l'appariement des phrases. L'alignement final est obtenu par une analyse combinant trois critères :

(1) taille similaire de deux phrases ;

(2) positions similaires des phrases dans l'ensemble du texte (enchaînement) ;

(3) similarités des mots contenus dans deux phrases (appel à un dictionnaire) [58]

L'alignement se construit à travers trois étapes :

- *construction* :
  établissement d'un maximum de liens potentiels entre les deux phrases ;
- *élimination* :
  implémentation de procédés de rapprochement syntaxique pour résoudre les ambiguï tés d'appariement d'une part, et écarter les appariements incorrects d'autre part ;
- *reconstruction* :
  augmentation de la résolution.

Lors de la première étape, l'appariement grossier des mots permet d'apparier les phrases. Considérons deux phrases S et C :

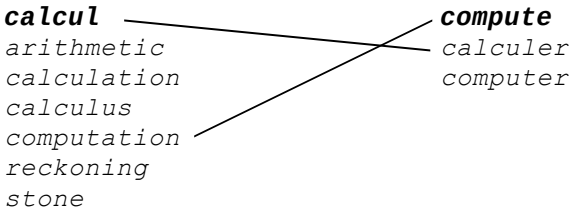
$$S = s_1, s_2 \ldots s_i \quad \ldots s_m$$
$$C = c_1, c_2 \ldots c_j \quad \ldots c_n$$

Sur la figure 2.3, la matrice associée à ce couple de phrases est obtenue par comparaison successive de chacun des mots de S à tous les mots de C. Les mots s_i et c_j appartenant respectivement à S et C se correspondent, si les conditions suivantes sont vérifiées :

- les mots des deux ensembles sont identifiés comme traductions mutuelles dans le dictionnaire ;
- les positions des mots dans les phrases respectives sont similaires ;
- les mots entretiennent le même type de relations syntaxiques avec les mots environnants.

[58] Le premier critère rappelle l'alignement par longueur de segments. La notion d'enchaînement séquentiel des unités dans la traduction est à la base du second critère. Cette notion est utilisée dans la plupart des algorithmes d'alignement des phrases. Cependant, la méthode de Debili et Sammouda [1992] ne traite pas l'enchaînement séquentiel de traduction comme un critère absolu. Les correspondances croisées des phrases sont admises dans une certaine mesure.

Exemple : On compare une par une toutes les traductions du mot français *calcul* au mot anglais *compute* et inversement [Debili *et al*., 1994, p. 9] :



L'implication des dictionnaires et des thésaurus bilingues pour l'alignement est relativement coûteuse en calcul et nécessite une lemmatisation cohérente du corpus. De plus, les ressources dictionnairiques adéquates ne sont pas toujours disponibles pour les textes issus de domaines spécialisés ou rédigés dans des langues rares.

L'algorithme d'alignement de Haruno et Yamazaki [1996] est fondé sur une utilisation combinée de méthodes statistiques et de ressources dictionnairiques. L'implication de méthodes statistiques permet d'établir les premières correspondances lexicales en corpus. Le recours au dictionnaire vient ensuite pour corriger, compléter et affiner les résultats de l'alignement. L'appel au dictionnaire rend possible la détection de correspondances de faible fréquence, tandis que l'implication de mesures statistiques permet de compenser le silence des dictionnaires lorsqu'il s'agit de termes spécifiques qui y sont absents.

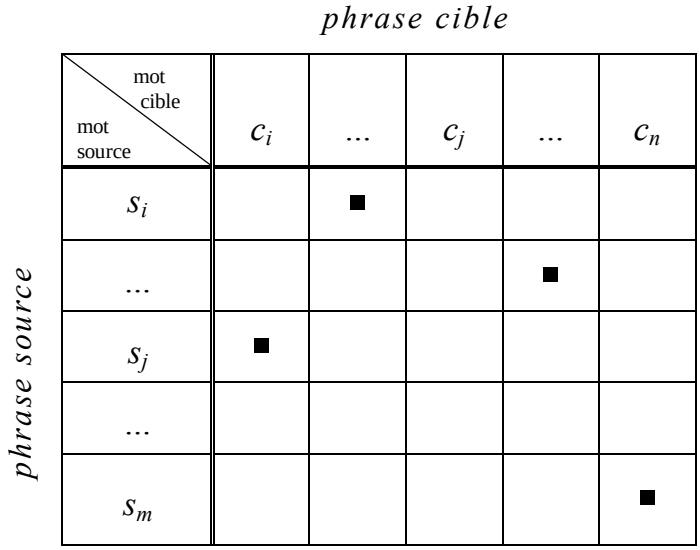


Figure 2.4 :
La comparaison matricielle des mots d'un couple de phrases lors de l'alignement à l'aide de dictionnaires bilingues [Debili et Sammouda, 1992]

***Guide de lecture* :** Lors de la recherche dictionnairique, chaque couple (s_j, c_i,) reçoit une note de proximité. Au plan algorithmique, Debili et Sammouda s'appuient notamment sur les ressemblances de paramètres des chaînes à comparer et n'examinent presque pas leurs différences.
Le calcul de la note est effectué de façon suivante : chaque traduction de s_j répertoriée dans le dictionnaire est comparée à c_i, et inversement. Le calcul de la note globale s'appuie sur la comparaison de la somme de tous les points représentant les meilleures correspondances mutuelles des mots d'un couple de phrases en question.
La mise en correspondance des mots est établie à l'aide d'un dictionnaire de transfert des mots simples. Cette technique permet d'appliquer la recherche dictionnairique non seulement aux mots entiers mais aussi aux sous-chaînes qu'ils contiennent. Elle permet de compenser le silence des dictionnaires de transfert de mots simples. Cette approche est particulièrement adaptée aux couples de langues dont les lexiques sont suffisamment proches (comme le français et l'anglais, par exemple).

2.3 L'alignement automatique des mots et des syntagmes

2.3.1 Problèmes et enjeux

L'alignement des unités qui se situent au-dessous du niveau de la phrase (syntagmes, mots, morphèmes, etc.) rend plus aisé l'accès aux corpus parallèles. Cela est particulièrement important pour améliorer les performances des applications informatisées à base de corpus alignés (concordanciers, systèmes de mémoire de traduction, etc.). Malheureusement, les correspondances entre unités textuelles qui se trouvent en bas de la hiérarchie (mots, morphèmes) sont beaucoup plus difficiles à établir même lorsqu'il s'agit de textes alignés manuellement : *« De manière générale, on observe que plus l'on descend dans la hiérarchie des unités, plus les correspondances tendent à se libérer aussi bien de l'ordre fixe que du un-pour-un. Ainsi, une traduction modifiera beaucoup plus facilement le nombre et l'ordre des mots que le nombre et l'ordre des sections. Le calcul automatique des correspondances est donc beaucoup plus simple dans le cas des unités de rang supérieur »* [Isabelle et Warwick-Armstrong, 1993, p. 291].

Les techniques d'ancrage lexical propres à l'alignement de phrases ne traitent que des occurrences isolées et ne permettent pas toujours de considérer les unités qui véhiculent le sens dans les phrases liées sur le plan de la traduction. Lors de l'alignement de ce type d'unités, il faut tenir compte de plusieurs phénomènes complexes liés, notamment, à la détection des emplois polysémiques de mots et de leur fonctionnement dans des séquences figées et locutions. Les unités à correspondances multiples posent également de nombreux problèmes notamment lorsqu'elles comportent des mots grammaticaux dont la traduction varie selon le contexte. L'*alignement lexical*, c'est-à-dire la mise en correspondance de mots et locutions entre les deux volets d'un corpus parallèle, demeure un problème difficile.

2.3.2 Développements récents

L'alignement lexical peut être décomposé en deux étapes : il s'agit d'abord de repérer les mots et expressions du texte source et du texte cible, puis de les mettre en correspondance. La force de cette approche réside dans la possibilité d'employer le savoir-faire précédemment acquis dans le traitement et l'extraction de ressources lexicales de corpus de textes monolingues. Sa principale faiblesse réside dans le fait que le repérage automatique des expressions du texte source ne peut pas se faire valablement sans une prise en compte des données de traduction dans la langue cible.

La question de la sélection des unités lexicales les mieux adaptées au traitement de corpus multilingues rejoint le débat déjà ancien sur les unités qui circulent dans le corpus de textes que l'on étudie. Comme le remarquent Lamalle et Salem [2002, p. 403], la diversité d'approches utilisées au sein de plusieurs communautés scientifiques montre qu'en l'état actuel des choses la question du choix des unités ne saurait être tranchée une fois pour toutes et pour tous les types d'études à venir. Il s'agit même d'une piste de recherche parmi les plus intéressantes et complexes dans le domaine des études textuelles automatisées.

Véronis [2000a] note que plusieurs méthodes de traitement automatique des langues sont employées actuellement à des fins d'automatisation de la sélection de mots et locutions candidats à l'appariement, cf. Ahrenberg *et al*. [2000] ; Choueka *et al.* [2000] ; Déjean *et al*. [2002] ; Fung [2000] ; Piperidis *et al.* [2000] ; Wu [2000]. Ces méthodes utilisent de façon combinée des approches quantitatives et des connaissances linguistiques [59]. Les unités lexicales sont identifiées à l'aide de *grammaires locales* [60], par les techniques de reconnaissance de *patrons syntaxiques*, en utilisant le langage des *expressions régulières* [61],

[59] Actuellement, les systèmes récents d'alignement lexical font l'objet d'une étude d'évaluation menée au sein du projet **EVALDA–ARCADE II** dans le cadre des *Actions de Recherche Concertée sur l'Alignement de Documents et son Evaluation*. Pour obtenir des renseignements sur ce projet, on consultera le portail *Technolangue* à l'adresse suivante : http://www.technolangue.net.

[60] Voir [Habert *et al.*, 1997, pp. 163-166].

[61] Voir [Desgraupes, 2001], Habert *et al.* [1998].

en faisant appel à des ressources dictionnairiques, par des méthodes statistiques [62] permettant la sélection d'expressions complexes au fil de corpus de textes, etc. Plusieurs algorithmes ont été proposés pour apparier les unités ainsi sélectionnées [63].

2.4 Conclusion du chapitre 2

Au fil de ce chapitre, nous avons tenté de mettre en évidence les connaissances théoriques et les outils spécifiques qui peuvent être employés pour mettre au point des systèmes d'alignement bi-textuel.

La complexité des méthodes de l'alignement automatique reflète pleinement les difficultés des recherches dans le domaine du *traitement automatique des langues* (TAL). Les études menées actuellement dans ce domaine intègrent de plus en plus de connaissances théoriques issues des disciplines variées, telles que la linguistique, l'informatique, la linguistique informatique, les statistiques, les mathématiques, l'intelligence artificielle. Les interactions avec tous les courants du TAL font de l'alignement automatique un champ d'expérience pour les théories développées dans toutes ces disciplines connexes.

L'analyse des méthodes d'alignement développées au cours de ces vingt dernières années montre que le perfectionnement des mécanismes de repérage automatique des équivalences lexicales occupe une place centrale dans les recherches actuelles. Dans les chapitres qui suivent, nous proposons de nouveaux moyens d'investigation pour aborder ce type de tâche à l'aide de méthodes quantitatives.

---

[62] Sur les questions d'automatisation de ces comptages, [Church et Gale, 1991] ; [Barlow, 2002] ; [Lamalle et Salem, 2002] ; [Martinez et Zimina, 2002] ; [Zimina, 2002].

[63] Voir, par exemple, Gale et Church [1991a] ; Dagan et Church [1997] ; Fung [2000] ; Déjean et Gaussier [2002] ; Brown P. *et al.* [1993] ; Gaussier [1998] ; Wu [2000].

Chapitre 3

La textométrie multilingue

Nous présentons dans ce chapitre les fondements de l'analyse textométrique des corpus multilingues. L'approche que nous avons élaborée utilise, à des fins d'extraction de ressources traductionnelles des corpus parallèles, les propriétés textométriques du matériau textuel repérées dans les études quantitatives des corpus de textes monolingues. Nous illustrons notre approche avec des échantillons de ressources traductionnelles obtenues à partir du corpus ***Convention***.

3.1 Le domaine de la textométrie

Les méthodes quantitatives d'analyse textuelle trouvent actuellement des applications de plus en plus nombreuses dans des domaines qui s'étendent de la *lexicographie* à l'*analyse du discours politique*, de la *recherche documentaire* à la recherche en *marketing*, la *linguistique computationnelle*, la *sociolinguistique*, etc. Dans le contexte multilingue, l'analyse quantitative procure une aide précieuse pour accéder à la description de divers phénomènes textuels relatifs

aux emplois d'unités linguistiques (graphèmes, formes, lemmes, lexies, systèmes de catégories grammaticales, séquences, etc.) [1].

Dans ce qui suit, on utilisera le terme ***textométrie*** [2] pour se référer à l'ensemble des méthodes quantitatives permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le *vocabulaire* [3] d'un corpus de textes, cf. Lebart et Salem, [1994, p. 314]. La spécificité de la démarche textométrique réside dans le statut privilégié conféré aux données textuelles qui ne prennent que très peu d'appui sur des savoirs *a priori* (linguistiques, pragmatiques, etc.). L'analyse textométrique s'appuie sur des comptages réalisés à partir du repérage des occurrences d'unités lexicales dans les différentes parties d'un corpus. Globalement, elle regroupe trois types de méthodes, cf. Salem [1987, p. 35] :

- Les méthodes documentaires qui opèrent une simple réorganisation de la surface textuelle.
- Les méthodes qui opèrent, pour des textes pris isolément, des comptages et des calculs d'indices statistiques.
- Les méthodes statistiques « contrastives » qui produisent des résultats portant sur le vocabulaire de chacun des textes par rapport à l'ensemble des textes réunis dans un corpus à des fins de comparaison.

[1] Gaussier [1999] remarque que l'analyse quantitative joue un rôle important compte tenu des limites auxquelles sont confrontés les systèmes actuels de traitement automatique des langues : *« Ni l'analyse syntaxique, ni l'analyse sémantique n'ont atteint aujourd'hui un seuil de développement suffisamment important pour être utilisées dans des applications s'intéressant à des domaines variés. Du côté de l'analyse syntaxique, on assiste à un essor de techniques dites de pré-syntaxe, ou analyse syntaxique de surface (de l'anglais shallow-parser), qui répondent à l'exigence d'une couverture importante et d'une certaine robustesse. Toutefois, ces analyseurs ne fournissent pas une analyse syntaxique complète des phrases étudiées, et un certain nombre de problèmes, comme le rattachement prépositionnel, ne reçoivent souvent qu'une solution partielle. Il en va de même au niveau de l'analyse sémantique, où des techniques d'assignation du sens des mots en contexte peuvent d'ores et déjà être utilisées, mais où la construction d'une représentation du sens d'un énoncé n'a pas encore reçu de solution acceptable pour les systèmes actuels. »*

[2] La présentation des outils de la textométrie peut être trouvée, par exemple, dans Heiden [2002]. On trouvera la présentation des axes de recherches de la textométrie sur le site de l'équipe « Analyse de corpus » (UMR « Interactions, Corpus, Apprentissages, Représentations ») de l'Ecole normale supérieure Lettres et Sciences humaines à l'adresse suivante : http://www.ens-lsh.fr/labo/corpus/pdf/projet.pdf

[3] Le *vocabulaire* d'un texte correspond à l'ensemble de formes distinctes qui y sont attestées.

Actuellement, ces méthodes sont largement utilisées dans l'étude statistique de textes monolingues pour la mise en évidence de phénomènes caractéristiques de variations du vocabulaire dans le temps ou selon les variables de l'étude [4]. Dans un contexte multilingue, la textométrie offre des perspectives de recherches prometteuses pour de multiples dimensions d'analyse de corpus dans des langues différentes (l'*alignement automatique,* la *vérification de l'alignement,* l'*extraction des lexiques bilingues,* l'*exploration intertextuelle,* la *synthèse de l'information bi-textuelle etc.*) [5].

L'extraction de ressources traductionnelles des corpus parallèles est un des thèmes de recherches en textométrie multiligue [6]. Pour présenter les développements récents dans ce domaine, nous ferons appel à des exemples tirés du corpus ***Convention***.

3.2 L'analyse textométrique du corpus bilingue *Convention*

3.2.1 Dépouillements en formes graphiques

L'utilisation de la textométrie pour l'analyse des corpus parallèles implique une tentative de mettre au point des méthodes particulières d'étude de données

---

[4] On consultera sur ces questions les *Actes de JADT'90* (Barcelone), *JADT'93* (Montpellier), *JADT'95* (Rome), *JADT'98* (Nice), *JADT'00* (Lausanne), *JADT'02* (Saint-Malo), *JADT'04* (Louvain-la-Neuve) : http://www.cavi.univ-paris3.fr/lexicometrica/jadt. Les JADT (*Journées internationales d'Analyse statistique des Données Textuelles*) réunissent statisticiens, linguistes, sociologues, spécialistes d'analyse du discours, informaticiens, spécialistes de lexicographie et d'exploration des données textuelles.

[5] La table-ronde « Lexicométrie et corpus multilingues » organisée dans le cadre des *7èmes Journées internationales d'Analyse statistique des Données Textuelles* (JADT'04) a montré le besoin d'adaptation des outils actuels de la statistique textuelle au terrain de corpus multilingues. La création du *Groupe d'Analyse des Données Textuelles* (GADT) – *Textométrie Multilingue* a permis de réunir des chercheurs de la communauté des statistiques textuelles intéressés par l'analyse des données textuelles multilingues. Sur les activités du groupe, on consultera le site à l'adresse suivante :
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/textometrie-multilingue.

[6] Voir Martinez et Zimina [2002] ; Zimina, [2000 ; 2002, 2004].

textuelles multilingues portant sur les distributions conjointes des vocabulaires dans les textes. Le recours à ce type de méthode nécessite que soient précisés les principes utilisés pour définir une *norme de dépouillement* en unités textuelles.

La norme de dépouillement : rappel

La notion de norme de dépouillement est imposée par une nécessité de rendre les unités comparables. Une fois définie, la norme est appliquée à tous les comptages ultérieurs. Pour un même texte, les différentes normes de dépouillement ne conduisent pas aux mêmes décomptes. Le choix des unités de décompte a été largement étudié dans le cadre de l'analyse statistique de corpus de textes monolingues. 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 [7].

Dans les études textométriques, les textes sont d'abord segmentés en occurrences de *formes graphiques* (chaînes de caractères bornées par deux caractères délimiteurs, en anglais – *token*). Une fois cette segmentation réalisée, on peut définir des règles d'*identification* entre des unités de segmentation ainsi obtenues. La formalisation de ces règles permet de produire des décomptes portant sur les occurrences d'un même type aux différents endroits d'un texte, cf. [Salem, 1987] ; [Lebart et Salem, 1994].

Le concept de *type généralisé* ***TGen*** permet de décrire des ensembles d'occurrences sélectionnés systématiquement dans le texte [Lamalle et Salem, 2002]. Ainsi, on peut recenser au-delà des occurrences des formes graphiques les occurrences d'un *segment répété* [8] [ex. : *démocratie apte à se*

[7] Il existe plusieurs types de segmentation en unités textuelles. Le choix des méthodes de segmentation s'appuie sur les objectifs de recherche et les objets à étudier. Pour un exposé de différents points de vue dans ce débat, on consultera [Muller, 1992, réimpression de l'édition 1977] ; [Lafon, 1984] ; [Labbé, 1990] ; [Salem, 1993] ; [Lebart et Salem, 1994] ; [Habert *et al.*, 1998] ; [Brunet, 2000, 2002] ; [Lamalle et Salem, 2002].

[8] Sur la méthode des *segments répétés* [Salem, 1987].

défendre], d'un *quasi-segment répété* [9] [ex. : (ang.) : *there had been a breach of the article* et *there had been a violation of the article*], la rencontre de deux formes (*cooccurrence*) [10] à l'intérieur d'une fenêtre de x formes graphiques ou d'une phrase [ex. : *démocratie +république*], d'un type constitué par les occurrences d'un ensemble de formes graphiques défini en raison de leur parenté sémantique dans le corpus [ex.: *démocratique, démocratie, démocratiques, démocrate*] [11], etc.

Si l'on effectue un dépouillement en *lemmes*, la définition d'un type de rattachement pour chaque occurrence nécessite le recours à un dictionnaire et, dans certains cas, un retour au contexte [12]. Il s'agit d'une opération de dépouillement plus complexe qui permet d'obtenir des unités plus élaborées du point de vue lexicographique [13].

---

[9] Les *quasi-segment répétés* permettent de repérer des séries de répétitions lexicales légèrement altérées par des modifications de l'un de leurs composants [Bécue et Peiro, 1993].

[10] Nous entendons par *cooccurrence* la présence simultanée, mais pas forcément contiguë, dans un fragment de texte (phrase, paragraphe, voisinage d'une occurrence, etc.) des occurrences de deux formes données. Sur le calcul des cooccurrences, [Church et Hanks, 1990] ; [Haruno *et al.*, 1996] ; [Lafon, 1981] ; [Martinez, 2000 ; 2003] ; [Grossmann et Tutin, 2003].

[11] Ces séries d'unités peuvent être définies à l'aide d'outils permettant l'accès au langage des *expressions régulières*, voir, par exemple, [Habert *et al.*, 1998].

[12] Le dépouillement en lemmes (ou *lemmatisation*) est une opération de regroupement sous une forme canonique (à partir d'un dictionnaire) des occurrences du texte. En règle générale, pour lemmatiser un texte en français, on ramène les formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier, les formes élidées à la forme sans élision. Les avantages de lemmatisation sont exposés dans les travaux de Charles Muller [1992, réimpression de l'édition 1977] ainsi que dans la préface qu'il a donnée à l'ouvrage de Pierre Lafon *Dépouillements et statistiques en lexicométrie* [1984]. Parmi les publications récentes sur les questions de lemmatisation, on consultera, par exemple, [Brunet, 2000, 2002] ; [Mellet, 2001].

[13] Dans l'absolu, il est toujours préférable de disposer d'un double réseau de décomptes (en formes graphiques et en lemmes). Comme le remarque Brunet [2000], l'utilisation simultanée de la forme et du lemme permet de cumuler les avantages des deux systèmes de décomptes et de neutraliser leurs inconvénients. Cependant, une lemmatisation complète, sur un corpus important, reste une opération coûteuse : *« /…/ indispensable dans un travail de recherche, elle est beaucoup moins justifiée s'il s'agit d'obtenir rapidement des visualisations ou des typologies de parties de corpus d'une certaine richesse lexicale. »* [Lebart et Salem, 1994, p. 226].

La segmentation du corpus

Dans son état initial, le corpus ***Convention*** correspond à un seul fichier texte où chaque couple de phrases équivalentes est introduit par un code (cf. *Chapitre 1 : Tableau 1.1*) [14]. L'appariement des deux volets du corpus est donné jusqu'au niveau du paragraphe. Le découpage en phrases au sein des paragraphes est correct à environ 90 %. Comme le montre la figure 3.1, les erreurs concernent notamment des phrases incluant une marque de ponctuation forte au milieu (point, deux points, etc.) où les fins des phrases n'ont pas été repérées correctement, ce qui a provoqué un décalage dans l'appariement.

Nous avons entrepris une série de transformations du corpus (cf. *Figure 3.2*) [15]. La *segmentation* [16] parallèle des volets français et anglais en formes graphiques a permis ensuite la mise en œuvre de calculs statistiques en allégeant la manipulation de la chaîne textuelle. Notre objectif consiste maintenant à vérifier si les règles formelles du dépouillement automatique peuvent aider à rapprocher des lexèmes candidats à l'appariement.

---

[14] Voir le fichier /fichiersCD/stmz/page1.htm sur le Cd-rom joint à ce volume.

[15] Les deux volets du corpus ont été séparés grâce à une petite procédure informatique programmée en langage *Perl*. Après cette première transformation, les codes correspondant à la numérotation des phrases ont été retirés dans chacun des volets du corpus. Les textes ont été ensuite transformés de manière à supprimer les majuscules pour confondre dans les mêmes décomptes les formes qui apparaissent en début de phrase et leurs homologues situées en l'intérieur d'une phrase.

[16] Le corpus n'a subi aucune annotation ni lemmatisation. 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. 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. Une description des principes de segmentation en formes graphiques peut être trouvée dans [Lebart et Salem, 1994, pp. 42-45] ; [Lafon, 1984, pp. 15-23].

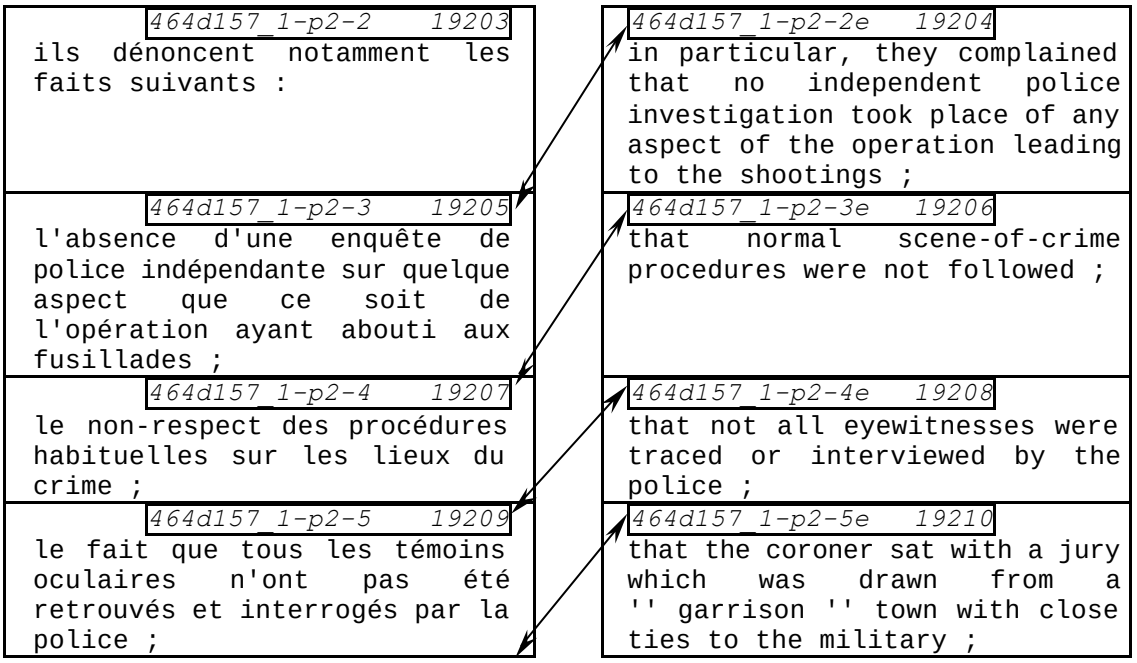


Figure 3.1 : Erreurs recensées dans l'appariement des phrases du corpus ***Convention***

Guide de lecture : Chaque couple de phrases alignées du corpus est introduit par un code. Les numéros indiquent le type de document, la partie de l'arrêt, le numéro de section et/ou du paragraphe ; le numéro de la phrase dans le corpus, précédé par la lettre « e » pour les phrases en anglais. Les flèches montrent le décalage dans l'appariement suite à la présence d'une phrase en français incluant une marque de ponctuation forte au milieu.

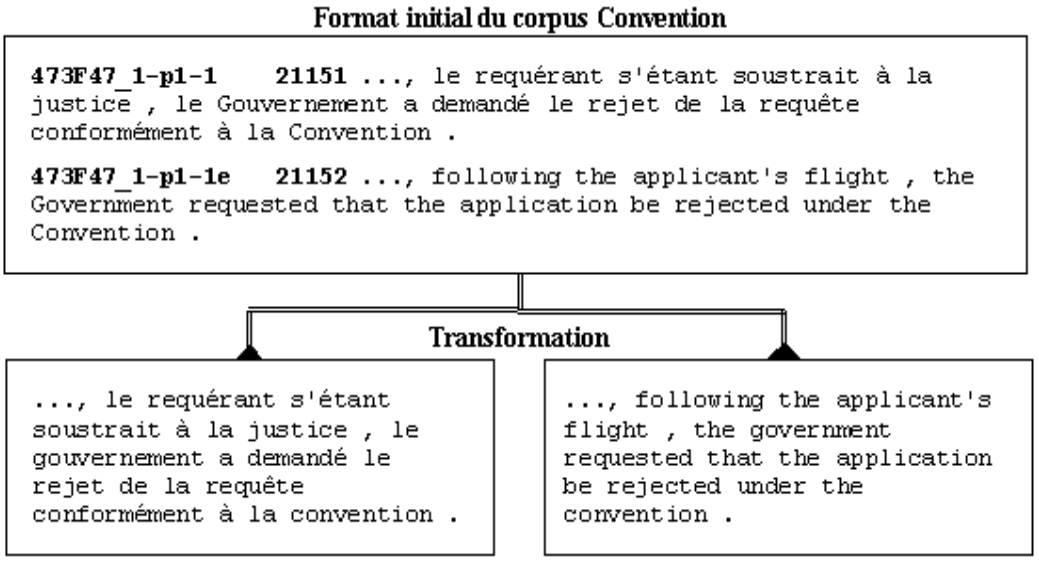


Figure 3.2 : Etat du corpus ***Convention*** après une série de transformations

Les premiers comptages réalisés sur le corpus segmenté en formes graphiques donnent un aperçu grossier des principales caractéristiques quantitatives des volets français et anglais (cf. *Tableau 3.3*).

Tableau 3.3 : Résultats de la segmentation du corpus ***Convention***

	occurrences	**formes**	**hapax**	**fmax**	
Français	296 396	12 913	4 959	*de*	17 572
Anglais	284 958	9 530	3 407	*the*	29 622
Liste de caractères-délimiteurs : **.,:;!?/_\ ’« »()[]{}§$**					

Nous constatons sur le *tableau 3.3* que le nombre total d’occurrences est plus important pour le volet *français* (296 396 contre 284 958 pour le volet *anglais*). Le volet français du corpus est également plus diversifié en terme de formes (12 913 > 9 530) et compte beaucoup plus d’*hapax* [17] (4959 > 3407). La fréquence maximale *the* en anglais est largement supérieure à celle du texte français *de* (29622 > 17572) [18]. Pour une comparaison plus précise des gammes des fréquences, on utilisera le *diagramme de Pareto* [19].

Gammes des fréquences – diagramme de Pareto

Le diagramme de Pareto présenté sur la *figure 3.4* nous permet de constater la plus grande diversité des formes employées dans le volet français. Les écarts entre les deux courbes concernent notamment les deux extrémités des gammes de

[17] Les *hapax* sont des formes qui ne sont attestées qu’une seule fois dans le texte.

[18] En français, l’article défini prend les formes suivantes : *le*, *la*, *les* et *l’* (devant un mot commençant par une voyelle ou un « h » muet). Cette diversité est absente en anglais où *the* est la seule forme de l’article défini. Cette différence explique la forte répétitivité de cette forme dans le volet anglais.

[19] 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, voir [Lebart et Salem, 1994] ; [Labbé *et al.*, 1988].

fréquences : les formes de faible fréquence et celles de fréquence maximale. Dans l'intervalle des fréquences moyennes (20 à 1 000 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 la présence de désinences qui change la forme graphique du mot [ex. : *contraignant* (F=2), *contraignante* (F=2), *contraignants* (F=2), *contraignantes* (F=2) / *binding* (F=12)].

La confrontation de contextes équivalents contenant des 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) [20].

La comparaison des dictionnaires des formes

L'analyse des dictionnaires des formes graphiques constitués à partir des deux volets du corpus permet de poursuivre la comparaison des caractéristiques fréquentielles (cf. *Tableau 3.5*). On constate que pour une partie des formes bilingues en correspondance de traduction, la confrontation des dictionnaires classés par *ordre lexicométrique* [21] fait apparaître des similitudes entre *rangs lexicaux* [22] (cf. *Tableau 3.6*). L'analyse sémantique de ces équivalences lexicales en corpus permet de comprendre les causes de ces correspondances souvent *biunivoques*.

[20] Le phénomène inverse est attesté dans le corpus à une échelle beaucoup plus modeste. Par exemple, lorsque la forme *de* (F=17 572) est employée en tant que préposition, elle peut être traduite en anglais par *of* (F=15 294), *for* (F=2 299), *as* (F=2 317), etc.

[21] *Ordre lexicométrique* : ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes ; les formes de même fréquence sont classées par ordre selon lequel elles apparaissent dans un dictionnaire (*ordre lexicographique*), voir Salem [1987, p. 316-317].

[22] Le numéro du rang lexical est attribué en fonction de la place occupée par la forme dans le dictionnaire trié par ordre lexicométrique.

Diagramme de Pareto

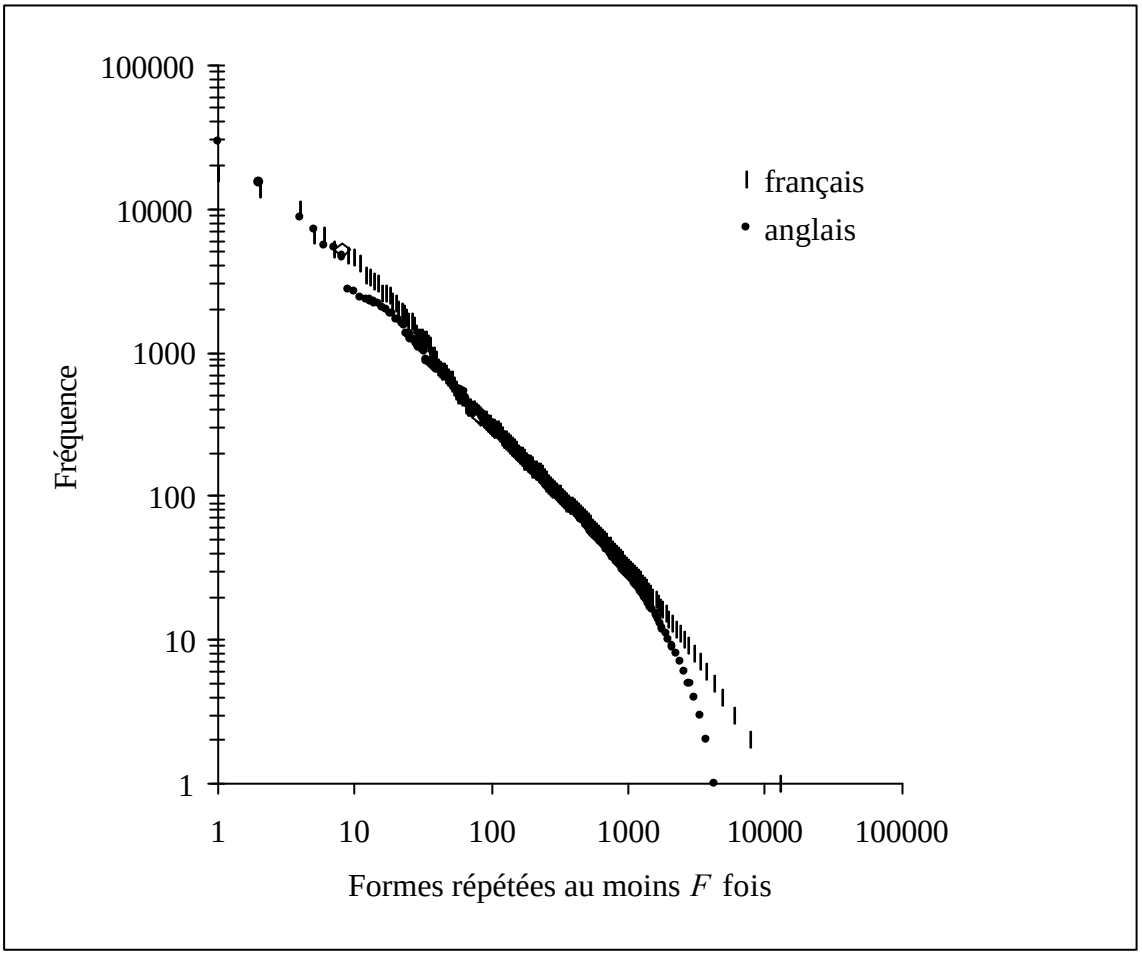


Figure 3.4 :
Diagramme de Pareto pour les volets français et anglais du corpus ***Convention***

Guide de lecture du Diagramme de Pareto :

Sur l'axe vertical (*Y*), gradué selon une échelle logarithmique, on lit la fréquence des formes *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}).

Tableau 3.5 : Corpus ***Convention*** : extraits des dictionnaires des formes graphiques

Fréquence	forme FRA	forme ANG	Fréquence
17572	de	the	29622
13409	la	of	15294
9896	l	to	8775
6629	à	in	7175
6435	le	and	5642
5232	et	a	5396
5134	d	that	4649
4695	les	court	2717
4666	en	by	2642
4122	des	was	2392
3477	du	as	2317
3289	que	for	2299
3119	une	be	2231
3003	a	it	2217
2585	il	not	2032
2544	un	on	1999
2514	par	had	1867
2304	au	is	1860
2223	dans	or	1716
1979	pour	s	1692
1913	cour	with	1600
1874	qu	which	1544
1758	pas	been	1362
1663	ne	this	1333
1645	sur	applicant	1244
1523	ou	convention	1228
1338	n	an	1149
1292	se	at	1101
1261	s	his	1078
1258	qui	he	1045
1223	convention	case	1009
1219	est	from	887
1143	elle	under	857
1111	son	have	856
1027	aux	commission	832
968	ce	see	808
942	droit	were	798
889	commission	any	789

Tableau 3.5 : (suite)

Fréquence	forme FRA	forme ANG	Fréquence
847	avait	proceedings	773
790	été	paragraph	772
748	sa	law	757
742	cette	no	746
740	paragraphe	government	740
721	gouvernement	its	739
700	si	shall	708
665	y	they	705
644	procédure	their	690
643	requérant	there	669
619	loi	judgment	668
603	était	has	640
567	devant	would	570
554	1	if	569
550	avec	1	560
537	comme	decision	557
534	peut	above	549
527	arrêt	are	546
508	décision	such	545
505	etat	order	537
499	même	before	536
498	fait	made	533
486	dessus	may	494
485	plus	other	477
481	ses	appeal	453
449	lui	right	449
436	leur	rights	444
427	ils	administrative	440
424	fin	application	432
418	après	also	428
418	droits	de	428
415	ont	only	423
410	ainsi	her	416
408	dont	one	411
405	sont	criminal	409
396	affaire	public	407
385	tribunal	hearing	406
384	violation	state	404
383	non	authorities	399
380	mais	against	396
363	tout	but	392
362	article	could	392

Tableau 3.6 : Rangs lexicaux des formes équivalentes

français	**rang lexical**	**fréquence totale**	**anglais**	**rang lexical**	**fréquence totale**
convention	31	1223	convention	26	1228
commission	38	889	commission	35	832
paragraphe	43	740	paragraph	40	772
gouvernement	44	721	government	43	740
décision	57	508	decision	54	557
droits	69	418	rights	65	444
mais	78	380	but	79	392
article	80	362	article	85	351
citation	83	352	citation	86	351
détention	96	314	detention	95	336

Guide de lecture : Pour une partie des formes en correspondance, les rangs lexicaux sont très proches. Le *tableau 3.6* fournit la liste de ces formes accompagnées de leurs rangs lexicaux. Le rang est attribué en fonction de la place occupée par la forme dans le dictionnaire des formes trié par ordre lexicométrique.

3.2.2 Rapports de correspondances lexicales multilingues

Les correspondances quasi-univoques

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

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

Exemple :

corpus français	**fréquence générale**	**corpus anglais**	**fréquence générale**
herbert	29	herbert	29
petzold	29	petzold	29
ii	77	ii	77
1991	13	1991	13
1992	11	1992	10

- **Les concepts désignés par les deux mots se correspondent car ils sont utilisés dans le corpus dans leur sens premier (ou étymologique)[23].**

Exemple :

corpus français	**fréquence générale**	**corpus anglais**	**fréquence générale**
père	29	father	27

père n.m. — I - **1.** Homme qui a engendré, qui a donné naissance à un ou plusieurs enfants. 2. *PÈRE DE FAMILLE*, qui a un ou plusieurs enfants qu'il élève. 3. Le parent mâle. 4. plur. littér. => aï eul, ancêtre, ascendant. 5. La première personne de la sainte Trinité. 6. Le père de qqch. => créateur, fondateur, inventeur. 7. Celui qui se comporte comme un père, est considéré comme un père. 8. Personnage âgé et solennel au théâtre. 9. Père abbé : religieux assurant la direction d'une communauté. 10. Désignant un homme mûr et de condition modeste. *Source : Dictionnaire LE ROBERT, édition 1996.*	**father** n. — **1.** male parent of child or animal 2. The man who began or invented (the stated thing) 3. [*usu. pl.*] a FOREFATHER. *Source : Longman Dictionary of Contemporary English, 1992.*

[23] D'après Seleskovitch [1968, p. 147], la production d'équivalences de type *mort / death*, *mère / mother*, etc. passe par la restitution du *sens premier* (ou sens élémentaire) des mots dans les deux langues. Il s'agit, en général, de concepts relativement universels, qui relèvent de l'expérience socioculturelle partagée. On peut qualifier ces équivalences de « traduction étymologique ».

- **L'équivalence est imposée par le respect d'usages terminologiques** [24]**.**

Exemple :

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

- **Le contexte réduit l'espace des sens possibles pour les deux mots.**

Exemple :

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

Le mot *article* est polysémique en anglais comme en français. Cependant, dans le corpus ***Convention*** ce terme est employé dans les deux langues pour désigner la structure de découpage des documents juridiques. Le contexte du corpus réduit l'espace des sens possibles pour les deux mots :

article n.m. — I - Pièce articulée des arthropodes. II – 1. Partie (numérotée ou non) qui forme une division (d'un texte légal, juridique, diplomatique, religieux, littéraire). 2. Partie (d'un écrit). 3. Écrit formant par lui-même un tout distinct, mais faisant partie d'une publication. III - Objet de commerce. IV – Dans certaines langues, Mot, qui placé devant un nom (ou l'adj. antéposé au nom) sert à le déterminer plus ou moins précisément, et peut prendre la marque du genre et du nombre. *Source : Dictionnaire LE ROBERT, édition 1996.*	**article** n. — 1. a particular or separate thing or object, esp. one of a group. 2. a separate piece of writing on a particular subject in a newspaper, magazine, etc., that is not fiction. 3. a complete separate part in a legal agreement, CONSTITUTION, etc. 4. tech a word used with a noun to show whether the noun refers to a particular example of something (the definite article – the in English) or to a general or not already mentioned example of something (the indefinite article - a or an in English). *Source : Longman Dictionary of Contemporary English, 1992*

[24] Nos observations convergent avec le classement d'équivalences proposé par Seleskovitch [1968, p. 150] qui distingue *« /…/ la traduction étymologique (celle du sens premier des mots) ; la traduction conventionnelle (celle qui fournit en guise d'équivalents agréés les termes usités dans un secteur déterminé), et la traduction contextuelle, réexpression créatrice qui fournit des équivalents linguistiques à validité unique dans un contexte donné. »*

Sur le plan fréquentiel, les *correspondances quasi-univoques* présentent parfois quelques écarts. Ces écarts s'expliquent notamment par l'utilisation des pronoms et de la paraphrase employée par le traducteur pour éviter la répétition d'un même lexème au sein de la même phrase ou dans les phrases voisines. Voici quelques illustrations qui expliquent l'écart dans les fréquences des formes *article* (F=362) / *article* (F=351) en équivalence de traduction dans le corpus :

français	**anglais**
*nul ne conteste que l'**article** de la convention s'applique à ces procédures. la question en cause devant la cour est celle de savoir si cet **article** a été violé et à quel degré.*	*it is common ground that the **article** of the convention applies to these proceedings. what is at issue before the court is whether and, if so, to what extent **it** was violated.*
*les membres de la commission et de la cour jouissent, pendant l'exercice de leurs fonctions, des privilèges et immunités prévus à l'**article** 40 du statut du conseil de l'europe et dans **les accords conclus en vertu de cet article**.*	*the members of the commission and of the court shall be entitled, during the discharge of their functions, to the privileges and immunities provided for in **article** 40 of the statute of the council of europe and in **the agreements made thereunder**.*

De la même manière, les formes *détention* (F=314) et *detention* (F=336) sont des traductions mutuelles dans le corpus. Lorsque l'équivalence *détention* / *detention* devient trop répétitive, le traducteur fait appel à des collocations dans une des langues pour se donner des moyens expressifs supplémentaires. Cette variation explique des écarts dans les fréquences des deux formes :

français	**anglais**
*selon le requérant, sa **mise sous écrou** extraditionnel n'aurait servi qu'à le garder en **détention** pour les besoins de l'instruction en france.*	*according to the applicant, his **detention** with a view to extradition had served solely to keep him in **custody** for the needs of the investigation in france.*

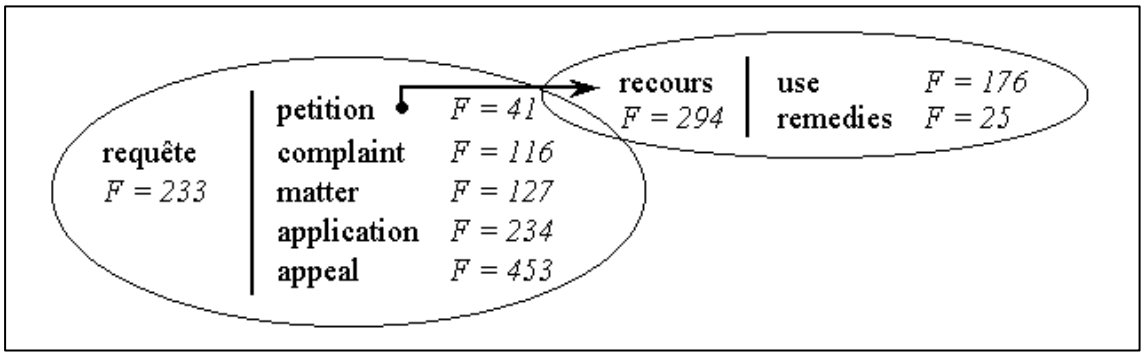
Les correspondances multiples

Le sens dans lequel le lexème est employé dans un contexte donné détermine sa traduction dans une autre langue. Prenons l'exemple du mot français *mort*. Dans la majorité des cas, la traduction de *mort* (F=44) en anglais est le mot *death* (F=26). Dans le corpus, on rencontre cette correspondance au sein des équivalences lexicales *la peine de mort – death penalty*, *les circonstances de la mort – the circumstances of the death*, *un danger de mort – a risk of death* et beaucoup d'autres. Cependant, l'équivalence *mort – death* n'est pas préservée dans les contextes suivants :

français	**anglais**
La mort *n'est pas considérée comme infligée en violation de cet article dans les cas où elle résulterait d'un recours à la force rendu absolument nécessaire.*	***Deprivation of life*** *shall not be regarded as inflicted in contravention of this Article when it results from the use of force which is no more than absolutely necessary.*
La Cour estime que les exceptions définies au paragraphe 2 montrent que l'article vise certes les cas où ***la mort*** *a été infligée intentionnellement, mais que ce n'est pas son unique objet.*	*The Court considers that the exceptions delineated in paragraph 2 indicate that this provision extends to, but is not concerned exclusively with, intentional* ***killing****.*

Dans les deux couples de phrases ci-dessus le mot *mort* est utilisé dans le sens « fin provoquée de la vie », répertorié, par exemple, par le dictionnaire *LE ROBERT* (édition 1996). L'absence de cette signification dans l'univers sémantique du mot anglais *death* explique la présence des équivalences *mort – deprivation of life*, *mort – killing* dans le corpus.

Lorsqu'il s'agit de mots dotés d'un large éventail de sens dans le corpus, les correspondances lexicales entre les deux volets forment un réseau complexe et la comparaison des fréquences globales des formes graphiques ne constitue pas toujours une bonne indication pour l'appariement. Ainsi, le mot français polysémique *requête* (F=233) reçoit plusieurs traductions (cf. *Figure 3.7*).



- introduire une *REQUETE* ~ to submit a *PETITION*
- le droit de *RECOURS* individuel ~ the right of individual *PETITION*
- épuiser les voies de *RECOURS* internes ~ to exhaust all domestic *REMEDIES*
- avoir droit d'un RECOURS effectif ~ to have an effective REMEDY
- avoir le droit d'introduire un *RECOURS* devant un tribunal ~ to be entitled to take *PROCEEDINGS*
- le *RECOURS* à la force ~ the *USE* of force
- *REQUETE* initial ~ original *APPLICATION*

Figure 3.7 : Univers lexical du terme *requête* : exploration bi-textuelle

Le retour au contexte montre que l'une des correspondances de *requête*, le terme anglais *petition* (F=41), est polysémique lui-même :

français	**anglais**
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 requête.*	*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 petition has been referred .*
/…/ *avec pour corollaire que l'etat pourrait limiter l'acceptation du droit de* ***recours*** *individuel à son territoire national, comme il l'a fait en l'occurrence.*	*/…/* *as a corollary, the state can limit acceptance of the right of individual* ***petition*** *to its national territory - as has been done in the instant case.*

De même pour le terme français *recours* (F=294) :

français	**anglais**
la mort n'est pas considérée comme infligée en violation de cet article dans les cas où elle résulterait d'un ***recours*** *à la force rendu absolument nécessaire: /…/*	*deprivation of life shall not be regarded as inflicted in contravention of this article when it results from the* ***use*** *of force which is no more than absolutely necessary : /…/*
la commission ne peut être saisie qu'après l'épuisement des voies de ***recours*** *internes /…/*	*the commission may only deal with the matter after all domestic* ***remedies*** *have been exhausted /…/*

Les équivalences contextuelles

Une partie des équivalences traductionnelles du corpus relèvent de la *traduction contextuelle*. Ces équivalences sont peu autonomes sur le plan lexical. Sur la *figure 3.8*, ce phénomène est illustré par des équivalences singulières recensées autour du mot français *monde* (F=9). L'équivalence « principale » *monde – world* (F=5) est dominante : elle a été utilisée près d'une fois sur deux dans la traduction. Les autres équivalences, plus rares, concernent notamment des expressions et des syntagmes complexes où le mot *monde* et son voisinage lexical immédiat sont liés au sein de la même unité de traduction, par exemple : *l'adhésion à une conception philosophique du monde ~ adherence to an ideology*. Il serait impossible d'isoler un élément correspondant au mot *monde* dans cette équivalence.

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

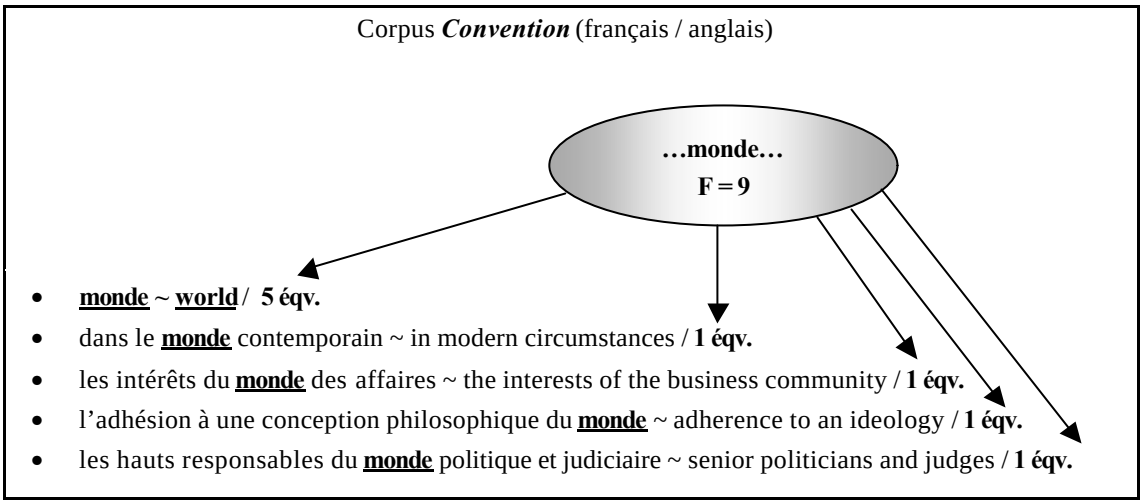


Figure 3.8 : Les équivalences recensées autour du mot *monde*

3.3 Recherche d'équivalences par la méthode de segments répétés

Nous avons montré que la comparaison des fréquences générales des formes isolées se révèle souvent insuffisante pour détecter des équivalences lorsqu'il s'agit des emplois polysémiques. L'étude des formes en contexte permet de lever certaines des ambiguï tés résiduelles ; au contact d'autres mots associés sur l'axe syntagmatique, différents composants du sens du mot sont activés et il devient possible d'en tenir compte lors de l'appariement. Nous allons ainsi examiner dans quelle mesure les comptages portant sur des unités textométriques plus larges, composées de plusieurs formes, sont susceptibles de fournir des indices pour l'extraction d'équivalences du corpus. Les résultats obtenus à partir de décomptes de formes graphiques isolées seront complétés par des comptages portant sur des objets textuels de type *segment répété* [25].

L'analyse de l'extrait de la concordance de la forme *respect* (cf. *Tableau 3.9*) permet de remarquer que les occurrences de cette forme sont prises dans des

[25] *Segment répété* (ou polyforme répétée) est suite de formes dont la fréquence est supérieure ou égale à deux dans le corpus [Salem, 1987, p. 319].

segments plus longs : *in this respect*, *obligation to respect*, *right to respect for*, *with respect to*, etc. Il existe des procédures de quantification qui peuvent éviter d'examiner l'ensemble des contextes de chacune des unités du corpus afin d'en cerner le sens. La méthode de *segments répétés* [26] permet de généraliser ce type de démarche à l'ensemble des formes du corpus par édition automatique des listes de *polyformes* [27], ou segments. Les techniques de repérage et décompte des segments répétés du corpus peuvent être utilisées avec profit pour la description locale de la concordance.

[26] Sur la pratique des segments répétés [Salem, 1987]. Sur les mesures de récurrences sur l'axe syntagmatique on consultera également [Habert *et al.*, 1997, pp. 191-193].

[27] En textométrie, la notion de *polyforme* désigne à la fois un type des occurrences d'un segment et une suite de formes non séparées par un délimiteur de séquence, qui n'est pas obligatoirement attestée dans le corpus, cf. [Salem, 1987, p. 318] ; [Lamalle *et al.*, 2003, p. 42].

Tableau 3.9 : Extrait de la concordance autour de la forme *respect* dans le volet anglais du corpus ***Convention***

er severed his links with them . in this	**respect**	the fact that he spends a lot of time ou
e be any difference of treatment in this	**respect**	according to the nature of the civil ser
not , there will have been a failure to	**respect**	their family life and the interference r
$ _ 4 (p4-2) on account of failure to	**respect**	her right to liberty of movement on fren
inges the right that every person has to	**respect**	for his private and family life if it is
footnote 1 $ _ article 1 - obligation to	**respect**	human rights $ _ article 2 - right to li
ng young people and to the obligation to	**respect**	personal honour . $ _ (3) there shall
rities were bound by their obligation to	**respect**	the right to life of the suspects to exe
ent without law $ _ article 8 - right to	**respect**	for private and family life $ _ article
s a continuing violation of her right to	**respect**	for her home contrary to of the conventi
authorities would infringe his right to	**respect**	for his family life and would breach of
xercise by the applicant of his right to	**respect**	for his family life . $ _ paragraph 2 of
rence with the enjoyment of his right to	**respect**	for family life resulting from the non-d
nations . $ _ everyone has the right to	**respect**	for his private and family life , his ho
ued . $ _ it would infringe the right to	**respect**	for family life and therefore constitute
in relation to enjoyment of the right to	**respect**	for family life under (see , mutatis mu
n relying on , invokes only the right to	**respect**	due to his family life . $ _ this approa
. $ _ their reflex action in this vital	**respect**	lacks the degree of caution in the use o
at ' ' their reflex action in this vital	**respect**	lacks the degree of caution . . . $ _ to
rence to ' ' reflex action in this vital	**respect**	' ' - underlining supplied . $ _ to be t
ce commercial debts . $ _ the court will	**respect**	the legislature ' s assessment in such m
on of such measures . $ _ the court will	**respect**	the legislature ' s judgment as to what
on of such measures . $ _ the court will	**respect**	the legislature ' s judgment as to what
able basis , it awards chf 40 , 000 with	**respect**	to items (a) , (b) and (c) and £70
t a state which is thus accountable with	**respect**	to a certain territory remains so even i
ected by the parliamentary assembly with	**respect**	to each high contracting party by a majo
should exercise judicial restraint with	**respect**	to issues of expediency (29) . $ _ jud
ty , as a rule there are safeguards with	**respect**	to third parties which are in no way to

Guide de lecture : Les contextes sont triés par ordre alphabétique de l'occurrence qui précèdent la forme-pôle. Ce type de tri permet de cerner les différents sens du mot *respect* dans le corpus.

Le terme *voisinage d'une occurrence* définit par Salem [1987, pp. 33-56, p. 139] désigne tout segment qui contient cette occurrence dans le corpus. Nous avons ainsi décrit le voisinage de la forme *respect* dans le volet anglais du corpus par les 30 segments répétés les plus fréquents recensés autour de cette forme (cf. *Figure 3.10*). Dans le texte, ces récurrences d'unités ont des origines diverses. Certaines d'entre elles correspondent à des syntagmes bien formés, d'autres résultent de la reprise partielle de fragments de phrases dont le statut syntaxique est intermédiaire. On remarque que les unités de ce type sont généralement dotées d'un sens que l'on ne peut pas déduire à partir des sens des mots isolés qui rentrent dans leur composition.

Long	*segment répété*	*Fréq*
2	**respect** of	250
3	in **respect** of	248
3	**respect** of the	46
4	in **respect** of the	45
3	in this **respect**	38
4	in **respect** of the	36
4	reservation in **respect** of	24
2	**respect** for	23
2	to **respect**	19
4	in **respect** of any	17
4	convention in **respect** of	15
2	**respect** to	15
5	the convention in **respect** of	14
4	in **respect** of costs	14
6	of the convention in **respect** of	13
5	in **respect** of costs and	13
3	with **respect** to	13
3	right to **respect**	13
2	**respect** the	13
7	reservation in **respect** of the convention	12
6	in **respect** of costs and expenses	12
3	to **respect** for	12
6	the convention in **respect** of the	11
5	in **respect** of the first	11
5	‘s reservation in **respect** of	11
4	right to **respect** for	11
7	of the convention in **respect** of the	10
6	in **respect** of the first applicant	10
4	in **respect** of a	10
4	in **respect** of facts	10

Figure 3.10 :
Les 30 segments répétés les plus fréquents recensés autour de la forme pivot *respect*

En dépit du statut intermédiaire de l'unité type *segment répété* du point de vue de la segmentation du corpus en unités de signification pertinentes, nous pensons que les inventaires de segments répétés recensés parallèlement dans les deux volets bilingues du corpus constituent une ressource précieuse pour l'analyse des éléments formels d'expression qui servent à maintenir l'équivalence au niveau des mots et des syntagmes (cf. *Tableau 3.11*) [28].

Plusieurs types de classements de segments répétés ont été élaborés dans le cadre de l'exploitation de corpus de textes monolingues, cf. Salem [1987, pp. 93-105]. L'ordre dans lequel sont classés les différents contextes de la forme joue un rôle déterminant dans la description formelle de son voisinage lexical.

Nous avons exploité plusieurs critères de sélection et de classements de segments répétés recensés parallèlement dans les deux volets bilingues du corpus afin de rapprocher automatiquement des polyformes équivalentes. Les *inventaires alphabétiques* dans lesquels les formes-pôles sont classées selon l'ordre lexicographique (celui des dictionnaires), se sont révélés peu adaptés au repérage des segments les plus longs ou des segments les plus fréquents qui présentent pourtant beaucoup d'intérêt pour la mise en correspondance des structures lexicales de deux textes parallèles.

La confrontation des *inventaires hiérarchiques* où les segments sont triés d'abord par ordre de longueur décroissante, les segments de même longueur étant rangés par ordre de fréquences décroissante permet une première approche des répétitions formelles contenues dans chacun des volets pour une longueur de segment donnée. (cf. *Tableau 3.11*). Cependant, ces résultats ne sont pas suffisamment précis pour les objectifs d'appariement fixés au départ. Comme le remarque Salem [1987, pp. 115] au sujet des corpus de textes monolingues, le principal défaut de ce type de comptage, opéré sur les familles de segments de différentes longueurs, tient au fait qu'ils ne sont pas indépendants mais au contraire fortement liés d'une classe de segments à l'autre.

[28] Les listes des segments répétés dont la fréquence est supérieures ou égale à 10 dans les volets français et anglais du corpus ***Convention*** sont présentées en annexe A.3.

Tableau 3.11 : Corpus ***Convention*** : extraits des inventaires de segments répétés [29] (français / anglais) de longueur de 6

L	F	Segment en français	Segment en anglais	F	L
6	61	au palais des droits de l	**the european commission of human rights**	47	6
6	56	il y a eu violation de	of the code of criminal procedure	46	6
6	56	secrétaire général du conseil de l	and the delegate of the commission	46	6
6	55	de la convention et du règlement	a violation of # of the convention	43	6
6	48	**commission européenne des droits**	the case was referred to the	37	6
6	42	sur le point de savoir si	to take part in the proceedings	37	6
6	42	y a pas eu violation de	in the presence of the registrar	37	6
6	41	**la juridiction obligatoire de la cour**	referred to the court by the	37	6
6	37	sur la violation alleguee de l	**the compulsory jurisdiction of the court**	36	6
6	36	dans le délai de trois mois	**done in english and in french**	36	6
6	36	**fait en français et en anglais**	a decision as to whether the	35	6
6	34	si les faits de la cause	in fine of the convention and	34	6
6	32	le nom des sept autres membres	**the full text of the commission**	33	6
6	32	figure en annexe au présent arrêt	it originated in an application against	33	6
6	31	**le texte intégral de son avis**	lodged with the commission under by	31	6
6	31	et le délégué de la commission	**had held a preparatory meeting beforehand**	30	6
6	30	**avait tenu auparavant une réunion préparatoire**	there has been a violation of	30	6
6	29	le désir de participer à l	there has been a breach of	29	6
6	25	accompagne figure en annexe au présent	that there has been a violation	26	6
6	24	et si le droit interne de	of human rights and fundamental freedoms	24	6
6	24	manifesté le désir de participer à	in accordance with # of the convention	24	6
6	23	la décision de la cour accorde	a breach of # of the convention	24	6
6	22	**etats membres du conseil de l**	**member states of the council of**	24	6
6	21	de sauvegarde des droits de l	the decision of the court shall	23	6
6	21	etat défendeur aux exigences de l	in accordance with the provisions of	21	6
6	20	la cour de sûreté de l	the commission declared the application admissible	20	6
6	20	au présent arrêt se trouve joint	there has been no violation of	20	6
6	19	le greffier a reçu le mémoire	within the meaning of # of the	20	6

[29] Au sein des segments répétés, le caractère # remplace un mot systématiquement absent dans le corpus qui nous a été fourni. Il s'agit, probablement, du mot « article ».

Tableau 3.12 : Corpus ***Convention*** : extraits des inventaires de segments répétés (français / anglais) de longueur de 7

L	F	Segment en français	Segment en anglais	F	L
7	52	de la convention et du règlement a	wished to take part in the proceedings	35	7
7	49	**il y a eu violation de l**	by the european commission of human rights	32	7
7	47	la commission européenne des droits de l	it originated in an application against the	32	7
7	37	a été déférée à la cour par	**recognised the compulsory jurisdiction of the court**	31	7
7	36	y a pas eu violation de l	referred to # and to the declaration whereby	30	7
7	35	dans le délai de trois mois qu	pursuant to the order made in consequence	29	7
7	35	sur le point de savoir si les	and if the internal law of the	24	7
7	34	**reconnaissant la juridiction obligatoire de la cour**	**afford just satisfaction to the injured party**	24	7
7	33	de la commission au sujet de l	allows only partial reparation to be made	24	7
7	32	le secrétaire général du conseil de l	**that there has been a breach of**	22	7
7	32	a tiré au sort le nom des	**that there has been a violation of**	22	7
7	29	la demande de la commission renvoie aux	been a violation of # of the convention	22	7
7	24	les débats se sont déroulés en public	convention for the protection of human rights	22	7
7	24	accompagne figure en annexe au présent arrêt	protection of human rights and fundamental freedoms	22	7
7	23	**à la partie lésée une satisfaction équitable**	application of article 50 of the convention	21	7
7	23	le texte intégral de son avis et	that the respondent state is to pay	20	7
7	22	la cour a entendu en leurs déclarations	member states of the council of europe	20	7
7	22	manifesté le désir de participer à l	there has been a breach of # of	19	7
7	18	a manifesté le désir de participer à	been a breach of # of the convention	18	7
7	18	au secrétaire général du conseil de l			
7	18	devant la commission a saisi la commission			
7	18	saisi la commission en vertu de l			

Tableau 3.13 : Segments répétés recensés autour des formes bilingues *cour/court*

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

Guide de lecture : Les segments répétés recensés autour des formes équivalentes cour/court sont triés par ordre décroissant des fréquences.

Par exemple, on constate sur les *tableaux 3.11-12* que le segment français *la juridiction obligatoire de la cour* (L=6) est contenu dans le segment *reconnaissant la juridiction obligatoire de la cour* (L=7). De la même manière, le segment équivalent en anglais *the compulsory jurisdiction of the court* fait partie d'un segment plus long *recognised the compulsory jurisdiction of the court* (L=7).

Les *inventaires des voisinages récurrents* (IVR) [30] édités parallèlement pour des couples de formes bilingues représentant des traductions mutuelles dans le corpus permettent de réaliser une étude plus fine des correspondances lexicales qui gravitent autour de pôles équivalents. Ces inventaires rendent compte des expansions récurrentes situées avant et après chaque forme-pôle. Sur le *tableau 3.13*, les segments recensés autour des formes équivalentes *cour/court* sont triés par ordre décroissant des fréquences, les segments de même fréquence étant triés par ordre de longueur décroissante. La confrontation des deux listes laisse apparaître des proximités dans les fréquences globales 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 correspondances sont d'autant plus mises en relief par le classement que le rang des fréquences des segments concernés est élevé [31].

Sur le *tableau 3.14*, la confrontation des listes des segments qui contiennent les occurrences des pôles bilingues *droits + homme* et *rights + human* renvoie à des locutions et à des associations syntagmatiques complexes présentant des rapports certains du point de vue de la traduction. Comme dans les exemples précédents,

[30] Sur les principes d'édition automatique des IVR, on consultera Salem [1987, pp. 139-142].

[31] Lorsque les effectifs sont faibles, il faut faire appel à des méthodes qui vont au-delà de la simple comparaison des fréquences globales. Les approches hybrides qui allient les méthodes de *topographie textuelle* et l'*analyse des spécificités* se révèlent adaptées à l'appariement des unités textuelles à basse fréquence. Ces approches font l'objet du chapitre 4.

la comparaison des fréquences totales des unités textuelles bilingues facilite le repérage des liens entre elles [32] :

Segment en français	**Fréq**	**Segment en anglais**	**Fréq**
droits de l *homme*	191	*human rights*	192 [33]
au palais des *droits* de l *homme*	61	in the *human rights* building	61
la commission européenne des *droits* de l *homme*	47	the european commission of *human rights*	47
sauvegarde des *droits* de l *homme*	23	protection of *human rights*	23

Nous avons vu plus haut comment il est possible de regrouper les segments liés sur le plan de la traduction par l'édition sélective des *inventaires des voisinages récurrents* (IVR). Il existe une autre approche formelle du paradigme constitué par l'ensemble des formes situées avant ou après une forme donnée. Les *inventaires distributionnels des expansions récurrentes avant/après une forme-pôle* (<-IDER/IDER->) permettent d'étudier l'environnement syntagmatique d'un couple de formes équivalentes. Pour ce type d'inventaire tous les segments contraints dans des segments plus longs (qui surchargeraient l'inventaire en apportant une information très redondante) ont été systématiquement écartés [34].

On peut voir sur les *tableaux 3.16ab* un inventaire distributionnel des expansions récurrentes droites des formes françaises *nécessaire(s)* et celui de la forme anglaise *necessary* (F=277).

---

[32] La mesure de longueur (nombre de formes dans le segment) peut servir de critère supplémentaire pour départager les polyformes en concurrence lors de l'appariement. Bien que des similitudes dans les longueurs de segments en équivalence ne soient pas absolues, il s'agit, néanmoins, d'un indice susceptible de compléter les informations fréquentielles.

[33] La différence entre les fréquences des segments *droits de l'homme* (F=191) et *human rights* (F=192) s'explique par la présence d'une citation en anglais dans le volet français du corpus. Cette citation mentionne un document juridique anglais dont le titre n'a pas été traduit :

français	***anglais***
on a remarqué à juste titre (voir andrew drzemczewski, ***the position of aliens in relation to the european convention on human rights****, conseil de l'europe, strasbourg, 1985, p-9)/…/*	*attention has rightly been drawn (see andrew drzemczewski,* ***''the position of aliens in relation to the european convention on human rights''****, council of europe, strasbourg, 1985 , p-9) /…/*

[34] Les techniques de comptages sur les surfaces sélectionnées (y compris la réalisation des *inventaires des expansions récurrentes*) sont exposées dans Salem [1987, pp. 134-142].

Tableau 3.14 : Segments répétés recensés autour des pôles bilingues *droits+homme* et *human+rights*

L	F	Segment en français	Segment en anglais	F	L
4	191	**droits** de l **homme**	**human rights**	192	2
5	182	des **droits** de l **homme**	of **human rights**	96	3
6	70	européenne des **droits** de l **homme**	the **human rights**	68	3
7	61	au palais des **droits** de l **homme**	in the **human rights**	62	4
7	48	commission européenne des **droits** de l **homme**	in the **human rights** building	61	5
8	47	la commission européenne des **droits** de l **homme**	european commission of **human rights**	48	5
9	32	par la commission européenne des **droits** de l **homme**	the european commission of **human rights**	47	6
5	26	**droits** de l **homme** et	by the european commission of **human rights**	32	7
9	24	des **droits** de l **homme** et des libertés fondamentales	and delivered at a public hearing in the **human rights** building	31	11
6	23	sauvegarde des **droits** de l **homme**	took place in public in the **human rights** building	29	9
7	21	de sauvegarde des **droits** de l **homme**	the hearing took place in public in the **human rights** building	28	11
10	21	sauvegarde des **droits** de l **homme** et des libertés fondamentales	**human rights** and	27	3
11	20	de sauvegarde des **droits** de l **homme** et des libertés fondamentales	**human rights** and fundamental freedoms	26	5
6	20	des **droits** de l **homme** à	of **human rights** and	25	4
11	18	la convention de sauvegarde des **droits** de l **homme** et des	of **human rights** and fundamental freedoms	24	6
9	18	au palais des **droits** de l **homme** à strasbourg	protection of **human rights**	23	4
7	15	cour européenne des **droits** de l **homme**	convention for the protection of **human rights**	22	7
11	13	de la convention de sauvegarde des **droits** de l **homme** et	protection of **human rights** and fundamental freedoms	22	7
9	11	au palais des **droits** de l **homme** à strasbourg	convention for the protection of **human rights** and fundamental freedoms	21	10
8	10	la cour européenne des **droits** de l **homme**	court of **human rights**	19	4

Tableau 3.14 : (suite)

L	F	Segment en français	Segment en anglais	F	L
9	8	de la commission européenne des **droits** de l **homme**	the convention for the protection of **human rights** and fundamental freedoms	17	11
11	7	la compétence de la commission européenne des **droits** de l **homme**	european court of **human rights**	15	5
8	7	la convention européenne des **droits** de l **homme**	of the convention for the protection of **human rights** and fundamental	13	11
10	5	avait saisi la commission européenne des **droits** de l **homme**	the european court of **human rights**	10	6
6	5	respect des **droits** de l **homme**	on **human rights**	10	3
			the european convention on **human rights**	9	6
			of the european commission of **human rights**	8	7
			the competence of the european commission of **human rights**	7	9
			for human **rights**	6	3
			the competence of the european commission of **human rights** to receive	5	11
			of the european convention for the protection of **human rights**	5	10
			lodged with the european commission of **human rights**	5	8
			of the european convention on **human rights**	5	7
			of the **human rights**	5	4
			human rights as	5	3

Guide de lecture : Les segments sont triés par ordre de fréquences décroissantes. Sur ce tableau, seuls les segments dont la fréquence dépasse le seuil de 4 sont représentés.

Tableau 3.15 : Segments répétés recensés autour des formes bilingues *démocratique(s)/democratic*

L	F	Segment en français	Segment en anglais	F	L
2	61	société **démocratique**	a **democratic**	63	2
3	57	une société **démocratique**	**democratic** society	61	2
4	52	dans une société **démocratique**	a **democratic** society	57	3
3	18	libéral et **démocratique**	in a **democratic**	53	3
5	15	nécessaire dans une société **démocratique**	in a **democratic** society	52	4
5	14	régime fondamental libéral et **démocratique**	necessary in a **democratic** society	45	5
6	10	le régime fondamental libéral et **démocratique**	the free **democratic**	20	3
9	8	libéral et **démocratique** au sens de la loi fondamentale	**democratic** constitutional	17	2
			the free **democratic** constitutional	15	4
			the free **democratic** constitutional system	14	5
			law and are necessary in a **democratic** society	11	8
			as are prescribed by law and are necessary in a **democratic**	9	11
			necessary in a **democratic** society in	8	6

Guide de lecture : Les segments sont triés par ordre de fréquences décroissantes. Sur ce tableau, seuls les segments dont la fréquence dépasse le seuil de 7 sont représentés.

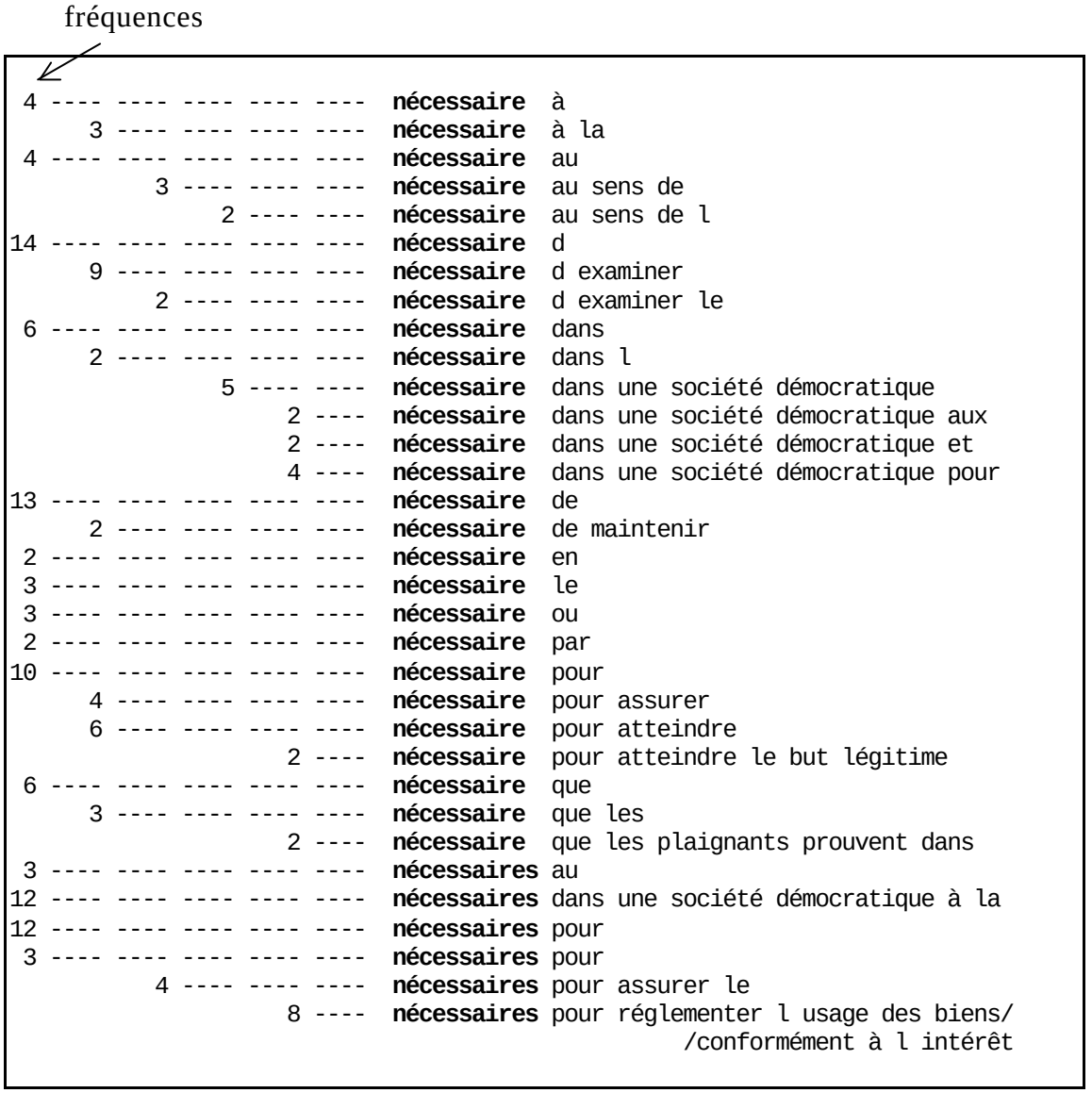


Figure 3.16a :
Inventaires distributionnels des expansions récurrentes après le pôle français *nécessaire(s)*

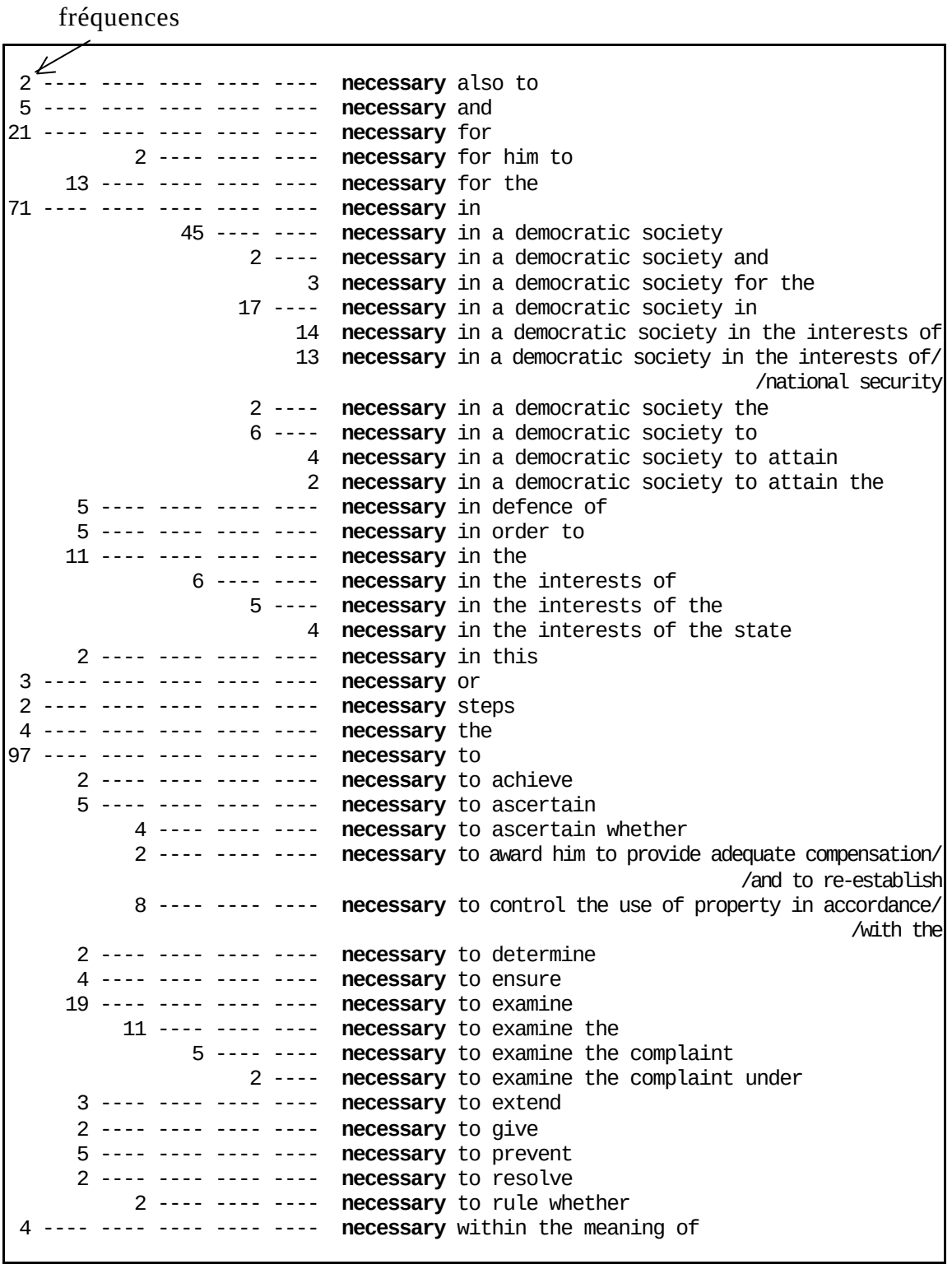


Figure 3.16b :

Inventaires distributionnels des expansions récurrentes après le pôle anglais*necessary*

Dans le corpus ***Convention***, la fréquence globale de la forme anglaise *necessary* (F=277) est supérieure à la fréquence cumulée des formes françaises *nécessaire* et *nécessaires* (F=199) car elle possède d'autres traductions dans le corpus.Une étude comparative des expansions récurrentes met en évidence un certain nombre de contextes où est attestée la correspondance *nécessaire(s)* / *necessary* (cf. *tableau 3.16ab*) :

Segment en français	**Fréq**	**Segment en anglais**	**Fréq**
nécessaire dans une société démocratique et	2	**necessary** in a democratic society and	2
nécessaire au sens de	3	**necessary** within the meaning of	4
nécessaire pour assurer	4	**necessary** to ensure	4
nécessaires pour réglementer l'usage des biens conformément à l'intérêt	8	**necessary** to control the use of property in accordance with the	8

Les différences entre les deux listes méritent également d'être étudiées avec attention. Elles peuvent servir de points de départ pour repérer des contextes où la correspondance des deux pôles bilingues n'est pas attestée. Par exemple, le segment *necessary to ascertain* n'a pas d'équivalent dans l'inventaire des expansions récurrentes recensées autour du pôle *nécessaire(s)*. Le retour au contexte a laissé apparaître des traductions telles que : *il échet de déterminer* / *it is necessary to ascertain* ; *il y a lieu de rechercher* / *it is necessary to ascertain*.

Dans l'exemple qui suit, les segments issus des inventaires du volet français présentés sur le *tableau 3.15* possèdent trois formes graphiques en commun :

libéral et **démocratique**	L = 3	F = 18
régime fondamental libéral et **démocratique**	L = 5	F = 14
le *régime fondamental libéral et* **démocratique**	L = 6	F = 10

Les segments du volet anglais qui s'en rapprochent sur le plan de l'équivalence traductionnelle n'ont pas la même structure :

the free **democratic**	L = 3	F = 20
democratic *constitutional*	L = 2	F = 17
the free **democratic** *constitutional*	L = 4	F = 15
the free **democratic** *constitutional* system	L = 5	F = 14

Le retour au contexte montre que les répétitions des formes graphiques n'ont pas forcément les mêmes origines dans les deux volets du corpus. Elles dépendent des contraintes syntaxiques propres à chacune des langues, de la variation lexicale, des choix sémantiques effectués par le traducteur, etc. :

français	**anglais**
*/…/ or ceux du dkp, tels qu'ils étaient énoncés dans le programme de mannheim (paragraphe 22 ci-dessous), allaient clairement à l'encontre **du régime fondamental libéral et démocratique** de la république fédérale d'allemagne.*	*/…/ the dkp's aims, as described in the mannheim programme of (see paragraph 22 below), were clearly opposed to **the free democratic constitutional system** of the federal republic of germany.*
*/…/ selon ces dispositions, un fonctionnaire devait constamment professer **le régime libéral et démocratique** au sens de la loi fondamentale et en défendre le maintien.*	*/…/ under those provisions, civil servants must at all times bear witness to **the free democratic constitutional system** within the meaning of the basic law and uphold that system.*
*/…/ elle signifie plutôt le devoir d'être prêt à s'identifier à l'idée de l'etat que le fonctionnaire doit servir, **au régime fondamental libéral et démocratique** de cet etat fondé sur la prééminence du droit et la justice sociale.*	*/…/ it means being prepared to identify with the idea of the state which the official has to serve and with **the free democratic constitutional order** of that state based on the rule of law and social justice.*

L'appel à des ressources dictionnairiques est envisageable pour préciser la mise en correspondance des mots et des syntagmes rapprochés automatiquement par la méthode des segments répétés. En outre, comme on le verra plus loin, l'analyse des distributions de segments au sein du corpus peut faciliter la description automatique d'équivalences traductionnelles entre les unités de ce même type [35].

[35] Sur le plan statistique, les ventilations des segments répétés au sein de corpus parallèles partitionnés offrent des indices supplémentaires pour l'appariement. Ces méthodes sont présentées au chapitre 5.

Nous constatons que la méthode de segments répétés constitue un outil adapté à la mise en évidence des récurrences d'unités textuelles en correspondance de traduction. En sélectionnant les segments bilingues dans lesquels sont attestées des formes équivalentes, il devient possible d'obtenir un « effet visuel » qui met en valeur des ressources traductionnelles du corpus lors de l'édition automatique des listes.

Les données recueillies à partir des comptages en segments répétés peuvent être utilisées avec profit pour préciser les résultats des analyses statistiques opérées à partir des formes graphiques. Les inventaires des segments répétés nous renseignent sur la structure des équivalences lexicales du corpus. Cependant, ils ne constituent pas à eux seuls un instrument d'appariement automatique. Dans le chapitre qui suit, nous montrerons que les constats produits par la méthode de segments répétés peuvent être affinés par la navigation textométrique dans le bi-texte.

Chapitre 4

L'exploration textométrique intertextuelle

Dans ce chapitre, nous présentons les résultats d'une série d'expériences consacrées au développement d'une nouvelle famille d'outils d'exploration textométrique intertextuelle, cf. Zimina [2002], [2004]. Fondée sur l'utilisation de la *topographie textuelle*, l'approche que nous développons permet d'extraire les *ressources traductionnelles* des corpus parallèles multilingues en utilisant la notion de *résonance textuelle*.

4.1 Les variations de fréquence dans les corpus parallèles

Nous avons montré au chapitre précédent que l'analyse des fréquences totales (*F*) des unités textuelles (formes, segments répétés) dans les volets bilingues d'un corpus parallèle aide à repérer des correspondances lexicales. Le recours à des comparaisons statistiques de la ventilation de ces mêmes unités dans le corpus fragmenté en parties offre des moyens plus sûrs pour explorer l'univers des équivalences traductionnelles. La partition des corpus parallèles permet de suivre l'évolution des rapports de correspondances entre les unités lexicales en fonction des variations thématiques au sein des parties. De façon générale, cette technique est employée lorsque l'on cherche à affiner la description sémantique des équivalences traductionnelles en corpus.

4.1.1 Les partitions de corpus

Le découpage simultané des différents volets d'un corpus parallèle en parties qui se correspondent peut être effectué à partir de la partition « naturelle » de ce corpus en *chapitres* ou *sections* issues des textes parallèles qui le composent. On parvient à affiner cette partition en segmentant les deux volets du corpus en *paragraphes* ou en *phrases*. Le découpage en paragraphes est souvent identique dans les textes originaux et leurs traductions. Les paragraphes sont identifiables dans les textes encodés sur support lisible par un ordinateur en s'appuyant sur la présence d'un caractère délimiteur (comme, par exemple, un *retour-chariot*).

L'alignement d'un corpus parallèle au niveau de la phrase facilite considérablement l'accès aux appariements lexicaux par le recours aux profils de ventilation des unités textuelles au sein des phrases équivalentes[1]. La localisation des unités textuelles au sein des phrases équivalentes sert aussi de point de départ à une deuxième étape qui verra la sélection des termes que leur *présence* (ou leur *absence*) rend caractéristiques pour un groupe de phrases sélectionné en fonction des objectifs spécifiques de l'étude.

4.1.2 Les statistiques par partie

Les deux volets d'un corpus parallèle partitionné se prêtent plus aisément à des comparaisons statistiques. Le point de départ des différentes méthodes statistiques qui permettent une étude comparative des parties du corpus est un tableau à double entrée que l'on appel *tableau lexical*. Il est établi en croisant les parties du corpus et les différents *types* qui constituent le système d'unités préalablement choisies. Les nombres à l'intersection des lignes et des colonnes de ce tableau correspondent aux *sous-fréquences* des unités textuelles dans les différentes parties du corpus. Un tableau lexical est généré pour chaque volet du corpus parallèle.

[1] Rappelons qu'il existe actuellement un certain nombre de méthodes permettant d'apparier les phrases d'un corpus parallèle avec un taux de réussite élevé, cf. Véronis [2000ab].

Pour apprécier les distributions des unités textuelles au sein du corpus parallèle partitionné, il faut les comparer avec l'ensemble des unités de même type. Selon notre hypothèse, les ventilations des unités textuelles en correspondance de traduction tendent à se ressembler. Dans le cas général, le rapprochement des unités bilingues avec des profils de distribution similaires est susceptible de laisser apparaître des équivalences lexicales. On peut vérifier cela, pour le corpus ***Convention***, en se reportant au graphique présenté sur la figure 4.3. Sur ce graphique, on a représenté simultanément la ventilation de la forme française *article* (F=351) et de la forme équivalente en anglais *article* (F=362). Pour cette expérimentation, les deux volets du corpus ont été découpés en vingt parties consécutives. Chacune d'entre elles compte le même nombre de paragraphes (cf. *Figure 4.1*). On constate sur le graphique des proximités importantes dans les distributions des formes équivalentes qui prouvent l'intérêt de cette comparaison pour l'appariement (cf. *Figure 4.2*)[2].

Lorsqu'une forme polysémique reçoit plusieurs traductions dans le corpus, il faut tenir compte d'un ensemble d'unités complexe. Ainsi, la confrontation simultanée du profil de ventilation de la forme anglaise *applicant* (F=1244) et du profil cumulé de la ventilation de ces traductions en français *requérant(e)* (F=971) et *intéressé(e)* (F=263) augmente considérablement la ressemblance des graphiques (cf. *Figures 4.3-4).*

[2] Comme nous le verrons au chapitre 5, la comparaison des profils de ventilation des unités textuelles bilingues d'un corpus parallèle peut être automatisée à base de la classification automatique des formes et des segments répétés.

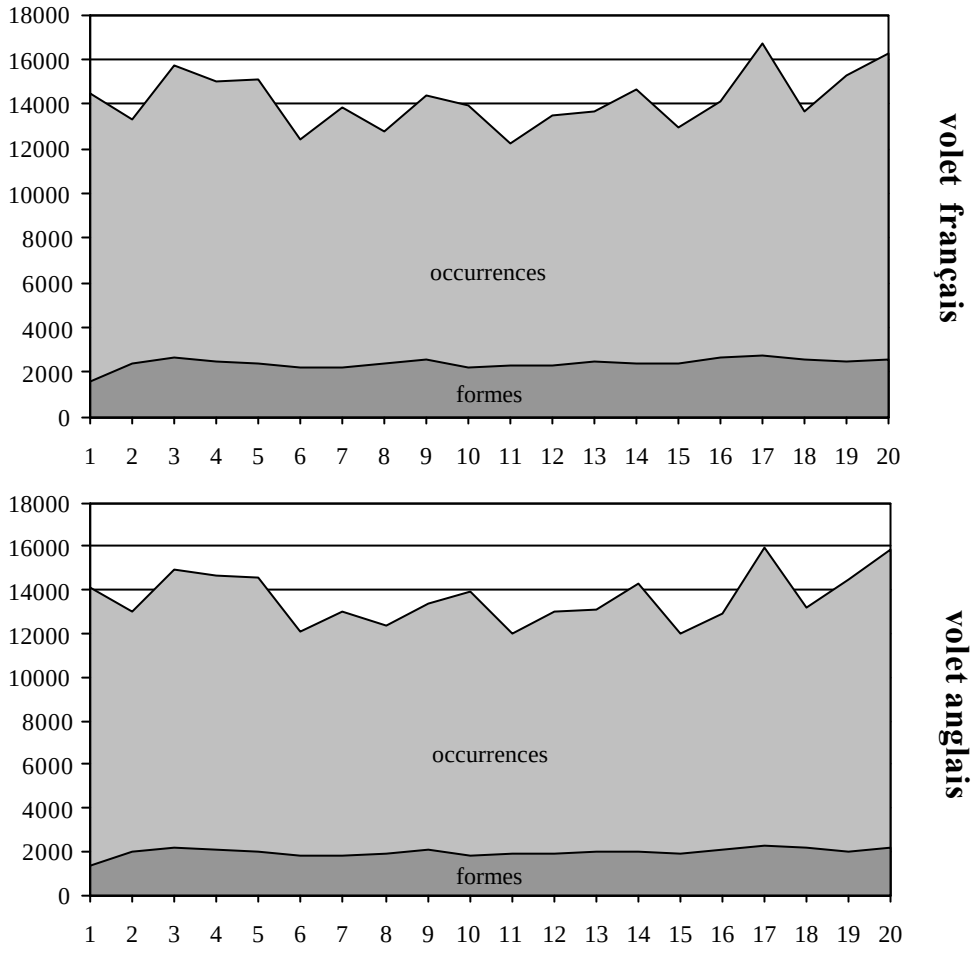


Figure 4.1 :
L'évolution parallèle du nombre d'*occurrences* et du nombre de *formes* dans les vingt parties consécutives du corpus bilingue ***Convention***

Guide de lecture : le corpus ***Convention*** a été divisé en vingt parties consécutives contenant le même nombre de paragraphes. Chaque partie ainsi obtenue compte approximativement 14 000 occurrences. Le graphique présenté sur la figure 4.4 permet de vérifier l'évolution parallèle du nombre des formes différentes par partie dans les deux volets du corpus, le volet français étant au total plus long en terme d'occurrences [296 396 contre 284 958 pour le volet *anglais*] et plus diversifié en formes employées [12 913 > 9 530] (cf. *Chapitre 3*).

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
article*	125	59	7	7	0	11	8	19	35	11	4	13	13	4	2	2	5	7	9	10
article	127	59	10	9	0	11	7	15	37	14	4	10	12	4	2	3	10	9	10	9

L'astérisque(*) indique la forme en anglais

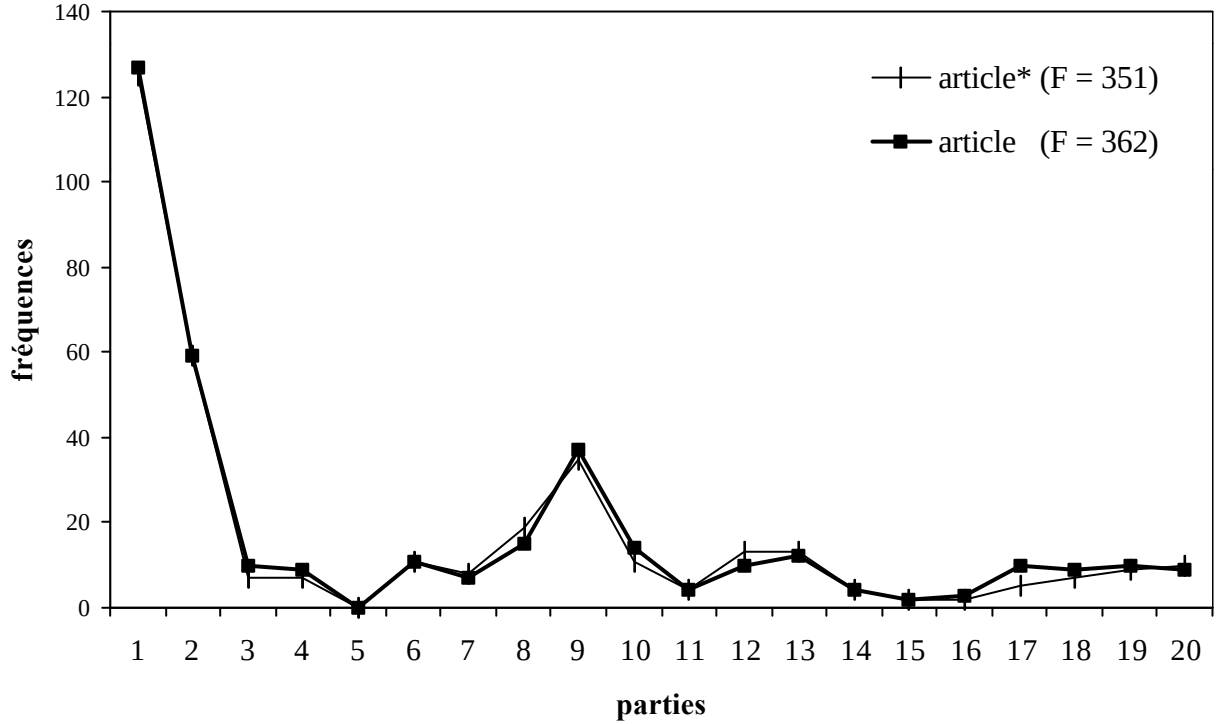


Figure 4.2 : Les ventilations des formes équivalentes *article* / article* dans les vingt parties consécutives du corpus ***Convention***

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
applicant	6	62	47	47	77	119	92	42	41	68	66	67	27	120	6	0	34	141	123	59
requérant(e)	6	49	42	44	58	96	80	35	39	44	46	49	23	80	3	0	29	103	103	42
intéressé(e)	15	11	2	2	8	17	7	16	15	26	19	16	4	20	5	2	5	24	16	33

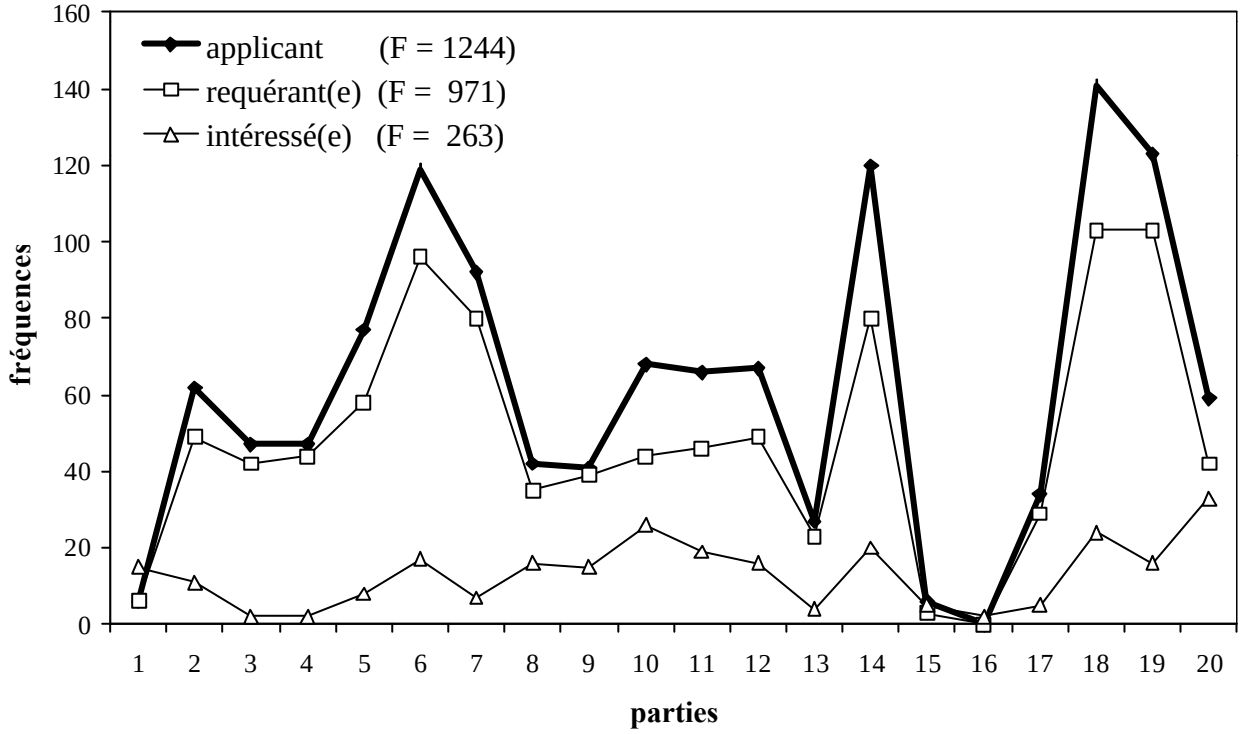


Figure 4.3 :
La ventilation de la forme anglaise *applicant* et des unités françaises équivalentes *requérant(e)* et *intéressé(e)* dans le corpus ***Convention***

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
applicant	6	62	47	47	77	119	92	42	41	68	66	67	27	120	6	0	34	141	123	59
requérant(e) / ***intéressé(e)***	31	60	44	46	66	113	87	51	54	70	65	65	27	100	8	2	34	127	119	75

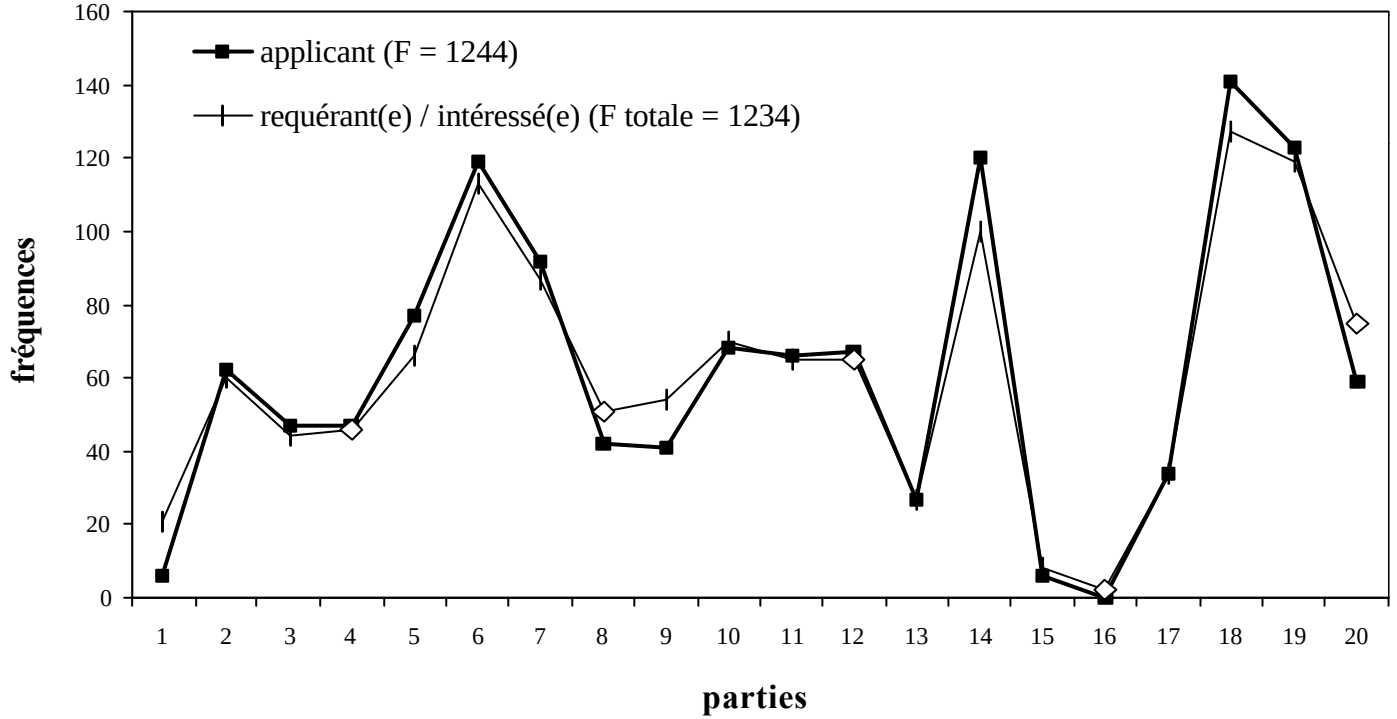


Figure 4.4 : La ventilation de la forme anglaise *applicant* et la ventilation cumulée des unités françaises équivalentes *requérant(e)* et *intéressé(e)* dans les vingt parties consécutives du corpus ***Convention***

4.2 Alignement lexical et résonance textuelle

Lors de l'étude simultanée des traductions d'un même texte, il est utile de considérer les variations conjointes de différentes unités textuelles dans les deux volets du corpus. Le terme de ***résonance textuelle*** [Salem, 2004] permet de décrire les perspectives de recherche d'un nouveau domaine de la textométrie lié à l'analyse de textes dont chacun entretient avec l'autre des rapports étroits [3].

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

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

- On repère, dans l'un des volets du corpus bilingue, des paragraphes (ou des phrases) dans lesquels la fréquence locale (*f*) d'une forme, d'un segment répété ou de toute autre unité textuelle dépasse un seuil fixé ;
- Le fragment sélectionné dans le premier volet provoque la sélection par *résonance* des paragraphes (ou des phrases) correspondants dans l'autre volet du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence, dans l'immense majorité des cas des expressions qui sont liées sur le plan de la traduction à la première expression.

Une ***sélection topographique*** se réalise à partir de considérations portant sur la localisation des unités bilingues dans les fragments textuels appariés (phrases, paragraphes, sections). L'ensemble de ces fragments est alors présenté sous

[3] Le schéma de la *résonance textuelle* permet de décrire des rapports de correspondance entre des ensembles textuels de nature différente : traductions mutuelles, tours de parole polémiques, données d'acquisition, etc. Voir, par exemple, Sansonetti [2004].

forme d'une *carte des sections parallèles*. La cartographie de la présence/absence des unités textuelle bilingues dans les sections appariées permet le repérage des liens de correspondances entre elles. La sélection topographique est particulièrement adaptée à l'étude des équivalences lexicales de basse fréquence.

Dans ce qui suit, nous allons présenter des applications du concept de résonance textuelle à l'exploration d'équivalences lexicales du corpus parallèle ***Convention***.

4.2.1 Recensement d'équivalences lexicales par seuillage

Rappel sur la méthode des spécificités

Les fréquences des unités varient fortement dans le corpus et le repérage des proximités entre profils est beaucoup trop complexe pour un expert humain. Fort heureusement, on dispose d'outils statistiques d'analyse qui permettent de porter un jugement sur les variations fréquentielles que les unités subissent tout au long du corpus. La *méthode des spécificités* [Lafon, 1984] met en évidence pour chaque unité de décompte les parties de corpus dans lesquelles l'unité possède de nombreuses occurrences (*spécificités positives*) ainsi que celles où son effectif est au contraire anormalement faible (*spécificités négatives*). En s'appliquant successivement à chacune des cases du tableau lexical (cf. *Figure 4.5*), on calcule le diagnostic de spécificité relatif à l'effectif constaté à base des paramètres suivants :

- k_{ij} - sous-fréquence de l'unité dans la partie ;
- F_i - fréquence de l'unité dans l'ensemble du corpus ;
- T_j - nombre des unités dans la partie ;
- T - nombre total des unités du corpus.

Un calcul probabiliste permet de porter un jugement sur l'effectif de la case du tableau analysée (k_{ij}) compte tenu des trois autres nombres (F_i, T_j, T). Si l'effectif k_{ij} se situe dans les limites de ce que le calcul permettait d'espérer, la répartition constatée est considérée « banale ». Si ce n'est pas le cas, on calcule un *indice de spécificité* de l'unité. Le *diagnostic* est fourni sous la forme **±*xx*** où :

- le signe (+ ou -) indique un *sur-emploi* ou un *sous-emploi* de l'unité dans la ou les partie(s) sélectionnée(s) par rapport à l'ensemble du corpus.
- ***xx*** est un indice de spécificité qui est d'autant plus élevé que la sous-fréquence analysée s'écarte d'une répartition « neutre » qui est sous-jacente au modèle des spécificités[4].

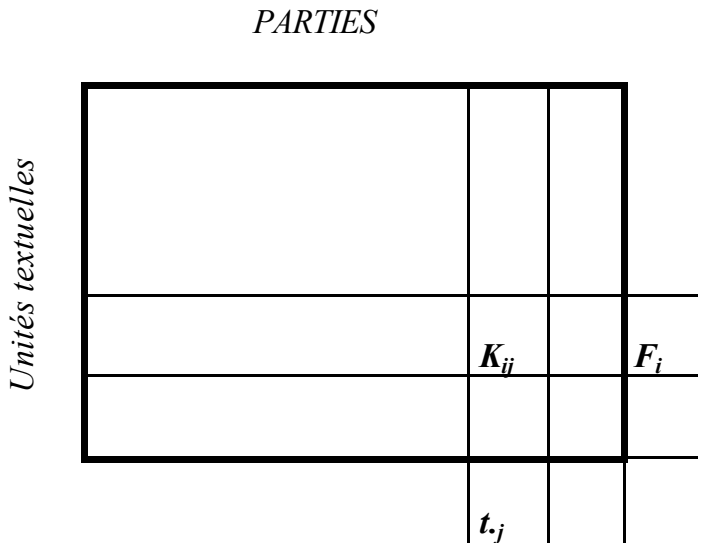


Figure 4.5 : Tableau lexical

Guide de lecture : A l'intersection de la ligne correspondant à l'unité i et de la colonne correspondant à la partie j, on trouve un nombre K_{ij} égal à la fréquence de l'unité j dans la partie i du corpus. La fréquence de l'unité i dans le corpus est égal à F_i. La longueur de la partie j est égal à t_j. Les totaux de chaque colonne représentent les nombres totaux d'unités utilisées dans chaque partie du corpus.

[4] Le modèle probabiliste utilisé ici pour évaluer la répartition est le *modèle hypergéométrique*, cf. Lafon [1984, pp. 54-68]. Sur la pratique du calcul des spécificités, on consultera également [Lebart et Salem, 1994, p. 172-176].

Les constats de spécificité établis pour une même unité à propos de chacune des parties du corpus permettent de décrire le comportement de cette unité au sein du corpus. Il est possible de considérer ces diagnostics parallèlement pour les unités issues des deux volets du corpus bilingue. Par exemple, on voit sur les graphiques présentés sur la figure 4.6 les diagnostics de spécificités obtenus sur chacun des volets du corpus ***Convention*** pour les types français/anglais *administr*+ et *administ*+ [5]. Chacun de ces types est constitué par l'ensemble d'occurrences des formes graphiques regroupées en raison de leur parenté sémantique dans le corpus :

administr+	**[483 occ.]**	***administ+***	**[484 occ.]**
administrant	1 occ.	administer	2 occ.
administrateur	6 occ.	administered	4 occ.
administrateurs	1 occ.	administering	1 occ.
administratieve	1 occ.	administratieve	1 occ.
administratif	90 occ.	administration	32 occ.
administratifs	21 occ.	administrations	1 occ.
administration	104 occ.	administrative	442 occ.
administrations	1 occ.	administrator	1 occ.
administrative	195 occ.		
administratives	58 occ.		
administre	2 occ.		
administrée	1 occ.		
administrer	2 occ.		

Les résultats du calcul indiquent qu'il existe des proximités importantes dans les répartitions des effectifs de ces types dans les deux volets du corpus (cf. *Figure 4.6*) [6]. On lit sur les graphiques que les types *administr*+ *et administ*+ sont *sur-représentés* dans la troisième partie du corpus (+21, +18). Ils sont au contraire *sous-représentés* dans la neuvième partie (-05). L'analyse de la

[5] Le langage des *expressions régulières* (cf. *Chapitre 3*: section 3.2) fournit des moyens formels pour mettre en évidence des ensembles de formes graphiques liées au plan lexical, telles que *administration*, *administratif*, *administrative*, *administrer*, *administrant,* 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.

[6] Pour cette première analyse, nous avons gardé le découpage initial du corpus en vingt parties consécutives de la même taille (cf. *Figure 4.1*).

traduction montre que les ventilations similaires des unités bilingues dans les zones correspondantes du bi-texte signale leur équivalence lexicale dans le corpus :

français	**anglais**
administr*ation de l'Etat*	***administ****ration of the State*
administr*ation de la justice*	***administ****ration of justice*
tribunal ***administr****atif*	***Administ****rative Court*
dossier ***administr****atif*	***administ****rative file*
administr*ation compétente*	*competent administrative authorities*
autorités compétentes chargées de l'enquête ***administr****ative ou de l'information judiciaire*	*independent administrative or prosecutorial authorities*
juridictions ***administr****atives*	***administ****rative courts*
avoir recours à la force	*to* ***administ****er the force*
contentieux ***administr****ative*	***administ****rative proceedings*

Nous constatons un léger écart dans les distributions des types *administr+ et administ+* dans la dixième partie du corpus (cf. *Figure 4.6*). Cet écart est en grande partie dû à la présence dans le corpus de l'équivalence français/anglais *l'****administr****ation des douanes – the customs*. Comme on le verra plus loin dans ce chapitre, l'approche topographique du bi-texte fournit une aide précieuse pour la découverte de rapports de correspondances lexicales entre les unités textuelles qui reçoivent plusieurs traductions dans le corpus.

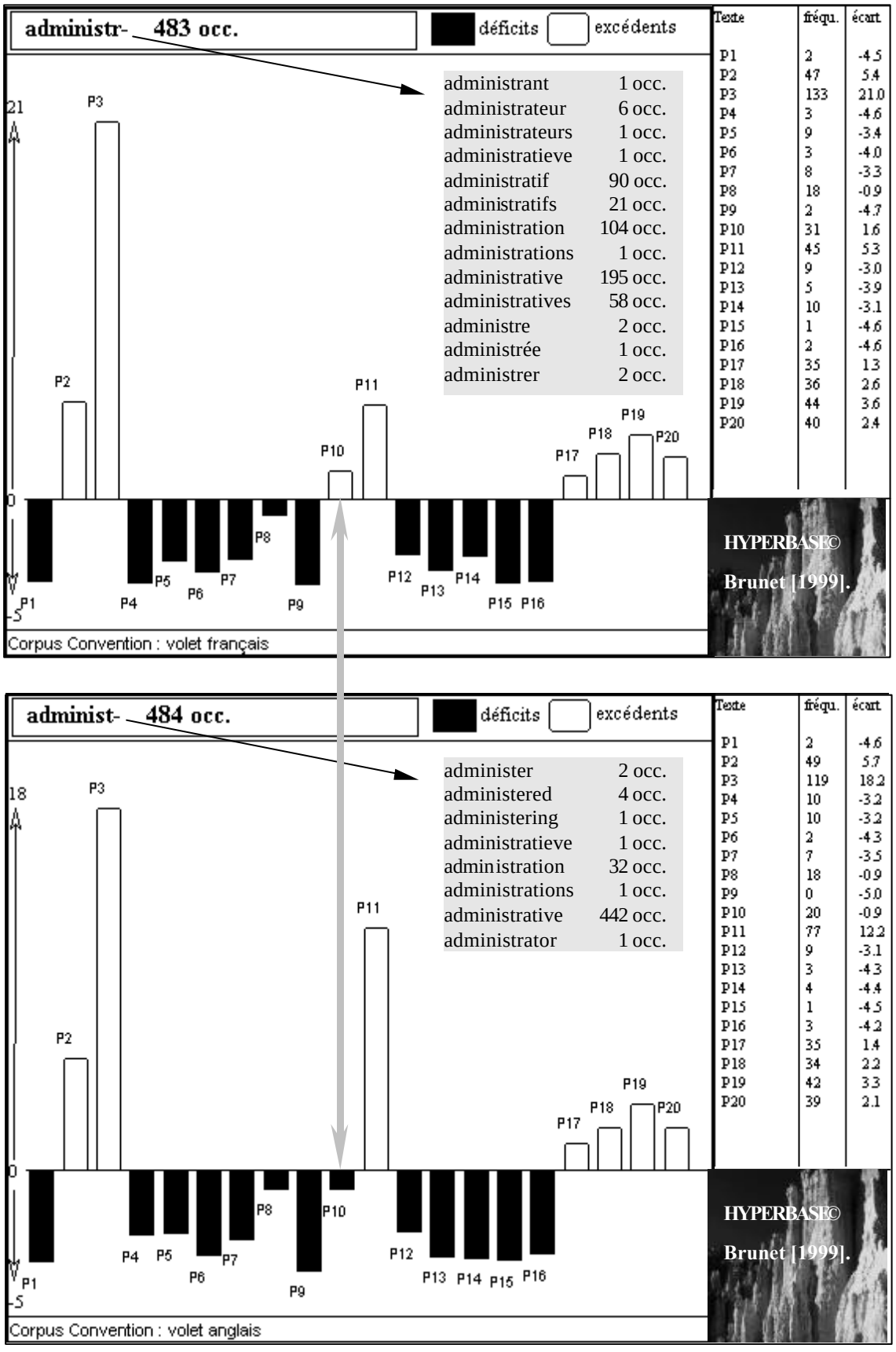


Figure 4.6 : Les types équivalents français/anglais ***administr-*** et ***administ-*** dans les vingt parties du corpus ***Convention***

Tableau 4.7 : Résultats du calcul des spécificités

Corpus ***Convention*** Partition = **20 parties**
Parties sélectionnées : **05, 06.**

volet français				*volet anglais*			
spec.	**F**	**f**	**forme**	**forme**	**f**	**F**	**spec.**
+E43	49	46	extradition	extradition	66	69	*+E63*
+E37	111	65	turquie	cyprus	49	58	*+E40*
+E31	44	37	chypre	declarations	42	59	*+E29*
+E28	13409	1676	la	turkey	59	122	*+E27*
+E21	328	92	requérante	international	56	127	*+E24*
+E20	19	19	extraditionnel	turkish	55	121	*+E24*
+E19	255	75	déclaration	circulation	20	20	*+E21*
+E18	1223	216	convention	declaration	73	225	*+E21*
+E18	83	39	restrictions	swiss	21	23	*+E19*
+E16	40	25	retrait	restrictions	42	93	*+E18*
+E15	20	17	suisse	withdrawal	22	27	*+E18*
+E14	13	13	écrou	convention	218	1228	*+E18*
+E13	12	12	valides	contracting	54	178	*+E14*
+E13	146	45	déclarations	detention	81	336	*+E14*
+E13	12	12	bluf	bluf	12	12	*+E13*
+E12	213	55	juridiction	clauses	12	12	*+E13*
+E12	30	19	internationale	association	30	73	*+E12*
+E12	68	28	contractantes	northern	24	51	*+E12*
+E11	10	10	territoriales	territorial	22	40	*+E12*
+E10	75	28	responsabilité	invalidity	10	10	*+E11*
+E10	314	68	détention	competence	28	73	*+E11*
+E10	16	12	cypriote	information	41	133	*+E11*
				continuing	15	24	*+E10*
				responsibility	20	40	*+E10*
				applicant	191	1244	*+E10*

Guide de lecture : Les unités spécifiques sont classées par ordre décroissant de spécificité. Chaque unité est accompagnée de sa fréquence totale (*F*) dans l'ensemble du corpus et de sa fréquence locale (*f*) dans les parties sélectionnées. Seules les *spécificités positives* majeures sont représentées.

Tableau 4.7 : (suite)

Corpus ***Convention*** Partition = **20 parties**
Parties sélectionnées : **09, 10.**

volet français				*volet anglais*			
spec.	**F**	**f**	**forme**	**forme**	**f**	**F**	**spec.**
+E43	143	81	accusation	indictment	82	112	*+E57*
+E32	58	44	contrainte	division	81	141	*+E44*
+E31	87	53	procureur	default	50	60	*+E40*
+E30	79	50	corps	prosecutor	58	107	*+E30*
+E27	50	37	trafic	imprisonment	65	140	*+E28*
+E26	48	36	stupéfiants	offences	54	104	*+E27*
+E23	76	43	accusé	judge	97	291	*+E27*
+E23	291	91	chambre	criminal	110	409	*+E22*
+E21	350	97	juge	trafficking	31	45	*+E21*
+E16	67	33	valeur	accused	50	112	*+E21*
+E15	80	35	infractions	drug	36	59	*+E21*
+E15	38	24	comparution	confiscation	35	63	*+E19*
+E14	51	27	condemnation	customs	32	64	*+E16*
+E14	119	43	instruction	remand	33	68	*+E16*
+E14	170	52	pénale	proceeds	17	21	*+E14*
+E13	34	21	athènes	value	23	38	*+E14*
+E13	51	26	entreprise	passages	18	23	*+E14*
+E13	17	15	drachmes	athens	21	34	*+E13*
+E13	105	39	confiscation	appear	36	95	*+E13*
+E12	167	48	ordonnance	drachmas	13	15	*+E12*
+E12	314	74	détention	maximum	17	24	*+E12*
+E11	26	17	passages	penalty	27	64	*+E11*
				fines	14	19	*+E11*

Guide de lecture : Les unités spécifiques sont classées par ordre décroissant de spécificité. Chaque unité est accompagnée de sa fréquence totale (*F*) dans l'ensemble du corpus et de sa fréquence locale (*f*) dans les parties sélectionnées. Seules les *spécificités positives* majeures sont représentées.

Les vocabulaires spécifiques des sections parallèles

Le regroupement des diagnostics de spécificité relatifs à des parties équivalentes du corpus parallèle ***Convention*** permet de porter un jugement sur les proximités traductionnelles de leurs *vocabulaires*. Le tableau 4.7 montre les résultats de la sélection par *seuillage* de formes caractéristiques issues des parties en correspondance. Les listes des spécificités obtenues fournissent une description thématique de ces fragments bi-textuels et permettent d'esquisser la structure sémantique de mise en rapport traductionnelle.

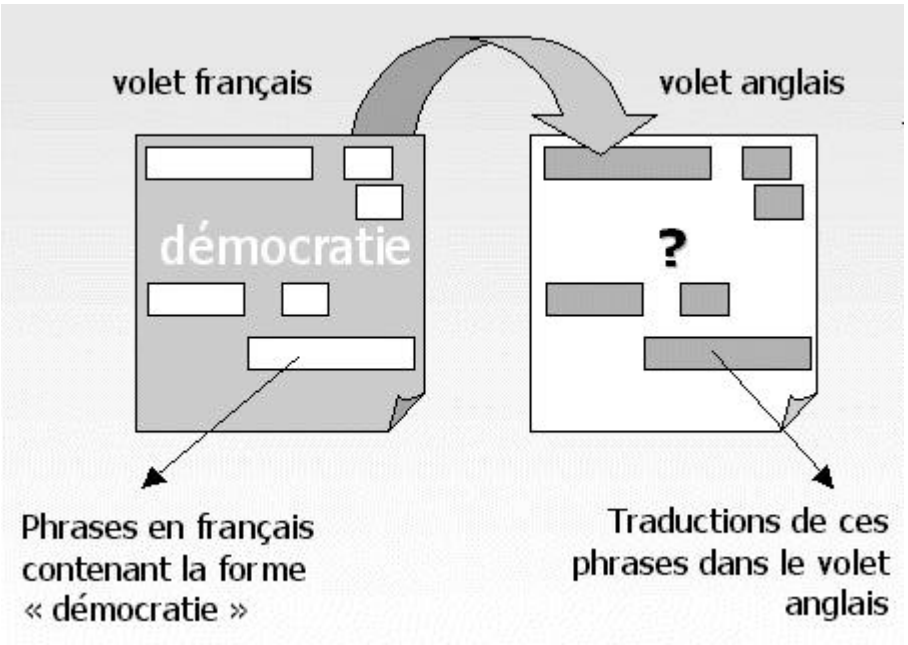
L'analyse des fragments caractéristiques du bi-texte

Nous avons montré que l'analyse des spécificités réalisée pour les parties consécutives découpées dans le corpus parallèle produit des indices permettant le repérage automatique d'équivalences lexicales. Les résultats sont particulièrement interprétables lorsque la fréquence des unités que l'on met en rapport est suffisamment élevée. Pour affiner la description d'équivalences parmi les unités de basse fréquence, on peut procéder au découpage de fragments textuels dont on souhaitera connaître le vocabulaire caractéristique selon des principes différents. Comme le montre le schéma sur la figure 4.8, les diagnostics portant sur la répartition des unités dans les fragments de texte équivalents choisis en raison de l'abondance relative des occurrences d'un *type* donné offrent des moyens supplémentaires pour l'appariement. Une fois repérées les unités dont les occurrences connaissent une abondance relative dans les fragments bi-textuels, on peut mettre en évidence des rapports de correspondance entre les unités textuelles bilingues par une série de comparaisons statistiques entre elles.

La localisation automatique de fragments bi-textuels équivalents peut être complétée par les résultats de l'alignement du corpus au niveau de la phrase. L'exploration débute par le marquage au fil du texte dans l'un des volets du bi-texte d'un sous-ensemble d'occurrences correspondant à un *type* quelconque [7]. Le repérage des phrases correspondantes dans l'autre volet permet de construire deux fragments de texte équivalents qui seront confrontés dans chacun des cas au

[7] On consultera Lamalle et Salem [2002] sur les principes de recensement automatique de *types généralisés* ***Tgen***.

reste du corpus à des fins de comparaison. Pour illustrer notre propos, nous considérerons les phrases du volet français du corpus ***Convention*** qui contiennent la forme *démocratie* (dont la fréquence totale *F* est égale à 9) et le sous-ensemble de phrases équivalentes dans le volet anglais (cf. *Figure 4.8*). Le calcul des spécificités permet de sélectionner parallèlement pour chacun de ces fragments une série de types caractéristiques de ces parties du corpus (cf. *Tableau 4.9*).



Le calcul des spécificités permet de sélectionner parallèlement pour chacun des fragments une série d'unités textuelles particulièrement caractéristiques

Figure 4.8 : Sélection des fragments caractéristiques du *bi-texte*

Guide de lecture : On repère dans le volet français du corpus bilingue ***Convention*** les phrases dans lesquelles sont attestées les occurrences d'une forme *démocratie*. Le fragment textuel sélectionné dans le volet français provoque la sélection par *résonance* des phrases correspondantes dans le volet anglais du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence des expressions qui sont liées sur le plan de la traduction à la forme *démocratie* (cf. *Figure 4.9*).

Les tableaux 4.9 présentent quelques-uns des types mis en évidence par la méthode statistique. On constate sur ces tableaux que les unités bilingues issues de l'exploration constituent bien des correspondances de traduction. Ainsi, la forme française *démocratie* (*spec.*+E27), ayant servi de point d'entrée pour la construction de l'échantillon de phrases pour cette analyse, peut être directement appariée avec la forme *democracy* (*spec.*+E27), la plus caractéristique du fragment anglais.

Lorsque l'équivalence de types bilingues est posée, la notion d'***Equivalence Traductionnelle Elémentaire*** **(ETE)** [8] permet de décrire la liste des adresses de leurs occurrences attestées dans les phrases équivalentes :

ETE : Equivalence Traductionnelle Elémentaire

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 unités 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 4.10 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) et le segment anglais *democracy capable of defending itself* (f=4, *spec.*+E12) se correspondent dans le corpus (cf. *Tableau 4.10*). Nous

[8] L'*équivalence traductionnelle élémentaire* (ETE) est le *type* d'unité bi-textuelle dont on peut recenser les occurrences dans un corpus parallèle. Sur le plan sémantique, il s'agit de l'unité d'équivalence attestée au niveau des mots et des syntagmes. La notion d'ETE est développée dans le *chapitre 7* (section 7.2.2).

observons que la valeur de l'indice de spécificité est liée au nombre de rencontres des types avec l'équivalence *démocratie*/*democracy* qui est la plus caractéristique du fragment.

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 4.9*).

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 cooccurrence des formes *démocratie* + *apte* + *défendre* (F=4) [9] 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 4.11*) [10].

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 4.12, lorsque la prise en compte des ventilations des types dans les phrases du fragment courant ne permet pas de départager les unités de même fréquence locale (*f*) [ex. *souhaitait* (F=8, f=1, *spec.*+E03), *fondant* (F=25, f=1, *spec*.+E03), *avoid* (F=13, f=1, *spec*.+E03),

[9] La fréquence d'une cooccurrence 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.

[10] 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 cooccurrences* (cf. *Chapitre 6*).

founding F=2, f=1, *spec*.+E03)], l'ambiguïté peut être levée si l'on prend soin de comparer les ventilations de ces mêmes unités dans tout le corpus.

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 4.12*). Notons que la mise en correspondance des hapax du corpus peut être facilitée si l'on fait appel aux ressources dictionnairiques.

Tableau 4.9 : Spécificités majeures

volet français					*volet anglais*				
orig.	**spec.**	**F**	**f**	**terme**	**terme**	**f**	**F**	**spec.**	**orig.**
*	*+E27*	9	9	démocratie	democracy	9	10	*+E26*	
*	*+E15*	5	5	de la démocratie	of democracy	4	4	*+E12*	*
*	*+E12*	4	4	démocratie apte à se défendre	democracy capable of defending itself	4	4	*+E12*	*
*	*+E12*	4	4	apte	of defending	4	5	*+E11*	
	+E11	22	5	défendre	defending	4	7	*+E11*	
	+E10	11	4	se défendre	capable of	4	34	*+E08*	
	+E08	29	4	à se	capable	4	34	*+E08*	
	+E06	3	2	instaurer une	nightmare	2	2	*+E06*	*
*	*+E06*	2	2	cauchemar	democracy and	2	3	*+E06*	
*	*+E06*	2	2	de la démocratie et	the nightmare of nazism	2	2	*+E06*	*
	+E06	3	2	la volonté d’	its constitution	2	3	*+E06*	
*	*+E06*	2	2	le cauchemar du nazisme	values of democracy	2	2	*+E06*	*
*	*+E06*	2	2	nazisme	being based	2	3	*+E06*	
					based on the principle	2	3	*+E06*	
					of democracy and	2	2	*+E06*	*
					on the principle	2	4	*+E06*	
					nazism	2	2	*+E06*	*
					constitution being based on the principle of	2	2	*+E06*	*

Tableau 4.9 : Spécificités majeures (suite)

orig.	spec.	F	f	terme
	+*E05*	5	2	volonté d’
	+*E05*	11	2	justifiant
	+*E05*	9	2	la volonté
	+*E05*	6	2	instaurer
	+*E05*	11	2	allemands
	+*E04*	15	2	valeurs
	+*E04*	38	2	allemagne
	+*E04*	27	2	idée
	+*E04*	20	2	mieux
	+*E04*	15	2	volonté
	+*E04*	33	2	particulière

terme	f	F	spec.	orig.
justifying	2	7	+*E05*	
founded	2	11	+*E05*	
values of	2	7	+*E05*	
imposed on	2	25	+*E04*	
values	2	15	+*E04*	
civil	4	302	+*E04*	
led to	2	23	+*E04*	
principle	3	103	+*E04*	
based on the	2	27	+*E04*	
itself	3	103	+*E04*	
germany	2	37	+*E04*	
led	2	30	+*E04*	
a special	2	14	+*E04*	
the principle of	2	35	+*E04*	
constitution1	3	18	+*E04*	
notion	2	20	+*E04*	

Guide de lecture du tableau : Un emploi caractéristique d’un type (forme, segment répété) dans les fragments de textes bilingues est indiqué par un *indice de spécificité*. Le fragment français correspond à l’échantillon des phrases où est attestée la forme *démocratie.* Le fragment anglais est composé des phrases équivalentes. 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. Seules les spécificités positives majeures sont représentées. L’astérisque (*) indique que le type n’est présent que dans le fragment de texte courant.

Tableau 4.10 : 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*.

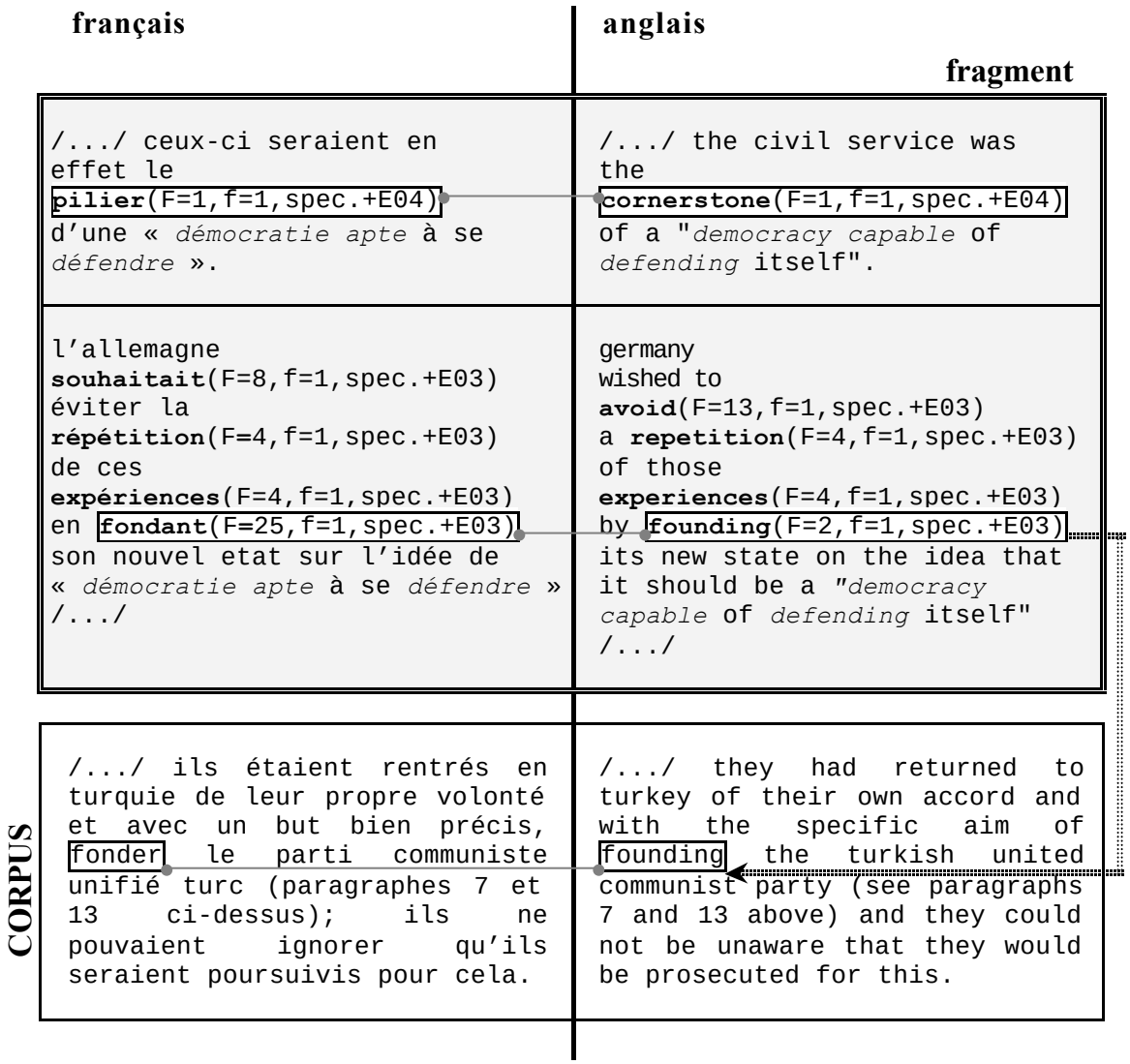
Tableau 4.11 : Cooccurrences multiples

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

Co-occurrents de ***democracy*** + ***defending*** + ***capable*** (*NbUC*=4)				
terme	**f**	**F**	**NbUC**	**spec.**
nazism	2	2	2	*+E07*
nightmare	2	2	2	*+E07*
itself	3	103	3	*+E05*
led	2	30	2	*+E05*
principle	2	103	2	*+E04*
republic	2	102	1	*+E04*
germany	2	37	2	*+E04*
founding	**1**	**2**	**1**	***+E04***
cornerstone	**1**	**1**	**1**	***+E04***
itself	1	1	1	*+E04*
its	3	739	3	*+E03*
being	2	283	2	*+E03*
based	2	130	2	*+E03*
constitution	2	118	2	*+E03*
notion	1	20	1	*+E03*
avoid	1	13	1	*+E03*
founded	1	30	1	*+E03*
idea	1	9	1	*+E03*
experiences	1	4	1	*+E03*
repetition	1	4	1	+E03

*_**NbUC**_ / Nombre d'unités de contextes / = nombre de phrases où est attestée la cooccurrence.

Tableau 4.12 : Retour au contexte



Guide de lecture : Le tableau permet de localiser dans le contexte quelques cooccurrents des pôles bilingues *démocratie + apte + défendre* et *democracy + defending + capable* (cf. *Tableau 4.9*). Un diagnostic est fourni pour chaque cooccurrent sous la forme (*F*,*f*, *spec*.) où *F* = fréquence totale, *f* = fréquence locale, *spec*. = indice de spécificité. 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.

4.2.2 Le repérage topographique d'équivalences de traduction

Construction de la carte de sections parallèles

Des comparaisons statistiques à base de corpus ont permis d'apparier les termes 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. Le repérage des équivalences peut être complété si l'on parvient à une description plus élaborée du pôle bilingue sur lequel s'appuie la sélection des phrases soumises au calcul des spécificités. Cette description peut être envisagée si l'on entreprend de donner une *représentation topographique* de corpus parallèles.

La *topographie textuelle* a pour objectif une localisation graphique des phénomènes mis en évidence par l'étude statistique. Dans le contexte de cette approche, l'expertise humaine de textes est appuyée de multiples outils de lecture et visualisation qui offrent de nouveaux moyens d'investigation de l'espace textuelle. Ainsi, l'étude de « l'organisation spatiale » d'occurrences d'équivalences lexicales peut être effectuée à partir de la *description cartographique* du *bi-texte.*

Fonctionnalités logicielles

Les fonctionnalités développées au sein du logiciel *Lexico3* [Lamalle *et al.*, 2003] permettent à l'utilisateur de visualiser une *carte des sections* [11] du logiciel *Lexico 3* (ex.: phrases ou paragraphes du corpus), puis de constituer une sélection arbitraire de sections dont on étudiera ensuite le vocabulaire spécifique.

L'utilisateur dispose d'un ensemble d'outils permettant de choisir (à partir du dictionnaire, du *garde-mots* [12], de la liste des *segments répétés*, etc.) un *type* sur lequel portera son exploration.

Après avoir sélectionné le type, il est possible de le faire glisser sur la carte (*glisser/déposer*). La ventilation du type étudié devient alors visible. Les sections dans lesquelles il est présent apparaissent en couleur [13]. Ce processus peut être réitéré.

Outils textométriques de navigation bi-textuelle

Sur la figure 4.13, la description cartographique des volets bilingues du corpus ***Convention*** divisé en phrases, traduit simultanément la ventilation des types équivalents français/anglais *démocrat+* et *democra+*. Chacun de ces types est constitué par l'ensemble des occurrences de formes graphiques liées au même thème au sein du corpus :

[11] La carte des sections permet une visualisation du corpus découpé en sections par la promotion d'un (ou de plusieurs) caractères particuliers au statut de délimiteurs de section (Lamalle *et al.*, 2003). Dans le cas des corpus parallèles, le découpage en sections peut être effectué parallèlement, en s'appuyant sur des codes attribués aux phrases en correspondance.

[12] Le *garde-mots* est une fonctionnalité de *Lexico3* permettant de mémoriser formes, segments, etc. pour une utilisation ultérieure. Pour stocker un type dans le *garde-mots* il suffit de le faire glisser sur l'icône de cette fonctionnalité.

[13] Il est possible d'obtenir un coloriage plus ou moins dense des sections en cochant d'abord la case *seuil*. Cette fonctionnalité permet de régler plusieurs seuils en probabilités qui entraîneront l'intensité variable du coloriage sur la carte.

démocrat+	[113 occ.]	***democra+***	[114 occ.]
démocratique	96 occ.	democratic	103 occ.
démocratie	9 occ.	democracy	10 occ.
démocratiques	7 occ.	democrat	1 occ.
démocrate	1 occ.		

Pour chaque volet du corpus, les carrés de couleurs sombres 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 4.13*).

La localisation des phrases équivalentes où sont présents parallèlement les types *démocrat+* et *democra+*, permet d'envisager une analyse 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 4.14*).

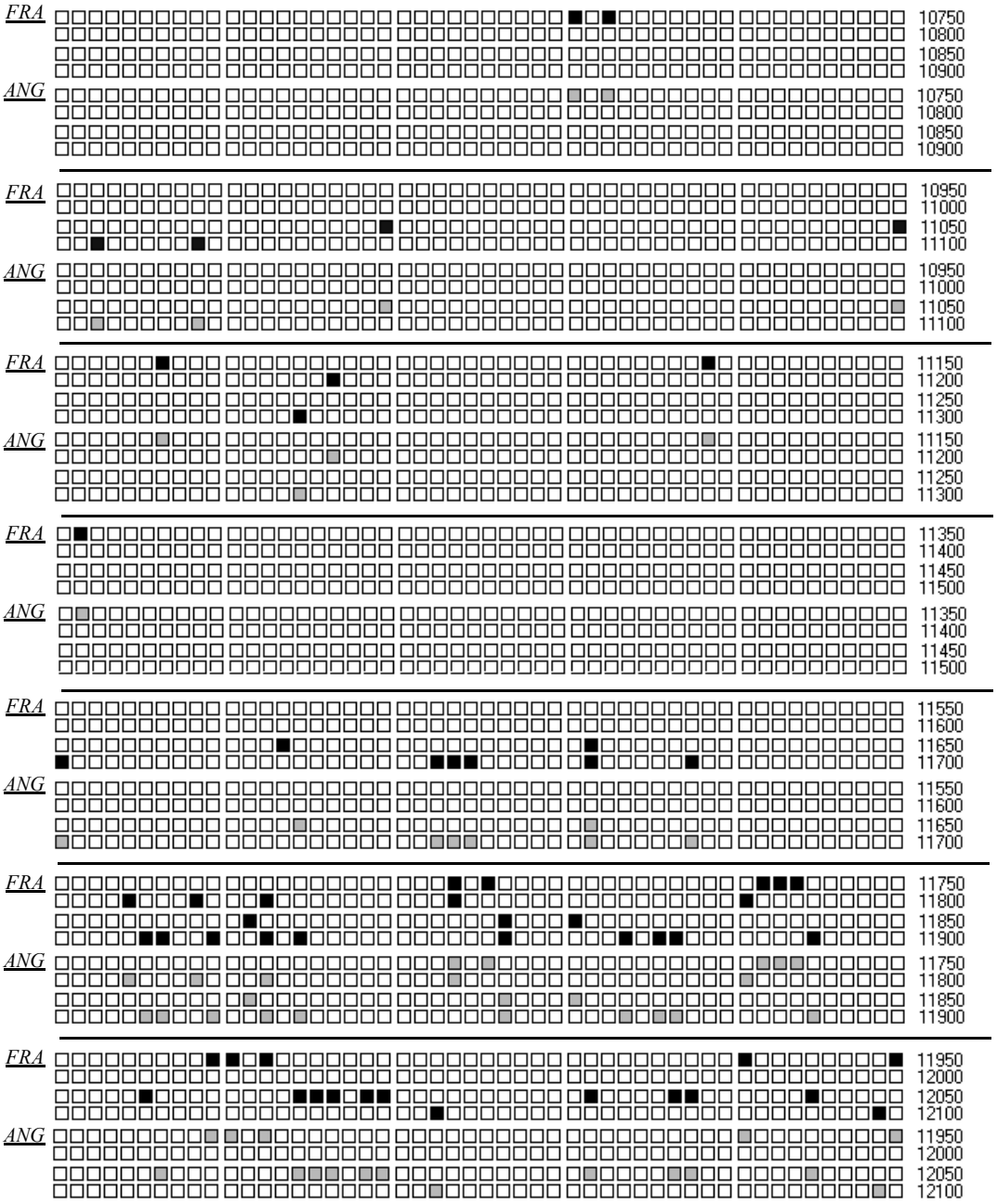


Figure 4.13 :
Les occurrences des types bilingues ■*démocrat*+ et □*democra*+ dans un extrait du corpus ***Convention***

Guide de lecture : La division du corpus en phrases est matérialisée par des carrés. Les carrés de couleurs sombres indiquent la présence du type concerné dans la phrase.

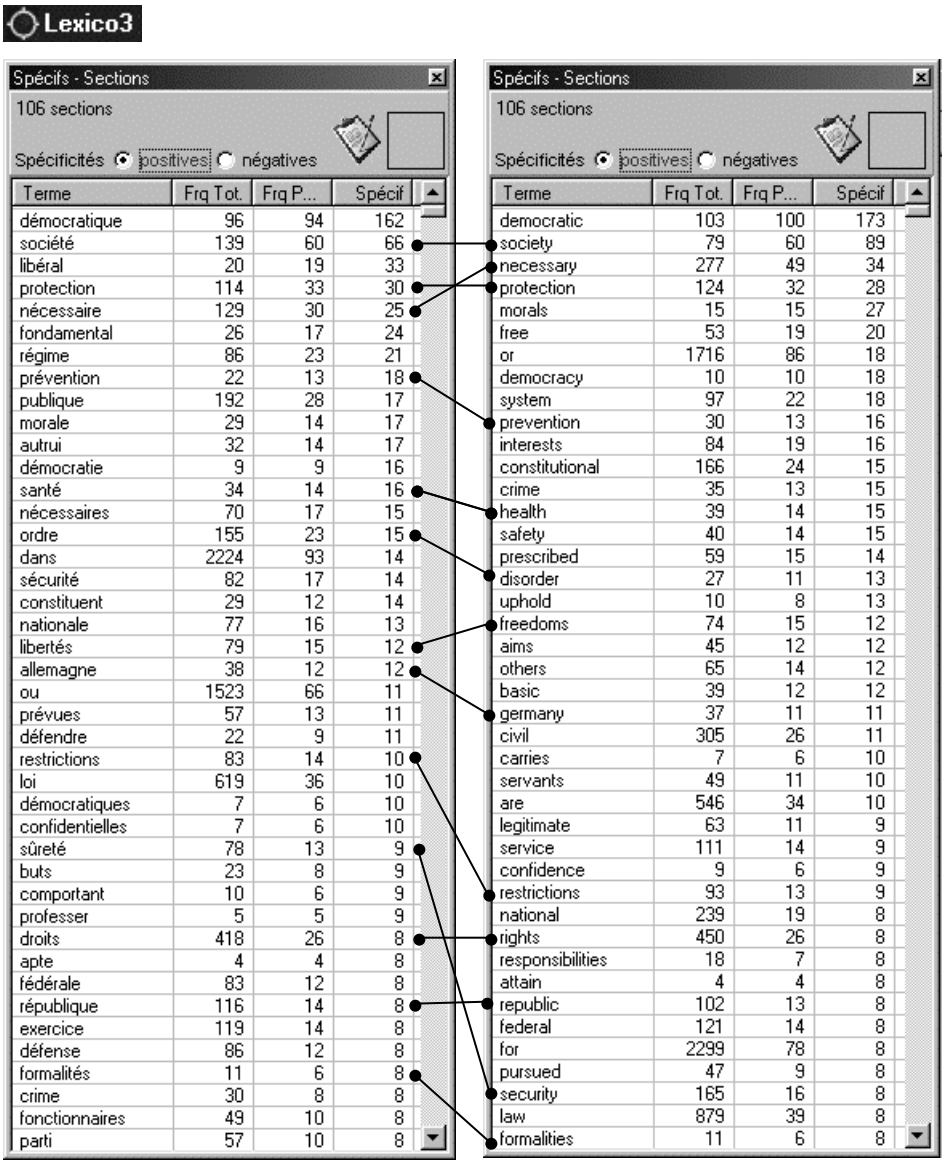


Figure 4.14 :
Vocabulaire caractéristique des phrases contenant le pôle *démocrat+/democra+*

Guide de lecture *:*

Grâce aux outils textométriques de la *carte des sections*, il est possible de sélectionner automatiquement l'ensemble des sections dans lesquelles le type étudié est présent. Sur la figure 4.14, le nombre des sections correspondant aux phrases contenant le pôle bilingue *démocrat+/democra+* apparaît en haut de la fenêtre (106 sections).

La méthode des *spécificités* permet de recenser parallèlement pour chacun des fragments ainsi constitués une série d'unités textuelles caractéristiques de ces fragments du corpus. Sur la figure, la confrontation des diagnostics obtenus pour chacun des volets du corpus ***Convention*** met en évidence des équivalences spécifiques de l'univers lexical du pôle bilingue *démocrat+/democra+*.

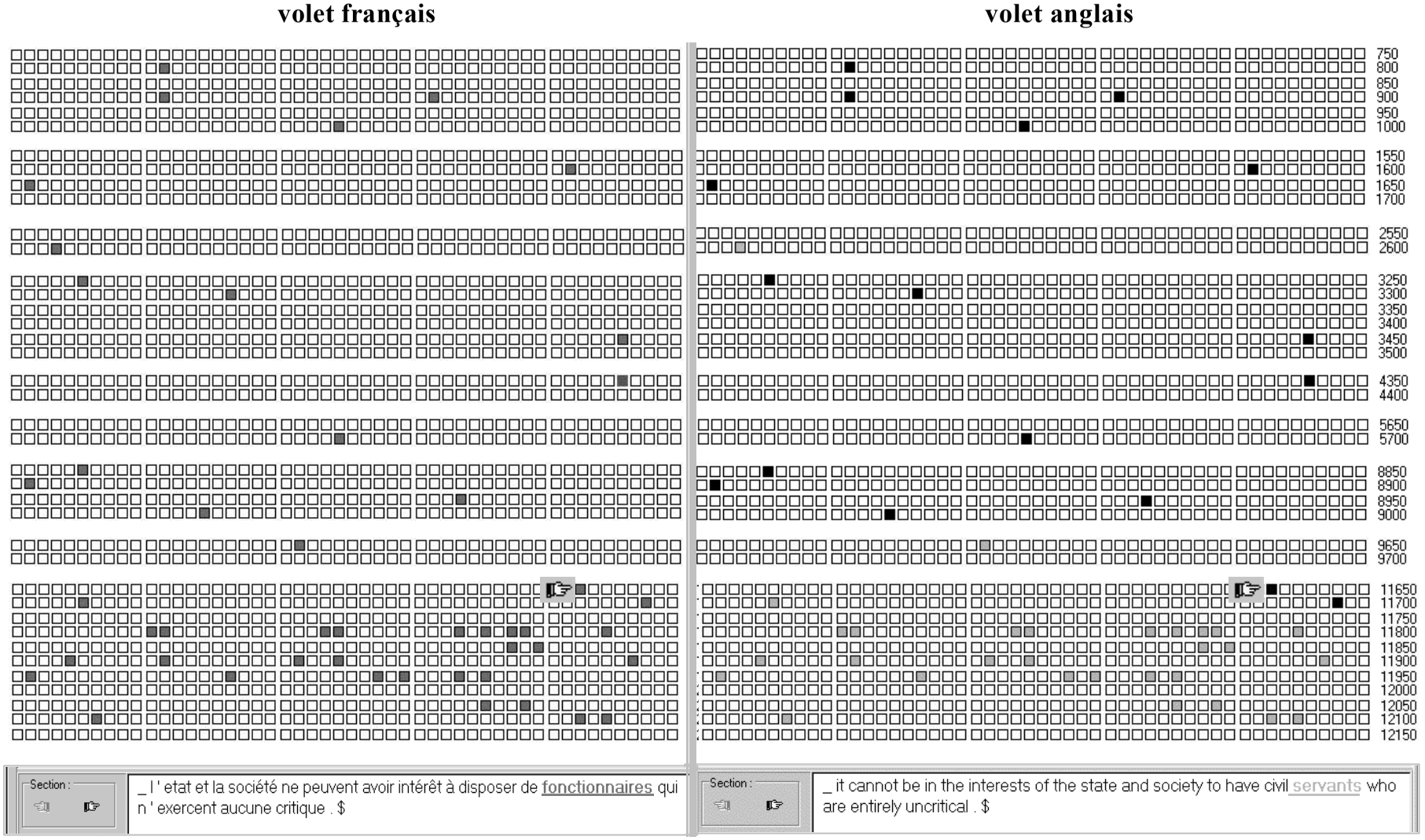
La mise en évidence des correspondances lexicales multiples

Dans cette section, nous allons décrire une méthode d'analyse permettant la découverte de correspondances lexicales entre les deux volets parallèles pour des mots qui possèdent plusieurs traductions au sein d'un même corpus [Zimina, 2004]. Une approche hybride qui allie la *topographie textuelle* et l'*analyse des spécificités* est à l'origine de cette méthode. Les fonctionnalités de la *carte des sections parallèles* auxquelles nous faisons référence prennent en compte l'état final d'un ensemble de procédures en cours de développement qui devraient être disponibles dans la prochaine version de *Lexico3*.

La cartographie des *présence-absence* de correspondances bilingues au sein des traductions fournit des moyens automatisés pour le recensement des équivalences, y compris pour des mots qui possèdent plusieurs traductions au sein d'un même corpus. Pour illustrer notre propos, nous considérerons les phrases du corpus ***Convention*** qui contiennent la forme *fonctionnaires* (F=49) et le sous-ensemble de phrases équivalentes en anglais.

Sur la figure 4.15, l'appariement des phrases du corpus est matérialisé par des carrés positionnés sur deux colonnes, la première correspond au texte en français et la deuxième au texte en anglais. Les carrés sont liés et toute sélection dans un volet est automatiquement répercutée dans l'autre. La ventilation de la forme française *fonctionnaires* est représentée par des carrés foncés. L'appariement des phrases étant représenté sur la carte, il est possible d'envisager l'activation de la sélection des phrases équivalentes dans le volet anglais [14].

[14] La réalisation informatique de la sélection parallèle des sections correspondantes est actuellement en cours de développement. Elle s'appuie sur les fonctionnalités existantes de la carte des sections de *Lexico3* et utilise des principes formels de la segmentation parallèle des corpus bilingues où chaque couple de phrases équivalentes est introduit par le même code. Nous avons simulé l'activation de cette fonctionnalité sur la carte de sections parallèles de *Lexico3* à l'aide de *Power Point.* Cette présentation est disponible sur le Cd-rom : /fichiersCD/stmz/page5_fichiers/JADT_2004F.ppt.



Lexico3

Figure 4.15 : Un extrait de la *carte des phrases* appariées issues du corpus ***Convention***

Le calcul des *spécificités* permet de sélectionner parallèlement pour chacun des fragments constitués une série d'unités textuelles caractéristiques. Les listes des spécificités s'affichent des deux côtés de la *carte des sections parallèles* (cf. *Figure 4.15*). Le tableau 4.16 présente les spécificités majeures calculées pour la sélection bi-textuelle. On constate sur ce tableau que les unités bilingues mises en évidence par cette méthode constituent bien des correspondances de traduction. Ainsi, la forme française *fonctionnaires* (*spec.*+E109), ayant servi de point d'entrée pour la construction de l'échantillon de phrases pour cette analyse, peut être appariée avec la forme *servants* (*spec.*+E55) et le segment répété *civil servants* (*spec.*+E52) qui se sont révélés les plus caractéristiques du fragment anglais (cf. *Tableau 4.15*).

Tableau 4.16 : Spécificités majeures

spec.	F	f	Forme / segment	Forme / segment	f	F	spec.
+E109	49	49	*fonctionnaires*	*servants*	31	50	+E55
+E31	14	14	les fonctionnaires	*civil servants*	29	46	+E52
+E31	14	14	des fonctionnaires	civil	41	304	+E40
+E22	36	14	de loyauté	loyalty	14	43	+E20
+E21	42	14	loyauté	duty	15	109	+E16
+E17	22	10	de loyauté politique	political loyalty	10	25	+E16
+E16	24	10	loyauté politique	of political	10	29	+E15
+E16	7	7	de fonctionnaires	duty of	11	45	+E15
+E14	15	8	obligation de loyauté politique	officers	10	38	+E14
+E13	21	8	obligation de loyauté	duty of political loyalty	9	23	+E14
				service	14	110	+E14
				civil service	11	58	+E13
				constitutional	14	146	+E13

Guide de lecture : Le tableau représente des spécificités majeures calculées pour l'échantillon des phrases où est attestée la forme *fonctionnaires* et le fragment correspondant en anglais.

La fréquence totale de la forme *fonctionnaires* (F=49) est supérieure à celle de la forme anglaise *servants* dans le fragment (f=31). Nous pouvons en conclure que la forme *fonctionnaires* reçoit d'autres traductions dans le corpus. Pour découvrir l'ensemble des équivalences lexicales appuyées sur la forme pôle, on soumet au calcul des spécificités les seules phrases du fragment anglais dans lesquelles la forme *servants* est absente. Sur la *Figure 4.15*, les sections correspondant à ces phrases sont de couleur gris claire. La réitération du calcul des spécificités dans ce sous-ensemble de phrases met en évidence une série d'unités les plus caractéristiques de ce nouveau fragment :

Forme / segment	**F**	**f**	**spec.**
officers	38	10	+E19
officials	16	7	+E16
senior	18	6	+E13
senior police	5	4	+E11
police	216	9	+E10
senior police officers	3	3	+E09

Le retour au contexte confirme que les unités *officers* (*spec.*+E19), *senior police officers* (*spec.*+E09), *officials* (*spec.*+E16), constituent bien des traductions de la forme *fonctionnaires* (au même titre que la forme *servants* et le segment *civil servants* découverts précédemment, cf. *Tableau 4.17*).

Fonctionnalités logicielles

L'exploration contextuelle s'appuie sur les outils de navigation textométrique de la *carte des sections*. Sur la figure 4.15, les boutons situés à gauche de la fenêtre de visualisation de la sélection (en forme de mains) permettent de passer, respectivement, à la section suivante/précédente ou l'occurrence suivante/précédente du type étudié. Pour explorer parallèlement les deux volets bilingues du corpus, les sections en correspondance sont liées. Toute sélection dans une fenêtre est automatiquement répercutée dans l'autre.

Analyse des résultats

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

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

(Corpus ***Convention***)

Dans ces dernières phrases, c'est l'expression *civil service* qui correspond à la forme *fonctionnaires*. Comme nous l'avons montré au chapitre 3, ce type de correspondance entre les unités lexicales relève de la notion d'*équivalence contextuelle*. Le segment *civil service* et son voisinage lexical immédiat sont liés au sein de la même locution : *the civil service (general principles) act*. Sur le plan sémantique, il s'agit d'unités traductionnelles singulières qui nécessitent un traitement particulier. Il appartient à l'expert humain de s'appuyer sur les blocs alignés pour examiner dans le détail les parallèles et les divergences entre ce type

de séquences : *la loi-cadre sur les fonctionnaires ~ the civil service (general principles) act*.

Tableau 4.17 : Retour au contexte

français	**anglais**
/…/ l'introduction de procédures disciplinaires à l'encontre de **fonctionnaires**, en raison de leur engagement politique /…/, violerait la convention de l'organisation internationale du travail (oit) /…/	/…/ the institution of disciplinary proceedings against civil **servants** on account of their political activities /…/ breached international labour organisation (ilo) convention /…/
il s'agissait en fait d'un document destiné aux agents du bvd (*binnenlandse veiligheidsdienst*) et d'autres **fonctionnaires** appelés à accomplir des missions pour lui.	it was in fact a document intended for bvd (*binnenlandse veiligheidsdienst*) staff and other **officials** who carried out work for the bvd.
il dénonce les propos tenus lors de la conférence de presse par le ministre de l'intérieur et les hauts **fonctionnaires** de police qui l'accompagnaient.	he complained of the remarks made by the minister of the interior and the senior police **officers** accompanying him at the press conference.

(Corpus ***Convention***)

4.3 Exemples d'extraction de ressources traductionnelles

La navigation textométrique dans le bi-texte a permis de mettre au point une série de pratiques d'exploration de corpus parallèles dans des langues différentes. Ces pratiques peuvent aider l'utilisateur des données textuelles multilingues (traducteur, lexicographe, terminologue etc.) dans l'extraction de ressources

traductionnelles à partir de corpus parallèles[15]. Dans ce qui suit, nous décrirons quelques explorations du corpus ***Convention*** réalisées à des fins d'extraction de ressources bi-textuelles.

Exemple n°1 – *cour*, *tribunal* / *court*

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

La technique de repérage des équivalences par *seuillage* permet de découvrir dans le volet français du corpus ***Convention*** la principale traduction du mot anglais *court*. Il s'agit du terme français *cour*. La localisation de l'équivalence *cour/court* dans les zones correspondantes du bi-texte montre qu'il existe d'autres traductions du mot *court* dans le corpus.

Sur la *figure 4.18*, les volets français et anglais du corpus ***Convention*** sont représentés sur une seule carte de sections[16]. Chaque carré représente un couple de phrases appariées. Le calcul des spécificités réalisé pour les sections contenant le mot anglais *court* a permis de signaler à l'utilisateur le mot français *cour* comme la principale traduction de ce terme dans le corpus.

Les fonctionnalités de *Lexico3* que nous avons décrites dans les sections précédentes, rendent possible une visualisation simultanée de la présence des formes *court/cour* dans les sections (cf. *Figure 4.18*). Les sections bicolores indiquent les sections où *court* est traduit par *cour*. La présence de sections monochromes sur la carte montre qu'il existe d'autres traductions du mot *court*

[15] Les expériences décrites dans cette section ont fait l'objet d'une série de cours « Corpus parallèles et textométrie » organisée par le Centre de Recherche en Ingénierie Multilingue (CRIM) pour les étudiants de DESS « Traductique et Gestion de l'Information » à l'Institut National des Langues et Civilisations Orientales (INaLCO) : http://www.cavi.univ-paris3.fr/ilpga/ed/student/stmz/ED268-PagePersoMZ_fichiers/stmz/page2.htm

[16] On peut envisager plusieurs représentations topographiques du bi-texte. La carte des sections parallèles peut être séparée en deux volets bilingues (cf. *Figure 4.15*). Il est possible également de représenter les deux volets du bi-texte sur une seule carte. Dans ce dernier cas, on peut représenter côte à côte les carrés correspondant à des sections appariées (paragraphes, phrases), par exemple : | (français|anglais).

dans le corpus. En cliquant sur un carré monochrome, il est possible de visualiser dans la fenêtre du bas le texte correspondant à la section où *court* n'est pas traduit par *cour*. L'itération du calcul des spécificités dans les sections monochromes permet de repérer le terme *tribunal* qui est la deuxième correspondance de *court* dans le corpus.

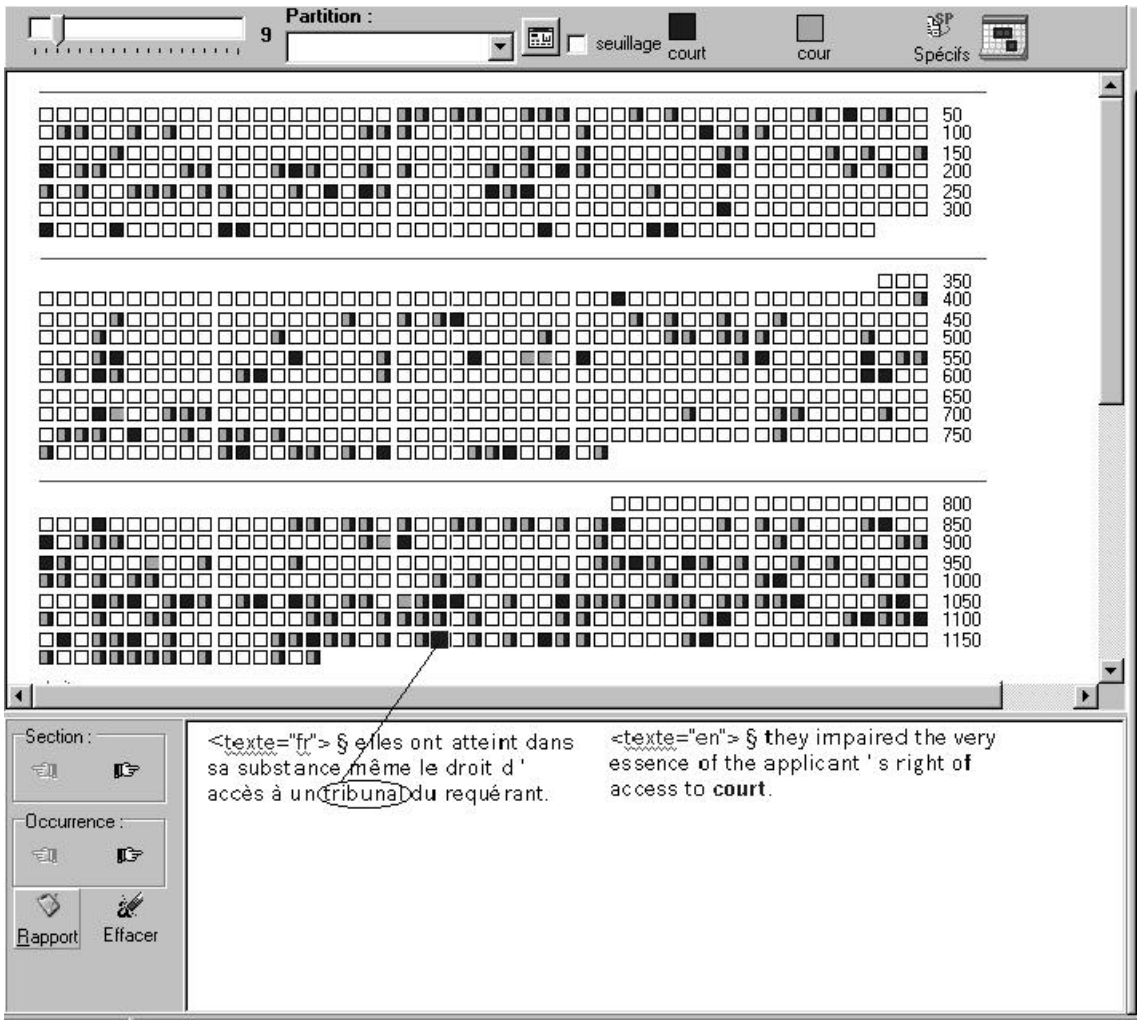


Figure 4.18 : Le repérage topographique des traductions du terme anglais *court*

Guide de lecture : Les sections bicolores de la carte correspondent aux sections du bi-texte où le terme anglais *court* est traduit par *cour* en français. La présence de sections monochromes sur la carte indique que le mot anglais reçoit d'autres traductions en français. Il s'agit notamment du mot français *tribunal.*

Exemple n°2 – *administr*+ □ *administ*+

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

Pour cette recherche, nous utiliserons la même méthode de navigation textométrique. Comme nous l'avons déjà mentionné, sous *Lexico3*, le langage des *expressions régulières* permet à l'utilisateur de constituer des groupes de mots correspondant au *type* de son choix et d'enregistrer la liste de ces unités pour une exploration ultérieure. Le repérage des sections de la carte en fonction de la présence/absence des *types* bilingues *administr*+ / *administ*+ (français/anglais) laisse apparaître des sections monochromes où sont attestées des équivalences traductionnelles originales telles que :

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

Notons que les résultats de cette exploration soulèvent également un certain nombre de problèmes liés à la détection de simplifications dans la traduction. Voici, par exemple, un couple de phrases en correspondance de traduction tirées du corpus :

français	**anglais**
toute autre lecture non seulement pécherait par manque de cohérence, mais surtout trahirait l'intention des autorités, lesquelles entendaient soustraire à l'emprise de la convention tout le système ***administr****atif*, y compris les dispositions de fond et de procédure du droit ***administr****atif* pénal.	any other construction would not only lack coherence.

Cet exemple montre que l'on peut envisager l'utilisation de la topographie bi-textuelle pour la vérification de l'*homogénéité traductionnelle*.

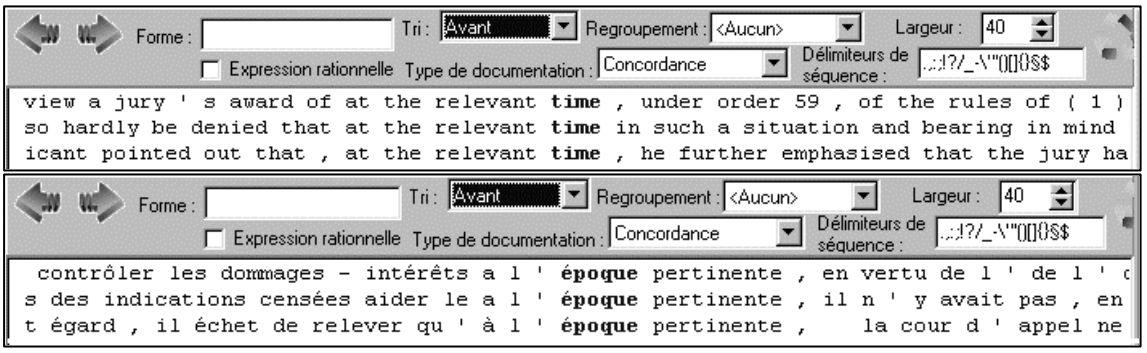
Exemple n°3 – *temps, time, etc.*

Objectif : rechercher les traductions des expressions anglaises avec le mot *time*. Recenser les variations dans les traductions dues au contexte.

L'exploration commence par le repérage dans le volet anglais du corpus des phrases dans lesquelles sont présentes les occurrences de la forme *time*. Le fragment sélectionné dans le volet anglais provoque la sélection par *résonance* des phrases correspondantes dans l'autre volet du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence des formes françaises *époque* et *temps* qui sont liées sur le plan de la traduction au terme anglais *time*. Pour mieux cerner le contexte autour des équivalences *time–temps* et *time–époque*, la navigation textométrique peut s'appuyer sur les résultats du calcul des segments répétés dans les deux volets bilingues. L'analyse topographique des ventilations des segments répétés recensés autour des formes françaises *temps* et celles des segments contenant la forme anglaise *time* laisse apparaître les équivalences suivantes :

français	**anglais**
à l'époque des faits à l'époque pertinente au fil du temps	at the material time at the relevant time in the course of time

La navigation textométrique à l'aide de la carte des sections peut être appuyée par l'analyse de concordances bilingues. Les concordances bilingues correspondant à chacune des équivalences établies autour des pôles traductionnels permet une meilleure analyse thématique des contextes :



4.4 Conclusion du chapitre 4

Au travers de plusieurs explorations du corpus ***Convention*** présentées dans ce chapitre, nous avons défini une série de méthodes qui permettent d'accéder à la description automatique de relations de correspondance entre des unités bilingues du corpus bi-textuel. Une approche hybride qui allie la *représentation topographique* du bi-texte et l'*analyse des spécificités* est à l'origine de ces méthodes. Cette approche repose entièrement sur les ressources construites à partir du corpus. Les fréquences et les distributions des unités textuelles servent de points de repère pour l'identification et l'extraction des correspondances.

Nous avons constaté que la localisation des différents types d'unités textuelles dans les phrases en correspondance apporte une plus grande précision dans l'analyse de leurs distributions. Ainsi, une description automatique de multiples relations d'équivalence entre des unités bilingues peut être obtenue par des comparaisons statistiques lorsque l'exploration du corpus tient compte de l'alignement des phrases.

L'utilisation de la topographie bi-textuelle dans le cadre de l'alignement automatique de corpus parallèles constitue une piste de recherche extrêmement prometteuse. Les applications ouvertes par cette approche concernent notamment 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.

Chapitre 5

Explorations multidimensionnelles des corpus de traduction

L'étude simultanée des distributions lexicales dans les deux volets d'un corpus parallèle peut être enrichie par des analyses et des visualisations appuyées sur les *méthodes de la statistique multidimensionnelle* [1]. Cette approche permet de produire des descriptions synthétiques des systèmes d'équivalences présents dans le bi-texte [2].

5.1 La synthèse de l'information bi-textuelle

5.1.1 L'approche factorielle du bi-texte

Les méthodes de statistique descriptive multidimensionnelle permettent de produire des descriptions synthétiques de *tableaux lexicaux* que nous avons décrits au chapitre 4 (cf. *Tableau 4.1*). Rappelons que l'individu statistique donnant lieu à des comptages pour chaque case de ce type de tableaux est

---

[1] Une étude systématique des *méthodes multidimensionnelles* appliquées à l'analyse exploratoire de données textuelles a été réalisée par Benzécri [1973], et développées dans les travaux de ses successeurs. Une description des modèles statistiques sous-jacents peut être trouvée, par exemple, dans [Lebart et Salem, 1994].

[2] Certaines expériences que nous décrivons dans ce chapitre constituent un prolongement de travaux déjà publiés dans [Zimina, 2000].

l'*occurrence* d'une unité textuelle : forme, segment répété etc. L'*analyse factorielle des correspondances* AFC[3] est une des méthodes d'analyse des données textuelles particulièrement adaptée à la mise en évidence des principales oppositions sous-jacentes au corpus. L'analyse factorielle permet de produire des représentations graphiques sur lesquelles les proximités géométriques entre *points-lignes* et *points-colonnes* traduisent des associations statistiques entre les lignes et entre les colonnes du tableau lexical.

Dans l'exemple qui suit nous décrivons à l'aide de l'AFC les tableaux de fréquences représentés sur le tableau 5.1. Ces tableaux croisent les formes issues du corpus ***Convention*** (en ligne), et (en colonne) neuf parties constituées par le texte officiel de la Convention de sauvegarde des Droits de l'Homme et ses huit protocoles intégraux[4]. On trouve au tableau ci-dessous les effectifs des parties ainsi constituées :

partie	code	volet français	volet anglais
texte de la Convention (1953)[5]	CONV	5 953 occ.	5 710 occ.
Protocole Additionnel (1954)	P-add	507 occ.	512 occ.
Protocole °2 (1970)	P°2	531 occ.	539 occ.
Protocole °4 (1968)	P°4	730 occ.	716 occ.
Protocole °6 (1985)	P°6	594 occ.	593 occ.
Protocole °7 (1988)	P°7	1 156 occ.	1 132 occ.
Protocole °9 (1994)	P°9	842 occ.	765 occ.
Protocole °10 (1992)	P°10	244 occ.	266 occ.
Protocole °11 (1998)	P°11	4 380 occ.	4 250 occ.

[3] L'analyse factorielle fait partie de méthodes statistiques d'analyse multidimensionnelle qui s'appliquent à des tableaux de nombres à double entrée. La méthode extrait des «facteurs » résumant le plus fidèlement possible en quelques nombres l'ensemble des informations contenues dans le tableau de départ. L'*analyse factorielle des correspondances* AFC est une des méthodes d'analyse factorielle qui s'applique à l'étude de tableaux à double entrée composés de nombres positifs. Elle est caractérisée par l'emploi d'une distance (ou métrique) particulière qui est une somme de carrés pondérés dite *distance du chi-2* (ou $\div^2$). Pour plus de détails sur l'utilisation de l'AFC dans les études textuelles, voir, par exemple, Salem [1987, pp. 309-310] ; Lebart et Salem [1996, p. 90, pp. 311-312].

[4] Cette analyse porte sur un fragment du corpus ***Convention*** constitué par le texte de la Convention des Droits de l'Homme proprement dit. Les arrêts trop nombreux pour une représentation visuelle simultanée dans le plan de l'analyse factorielle des correspondances ont été volontairement exclus. On trouvera à l'annexe A la description de la structure du corpus ***Convention***.

[5] La date correspond à l'entrée en vigueur du document.

La méthode fournit des représentations approchées des distances calculées entre ces neuf parties textuelles et l'ensemble des unités par lesquelles elles sont décrites [6]. L'analyse est réalisée sur chacun des tableaux lexicaux générés séparément pour les volets anglais et français du corpus (cf. *Tableau 5.1*). Cette démarche constitue une tentative de mettre au point des méthodes d'automatisation de la recherche de proximités dans les variations fréquentielles que les individus bilingues subissent au sein du corpus parallèle.

5.1.2 Les dimensions parallèles du bi-texte

Les graphiques présentés sur la *figure 5.2ab* donnent des représentations visuelles des associations entre lignes et colonnes des tableaux lexicaux issus des volets français et anglais du corpus ***Convention***. Sur chacun des graphiques, le plan factoriel engendré par les deux premiers *axes* de l'analyse factorielle des correspondances fournit une représentation schématique des informations contenues dans ces tableaux [7]. Pour gagner en lisibilité, nous avons volontairement affiché seulement une partie des libellés des *individus actifs* [8] (formes). Les individus dont les libellés sont affichés dans le plan factoriel ont été sélectionnés en raison de leurs positions excentrées, et donc caractéristiques. Pour les formes représentées, les distances à l'origine des axes (DISTO) sont supérieures ou égales à 1 (cf. *Tableaux 5.3ab*).

Nous constatons que les deux représentations se ressemblent nettement. Les proximités entre les *points-lignes* et les *points-colonnes* correspondants ainsi que leurs positions respectives par rapport à l'origine des axes sont similaires sur

---

[6] La comparaison effectuée entre deux *profils-lignes* va nous renseigner sur la façon dont les unités textuelles correspondantes s'associent dans parties du corpus, tandis que la comparaison de deux *profils-colonnes* nous renseigne sur les proximités existant entre les différentes parties vis-à-vis du vocabulaire employé.

[7] Les *individus actifs* correspondent à l'ensemble des individus du tableau lexical servant de base au calcul des axes factoriels.

[8] Les *axes factoriels* sont construits par les techniques d'analyse factorielle pour décrire brièvement les informations contenues dans le tableau lexical de départ. Représentés sur les graphiques-plans, ils fournissent des représentations bidimensionnelles approximatives des données textuelles.

les deux graphiques. Les valeurs des paramètres repris sur les tableaux 5.3ab contribuent à l'interprétation des résultats résumés sur les deux graphiques. Nous pouvons remarquer que les coordonnées des individus sur les axes factoriels sont proches lorsqu'il s'agit des unités en correspondance de traduction (cf. *figure 5.2ab*).

L'approche factorielle de l'analyse des corpus parallèles confirme une organisation textuelle semblable au sein des deux volets bilingues. Les graphiques obtenus avec ce type de méthodes produisent des images sommaires des correspondances traductionnelles entre les deux volets.

Les résultats de nos explorations montrent que les méthodes factorielles ouvrent des perspectives de recherche prometteuses pour la *synthèse globale* de l'information bi-textuelle. En revanche, leur utilité reste limitée lorsqu'il s'agit d'affiner la description de structures lexicales parallèles dans les corpus de traduction.

Pour accéder à des lectures plus nuancées de l'information bi-textuelle, il est nécessaire d'utiliser conjointement avec ces méthodes des méthodes de *classification* présentées dans la section suivante. Cette dernière famille de méthodes de la statistique descriptive multidimensionnelle permet de réaliser des explorations *locales* des correspondances lexicales. Elle convient particulièrement à des fins d'extraction de ressources traductionnelles des corpus parallèles.

Guide de lecture des tableaux 5.3ab :

- les *poids relatifs* désignent les marges des lignes du tableau lexical ; les marges sont obtenues en divisant chaque élément du tableau par la somme de la colonne correspondante ;
- les *distances à l'origine des axes* (DISTO) montrent les distances de chaque profil au profil moyen ;
- les *coordonnées* sont celles des points représentés sur les graphiques de la figure 5.2ab ; les contributions traduisent l'importance des éléments dans la construction de chaque axe [Lebart et Salem, 1994, pp. 82-92].

Les formes des tableaux sont triées par rapport à leurs distances à l'origine des axes(DISTO). Le profil des formes représentées sur les tableaux est supérieur ou égal à 1.

Tableau 5.1ab : Extraits des tableaux lexicaux : croisement (formes – 9 parties du texte intégral de la Convention des Droits de l'Homme)

a)

forme	CONV	P-add	P°2	P°4	P°6	P°7	P°9	P°10	P°11
de	367	31	35	47	42	65	49	18	281
la	332	16	29	32	22	45	45	8	218
l	168	17	20	18	17	35	32	8	160
à	161	16	11	22	16	30	26	7	113
le	95	13	11	21	20	30	24	8	80
des	126	13	17	12	6	12	9	3	95
les	110	17	8	16	10	17	13	4	94
ou	108	8	10	11	11	30	18	7	55
d	101	3	13	10	9	21	21	13	64
du	68	11	14	12	22	25	16	7	61
et	97	10	6	9	4	14	7	0	79
en	73	6	7	12	11	23	14	4	63
une	103	4	1	10	4	17	15	1	49
article	32	2	5	5	3	11	10	4	111
par	69	2	7	10	8	21	14	2	48
convention	48	7	9	10	8	10	15	4	46
au	52	7	6	8	8	13	4	1	29
cour	35	0	11	1	0	1	8	0	68
un	70	1	3	3	2	10	2	0	32
dans	61	7	1	8	5	9	2	0	21
est	42	1	6	5	4	9	7	2	38
qui	41	8	2	10	7	11	10	4	17
protocole	0	6	9	12	15	18	10	6	33
présent	9	5	8	9	12	17	8	5	28
conseil	32	4	5	4	10	8	7	4	15
/.../									

b)

forme	CONV	P-add	P°2	P°4	P°6	P°7	P°9	P°10	P°11
the	596	58	60	70	68	104	88	28	448
of	448	46	51	69	59	95	54	25	319
to	158	19	19	23	16	31	26	12	132
shall	147	13	14	17	16	23	21	5	118
in	120	12	16	17	17	29	11	3	88
a	131	5	4	16	5	29	16	2	71
or	118	8	12	12	12	30	16	7	58
be	101	7	7	13	12	20	13	5	84
and	98	9	5	10	4	15	6	1	76
article	32	2	5	5	3	11	10	4	109
by	70	4	4	9	7	21	9	4	45
convention	44	7	9	10	8	10	13	6	47
any	61	8	7	9	11	12	6	3	34
this	52	4	9	11	12	21	7	6	29
court	42	0	12	1	0	1	8	0	69
with	44	5	8	5	4	9	9	4	42
for	64	4	6	8	3	6	4	2	28
as	41	8	8	9	4	9	9	3	26
which	49	5	1	7	3	11	8	2	31
protocol	0	5	8	12	15	17	9	7	33
may	40	2	3	3	6	8	3	1	23
council	28	4	5	4	10	8	6	5	15
it	42	6	2	4	3	5	6	1	15
on	27	1	3	4	6	9	4	3	27
europe	28	4	4	4	9	8	6	4	14
/.../									

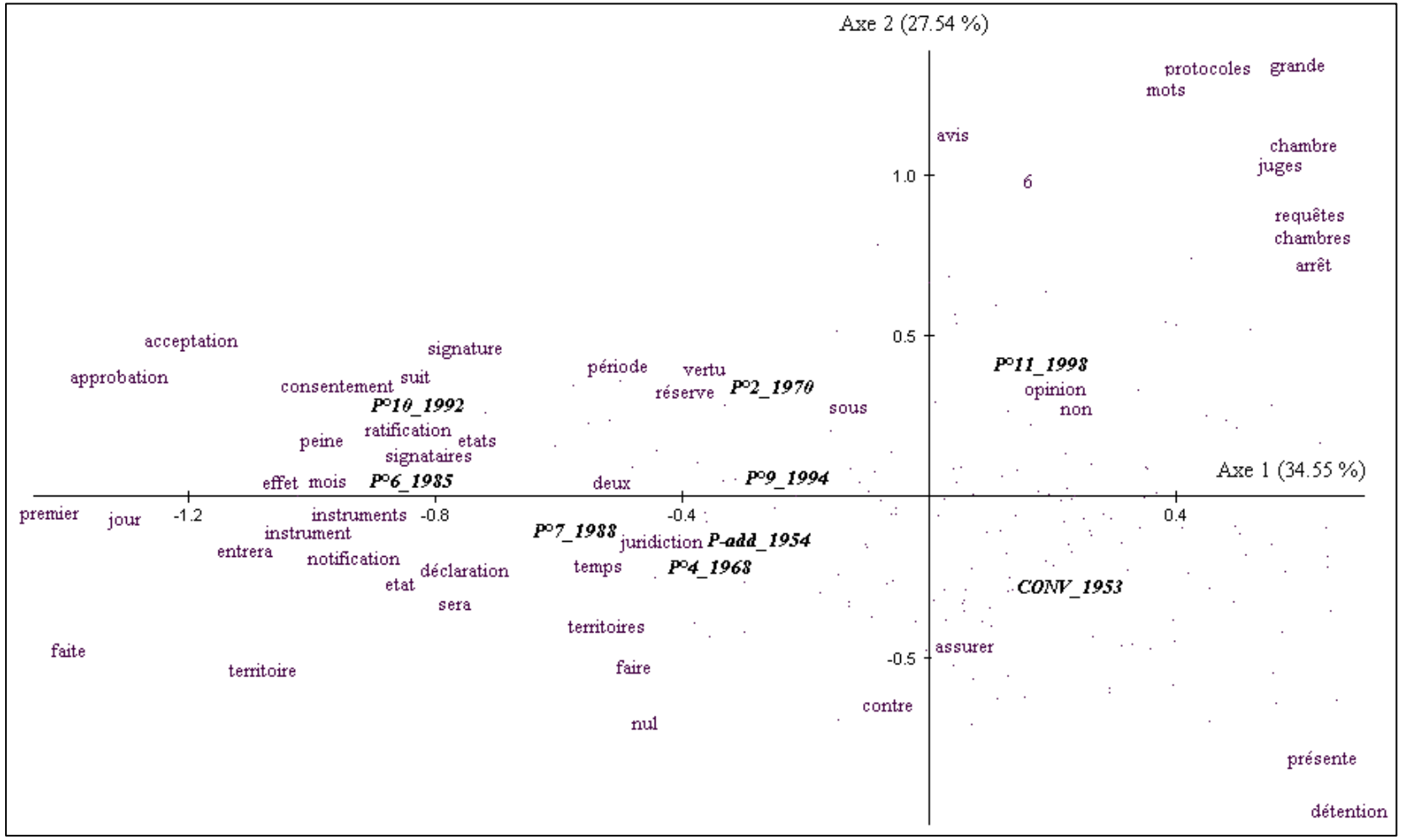


Figure 5.2a : Proximités entre formes et entre 9 parties de l'extrait du corpus ***Convention***. Analyse des correspondances du tableau 5.1(a)

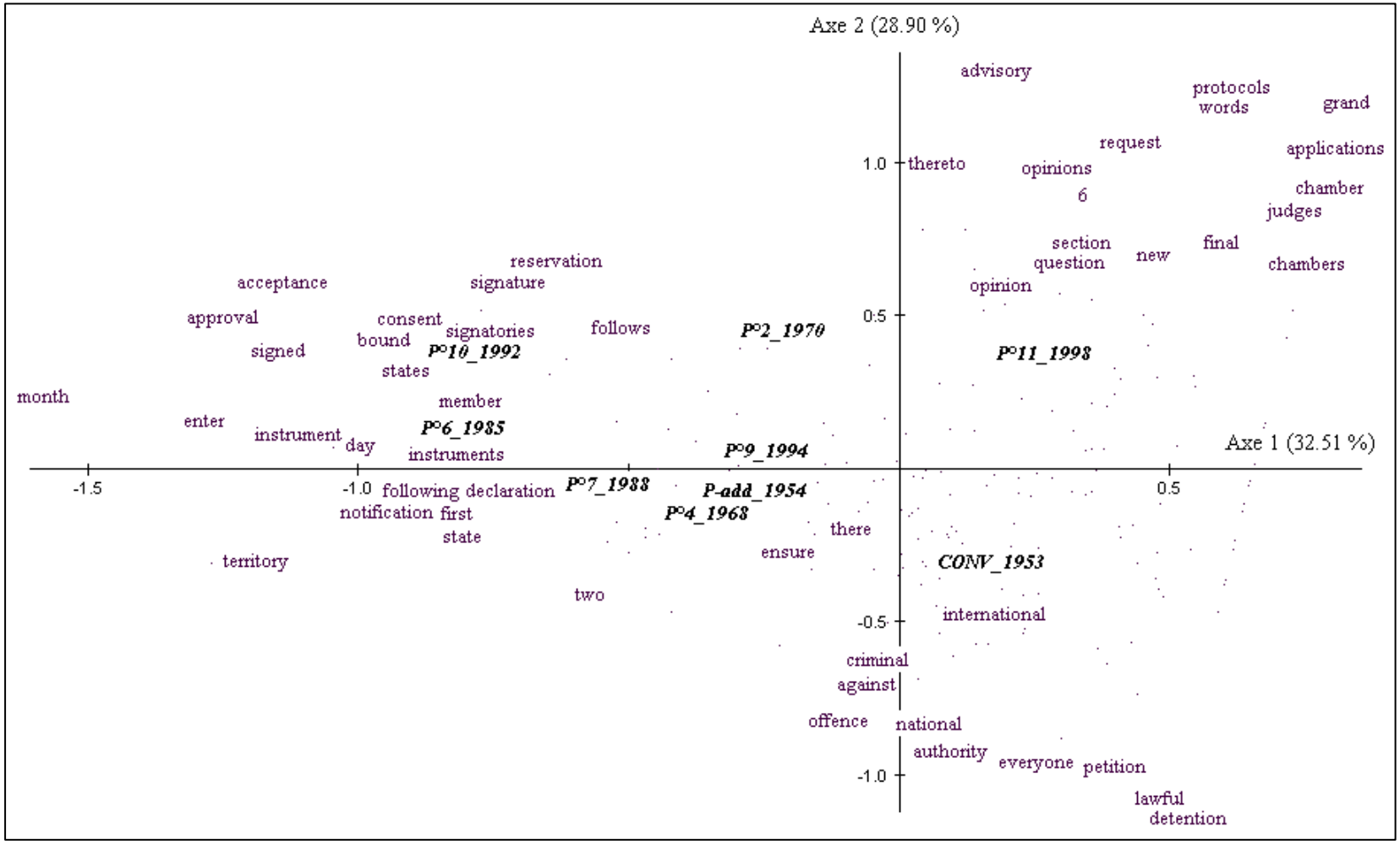


Figure 5.2b : Proximités entre formes et entre 9 parties de l'extrait du corpus ***Convention***. Analyse des correspondances du tableau 5.1(b)

Tableau 5.3a :

Coordonnées factorielles et distances à l'origine des axes (DISTO) pour des formes issues de l'analyse des correspondances du tableau 5.1(a)

identificateur	P.REL	DISTO	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
avis	0,18	4,78	0,03	1,13	0,76	1,44	-0,70	0,43	-0,41	-0,07
faite	0,11	4,39	-1,39	-0,48	-1,32	0,29	0,12	0,59	0,02	-0,15
peine	0,10	4,21	-0,99	0,17	-0,07	-0,69	-0,86	-1,14	-0,75	0,34
approbation	0,24	3,17	-1,32	0,37	0,62	-0,79	0,21	-0,04	0,49	0,02
premier	0,12	2,64	-1,43	-0,04	0,05	-0,68	-0,19	-0,13	-0,24	0,16
territoire	0,17	2,62	-1,08	-0,54	-0,91	0,11	-0,09	0,26	-0,15	0,47
jour	0,12	2,58	-1,30	-0,06	0,22	-0,80	-0,30	-0,19	-0,25	-0,04
acceptation	0,31	2,55	-1,19	0,49	0,88	0,17	-0,11	0,21	0,18	0,06
grande	0,18	2,44	0,60	1,34	-0,46	-0,24	-0,03	-0,09	0,04	0,01
consentement	0,12	2,31	-0,96	0,34	0,69	-0,70	0,50	-0,05	0,24	0,04
mots	0,12	2,24	0,38	1,31	-0,23	-0,28	-0,01	-0,11	0,47	0,07
protocoles	0,13	2,05	0,42	1,31	-0,09	0,28	-0,23	0,07	-0,10	-0,01
sous	0,10	1,99	-0,13	0,28	1,06	0,78	-0,24	0,26	0,19	0,04
reserve	0,15	1,84	-0,39	0,33	1,08	0,54	-0,17	0,06	0,25	0,12
chambre	0,32	1,77	0,61	1,09	-0,37	-0,22	-0,05	-0,09	0,05	0,02
instruments	0,10	1,75	-0,96	-0,04	-0,14	0,65	0,42	-0,29	0,37	-0,05
nul	0,10	1,75	-0,46	-0,71	-0,72	0,54	0,26	-0,34	0,09	0,18
faire	0,11	1,74	-0,48	-0,53	-0,62	-0,18	-0,23	0,87	0,05	-0,03
detention	0,09	1,61	0,68	-0,98	0,30	-0,09	-0,27	-0,09	0,11	0,07
mois	0,23	1,58	-0,98	0,03	0,18	-0,70	-0,07	0,23	-0,01	-0,20
notification	0,11	1,58	-0,93	-0,20	0,22	-0,65	-0,40	-0,2	-0,01	0,00
temps	0,09	1,55	-0,54	-0,22	-0,84	0,47	0,26	-0,46	0,06	-0,10
juges	0,27	1,49	0,57	1,04	-0,27	-0,09	-0,11	-0,05	0,01	0,01
6	0,11	1,43	0,16	0,98	-0,37	-0,39	0,08	0,30	-0,09	-0,20

Tableau 5.3a : (suite)

Coordonnées factorielles et distances à l'origine des axes (DISTO) pour des formes issues de l'analyse des correspondances du tableau 5.1(a)

IDENTIFICATEUR	P,REL	DISTO	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
période	0,11	1,42	-0,51	0,41	0,02	-0,58	0,19	0,66	0,25	-0,35
etat	0,38	1,4	-0,83	-0,25	-0,62	-0,23	-0,24	0,32	-0,21	0,11
déclaration	0,40	1,39	-0,79	-0,24	-0,81	0,17	0,10	0,00	-0,06	-0,06
assurer	0,11	1,37	0,00	-0,47	-0,52	0,47	0,27	-0,39	0,25	-0,61
entrera	0,15	1,33	-1,1	-0,11	-0,06	0,12	0,14	-0,25	-0,04	-0,07
requêtes	0,13	1,31	0,62	0,88	-0,30	-0,21	-0,08	-0,09	0,05	0,02
effet	0,10	1,25	-1,02	0,02	0,29	-0,09	-0,23	0,01	0,06	0,27
signature	0,31	1,23	-0,75	0,47	0,51	0,29	0,16	-0,01	0,27	0,01
non	0,09	1,23	0,24	0,28	0,43	-0,37	0,80	0,08	-0,37	-0,01
présente	0,31	1,21	0,64	-0,81	0,30	-0,12	-0,16	-0,07	0,06	0,06
suit	0,28	1,20	-0,84	0,37	0,26	-0,39	0,30	-0,08	-0,19	-0,04
sera	0,12	1,19	-0,77	-0,33	-0,35	0,08	0,42	-0,31	-0,28	-0,11
vertu	0,22	1,18	-0,35	0,37	-0,88	0,05	0,06	0,30	0,02	0,23
signataires	0,15	1,17	-0,81	0,08	0,25	0,45	0,41	-0,24	-0,07	0,01
chambres	0,11	1,17	0,62	0,81	-0,28	-0,20	-0,08	-0,09	0,06	0,03
opinion	0,09	1,13	0,21	0,40	0,22	0,83	-0,24	-0,08	-0,08	-0,34
instrument	0,14	1,11	-1,01	-0,10	0,18	-0,11	-0,05	-0,16	-0,10	-0,05
contre	0,09	1,09	-0,05	-0,64	0,33	-0,41	0,27	0,47	-0,23	-0,25
juridiction	0,11	1,08	-0,42	-0,13	-0,53	-0,19	0,10	0,75	-0,06	0,05
etats	0,33	1,05	-0,80	0,18	0,51	-0,13	0,00	-0,02	-0,28	-0,13
arrêt	0,13	1,04	0,62	0,72	-0,25	-0,20	-0,09	-0,09	0,06	0,03
deux	0,17	1,02	-0,51	0,03	0,16	-0,28	-0,45	0,38	0,52	-0,19
ratification	0,48	1,01	-0,82	0,23	0,21	0,42	0,16	-0,04	0,22	0,01
territoires	0,14	1,00	-0,52	-0,40	-0,69	0,06	-0,06	0,28	0,00	0,09

Tableau 5.3b :

Coordonnées factorielles et distances à l'origine des axes (DISTO) pour des formes issues de l'analyse des correspondances du tableau 5.1(b)

IDENTIFICATEUR	P.REL	DISTO	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
advisory	0,16	4,55	0,20	1,31	0,82	1,28	0,53	-0,41	-0,18	0,02
reservation	0,09	4,01	-0,63	0,67	1,41	1,00	0,39	-0,02	0,05	-0,12
month	0,11	3,72	-1,58	0,24	-0,13	-0,82	0,60	0,15	-0,29	0,06
territory	0,22	3,31	-1,27	-0,29	-1,05	0,20	0,09	-0,52	0,01	-0,46
approval	0,23	3,04	-1,25	0,50	0,40	-0,91	-0,03	0,27	0,41	0,06
opinions	0,13	2,78	0,29	0,98	0,65	1,00	0,45	-0,29	-0,13	0,01
grand	0,19	2,42	0,83	1,20	-0,47	-0,26	-0,02	0,09	-0,01	0,00
bound	0,12	2,30	-0,91	0,42	0,56	-0,93	-0,26	0,10	0,18	-0,01
consent	0,12	2,30	-0,91	0,42	0,56	-0,93	-0,26	0,10	0,18	-0,01
ensure	0,09	2,26	-0,20	-0,27	-0,61	0,56	-0,78	0,18	-0,21	0,88
acceptance	0,32	2,22	-1,14	0,61	0,66	0,03	0,16	-0,13	0,28	-0,05
opinion	0,16	2,20	0,19	0,60	0,67	1,08	0,31	-0,18	-0,17	0,17
words	0,12	2,20	0,60	1,20	-0,30	-0,30	-0,09	0,27	0,37	0,03
signatories	0,09	2,11	-0,83	0,46	0,61	-0,02	-0,69	0,35	0,49	-0,01
request	0,10	2,09	0,43	1,08	0,36	0,68	0,33	-0,19	-0,10	0,00
protocols	0,12	2,02	0,62	1,24	-0,03	0,26	0,17	-0,08	-0,07	0,01
applications	0,14	2,01	0,81	1,06	-0,43	-0,24	0,00	0,10	0,00	-0,01
enter	0,14	1,95	-1,29	0,16	-0,23	0,15	-0,39	-0,01	-0,18	-0,08
day	0,14	1,81	-1,04	0,09	-0,01	-0,63	0,43	0,27	-0,24	-0,05
two	0,10	1,81	-0,57	-0,41	-0,44	-0,33	0,86	-0,34	0,05	0,39
signed	0,15	1,72	-1,15	0,38	0,10	0,20	-0,25	-0,14	0,32	-0,16
chamber	0,35	1,67	0,79	0,92	-0,38	-0,22	0,01	0,11	0,01	-0,01
detention	0,11	1,64	0,51	-1,07	0,29	0,05	0,24	0,25	0,12	-0,09
lawful	0,10	1,64	0,51	-1,07	0,29	0,05	0,24	0,25	0,12	-0,09
notification	0,11	1,54	-0,94	-0,09	0,01	-0,50	0,48	0,39	-0,05	-0,05
offence	0,09	1,42	-0,11	-0,82	-0,15	-0,17	0,65	-0,31	0,21	0,34
6	0,11	1,42	0,34	0,90	-0,39	-0,46	0,09	-0,30	-0,03	0,19

Tableau 5.3b : (suite)

Coordonnées factorielles et distances à l'origine des axes (DISTO) pour des formes issues de l'analyse des correspondances du tableau 5.1(b)

IDENTIFICATEUR	P.REL	DISTO	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8
following	0,17	1,40	-0,90	-0,04	0,07	-0,64	0,36	0,08	-0,22	0,01
states	0,30	1,35	-0,90	0,34	0,53	-0,27	0,25	0,05	-0,08	-0,02
judges	0,28	1,35	0,73	0,85	-0,26	-0,09	0,06	0,08	0,00	-0,01
follows	0,15	1,34	-0,47	0,45	0,50	-0,41	-0,62	-0,27	-0,19	0,03
petition	0,20	1,33	0,40	-0,97	0,44	-0,13	0,06	0,09	0,00	-0,09
instrument	0,13	1,32	-1,11	0,11	0,09	-0,14	0,04	0,12	-0,20	0,02
instruments	0,10	1,32	-0,82	0,03	-0,04	0,46	-0,63	0,11	0,11	0,14
state	0,41	1,32	-0,80	-0,17	-0,68	-0,07	0,28	-0,30	-0,08	-0,03
there	0,20	1,32	-0,09	-0,20	0,51	-0,74	-0,26	-0,57	-0,26	0,11
declaration	0,36	1,31	-0,77	-0,08	-0,77	0,17	-0,17	-0,17	-0,17	0,04
signature	0,29	1,27	-0,72	0,63	0,49	0,11	-0,23	0,02	0,21	0,07
authority	0,09	1,19	0,10	-0,90	0,00	-0,10	0,51	-0,12	0,18	0,20
final	0,11	1,16	0,60	0,74	-0,41	-0,24	0,15	-0,02	0,04	0,09
chambers	0,11	1,16	0,75	0,67	-0,30	-0,18	0,04	0,13	0,02	-0,02
against	0,12	1,14	-0,06	-0,69	0,46	-0,51	-0,01	-0,40	-0,12	0,11
question	0,09	1,13	0,31	0,69	0,55	0,35	0,11	-0,28	-0,19	-0,02
everyone	0,16	1,11	0,22	-0,95	-0,02	0,16	0,09	-0,03	0,21	-0,28
new	0,13	1,11	0,47	0,71	-0,47	-0,27	0,23	-0,12	0,07	0,16
thereto	0,18	1,09	0,03	1,00	0,11	0,20	0,08	-0,07	0,08	-0,11
first	0,19	1,08	-0,84	-0,11	0,00	-0,51	0,25	0,03	-0,16	-0,11
member	0,34	1,07	-0,79	0,23	0,41	-0,35	0,23	0,16	-0,11	-0,07
national	0,15	1,03	0,03	-0,83	-0,05	0,12	-0,16	-0,25	0,17	-0,44
criminal	0,09	1,02	-0,03	-0,64	-0,18	-0,18	0,59	-0,27	0,19	0,31
section	0,09	1,01	0,34	0,72	-0,29	0,44	-0,29	-0,03	-0,07	0,07
international	0,09	1,00	0,17	-0,48	-0,28	0,43	-0,54	0,26	-0,09	0,34

5.2 Descriptions locales pour l'extraction de lexiques bilingues

Les *méthodes de classification* constituent la deuxième grande famille de techniques d'analyse de données. Ces méthodes permettent d'obtenir à partir d'un ensemble d'individus décrits par des variables (qui correspondent la plupart du temps aux différentes parties d'un corpus) des regroupements en classes (ou en familles de classes hiérarchisées)[9]. Ces méthodes sont utilisées notamment lorsque le nombre d'éléments à décrire est trop important pour une représentation simultanée sur des plans factoriels.

5.2.1 Classification automatique des unités lexicales

Il existe une grande variété de méthodes de classification automatique. La *classification ascendante hiérarchique* (CAH) permet de regrouper les individus d'un tableau lexical en classes en minimisant, autant que possible la perte d'information[10]. Dans le domaine de l'alignement, cette approche offre des moyens formels pour repérer les proximités et les différences dans les profils de ventilation des unités textuelles au sein du corpus parallèle.

Sur la *figure 5.4*, nous avons présenté un extrait du tableau lexical formé par la superposition des deux tableaux lexicaux, français et anglais, qui ont été générés séparément pour chaque volet du corpus ***Convention*** (cf. *Tableau 5.1*). Les données de ces deux tableaux ont été fusionnées en un seul tableau afin de pouvoir analyser simultanément l'ensemble des unités textuelles du corpus

---

[9] On peut soumettre à la classification soit l'ensemble des colonnes du tableau lexical soit l'ensemble des lignes de ce même tableau. On trouve un panorama des applications de ces méthodes à l'analyse de corpus textuels dans [Lebart et Salem, 1994, pp. 111-133].

[10] On part d'un ensemble de *n* éléments, affectés chacun d'un poids proportionnel à leur importance dans l'ensemble, et entre lesquels on a calculé des distances. L'agrégation des deux éléments les plus proches permet de constituer un nouvel élément dont on peut recalculer à la fois le poids et la distance par rapport à chacun des éléments qui restent à classer (la *méthode des voisins réciproques* proposée par J. Juan [1982] constitue l'une des multiples façons d'effectuer ce calcul). A l'issue de cette étape, le problème est ramené à celui de la classification de *n-1 éléments*. On agrège à nouveaux les deux éléments les plus proches. Le processus est réitéré jusqu'à épuisement de l'ensemble des éléments [Habert *et al.*, 1997, pp. 198-199].

parallèle. Le tableau sur la *figure 5.4* correspond à la réunion des formes dont la fréquence est égale, ou supérieure, à 10. Les formes issues de chaque volet bilingue sont triées par *ordre lexicométrique*. Les individus sont identifiés par des numéros permettant de départager les deux langues : les formes en français sont numérotées de F1 à F958 et celles en anglais de A1 à A835 [11]. Les colonnes de ce même tableau représentent les neuf parties constituées par le texte officiel de la Convention de sauvegarde des Droits de l'Homme et ses huit protocoles intégraux.

Appliquée aux lignes du tableau lexical, la CAH décrit leurs proximités en les regroupant en *classes* [12]. On détermine a priori le nombre des classes dans lesquelles on désire répartir les éléments du tableau. En fixant ce nombre à la moitié du nombre total d'individus (cf. *Figure 5.4*), on tente de se rapprocher d'une classification constituée par de petites classes de deux éléments dont les profils sont très similaires (voir identiques) et dont chacun appartient à l'un des volets du corpus. Sur la *figure 5.5*, la représentation de la classification sous forme d'arbre hiérarchique ou *dendrogramme* aide à interpréter les résultats de la hiérarchie des classes obtenues. L'histogramme sur la *figure 5.6* permet de retrouver l'ordre des regroupements effectués.

[11] Cette numérotation des individus bilingues du tableau lexical a servi par la suite à départager les homographes (*article / article*).

[12] Chacun des regroupements effectués en suivant la méthode de la CAH s'appelle un *noeud*. L'ensemble des éléments terminaux rassemblés dans un noeud est une *classe*. La classification produit une hiérarchie de classes partiellement emboîtées les unes dans les autres.

		CONV	P-add	P°2	P°4	P°6	P°7	P°9	P°10	P°11
/.../										
F198	avoir	6	.	1	.	1	1	.	.	1
F199	contre	6	.	.	.	.	2	2	.	.
F200	désignés	5	1	.	1	.	.	.	.	3
F201	détention	10	.	.	.	.	.	.	.	.
F202	faits	6	.	.	.	.	1	.	.	3
F203	fin	6	1	.	1	.	.	.	.	2
F204	lui	4	.	1	.	.	1	1	.	3
F205	nécessaires	5	1	.	1	.	1	.	.	2
F206	non	3	.	.	.	.	.	3	.	4
F207	obligatoire	6	.	.	1	.	1	1	.	1
F208	opinion	3	1	2	.	.	.	.	.	4
F209	prendre	4	1	.	1	.	2	.	.	2
F210	restrictions	6	.	.	2	.	.	.	.	2
F211	temps	2	2	.	2	1	1	.	.	2
F212	violation	6	.	.	.	.	.	1	.	3
/.../										
A201	already	4	1	.	1	.	1	.	.	3
A202	authority	8	.	.	.	.	2	.	.	.
A203	deal	5	.	1	.	.	.	.	.	4
A204	ensure	3	3	.	1	.	1	.	.	2
A205	european	4	.	1	.	.	.	.	.	5
A206	he	4	.	.	2	.	2	.	.	2
A207	hold	7	1	.	.	.	.	.	.	2
A208	individuals	3	.	.	.	.	.	2	.	5
A209	legal	5	1	1	.	.	.	.	.	3
A210	offence	7	.	.	.	.	3	.	.	.
A211	part	4	.	2	.	.	.	.	.	4
A212	question	2	.	2	.	.	.	1	.	5
A213	receive	5	.	.	.	.	.	.	.	5
A214	reservation	2	.	4	.	1	.	1	1	1
A215	transmitted	4	.	.	.	.	.	2	.	4
/.../										

identifiant

Figure 5.4 :
Extrait du tableau lexical formé par la superposition des deux tableaux lexicaux, français et anglais (cf. *Tableau 5.1*)

Guide de lecture : Sur la figure 5.4, l'extrait du tableau lexical bilingue correspond à la réunion des formes dont la fréquence est égale, ou supérieure, à 10. Les individus issus des volets français et anglais du corpus sont identifiés par des numéros permettant de départager les deux langues.

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
/.../						
3105	1662	2667	2	45.00	0.00003	*territoire / territory
3106	3033	1631	3	80.00	0.00003	*
3107	2991	2998	5	45.00	0.00003	*
3108	2123	454	18	42.00	0.00003	*
3109	2973	2361	10	48.00	0.00003	*
3110	1349	799	2	52.00	0.00003	*
3111	2980	2929	8	53.00	0.00003	*
/.../						
3541	3523	3484	17	656.00	0.00175	**
3542	2903	3501	237	1018.00	0.00178	**
3543	3500	3494	37	1454.00	0.00181	**
3544	3533	3524	54	559.00	0.00186	**
3545	3507	3478	47	180.00	0.00193	**
3546	3519	3528	41	9372.00	0.00196	**
3547	3532	3525	51	324.00	0.00197	**
3548	3516	3456	37	212.00	0.00202	**
/.../						
3559	3497	3424	22	284.00	0.00273	***
3560	3543	3537	65	3945.00	0.00313	***
3561	3535	3548	60	558.00	0.00321	****
3562	3506	3531	56	181.00	0.00354	****
3563	3545	2899	60	208.00	0.00367	****
3564	3556	3534	64	1289.00	0.00387	****
3565	3546	3560	106	13317.00	0.00440	*****
3566	3540	3553	75	2972.00	0.00446	*****
3567	3549	3536	124	1533.00	0.00504	*****
3568	3547	3555	90	468.00	0.00521	*****
3569	3568	3512	110	515.00	0.00677	*******
3570	3564	3559	86	1573.00	0.00743	********
3571	3542	3550	310	1865.00	0.00797	********
3572	3570	3505	101	1631.00	0.00818	********
3573	3539	3551	115	1456.00	0.00863	*********
3574	3571	3558	361	2123.00	0.00984	**********
3575	3566	3565	181	16289.00	0.01045	**********
3576	3552	3554	429	2716.00	0.01076	**********
3577	3561	3563	120	766.00	0.01083	**********
3578	3487	3567	259	2297.00	0.01279	**********
3579	3573	3576	544	4172.00	0.01920	**********
3580	3572	3562	157	1812.00	0.01977	**********
3581	3574	3579	905	6295.00	0.02421	**********
3582	3557	3578	321	2607.00	0.02728	**********
3583	3577	3569	230	1281.00	0.02878	**********
3584	3581	3575	1086	22584.00	0.03710	**********
/.../						

Figure 5.5 : Histogramme des indices de niveau

Guide de lecture : La première ligne du tableau indique qu'un élément artificiel (*nœud*) est obtenu par fusion des éléments avec les identifiants 1662 et 2667 (*aîné* et *benjamin*) qui sont la forme française *territoire* et la forme anglaise *territory.* Cet élément portera le numéro 3105. Il sera caractérisé par le profil moyen de ces deux composants. La valeur de l'indice (0.00003) permet de situer la classe constituée par rapport à l'ensemble des regroupements effectués. Elle décrit la plus petite distance entre les deux *profils-lignes* correspondants. La masse du nouvel élément (45.00) indique la somme des effectifs. On classe ainsi tous les éléments jusqu'à ce qu'il n'y en ait plus qu'un seul.

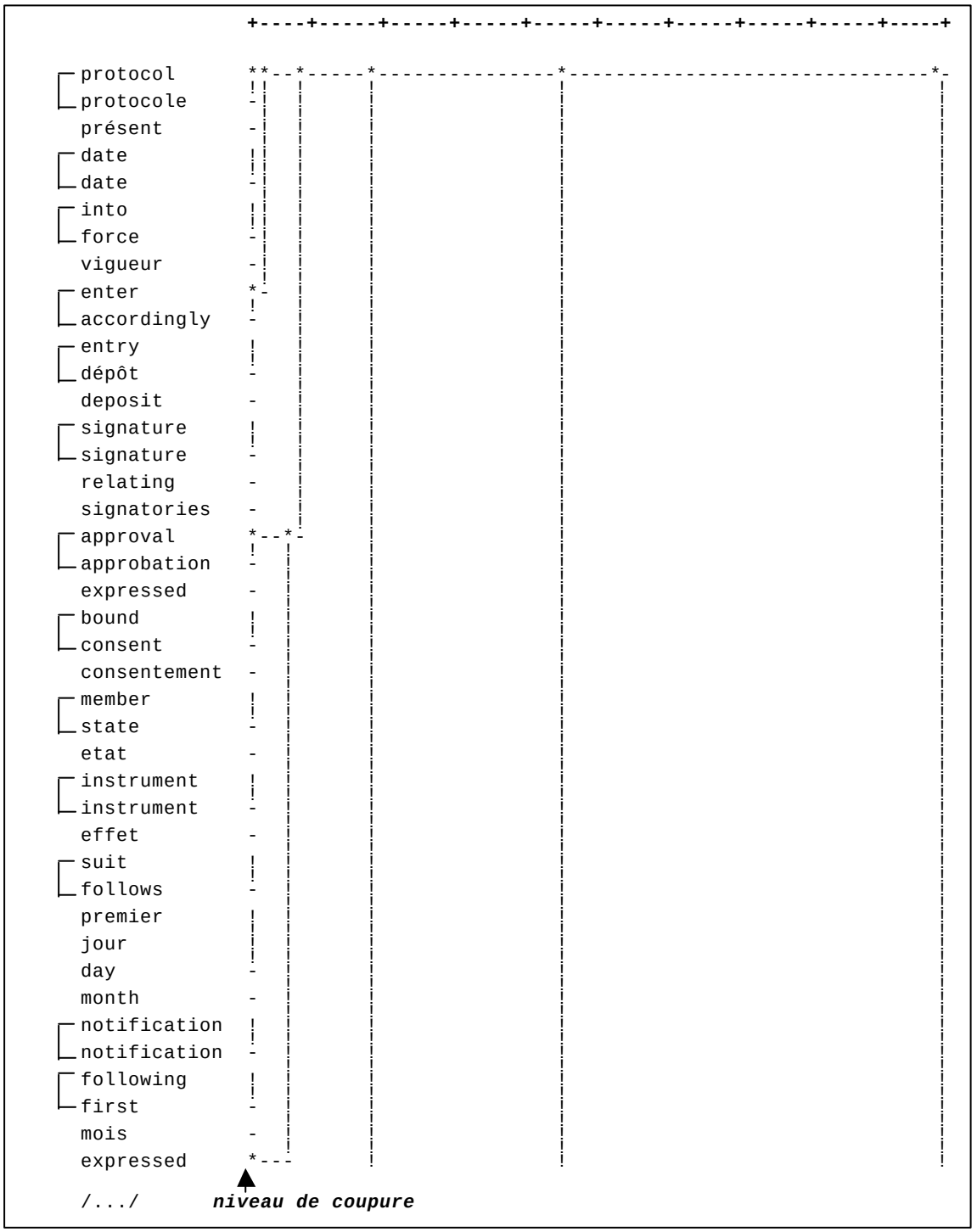


Figure 5.6 : Dendrogramme décrivant les proximités entre lignes du tableau lexical représenté sur la figure 5.4 (extrait)

Guide de lecture : La longueur des branches de l'arbre est proportionnelle aux valeurs de l'indice (cf. *Figure 5.5*). Si le nombre des éléments à classer est important, la représentation complète de l'arbre devient difficile à étudier. Une solution pratique consiste à définir un niveau de coupure qui correspond à un intervalle entre deux itérations du processus de classification ascendante. La coupure permet de considérer une classification résumée aux classes inférieures de la hiérarchie, qui ne regroupent que des éléments extrêmement proches.

5.2.2 Premiers résultats de la classification sur les formes du corpus **Convention**

Toutes les classes issues de la CAH réalisée à partir de l'ensemble des unités du tableau lexical (cf. *Figure 5.4*) ne présentent pas le même intérêt pour l'alignement. Les classes supérieures de la hiérarchie sont souvent constituées de nombreux individus et résument les principales oppositions observées dans les premiers plans factoriels. Elles ne fournissent pas d'indices directs sur les correspondances lexicales entre les deux volets. En revanche, les associations réalisées aux tous premiers niveaux de la classification nous informent sur des propriétés d'alignement. Les classes du niveau inférieur de la hiérarchie sont constituées par des agrégations d'individus correspondant à un indice très bas (cf. *Figures 5.5-6*). Elles regroupent des unités textuelles dont les profils sont très similaires dans les deux volets bilingues du corpus (cf. *Figure 5.8*).

A l'issue de cette première expérimentation, nous constatons que la grande majorité des classes composées de deux individus appartenant à des volets différents du corpus regroupent des formes correspondant à des équivalences de traduction dans le corpus ***Convention*** (cf. *Tableau 5.7*). Les confusions résultent essentiellement du grain du découpage en parties qui se révèle, dans certains cas, trop grossier pour différencier les profils de ventilation d'individus. Comme le montre la *figure 5.8*, la forme française *mandats* et la forme anglaise *take* ne se retrouveront plus dans la même classe si l'on parvient à préciser leurs distributions dans la première partie du corpus [13].

La procédure de la classification ascendante hiérarchique appliquée aux individus du tableau lexical composés de formes graphiques a permis d'obtenir des regroupements d'unités bilingues pertinents du point de vue de l'appariement. Ces premiers résultats ont encouragé des expérimentations sur la mise en correspondance des unités lexicales plus élaborées du point de vue sémantique.

[13] Cette approche se révèle problématique lorsqu'il s'agit de formes en relation de cooccurrence à l'intérieur d'une phrase. La prise en compte de segments répétés et l'étude des attractions lexicales qui s'opèrent sur l'axe syntagmatique dans les deux volets du corpus permettent d'augmenter la précision des résultats.

La section suivante prolonge la description de techniques d'alignement lexical à base de procédures de la classification automatique de segments répétés.

Tableau 5.7 : Exemples de classes composées de deux formes bilingues

NUM.	INDICE	AINE	BENJ	EFF.	DESCRIPTION DES CLASSES	
2880	0.000	1400	124	2	suit	follows
2746	0.000	1489	164	2	période	month
2701	0.000	172	1446	2	without	sans
2689	0.000	1333	13	2	au	any
2688	0.000	86	1404	2	human	homme
2676	0.000	69	1398	2	date	date
2667	0.000	1317	1	2	de	the
2666	0.000	145	1401	2	with	dont
2660	0.000	12	1332	2	convention	convention
2650	0.000	1354	63	2	tout	other
2646	0.000	68	1520	2	an	lui
2633	0.000	96	1437	2	territory	territoire
2623	0.000	1428	115	2	mesure	public
2619	0.000	1372	31	2	pas	not
2606	0.000	1360	37	2	droits	rights
2604	0.000	20	1339	2	protocol	protocole
2591	0.000	7	1324	2	or	ou
2586	0.000	1473	132	2	jour	day
2578	0.000	1391	66	2	acceptation	acceptance
2570	0.000	1407	51	2	été	been
2567	0.000	9	1327	2	and	et
2566	0.000	45	1370	2	provisions	dispositions
2558	0.000	49	1373	2	states	etats
2549	0.000	248	1523	2	compulsory	obligatoire
2546	0.000	1515	147	2	contre	against
2538	0.000	38	1365	2	general	général
2532	0.000	74	1387	2	articles	articles
2526	0.000	1500	178	2	instruments	instruments
2519	0.000	1425	118	2	avis	advisory
2517	0.000	1371	53	2	déclaration	declaration
2509	0.000	1456	142	2	instrument	instrument
2505	0.000	1467	167	2	même	same
2498	0.000	1526	238	2	restrictions	restrictions
2494	0.000	1393	75	2	signature	signature
2493	0.000	1402	79	2	juge	judge
2491	0.000	137	1443	2	judges	juges
2488	0.000	1502	417	2	introduite	alleged
2485	0.000	1386	67	2	après	after
2475	0.000	1491	326	2	protection	forth
2473	0.000	54	1384	2	chamber	chambre
2470	0.000	1469	155	2	protocoles	protocols
2467	0.000	1364	39	2	parties	parties
2464	0.000	10	1330	2	article	article
2458	0.000	134	1483	2	election	durée
2454	0.000	1466	169	2	mandat	take
1863	0.000	41	1362	2	ratification	ratification
1841	0.000	193	1503	2	majorité	majority
1833	0.000	173	1493	2	25	25
1828	0.000	192	1504	2	list	mandats
1818	0.000	1383	64	2	paragraphe	paragraph

	CONV	P-add	P°2	P°4	P°6	P°7	P°9	P°1.	P°11
dispositions	11	4	2	3	6	4	2	1	14
provisions	11	5	2	4	7	5	2	1	14
obligatoire	6	.	.	1	.	1	1	.	1
compulsory	5	.	.	1	.	1	1	.	.
suit	1	1	1	1	5	5	7	2	9
follows	1	1	1	1	1	1	5	1	5
jour	2	.	.	.	4	4	2	1	1
day	3	.	.	.	4	4	2	1	2
mesure	12	1	.	3	.	1	.	.	3
public	11	1	.	4	.	2	.	.	3
mandats	6	.	.	.	.	.	.	.	5
take	7	1	.	1	.	1	.	.	3
protection	9	.	.	2	.	.	.	.	2
forth	4	.	.	1	.	.	.	.	1

	CONV-1	CONV-2	CONV-3	CONV-4
mesure	2	3	2	4
public	5	7	.	.
mandats	.	3	.	3
take	2	1	3	1
protection	3	6	.	.
forth	.	3	1	.

Figure 5.8 :
Influence de la variation du découpage en parties sur les résultats de la classification

Guide de lecture : Les individus bilingues agrégés dans les mêmes classes ont des sous-fréquences similaires. La variation du découpage influe de manière importante sur la qualité des résultats. La *figure 5.8* montre que l'on peut préciser les profils des individus à classer si le découpage en parties est affiné. Dans l'exemple ci-dessus, la première partie correspondant au texte de base de la Convention de sauvegarde des droits de l'homme (**Conv**) a été fragmentée en quatre sous-parties (**Conv-1**, **Conv-2**, **Conv-3**, **Conv-4**) par des regroupements successifs de ces articles. L'étude des ventilations des formes issues des classes de non-correspondances au sein de ces quatre nouvelles sous-parties montre que ce découpage permet alors de différencier les profils précédemment confondus.

Tableau 5.9 : Classes de segments répétés

NUM.	AINE	BENJ	EFF.	INDICE	DESCRIPTION DES CLASSES	
9518	4595	1904	2	0.00002	protocole entrera en vigueur dès	protocol shall enter into force
9470	903	3864	2	0.00001	the date of entry into force of this protocol	la date d'entrée en vigueur du présent protocole
9322	4485	1345	2	0.00001	au secrétaire général du conseil de l'europe	to the secretary general of the council of europe
9341	3887	975	2	0.00001	la haute partie contractante	the high contracting party
9299	2465	5039	2	0.00001	months after the date	mois après la date
6990	5242	870	2	0.00000	assemblée consultative	the consultative assembly
6999	4770	1766	2	0.00000	partie contractante	any high contracting party
7309	2605	5342	2	0.00000	addressed to the secretary general of the council of europe	adressée au secrétaire général du conseil de l'europe
9113	1016	4158	2	0.00000	the members of the commission shall be elected	les membres de la commission sont élus
6251	4438	1731	2	0.00000	par le comité des ministres	by the committee of ministers
6742	2400	5364	2	0.00000	national security	sécurité nationale
8098	353	4067	2	0.00000	two-thirds	des deux tiers
8180	1101	4927	2	0.00000	the states concerned	etats intéressés
8214	3810	2242	2	0.00000	de trois mois	period of three months
6362	4517	2108	2	0.00000	un ou plusieurs	one or more
6499	1108	4025	2	0.00000	the terms of office	le mandat
6868	1286	3808	2	0.00000	of three years	de trois ans
7102	5046	1563	2	0.00000	trois mois	a period of three
9120	2096	4704	2	0.00000	members of the court	membres de la cour
7643	1998	5119	2	0.00000	have signed	ont signé
7329	4724	1995	2	0.00000	sont convenus de ce qui suit	have agreed as follows
7611	2266	5142	2	0.00000	their consent to be bound by	leur consentement à être liés par
8157	4954	2190	2	0.00000	après la date	after the date
9197	2094	4703	2	0.00000	members of the commission	membres de la commission
9208	2167	5255	2	0.00000	no one shall be	nul ne peut être
7602	5211	2441	2	0.00000	consentement à être liés par	consent to be bound by
7623	5041	2289	2	0.00000	trois ans	three years

NUM.	AINE	BENJ	EFF.	INDICE	DESCRIPTION DES CLASSES	
7957	4031	1944	2	0.00000	le premier jour du mois qui suit l’	on the first day of the month following
7314	2380	5140	2	0.00000	entry into force of this protocol	entrée en vigueur du présent protocole
7681	5131	973	2	0.00000	la haute partie contractante	the high contracting partie
7794	5233	2470	2	0.00000	notification adressée au secrétaire général	a notification addressed to the secretary general
8272	4876	1113	2	0.00000	etat dont il est le ressortissant	the territory of the state of which he is a national
7613	4981	2187	2	0.00000	acceptation ou	acceptance or approval
5698	994	3828	2	0.00000	the jurisdiction of the court	la compétence de la cour
6859	5376	2375	2	0.00000	élu au titre	elected in respect of
7041	1267	5202	2	0.00000	of the same state	même etat
7179	1569	4364	2	0.00000	a petition submitted under article	une requête introduite en vertu de l’article
7530	4449	2471	2	0.00000	notification adressée au secrétaire général	notification by the secretary general
8293	913	3863	2	0.00000	the date on which all parties to the convention have expressed	la date à laquelle toutes les parties à la convention
8747	2461	5289	2	0.00000	detention of a person	détention régulière
8989	5175	2221	2	0.00000	réserve de ratification ou d’	signature with reservation in respect of ratification or acceptance
9115	971	4989	2	0.00000	the high contracting parties	hautes parties contractantes
5598	2139	3985	2	0.00000	from the date on which	à partir de la date à laquelle
5690	5262	982	2	0.00000	l’interprétation ou	the interpretation or application of
7364	3890	869	2	0.00000	la juridiction obligatoire de la cour	the compulsory jurisdiction of the court
8241	4952	2191	2	0.00000	après la date à laquelle	after the date on which
8606	5044	2287	2	0.00000	trois mois après la date	three months after the date
8973	3907	916	2	0.00000	la peine de mort	the death penalty
9199	4780	1211	2	0.00000	aux dispositions de la convention	of the provisions of the convention
8220	5229	1539	2	0.00000	faite conformément au	a declaration made in accordance with this article shall be deemed

5.2.3 Extension aux segments répétés

L'application de la classification ascendante hiérarchique au tableau lexical composé de segments répétés du corpus a produit une série de classes d'individus appartenant à des volets différents qui regroupent de fait des segments qui se correspondent au plan traductionnel. Comme dans le cas des formes graphiques, les classes les plus cohérentes du point de vue de l'appariement automatique se trouvent au niveau inférieur de la hiérarchie et rassemblent des agrégations d'individus correspondant à un indice très bas (cf. *Tableau 5.9*).

Particularités de l'agrégation de segments en classes

A l'issue de l'analyse des résultats de la *CAH*, nous avons constaté que les classes des segments répétés possèdent un certain nombre de particularités dont il faut tenir compte lors de l'alignement (cf. *Tableau 5.9* ; *Annexe B.3*) :

⇒ Lorsque des séquences récurrentes emboîtées dans le corpus sont mises en relation dans les deux volets bilingues, l'agrégation en classes produit des correspondances univoques au plan traductionnel :

Exemple :

classe 8157
après la date
after the date

classe 8241
après la date à laquelle
after the date on which

Retour au contexte :

l'expiration d'une période de trois mois **après la date *à laquelle*** dix etats membres du
l'expiration d'une période de deux mois **après la date *à laquelle*** sept etats membres du
l'expiration d'une période de trois mois **après la date *à laquelle*** toutes les parties à
suit l'expiration d'une période d'un an **après la date *à laquelle*** toutes les parties à
l'expiration d'une période de trois mois **après la date** de la signature ou du dépôt de
devant la grande chambre ; ou trois mois **après la date** de l'arrêt, si le renvoi de
l'expiration d'une période de deux mois **après la date *de réception*** de la déclaration par
l'expiration d'une période de deux mois **après la date *de réception*** de la notification
l'expiration d'une période de deux mois **après la date** du dépôt de l'instrument de

expiration of a period of three months **after the date *on which*** ten member states of the
expiration of a period of two month, **after the date *on which*** seven member states of
expiration of a period of three months **after the date *on which*** all parties to the
the expiration of a period of one year **after the date *on which*** all parties to the
expiration of a period of three months **after the date** of signature or of the deposit
to the grand chamber ; or three months **after the date** of the judgment, if reference
the expiration of a period of two months **after the date *of receipt*** by the secretary general
the expiration of a period of two months **after the date *of receipt*** of such notification
the expiration of a period of two months **after the date** of the deposit of the instrument

⇒ Au sein des classes, la longueur de segments répétés en correspondance de traduction peut être différente [14] :

Exemple :

```
classe 7613
acceptation ou
acceptance or approval
```

Retour au contexte:

il sera soumis à ratification, **acceptation ou** approbation. un etat membre du conseil
en vertu de sa ratification, de son **acceptation ou** de son approbation par ledit etat, et
il sera soumis à ratification, **acceptation ou** approbation. un etat membre du conseil

shall be subject to ratification, **acceptance or approval**. a member state of the council
applies by virtue of ratification, **acceptance or approval** by that state, and each
it is subject to ratification, **acceptance or approval**. a member state of the council

Exemple :

```
classe 7041
même etat
of the same state
```

Retour au contexte:

comprendre plus d'un ressortissant du **même etat**. la commission siège en séance plénière
comprendre plus d'un ressortissant d'un **même etat**. les membres de la cour sont élus
puni pénalement par les juridictions du **même etat** en raison d'une infraction pour laquelle

the commission may be nationals **of the same state**. the commission shall sit in plenary
no two judges may be nationals **of the same state**. the members of the court shall be
proceedings under the jurisdiction **of the same state** for an offence for which he has already

Exemple :

```
classe 8645
de ce droit
exercise of this right
```

Retour au contexte:

aucune mesure ***l'exercice*** efficace **de ce droit** . conditions de recevabilité la cour ne
aucune mesure ***l'exercice*** efficace **de ce droit** . ces déclarations peuvent être faites pour
autorité publique dans ***l'exercice*** **de ce droit** que pour autant que cette ingérence est

any way the effective **exercise of this right** . admissibility criteria the court may only
any way the effective **exercise of this right** . such declarations may be made for a specific
authority with the **exercise of this right** except such as is in accordance with the

Dans les exemples ci-dessus, la qualité de l'appariement peut être améliorée si l'on tient compte des résultats de l'extraction des *quasi-segments répétés* [Bécue et Peiro, 1993] dans chaque volet bilingue du corpus ***Convention***. Pour la

[14] Notons que la *variation syntagmatique* qui explique la différence dans la longueur des segments équivalents issus des deux volets du corpus ***Convention*** est plus souvent attestée dans le volet français.

classe 8645, par exemple, cette approche permettrait de poser l'équivalence du quasi-segment répété français *l'exercice (efficace) de ce droit* et du segment répété anglais *exercice of this right*

Les techniques de filtrage

Toutes les classes composées de segments répétés ne présentent pas le même intérêt pour l'alignement [15]. Plusieurs approches peuvent être envisagées pour tenter de sélectionner, dans les produits de la classification, un nombre restreint de classes jugées pertinentes pour l'appariement lexical :

⇒ La sélection de classes pour l'appariement peut être effectuée sur la base de comparaison des *fréquences totales* (*F*) des individus agrégés ensemble :

segments agrégés en classes	**CONV**	**P-add**	**P°2**	**P°4**	**P°6**	**P°7**	**P°9**	**P°10**	**P°11**	**F**	***classe retenue***	***classe rejetée***
de cette	5	0	0	0	0	1	0	0	1	**7**		
before the	8	0	0	0	0	2	1	0	4	**15**		✓
la compétence de	3	0	1	0	0	0	0	0	5	**9**		
in article	4	0	1	0	0	1	0	0	9	**15**		✓
nul ne peut être	5	1	0	3	1	1	0	0	0	**11**		
no one shall be	6	1	0	3	1	1	0	0	0	**12**	✓	
declaration made	0	1	0	3	1	4	0	0	0	**9**		
declaration faite	0	1	0	2	1	3	0	0	0	**7**	✓	

Dans certains cas, les fréquences totales des unités textuelles agrégées dans les mêmes classes ne sont pas suffisantes pour réaliser le filtrage décrit plus haut. Le filtrage sur les fréquences totales montre ses limites lorsqu'il s'agit de segments en relation de cooccurrence dans les phrases équivalentes du corpus partitionné :

[15] La phase du filtrage des sorties de la classification des segments répétés est, généralement, plus lourde que dans le cas des formes graphiques. En revanche, les classes de segments permettent de cerner des équivalences plus élaborées au plan sémantique dont on peut se servir pour approfondir l'analyse automatique de la traduction. Les résultats de la classification ascendante hiérarchique réalisée à partir des segments répétés du corpus ***Convention*** sont regroupés en annexe B.

Exemple :

classe 4066
aucune derogation (F=3)
shall be made under (F=3)

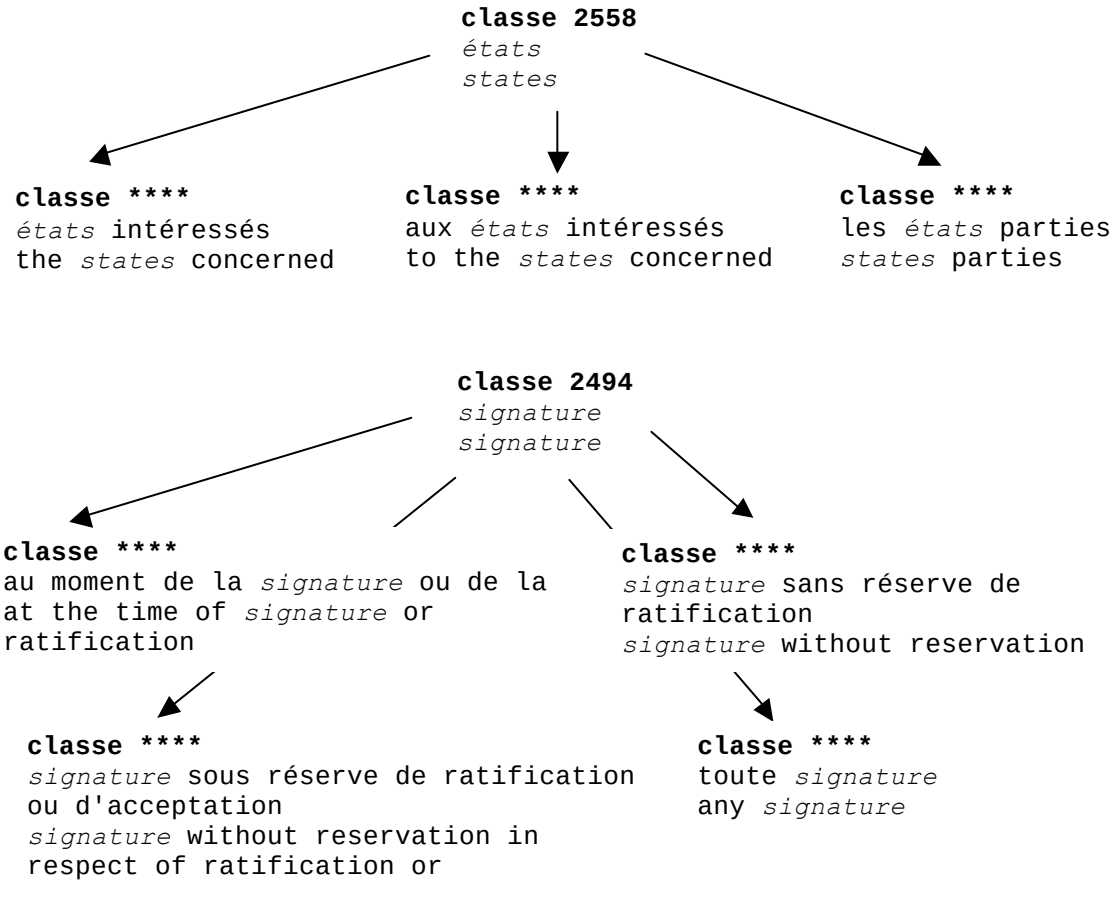
Retour au contexte:

aucune dérogation *n'est autorisée au* présent article au titre de l'article 15 de la convention.

no derogation from this article **shall be made under** article 15 of the convention.

L'exploration des *réseaux de cooccurrences* qui s'opèrent parallèlement dans les deux volets du corpus bilingue offre des moyens automatisés pour cerner des rapports syntagmatiques entre ce type de correspondances lexicales. Nous présenterons cette approche au chapitre 6.

⇒ On parvient à identifier des classes de segments répétés qui se correspondent au plan traductionnel lorsque l'on tient compte des résultats de la classification des formes graphiques (cf. *Tableaux 5.7*, *5.9* ; *Annexe B*) :



L'analyse des résultats de la classification ascendante hiérarchique des formes et des segments répétés prouve, sans nul doute, l'intérêt de cette méthode pour l'alignement. L'avantage de la méthode réside dans son degré de flexibilité : elle n'impose aucune restriction dans le processus d'identification des correspondances de traduction. Ces dernières sont recherchées librement dans l'ensemble de deux textes. La recherche des points d'ancrage pour l'alignement se base sur les proximités de profils de formes et de segments répétés dans les deux corpus. Cette recherche peut être entièrement automatisée.

Deux pistes de développement se révèlent intéressantes. Premièrement, il est possible d'utiliser les classes d'individus agrégés pour aligner les phrases correspondantes. Deuxièmement, on peut envisager l'intégration de cette méthode aux systèmes d'alignement des phrases fondés sur d'autres critères pour augmenter la résolution et procéder à l'extraction de correspondances lexicales *à base de corpus*.

De façon générale, l'affinement de la partition du corpus augmente considérablement la qualité globale des sorties de la classification. Avec la puissance grandissante de systèmes informatiques, il serait envisageable d'analyser les tableaux lexicaux croisant en lignes les formes et les segments et en colonnes l'ensemble de phrases issues de chaque volet d'un corpus parallèle. Les résultats d'explorations de ce type effectuées à l'échelle réduite confirment cette hypothèse.

Sur la *figure 5.9*, nous avons présenté les profils de ventilation du segment répété français *membres de la cour* et celui du segment anglais *the members of the court* dans les premières 212 phrases du corpus ***Convention***. Globalement, la prise en compte de l'appariement des phrases apporte une plus grande précision dans l'agrégation des individus en classes.

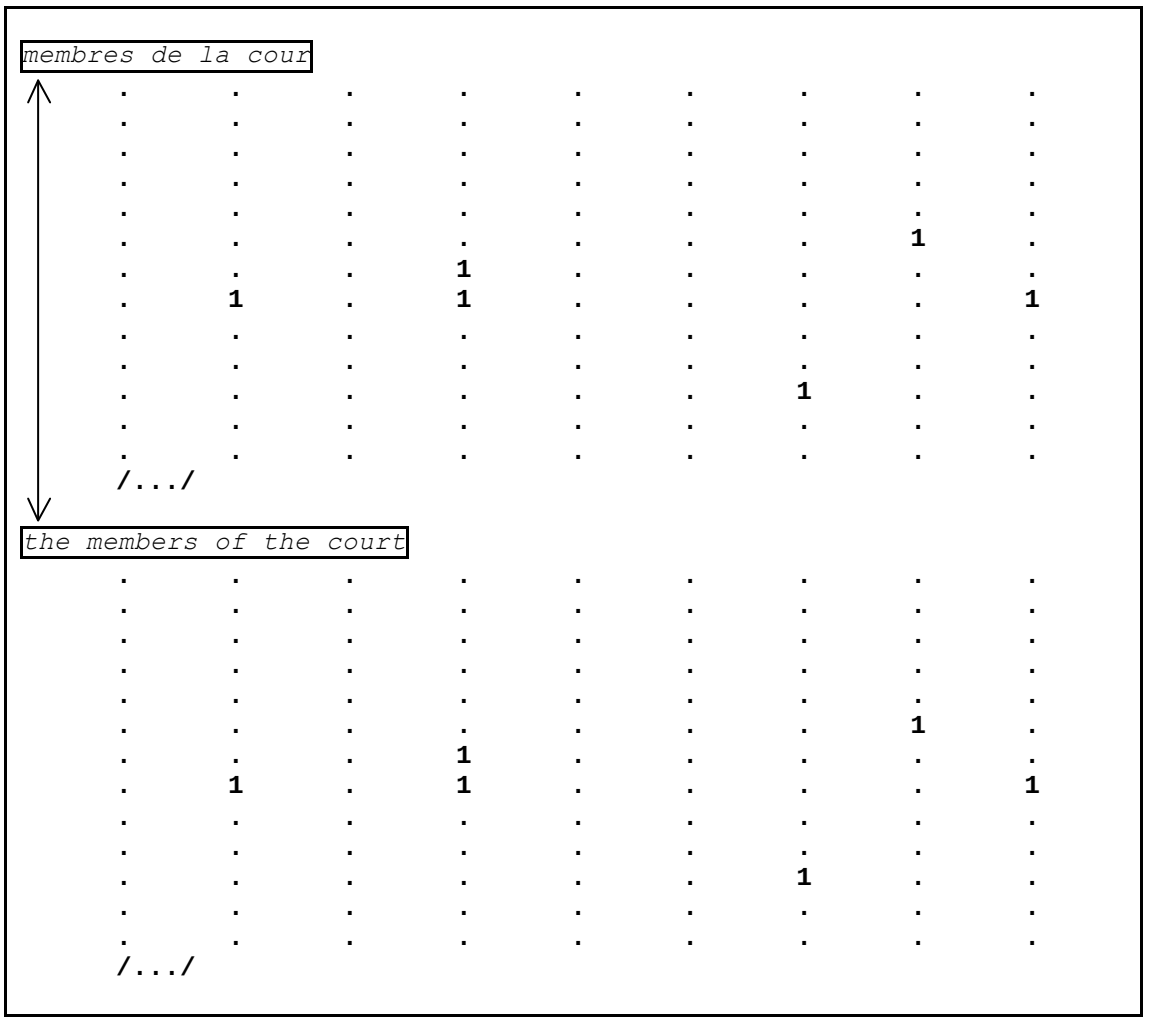


Figure 5.10 :
Les ventilations des segments répétés équivalents dans les 212 premières phrases du corpus ***Convention*** (extrait)

5.3 La vérification de l'alignement phrastique

Lorsque la correspondance de fragments de deux textes dont l'un constitue une traduction de l'autre est posée, il se révèle nécessaire de procéder à la vérification de l'alignement. Les unités alignées peuvent correspondre à des sections, paragraphes ou phrases d'un corpus parallèle. Pour tester la qualité des

appariements établis, il est envisageable de faire appel à la classification automatique.

Des balises peuvent être insérées dans chacun des volets d'un corpus parallèle pour permettre de découper chacun des textes en un ensemble de parties que l'on souhaite prendre en compte pour la vérification de l'appariement (chapitres, sections, paragraphes, phrases, etc.). Grâce à l'insertion de telles balises, il devient possible d'étudier les ventilations des unités textuelles au sein des parties.

Prenons l'exemple du corpus ***Convention*** aligné au niveau de la phrase. Nous allons prélever dans le corpus les premières 212 phrases (correspondant au texte de base de la Convention de sauvegarde des droits de l'homme). Pour procéder à la vérification de l'alignement de ces phrases, nous allons construire deux tableaux lexicaux (français et anglais) croisant en lignes les unités textuelles (formes et segments répétés) et, en colonnes, les 212 parties correspondant aux phrases appariées.

La classification ascendante hiérarchique appliquée aux colonnes de ces tableaux lexicaux permet de produire des regroupements des phrases similaires en termes de vocabulaire employé. L'agrégation en classes est faite séparément pour les phrases issues de chaque volet du corpus parallèle. La figure 5.11 permet de confronter les extraits des deux dendrogrammes obtenus pour les volets français et anglais. Bien que l'ordre des regroupements effectués diffère dans chaque volet du corpus, nous constatons que les phrases équivalentes (indiquées par des numéros identiques de *ph1* à *ph212*) subissent le même classement en français et en anglais (cf. *Figure 5.11*). Par exemple, les phrases qui portent les numéros 61 et 206 sont agrégées dans les mêmes classes en anglais comme en français. L'analyse simultanée de la structure des regroupements des phrases effectués dans chaque volet du corpus ***Convention*** confirme les résultats de l'alignement pour les 212 premières phrases.

Au cours de nos expérimentations ultérieures, l'utilisation de la classification automatique a permis de relever des erreurs dans la mise en correspondance des phrases issues des arrêts juridiques du corpus ***Convention***. Comme nous l'avons montré au chapitre 3, il s'agit notamment des phrases incluant une marque de

ponctuation forte au milieu (point, deux points, etc.). Les divergences dans la structure des arbres de la classification automatique obtenus pour ces phrases ont servi de points de repère pour identifier le décalage dans l'appariement.

5.4 Conclusion du chapitre 5

Au cours de ce chapitre, nous avons montré à plusieurs reprises que les renseignements obtenus à partir de la classification automatique peuvent être utilisés avec profit pour le traitement des données textuelles multilingues[16]. L'application des procédures de la classification automatique aux unités textuelles permet la conception de systèmes destinés à produire des appariements lexicaux. Les classes des segments répétés apportent des indices particulièrement utiles pour l'alignement et permettent de se rapprocher de la notion d'unité de traduction.

Pour parvenir à recenser des correspondances plus fines au niveau des mots et des syntagmes, il est souhaitable d'appliquer la classification automatique à des tableaux lexicaux croisant en lignes les formes et les segments d'un corpus parallèle et en colonnes les fragments textuels correspondant à des phrases appariées de ce corpus.

A la suite de nos expériences sur les unités textuelles, nous avons montré qu'une classification appliquée à des phrases issues de textes multilingues peut servir de support à l'automatisation de la vérification de l'appariement phrastique. Cette constatation devrait servir de point de départ à la conception de procédures destinées à repérer et corriger des erreurs dans la mise en correspondances des fragments textuels (phrases, paragraphes, sections) des deux volets d'un corpus parallèle.

[16] On trouvera des renseignements supplémentaires sur ces questions sur le Cd-rom joint : /fichierCD/stmz/page5_fichiers/JADT_2000.ppt.

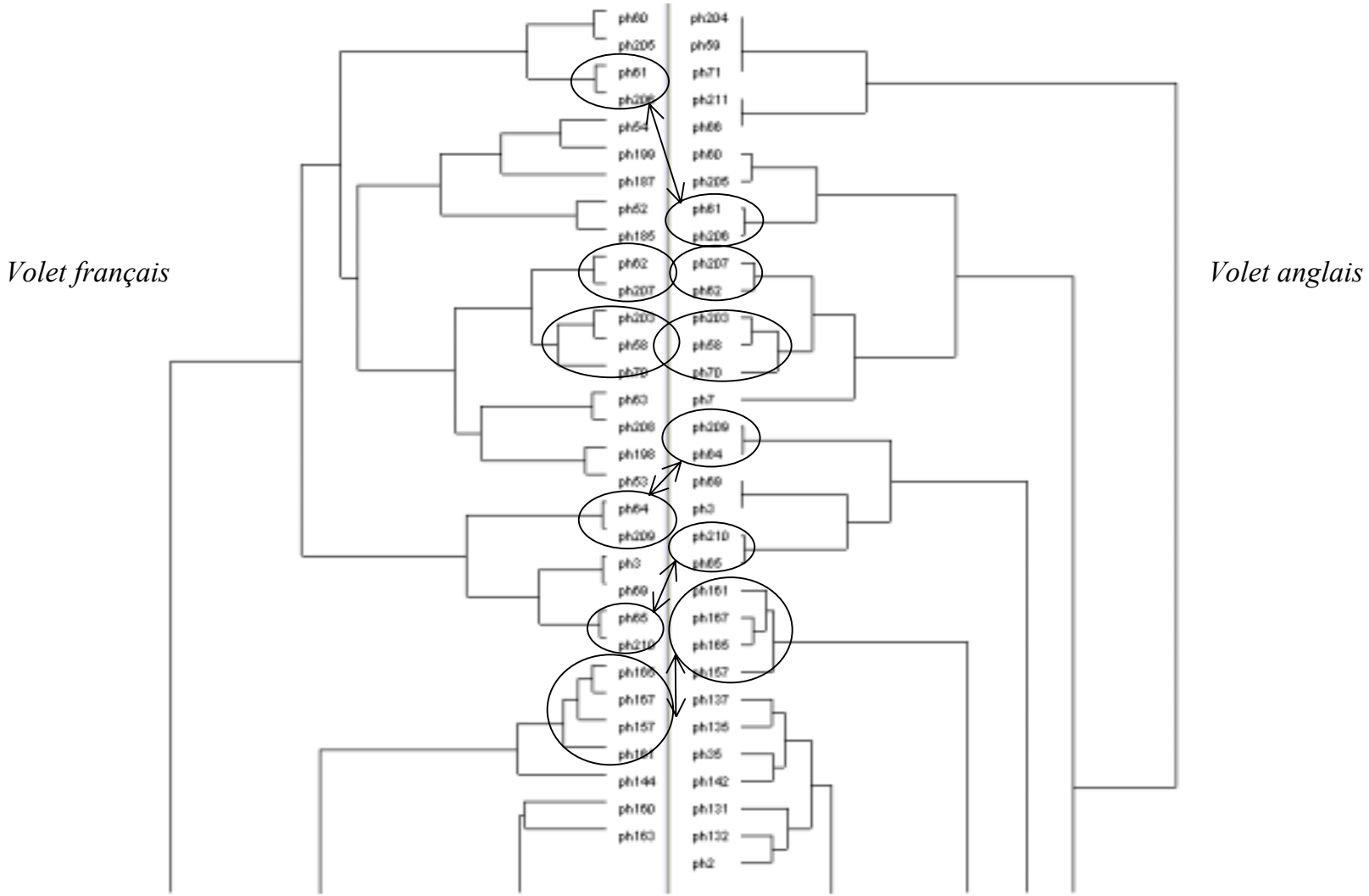


Figure 5.11 : Validation de l'alignement des phrases du corpus ***Convention*** à base de la *CAH*

« Un même mot n'est pas forcément employé dans le même sens selon les parties d'un corpus. Ces écarts se manifestent en particulier par des associations (syntaxiques ou de simple co-présence) avec des mots différents d'une partie à l'autre. »

[Habert, 2000, p. 1] [1]

Chapitre 6

Les univers lexicaux parallèles

Nous avons montré au cours des chapitres précédents que l'analyse des unités composées de type *segment répété* fournit des indices pour l'étude des équivalences lexicales dans les contextes parallèles. Pour mettre en évidence des régularités dans la structuration des deux discours il est nécessaire de convoquer des méthodes qui vont au-delà de la comparaison des formes et de leurs associations immédiates. Dans ce chapitre, nous proposons de nouveaux moyens automatisés par mettre en valeur les correspondances lexicales à l'aides des réseaux parallèles de cooccurrences [Martinez et Zimina, 2002].

6.1 Les attractions lexicales binaires

L'utilisation de la méthode des *cooccurrences* [2] pour l'étude des correspondances lexicales du bi-texte est fondée sur une hypothèse qu'il existe pour les formes en correspondance de traduction un parallélisme traductionnel entre univers

[1] Habert Benoît (2000). « Convergences et divergences de contextes normalisés : CFTC 1945-1964, CFDT 1965-1985 et CFTC 1965-1985. » Polycopié du séminaire de DEA *Langages sociaux et politiques. Analyse du discours & lexicométrie*. 3 mai 2000, ILPGA, Université de la Sorbonne Nouvelle – Paris 3.

[2] Sur les méthodes des cooccurrences, on consultera [Lafon, 1984] ; [Labbé *et al.* 1988] ; [Church et Hanks, 1990] ; [Haruno *et al.* 1996] ; Martinez [2000 ; 2003].

lexicaux. Selon ce principe, deux formes pôles représentant des traductions mutuelles doivent être en relation de cooccurrence avec les formes qui sont également en correspondance de traduction. Le repérage des unités qui ont tendance à faire partie du voisinage lexical d'un couple de formes équivalentes offre ainsi de nouveaux moyens formels pour la mise en évidence des rapports de correspondances traductionnelles au niveau des mots et des syntagmes.

6.1.1 Les voisinages lexicaux de pôles équivalents

Fondé sur le *modèle hypergéométrique*[3], la méthode des cooccurrences spécifiques mesure les attractions lexicales les plus intenses autour d'un pôle donné et livre des résultats sous la forme d'une liste hiérarchisée. Pour le recensement automatique des cooccurrences il est impératif de définir une unité de contexte (ou *voisinage*) à l'intérieur de laquelle on considérera que deux formes sont cooccurrentes. L'appariement des phrases d'un corpus parallèle offre des moyens formels permettant la fragmentation des deux volets bilingues en unités de contexte comparables. Dans nos expérimentations, la fenêtre de l'exploration contextuelle sera la phrase dans le cas où il existe une correspondance phrase pour phrase entre les deux volets du corpus. Lorsque deux phrases d'un volet correspondent à une phrase de l'autre volet, nous considérerons ces deux phrases comme une seule unité de contexte. Pour chaque forme pôle, le calcul des cooccurrences débute par le repérage de l'ensemble des phrases du corpus où elle est attestée. Les cooccurrents du pôle sont des formes dont les sous-fréquences sont spécifiques du sous-ensemble de phrases sélectionnées.

Comme nous l'avons vu au chapitre 5, parmi les formes agrégées dans les mêmes classes par la classification automatique, nous retrouvons l'équivalence des formes *droits/rights*. La confrontation des fréquences locales de ces formes

[3] L'exploitation du modèle hypergéométrique pour le calcul des cooccurrences spécifiques s'inspire de la méthode élaborée par 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 $\pm Exx$ où le signe indique un *sur-emploi* ou un *sous-emploi* de la forme et la valeur indique son degré de spécificité.

permet de vérifier leur évolution parallèle dans les deux volets du corpus *Convention* découpé en vingt parties (cf. *Figure 6.1*). Cependant, les ressemblances dans les profils de ventilation de ces formes n'apportent qu'une indication superficielle sur les correspondances entre univers lexicaux. 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.

Le tableau 6.2 montre une partie des cooccurrences spécifiques calculées pour chacun des pôles bilingues *droits* et *rights*. 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 sur lesquels repose l'équivalence traductionnelle au niveau lexical. Cette information permet d'apparier non plus des pôles isolés mais deux ensembles de formes cooccurrentes [4].

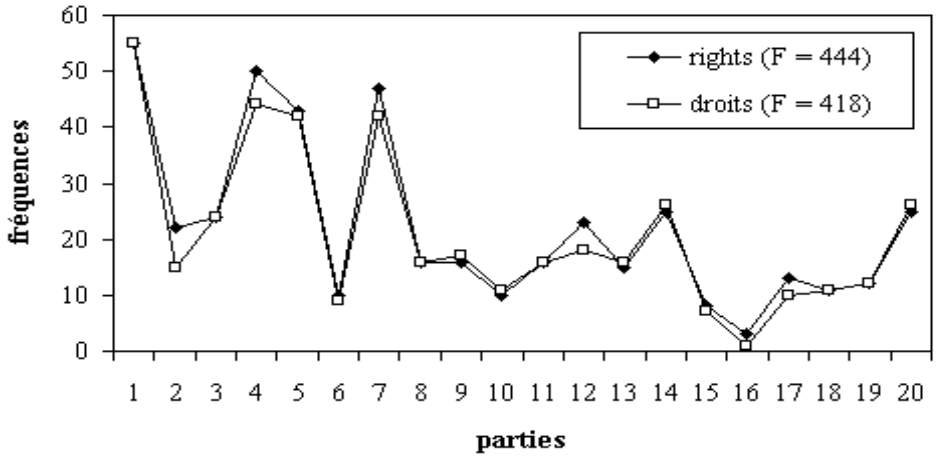


Figure 6.1 : Profils de répartition des formes *droits* et *rights* au sein du corpus ***Convention*** découpé en vingt parties consécutives

[4] On trouvera en annexe C d'autres exemples d'explorations réalisées à l'aide de la méthode des cooccurrences spécifiques.

Tableau 6.2 :
Principaux cooccurrents binaires des formes *droits* et *rights* dans l'unité contextuelle de la phrase jusqu'au seuil *+E15*

cooccurrents de *droits*				cooccurrents de *rights*			
FRA	**F tot.**	**co-fréq**	**spéc.**	**ANG**	**F tot.**	**co-fréq**	**spéc.**
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 bilingue *droits/rights* dans une fenêtre d'exploration contextuelle donnée (ici, la phrase). A partir des fréquences totales 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 formes-pôles sont très ressemblants.

6.1.2 Limites des attractions lexicales binaires

L'exploitation des résultats issus du calcul des cooccurrences binaires reste limitée dans la mesure où la méthode en question ne fournit un indice précis que

pour l'association d'un pôle avec chacun de ses cooccurrents. Le schéma sur la figure 6.3 décrit le cas général des rencontres entre cooccurrents binaires dans chaque contexte multilingue :

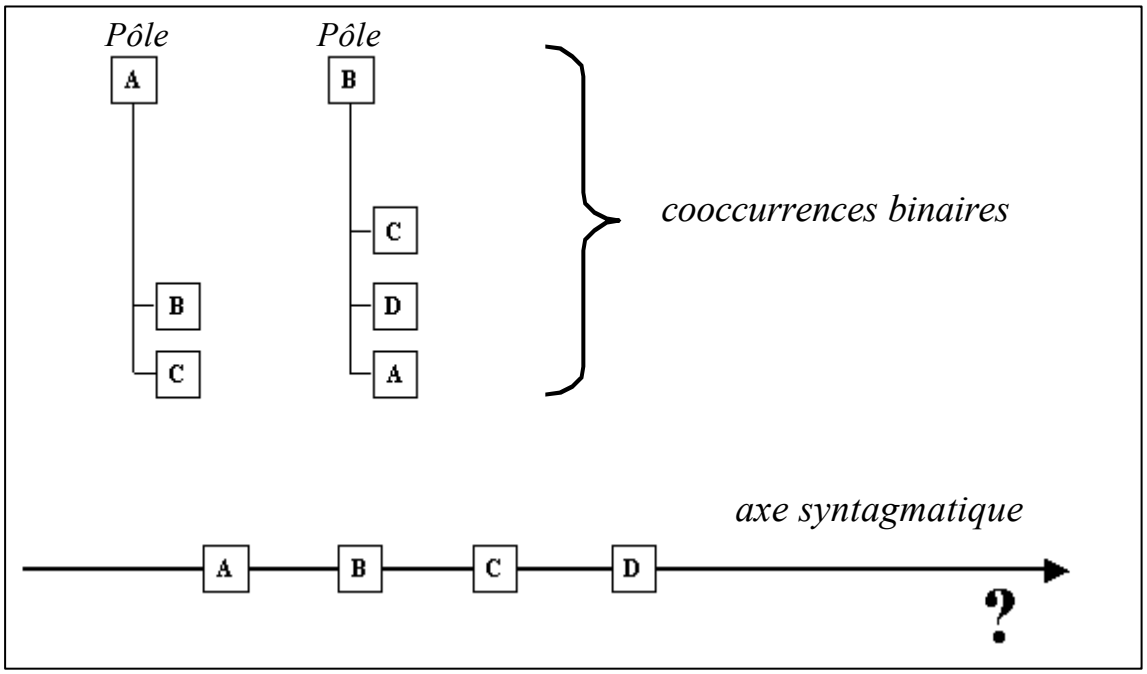


Figure 6.3 : Phénomènes de cooccurrences binaires

Guide de lecture : Sur le schéma, *A* (dont les cooccurrents binaires sont *B* et *C*) et *B* (dont les cooccurrents binaires sont *C*, *D*, et *A*) sont des formes dans un contexte monolingue. Par ailleurs, elles se rencontrent ensemble dans certaines phrases.

Le schéma représenté sur la figure 6.3 permet de constater que le calcul des cooccurrences binaires n'offre aucune certitude quant aux attractions multiples entre les formes. Le retour au contexte est inévitable pour vérifier toute association multiple qui pourrait être suggérée par les résultats.

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. L'approche suggérée par Martinez [2000 ; 2003] permet de mesurer la cooccurrence en tant que

phénomène global en considérant les formes cooccurrentes comme parties intégrantes d'un système dont les différents composants entretiennent une relation de dépendance simultanée. Les résultats de l'exploitation de cette nouvelle approche pour l'appariement font l'objet de la section qui suit.

6.2 Les attractions lexicales multiples

La méthode des *cooccurrences multiples* permet d'explorer les contextes spécifiques d'un pôle et révèle à chaque étape du calcul un élément supplémentaire du réseau de cooccurrences qui s'élabore à partir de ce pôle. Le calcul itératif qui est à la base de cette méthode permet d'aller au-delà des attractions binaires, *forme à forme* et laisse apparaître des associations plus élaborées sur l'axe syntagmatique qui lient le pôle à plusieurs formes cooccurrentes. La méthode permet d'aboutir à l'extraction automatique de réseaux structurés des poly-cooccurrences (cf. *Figures 6.4-6.5a* et *6.5b*).

6.2.1 Mise en évidence des structures lexicales parallèles sur l'axe syntagmatique

La méthode de prospection contextuelle qui est à la base du calcul des cooccurrences multiples se distingue par deux caractéristiques principales : la *récursivité* et la *profondeur de l'exploration* [Martinez, 2003, p. 293].

1) *Récursivité* : La figure 6.4 décrit de manière schématique l'algorithme permettant la réitération du calcul des cooccurrences autour d'un pôle donné. Sur ce schéma, le signe '+' symbolise la notion de *cooccurrence chaînée* qui est à la base de la procédure [5].

2) *Profondeur de l'exploration* : En plus de l'épuisement naturel de l'exploration, la méthode élaborée par Martinez permet de contraindre la prospection

[5] Sur la description de réitération du dépouillement statistique à de multiples niveaux d'exploration, on consultera le fichier suivant sur le Cd-rom : /fichiersCD/stmz/page5_fichiers/JADT_2002.ppt.

contextuelle par différents paramètres tels qu'une *co-fréquence minimum*, un *seuil de spécificité* de la rencontre ainsi *qu'un nombre minimum de contextes* dans lesquels se produit la cooccurrence.

En plus de seuils définis en amont de la détection des cooccurrences spécifiques, Martinez propose de procéder après leur recensement à un filtrage des résultats. Il distingue trois types de *chemins de cooccurrence* de pertinences différentes [Martinez, 2003, p. 294] :

- les *chemins de cooccurrences explorés*[6] qui correspondent aux résultats exhaustifs de la prospection contextuelle ;

- les *chemins de cooccurrences retenus* (par l'exploration) qui répondent aux seuils fixés en amont par l'utilisateur ;

- les *chemins de cooccurrences originaux* qui réduisent l'information rapportée par la méthode en excluant les poly-cooccurrences qui se contiennent (cf. *Figure 6.4*).

[6] Ces chemins comprennent les poly-cooccurrences dont la *co-fréquence*, la *spécificité* et le *nombre de contextes* ne satisfont pas les seuils définis au début de l'exploration contextuelle. Pour écarter ces *chemins de cooccurrence*, il faut d'abord identifier tous les chemins en les parcourant.

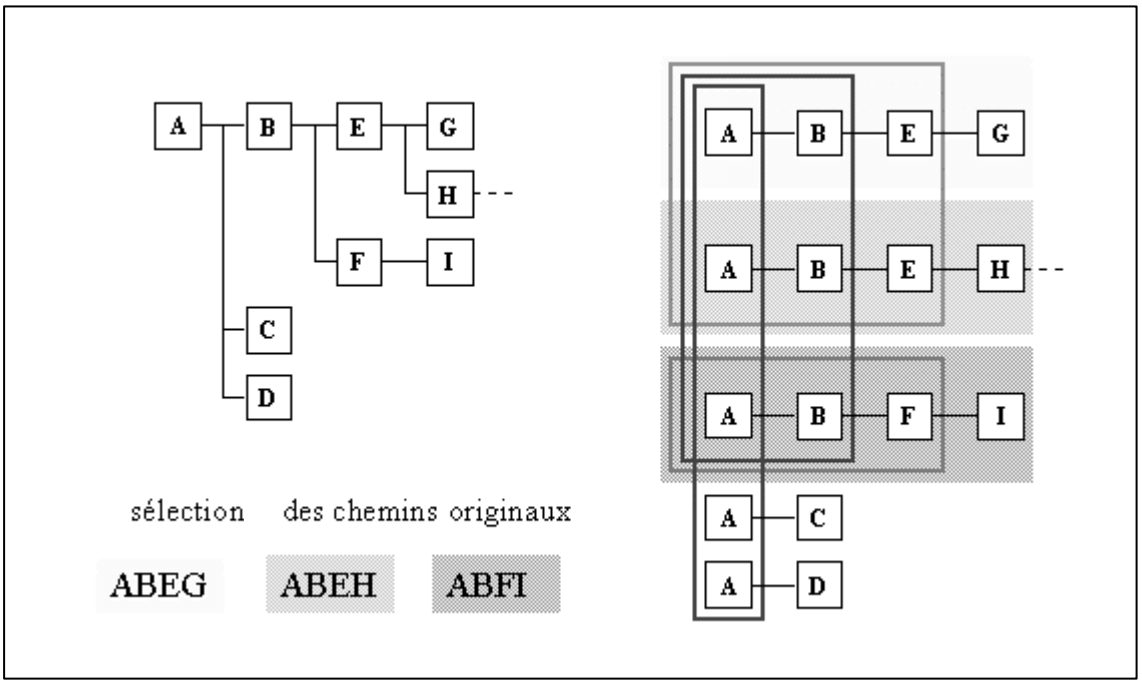


Figure 6.4 : Phénomènes de cooccurrences multiples

Guide de lecture de la figure 6.4 :

Le schéma sur la figure 6.4 décrit l'algorithme de calcul de *réseaux de poly-cooccurrences* :

Étape 1 : Le pôle *A* a pour cooccurrents spécifiques *B*, *C* et *D*

Étape 2 : Le pôle *A+B* a pour cooccurrents spécifiques *E* et *F*

Étape 3 : Le pôle *A+B+E* a pour cooccurrents spécifiques *H* et *G*

Étape 4 : Le pôle *A+B+E+H* n'a pas de cooccurrent spécifique
alors l'exploration s'interrompt pour ce chemin

Étape 5 : Le pôle *A+B+F* a pour cooccurrents spécifiques *I*, etc.

Au cours du calcul, différents filtrages limitent le nombre des explorations contextuelles et réduisent le bruit. En plus de l'épuisement naturel des explorations de chaque suite de coocurrents, des filtres comme la fréquence et la spécificité minimale du cooccurrent ainsi que le nombre de contextes où apparaît le phénomène réduisent le champ d'investigation. A l'issue du calcul, on sélectionne parmi les résultats exhaustifs les chemins originaux en écartant les chemins qui :

- se recouvrent partiellement, comme *AB*, *ABC* et *ABCD*, contenus dans *ABCDE*

- se répètent, comme *ACB*, *BAC*, *BCA*, *CAB* et *CBA*, contenus dans *ABC*.

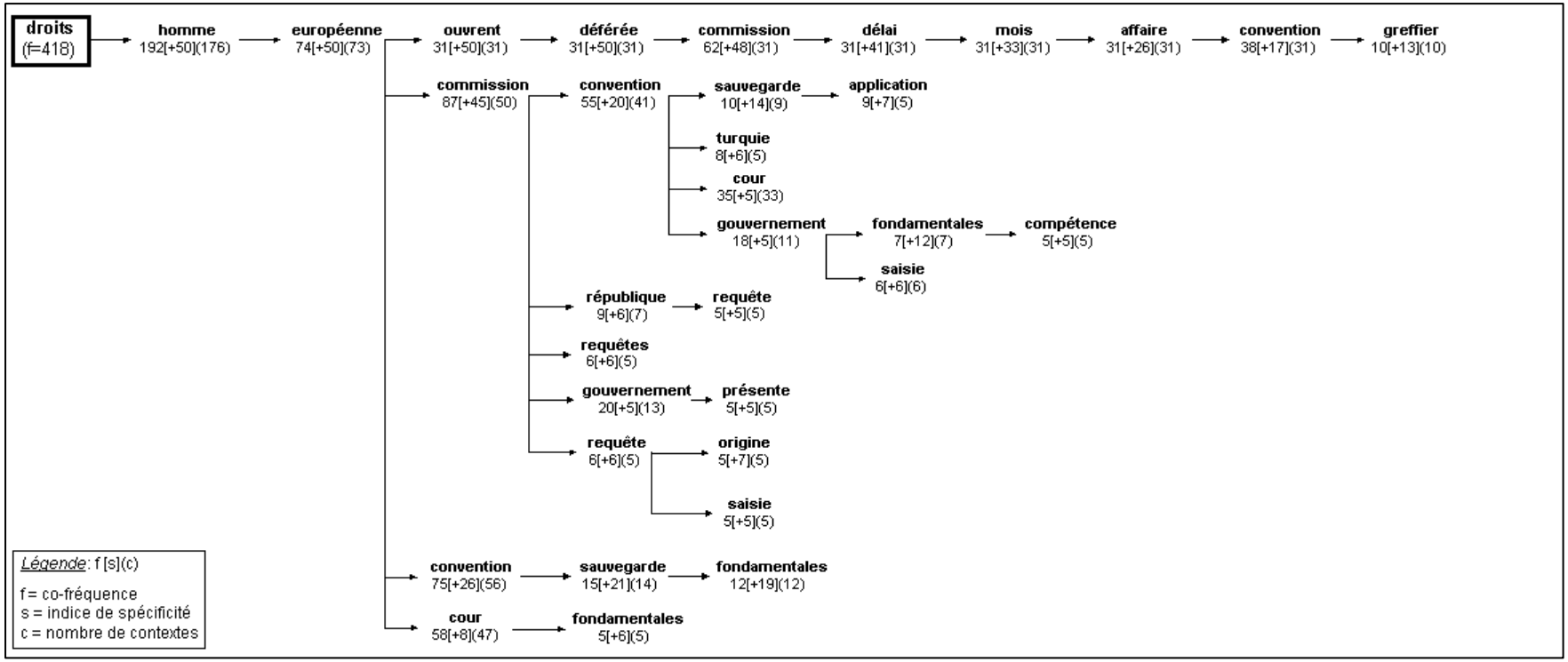


Figure 6.5a : Vue partielle du réseau de cooccurrences élaboré à partir du pôle *droits*

Guide de lecture de la figure **:** Fondée 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 *droits*. Sur la première ligne qui correspond à la branche la plus spécifique du réseau, est rapportée une forte cooccurrence du pôle avec la forme *homme :* 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 *européenne* dans 74 phrases. En répétant ainsi le processus de comptage jusqu'à épuisement on détermine des chemins de cooccurrence correspondant à des "squelettes" de phrases avérées dans le corpus.

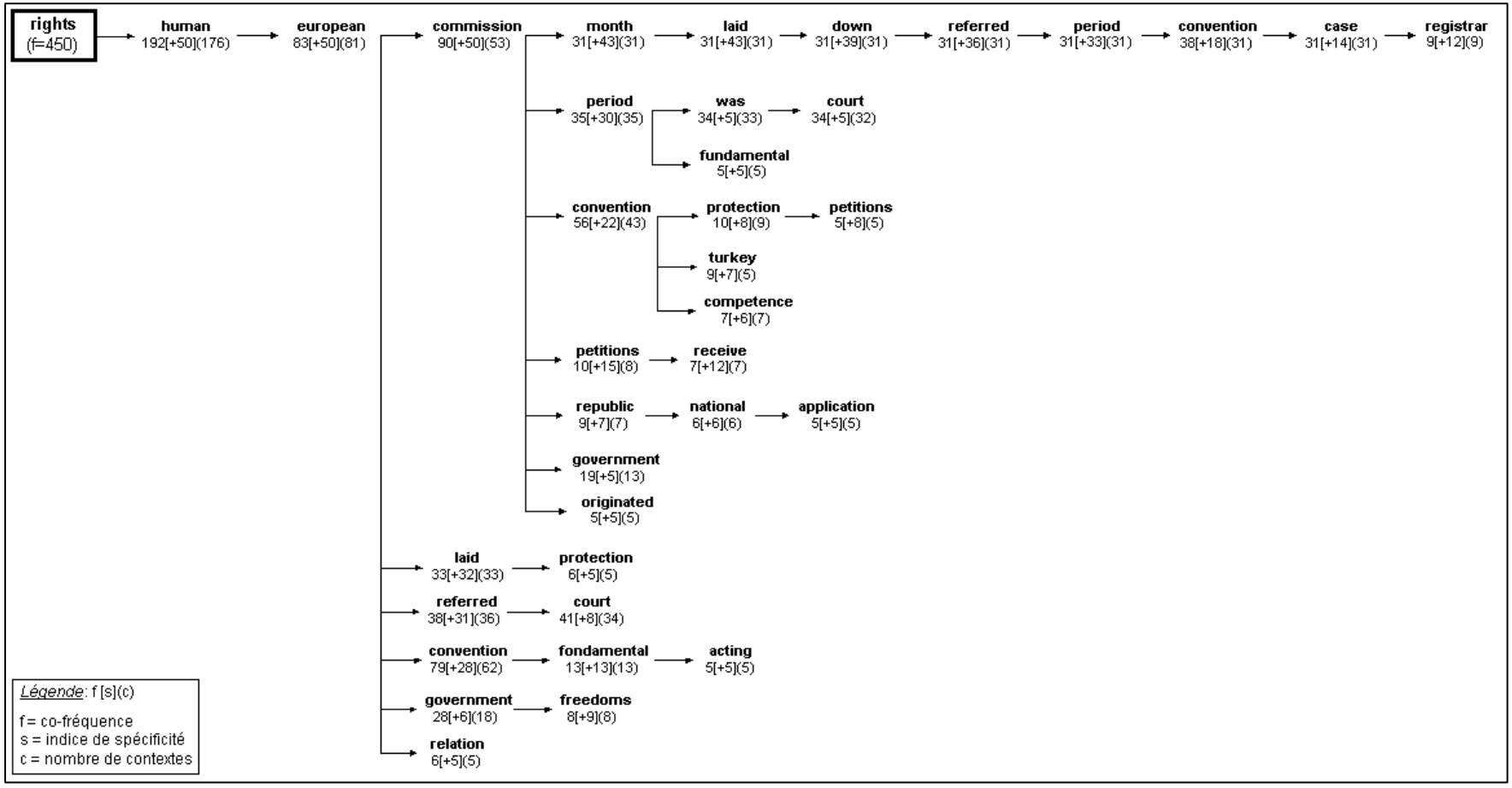


Figure 6.5b : Vue partielle du réseau de cooccurrences élaboré à partir du pôle *rights*

Tableau 6.6 : Mise en évidence de contextes spécifiques

volet français	volet anglais
1. Ratio : 0.41 *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 #[7] de la **convention**.	1. Ratio : 0.48 *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 *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 *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 *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 *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 *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 *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 *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 *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**.

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

Tableau 6.6 : Mise en évidence de contextes spécifiques (suite)

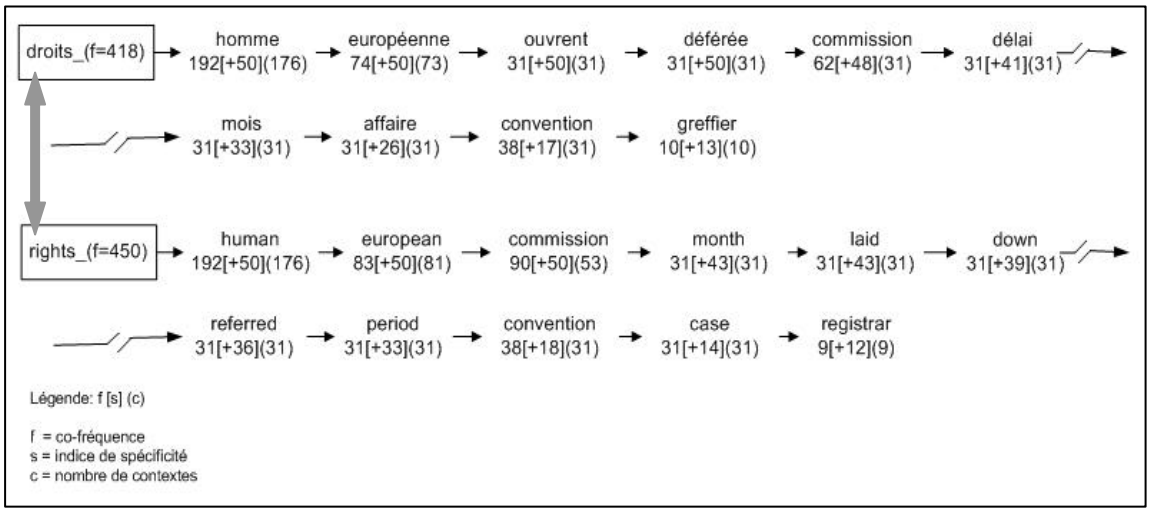
volet français	volet anglais
6. Ratio : 0.41 *n° 9349 :* l'**affaire** a été **déférée** à la cour par la **commission européenne** des **droits** de l'**homme** ("la **commission**"), dans le **délai** de trois **mois** qu'**ouvrent** les # de la **convention**.	6. Ratio : 0.48 *n° 9593 :* the **case** 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 *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 *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 *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 *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 *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 *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**.

6.2.2 Le calcul des réseaux parallèles de correspondances de traduction

A partir de la liste des principaux cooccurrents des formes *rights/droits* dont nous avons supprimé les *mots-outils* [8], il est possible de construire les univers lexicaux

[8] Cette dénomination recouvre les formes très fréquentes dans le corpus qui remplissent des fonctions essentiellement syntaxiques tels que les articles, les conjonctions, que l'on distingue par rapport aux mots sémantiquement pleins.

des pôles tels qu'ils s'élaborent au fur et mesure dans les deux volets bilingues du corpus. L'application de la méthode des cooccurrences multiples parallèlement aux corpus français et anglais révèle une correspondance presque totale entre les principaux chemins (les plus spécifiques) élaborés à partir des pôles (cf. *Figures 6.5a-b*) :



L'analyse de l'ensemble de réseaux de cooccurrents permet de saisir la totalité de formes qui entretiennent des liens sémantico-syntaxiques avec les pôles bilingues dans les deux volets du corpus. Pour faciliter la visualisation des réseaux les résultats du calcul sont présentés sous la forme d'arborescences (cf. *Figures 6.5a-b*). 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 contextes spécifiques où se réalisent ces réseaux de cooccurrences (cf. *Tableaux 6.6ab*). En comparant le nombre de cooccurrents spécifiques et le nombre de mots dans l'unité contextuelle (ici la phrase), le ratio mesure la densité de chaque contexte par rapport au nombre de formes spécifiques qui y figurent. La présentation des deux listes en vis-à-vis laisse apparaître une symétrie complète entre les corpus français et anglais.

On remarque sur la figure 6.5ab que les attractions lexicales simultanées des formes constituent un facteur de désambiguï sation. Ainsi, l'examen des contextes

où apparaît le mot anglais *case* montre qu'il possède des sens différents dans le corpus ***Convention*** :

français	**anglais**
/.../auquel **cas** la première incarcération doit toujours être/.../	/.../in which **case** the first period of imprisonment must always be /.../
/.../toute partie à l'**affaire** peut, dans des cas exceptionnels , demander le/.../	/.../any party to the **case** may , in exceptional cases , request that/.../
/.../qu'il ait eu accès au **dossier** en possession de la cour suprême/.../	/.../he had access to the **case** file/.../
/.../laquelle ne saurait parer à toute **éventualité**/.../	/.../which cannot in any **case** provide for every eventuality /.../
/.../l'**affaire** a été déférée à la cour/.../	/.../the **case** was referred to the court/.../

En revanche, dans les contextes où se réalisent les réseaux de cooccurrences établis autour des pôles *droits/rights*, le mot *case* est toujours traduit en français par *affaire* (cf. *Tableau 6.4*)[9]. Contrairement aux fréquences totales dans le corpus, les co-fréquences des formes *case/affaire* dans les contextes où sont attestés les pôles *droits/rights* sont identiques. On peut ainsi procéder à l'appariement de ces formes :

	Fréquence totale (F)	*co-fréquence (f)*
case	209	31
affaire	93	31

Comme on le voit, la structure des correspondances lexicales entre les deux volets d'un corpus parallèle est complexe. Sur le plan fréquentiel, cette complexité se manifeste par des écarts entre les fréquences totales de mots qui ne constituent des équivalences traductionnelles que dans certains contextes.

[9] Le dictionnaire *Longman Dictionary of Contemporary English* [1987] définit ce sens du mot *case* comme : *"/.../ a question to be decided in a court of law"*.

Globalement, on peut effectuer le rapprochement des unités textuelles bilingues en s'appuyant sur des indices fréquentiels si l'on tient compte de leurs occurrences attestées dans les portions de textes équivalentes au plan traductionnel.

Comme nous l'avons montré au chapitre 4, la forme française *fonctionnaires* (F=49) possède plusieurs traductions dans le volet anglais du corpus [10]. Pour analyser l'univers sémantique de cette forme, nous allons réaliser le calcul des réseaux des cooccurrences séparément pour chaque sous-ensemble d'occurrences ayant reçu des traductions différentes dans le volet anglais. Cette sélection est envisageable grâce à l'*approche topographique* du bi-texte que nous avons décrite au chapitre 4.

En ayant recours aux fonctionnalités de la *carte des sections* (cf. *Chapitre 4, section 4.2.2*), nous allons rassembler au sein de ***trois pôles distincts*** :

⇒ les occurrences de la forme française *fonctionnaires* traduites en anglais par *civil servants* ;

⇒ les occurrences qui correspondent à la forme anglaise *officers* ;

⇒ les occurrences qui sont des traductions de la forme *officials* en anglais.

Le calcul des réseaux des cooccurrences sera ainsi effectué à partir de trois pôles bilingues composés des sous-ensembles d'occurrences attestées dans les contextes équivalents :

Pôle N°	**forme (FRA)**	**traduit par (ANG)**	***NbUC*** [11]
1	*fonctionnaires*	*servants*	29
2	*fonctionnaires*	*officers*	9
3	*fonctionnaires*	*officials*	7

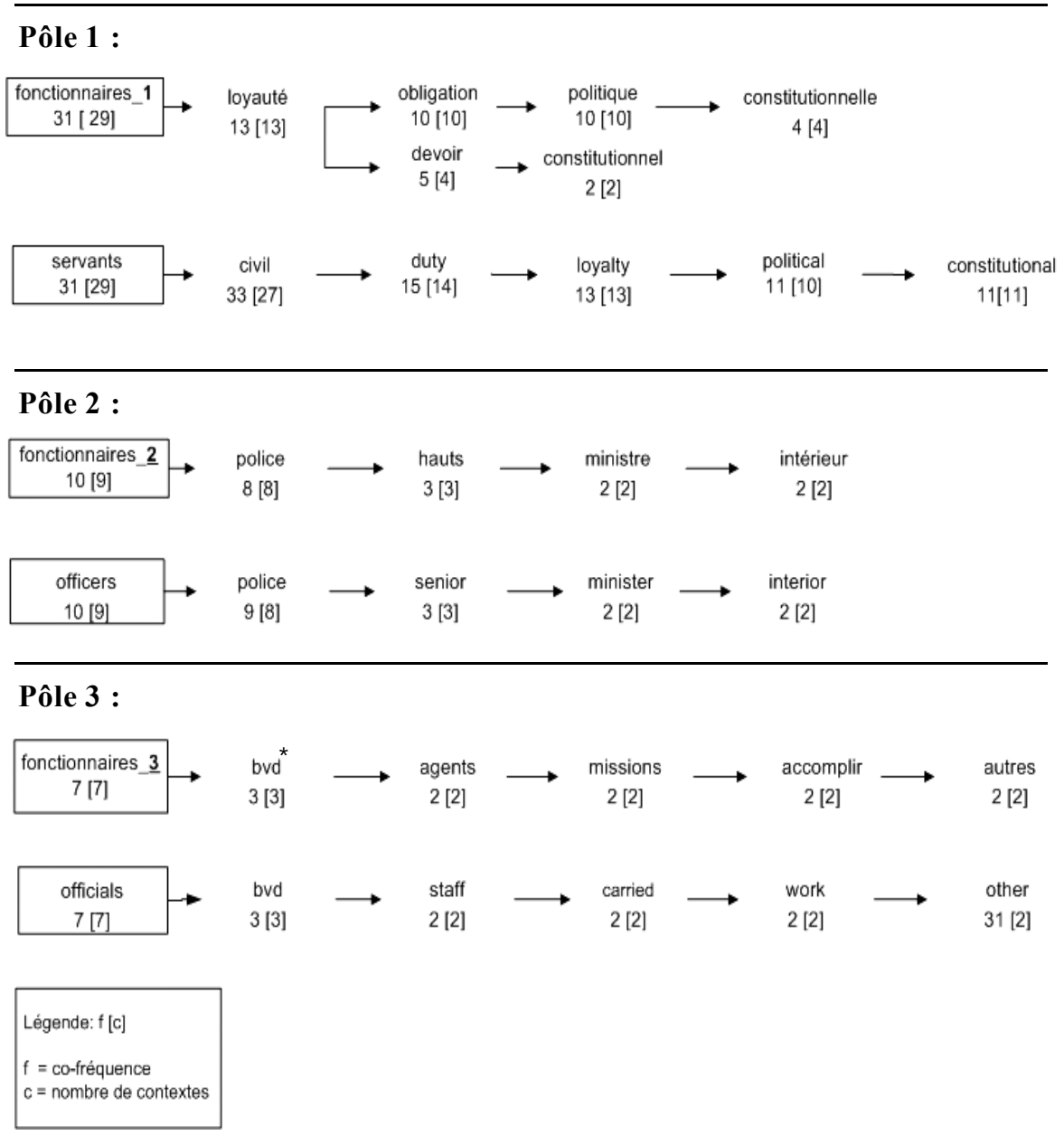
[10] Nous avons montré qu'il possible de réaliser une exploration topographique du corpus parallèle ***Convention*** pour mettre en évidence l'ensemble de lexèmes qui correspondent au mot français *fonctionnaire* dans le volet anglais du corpus. Les principales étapes de cette exploration sont présentées sur le Cd-rom : /fichierCD/stmz/page5_fichiers/JADT_2004F.ppt. La description sommaire de l'exploration topographique est disponible également dans les annexes à la fin de ce volume.

[11] *NbUC* : Nombre d'unités de contextes où est attesté le pôle bilingue.

L'analyse des résultats de cette expérimentation présentés sur la figure 6.7, permet de constater que les chemins les plus spécifiques recensés à partir des pôles bilingues sélectionnés selon ces critères sont sensiblement différents. Dans les corpus parallèles, cette approche du calcul des réseaux de cooccurrences produit des indices plus précis pour l'appariement. Il permet de nuancer les univers sémantiques des formes-pôles et de découvrir de nouvelles attractions lexicales qui reflètent d'avantage le parallélisme de ces univers (cf. *Tableau 6.8*).

A l'issue de nos explorations, nous avons constaté que la représentation des attractions lexicales sous forme de réseaux de cooccurrences facilite la constitution de ressources bi-textuelles appariées. Les univers lexicaux parallèles découverts grâce à la méthode des cooccurrences multiples pourraient servir à des études de corpus parallèles dans des domaines variés, tels que la lexicographie bilingue, la synthèse et l'extraction de l'information multilingue sur le Web, etc.

Les réseaux parallèles des cooccurrences donnent accès à des ressources traductionnelles particulièrement riches lorsque l'on dispose de corpus parallèles étiquetés. Dans le chapitre qui suit, nous monterons quelques exemples d'exploration textométrique du bi-texte catégorisé.



* BVD (Binnenlandse Veiligheidsdienst) est le nom du service de sécurité intérieure de l'Allemagne cité dans le corpus ***Convention***.

Figure 6.7 :
Vue partielle de réseaux de cooccurrences élaborés à partir de trois pôles de correspondances traductionnelles de la forme française *fonctionnaires* (F=49)

Guide de lecture : Pour comparer l'univers sémantique du mot français *fonctionnaire* à celui qui lui correspond dans le volet anglais, on calcule les réseaux des cooccurrences séparément pour chaque sous-ensemble d'occurrences du mot *fonctionnaire* ayant reçu des traductions différentes dans le volet anglais.

Tableau 6.8 :
Retour au contexte pour les réseaux de cooccurrences élaborés à partir des pôles bilingues représentés sur la figure 6.7

Pôle 1 : ***fonctionnaires / servants***	
volet français	**volet anglais**
l'**obligation** de **loyauté politique**, qui certes restreignait les droits fondamentaux des **fonctionnaires**, figurait parmi les principes traditionnels de la fonction publique et avait valeur **constitutionnelle** /…/	the **duty** of **political loyalty**, which admittedly restricted **civil servants**' fundamental rights, was one of the traditional principles of the civil service and had **constitutional** status /…/
Pôle 2 *:* ***fonctionnaires / officers***	
volet français	**volet anglais**
il dénonce les propos tenus lors de la conférence de presse par le **ministre** de l'**intérieur** et les **hauts fonctionnaires** de **police** qui l'accompagnaient.	he complained of the remarks made by the **minister** of the **interior** and the **senior police officers** accompanying him at the press conference.
Pôle 3 : ***fonctionnaires / officials***	
volet français	**volet anglais**
daté de 1981 et classé « confidentiel » , ce document visait principalement à informer de ses activités les **agents** du **bvd** et d'**autres fonctionnaires** appelés à **accomplir** des **missions** pour lui .	dated 1981 and marked "confidential " , it was designed mainly to inform **bvd staff** and **other officials** who **carried** out **work** for the **bvd** about the organisation's activities .

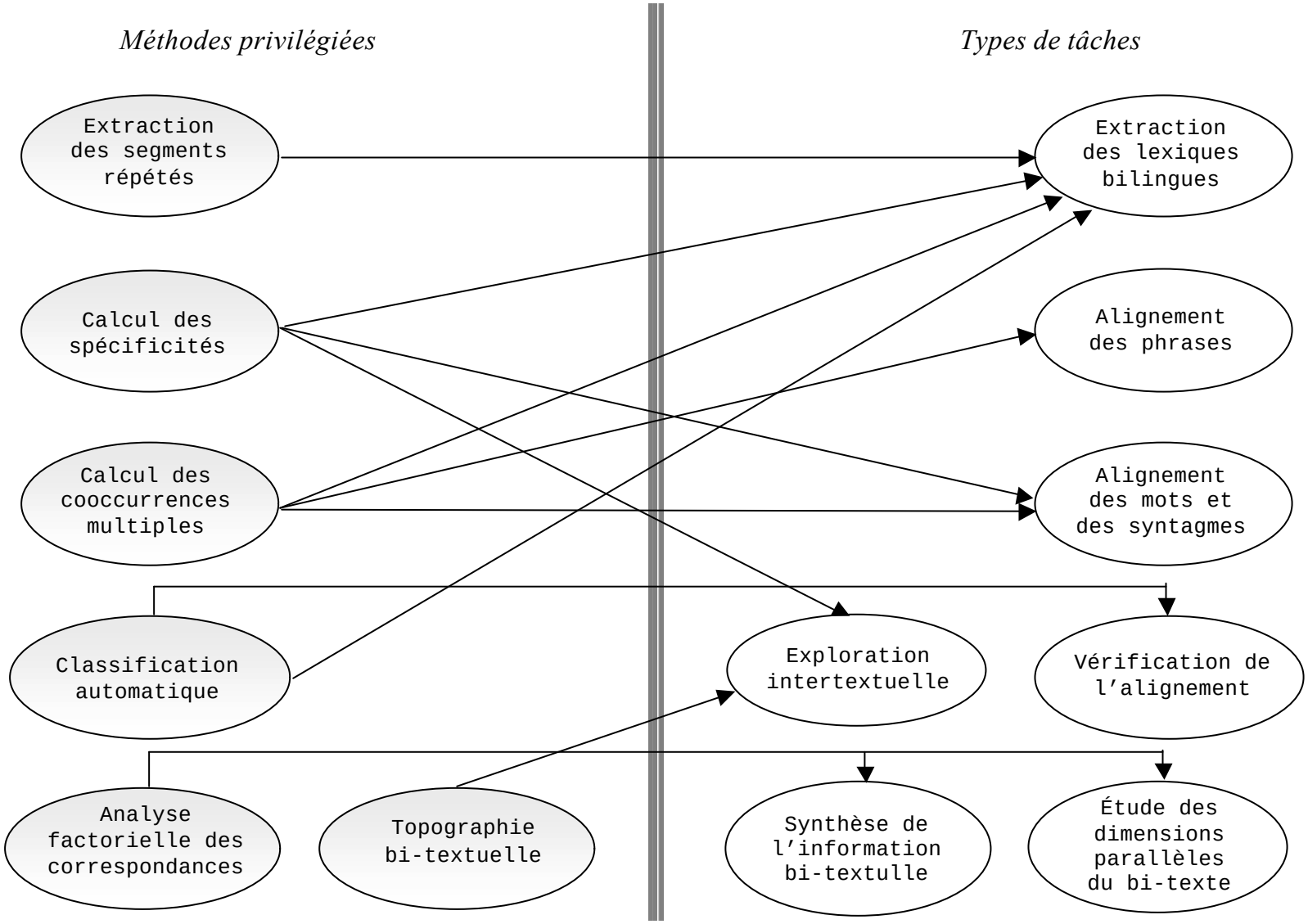


Figure 6.9 : Les utilisations des méthodes de la textométrie pour l'analyse de corpus parallèles

6.3 Conclusion sur les méthodes d'alignement

Dans les chapitres précédents, nous avons montré que les méthodes de la statistique textuelle peuvent servir à l'analyse et l'extraction de ressources bi-textuelles des corpus parallèles. Développées dans le contexte monolingue, les pratiques de l'analyse textométrique de corpus se révèlent particulièrement adaptées à la recherche automatique des équivalences du bi-texte.

Dans le cas des corpus parallèles bilingues, la textométrie aide à mettre en relation automatiquement différents *types* d'unités textuelles entre les deux volets. L'approche quantitative permet d'établir des correspondances aussi bien entre les paragraphes et les phrases, qu'au niveau lexical. Grâce à cette approche, on parvient à mette en relation des *formes graphiques* isolées, des *lexèmes*, des *structures lexicales récurrentes* sur l'axe syntagmatique, etc. Comme nous l'avons montré au cours de nos recherches, il appartient à l'expert humain de préciser les normes de dépouillements en unités de décomptes servant de base pour la recherche et l'extraction des équivalences du bi-texte.

Certaines méthodes textométriques sont particulièrement adaptées à la réalisation de tâches précises que l'on retrouve souvent dans l'étude informatisée de corpus parallèles. Le schéma sur la figure 6.4 établit un rapport entre le type de tâche envisagée et la (ou les) méthode(s) privilégiée(s) pour l'accomplir. Par exemple, on lit sur ce schéma que le *calcul des cooccurrences multiples* contribue à l'alignement phrastique et lexical des corpus parallèles, tandis que *l'analyse factorielle des correspondances* est plus adaptée à la mise en évidence des dimensions parallèles du bi-texte et à la synthèse de l'information bi-textuelle.

Plusieurs méthodes textométriques peuvent être utilisées conjointement. Ainsi, comme nous l'avons montré au cours des chapitre 3-6, on peut mobiliser la méthode des *segments répétés*, les *cooccurrences multiples*, le calcul des *spécificités* et la *classification automatique* des formes et segments répétés pour l'extraction des *lexiques bilingues* des corpus parallèles (cf. *Figure 6.9*).

Chapitre 7

Les perspectives de la textométrie multilingue

Les expérimentations présentées dans les chapitres précédents ont abouti à la création de nouvelles stratégies d'analyse de corpus parallèles sur des bases quantitatives. L'approche textométrique du bi-texte a permis de construire semi-automatiquement un ensemble de ressources traductionnelles à partir du corpus bilingue ***Convention***. Dans ce chapitre, nous tâcherons d'esquisser quelques pistes qui ont été encore peu explorées mais qui constituent pour nous des perspectives de recherche prometteuses pour l'analyse de corpus parallèles.

7.1 L'exploration textométrique du bi-texte catégorisé

Les dépouillements en formes graphiques fournissent des moyens simples pour constituer des unités textuelles à partir d'un corpus de texte. A l'issue des expériences sur le corpus ***Convention***, nous avons constaté que la segmentation de corpus parallèles en occurrences de formes graphiques constitue une base efficace pour dégager des ressources traductionnelles de ces corpus. Cependant, il est clair que la prise en compte des informations relatives au statut morphosyntaxique des formes permet de préciser des rapports de

correspondances au niveau des mots et des syntagmes. L'étiquetage automatique de corpus parallèles peut être utilisé pour ce type de tâche [1].

L'étiquetage d'un segment de texte (d'un mot, mais aussi d'un groupe de mot, d'une phrase, d'un paragraphe, etc.) est une opération qui consiste à lui associer des informations arbitrairement complexes qui peuvent se situer à plusieurs niveaux de l'analyse linguistique : morphologie, syntaxe, sémantique, pragmatique, sans se limiter d'ailleurs aux aspects linguistiques [Habert *et al.*, 1997, p. 21]. Lorsque l'on parle d'un corpus étiqueté dans la communauté du traitement automatique des langues (TAL), on fait référence le plus souvent à un document où chaque mot possède une étiquette morphosyntaxique et une seule.

Nous utiliserons également le terme *étiquetage* pour décrire les résultats de catégorisations morphosyntaxiques produits par un programme que l'on appelle un *étiqueteur* (ou *tagger*, en anglais). Pour étiqueter le corpus ***Convention***, nous avons utilisé le programme ***TreeTagger*** [2].

7.1.1 Hétérogénéité des étiquetages bilingues

TreeTagger [Schmid, 1994] est un programme de catégorisation à base de *méthodes probabilistes* [3]. Comme le remarquent Habert *et al.* [1997, p 167], la désambiguïsation probabiliste s'appuie sur le caractère positionnel des langues (en l'occurrence le français et l'anglais) qui fournit des contraintes locales. Parmi l'ensemble d'étiquettes possibles pour chacun des mots, le programme privilégie

---

[1] Actuellement, les corpus parallèles étiquetés sont de plus en plus utilisés pour l'analyse et l'extraction de ressources traductionnelles multilingues. Sur ces questions, on consultera, par exemple, [Ahrenberg *et al.*, 2000] ; [Blank, 2000] ; [Gaussier *et al.*, 2000] ; [Piperidis *et al.*, 2000].

[2] Nous remercions Serge Fleury (Maître de conférences à l'Université de la Sorbonne Nouvelle – Paris 3) pour l'aide qu'il nous a prodiguée dans la préparation de la version étiquetée du corpus ***Convention***. Le corpus catégorisé peut être consulté sur le Cd-rom qui accompagne ce volume : /fichiersCD/stmz/page1_fichiers/CONV_cat.txt. L'annexe D présente un extrait des résultats d'étiquetage.

[3] Pour la description du *TreeTagger*, on consultera le site dédié à cet outil à l'adresse suivante : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

l'étiquette la plus probable. Le choix de cette étiquette se fait au regard de l'historique des dernières étiquettes attribuées. Dans la pratique, cet historique se limite souvent aux deux ou trois dernières étiquettes attribuées. Le calcul se fait à partir des chaînes de Markov[4].

Dans notre cas, une catégorie a été attachée à chaque mot des volets anglais et français du corpus ***Convention*** à la fin de la procédure d'étiquetage. Les jeux d'étiquettes utilisés par le programme *TreeTagger* pour le français et l'anglais sont présentés en annexe D.1-2. L'analyse attentive des résultats de l'étiquetage montre que le logiciel commet des erreurs. Dans certains cas, les étiquettes attachées aux mots du corpus sont incorrectes ou imprécises. Il s'agit, le plus souvent, de limites purement techniques de l'étiqueteur. Certaines suites de catégories sont mal reconnues, par exemple [5] :

français		**anglais**	
Le	{Déterminant}	*Relevant*	{Adjective}
droit	{Nom}	*domestic*	{Adjective}
et	{Conjonction}	*law*	{Noun}
la	{Déterminant}	*and*	{Conjunction}
pratique	**{Adjectif}**	*practice*	{noun}
internes	{Adjective}		
pertinents	**{Nom}**		

Les jeux d'étiquettes utilisés pour les volets anglais et français du corpus ***Convention*** ne sont pas univoques (cf. *Annexe D*). Une homogénéisation de l'étiquetage des deux volets bilingues du corpus parallèle se révèle nécessaire avant l'exploration bi-textuelle. Par exemple, le jeu de catégories utilisé par *TreeTagger* pour l'étiquetage des noms est plus précis en anglais :

français	**anglais**
Nom	Noun singular or uncountable Noun plural
Nom propre	Proper noun singular Proper noun plural

[4] Sur les chaînes de Markov, voir, par exemple, [Rabiner et Juang, 1986].

[5] Les erreurs d'étiquetage apparaissent ci-dessous en gras.

Pour la comparaison de corpus parallèles, on souhaiterait des catégoriseurs utilisant les mêmes types d'étiquettes pour chacun des volets bilingues. Cependant, les différences d'information ne constituent pas un obstacle pour l'analyse automatique du bi-texte. Par exemple, les étiquettes des catégories *Noun singular / uncountable* et *Noun plural* (attribuées par *TreeTagger* aux noms communs du volet anglais du corpus ***Convention***) peuvent être remplacées par une seule étiquette *Noun*. Cette simplification permettrait le rapprochement avec le volet français où il n'existe qu'une seule catégorie *Nom* pour tous les noms communs. Cette homogénéisation implique donc des pertes d'informations.

Lorsqu'il existe des différences significatives dans les structures grammaticales des deux langues, l'homogénéisation des étiquetages se révèle encore plus délicate. Ainsi, le système des verbes du français est plus complexe que celui de l'anglais. Par exemple, l'impératif en anglais a la même forme graphique que l'infinitif. Le jeu d'étiquettes utilisé par *TreeTagger* pour l'anglais confond l'infinitif et l'impératif au sein de la même catégorie *Verb base form* :

français	**anglais**

`Verbe à l'infinitif`
`Verbe à l'impératif`
`Verbe au présent`
`Verbe au futur`
`Verbe au conditionnel`
`Verbe au participe présent`
`Verbe au participe passé`
`Verbe au passé simple`
`Verbe à l'imparfait`
`Verbe au subjonctif imparfait`
	`Verb base form`
	`Verb non 3rd person singular present`
	`Verb 3rd person singular present`
	`Verb gerund or present participle`
	`Verb past tense`
	`Verb past participle`

En comparant les jeux d'étiquettes utilisés par *TreeTagger* pour le français et l'anglais, on constate que les chercheurs qui les ont élaborés ont suivi des chemins différents (cf. Santorini, 1990 ; Schmid, 1994). Les jeux d'étiquettes définis pour chaque langue dépendent étroitement des méthodologies de recherche adoptées, des objectifs expérimentaux fixés au départ, etc.

Nous sommes consciente que les problèmes d'homogénéisation d'étiquetage de corpus parallèles méritent une analyse approfondie. Des étiqueteurs conçus

spécifiquement pour l'analyse de corpus multilingues, permettraient une meilleure exploitation de ressources traductionnelles du bi-texte catégorisé.

7.1.2 Perspectives de recherche utilisant les étiquetages bilingues

Certaines catégories sont employées conjointement. Par exemple, les prépositions contractées (*au*, *aux*, *du*, *des*) en français sont souvent suivies de noms. Ces suites de catégories composent les patrons syntaxiques que l'on parvient à recenser automatiquement dans les corpus à l'aide de calculs statistiques. La méthode des segments répétés que nous avons décrite au chapitre 3 peut être utilisée avec succès pour détecter nombre de patrons syntaxiques récurrents. La *figure 7.1* présente les quinze patrons les plus employés dans les volets français et anglais du corpus ***Convention***. En haut de la liste, nous retrouvons essentiellement des patrons liés à l'environnement syntaxique des noms des deux langues. Notons également les fréquences similaires des patrons français et anglais correspondant à la suite des catégories *Verbes au participe passé + Préposition* [6].

Les patrons syntaxiques recensés dans les deux volets du corpus ***Convention*** permettent de mieux cerner la structure des équivalences traductionnelles. L'accès à la catégorie morphosyntaxique de chaque mot permet de réaliser de nouveaux types d'exploration topographique du bi-texte.

[6] Les listes complètes des patrons syntaxiques en français et en anglais calculés à partir du corpus ***Convention*** sont disponibles sur le Cd-rom.

F	Patron syntaxique français	Patron syntaxique anglais	F
31887	Déterminant + Nom	Determiner+ Noun	24273
22321	Nom + Préposition	Noun + Preposition	21683
19650	Préposition + Déterminant	Preposition + Determiner	21019
16313	Préposition + Déterminant + Nom	Preposition + Determiner + Noun	12752
10704	Déterminant + Nom + Préposition	Adjective + Noun	10955
10504	Préposition + Nom	Noun + Preposition + Determiner	10122
10123	Nom + Adjectif	Determiner + Noun + Preposition	10072
9211	Nom + Préposition + Déterminant	Preposition + Noun	7879
9159	Préposition contractée + Nom	Determiner + Adjective	6626
7356	Nom + Préposition + Déterminant + Nom	Determiner +Proper Noun	6452
6614	Nom + Préposition + Nom	Noun + Preposition + Determiner Noun	6125
5383	Nom + Préposition contractée	Determiner + Adjective + Noun	5599
5323	Déterminant + Nom + Adjectif	Preposition + Determiner + Noun + Preposition	5076
5287	Verbe au participe passé + Préposition	Verb past participle + Preposition	4993
5246	Préposition + Déterminant + Nom Préposition	Determiner + Noun + Preposition + Determiner	4759

Figure 7.1 :
Les quinze patrons syntaxiques les plus employés dans les volets français et anglais du corpus ***Convention***

Nous avons montré au chapitre 4 que les fonctionnalités du logiciel ***Lexico3*** permettent de colorier la carte des sections parallèles d'un corpus bilingue en fonction de la présence/absence d'une unité textuelle dans un volet du corpus, puis la sélection *par résonance* des sections équivalentes dans l'autre volet. Généralement, l'unité textuelle la plus caractéristique de cette deuxième sélection est la traduction de l'unité de départ. Si l'on dispose d'un corpus bilingue catégorisé, on peut colorier la carte des sections en fonction de la présence/absence d'une catégorie ou d'un patron syntaxique dans un volet, repérer les sections équivalentes dans l'autre volet, puis calculer les catégories et les patrons particulièrement caractéristiques de cette deuxième sélection (cf. *Figure 7.2*).

Exploration topographique de corpus parallèles catégorisés

Étape 1 : On repère dans l'un des volets du corpus bilingue des sections (paragraphes, phrases, etc.) dans lesquels les occurrences d'un *patron syntaxique* dépasse un seuil fixé ;

Étape 2 : Le fragment sélectionné dans le premier volet provoque la sélection par *résonance* des sections correspondantes dans l'autre volet du corpus [7]. Les patrons syntaxiques particulièrement fréquents dans cette seconde sélection laissent apercevoir les éléments des structures formelles de l'autre langue qui servent à maintenir l'équivalence traductionnelle [8].

Étape 3 : Lorsque la correspondance de patrons syntaxiques est établie, on procède à l'appariement des syntagmes correspondants à ces patrons dans les phrases où se réalisent les appariements.

Figure 7.2 : Approche topographique du bi-texte catégorisé

[7] Ce type de sélection peut être réalisé à l'aide de la *carte des sections parallèles* présentée au chapitre 4. Voir aussi la description sommaire du Cd-rom en annexes.

[8] Certains patrons syntaxiques apparaissent conjointement dans les phrases en correspondance de traduction. Par exemple, les noms sont généralement en cooccurrence avec les verbes. Les relations de cooccurrences entre les catégories morphosyntaxiques compliquent le repérage automatique des correspondances entre les patrons syntaxiques particulièrement productifs. Pour contourner ce problème, on peut envisager deux solutions : 1) augmenter les *seuils de probabilité* pour amorcer le processus de résonance ; 2) réaliser le filtrage des résultats mis en évidence par l'exploration topographique par comparaison des *fréquences totales* des patrons bilingues sélectionnés par notre méthode.

Pour le corpus ***Convention***, cette approche nous a permis de poser automatiquement l'équivalence des patrons français/anglais suivants :

→ *Verbe au participe passé* + *Préposition* (F = 12 059)
→ *Verb past participle* + *Preposition* (F = 10 181).

Comme le montrent les extraits des *inventaires distributionnels des segments répétés*[9] sur la figure 7.3, ces catégories ont des voisinages similaires dans les deux volets bilingues du corpus ***Convention***. La comparaison des fréquences globales des patrons bilingues facilite le repérage des liens de correspondance entre eux. Lorsque l'équivalence des patrons morphosyntaxique est posée, il est possible de procéder à l'appariement et l'extraction de syntagmes qui se correspondent mutuellement dans le corpus.

Par exemple, la correspondance des patrons représentés ci-dessous :

→ *Verbe au participe passé* + *Préposition* + *Déterminant* + *Nom* (F = 2 228)
→ *Verb past participle* + *Preposition* + *Determiner* + *Noun* (F = 1 559)

permet de procéder à l'appariement des syntagmes suivants (cf. *Figure 7.2*) :

français	**anglais**
accusée d'une infraction	charged with a criminal offence
désignés dans la notification	named in the notification
énoncés dans la déclaration	stated in the declaration
exclut de la compétence	excluded from the jurisdiction
envoyée par la requérante	published by the applicant
dressée par le bureau	drawn up by the bureau
calqué sur le modèle	fashioned after the model
protégés par la convention	protected under the convention

[9] Sur la description des *inventaires distributionnels* des segments répétés, on consultera *infra* la section 3.2.3.

Volet français

4 964 ---- ---- ---- ---- ---- ***Verbe au participe passé + Préposition***

dont 2 668 ---- ---- ---- ---- ***Verbe au participe passé + Préposition*** + *Déterminant*

dont 2 228 ---- ---- ---- ***Verbe au participe passé + Préposition*** + *Déterminant* + *Nom*

Volet anglais

4 989 ---- ---- ---- ---- ---- ***Verb past participle + Preposition***

dont 2 529 ---- ---- ---- ---- ***Verb past participle + Preposition*** + *Determiner*

dont 1 559 ---- ---- ---- ***Verb past participle + Preposition*** + *Determiner* + *Noun*

Figure 7.3 :
Les extraits des inventaires distributionnels des expansions récurrentes après les patrons équivalents français/anglais
Verbe au participe passé + Préposition / *Verb past participle + Preposition*

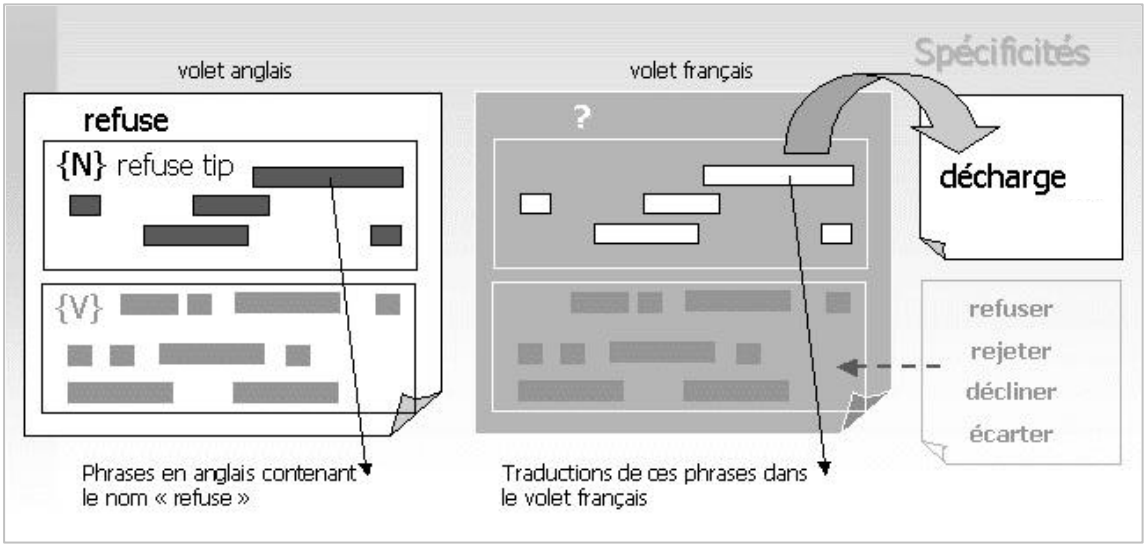


Figure 7.4 : L'approche topographique du bi-texte catégorisé

Guide de lecture : On repère dans le volet anglais du corpus bilingue ***Convention*** les phrases dans lesquelles sont attestées les occurrences du nom *refuse*. Le fragment textuel sélectionné dans le volet anglais provoque la sélection par *résonance* des phrases correspondantes dans le volet anglais du corpus. Le nom *décharge* particulièrement fréquent dans cette seconde sélection est mis en évidence par le calcul des spécificités. Une exploration similaire est réalisée pour le verbe *refuse*, elle laisse apparaître les verbes français liés à cette forme par une équivalence de traduction : *refuser*, *rejeter*, *décliner*, *écarter*, etc.

Nous avons montré que la catégorisation de corpus parallèles rend possible la mise en évidence des liens de correspondance entre les patrons syntaxiques bilingues. L'accès à la catégorie permet également de lever des ambiguïtés touchant au statut morphosyntaxique de chaque forme. Ainsi, pour rechercher les correspondances traductionnelles de la forme anglaise *refuse* dans le volet français du corpus ***Convention***, il devient possible d'analyser séparément les contextes correspondant à *refuse*{Nom} et ceux de *refuse*{Verbe} (cf. *Figure 7.4*).

A l'aide de la carte des sections parallèles (cf. *Chapitre 4*), on repère dans le volet anglais du corpus de travail les seules sections (phrases, paragraphes) dans lesquelles sont attestées les occurrences du nom anglais *refuse.* Le fragment textuel sélectionné dans le volet anglais provoque la sélection par *résonance* des sections correspondantes dans le volet français du corpus. La forme française

décharge particulièrement fréquente dans cette seconde sélection permet de repérer l'équivalence traductionnelle *refuse tip / décharge* (cf. *Figure 7.4*).

Ces mêmes étapes sont répétées pour les sections dans lesquelles sont attestées les occurrences du verbe *refuse* (cf. *Figure 7.4*). Cette deuxième exploration laisse apparaître une série de verbes français équivalents à *refuse* : *refuser*, *rejeter*, *décliner*, *écarter* l'équivalence *refuse / refuser* étant la plus caractéristique de ce fragment bi-textuel.

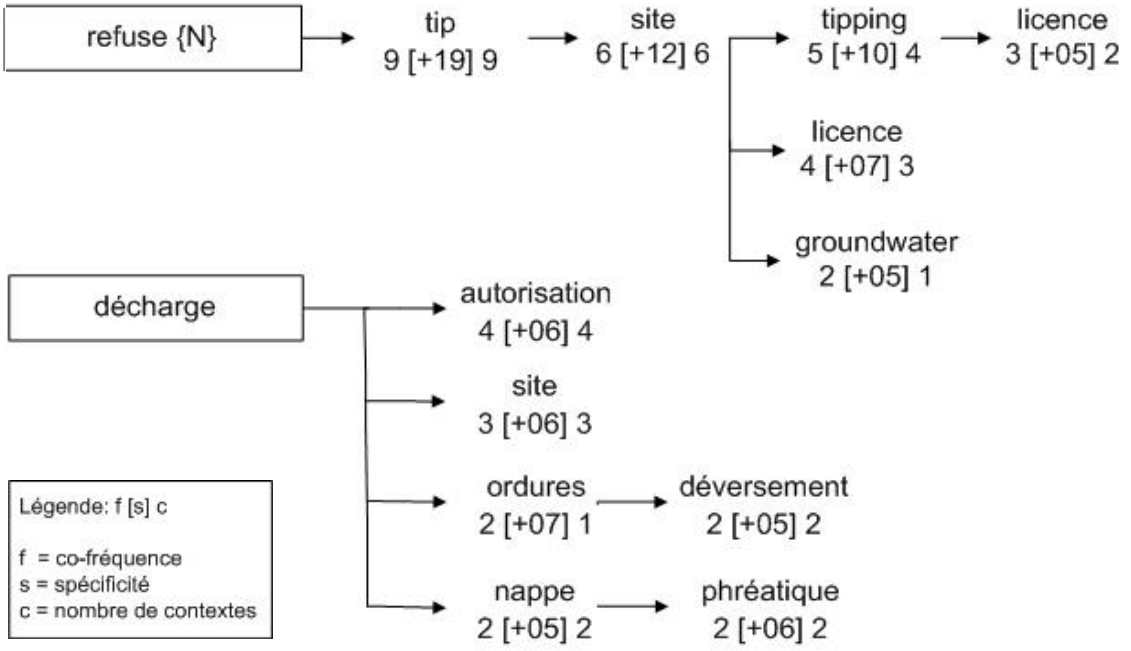
Dans le corpus ***Convention***, le nom *refuse* n'évolue pas dans le même univers sémantique que le verbe homographe. La catégorisation ouvre de nouvelles voies pour l'exploration des réseaux de cooccurrences qui s'élaborent à partir de formes-pôles désambiguïsées par l'étiquetage morphosyntaxique (cf. *Figures 7.5-6*). Nous avons montré au chapitre 6 qu'il est possible de réaliser le calcul des réseaux des cooccurrences séparément pour chaque sous-ensemble d'occurrences de la forme *fonctionnaires* ayant reçu des traductions différentes dans le volet anglais du corpus ***Convention*** [10].

De la même façon, on peut procéder au calcul des réseaux de cooccurrences dans le volet anglais du corpus à partir des occurrences du nom *refuse*, puis à partir celles du verbe ayant la même forme graphique. Dans le volet français, l'exploration sera réalisée symétriquement à partir de la forme *décharge* correspondant au nom *refuse*, puis à partir du verbe *refuser* utilisé le plus souvent pour traduire le verbe anglais *refuse* (cf. *Figure 7.4*).

Les schémas sur les figures 7.5-6 montrent que les réseaux bi-textuels mis en évidence par ces deux séries d'explorations sont distincts dans chacun des volets et se correspondent au plan traductionnel. Par exemple, l'analyse des univers sémantiques des correspondances traductionnelles *refuse tip* / *décharge* permet de poser l'équivalence de leurs cooccurrents, les formes *licence* et *autorisation* (anglais / français). De la même façon, les correspondances traductionnelles *refuse* / *refuser* sont co-occurrentes avec les formes équivalentes (anglais / français) *hearing* et *débats* (cf. *Figures 7.5-6*).

[10] Rappelons qu'une telle sélection est possible grâce à l'approche topographique du bi-texte (cf. *Chapitre 4*).

Pôle N°	forme (ANG)	traduit par (FRA)	*NbUC*[11]
1	*refuse* **{Nom}**	*décharge*	12



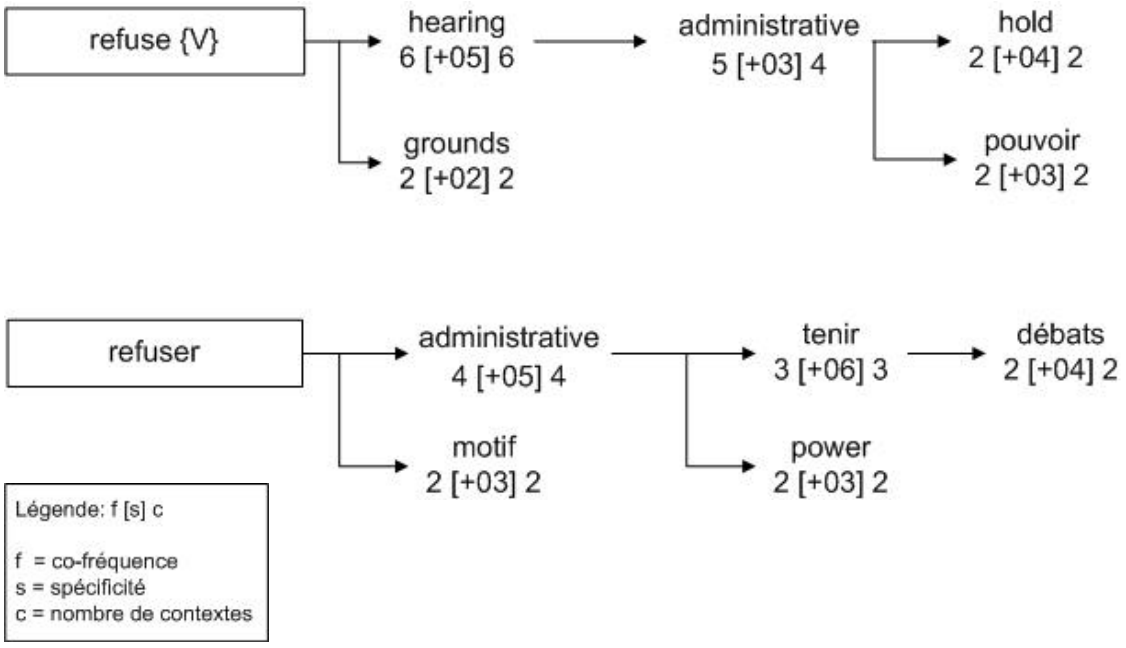
Retour au contexte :

volet français	volet anglais
dans la décision objet de l'appel , le **site** de la **décharge** en tant que tel était pareillement qualifié de problématique vu les expertises : l'avis d'expert reproduit dans la décision entreprise déclare qu'il est hors de question d'utiliser comme **site** de **décharge** des terrains où existent des sources d'eaux souterraines pouvant, par leur quantité et leur qualité, servir à la consommation d'eau potable ; il indique aussi que le **déversement** de déchets mettant en danger la **nappe phréatique** s'expliquerait, notamment, par le 'libellé imprécis de l'**autorisation**' accordée en 1973.	in the decision appealed against, the **site** of the **tip** as such was likewise described as problematical on the basis of the experts' reports - in the expert opinion that is reproduced in the impugned decision it is stated that it was out of the question that areas where there were sources of **groundwater** suitable in quantity and quality for use as a water supply should be used for **refuse tips** - and the **tipping** of waste that endangered the **groundwater** was attributed to among other things the 'imprecise wording of the **licence**' of 1973 .

Figure 7.5 : Réseaux de cooccurrences élaborés à partir du pôle anglais / français *refuse*{N} / *décharge*

[11] ***NbUC*** : Nombre d'unités de contextes où est attesté le pôle bilingue.

Pôle N°	forme (ANG)	traduit par (FRA)	*NbUC*
2	*refuse* **{Verbe}**	*refuser*	9



Retour au contexte :

volet français	volet anglais
le **motif** ajouté en 1982 permet dorénavant à la cour **administrative**, après examen des mémoires et d'autres documents versés au dossier, de **refuser** de **tenir** des **débats** pour des raisons touchant au fond de l'affaire, dans des cas où l'appel pourrait être rejeté.	the ground added in 1982 made it possible for the first time for the **administrative** court, after considering the written pleadings and other documents in the file, to **refuse** an oral **hearing** on **grounds** pertaining to the merits of the case, in instances where the appeal fell to be dismissed.
comme le souligne à juste titre la commission l'introduction du paragraphe (2) (6) a de fait considérablement élargi le **pouvoir** de la cour **administrative** de **refuser** de **tenir** des **débats**.	as the commission rightly pointed out, the introduction of subsection (2) (6) in effect considerably extended the **administrative** court's **power** to **refuse** to **hold** a public **hearing**.

Figure 7.6 : Réseaux de cooccurrences élaborés à partir du pôle anglais / français *refuse* {V} / *refuser*

Les premières explorations réalisées à partir de la version catégorisée du corpus ***Convention*** nous ont convaincue que l'étiquetage morphosyntaxique offrent des indices précieux pour l'extraction de ressources traductionnelles du bi-texte. Les stratégies d'analyse textométrique de corpus multilingues catégorisés seront sans doute affinées dans les années à venir.

7.2 Le stockage informatique des correspondances bi-textuelles

7.2.1 Le balisage XML

La constitution d'un ensemble d'équivalences traductionnelles à partir de corpus parallèles suppose le développement d'un système de stockage informatique des correspondances repérées. La conception de structures de données pour l'exploration bi-textuelle est un enjeu particulièrement important pour le traitement automatique de corpus multilingues.

Actuellement, le langage de balisage XML (eXtensible Markup Language) est de plus en plus utilisé pour le stockage informatique des données textuelles multilingues. Plusieurs projets de recherche en cours visent à produire des balisages à base XML pour représenter l'information contenue dans les corpus de traduction[12]. Cette représentation a pour objectif d'augmenter les possibilités de traitement automatique de ces corpus et de faciliter l'extraction de leurs ressources traductionnelles[13].

---

[12] On consultera sur ces questions les articles de Melby [2000], Romary et Bonhomme [2000], Sharoff [2002], ainsi que les sites suivants :
TEI - Text Encoding Initiative : http://www.mutu-xml.org/xml-base/shared/KEY-TEI.html
LISA - TMX (Translation Memory eXchange) standard : http://www.lisa.org/tmx/
OASIS XLIFF (XML Localisation Interchange File Format) : http://www.oasis-open.org/

[13] Pour plus d'informations sur le langage XML ainsi que sur les outils de manipulation de corpus balisés, on consultera, par exemple, la présentation de Serge Fleury sur le site TAL de l'Université de la Sorbonne nouvelle – Paris 3 :
http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/parcours/slides/slidesxml/

La *figure 7.4* montre un extrait du corpus ***Convention*** catégorisé au format XML[14]. Nous avons représenté sur la *figure 7.5* le schéma de la structure formelle du corpus balisé. Sur le schéma, chaque alignement phrastique est référencé par un élément ***P*** (partie) de l'arbre XML. Cet élément contient deux sous-éléments : la phrase en français et sa traduction en anglais ; ces mêmes sous-éléments contiennent aussi la version étiquetée de la partie visée : *forme-catégorie-lemme*.

Si l'alignement phrastique se prête assez facilement à l'encodage au format XML, le balisage des correspondances lexicales est beaucoup plus complexe. Nous avons montré au chapitre 2 qu'il est difficile de produire une description formelle des rapports d'équivalence qui s'établissent entre les mots de textes source et cible.

La *figure 7.6* décrit grossièrement la structure des correspondances traductionnelles du *bi-texte*. Les correspondances croisées, les équivalences lexicales à composants discontinus, le chevauchement entre les équivalences posent de nombreux problèmes lorsque l'on cherche à décrire les données de traduction sous forme d'une structure arborescente[15]. Il nous semble que les systèmes d'*index* informatisés utilisés dans le traitement textométrique des données textuelles conviennent mieux à ce type de tâches. On peut envisager leur intégration dans les structures de données des futurs systèmes de mémoires de traduction.

[14] Nous remercions Serge Fleury (Maître de conférences à l'Université de la Sorbonne Nouvelle – Paris 3) qui nous a aidé à préparer la version XML du corpus *Convention*. L'ensemble du corpus au format XML et sa DTD (Définition du type de document) peuvent être consultés sur le Cd-rom qui accompagne ce volume : /fichiersCD/stmz/page1.htm.

[15] Quelques solutions techniques à ces problèmes sont exposées dans les articles de Bird et Liberman [2001] ; Romary et Bonhomme [2000].

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<corpusconvention>
<head>
<corpus name="Convention"/>
<presentation>Le corpus Convention a été constitué à partir des
documents contenus dans la Convention européenne des Droits de
l'Homme (y copris les protocoles intégraux) , et d'une série d'arrêts
rendus par la Cour européenne des Droits de l'Homme de Strasbourg en
1995</presentation>
<volet langue="français">
 (296 396 occ.)
Convention européenne des Droits de l'Homme (5 953 occ.)
Protocoles intégraux de la Convention (8 984 occ.)
Protocole Additionnel
Protocoles 2 , 4, 6, 7, 9, 10, 11.
Arrêts de la Cour Européenne des Droits de l'Homme (281 459 occ.)
</volet>
<volet langue="anglais">
 (284 958 occ.)
Convention européenne des Droits de l'Homme (5 710 occ.)
Protocoles intégraux de la Convention (8 773 occ.)
Protocole Additionnel
Protocoles 2 , 4, 6, 7, 9, 10, 11.
Arrêts de la Cour Européenne des Droits de l'Homme (274 475 occ.)
</volet>
</head>
<textes>
<p>
<source type="conv_a0_p1-1">
<ref number="1"/>
<texte traduction="fr">Les gouvernements signataires , membres du
Conseil de l'Europe , </texte>
<texteandtag traduction="fr"><w><f>Les</f><c>DET:ART</c><l>le</l></w>
<w><f>gouvernements</f><c>NOM</c><l>gouvernement</l></w>
<w><f>signataires</f><c>NOM</c><l>signataire</l></w><w><f>,</f>
<c>PUN</c><l>,</l></w><w><f>membres</f><c>NOM</c><l>membre</l></w>
<w><f>du</f><c>PRP:det</c><l>du</l></w><w><f>Conseil</f><c>NOM</c>
<l>conseil</l></w><w><f>de</f><c>PRP</c><l>de</l></w><w><f>l'</f>
<c>DET:ART</c><l>le</l></w><w><f>Europe</f><c>NAM</c><l>Europe</l>
</w><w><f>,</f><c>PUN</c><l>,</l></w></texteandtag>
</source>
<source type="conv_a0_p1-1e">
<ref number="2"/>
<texte traduction="en">The governments signatory hereto , being
members of the Council of Europe , </texte>
<texteandtag traduction="en"><w><f>The</f><c>DT</c><l>the</l></w><w>
<f>governments</f><c>NNS</c><l>government</l></w><w><f>signatory</f>
<c>NN</c><l>signatory</l></w><w><f>hereto</f><c>RB</c><l>hereto</l>
</w><w><f>,</f><c>,</c><l>,</l></w><w><f>being</f><c>JJ</c>
<l>being</l></w><w><f>members</f><c>NNS</c><l>member</l></w><w>
<f>of</f><c>IN</c><l>of</l></w><w><f>the</f><c>DT</c><l>the</l></w>
<w><f>Council</f><c>NP</c><l>Council</l></w><w><f>of</f><c>IN</c>
<l>of</l></w><w><f>Europe</f><c>NP</c><l>Europe</l></w><w><f>,</f>
<c>,</c><l>,</l></w></texteandtag>
</source>
</p>...</textes>
</corpusconvention>
```

Figure 7.7 : Le balisage XML du corpus ***Convention***

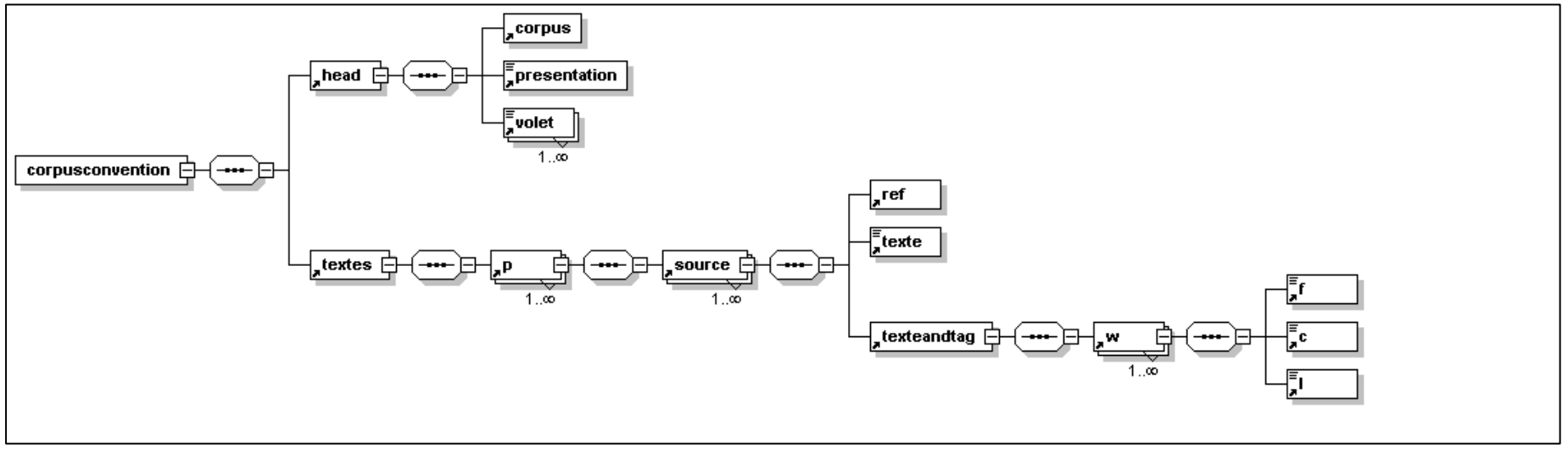


Figure 7.8 : L'arbre associé au corpus ***Convention*** balisé au format XML

Guide de lecture : Le corpus est représenté par une arborescence logique. Comme le montre la *figure 7.4*, les composants du document sont identifiés par le biais de balises.

Le corpus (*corpusconvention*) est constitué de l'en-tête (*head*) qui regroupe des renseignements généraux concernant les documents qui en font partie et du corps (*textes*) correspondant au texte de ces documents. Le texte du corpus est constitué d'éléments *p* qui représentent chacun un alignement phrastique. Chaque élément *p* comprend des sous-éléments (*sources*) qui correspondent aux phrases alignées en français et en anglais. Chacune des phrases a un identifiant unique relatif au type de document du corpus dont elle fait partie (protocole, arrêt, etc.). Elle est représentée par sa version originale au format texte brut (*texte*) et sa version catégorisée (*textandtag*) où chaque mot (*w*) reçoit une étiquette relative à sa forme graphique (*f*), sa catégorie(*c*) et son lemme(*l*).

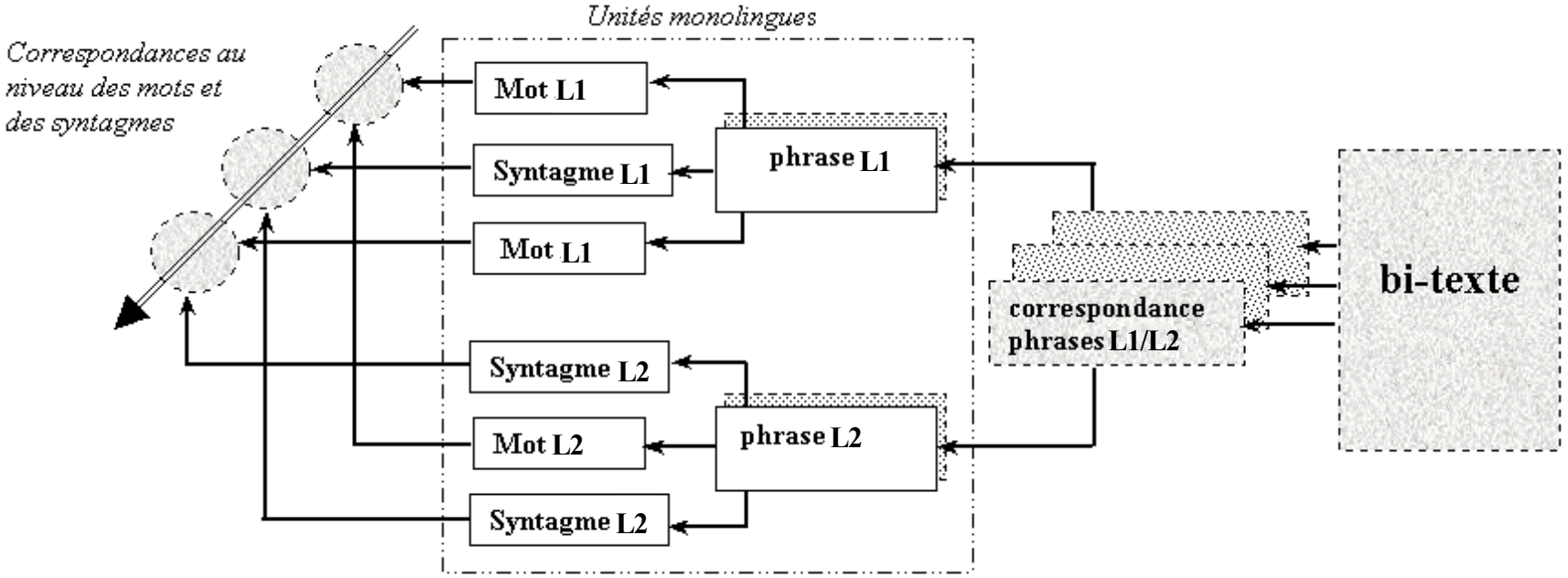


Figure 7.9 : La structure des correspondances traductionnelles du *bi-texte*

7.2.2 Les index bi-textuels

En textométrie, les *index* permettent de repérer immédiatement, pour chacune des unités textuelles, tous les endroits du corpus où sont situées ses occurrences. L'index d'un corpus de textes monolingues se base sur les résultats de la *numérisation textométrique* de la séquence textuelle segmentée en occurrences de formes graphiques. Rappelons que pendant l'étape de calcul, la technique de numérisation permet de faire abstraction du graphisme des formes décomptées pour ne retenir qu'un numéro d'ordre qui sera associé à toutes les occurrences d'une forme donnée dans le corpus. Un *dictionnaire* permet de stocker les formes et les numéros qui y sont associer afin de pouvoir reconstituer au besoin leur graphisme. Les index monolingues servent à repérer immédiatement pour chacune des formes, les endroits du corpus où sont localisées ses occurrences. Cette réorganisation de la séquence textuelle permet de retourner plus facilement au document d'origine grâce à un système d'*adresses* (ou coordonnées numériques) comme le tome, la page, la ligne, la position de l'occurrence dans la ligne, etc.

Pour localiser automatiquement les équivalences lexicales dans un corpus parallèle, on peut utiliser un système d'adresses similaire à celui décrit plus haut que l'on appellera *index bi-textuel*. Comme le montrent la *figure 7.7*, l'*index bi-textuel* a une structure parallèle associant deux index monolingues (un pour chaque volet du corpus). Il permet de regrouper les références de formes et polyformes des deux volets bilingues, mises en correspondance au cours de l'appariement. L'index bi-textuel représente un système de coordonnées numériques liées à l'édition de textes source et cible qui permet la localisation dans les deux volets du corpus de chaque composant d'une équivalence traductionnelle.

Dans ce qui suit, nous appellerons ***Equivalence Traductionnelle Elémentaire* (ETE)**, le *type* d'unité bi-textuelle dont on peut recenser les occurrences dans un corpus parallèle. Sur le plan sémantique, il s'agit de l'unité d'équivalence attestée au niveau des mots et des syntagmes. Lamalle et Salem [2002, p. 404] définissent le concept de *type* (ou *type généralisé* ***Tgen***) comme « *ensemble d'occurrences*

sélectionnées parmi les occurrences du texte ». Dans le cas des unités type *ETE*, la sélection des occurrences est effectuée parallèlement dans les deux volets du bi-texte.

Le marquage des *ETE(s)* dans un corpus parallèle s'appuie sur les résultats de la segmentation automatique de chaque volet en occurrences (cf. *Figure 7.7*). La segmentation est suivie de l'identification de différents types d'unités monolingues (*formes*, *lemmes*, *segments répétés*, *cooccurrences* etc.) puis de la phase de l'appariement automatique des unités mises en relation.

Par exemple, la correspondance des segments répétés d*émocratie apte à se défendre* (F=4) / *democracy capable of defending itself* (F=4) peut être considérée comme une *ETE*. Cette *ETE* compte quatre occurrences dans le corpus. Notons que les fréquences globales des types appariés au sein d'une *ETE* ne sont pas forcement identiques, surtout lorsqu'il s'agit des unités à correspondances multiples. Ainsi, comme nous l'avons montré au chapitre 4 (*section 4.2.2*), la forme française *fonctionnaires* (F=49) reçoit plusieurs traductions dans le volet anglais du corpus ***Convention*** : *officers* (F=38), *officials* (F=16) et *servants* (F=50), *civil servants* (F=46) et *civil service* (F=58). Pour cette unité, nous allons enregistrer cinq *ETE(s) différentes* :

fonctionnaires / civil servants	(29 occurrences)
fonctionnaires / civil service	(1 occurrences)
fonctionnaires / servants	(2 occurrences)
fonctionnaires / officials	(7 occurrences)
fonctionnaires / officers	(10 occurrences)

Une *ETE* correspond ainsi à une liste d'adresses d'un sous-ensemble d'occurrences des types bilingues appariés. Les adresses sont des coordonnées numériques qui indiquent pour chaque élément monolingue de l'équivalence sa position dans la section, paragraphe, phrase, ligne, etc. du volet. Ces renseignements qui permettent de retourner aux documents bilingues d'origine sont les *références* associées à chacune des *ETE(s)* (cf. *Figure 7.7*).

Lorsque la correspondance de types bilingues est posée, il devient possible de répertorier systématiquement dans le corpus toutes les occurrences de cette unité de traduction dans les phrases en correspondance. Nous appellerons ainsi ***Occurrence Traductionnelle Elémentaire*** **(OTE)** la présence simultanée dans un couple de phrases équivalentes des unités textuelles appariées au sein d'une *ETE*[16].

La structure de l'index bi-textuel offre beaucoup de souplesse dans l'organisation du stockage informatique des équivalences car les occurrences des unités qui composent chaque *ETE* sont explicitement référencées grâce à leurs adresses dans le volet correspondant du corpus. Nous pouvons ainsi tenir compte du phénomène d'emboîtement des unités lexicales liées sur le plan de la traduction ainsi que des correspondances lexicales discontinues. L'exploration de liens entre les systèmes de stockages des données textuelles utilisées en textométrie et le balisage XML fera l'objet de recherches à venir.

[16] L'Occurrence Traductionnelle Elémentaire (OTE) correspond à un groupe d'occurrences en équivalence de traduction appartenant à chacun des volets du bi-texte. Il s'agit d'une occurrence d'une Equivalence Traductionnelle Elémentaire (ETE).

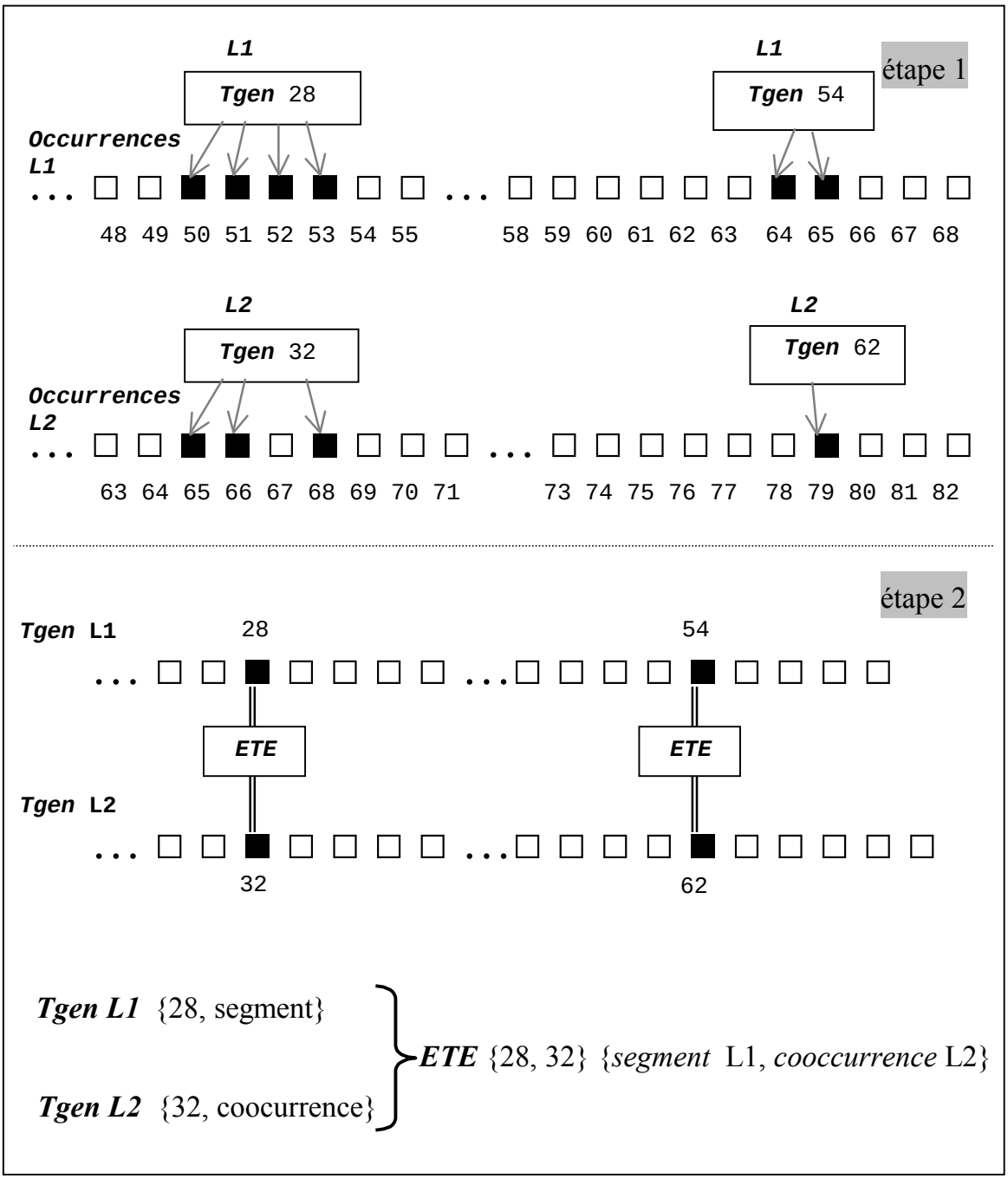


Figure 7.10 :
Les structures de données pour l'exploration textométrique bi-textuelle

Guide de lecture de la figure 7.7 :

Tgen = Type généralisé (forme, lemme, segment répété, cooccurrence, etc.)

ETE = Equivalence Traductionnelle Elémentaire

En textométrie, le traitement informatisé des données textuelles multilingues commence par la segmentation de chaque volet du bi-texte (***L1*** et ***L2***) en occurrences. Chacun des éléments découpé par un algorithme de segmentation automatique est identifié par son numéro d'ordre dans le volet.

Les occurrences identiques sont ensuite regroupées sous un même *type* (***Tgen***) référencé dans l'index (cf. **étape 1**). Les regroupements des unités identiques au sein des *Tgen(s)* peuvent correspondre à un *lemme*, à un *segment répété*, à une *cooccurrence*, etc.

Enfin, on peut constituer des unités textuelles bilingues ***ETE(s)*** par le regroupement des *Tgen(s)* liés sur le plan de la traduction dans les deux volets *L1* et *L2*. Pour chaque *ETE*, ce sont les coordonnées des *Tgen(s)* équivalents qui servent de références permettant sa localisation dans les deux volets du bi-texte (cf. **étape 2**).

Conclusions

Le bilan que nous tentons de dresser ne peut se faire qu'en liaison étroite avec l'état de développement actuel de la textométrie multilingue.

Du côté de l'*alignement de corpus parallèles*, l'efficacité pratique de techniques relativement simples (telles que la prise en compte des corrélations entre la longueur des segments en correspondance de traduction, par exemple) montre que le processus d'alignement au niveau de la phrase est désormais bien maîtrisé. Nous avons montré que certaines tâches d'alignement automatique peuvent être simplifiées par l'utilisation d'indices statistiques. Dans ce domaine, la textométrie peut être utile dans les phases de pré-traitement ou de post-traitement pour obtenir une plus grande précision, pour vérifier et rectifier les appariements fautifs.

L'évolution actuelle du domaine de l'alignement automatique peut conduire à séparer les problèmes de l'appariement des phrases et ceux qui concernent l'expérimentation de formalismes plus élaborés pour l'alignement lexical. Les travaux de recherches les plus récents laissent transparaître un désenchantement certain en ce qui concerne la conception d'outils d'alignement lexical intégral. La frontière entre ce qu'il est possible d'apparier et ce qui relève de compétences humaines qui ne peuvent pas être totalement formalisées apparaît plus clairement. Les pratiques d'*extraction de ressources traductionnelles à base de corpus parallèles* ou *comparables* occupent progressivement une place centrale dans les recherches sur le traitement automatique des corpus multilingues.

Au cours de nos recherches, nous avons montré que l'analyse textométrique offre une base précieuse pour l'acquisition de *ressources traductionnelles* à base de corpus. La textométrie permet d'accéder à de nouvelles dimensions d'analyse des

données de traduction entre des phénomènes langagiers relavant de différents niveaux de l'analyse linguistique.

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

L'extraction des ressources traductionnelles à partir de corpus multilingues sur des bases quantitatives ne peut être menée en dehors d'une réflexion sur les problèmes de la segmentation parallèle de textes source et cible. Nous avons été confrontée tout au long de cette recherche à l'impossibilité de proposer un modèle de segmentation parallèle de corpus multilingues qui constitue directement des listes de correspondances traductionnelles et fournisse des représentations sémantiques et syntaxiques indépendantes du corpus. La détermination des mécanismes formels permettant d'automatiser cette segmentation nous semble d'ailleurs une perspective peu envisageable pour des raisons qui tiennent aux différences de structures des langues à tous les niveaux de l'analyse linguistique.

Ce constat a orienté notre travail vers la recherche de nouvelles stratégies d'analyse exploratoire de corpus multilingues. A côté des systèmes d'extraction automatique de lexèmes bilingues – candidats à l'appariement – présentés sous forme de listes figées, il nous parait indispensable de fournir à l'utilisateur des outils de navigation dans les corpus multilingues, modulables en fonction de besoins particuliers. Dans cette perspective, nous avons largement utilisé la

cartographie d'équivalences traductionnelles, fondée sur la représentation *topographique* du *bi-texte*, pour l'analyse textométrique de corpus de traduction.

Pour mieux cerner les rapports de correspondances entre les lexèmes en fonction des variations contextuelles, nous avons fait appel à la notion de *résonance textuelle*. Le processus de *résonance textuelle* amorcé par la sélection dans le texte source des sections dans lesquelles les occurrences d'une unité textuelle (*forme*, *segment répété*, *patron morpho-syntaxique*) dépasse un seuil fixé, induit une sélection topographique correspondante dans le texte cible et met en évidence des séquences, liées à l'unité de départ, sur le plan de la traduction. Le processus de résonance textuelle peut être engagé par localisation *topographique* de fragments thématiques du *bi-texte*. Cette exploration topographique s'enrichit des résultats de l'alignement des deux volets bilingues du corpus au niveau de la phrase. Une description automatique des multiples relations d'équivalence entre unités bilingues peut être obtenue par le biais d'appariements statistiques lorsque l'exploration du corpus s'appuie sur un alignement des phrases. Cette approche peut être utilisée pour le repérage des équivalences lexicales y compris dans le cas où leurs fréquences dans le corpus sont peu élevées.

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

Dans le domaine de l'*ingénierie linguistique*, la croissance spectaculaire des données textuelles multilingues appelle l'attention sur l'importance de la création d'outils de traitement automatique de corpus dans des langues différentes. Dans ce contexte, le succès pratique des méthodes d'exploration que nous avons élaborées au fil de nos recherches, nous a incitée à produire des maquettes des logiciels d'exploration textométrique intertextuelle. Nous nous sommes attachée

à décrire des procédures et des objets informatiques nécessaires pour l'acquisition de ressources traductionnelles à base de corpus. Ces maquettes sont fournies sur le Cd-rom qui accompagne ce volume.

Le corps de méthodes développées dans ce travail nous semble pouvoir être étendu à l'analyse de *corpus multilingues comparables* dans des couples de langues variables. Il ne fait nul doute pour nous que la comparaison simultanée de traductions de textes écrits dans un grand nombre de langues continuera d'ouvrir aux linguistes de chantiers de recherches passionnants.

Liste des illustrations

Chapitre 1

Chapitre 2

Chapitre 3

Chapitre 4

Chapitre 5

Chapitre 6

Chapitre 7

Glossaire pour la textométrie multilingue

Note : Nous avons emprunté les définitions des termes du domaine de la statistique textuelle à Salem [1987] et Lebart et Salem [1994]. Les sites Web ayant contribué à la création du glossaire sont recensés dans la *CyberBibliographie* sur le Cd-rom qui accompagne cette thèse.

Dans le glossaire, les astérisques renvoient à une entrée de ce même glossaire. Les abréviations qui suivent entre parenthèses précisent le domaine auquel s'applique plus particulièrement la définition.

Abréviations :

afc	Analyse factorielle des correspondances
cla	Classification
sp	Méthode des Spécificités
sr	Analyse des segments répétés
ling	Linguistique
stat	Statistique
sa	Segmentation automatique
tal	Traitement Automatique des Langues
tml	Textométrie Multilingue
trad	Traductologie

accroissement spécifique – (*sp*) spécificité* calculée pour une partie d'un corpus* par rapport à une partie antérieure.

algorithme – ensemble des règles opératoires propres à un calcul.

alignement – (*tal*, *tml*) processus qui consiste à aligner, c'est-à-dire à poser comme équivalents des unités textuelles qui se correspondent au sein des corpus de textes bilingues ou multilingues. Les unités en correspondance se positionnent à plusieurs niveaux : mots*, syntagmes*, phrases*, paragraphes*, etc.

alignement automatique – (*tal*, *tml*) ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à l'appariement des unités textuelles en correspondance de traduction.

analyse des correspondances – (*stat*) méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou χ^2).

analyse factorielle – (*stat*) famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des « facteurs » résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

bi-texte – (*tal*, *tml*) corpus de traduction réunissant un texte original en langue source* et sa traduction en langue cible* appuyé sur un système de mise en correspondance interactive entre les segments équivalents des deux volets bilingues.

caractère – (*sa*) signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

caractères délimiteurs / non-délimiteurs – (*sa*) distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

carte des sections – (*tml*) visualisation du corpus découpé en sections (phrases, paragraphes, etc.) réalisée par la promotion d'un (ou de plusieurs) caractère particulier (paragraphe, point, etc.) au statut de délimiteur* de section.

cible – (*trad*) qui a trait à la langue de destination de l'opération de traduction (par opposition à *source*).

classification – (*stat*) technique statistique permettant de regrouper des individus ou observations entre lesquels a été définie une distance.

classification hiérarchique – (*cla*) technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.

compositionnalité de traduction – (*trad*) principe d'analyse selon lequel l'équivalence traductionnelle globale est décomposable en une hiérarchie structurée de correspondances traductionnelles*. La décomposition commence par constituer des unités larges (sections, paragraphes, etc.) pour descendre ensuite vers des unités de plus en plus petites (syntagmes, mots, morphèmes).

concordance – (*sa*) ensemble de lignes de contexte se rapportant à une même forme-pôle.

concordance bilingue – (ou concordance parallèle) liste ordonnée de toutes les occurrences d'un terme et de ses traductions dans une paire de textes source et cible, à l'intérieur de laquelle chaque occurrence est accompagnée de son contexte dans chacun des volets du corpus.

contribution absolue (ou contribution) – (*ac*) contribution apportée par un élément au facteur. Pour un facteur donné, la somme des contributions sur les éléments de chacun des ensembles mis en correspondance est égale à 100.

contribution relative (ou cosinus carré) – (*ac*) contribution apportée par le facteur à un élément. Pour un élément donné, la somme des contributions relatives sur l'ensemble des facteurs est égale à 1.

cooccurrence – (*sa*) présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux formes données.

cooccurrences multiples – (*sa*) associations lexicales complexes qui lient une forme pôle à plusieurs formes cooccurrentes.

corpus – (*ling*) ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

– (*textométrie*) ensemble de textes réunis à des fins de comparaison ; servant de base à une étude quantitative.

corpus comparables – (*tal*) ensemble de textes entre lesquels on peut établir des correspondances partie à partie et dont les corpus parallèles constituent un cas particulier.

corpus parallèles – (*tal*) corpus comportant plusieurs volets en relation de traduction qui correspondent chacun à une version d'un même texte dans deux ou plusieurs langues différentes.

correspondance traductionnelle – (*trad*, *tml*) ensemble d'unités textuelles en relation d'équivalence traductionnelle*. On distingue des correspondances traductionnelles quasi-univoques* et des correspondances multiples*.

correspondances traductionnelles multiples – (*tml, trad*) ensemble d'unités lexicales en relation d'équivalence traductionnelle dans lequel chaque segment source* possède plusieurs équivalents dans le texte cible* et réciproquement.

correspondances traductionnelles quasi-univoques – (*tml*, *trad*) couple d'unités lexicales en relation d'équivalence traductionnelle dans lequel chaque segment source* possède presque toujours un seul équivalent dans le texte cible* et réciproquement.

délimiteurs de séquence – (*sa*) sous-ensemble des caractères délimiteurs* de forme* correspondant aux ponctuations faibles et fortes (en général – le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses).

On distingue parmi les caractères délimiteurs :

- les caractères délimiteurs d'occurrence (encore appelés « délimiteurs de forme ») qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.
- les caractères délimiteurs de séquence : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.
- les caractères séparateurs de phrase : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

dendrogramme – (*cla*) représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.

distance du chi-2 – distance entre profils* de fréquence utilisée en analyse des correspondances* et dans certains algorithmes* de classification*.

éditions de contextes – (*sa*) éditions de type concordanciel dans lesquelles les occurrences d'une forme sont accompagnées d'un fragment de contexte pouvant contenir plusieurs lignes de texte autour de la forme-pôle. La longueur de ce contexte est définie en nombre d'occurrences avant et après chaque occurrence de la forme-pôle.

éléments d'un segment – (*sr*) chacune des formes correspondant aux occurrences qui entrent dans sa composition. ex : A, B, C sont respectivement les premier, deuxième et troisième éléments du segment ABC.

équivalence contextuelle – (*trad*) couple d'unités lexicales en relation d'équivalence traductionnelle dans un corpus particulier. Sur le plan sémantique, ce type d'équivalences singulières ne fournit pas de correspondances lexicales « stables », susceptibles d'être utilisées avec profit dans d'autres contextes.

équivalence traductionnelle – (*trad*) ensemble de rapports d'équivalence qui s'établissent entre les segments de textes source* et cible*. L'équivalence sémantique globale est fondée sur le principe de compositionnalité de traduction*.

équivalence traductionnelle élémentaire (ETE) – (*tml*) *type* bi-textuel composé de deux types appartenant chacun à l'un des volets du corpus bilingue et entretenant des rapports de traduction dans le corpus.

étiqueteur morpho-syntaxique – (*tal*) outil permettant d'associer des informations morphologiques et syntaxiques aux occurrences d'un corpus de textes.

expansion contrainte – (*sr*) terme dont les occurrences constituent chaque fois les expansions (du même côté) d'un même terme ayant plusieurs occurrences dans le corpus.

expansion d'un segment – (*sr*) segment situé immédiatement avant (expansion gauche) ou après (expansion droite) d'un segment donné, non séparé de ce segment par un délimiteur de séquence.

expansion récurrente d'un terme – (*sr*) terme dont les occurrences constituent plusieurs fois l'expansion des occurrences d'un terme donné.

expressions régulières – (*tal*, *tml*) ensemble d'opérateurs (méta-caractères) offrant la possibilité de représenter des portions de texte de manière générique.

facteur – (*ac*) variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.

forme – (*sa*) ou « forme graphique » archétype correspondant aux occurrences* identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

forme banale – (*sp*) pour une partie du corpus donnée, forme ne présentant aucune spécificité (ni positive ni négative) dans cette partie.

forme caractéristique – (*sp*) (d'une partie) synonyme de spécificité positive*.

forme commune – (*sp*) forme attestée dans chacune des parties du corpus.

forme originale – (*sp*) (pour une partie du corpus) forme trouvant toutes ses occurrences dans cette seule partie.

fréquence – (*sa*) (d'une unité textuelle) le nombre de ses occurrences dans le corpus*.

fréquence d'un segment – (*sr*) (ou d'une polyforme) le nombre des occurrences de ce segment, dans l'ensemble du corpus.

fréquence maximale – (*sa*) fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition « de »).

fréquence relative – (*sa*) la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

gamme des fréquences – (sa) suite notée V_k, des effectifs correspondant aux formes de fréquence *k*, lorsque *k* varie de 1 à la fréquence maximale.

hapax – gr. *hapax* (legomenon), « chose dite une seule fois », – (sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).

identification – (*stat*, *ling*, *sa*) reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.

index – (*sa*) liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regrouper les références* relatives à l'ensemble des occurrences d'une même forme.

index alphabétique (*sa*) – index* dans lequel les formes-pôles* sont classées selon l'ordre lexicographique* (celui des dictionnaires).

index hiérarchique (*sa*) – index* dans lequel les formes-pôles* sont classées selon l'ordre lexicométrique*.

index par parties – ensemble d'index (hiérarchiques ou alphabétiques) réalisés séparément pour chaque partie d'un corpus.

lemmatisation – regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

lexical – (*ling*) qui concerne le lexique* ou le vocabulaire*.

lexique – (*ling*) ensemble virtuel de mots d'une langue (auxquels peuvent être associées des informations linguistiques).

lexiques bilingues – (*ling*) ensemble de correspondances traductionnelles bi- ou pluri langue. Dans sa forme la plus simple, il s'agit d'une table d'équivalences 1:1. La complexité des langues et de leur usage impose souvent la création de bases de données lexicales plus élaborées permettant, par exemple, le traitement des synonymes, des abréviations, etc.

longueur – (*ling*) (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus (resp. partie, fragment, etc.). Synonyme de taille*.

On note : T la longueur du corpus ; t_j celle de la partie (ou tranche) numéro j du corpus.

longueur d'un segment – (*sr*) le nombre des occurrences entrant dans la composition de ce segment.

mémoire de traduction – (*trad, tal*) base de données permettant de proposer aux traducteurs des expressions candidates dans la langue cible à partir de traductions déjà effectuées dans des contextes similaires. Le recours à ce type d'aide débouche sur un concept plus large de poste de travail du traducteur* (PTT).

mots apparentés (ou cognats) – (*tal, trad*) mots étymologiquement reliés, suites de chiffres ou noms propres préservés dans la traduction.

navigation textométrique – (*tml*) utilisation de procédures informatiques qui permettent de se déplacer parmi les résultats produits par les différentes méthodes textométriques et le texte initial.

norme de dépouillement – (*sa*) normes d'indentification des unités textuelles. On distingue, par exemple :

- le dépouillement en formes graphiques (identification automatique des occurrences d'une même chaîne de caractères) ;
- le dépouillement en lemmes (regroupement sous une forme canonique à partir d'un dictionnaire) des occurrences du texte ;
- regroupement d'occurrences qui peuvent être rapportés à une même racine ou n-gramme.

numérisation textométrique – (*sa*) technique qui consiste à faire abstraction, pendant l'étape des calculs, du graphisme des formes décomptées pour ne retenir qu'un numéro d'ordre qui sera associé à toutes les occurrences de cette forme. Ces numéros stockés dans un dictionnaire de formes, propre à chaque exploitation, permettent de reconstituer le graphisme des formes du texte mises en évidence par les calculs statistiques.

occurrence – (*sa*) suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs* de forme.

occurrence traductionnelle élémentaire (OTE) – (*tml*) groupe d'occurrences en équivalence de traduction appartenant à chacun des volets du bi-texte. Occurrence d'une équivalence traductionnelle élémentaire* (ETE).

ordre lexicographique – (*sa*)

pour les formes graphiques :

- ordre selon lequel les formes sont classées dans un dictionnaire.

  Note : les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple, rangées dans cet ordre, les formes : *mais, maï s, maison, maître*.

pour les polyformes :

- ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante, les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, etc.

ordre lexicométrique – (*sa*)

pour les formes graphiques :

- ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes ; les formes de même fréquence sont classées par ordre lexicographique.

pour les polyformes :

- ordre résultant d'un tri par ordre de longueur décroissante des segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.

parallélisme textuel – (*ling*)

dans le contexte multilingue :

- rapports de correspondances entre les textes en relation d'équivalence traductionnelle* ou entre les textes qui sont proches dans leurs contenus (corpus comparables*).

dans le contexte monolingue :

- rapports de correspondances entre réécritures et paraphrases de textes susceptibles d'autoriser une comparaison.

partie – (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition – (*stat*, *tml*) (d'un corpus de textes) division d'un corpus en *parties** constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

patron syntaxique – (*tal*) suite de catégories morphosyntaxiques. Les patrons syntaxiques permettent de rapprocher des séquences dont le contenu lexical peut être variable.

phrase – (*sa*) fragment de texte compris entre deux séparateurs* de phrase.

points d'ancrage – (*tal*) couple d'unités textuelles dont l'appariement est considéré comme fiable.

polyforme – (*sr*) archétype des occurrences d'un segment ; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.

ponctuation – système de signes servant à indiquer les divisions d'un texte et à noter certains rapports syntaxiques et/ou conditions d'énonciation.

– (*sa*) caractère (ou suite de caractères) correspondant à un signe de ponctuation.

poste de travail du traducteur (PTT) – ensemble d'outils d'aide à la traduction, à la lecture, à la rédaction, à la gestion et à la transmission de textes multilingues comportant des fonctions de traitement de texte, vérification d'orthographe, dictionnaire électronique, bases de données terminologiques, mémoires de traduction*, etc.

précision – (*tal*) (mesure de) calcul statistique qui reflète la proportion d'appariements exacts dans l'ensemble des appariements réalisés automatiquement dans le bi-texte*. On suppose qu'il existe un corpus de référence aligné par un expert humain, et que c'est par rapport à cet alignement que sont évalués les appariements fournis par un système automatique.

profil – (*stat*, *ac*) (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).

quasi-segment répété – (*stat*, *ac*) série de répétitions lexicales légèrement altérées par des modifications de l'un de leurs composants.

rappel – (*tal*) (mesure de) calcul statistique qui reflète la proportion d'appariements de référence qui ont été trouvés dans l'ensemble d'appariements réalisés automatiquement dans le bi-texte*. On suppose qu'il existe un corpus de référence aligné par un expert humain, et que c'est par rapport à cet alignement que sont évalués les appariements fournis par un système automatique.

références – (*sa*) système de coordonnées numériques permettant de repérer dans le texte d'origine chacune des occurrences issues de la segmentation (ex : le tome, la page, la ligne, la position de l'occurrence dans la ligne) ou de situer rapidement cette occurrence parmi des catégories prédéfinies (auteur, année de parution, citation, mise en valeur, etc.).

répartition – (*sa*) (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.

résolution de l'alignement – (*tal*) nature des unités ayant servi à la segmentation bi-textuelle. Pour une résolution donnée, un alignement correct est maximum, s'il est composé du plus petit des couples de segment possible. Par exemple, un alignement qui met en correspondance des paragraphes peut être considéré comme un alignement de résolution moindre que celui qui montrera une correspondance au niveau des phrases.

résonance textuelle – (*tml*) processus amorcé par la sélection d'un sous-ensemble d'unités dans un des volets du bi-texte* laquelle induit une sélection correspondante dans l'autre volet.

ressources traductionnelles – ensemble de données linguistiques constituées à partir de traductions existantes présentées sous forme électronique que l'on peut utiliser pour produire de nouvelles traductions.

segment – (*sr*) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur* de séquence est un segment du texte.

segment répété – (*sr*) (ou polyforme répétée) suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentation – opération qui consiste à délimiter des unités minimales* dans un texte.

segmentation automatique – ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales*.

séparateurs de phrases – (*sa*) sous-ensemble des caractères délimiteurs* de séquence* correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).

séquence – (*sa*) suite d'occurrences du texte non séparées par un délimiteur* de séquence.

seuil – (*stat*) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

seuil d'absence spécifique – (*sp*) pour un seuil fixé, pour une partie du corpus fixée, fréquence *A*(*j*) pour laquelle toute forme de fréquence supérieure à *A*(*j*) dans le corpus et absente dans la partie *j* est spécifique (négative) pour la partie *j*.

seuil de présence spécifique – (*sp*) pour un seuil fixé, pour une partie du corpus fixée, fréquence B(*j*) pour laquelle toute forme de fréquence inférieure à A(*j*) dans le corpus et présente dans la partie *j* est spécifique (positive) pour la partie *j*.

source – (*trad*) qui a trait à la langue de départ de l'opération de traduction (par opposition à *cible*).

sous-fréquence – (*sa*) (d'une unité textuelle dans une partie, tranche, etc.) nombre des occurrences de cette unité dans la seule partie (resp. tranche, etc.) du corpus.

sous-segments – (*sr*) pour un segment donné, tous les segments de longueur inférieure et compris dans ce segment sont des sous-segments. ex : AB et BC sont deux sous-segments du segment ABC.

spécificité négative – (*sp*) pour un seuil de spécificité fixé, une forme *i* et une partie *j* données, la forme *i* est dite spécifique négative de la partie *j* si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

spécificité positive – (*sp*) pour un seuil de spécificité fixé, une forme *i* et une partie *j* données, la forme *i* est dite spécifique positive de la partie *j* (ou forme caractéristique* de cette partie) si sa sous-fréquence est « anormalement élevée » dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique* pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

stock distributionnel du vocabulaire – (*sp*) (d'un fragment de texte) le vocabulaire* de ce fragment assorti de comptages de fréquence pour chacune des formes entrant dans sa composition.

syntagmatique – (*sa*) qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.

syntagme – (*ling*) groupe de mots en séquence formant une unité à l'intérieur de la phrase.

tableau des segments répétés (TSR) – tableau à double entrée dont les lignes sont constituées par les ventilations* des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique* des segments. (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).

tableau lexical – tableau à double entrée résultant du TLE par suppression de certaines lignes (par exemple celles qui correspondent à des formes dont la fréquence est inférieure à un seuil donné).

tableau lexical entier (TLE) – tableau à double entrée dont les lignes sont constituées par les ventilations* des différentes formes dans les parties du corpus. Le terme générique *k(i,j)* du TLE est égal au nombre de fois que la forme *i* est attestée dans la partie *j* du corpus. Les lignes du TLE sont triées selon l'ordre lexicométrique* des formes correspondantes.

taille – (*sa*) (d'un corpus) sa longueur* mesurée en occurrences (de formes simples).

terme – (*sr*) nom générique s'appliquant à la fois aux formes* et aux polyformes*. Dans le premier cas on parlera de termes de longueur 1. Les polyformes sont des termes de longueur 2, 3, etc.

termes contraints / **termes libres** – (*sr*) un terme *S1* est contraint dans un autre terme *S2* de longueur supérieure si toutes ses occurrences* sont des sous-segments* de segments correspondant à des occurrences du segment *S2*. Si au contraire un terme possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, c'est un terme libre.

textométrie – (*tml*) ensemble des méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur les unités d'un corpus de textes.

topographie textuelle – représentation graphique des phénomènes mis en évidence par l'étude statistique qui permet d'apprécier leurs positions relatives dans le texte.

traduction assistée par ordinateur (TAO) – mise en œuvre d'un ensemble cohérent d'outils informatiques visant à assister le traducteur. Les plus importants de ces outils gèrent, d'une part, la terminologie spécifique du domaine de travail (dictionnaires, lexiques bilingues*, banques de termes, etc.) et, d'autre part, les mémoires de traduction*.

traduction automatique (TA) – opération réalisée au moyen de procédures informatiques, sans assistance d'un traducteur humain, qui consiste à analyser la structure de chaque énoncé, ou phrase, du texte à traduire (texte source*), décomposer cette structure en éléments aisément traduisibles, et recomposer un énoncé de même structure dans la langue cible*.

traduction semi-automatique – activité associant la traduction humaine et le recours à la machine comme un appui pour l'accomplissement des tâches périphériques.

type généralisé (TGen) – (*tml*) sous-ensemble d'occurrences du texte (segment répété*, quasi-segment répété*, cooccurrence*, ensemble de formes graphiques défini à l'aide des expressions régulières*, surlignage sélectif réalisé par un chercheur en fonction des besoins de l'étude, etc.).

unité de contexte – (*sa*) fragment de contexte pouvant contenir plusieurs lignes de texte autour de la forme-pôle. La longueur de ce contexte peut être définie en nombre d'occurrences avant et après chaque occurrence de la forme-pôle ou délimitée par les caractères séparateurs de phrase (sous-ensemble des délimiteurs* de séquence) qui correspondent, en général, aux seules ponctuations fortes.

unité de traduction (UT) – (*trad*) unité segmentale bi-textuelle composée d'un segment source* et d'un segment cible*.

unités minimales (pour un type de segmentation) – unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent).

variables actives – (*ac, cla*) variables utilisées pour dresser une typologie, soit par une analyse factorielle, soit par classification. Les typologies dépendent du choix et des poids de variables actives, qui doivent de ce fait constituer un ensemble homogène.

ventilation – (*sa*) (des occurrences d'une unité dans les parties du corpus) La suite des *n* nombres (*n* = nombre de parties du corpus) constituée par la succession des sous-fréquences* de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire – (*sa*) ensemble des formes* attestées dans un corpus de textes.

vocabulaire commun – (*sa*) l'ensemble des formes attestées dans chacune des parties du corpus.

vocabulaire de base – (*sp*) ensemble des formes du corpus ne présentant, pour un seuil fixé, aucune spécificité (négative ou positive) dans aucune des parties, (i.e. l'ensemble des formes qui sont « banales » pour chacune des parties du corpus).

vocabulaire original – (*sa*) (pour une partie du corpus) l'ensemble des formes* originales* pour cette partie.

voisinage d'une occurrence – (*sa*) pour une occurrence donnée du texte, tout segment (suite d'occurrences consécutives, non séparées par un délimiteur de séquence) contenant cette occurrence.

voisinages d'une forme – (*sa*) ensemble constitué par les voisinages de chacune des occurrences correspondant à la forme donnée.

Table des sigles et abréviations

AFC	**A**nalyse **F**actorielle des **C**orrespondances
CAH	**C**lassification **A**scendante **H**iérarchique
Convention	***Convention*** de Sauvegarde des Droits de l'Homme
DISTO	**DIST**ances à l'**O**rigine des axes (AFC)
DTD	**D**ocument **T**ype **D**efinition
ETE	**E**quivalence **T**raductionnelle **E**lémentaire
IVR	**I**nventaires des **V**oisinages **R**écurrents
JADT	**J**ournées d'**A**nalyse statistique des **D**onnées **T**extuelles
NbUC	**N**om**B**re d'**U**nités de **C**ontextes
OTE	**O**ccurrence **T**raductionnelle **E**lémentaire
PTT	**P**oste de **T**ravail du **T**raducteur
TA	**T**raduction **A**utomatique
TAL	**T**raitement **A**utomatique des **L**angues
TAO	**T**raduction **A**ssistée par **O**rdinateur
Tgen	**T**ype **gén**éralisé
TM	**T**ranslation **M**emory (Mémoire de traduction)
TMX	**T**ranslation **M**emory e**X**change standard (Standard d'échange pour Mémoires de traduction)
UT	**U**nité de **T**raduction
XML	e**X**tensible **M**arkup **L**anguage (Langage de balise étendu)

Index des auteurs

W

Y

Z

Index des matières

Bibliographie générale [1]

[1] AHRENBERG Lars, ANDERSSON Mikael, MERKEL Magnus (2000). "A knowledge-lite approach to word alignment." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 97-116.

[2] AIJMER Karin, ALTENBERG Bengt (1991). "Introduction." In Aijmer K., Altenberg B. (Eds.), *English Corpus Linguistics*: *Studies in Honour of Jan Svartvik.* London: Longman, pp. 1-6.

[3] BALLARD Michel (1993). « L'unité de traduction. Essai de redéfinition d'un concept. » In Ballard M. (Ed.), *La traduction à l'université. Recherches et propositions didactiques.* Lille : PUL, pp. 223-262.

[4] BAR-HILLEL Yehoshua (1960). "The present status of automatic translation of languages." *Advances in Computers* 1, pp. 91-163.

[5] BARKHUDAROV Leonid (1993). "The Problem of the Unit of Translation." In Zlateva, P. (Ed./trans), *Translation as Social Action*. London: Routledge.

[6] BARLOW Michael (1995). "ParaConc: A Concordancer for Parallel Texts." In *Computers and Texts* 10, Oxford: Oxford University, pp. 14-16.
Disponible sur : http://www.ruf.rice.edu/~barlow/para-ox.html

[7] BARLOW Michael (1996). "Analysing Parallel Texts with ParaConc." *ALLC/ACH Conference Abstracts*, Bergen, Norway, 1996, pp. 25-27.
Disponible sur : http://www.ruf.rice.edu/~barlow/norway.html

[8] BARLOW Michael (1999). "MonoConc1.5 and ParaConc." *International Journal of Corpus Linguistics* 4(1), pp. 173-184.

[9] BARLOW Michael (2002). "ParaConc: Concordance software for multilingual parallel corpora." In *Proceedings of the LREC-2002 Workshop 'Language Resources for Translation Work and Research'*, Canary Islands, Spain, 2002, pp. 20-24.

[1] La *Cyberbibliographie* enregistrée sur le Cd-rom complète cette liste de références bibliographiques : /fichiersCD/stmz/page3.htm.

[10] BÉCUE Monica, PAGÈS Jérôme, PARDO Campo-Elias (2004). "Analysis of multilingual free responses." In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, 2004, pp. 119-127.
Disponible prochainement sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/

[11] BÉCUE Monica, PEIRO R. (1993). « Les quasi-segments pour une classification automatique des réponses ouvertes. » In *Actes des 2es Journées Internationales d'analyse des données textuelles*, Montpellier, 1993, ENST, Paris, pp. 310-325.

[12] BENNETT Paul (1994). "The Translation Unit in Human and Machine." *Babel* 40, pp. 12-20.

[13] BENSON Morton, BENSON Evelyn, ILSON Robert (1986). *The BBI Combinatory Dictionary of English.* Amsterdam-Philadelphia: John Benjamins Publishing Company, 286p. [Special Edition: Moscow, Russki Yazik, 1990].

[14] BENZÉCRI Jean-Paul & coll. (1973). *La taxinomie*, Vol. I ; *L'analyse des correspondances*. Dunod : Paris.

[15] BEUST Pierre (2002). «Un outil de coloriage de corpus pour la représentation de thèmes. » In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, 2002, pp. 161-172.
Disponible sur :
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/beust.pdf

[16] BIBER Douglas (1994a). "Using Register-Diversified Corpora for General Language Studies." *Computational Linguistics* 19(2), pp. 219-241.
Disponible sur : http://acl.ldc.upenn.edu//J/J93/J93-2001.pdf

[17] BIBER Douglas (1994b). "Co-Occurrrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition." *Computational Linguistics* 19(3), pp. 531-538.

[18] BIRD Steven, LIBERMAN Mark (2001). "A Formal Framework for Linguistic Annotation." *Speech Communication* 33(1,2), pp 23-60.

[19] BLANK Ingeborg (2000). "Terminology extraction from parallel technical texts." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 237-252.

[20] BOUILLON Pierrette (Ed.) (1998). *Traitement automatique des langues naturelles*. Aupelf-Uref – Editions Duculot, 245 p.

[21] BOUILLON Pierrette, CLAS André (Eds.) (1993). *La Traductique: Études et Recherches de traduction par ordinateur.* Montréal: Les Presses de l'Université de Montréal, 507 p.

[22] 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.
Disponible sur : http://www.cfwb.be/franca/termin/charger/rint19.pdf

[23] BOURIGAULT Didier, SLODZIAN Monique (1999). « Pour une terminologie textuelle. » In *Actes de la troisième conférence 'Terminologie et Intelligence Artificielle'*, Nantes, 1999, pp. 29-32.
Disponible sur : http://www.cfwb.be/franca/termin/charger/rint19.pdf

[24] BRACHMAN Ronald, SCHMOLZE Jim. (1985). "An overview of the KL-ONE Knowledge Représentation System." *Cognitive Science* 9(2), pp. 171-216.

[25] BREW Chris, McKELVIE David (1996). "Word-pair extraction for lexicography." In Oflazer K., Somers H. (Eds.), Proceedings of the Second International Conference on New Methods in Language Processing, Ankara, 1996, pp. 45-55.
Disponible sur :
http://www.ltg.ed.ac.uk/~chrisbr/papers/nemlap96/nemlap96.html

[26] BROWN Peter, DELLA PIETRA Stephen, DELLA PIETRA Vincent, MERCER Robert (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics* 19(2), pp. 263-311.
Disponible sur : http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf

[27] 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.

[28] BROWN Ralf (1998). "Automatically-Extracted Thesauri for Cross-Language IR: When Better is Worse." In *Proceedings of the First Workshop on Computation Terminology*, Montreal, 1998, pp. 15-21.

[29] BROWN Ralf, CARBONELL Jaime, YANG Yiming (2000). "Automatic dictionary extraction for cross-language information retrieval." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 275-298.

[30] BRUNET Etienne (1999). Logiciel *HYPERBASE*. Manuel d'utilisation. Commercialisé et diffusé par l'InaLF (Nice) et les Éditions Champion, Paris. Disponible sur : http://ancilla.unice.fr

[31] BRUNET Etienne (2000). « Qui lemmatise, dilemme attise. » Revue électronique *Lexicometrica* 2.
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2.htm

[32] BRUNET Etienne (2002). « Le lemme comme on l'aime. » In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, 2002, pp. 221-232.
Disponible sur :
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF 2002/brunet.pdf

[33] CASILLAS Arantza, ABAITUA Joseba, MARTINEZ Raquel (2000). "DTD-Driven Bilingual Document Generation." In *Proceedings of the NAACL-ANLP Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems*, Seattle, 2000, pp. 32-38.
Disponible sur : http://www.cs.bgu.ac.il/~nlg2000/final/arantza/origin.ps.gz

[34] CATIZONE Roberta, RUSSEL Graham, WARWICK Susan (1989). "Deriving Translation Data from Bilingual Texts." In *Proceedings of the First International Acquisition Workshop*, Detroit, 1989.

[35] CHAPMAN Robert (Ed.) (1977). *Roget's International Thesaurus* (Fourth Edition). New York: Harper & Row Publishers, Inc., 1317 p.

[36] CHEN Stanley (1993). "Aligning Sentences in Bilingual Corpora using Lexical Information." In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Ohio, 1993, pp. 9-16.
Disponible sur :
http://www-2.cs.cmu.edu/afs/cs/user/sfc/www/papers/aclrev.ps.gz

[37] CHIAO Yun-Chuang, ZWEIGENBAUM Pierre (2002). "Looking for French-English translations in comparable medical corpora." *Journal of the American Medical Informatics Association* 8(suppl), pp. 50-154.
Disponible sur :
http://www-test.biomath.jussieu.fr/~pz/FTPapiers/Chiao:AMIA2002.pdf

[38] CHIBOUT Karim, MARIANI Joseph, MASSON Nicolas *et al.* (Eds.) (2000). *Ressources et évaluation en ingénierie des langues.* Bruxelles: De Boeck & Larcier s.a., 677 p.

[39] 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.

[40] CHURCH Kenneth (1993). "Char_align: A Program for Aligning Parallel Texts at the Character Level." In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, Ohio, 1993, pp. 1-8.
Disponible sur : http://acl.ldc.upenn.edu//P/P93/P93-1001.pdf

[41] CHURCH Kenneth, GALE William (1991). "Concordances for Parallel Text." In *Proceedings of the Seventh Annual Conference of UW Centre for the New OED and Text Research*, Oxford, 1991, pp. 40-62.

[42] CHURCH Kenneth, HANKS Peter (1990). "Word association norms, mutual information, and lexicography." *Computational Linguistics* 16(1), pp. 22-29.
Disponible sur : http://acl.ldc.upenn.edu//P/P89/P89-1010.pdf

[43] DAGAN Ido, CHURCH Kenneth. (1997). "Termight: Coordinating man and machine in bilingual terminology acquisition." *Machine Translation* 12(1-2), pp. 89-107.
Disponible sur : www.cs.biu.ac.il/~dagan/publications/dagan_church_97.ps

[44] DAILLE Béatrice, GAUSSIER Éric, LANGÉ Jean-Marc (1994). "Towards Automatic Extraction of Monolingual and Bilingual Terminology." In *Proceedings of the 14th International Conference on Computational Linguistics*, Kyoto, 1994, pp. 515-521.
Disponible sur : http://acl.ldc.upenn.edu/C/C94/C94-1084.pdf

[45] 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.

[46] DEBILI Fathi, SAMMOUDA Elyès (1992). "Appariement des phrases de textes bilingue francais-anglais et francais-arabe." In *Proceedings of the 14th International Conference of Computational Linguistics 'COLING-92'*, Nantes, 1992, vol. 2, pp. 517-524.
Disponible sur : http://www.ercim.org/medconf/papers/debili1.html

[47] DEBILI Fathi, SAMMOUDA Elyès, ZRIBI Adnane (1994). « De l'appariement des mots à la comparaison de phrases : un algorithme pour la reconnaissance de la paraphrase et de la traduction. » In *Actes du 9ème Congrès 'Reconnaissance des Formes et Intelligence Artificielle'*, Paris, 1994.

[48] 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.

[49] DÉJEAN Hervé, GAUSSIER Éric (2002). « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. » In Véronis J. (Ed.), Revue électronique *Lexicometrica*, n° spécial « Corpus alignés ». :
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm

[50] DÉJEAN Hervé, GAUSSIER Éric, SADAT Fatia (2002). "An approach based on multilingual thesauri and model combination for bilingual lexicon extraction." In *Proceedings of the 19th International Conference on Computational Linguistics 'COLING-02'*, Taipei, Taiwan, pp. 218-224.
Disponible sur : http://acl.ldc.upenn.edu/C/C02/C02-1166.pdf

[51] DENNETT Gerald (1995). *Translation Memory: Concept, products, impact and prospects.* MSc Major project report, South Bank University, London, 56p.
Disponible sur :
http://www.star-uk.co.uk/About_us/People/Gerald_Dennett/msc.pdf

[52] DEROUBAIX Jean-Claude (1998). « Deux langues pour une même politique : étude d'un corpus bilingue parallèle de textes politiques. » In *Actes des 4es Journées internationales d'Analyse statistique des Données Textuelles*, Nice, 1998, pp. 253-178.
Disponible sur :
www.cavi.univ-paris3.fr/lexicometrica/ jadt/jadt1998/deroubai.htm

[53] DEROUBAIX Jean-Claude (2004). « Que faire des corpus multilingues parallèles ? Une expérience. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, 2004, pp. 295-303.
Disponible prochainement sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/

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

[55] DIAB Mona (2000). "An Unsupervised Method for Multilingual Word Sense Tagging Using Parallel Corpora." In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics Workshop on Word Senses and Multi-linguality*, Hong Kong, 2000, pp. 1-9.
Disponible sur : http://acl.ldc.upenn.edu/W/W00/W00-0801.pdf

[56] DICE Lee (1945). "Measures of the Amount of Ecologic Associations between Species." *Journal of Ecology* 26(3), pp. 297-302.

[57] EAGLES (Expert Advisory Group on Language Engineering Standards) (1995). "Benchmarking translation memories." *Final Report on Evaluation of Natural Language Processing Systems.* Doc EAG-EWG-PR.2.
Disponible sur : http://issco-www.unige.ch/ewg95

[58] FUCHS Catherine, LE GOFFIC Pierre (1992). *Les Linguistiques Contemporaines*. Paris : Librairie Hachette, 158 p.

[59] FUNG Pascale (1995). "Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus." In *Proceedings of the Third Workshop on Very Large Corpora*, The Association for Computational Linguistics, 1995, Cambridge: Massachusetts, 1995, pp. 173-183.
Disponible sur : http://acl.ldc.upenn.edu/W/W95/W95-0114.pdf

[60] FUNG Pascale (2000). "A Statistical View on bilingual lexicon extraction: From Parallel corpora to non-parallel corpora." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 219-236.

[61] FUNG Pascale, CHURCH Kenneth (1994). "K-vec: A New Approach for Aligning Parallel Texts". In *Proceedings of the 15th International Conference on Computational Linguistics 'COLING-94'*, Kyoto, 1994, pp. 1096-1104.
Disponible sur : http://acl.ldc.upenn.edu/C/C94/C94-2178.pdf

[62] FUNG Pascale, McKeown Kathleen (1994). "Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping." In *Proceedings of the First Conference of the Association for Machine Translation in the Americas 'AMTA-94'*, Columbia, Maryland, 1994, pp. 81-88.
Disponible sur : http://arxiv.org/PS_cache/cmp-lg/pdf/9409/9409011.pdf

[63] GALE William, CHURCH Kenneth (1991a). "Identifying word correspondences in parallel texts." In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, 1991, pp. 152-157.

[64] GALE William, CHURCH Kenneth (1991b). "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.
Disponible sur : http://acl.ldc.upenn.edu//P/P91/P91-1023.pdf

[65] 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.
Disponible sur : http://acl.ldc.upenn.edu/P/P98/P98-1074.pdf

[66] GAUSSIER Eric (1999). « Le traitement automatique des langues au service de la terminologie. » In *Actes de la Conférence sur la coopération dans le domaine de la terminologie en Europe*, Association Européenne de Terminologie (AET), Paris, 1999.
Disponible sur : http://www.eaft-aet.net/actes/GAUSSIER.htm

[67] GAUSSIER Éric, HULL David, AÏT-MOKHTAR Salah (2000). "Term alignment in use: Machine-aided human translation." In Véronis J. (Ed.) *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 253-274.

[68] GAUSSIER Eric, LANGÉ Jean-Marc (1995). « Modèles statistiques pour l'extraction de lexiques bilingues. » *T.A.L.* 36(1-2), pp. 133-155.

[69] GAVRICHINA K. S., SAZONOV M. A., GAVRICHINA I.N. (1993). *Dictionnaire commercial et financier.* Moscou : VIKRA, 792p.

[70] GÉMARD Jean-Claude (1999). « Les enjeux de la traduction juridique : principes et nuances. » In *Übersetzung von Rechtstexten : Probleme und Methoden*, Bern : ASTTI, 1999, pp. 45-64.
Disponible sur : http://www.tradulex.org/Actes1998/Gemar.pdf

[71] GHORBEL Hatem, *et al.* (2002). « L'alignement multicritères des documents médiévaux. » In Véronis J. (Ed.), Revue électronique *Lexicometrica*, n° spécial « Corpus alignés ».
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm

[72] GREFENSTETTE Gregory (1994). *Explorations in Automatic Thesaurus Discovery.* Dordrecht: Kluwer Academic Publishers, 320 p.

[73] GREFENSTETTE Gregory (1992). "Use of syntactic context to produce term association lists for text retrieval." In *Proceedings of the 15th International Conference on Research and Development in Information Retrieval*, Copenhagen, pp. 89-97.

[74] GREFENSTETTE Gregory (Ed.) (1998). *Cross-Language Information Retrieval.* Dordrecht: Kluwer Academic Publishers, 182 p.

[75] GROSSMANN Francis, TUTIN Agnès (Eds.) (2003). *Les collocations : analyse et traitement.* Amsterdam : Editions "De Werelt", 142 p.

[76] HABERT Benoît, BERTRAND-GASTALDY Suzanne, NAZARENKO Adeline *et al.* (1997). « Recyclage d'analyses syntaxiques automatiques pour le repérage de variantes de termes. » In *Fifth RIAO conference 'Computer-Assisted Information Searching on Internet'*, Montreal, 1997, pp. 751-760.
Disponible sur : http://www.limsi.fr/Individu/habert/Publications/Fichiers/habert-et-al97b/habert-et-al97b.html

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

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

[79] HARRIS Brian (1988). "Bi-text: A New Concept in Translation Theory." *Language Monthly* 54, pp. 8-10.

[80] HARRIS Zellig (1988). *Language and Information*. New York: Columbia University Press, 120 p.

[81] HARUNO Masahiko, IKEHARA Satoru, YAMAZAKI Takefumi (1996). "Learning Bilingual Collocations by Word-Level Sorting." In *Proceedings of the 16th International Conference on Computational Linguistics 'COLING-96'*, Copenhagen, 1996, pp. 525-530.

[82] HARUNO Masahiko, YAMAZAKI Takefumi (1996). "High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, 1996, pp. 131-138.
Disponible sur : http://acl.ldc.upenn.edu//P/P96/P96-1018.pdf

[83] HEATHER Michael, ROSSITER Nick (1990). "Syntactical Relations in Parallel Text". In Choueka Y. (Ed.), *Proceedings of the 15th International Conference of Association for Literary* & *Linguistic computing*, Jerusalem, 1988, pp. 197-214.

[84] HEIDEN Serge (2002). *Weblex* : *Manuel Utilisateur. Version 4.1 (intermédiaire).* UMR 8503, ENS Lettres et Sciences humaines.
Disponible sur : http://lexico.ens-lsh.fr/doc/weblex.pdf

[85] HMSO Technical Services (1981). *General Specification of HMSO Data Tapes for Information Retrieval*. London: HMSO.

[86] HOFLAND Knut, JOHANSSON Stig (1998). "The Translation Corpus Aligner: A program for automatic alignment of parallel texts." In Johansson S., Oksefjell S. (Eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, Amsterdam: Rodopi, pp. 87-100.

[87] HUMBLEY John (1999). "Terminologie et documentation." In *Actes de la Conférence sur la coopération dans le domaine de la terminologie en Europe*, Association Européenne de Terminologie (AET), Paris, 1999.
Disponible sur : http://www.eaft-aet.net/actes/HUMBLEY.htm

[88] HUTCHINS John (1994). « Vers une nouvelle époque en traduction automatique. » In Clas, A., Bouillon, P. (Eds.), *TA-TAO*: *recherches de pointe et applications immédiates*. Actes des Troisièmes Journées Scientifiques du réseau thématique 'Lexicologie, Terminologie, Traduction', Montréal, 1993. AUPELF/UREF, pp. 3-16.

[89] HUTCHINS John (1998). "The origins of the translator's workstation." *Machine Translation* 13(4), pp. 287-307.
Disponible sur :
http://ourworld.compuserve.com/homepages/wjhutchins/MTJ1998.pdf

[90] HUTCHINS John (2001). "Machine translation over fifty years." *Histoire Epistémologie Langage* 23(1), pp. 7-31.
Disponible sur :
http://ourworld.compuserve.com/homepages/WJHutchins/HEL.pdf

[91] HUTCHINS John, SOMERS Harold (1992). *An Introduction to Machine Translation*. London: Academic Press, 362p.
Disponible sur :
http://ourworld.compuserve.com/homepages/WJHutchins/IntroMT-TOC.htm

[92] ISABELLE Pierre (1992a). « La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. » *META* 37(4), pp. 721-737. Disponible sur : http://www.erudit.org/revue/meta/1992/v37/n4/003228ar.pdf

[93] ISABELLE Pierre (1992b). "Bi-textual Aids for Translators." In *Proceedings of the Eighth Annual Conference of UW Centre for the New OED and Text Research*. Waterloo, 1992, pp. 76-89.

[94] ISABELLE Pierre, MACKLOVITCH Elliott (1990). « Où en est la traduction automatique. » In *Actes du colloque annuel CIPS/CATA*, Ottawa, 1990, 11p. Disponible sur : http://www-rali.iro.umontreal.ca/Publications.en.html

[95] ISABELLE Pierre, SIMARD Michel (1996). « Propositions pour la représentation et l'évaluation des alignements de textes parallèles. » *Projet ARCADE*. Disponible sur : http://www-rali.iro.umontreal.ca/arc-a2/PropEval

[96] ISABELLE Pierre, WARWICK-ARMSTRONG Susan (1993). « Les corpus bilingues : une nouvelle ressource pour le traducteur. » In Bouillon P., Clas A. (Eds.), *La Traductique: Études et Recherches de traduction par ordinateur.* Montréal: Les Presses de l'Université de Montréal, pp. 288-306.

[97] JANICIJEVIC Tatjana (1997). « L'approche informatisée du dépistage des unités de traduction. » In *Actes du colloque interdisciplinaire 'L'informatique dans les études françaises'*, Kingston, 1997. Disponible sur : http://qsilver.queensu.ca/french/Confs/Lil97/TatjanaJanicijevic.html

[98] JENNINGS Andrew, EDGE Colin, STERNBERG Michael (2001). "An approach to improving multiple alignments of protein sequences using predicted secondary structure." *Protein Engineering* 14(4), Oxford University Press, pp. 227-231. Disponible sur : http://peds.oupjournals.org/cgi/content/full/14/4/227

[99] JOHANSSON Stig, EBELING Jarle, HOFLAND Knut (1996). "Coding and aligning the English-Norwegian parallel corpus". In Aijmer K., Altenberg B., Johansson M. (Eds.), *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*, Lund, 1994. Lund: Lund University Press, pp. 87-112. Disponible sur : ftp://www.hd.uib.no/pub/corpora/enpc.lund.ps

[100] JOHANSSON Stig, HOFLAND Knut (1994). "Towards an English-Norwegian parallel corpus." In Fries, U., Tottie, G., Schneider P. (Eds.), *Creating and using English language corpora*. Amsterdam: Rodopi, pp.25-37.

[101] JONES Daniel, SOMERS Harold (1995). "Bilingual vocabulary estimation from noisy parallel corpora using variable bag estimation." In Bolasco S., Lebart L., Salem A. (Eds.), *Actes des 3es Journées internationales d'Analyse statistique des Données Textuelles*, Rome, 1995, vol. 1, pp. 255-262.

[102] KAY Martin, RÖCHEISEN Martin. (1993). "Text-Translation Alignment." *Computational Linguistics* 19(1), 121-142. [Première parution : Kay Martin, Röscheisen Martin (1988). *Text-Translation Alignment*. Technical Report, Xerox Palo Alto Research Center.]
Disponible sur : http://acl.ldc.upenn.edu//J/J93/J93-1006.pdf

[103] KILGARRIFF Adam, GREFENSTETTE Gregory (2004). "Introduction to the Special Issue on the Web as Corpus." *Computational Linguistics – Special Issue on the Web as Corpus*, 29(3), pp. 333-348.

[104] KING Philip, WOOLLS David (1996). "Creating and Using a multilingual parallel concordancer." *Translation and Meaning* Part 4, 1996, pp. 459-466.
Disponible sur : http://www.oasis-open.org/cover/kingCreatingConcordancer.html

[105] KLAVANS Judith, TZOUKERMANN Evelyne (1990). "The BICORD System: Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries." In *Proceedings of the 13th International Conference on Computational Linguistics 'COLING-90'*, Helsinki, 1990, pp. 174-179.
Disponible sur : http://www.ldc.upenn.edu/acl/C/C90/C90-3031.pdf

[106] KLEVBACKE Anders (2001). "Parallel Text Alignment Algorithms." *Algorithms - Advanced Course Project Report*, Chalmers University of Technology, 18 p.

Disponible sur : http://www.dtek.chalmers.se/~d95ankle/algorithms-ac-project.html

[107] KONDRAK Grzegorz (2000). "A New Algorithm for the Alignment of Phonetic Sequences." In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, 2000, pp. 288-295. Disponible sur : http://web.cs.ualberta.ca/~kondrak/naacl00.pdf

[108] KONDRAK Grzegorz (2001). *"Identifying Cognates by Phonetic and Semantic Similarity." In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, 2001, pp. 103-110.
Disponible sur : http://web.cs.ualberta.ca/~kondrak/naacl01.pdf

[109] KRAIF Olivier (2001). *Constitution et exploitation de bi-textes pour l'Aide à la traduction*. Thèse pour obtenir le grade de Docteur de l'Université de Nice Sophia Antipolis, 679 p.

Disponible sur : http://web.cs.ualberta.ca/~kondrak/naacl01.pdf

[110] KRAIF Olivier (2002). « Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné" » In Véronis J. (Ed.), Revue électronique *Lexicometrica*, n° spécial « Corpus alignés ».

Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm

[111] KUNIN A. (1984). *English-Russian Phraseological Dictionnary*, Forth Edition, Moscow: Russky Yazyk, 942 p.

[112] KUPIEC Julian (1993). "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora." In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, Ohio, 1993, pp. 17-22.
Disponible sur : http://acl.eldoc.ub.rug.nl/mirror/P/P93/P93-1003.pdf

[113] LABBÉ Dominique, THOIRON P, SERANT D. (Eds.) (1988). *Etudes sur la richesse et la structure lexicales*. Paris-Genève : Slatkine-Champion, 172p.

[114] LAFON Pierre (1981). « Analyse lexicométrique et recherche des cooccurrences. » *MOTS* 3, pp. 95-148.

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

[116] LAMALLE Cédric, MARTINEZ William, FLEURY Serge, SALEM André *et al.* (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation.* SYLED–CLA2T, Université de la Sorbonne nouvelle – Paris 3, 48 p.
Disponible sur :
http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/manuelsL3/

[117] 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.
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF 2002/lamalle_salem.pdf

[118] LAMBERT T., LEBART Ludovic, MORINEAU Alain, PLEUVRET Philippe (1996). *Manuel de référence de SPAD*. CISIA-CERESTA, Saint-Mandé.

[119] LANGÉ Jean-Marc, GAUSSIER Eric (1995). « Alignement de corpus multilingues au niveau des phrases ». *TAL* 36(1-2), pp. 67-80.

[120] 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.

[121] 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.

[122] LE GRAND ROBERT ELECTRONIQUE (1995). *Outils d'aide à la rédaction sur la base du Grand Robert de la langue française en 9 volumes*. Paris : Liris Interactive.

[123] LE ROBERT Quotidien (1996). Paris : Dictionnaires LE ROBERT.

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

[125] LEECH Geoffrey (1987). "General Introduction." In Garside R., Leech G., Sampson G., (Eds.), *The Computational Analysis of English*. London and New York: Longman, pp. 1-15.

[126] LIXUN Wang (2001). "Exploring Parallel Concordancing in English and Chinese." *Language Learning & Technology* 5(3), pp. 174-184.
Disponible sur: http://llt.msu.edu/vol5num3/pdf/wang.pdf

[127] LONGMAN Dictionary of Contemporary English (Second Edition) (1987). London: Longman Group Limited.

[128] LONSDALE Deryle (1994). « Extraction d'un vocabulaire bilingue : outils et méthodes. » In Clas A., Bouillon P. (Eds.), *TA-TAO: Recherches de pointe et applications immédiates. Actes du Colloque de Montréal 1993*. Beirut: FMA, pp. 241-253.

[129] MACKLOVITCH Elliott (1992). "Corpus-based tools for translators." In Losa, E.F (Ed.), *Frontiers: Proceedings of the 33rd Annual Conference of the American Translators Association*. Medford, New Jersey: Learned Information Inc., pp. 317-325.
Disponible sur : http://www-rali.iro.umontreal.ca/Publications/tmi92.ps

[130] MACKLOVITCH Elliott (1993). « Le PTT, ou les aides à la traduction. » In Bouillon P., Clas A. (Eds.), *La traductique : Études et Recherches de traduction par ordinateur.* Montréal : Les Presses de l'Université de Montréal, pp. 281-287.

[131] MACKLOVITCH Elliott, HANNAN Marie-Louise (1996). "Line'Em Up: Advances in Alignment Technology And Their Impact On Translation Support Tools." In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montréal, 1996, pp. 145-156.
Disponible sur : http://www-rali.iro.umontreal.ca/Publications/mhAMTA96.ps

[132] MALFRÈRE Fabrice, DUTOIT Thierry (2000). « Alignement automatique du texte sur la parole et extraction de caractéristiques prosodiques. » In Chibout K., Mariani J., Masson N. *et al.* (Eds.), *Ressources et évaluation en ingénierie des langues.* Bruxelles : De Boeck & Larcier s.a., pp. 541-552.

[133] MANGEOT-LEREBOURS Mathieu (2001). *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse pour obtenir le grade de Docteur de l'Université Joseph Fourier.
Disponible sur : http://www-clips.imag.fr/geta/mathieu.mangeot/MM-These/plan-these.html

[134] MARTIN Robert (2002). *Comprendre la linguistique*. Paris : PUF, 190 p.

[135] MARTINEZ William (2000). « Mise en évidence de rapports synonymiques par la méthode des cooccurrences. » In *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, Lausanne, 2000, pp. 197-203.
Disponible sur :
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/78/78.pdf

[136] MARTINEZ William (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. Thèse pour obtenir le grade de Docteur de l'Université de la Sorbonne nouvelle – Paris 3, 468 p.

[137] 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.
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/martinez_zimina.pdf

[138] McENERY Tony, OAKES Michael (1995). "Sentence and word alignment in the CRATER project: methods and assessment." In *Proceedings of the EACL-SIGDAT Workshop 'From texts to Tags: Issues in Multilingual Language Analyses'*, Dublin, 1995, pp. 77-86.

[139] MELAMED Dan (1997). "A Portable Algorithm for Mapping Bitext Correspondence." In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, 1997, pp. 305-312.
Disponible sur : http://acl.ldc.upenn.edu//P/P97/P97-1039.pdf

[140] MELAMED Dan (1998). "Models of Co-occurrence." *Institute for Research in Cognitive Science Technical Report #98-05.* University of Pennsylvania, Philadelphia, 34 p.
Disponible sur : http://cs.nyu.edu/~melamed/ftp/papers/coocmod.ps.gz

[141] MELAMED Dan (1999). "Bitext Maps and Alignment via Pattern Recognition." *Computational Linguistics* 25(1), pp. 107-130.
Disponible sur : http://acl.ldc.upenn.edu//J/J99/J99-1003.pdf

[142] MELAMED Dan (2000a). "Models of Translational Equivalence among Words." *Computational Linguistics* 26(2), pp. 221-250.
Disponible sur : http://cs.nyu.edu/~melamed/ftp/papers/clmote.pdf

[143] MELAMED Dan (2000b). "Pattern recognition for mapping bitext correspondence." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers, pp. 25-47.

[144] MELBY Alan (1981). "Translators and Machines - Can they cooperate ?" *META* 26(1), pp. 23-34.
Disponible sur : http://www.erudit.org/revue/meta/1981/v26/n1/003619ar.pdf

[145] MELBY Alan (1982). "A Bilingual Concordance System and its Use in Linguistic Studies." In Gutwinski W., Jolly G. (Eds.), *Proceedings of the Eighth meeting of the Linguistics Association of Canada and the United States*, Toronto, 1981. Eighth LACUS forum, Hornbeam Press, pp. 541-549.

[146] 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.

[147] MELLET Sylvie (2001). « Lemmatisation et encodage grammatical : un luxe inutile ? » Revue électronique *Lexicometrica* 3.
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3.htm

[148] MILLER Philip, TORRIS Thérèse (Eds.) (1990). *Formalismes syntaxiques pour le traitement automatique du langage naturel.* Paris : Hermès, 359 p.

[149] MINYAR-BELOROUCHEV Rurik (1996). *Teoriya i Metodi perevoda*. Moscou : Moscovski litzei, 208p. [МИНЬЯР-БЕЛОРУЧЕВ Рюрик (1996). *Теория и методы перевода.* Москва: Московский лицей, 208 p.].

[150] MOUNIN George (1976). *Les problèmes théoriques de la traduction*. Paris : Gallimard, 296 p.

[151] MULLER Charles (1992). *Principes et méthodes de statistique lexicale.* Paris : Editions Champion, 210p. [réimpression de l'édition 1977].

[152] OLSSON Leif-Jöran, BORIN Lars (2000). "A web-based tool for exploring translation equivalents on word and sentence level in multilingual parallel corpora." In *Publications of the Research Group for LSP and Theory of Translation at the University of Vaasa* 27, pp. 76-84.
Disponible sur : http://www.ling.uu.se/lars/pblctns/VAKKI00.pdf

[153] PAQUIN Louis-Claude, BEAUCHEMIN Jacques (1989). « Apport de l'ordinateur à l'analyse des données textuelles. » In *Actes du colloque 'La description des langues naturelles en vue d'applications linguistiques'*, Québec: Centre international de recherche sur le bilinguisme, 1989, pp. 21-31.
Disponible sur :
http://www.ling.uqam.ca/sato/publications/bibliographie/Apport.htm

[154] PERRON Jean (1990). « Présentation du progiciel de dépouillement terminologique assisté par ordinateur : TERMINO. » In *Actes du Colloque international 'Les industries de la langue: Perspectives des années 1990'*, Montréal, 1991, pp. 715-755.

[155] PIERREL Jean-Marie (Ed.) (2000). *Ingénierie des langues*. Hermès Science Publications, 354p.

[156] PIERREL Jean-Marie, SLODZIAN Monique (Dir.) (2004). *Techniques informatiques et structuration de terminologies*. Numéro spécial de la revue *RSTI-Revue d'Intelligence Artificielle*, 18 (1/2004). Paris : Hermès/Lavoisier.

[157] PIPERIDIS Stelios, PAPAGEORGIUO Harris, BOUTSIS Sotiris (2000). "From sentences to words and clauses." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 117-138.

[158] POLGUÈRE Alain (2003*)*. *Lexicologie et sémantique lexicale*. Les Presses de l'Université de Montréal, 260 p.

[159] RABINER Lawrence, JUANG Biing-Hwang (1986). "An introduction to hidden markov models." *IEEE Magazine on Accoustics, Speech and Signal Processing* 3(1), pp. 4-16.

[160] RASTIER François (1989). *Sens et textualité*. Paris : Hachette, 287 p.

[161] RESNIK Philip (1998). "Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual text." In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, Langhorne, 1998, pp. 72-82.
Disponible sur : http://umiacs.umd.edu/~resnik/pubs/amta98.ps.gz

[162] RESNIK Philip, MELAMED Dan (1997). "Semi-automatic acquisition of domain-specific translation lexicons." In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, 1997, pp. 340-347.
Disponible sur : http://xxx.lanl.gov/PS_cache/cmp-lg/pdf/9703/9703005.pdf

[163] RESNIK Philip, SMITH Noah (2004). "The Web as a Parallel Corpus." *Computational Linguistics – Special Issue on the Web as Corpus*, 29(3), pp. 349-380.
Disponible sur : http://mitpress.mit.edu/journals/pdf/coli_29_3_349_0.pdf

[164] RIBEIRO António, DIAS Gaël, LOPES Gabriel, MEXIA João (2001). "Cognates Alignment." In *Proceedings of the Eighth Machine Translation Summit*, Santiago de Compostela, 2001, pp. 287-292.
Disponible sur : http://www.eamt.org/summitVIII/papers/ribeiro.pdf

[165] ROMARY Laurent (2000). « Outils d'accès à des ressources linguistiques. » In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Paris : Editions Hermès, pp. 193-211.

[166] ROMARY Laurent, BONHOMME Patrice (2000). "Parallel alignment of structured documents." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 201-217.

[167] ROMARY Laurent, MEHL Nathalie, WOOLLS David (1995). "The Lingua Parallel Concordancing Project: Managing Multilingual Texts for Educational Purpose." *Text Technology* 5(3), p. 206-220.
Disponible sur : http://citeseer.ist.psu.edu/63642.html

[168] SALEM André (1987). *Pratique des segments répétés : essai de statistique textuelle.* Paris : Klincksieck, 333 p.

[169] SALEM André (1993). *Méthodes de la statistique textuelle*. Thèse pour le Doctorat d'Etat ès letters. Université de la Sorbonne nouvelle - Paris 3, 819p.

[170] SALEM André (2004). «Introduction à la résonance textuelle. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, 2004, pp. 986-992.
Disponible prochainement sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/

[171] SANSONETTI Luiggi (2004). « Apports de la statistique textuelle pour le repérage des reprises et reformulations dans les corpus d'interaction verbale entre un adulte et un enfant. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, 2004, pp. 993-999.
Disponible prochainement sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/

[172] SANTORINI Beatrice. (1990). *Part-of-speech Tagging Guidelines for the Penn Treebank Project*. Technical report, Department of Computer and Information Science, University of Pennsylvania, 34p.
Disponible sur : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps

[173] SANTOS Diana (2000). "The translation network. A model for a fine-grained description of translations." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 169-186.

[174] SCHMID Helmut (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, 1994.
Disponible sur : http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

[175] SELESKOVITCH Danica (1968). *L'interprète dans les conférences internationales*. Paris : Lettres modernes, Minard, 261 p.

[176] SHAROFF Serge (2002). "Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics." In *Proceedings of Third International Conference on Language Resources and Evaluation 'LREC-02'*, Las Palmas, 2002, pp. 447-452.
Disponible sur : http://www.comp.leeds.ac.uk/ssharoff/texts/lrec-02.pdf

[177] SIMARD Michel (2000). "Multilingual text alignment." In Véronis J. (Ed.), *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, pp. 49-67.

[178] 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.
Disponible sur :
http://portal.acm.org/citation.cfm?id=962411&dl=ACM&coll=portal

[179] SIMARD Michel, FOSTER George, PERRAULT François (1993). *TransSearch: a bilingual concordance tool.* Technical Report. Laval, Canada: Centre for Information Technology Innovation (CITI), 19 p.
Disponible sur : http://rali.iro.umontreal.ca/Publications/

[180] SIMARD Michel, PLAMONDON Pierre (1996). "Bilingual Sentence Alignment: Balancing Robustness And Accuracy." In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montréal, 1996, pp. 135-144.
Disponible sur : http://rali.iro.umontreal.ca/Publications/

[181] SLODZIAN Monique, SOUILLOT Jacques (Eds.) (1997). *Compréhension multilingue en Europe, Actes du Séminaire de Bruxelles 10 et 11 mars 1997.* Paris : Centre de Recherches en Ingénierie Multilingue. INaLCO.
Disponible sur : http://crim.inalco.fr/recomu/colloque/

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

[183] SOMERS Harold (1998b). "Similarity metrics for aligning children's articulation data." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics 'COLING–ACL-98'*, Montreal, 1998, pp. 1227-1232.
Disponible sur : http://acl.ldc.upenn.edu/P/P98/P98-2200.pdf

[184] SOMERS Harold (2001). "Bilingual Corpora and Language Engineering." In *Proceedings of Anglo-Indian workshop 'Language Engineering for South-Asian languages'(LESAL)*, Mumbai, April 2001.
Disponible sur : http://www.emille.lancs.ac.uk/lesal.htm

[185] STEVENSON Mark, WILKS Yorick (2001). "The Interaction of Knowledge Sources in Word Sense Disambiguation." *Computational Linguistics* 27(3), pp. 321-350.
Disponible sur : http://acl.ldc.upenn.edu/J/J01/J01-3001.pdf

[186] SYNAPSE Développement (1999). *CORDIAL 6 : Correcteur global de la langue française* (édition 2000). Toulouse : Synapse Développement, 320 p.

[187] TAURITZ Daniel, SPRINKHUIZEN-KUYPER Ida (2000). "Adaptive information filtering: evolutionary computation and *n*-gram representation." In van den Bosch A., Weigand H. (Eds.), *Proceedings of the Twelfth Belgium-Netherlands Artificial Intelligence Conference*, Tilburg, 2000, pp. 157-164.
Disponible sur : http://citeseer.ist.psu.edu/381832.html

[188] VEGLIANTE Jean-Charles (1996). « Quelle théorie, pour quelle traduction ? » In Vegliante J.-Ch., *D'écrire la traduction*. Paris : Presses de la Sorbonne Nouvelle, pp. 39-62.

[189] VÉRONIS Jean (2000a). « Alignement de corpus multilingues. » In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Paris : Editions Hermès, pp. 151-171.
Disponible sur : http://www.up.univ-mrs.fr/~veronis/pdf/2000hermes6.pdf

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

[191] 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.

[192] WEBB Lynn (1999). *Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis.* Master of Arts Project, San Francisco State University, Monterey, 1992.
Disponible sur : http://www.cicenter.com/a_memory.htm

[193] WOOLLS David, KING Philip, JOHNS Tim (1993). *Multiconcord: the Lingua Multilingual Parallel Concordancer for Windows.* Lingua project ndeg.93-09/1245/F-VB.
Disponible sur : http://web.bham.ac.uk/johnstf/lingua.htm

[194] WU Dekai (2000a). "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.

[195] WU Dekai (2000b). "Alignment." In Dale R., Moisl H., Somers H. (Eds.), *Handbook of Natural Language Processing.* New York: Marcel Dekker, pp. 415-458.

[196] YAROWSKI David (1992). "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora." In *Proceedings of the 14th International Conference on Computational Linguistics 'COLING-92'*, Nantes, 1992, pp. 454-460.
Disponible sur : http://citeseer.ist.psu.edu/39762.html

[197] 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.
Disponible sur :
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/77/77.pdf

[198] ZIMINA Maria (2002). « Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. » In Véronis J. (Ed.), Revue électronique *Lexicometrica*, n° spécial « Corpus alignés ».
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm

[199] ZIMINA Maria (2004). « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, 2004, pp. 1195-1202.
Disponible prochainement sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/

Annexe A

Le corpus *Convention*

A.1 Convention de sauvegarde des Droits de l'Homme et des libertés fondamentales

Signée à Rome le 4 novembre 1950, la *Convention* permet de protéger les droits individuels de 800 millions de femmes et hommes dans 41 pays[1]. Un système international de sauvegarde, sans précédent dans l'histoire, offre à toute personne qui réside dans un Etat membre du Conseil de l'Europe, un ultime recours en cas de violation de ses droits fondamentaux : la possibilité de s'adresser à la Cour européenne des Droits de l'Homme à Strasbourg. Les juges de la Cour, totalement indépendants sont élus par l'Assemblée parlementaire.

En adoptant, dès 1950, la Convention européenne des Droits de l'Homme, l'Europe s'est dotée du système de protection des libertés fondamentales le plus élaboré au monde. Plus de 40 000 requêtes individuelles et une douzaine de requêtes étatiques ont été examinées par les organes de contrôle établis par la Convention. Elles ont souvent conduit les Etats membres à modifier leurs législations ou leurs pratiques[2]. Le catalogue des Droits de l'Homme et des libertés fondamentales est à présent identique dans toute l'Europe.

La *Convention* et ses protocoles garantissent :

- le droit à la vie, à la liberté et à la sûreté de l'individu,
- le droit à un procès équitable en matière civile et pénale,
- le droit de vote et le droit de se présenter à des élections,

---

[1] Le prédécesseur direct de la Convention européenne des Droits de l'Homme est la Déclaration universelle des droits de l'homme, adoptée par les Nations unies en 1948.

[2] Tous les États contractants, à l'exception de l'Irlande et de la Norvège, ont intégré la *Convention* dans leur législation, de telle sorte que les juridictions internes prennent ses dispositions entièrement en compte lorsqu'elles statuent sur un grief. Lorsque des États souverains ont accepté qu'une juridiction supranationale remette en cause les décisions des juridictions internes, une étape historique a été franchie dans le développement du droit international. La théorie selon laquelle les droits de l'homme ont un caractère fondamental les plaçant au-dessus des législations et des pratiques nationales a été appliquée.

- la liberté de pensée, de conscience et de religion,
- la liberté d'expression (notamment celle des médias),
- le droit au respect de ses biens.

Ils interdisent :

- la torture et les peines ou traitements inhumains ou dégradants,
- la peine de mort,
- les discriminations dans la jouissance des droits et libertés reconnus par la Convention,
- l'expulsion ou le refoulement par un État de ses propres ressortissants,
- l'expulsion collective d'étrangers.

Le texte de la *Convention* dans les langues des États signataires ainsi que des informations générales sur les divers aspects historiques et juridiques concernant sa création sont disponibles via le portail du Conseil de l'Europe ainsi que sur le site officiel de la Cour Européenne des Droits de l'Homme :

- http://www.coe.int
- http://www.echr.coe.int.

A.2 Structure du corpus Convention

Le corpus ***Convention*** a été constitué à partir des documents contenus dans la Convention européenne des Droits de l'Homme (y compris les protocoles intégraux), et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995.

Note : Le texte de la **Convention** avait été amendé conformément aux dispositions du Protocole n°3 (STE n°45), entré en vigueur le 21 septembre 1970, du **Protocole n° 5** (STE n°55), entré en vigueur le 20 décembre 1971, et du **Protocole n°8** (STE n°118), entré en vigueur le 1er janvier 1990, et comprenait en outre le texte du Protocole n°2 (STE n°44) qui, conformément à son article 5, paragraphe 3, avait fait partie intégrante de la Convention depuis son entrée en vigueur le 21 septembre 1970. Toutes les dispositions qui avaient été amendées ou ajoutées par ces **Protocoles** sont remplacées par le **Protocole n°11** (STE n°155), à compter de la date de son entrée en vigueur le 1er novembre 1998. A compter de cette date, le **Protocole n°9** (STE n°140), entré en vigueur le 1er octobre 1994, est abrogé et le **Protocole n° 10** (STE n° 146) est devenu sans objet.

Source : le texte officiel de la Convention européenne des Droits de l'Homme.

a) Description sommaire du corpus *Convention* :

Volet français (296 396 occ.)

- Convention européenne des Droits de l'Homme (5 953 occ.)
- Protocoles intégraux de la Convention (8 984 occ.)
  - *Protocole Additionnel*
  - *Protocoles °2* , °4, °6, °7, °9, °10, °11.
- Arrêts de la Cour Européenne des Droits de l'Homme (281 459 occ.)

Volet anglais (284 958 occ.)

- Convention européenne des Droits de l'Homme (5 710 occ.)
- Protocoles intégraux de la Convention (8 773 occ.)
  - *Protocole Additionnel*
  - *Protocoles °2* , °4, °6, °7, °9, °10, °11.
- Arrêts de la Cour Européenne des Droits de l'Homme (274 475 occ.)

b) Extrait du texte de base de la *Convention*

```
conv_a0_p1-1    1  Les gouvernements signataires , membres du Conseil de l'Europe ,
conv_a0_p1-1e   2  The governments signatory hereto , being members of the Council of Europe ,
/.../
conv_a0_p7-1    13 Sont convenus de ce qui suit :
conv_a0_p7-1e   14 Have agreed as follows :
conv_a1_p1-1    15 Les Hautes Parties contractantes reconnaissent à toute personne relevant de leur juridiction
                   les droits et libertés définis au titre I de la présente Convention :
conv_a1_p1-1e   16 The High Contracting Parties shall secure to everyone within their jurisdiction the rights and
                   freedoms defined in Section I of this Convention .
conv_a2.1_p1-1  17 Le droit de toute personne à la vie est protégé par la loi .
conv_a2.1_p1-1e 18 Everyone's right to life shall be protected by law .
conv_a2.1_p1-2  19 La mort ne peut être infligée à quiconque intentionnellement , sauf en exécution d'une sentence
                   capitale prononcée par un tribunal au cas où le délit est puni de cette peine par la loi .
conv_a2.1_p1-2e 20 No one shall be deprived of his life intentionally save in the execution of a sentence of a
                   court following his conviction of a crime for which this penalty is provided by law .
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 recours 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 use of force which is no more than absolutely necessary :
conv_a2.2_p2-1  23 pour assurer la défense de toute personne contre la violence illégale ;
conv_a2.2_p2-1e 24 in defence of any person from unlawful violence ;
conv_a2.2_p3-1  25 pour effectuer une arrestation régulière ou pour empêcher l'évasion d'une personne
                   régulièrement détenue ;
conv_a2.2_p3-1e 26 in order to effect a lawful arrest or to prevent the escape of a person lawfully detained ;
conv_a2.2_p4-1  27 pour réprimer , conformément à la loi , une émeute ou une insurrection .
conv_a2.2_p4-1e 28 in action lawfully taken for the purpose of quelling a riot or insurrection .
conv_a3_p1-1    29 Nul ne peut être soumis à la torture ni à des peines ou traitements inhumains ou dégradants .
conv_a3_p1-1e   30 No one shall be subjected to torture or to inhuman or degrading treatment or punishment .
conv_a4.1_p1-1  31 Nul ne peut être tenu en esclavage ni en servitude .
conv_a4.1_p1-1e 32 No one shall be held in slavery or servitude .
conv_a4.2_p1-1  33 Nul ne peut être astreint à accomplir un travail forcé ou obligatoire .
conv_a4.2_p1-1e 34 No one shall be required to perform forced or compulsory labour .
/.../
```

c) Protocole °10 de la *Convention* (extrait)

```
#pro10_a0_p1   769 Les Etats membres du Conseil de l'Europe , signataires du présent Protocole à la Convention de
                   sauvegarde des Droits de l'Homme et des Libertés fondamentales signée à Rome le 4 novembre 1950
                   ( ci-après dénommée ''la Convention'' ) ,
#pro10_a0_p1   770 The member States of the Council of Europe , signatories to this Protocol to the Convention for
                   the Protection of Human Rights and Fundamental Freedoms , signed at Rome on 4 November 1950 (
                   hereinafter referred to as '' the Convention '' ) ,
pro10_a0_p2-1  771 Considérant qu'il convient d'amender l'article 32 de la Convention en vue de réduire la
                   majorité des deux tiers qui y est prévue ,
pro10_a0_p2-1e 772 Considering that it is advisable to amend Article 32 of the Convention with a view to the
                   reduction of the two-thirds majority provided therein ,
pro10_a0_p3-1  773 Sont convenus de ce qui suit :
pro10_a0_p3-1e 774 Have agreed as follows :
pro10_a1_p1-1  775 Les mots ''des deux tiers'' sont supprimés du paragraphe 1 de l'article 32 de la Convention .
pro10_a1_p1-1e 776 The words '' of two-thirds '' shall be deleted from paragraph 1 of Article 32 of the
                   Convention.
/.../
pro10_a2.2_p1-1    785  Les instruments de ratification , d'acceptation ou d'approbation seront déposés près le
                   Secrétaire Général du Conseil de l'Europe .
pro10_a2.2_p1-1e   786  Instruments of ratification , acceptance or approval shall be deposited with the Secretary
                   General of the Council of Europe .
pro10_a3_p1-1  787 Le présent protocole entrera en vigueur le premier jour du mois qui suit l'expiration d'une
                   période de trois mois après la date à laquelle toutes les parties à la Convention auront
                   exprimé leur consentement à être liées par le protocole conformément aux dispositions de
                   l'article 2 .
pro10_a3_p1-1e 788 This Protocol shall enter into force on the first day of the month following the expiration of
                   a period of three months after the date on which all Parties to the Convention have expressed
                   their consent to be bound by the Protocol in accordance with the provisions of Article 2 .
/.../
```

d) Extrait des arrêts de la Cour des Droits de l'Homme

475P1_1-p1-1	**3143**	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 .
475P1_1-p1-1e	**3144**	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 .
475P1_1-p1-2	**3145**	A son origine se trouve une requête dirigée contre la République d'Autriche et dont un citoyen de cet Etat , , avait saisi la Commission en vertu de l' .
475P1_1-p1-2e	**3146**	It originated in an application against the Republic of Austria lodged with the Commission under by an Austrian national , , .
475P1_1-p2-1	**3147**	La demande de la Commission renvoie aux ainsi qu'à la déclaration autrichienne reconnaissant la juridiction obligatoire de la Cour ( ) .
475P1_1-p2-1e	**3148**	The Commission's request referred to and to the declaration whereby Austria recognised the compulsory jurisdiction of the Court ( ) .
475P1_1-p2-2	**3149**	Elle a pour objet d'obtenir une décision sur le point de savoir si les faits de la cause révèlent un manquement de l'Etat défendeur aux exigences des de la Convention .
475P1_1-p2-2e	**3150**	The object of the request was to obtain a decision as to whether the facts of the case disclosed a breach by the respondent State of its obligations under of the Convention .
475P2_1-p1-1	**3151**	En réponse à l'invitation prévue à l' d) du règlement A , le requérant a exprimé le désir de participer à l'instance et a désigné son conseil ( ) .
475P2_1-p1-1e	**3152**	In response to the enquiry made in accordance with (d) of Rules of Court A , the applicant stated that he wished to take part in the proceedings and designated the lawyer who would represent him ( ) .
475P3_1-p1-1	**3153**	La chambre à constituer comprenait de plein droit , juge élu de nationalité autrichienne (de la Convention) , et , président de la Cour (b) du règlement A) .
475P3_1-p1-1e	**3154**	The Chamber to be constituted included ex officio , the elected judge of Austrian nationality (of the Convention) , and , the President of the Court ((b)) .

/.../

A.3 Inventaires des segments répétés

a) Les segments les plus fréquents du volet français du corpus *Convention*

Les segments répétés issus du volet français du corpus ***Convention*** sont triés par ordre de longueur décroissante puis par ordre de fréquence décroissante. Seuls les segments répétés de longueur supérieure ou égale à 5 sont représentés :

Longueur	Segment répété (volet français)	Fréquence
11	sur le point de savoir si les faits de la cause	33
11	si les faits de la cause révèlent un manquement de l'	33
11	obtenir une décision sur le point de savoir si les faits	32
11	puis prononcé en audience publique au palais des droits de l'	31
11	affaire a été déférée à la cour par la commission européenne	30
11	celui-ci a tiré au sort le nom des sept autres membres	28
11	et si le droit interne de ladite partie ne permet qu'	23
11	une décision prise ou une mesure ordonnée par une autorité judiciaire	23
11	une partie contractante se trouve entièrement ou partiellement en opposition avec	23
11	a son origine se trouve une requête dirigée contre la république	20
11	requérant et le délégué de la commission au sujet de l'	16
11	toute personne a droit à ce que sa cause soit entendue	15
11	du requérant et le délégué de la commission au sujet de	13
10	a tiré au sort le nom des sept autres membres	29
10	et le délégué de la commission au sujet de l'	28
10	effacer les conséquences de cette décision ou de cette mesure	23
10	a son origine se trouve une requête dirigée contre la	22
10	à la protection de la santé ou de la morale	13
10	le requérant a manifesté le désir de participer à l'	10
10	des contestations sur ses droits et obligations de caractère civil	10
10	a son origine se trouve une requête dirigée contre le	10
9	a son origine se trouve une requête dirigée contre	33
9	obtenir une décision sur le point de savoir si	33
9	in fine de la convention et du règlement a	32
9	la demande de la commission renvoie aux ainsi qu'	24
9	conformément aux # [3] de la convention et du règlement a	19
9	herbert petzold greffier au présent arrêt se trouve joint	16
9	article 50 de la convention aux termes de l'	15

[3] Le caractère # remplace les mots systématiquement absents dans le corpus.

Longueur	Segment répété (volet français)	Fréquence
9	de la convention de sauvegarde des droits de l'	13
9	le requérant a manifesté le désir de participer à	11
9	la compétence de la commission et de la cour	10
8	affaire a été déférée à la cour par	36
8	sur le point de savoir si les faits	34
8	in fine de la convention et du règlement	34
8	en sa qualité de président de la chambre	32
8	par la commission européenne des droits de l'	32
8	la chambre à constituer comprenait de plein droit	31
8	la cour avait tenu auparavant une réunion préparatoire	29
8	celui-ci a tiré au sort le nom des	29
8	si la décision de la cour déclare qu'	23
8	conformément aux de la convention et du règlement	20
8	la convention de sauvegarde des droits de l'	18
8	a manifesté le désir de participer à l'	17
8	procédure devant la commission a saisi la commission	17
8	la demande de satisfaction équitable pour le surplus	15
8	le texte intégral de son avis et des	15
8	conclusions présentées a la cour par le gouvernement	14
8	avait saisi la commission en vertu de l'	13
8	y avait invitée sur les instructions du président	13
8	le greffier a reçu le mémoire du gouvernement	12
8	adressée au secrétaire général du conseil de l'	10
7	de la convention et du règlement a	52
7	il y a eu violation de l'	49
7	la commission européenne des droits de l'	47
7	a été déférée à la cour par	37
7	y a pas eu violation de l'	36
7	dans le délai de trois mois qu'	35
7	sur le point de savoir si les	35
7	reconnaissant la juridiction obligatoire de la cour	34
7	de la commission au sujet de l'	33
7	le secrétaire général du conseil de l'	32
7	a tiré au sort le nom des	32
7	la demande de la commission renvoie aux	29
7	les débats se sont déroulés en public	24
7	accompagne figure en annexe au présent arrêt	24
7	à la partie lésée une satisfaction équitable	23
7	le texte intégral de son avis et	23
7	la cour a entendu en leurs déclarations	22
7	manifesté le désir de participer à l'	22
7	a manifesté le désir de participer à	18
7	au secrétaire général du conseil de l'	18
7	devant la commission a saisi la commission	18

Longueur	Segment répété (volet français)	Fréquence
7	saisi la commission en vertu de l'	18
7	de la convention aux termes de l'	16
7	de la loi sur la cour administrative	16
7	de la commission et de la cour	15
7	de la santé ou de la morale	14
7	le greffier a reçu le mémoire du	14
7	le premier jour du mois qui suit	12
7	ses droits et obligations de caractère civil	11
7	toute personne a droit à la liberté	11
7	la cour européenne des droits de l'	10
7	et dont un ressortissant de cet etat	10
7	saisi la commission en vertu de l'	10
6	au palais des droits de l'	61
6	il y a eu violation de	56
6	secrétaire général du conseil de l'	56
6	de la convention et du règlement	55
6	commission européenne des droits de l'	48
6	sur le point de savoir si	42
6	y a pas eu violation de	42
6	la juridiction obligatoire de la cour	41
6	sur la violation alléguée de l'	37
6	dans le délai de trois mois	36
6	fait en français et en anglais	36
6	si les faits de la cause	34
6	le nom des sept autres membres	32
6	figure en annexe au présent arrêt	32
6	le texte intégral de son avis	31
6	et le délégué de la commission	31
6	avait tenu auparavant une réunion préparatoire	30
6	le désir de participer à l'	29
6	accompagne figure en annexe au présent	25
6	et si le droit interne de	24
6	manifesté le désir de participer à	24
6	la décision de la cour accorde	23
6	etats membres du conseil de l’	22
6	de sauvegarde des droits de l'	21
6	etat défendeur aux exigences de l'	21
6	la cour de sûreté de l'	20
6	au présent arrêt se trouve joint	20
6	le greffier a reçu le mémoire	19
6	droits et obligations de caractère civil	18
6	de satisfaction équitable pour le surplus	17
6	devant les organes de la convention	17
6	la commission a retenu la requête	16

Longueur	**Segment répété (volet français)**	**Fréquence**
6	la compétence de la cour administrative	16
6	à ce que sa cause soit	16
6	cour européenne des droits de l'	15
6	entrée en vigueur du présent protocole	14
6	non-épuisement des voies de recours internes	14
6	de la procédure suivie devant elle	13
6	arrêt de la cour de cassation	13
6	le recours à la force meurtrière	12
6	le secrétaire de la commission l'	12
6	droit au respect de ses biens	12
6	usage des biens conformément à l'	12
6	à la compétence de la cour	11
6	instance et a désigné son conseil	11
6	le régime fondamental libéral et démocratique	10
6	dans le cadre de la procédure	10
6	etat défendeur doit verser au requérant	10
/.../		

b) Les segments les plus fréquents du volet anglais du corpus *Convention*

Les segments répétés issus du volet anglais du corpus ***Convention*** sont triés par ordre de longueur décroissante puis par ordre de fréquence décroissante. Seuls les segments répétés de longueur supérieure ou égale à 6 sont représentés :

Longueur	Segment répété (volet anglais)	Fréquence
11	within the three-month period laid down by # and # of the convention	35
11	and the delegate of the commission on the organisation of the	33
11	was to obtain a decision as to whether the facts of	33
11	as to whether the facts of the case disclosed a breach	33
11	the president drew by lot the names of the other seven	32
11	and delivered at a public hearing in the human rights building	31
11	the case was referred to the court by the european commission	30
11	the hearing took place in public in the human rights building	28
11	contained in the report is reproduced as an annex to this	26
11	the object of the request was to obtain a decision as	24
11	and if the internal law of the said party allows only	23
11	if the court finds that a decision or a measure taken	23
11	wished to take part in the proceedings and designated the lawyer	22
11	in accordance with of the convention and of rules of court	20
11	lawyer and the delegate of the commission on the organisation of	18
11	the convention for the protection of human rights and fundamental freedoms	17
11	registrar in accordance with of the convention and of rules of	17
11	opinions contained in the report is reproduced as an annex to	17
11	the applicant stated that he wished to take part in the	16
11	that he wished to take part in the proceedings and designated	16
11	's lawyer and the delegate of the commission on the organisation	16
11	herbert petzold registrar in accordance with of the convention and of	16
11	the secretary to the commission informed the registrar that the delegate	13
11	of the convention for the protection of human rights and fundamental	13
11	to control the use of property in accordance with the general	12
11	shall be deposited with the secretary general of the council of	12
11	herbert petzold registrar the case was referred to the court by	11
11	application of article 50 of the convention under of the convention	10
11	protocol shall enter into force on the first day of the	10
11	separate opinions contained in the report is reproduced as an annex	10
10	the president drew by lot the names of the other	33
10	wished to take part in the proceedings and designated the	30
10	the object of the request was to obtain a decision	25

Longueur	Segment répété (volet anglais)	Fréquence
10	convention for the protection of human rights and fundamental freedoms	21
10	that he wished to take part in the proceedings and	17
10	that there has been a breach of # of the convention	17
10	that there has been a violation of # of the convention	13
10	that there has been no violation of # of the convention	11
10	that the delegate would submit his observations at the hearing	10
9	the case was referred to the court by the	36
9	was to obtain a decision as to whether the	34
9	in response to the enquiry made in accordance with	32
9	's request referred to and to the declaration whereby	29
9	took place in public in the human rights building	29
9	of the convention and of rules of court a	20
9	that he wished to take part in the proceedings	19
9	that the respondent state is to pay the applicant	16
9	to the secretary general of the council of europe	15
9	with the secretary general of the council of europe	15
9	there has been a violation of # of the convention	15
9	that there has been no violation of # of the	13
9	on the first day of the month following the	12
9	there has been no violation of # of the convention	12
9	in the determination of his civil rights and obligations	11
8	the secretary general of the council of europe	52
8	as to whether the facts of the case	34
8	is reproduced as an annex to this judgment	34
8	the chamber to be constituted included ex officio	31
8	wished to take part in the proceedings and	31
8	the court had held a preparatory meeting beforehand	29
8	of the convention and of rules of court	21
8	he wished to take part in the proceedings	20
8	as requested by the registrar on the president	19
8	that there has been a breach of # of	18
8	that the respondent state is to pay the	17
8	the remainder of the claim for just satisfaction	16
8	that there has been no violation of # of	15
8	the commission informed the registrar that the delegate	14
8	proceedings before the commission applied to the commission	14
8	there has been no violation of # of the	14
8	application of article 50 of the convention under	14
8	in their memorial the government asked the court	13
8	to the court in their memorial the government	12
8	law and are necessary in a democratic society	11
8	produced the file on the proceedings before it	11
8	the member states of the council of europe	10
8	's reservation in respect of # of the convention	10

Longueur	Segment répété (volet anglais)	Fréquence
8	opinion that there had been a violation of	10
8	date of entry into force of this protocol	10
8	holds that the respondent state is to pay	10
8	final submissions to the court in their memorial	10
7	wished to take part in the proceedings	35
7	by the european commission of human rights	32
7	it originated in an application against the	32
7	recognised the compulsory jurisdiction of the court	31
7	referred to # and to the declaration whereby	30
7	pursuant to the order made in consequence	29
7	and if the internal law of the	24
7	afford just satisfaction to the injured party	24
7	allows only partial reparation to be made	24
7	that there has been a breach of	22
7	that there has been a violation of	22
7	been a violation of # of the convention	22
7	convention for the protection of human rights	22
7	protection of human rights and fundamental freedoms	22
7	application of article 50 of the convention	21
7	that the respondent state is to pay	20
7	member states of the council of europe	20
7	there has been a breach of # of	19
7	been a breach of # of the convention	18
7	there has been no violation of # of	17
7	that there has been no violation of	16
7	there has been a violation of # of	16
7	within the meaning of # of the convention	16
7	been no violation of # of the convention	15
7	the member states of the council of	14
7	the delegate of the commission did not	13
7	the file on the proceedings before it	13
7	for the protection of health or morals	13
7	expressed the opinion that there had been	13
7	that there had been no violation of	12
7	reservation in respect of # of the convention	12
7	informed the registrar that the delegate would	12
7	the second paragraph of # of protocol no	11
7	for the prevention of disorder or crime	11
7	would submit his observations at the hearing	11
7	opinion that there had been a violation	11
7	republic lodged with the commission under by	11
7	the court does not consider it necessary	10
7	of the convention in respect of the	10
7	that there had been a breach of	10

Longueur	Segment répété (volet anglais)	Fréquence
7	that there had been a violation of	10
7	that there had been no violation of	10
7	no one shall be deprived of his	10
7	opinion of is annexed to this judgment	10
6	the european commission of human rights	47
6	of the code of criminal procedure	46
6	and the delegate of the commission	46
6	a violation of # of the convention	43
6	the case was referred to the	37
6	to take part in the proceedings	37
6	in the presence of the registrar	37
6	referred to the court by the	37
6	the compulsory jurisdiction of the court	36
6	done in english and in french	36
6	a decision as to whether the	35
6	in fine of the convention and	34
6	the full text of the commission	33
6	it originated in an application against	33
6	lodged with the commission under by	31
6	had held a preparatory meeting beforehand	30
6	there has been a violation of	30
6	there has been a breach of	29
6	that there has been a violation	26
6	of human rights and fundamental freedoms	24
6	in accordance with of the convention	24
6	a breach of # of the convention	24
6	member states of the council of	24
6	the decision of the court shall	23
6	in accordance with the provisions of	21
6	the commission declared the application admissible	20
6	there has been no violation of	20
6	within the meaning of # of the	20
6	been no violation of # of the	17
6	informed the registrar that the delegate	16
6	the right to freedom of expression	15
6	holds that there has been a	15
6	the opinion that there had been	14
6	in the interests of national security	14
6	detention with a view to extradition	14
6	alleged violation of article 6 para	14
6	entry into force of this protocol	14
6	of the convention in respect of	13
6	to strike the case out of	13
6	to the court in their memorial	13

Longueur	Segment répété (volet anglais)	Fréquence
6	by the minister of the interior	13
6	there had been no violation of	13
6	opinion that there had been a	13
6	the court does not consider it	12
6	in respect of costs and expenses	12
6	the convention in respect of the	11
6	the requirements of # of the convention	11
6	to the court by the government	11
6	that there had been a violation	11
6	that there had been no violation	11
6	an appeal on points of law	11
6	no one shall be deprived of	11
6	there had been a violation of	11
6	there had been no violation of	11
6	violation of # of the convention in	11
6	period to be taken into consideration	11
6	does not consider it necessary to	11
6	the case out of the list	10
6	the european court of human rights	10
6	the government of the republic of	10
6	in respect of the first applicant	10
6	in the circumstances of the case	10
6	in the proceedings before the court	10
6	there has been no breach of	10
6	members of the council of europe	10
6	strike the case out of the	10
/…/		

Annexe B

La classification automatique et l'alignement lexical

B.1 Résultats de l'agrégation des formes et des segments répétés en classes

Les exemples regroupés dans cette annexe illustrent les modalités pratiques d'utilisation de méthode de la classification automatique pour l'analyse du corpus parallèle ***Convention*** et rendent compte de leur contribution à l'alignement lexical.

Appliquée aux formes et segments répétés issus du corpus ***Convention***, la classification automatique a dégagé un nombre important de classes qui regroupent des unités textuelles en correspondances de traduction. Il s'agit, essentiellement, des associations réalisées aux premiers niveaux de l'agrégation (cf. *Chapitre 5*). Les unités en correspondance de traduction qui se trouvent dans ces mêmes classes ont des fréquences générales et des sous-fréquences très similaires.

La visualisation de l'agrégation hiérarchisée sous forme d'un dendrogramme montre que les correspondances de traduction se trouvent parfois dans les classes voisines. La coupure du dendrogramme détermine le nombre de classes dans lesquelles on répartit l'ensemble des individus. Elle doit permettre de regrouper dans les mêmes classes, situées avant la coupure, les individus suffisamment proches. Cependant, une partition optimale est relativement difficile à obtenir. La classification ne fournit pas de critères permettant d'effectuer un découpage en classes optimal du point de vue des correspondances de traduction. Un filtrage des résultats de classification se révèle alors nécessaire [Zimina, 2000].

a) Classes de formes et segments répétées en correspondance de traduction

1. CLASSE 2692
égalité des
equality of

2. CLASSE 2862
respectives de
différence de

3. CLASSE 1103
difference in

4. CLASSE 2659
absolutely necessary
absolument nécessaire
absolument nécessaires
homicide légal

5. CLASSE 2379
responsables de l
ouvert le feu
hit the
responsables de l

6. CLASSE 2526
replies to
réponses à

7. CLASSE 2308
signatories to
dûment autorisés à cet effet
signataires du présent protocole
signatories to the convention

8. CLASSE 2267
speculate as to
spéculer sur

9. CLASSE 2265
privilèges et immunités
privileges and immunities
régulièrement sur le territoire

10. CLASSE 2320
generally recognised
généralement reconnus

11. CLASSE 2797
tenant compte du
despite the

12. CLASSE 2422
37 above
37 ci

13. CLASSE 2594
differences in
différences entre

14. CLASSE 2472
transfert de
transfert de propriété
transfer of

15. CLASSE 1576
transfer of ownership

16. CLASSE 2708
accepts that
admet que

17. CLASSE 2656
né en
born in

18. CLASSE 2966
distance of
inspecteur principal
explosive device
tunnel de landport
landport tunnel
q and
q et
winston churchill avenue
winston churchill
winston churchill

19. CLASSE 2523
côté de la
côté de
side of
side of the

20. CLASSE 2823
décision
decision

21. CLASSE 2256
the judges
des juges

Classes de formes et segments répétées en correspondance de traduction (suite)

22. CLASSE 2534
voted with the majority
voted with
voté avec
voté avec la majorité

23. CLASSE 2026
official gazette
journal officiel

24. CLASSE 2669
contributed to
contribué à

25. CLASSE 2657
considerations of
considérations d

26. CLASSE 2371
verdict d
verdict of

27. CLASSE 2848
hundred thousand
cent mille
100 000

28. CLASSE 2158
lecture des
reading out

29. CLASSE 2690
proportionnée au
proportionnée au but

30. CLASSE 2173
proportionate to
proportionate to the aim

31. CLASSE 2676
gouvernemental ou
groupe de particuliers
governmental organization or/
/group of ndividuals claiming
governmental organization or/
/group of individuals
gouvernementale ou tout groupe/
/de particuliers

32. CLASSE 2702
published in
publié dans

33. CLASSE 2973
prévention des
prevention of crime
prévention des infractions
pénales

34. CLASSE 2746
prevention of
prevention of disorder

35. CLASSE 2575
sirène de
sirène de police
tirs sur
opened fire

36. CLASSE 2965
whereas the commission
/accepted it
whereas the commission
tandis que la
tandis que la commission/
/y souscrit

37. CLASSE 2007
onze mois
eleven months

38. CLASSE 2273
trente jours
thirty days

39. CLASSE 2048
conformément à
conformément à l
in accordance with
in accordance with the

40. CLASSE 2536
objective and
objective and reasonable
objective et raisonnable

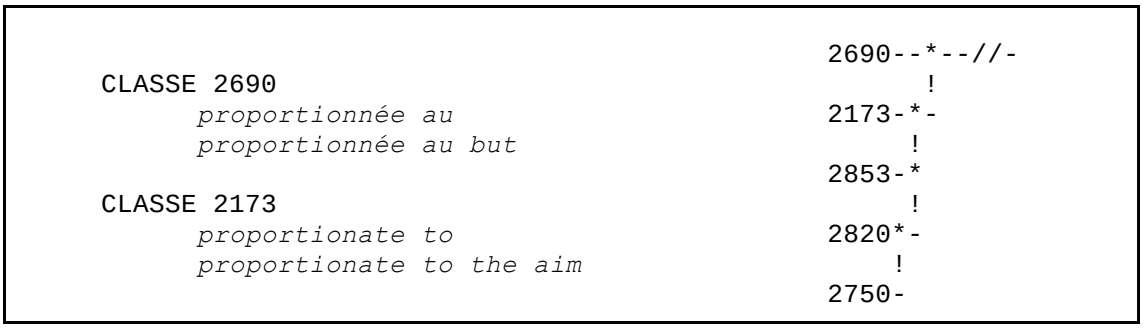
b) Retour au contexte

```
/heard addresses by,, mr jäckel, and, and replies to a question put by it. partic/
/ions,, me jäckel, et, ainsi qu' en leurs réponses à sa question. les circonstanc/

/ir basil hall, lord lester and, and also replies to questions put by one of its /
/asil hall, lord lester et, ainsi que des réponses à des questions posées par un /

/rative procedure, bgbl[ federal official gazette] no. 172/ 1950, subject to revi/
/lois de procédure administrative, bgbl.[ journal officiel fédéral], concernant l/

/rative procedure, bgbl[ federal official gazette] no. 172/ 1950, subject to revi
/lois de procédure administrative, bgbl.[ journal officiel fédéral], concernant l/

/ceive applications from any person, , non- governmental organisation or group of individuals claiming to be the/
/rsonne physique, toute organisation non- gouvernementale ou tout groupe de particuliers qui se prétend victime /

/council of europe from any person, , non- governmental organisation or group of individuals claiming to be the/
/ersonne physique, toute organisation non gouvernementale ou tout groupe de particuliers, qui se prétend victime/

/ion. the government contested this view, whereas the commission accepted it. the/
/ion. le gouvernement combat cette thèse, tandis que la commission y souscrit en /

/at the vendor exercised his right within thirty days of delivery to the purchase/
/e le vendeur exerçât ses droits dans les trente jours de la livraison à l' achet/

/ont à la charge du conseil de l' europe. privilèges et immunités des juges les j/
/shall be borne by the council of europe. privileges and immunities of judges the/

/dant l' exercice de leurs fonctions, des privilèges et immunités prévus à l' art/
/the exercise of their functions, to the privileges and immunities provided for /

/is quite distinct from the authorities' generally recognised discretion to make/
/ent distincte du pouvoir discrétionnaire généralement reconnu à l' administratio/

/ut also whether they had duly observed'' generally recognised legal and administ/
/s principes juridiques et administratifs généralement reconnus''( 45). une derni/
```

c) Profils de ventilation de segments répétés aggrégés dans les mêmes classes

réponses à	0	0	1	1	0	1	0	0	1	1	1	0	0	2	0	0	0	2	0	1
replies to	0	0	1	1	0	1	0	0	1	2	2	0	1	1	0	0	0	2	0	1
des juges	11	0	0	2	0	0	0	1	4	1	0	0	0	0	2	0	0	2	0	0
the judges	15	0	0	1	2	0	0	3	5	2	0	0	0	0	2	0	0	0	0	0
verdict d	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	6	1	0	3	0
verdict of	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	1	3	0

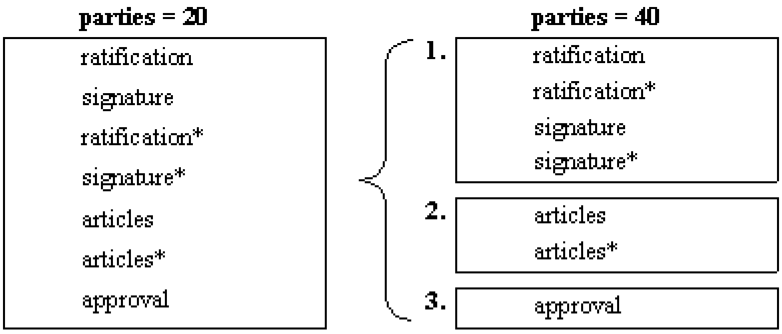
d) L'extrait du dendrogramme pour les classes voisines des correspondances



B.2 Procédures permettant d'augmenter la précision de l'appariement

a) L'affinement de la partition du corpus

Nous avons effectué une série de partition du corpus afin d'obtenir des parties de taille équivalente. Nous avons obtenu des fragments de texte consécutifs n'ayant pas d'intersection. Nous constatons que la variation du découpage influe sur la qualité des résultats. Plus la partition est fine, plus les résultats sont fiables :

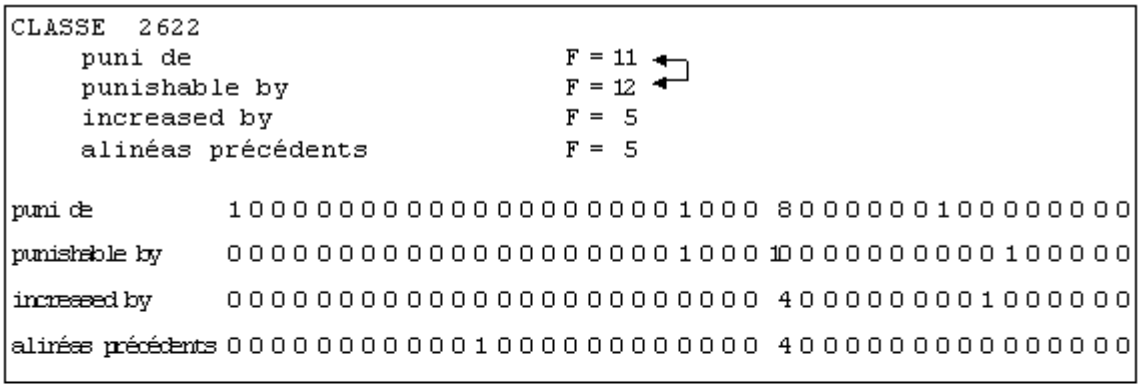


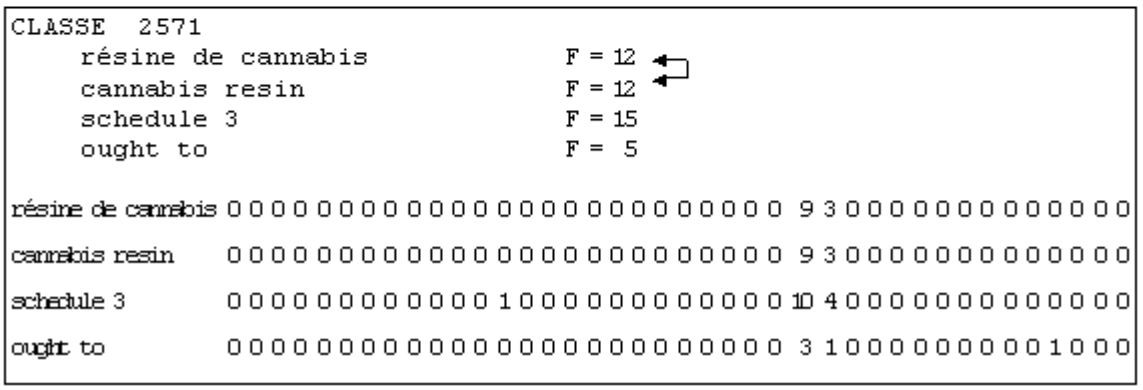
signature	14 16 5 0 1 0 0 0 2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0
signature	30 5 1 0 2 1 0 0 0 0 0 0 0 0 1 0 0 1 1 0
*signature**	13 16 5 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
*signature**	29 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
articles	15 21 1 0 0 1 0 0 0 1 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
articles	36 1 1 0 1 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0
*articles**	13 21 1 0 0 1 0 0 0 0 1 0 0 1 0 1 2 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0
*articles**	34 1 1 0 1 1 2 1 0 0 1 0 0 0 0 1 0 0 0 1

Note : L'astérisque (*) indique la forme en anglais.

b) Le filtrage des individus par comparaison de leurs fréquences totales

On augmente considérablement la qualité des résultats de la classification, si l'on prend en compte les fréquences des individus calculées dans l'ensemble du corpus. On sélectionne, par exemple, les individus dont les fréquences ne présentent pas d'écart important par rapport aux autres membres de la classe. Le filtrage sur ce paramètre regroupe avec plus de netteté les correspondances de traduction :





B.3 Lexique bilingue obtenu à partir de la classification sur les segments répétés

FRA	ANG
a	having
à 5	to 5
à partir de la date à laquelle	from the date on which
à prendre	to take
acceptation ou	acceptance or approval
adressée au secrétaire général	addressed to the secretary general
adressée au secrétaire général du conseil de l'europe	addressed to the secretary general of the council of europe
appliquent	protocol shall apply
approbation	approval
après	after
après la date	after the date
après la date à laquelle	after the date on which
arrestation	arrested
article	article
article 3	article 3
article 5	article 5
article 6	article 6
articles	articles
articles 1	articles 1
articles 1 à	articles 1 to
articles 1 à 5	of articles 1 to 5 of this protocol
articles 2	in articles 2
assemblée	assembly
au comité des ministres	to the committee of ministers
au moins	at least
au moment	at the time
au moment de la signature ou de la	at the time of signature or ratification
au secrétaire	to the secretary general of the council of europe
au secrétaire général	to the secretary general of the council of
au territoire	to the territory
autrui	of others
aux dispositions de la convention	of the provisions of the convention
aux etats intéressés	to the states concerned
ayant trait au présent protocole	relating to this protocol
candidats	candidates
ce droit	this right
chambre	chamber
chambres	chambers
cinq	five
circonstances	circumstances
collective	collective
commission	commission
concernant	concerning

FRA	ANG
condamnation	tribunal
consentement	consent
consentement à être liés par	consent to be bound by
considérant	considering
considérant que	considering that it
considérant qu'il	considering that it is
considère	it considers
consultative	consultative
convention	convention
convenus	agreed
cour européenne des droits de l'homme	european court of human rights
dans le	with a
dans une société démocratique	in a democratic society
date d'entrée en vigueur du présent protocole	date of entry into force of this protocol
day of	jour
de caractère	who is
de ce droit	exercise of this right
de cinq	of five
de guerre	of war
de la commission	of the commission
de la cour	of the court
de la signature ou	of signature or
de l'article 25	under article 25
de particuliers	of individuals
de sa	of his
de tout	any other
de trois	of three
de trois ans	of three years
de trois mois	period of three months
déclaration	declaration
déclaration faite	declaration made
dépôt de	the deposit of
dépôt de l'	of the deposit of
dérogation	derogation
des deux tiers	two-thirds
des juges	the judges
des membres	the members of the council of europe
des ministres	the committee of ministers
détention régulière	detention of a person
dix	ten
dont	from
dont ils	which they
droits	rights
du paragraphe précédent	the preceding paragraph
élu au titre	elected in respect of
en même temps	at the same time
en vigueur	into force
enseignement	teaching
entrée en vigueur du présent protocole	entry into force of this protocol
et	and
etat	state
etat dont il est le ressortissant	the territory of the state of which he is a national
etat peut	state may

FRA	ANG
etats	states
etats intéressés	the states concerned
être expulsé	be expelled
expulsions	collective expulsion of aliens
faite conformément au	a declaration made in accordance with this article shall be deemed
faits	facts
faits et	the facts and
faveur	favour
fin au	expire at the end of
grave	serious
groupes	from individuals
haute partie contractante qui a	high contracting party which
hautes parties contractantes	the high contracting parties
impartialité	impartiality
international	international law
international relations	relations internationales
introduite en application de l'artcle	application of article of the convention
judges	juges
juge	judge
la commission	the commission
la compétence de la cour	the jurisdiction of the court
la cour	the court
la cour décide si la demande d'avis	competence to give advisory opinions
la cour peut	the court may
la date	the date
la date à laquelle toutes les parties à	the date on which all parties to the
la date à laquelle toutes les parties à la convention	the date on which all parties to the convention have expressed
la haute partie contractante	the high contracting party
la juridiction obligatoire de la cour	the compulsory jurisdiction of the court
la majorité	majority
la majorité des deux tiers	of two-thirds
la peine de mort	the death penalty
l'article 32	of article 32
l'article 48	article 48
l'assemblée consultative	the consultative assembly
l'avis de la cour est transmis au comité	give advisory opinions
le juge	the judge
le juge élu	judge elected
le mandat	the terms of office
le premier jour du mois qui suit	on the first day of the month following
le rapport	report shall be
le territoire	the territory
les articles 1 à 5	the provisions of articles 1 to 5
les circonstances	the circumstances
les etats parties	states parties
les hautes parties contractantes	the high contracting parties
les hautes parties contractantes s'engagent à	the high contracting parties undertake
les membres de la commission sont élus	the members of the commission shall be elected

FRA	ANG
les membres de la cour	the members of the court shall
les mots	the words
les noms	the names of
l'etat concerné	of the state concerned
leur consentement à être liés par	their consent to be bound by
l'interprétation ou	the interpretation or application of
lois	laws
membre	a member
membres de la commission	members of the commission
membres de la cour	members of the court
même etat	of the same state
ministres	the committee of ministers
mois après la date	months after the date
mois après la date de	months after the date of
n	not
neuf	nine
noms	names
notification	notification
notification adressée au secrétaire général	a notification addressed to the secretary general
notification or communication relating to this protocol	notification ou communication ayant trait au présent protocole
notification ou déclaration ayant trait au présent protocole	notification or declaration relating to this protocol
notifiera à tous les	notify all
nul ne peut être	no one shall be
obligations	obligations
obligatoire	compulsory
ont signé	have signed
ou	or
ou en	or in
où la haute partie contractante	the high contracting party concerned has
par le	on the
par le comité des ministres	by the committee of ministers
par le secrétaire général	by the secretary general
par notification adressée au secrétaire général	notification by the secretary general
paragraphe	paragraph
paragraphe 1	paragraph 1
paragraphe 1 de	paragraph 1 of
paragraphe 1 de l'article	paragraph 1 of article
paragraphe 2	paragraph 2
paragraphes	paragraphs
paragraphes précédents	preceding paragraphs
partie	party
partie contractante	any high contracting party
partie contractante peut	any high contracting party may
parties	parties
parties à la convention	parties to the convention
parties contractantes	high contracting parties
peine de mort	death penalty
plénière	plenary
présent article	this article
présent protocole	this protocol
président	president
protocole	protocol

FRA	ANG
protocole entrera en vigueur dès	protocol shall enter into force
protocole ou	of which
publique	public
qui peuvent exprimer leur consentement à être liés	which may express their consent to be bound
ratification	ratification
règlement amiable	friendly settlement
régulièrement	lawfully
religion	religion
requérant	applicant
requête introduite	petition submitted
réserve de ratification ou d'	signature with reservation in respect of ratification or acceptance
résolus à	being resolved to
restrictions	restrictions
sans	without
se lit comme suit	shall read
se trouve	is a
secrétaire général	the secretary general
secrétaire générale du conseil de l'europe	the secretary general of the council of europe
sécurité nationale	national security
sept	seven
signature	signature
signature sans réserve de ratification	signature without reservation
signature sous réserve de ratification ou d'acceptation	signature without reservation in respect of ratification or
signé	signed
sont convenus de ce qui suit	have agreed as follows
sont élus	shall be elected
suivie	followed
suivie de ratification	followed by ratification
sur une liste de	from a list of
territoire	territory
territoire de l'etat dont il est le ressortissant	the territory of the state of which he
territoires	territories
territoriale	territorial
tout etat	any state
toute signature	any signature
travail	trial
trois	three
trois ans	three years
trois mois	a period of three
trois mois après la date	three months after the date
ultérieure	as far as possible
ultérieurement	subsequently
un ou plusieurs	one or more
une chambre	a chamber
une déclaration	a declaration
une durée	term of office
une requête introduite en vertu de l'article	a petition submitted under article
une telle	relevant
une telle peine	such penalty
victime	victim

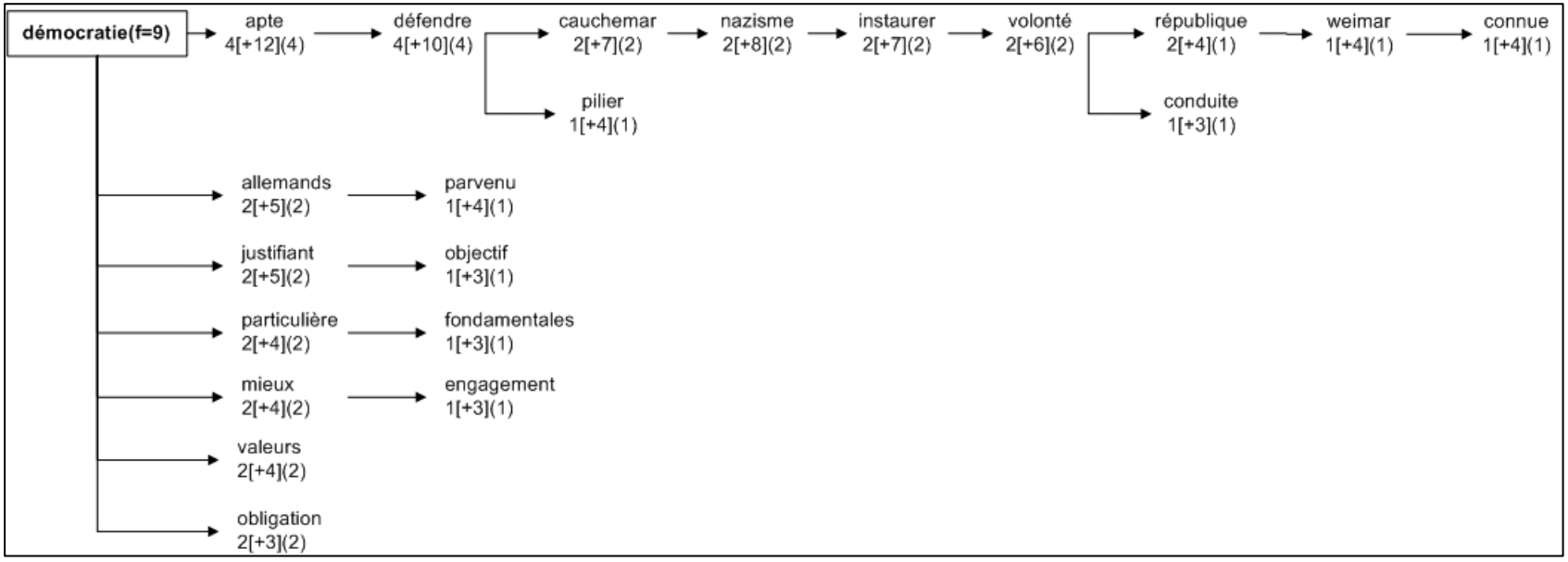
Annexe C

Utilisations des cooccurrences spécifiques pour l'alignement

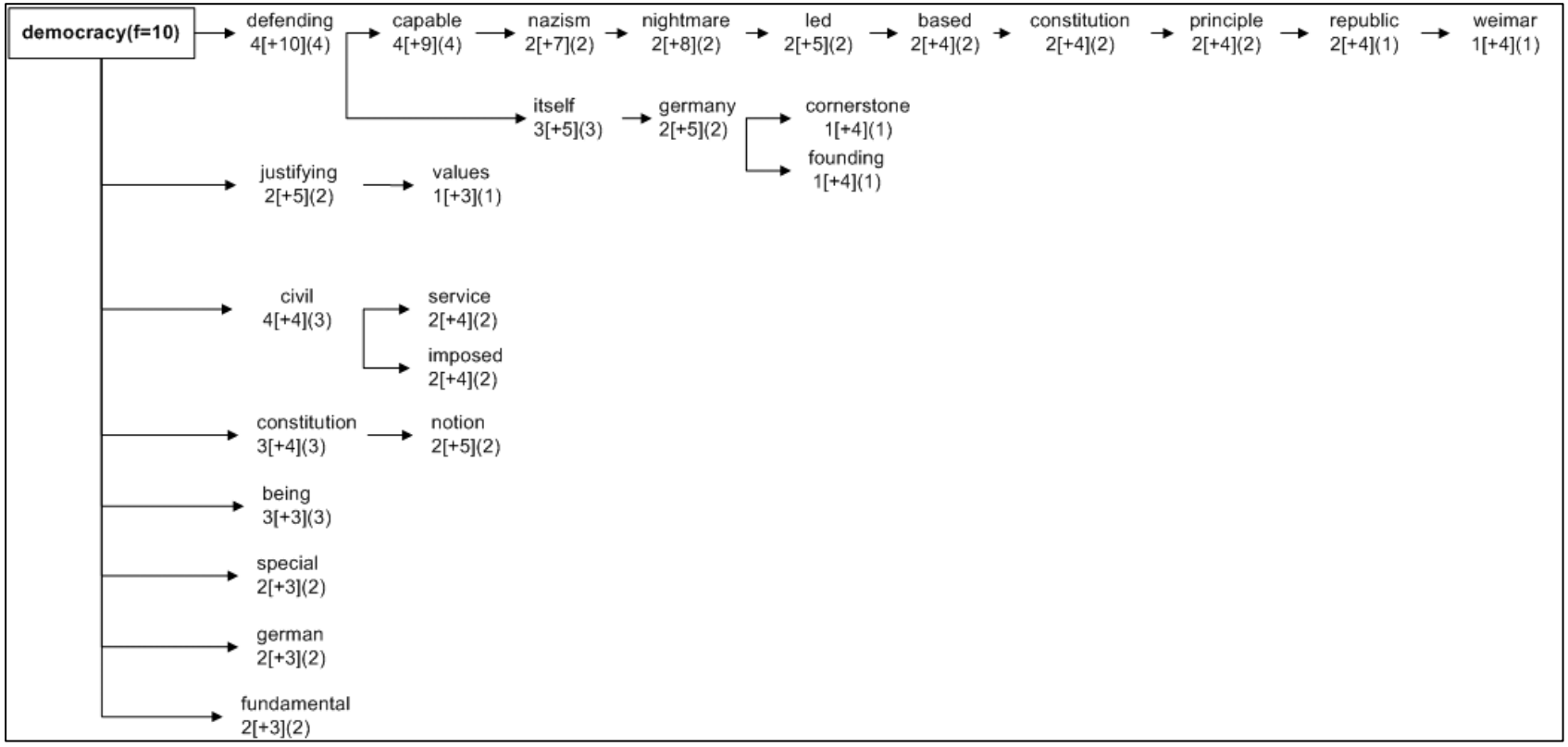
a) La liste des principaux cooccurrents spécifiques des formes *convention*/*convention* (français/anglais) dans l'unité contextuelle de la phrase

cooccurrent FRA	Fréq totale	co-fréq	indice spécif	cooccurrent ANG	Fréq totale	co-fréq	indice spécif
violation	376	166	+E47	violation	290	132	+E41
droits	418	150	+E31	breach	186	91	+E32
européenne	116	69	+E31	human	202	95	+E31
ouvrent	35	35	+E31	rights	444	152	+E30
plein	44	39	+E29	european	173	85	+E30
homme	222	93	+E25	three-month	36	35	+E30
révèlent	47	38	+E25	institutions	52	42	+E28
fine	42	36	+E25	freedoms	74	48	+E24
comprenait	38	34	+E25	officio	40	34	+E24
règlement	284	107	+E24	holds	108	56	+E22
décision	508	157	+E24	object	60	41	+E22
partie	291	108	+E24	nationality	51	36	+E21
libertés	78	49	+E24	names	48	35	+E21
contractante	75	48	+E24	lot	46	34	+E21
déférée	49	38	+E24	allows	30	27	+E21
manquement	69	44	+E22	article	351	112	+E20
organes	58	40	+E22	ex	47	34	+E20
élu	40	33	+E22	drew	44	33	+E20
effacer	24	24	+E21	partial	29	26	+E20
imparfaitement	24	24	+E21	constituted	78	44	+E19
turquie	111	55	+E20	disclosed	50	34	+E19
constituer	51	36	+E20	injured	28	25	+E19
lésée	25	24	+E20	protocol	264	90	+E18
protocole	263	93	+E19	reparation	41	29	+E17
exigences	105	53	+E19	internal	31	25	+E17
déclare	49	34	+E19	contracting	178	67	+E16
savoir	209	79	+E18	obligations	102	47	+E16
découlant	29	25	+E18	elected	57	34	+E16
tiré	83	44	+E17	afford	36	26	+E16
protocoles	19	19	+E17	protocols	18	18	+E16
article	362	109	+E16	turkey	122	52	+E15
sort	54	33	+E16	presence	63	35	+E15
fondamentales	33	25	+E16				
application	223	77	+E15				
nationalité	60	34	+E15				
partiellement	36	26	+E15				
sauvegarde	31	24	+E15				

b) Vue partielle du réseau de cooccurrences élaboré à partir du pôle *démocratie*



c) Vue partielle du réseau de cooccurrences élaboré à partir du pôle *democracy*



d) Les contextes spécifiques où se réalisent les réseaux de cooccurrences calculés à partir des pôles *démocratie* – *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 .

Annexe D

L'étiquetage morphosyntaxique du corpus *Convention* par *TreeTagger*

D.1 Jeu d'étiquettes utilisé pour le volet français du corpus Convention

Etiqueteur : **TreeTagger** [Schmid, 1994]

Références : http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html

ABR	abréviation
ADJ	adjective
ADV	adverbe
DET:ART	déterminant
DET:POS	pronom possessif(*ma*, *ta*, ...)
INT	interjection
KON	conjonction
NAM	nom propre
NOM	nom
NUM	nombre
PRO	pronom
PRO:DEM	pronom démonstratif
PRO:IND	pronom indéfini
PRO:PER	pronom personnel
PRO:POS	pronom possessif(*mien*, *tien*, ...)
PRO:REL	pronom relatif
PRP	préposition
PRP:det	préposition contractée(*au*,*du*,*aux*,*des*)
PUN	ponctuation
PUN:cit	ponctuation de citation
SENT	phrase
SYM	symbole
VER:cond	verbe au conditionnel
VER:futu	verbe au futur
VER:impe	verbe à l'impérative
VER:impf	verbe à l'imparfait
VER:infi	verbe à l'infinitive
VER:pper	verbe au participe passé
VER:ppre	verbe au participe présent
VER:pres	verbe au présent
VER:simp	verbe au passé simple
VER:subi	verbe au subjonctive imparfait
VER:subp	verbe au subjonctive présent

D.2 Jeu d'étiquettes utilisé pour le volet anglais du corpus Convention

Etiqueteur : **TreeTagger** [Schmid, 1994]

Références : Santorini [1990]

CC	Coordinating Conjunction
CD	Cardinal Number
DT	Determiner
EX	Existential *_there*
FW	Foreign word
IN	Preposition or Subordinating Conjunction
JJ	Adjective
JJR	Adjective comparative
JJS	Adjective superlative
LS	List item marker
MD	Modal Verb
NN	Noun singular or uncountable
NNS	Noun plural
NP	Proper Noun singular
NPS	Proper Noun plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP	Possessive Pronoun
RB	Adverb
RBR	Adverb comparative
RBS	Adverb superlative
RP	Particle
PROBLEMATIC CASES	
SYM	Symbol
TO	*to*
UH	Interjection
VV/VB	Verb base form
VBD	Verb past tense
VBG	Verb gerund or present participle
VBN	Verb past participle
VBP	Verb non 3rd person singular present
VBZ	Verb 3rd person singular present
WDT	*Wh*_determiner
WP	*Wh*_pronoun
WP	Possessive *wh*_pronoun
WRB	*Wh*_adverb

D.3 Extrait du corpus Convention *étiqueté*

```
<texte="fr">§ Les gouvernements signataires , membres du Conseil de l'Europe ,
<cat="fr">§ DET:ART NOM NOM PUN NOM PRP:det NOM PRP DET:ART NAM PUN
<lemm="fr">§ le gouvernement signataire , membre du conseil de le Europe ,
<texte="en">§ The governments signatory hereto , being members of the Council of Europe ,
<cat="en">§ DT NNS NN RB , JJ NNS IN DT NP IN NP ,
<lemm="en">§ the government signatory hereto , being member of the Council of Europe ,

<texte="fr">§ Considérant la Déclaration universelle des Droits de l'Homme , proclamée par l'Assemblée générale
des Nations Unies le 10 décembre 1948 ;
<cat="fr">§ VER:ppre DET:ART NOM ADJ PRP:det NOM PRP DET:ART NAM PUN VER:pper PRP DET:ART NOM ADJ
PRP:det NOM VER:pper DET:ART NUM NOM NUM :
<lemm="fr">§ considérer le déclaration universel du droit de le Homme , proclamer par le assemblée
général du nation unir le @card@ décembre @card@ ;
<texte="en">§ Considering the Universal Declaration of Human Rights proclaimed by the General Assembly of the
United Nations on 10th December 1948 ;
<cat="en">§ VVG DT NP NP IN NP NPS VVD IN DT NP NP IN DT NP NPS IN JJ NP CD :
<lemm="en">§ considering the Universal Declaration of Human Rights proclaim by the General Assembly
of the United Nations on 10th December @card@ ;

<texte="fr">§ Considérant que cette déclaration tend à assurer la reconnaissance et l'application universelles et
effectives des droits qui y sont énoncés ;
<cat="fr">§ VER:ppre KON PRO:DEM NOM VER:pres PRP VER:infi DET:ART NOM KON DET:ART NOM ADJ KON ADJ
PRP:det NOM PRO:REL PRO:PER VER:aux:pres VER:pper :
<lemm="fr">§ considérer que ce déclaration tendre à assurer le reconnaissance et le application
universel et effectif du droit qui y être énoncer ;
<texte="en">§ Considering that this Declaration aims at securing the universal and effective recognition and
observance of the Rights therein declared ;
<cat="en">§ VVG IN DT NP VVZ IN VVG DT JJ CC JJ NN CC NN IN DT NPS RB VVD :
<lemm="en">§ considering that this Declaration aim at secure the universal and effective recognition
and observance of the Rights therein declare ;
/.../
```

Annexes électroniques sur le Cd-rom

Table des matières :

Ce Cd-rom présente les documents annexes aux travaux de recherche exposés dans ce volume. Les documents électroniques y sont regroupés au sein des rubriques distinctes référencées dans le fichier **index.html**. Ce fichier permet d'avoir accès à l'ensemble des informations enregistrées sur le Cd-rom[1]. Par exemple, si l'on sélectionne le lien correspondant à la rubrique **Corpus parallèle *Convention***, on a accès au texte intégral du corpus en français et en anglais. Trois versions électroniques du corpus ***Convention*** sont accessibles sur le Cd-rom :

- version initiale au format TXT (corpus aligné au niveau de la phrase)
- version catégorisée au format TXT
- version catégorisée au format XML (avec la DTD correspondante).

Les fichiers contenant les listes des *segments répétés* et celles des *patrons syntaxiques* récurrents du corpus ***Convention*** sont également disponibles sur le Cd-rom.

La rubrique **Présentations** permet de consulter des *diapositives* qui nous ont servi à faire des exposés sur les thèmes des différents chapitres de ce travail de thèse. Les documents sont présentés sous forme de fichiers *PowerPoint* avec des animations permettant de simuler le fonctionnement des outils d'exploration textométrique de corpus parallèles proposés dans la thèse.

Les rubriques **Publications** et **TPs « Corpus Parallèles »** concernent les articles et les cours consacrés à l'étude des corpus parallèles que nous avons réalisés durant nos recherches.

Le lien vers le site du groupe **GADT – *Textométrie multilingue*** à l'animation duquel nous participons, donne une vision sur des recherches actuelles dans le domaine de l'analyse quantitative des données textuelles multilingues. On y trouvera des publications récentes sur l'étude de corpus multilingues à l'aide des méthodes de la statistique textuelle.

La dernière rubrique **CyberBibliographie** regroupe des liens vers les sites Internet consacrés aux publications, projets et outils liés aux corpus parallèles multilingues.
Les pages qui suivent montrent des extraits de ces différentes annexes.

---

[1] La connexion à Internet est requise pour consulter certains liens enregistrés sur le Cd-rom.

/index.html

La page d'accueil

Doctorat en Sciences du langage :
Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles
soutenu le 26 novembre 2004

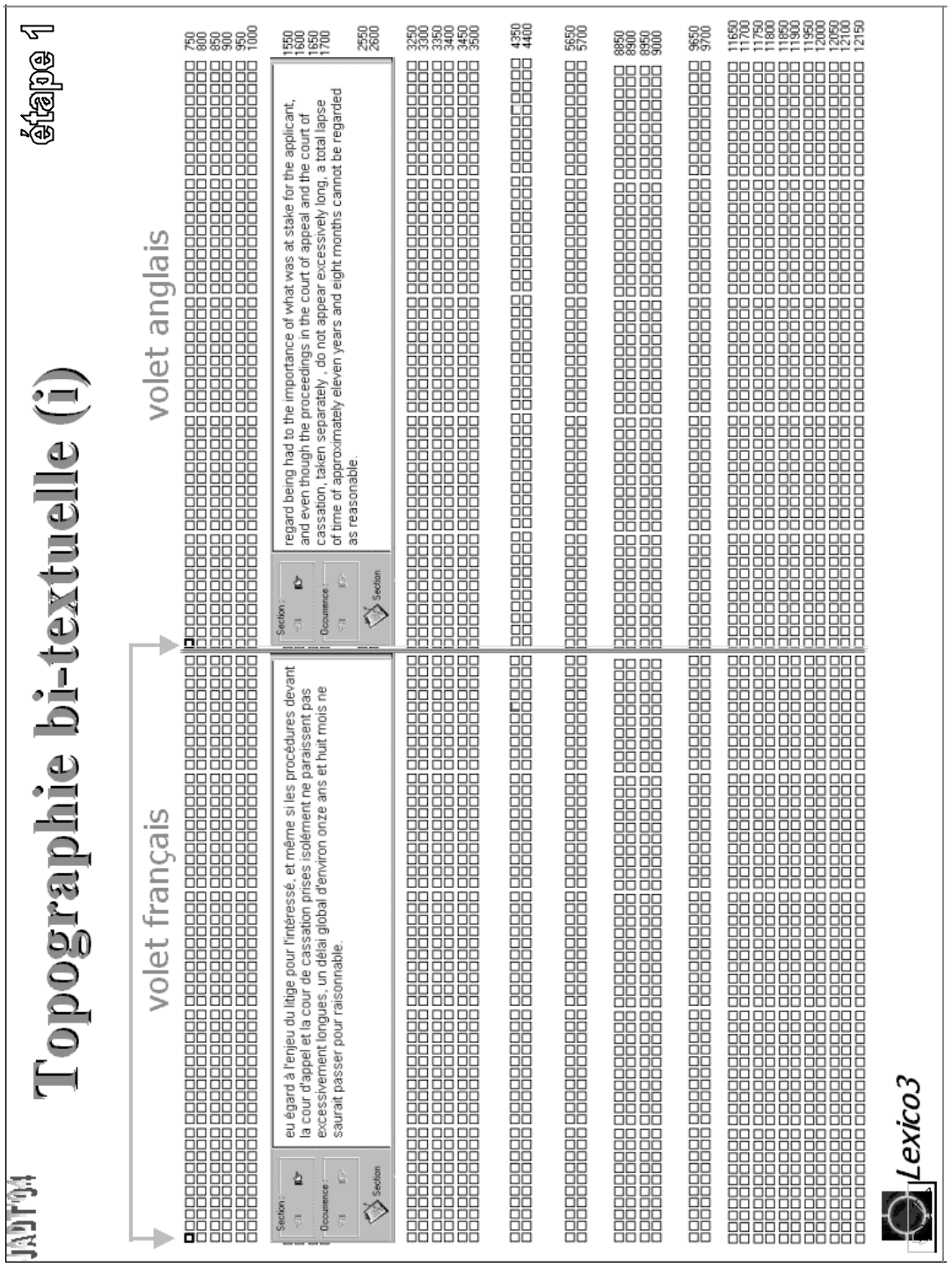
Maria Zimina-Poirot

Corpus parallèle CONVENTION

version initiale (format TXT)	Le corpus est aligné au niveau de la phrase. Chaque couple de phrases équivalentes est introduit par un code. Les numéros indiquent (dans l'ordre) : le type de document (*convention*, *protocole*, etc.) ; le numéro de l'arrêt ; la partie de l'arrêt ; le numéro de section et/ou du paragraphe ; le numéro de la phrase dans le corpus, précédé par la lettre "e" pour les phrases en anglais. Segments répétés français. Segments répétés anglais.
version catégorisée (format TXT)	Le corpus est catégorisé avec TreeTagger. Le jeu d'étiquettes utilisé pour le volet français est disponible ici. Le jeu d'étiquettes utilisé pour le volet anglais est disponible ici. Patrons syntaxiques français. Patrons syntaxiques anglais.
version catégorisée (format XML)	Le corpus catégorisé est balisé au format XML. Le schéma de la structure formelle du corpus balisé est disponible ici. Sur le schéma, chaque alignement phrastique est référencé par un élément *P* (partie) de l'arbre XML. Cet élément contient deux sous-éléments : la phrase en français et sa traduction en anglais ; ces mêmes sous-éléments contiennent aussi la version étiquetée de la partie visée : *forme-catégorie-lemme*. La DTD (Document Type Defintion) du corpus balisé est disponible ici.

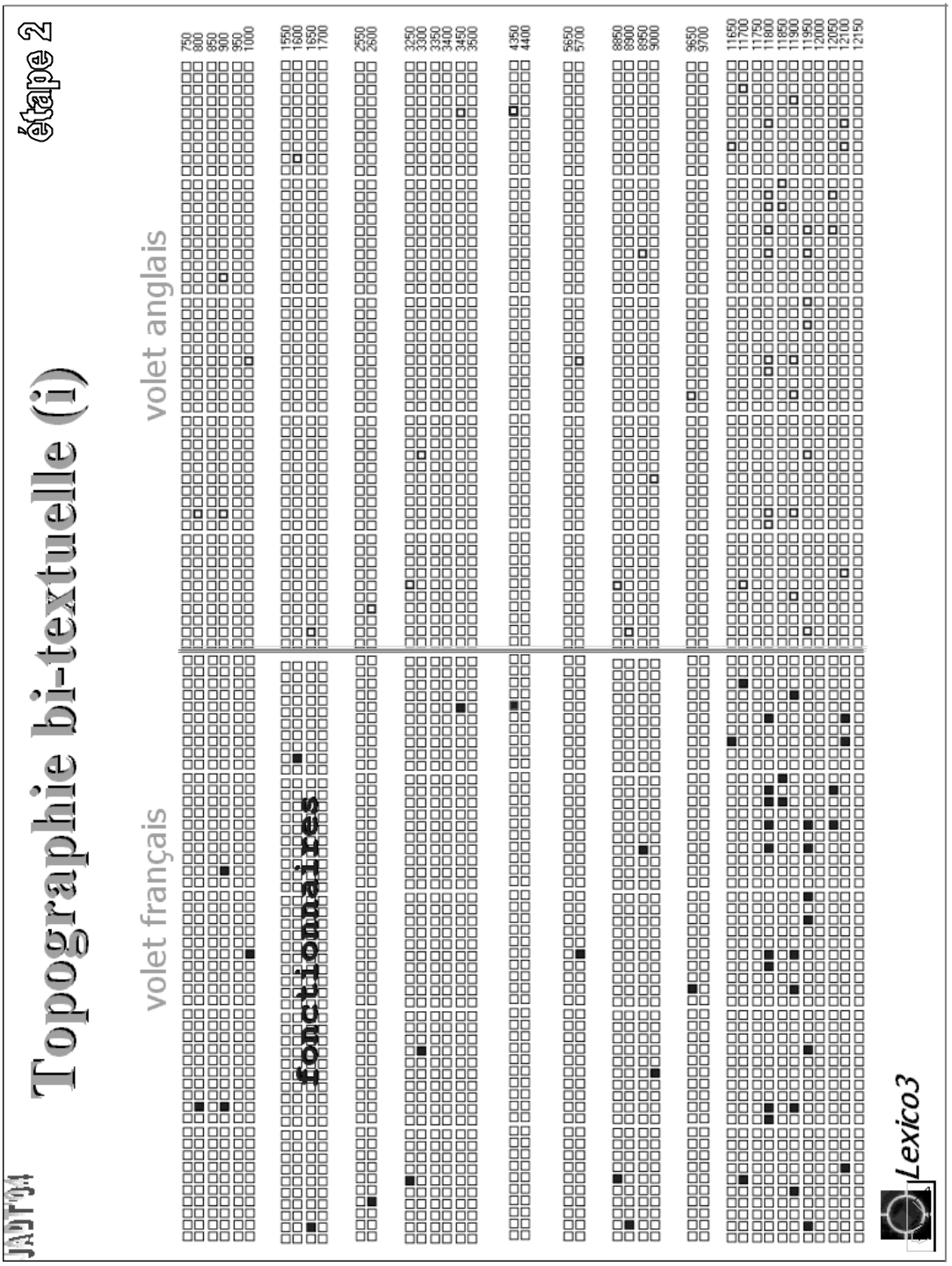
Maquette de visualisation bi-textuelle : étape 1

/fichierCD/stmz/page5_fichiers/JADT_2004F.ppt



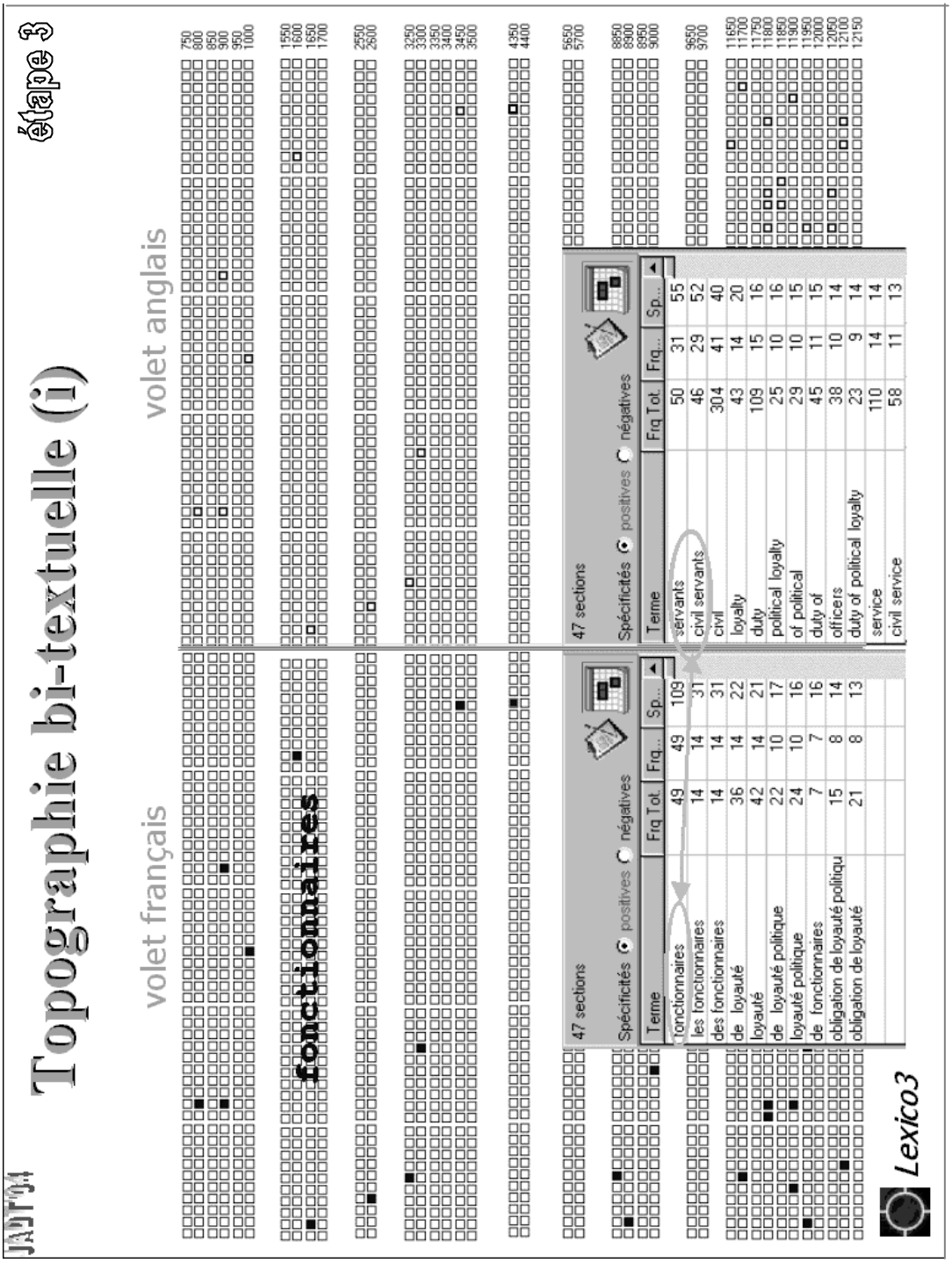
Maquette de visualisation bi-textuelle : étape 2

/fichierCD/stmz/page5_fichiers/JADT_2004F.ppt



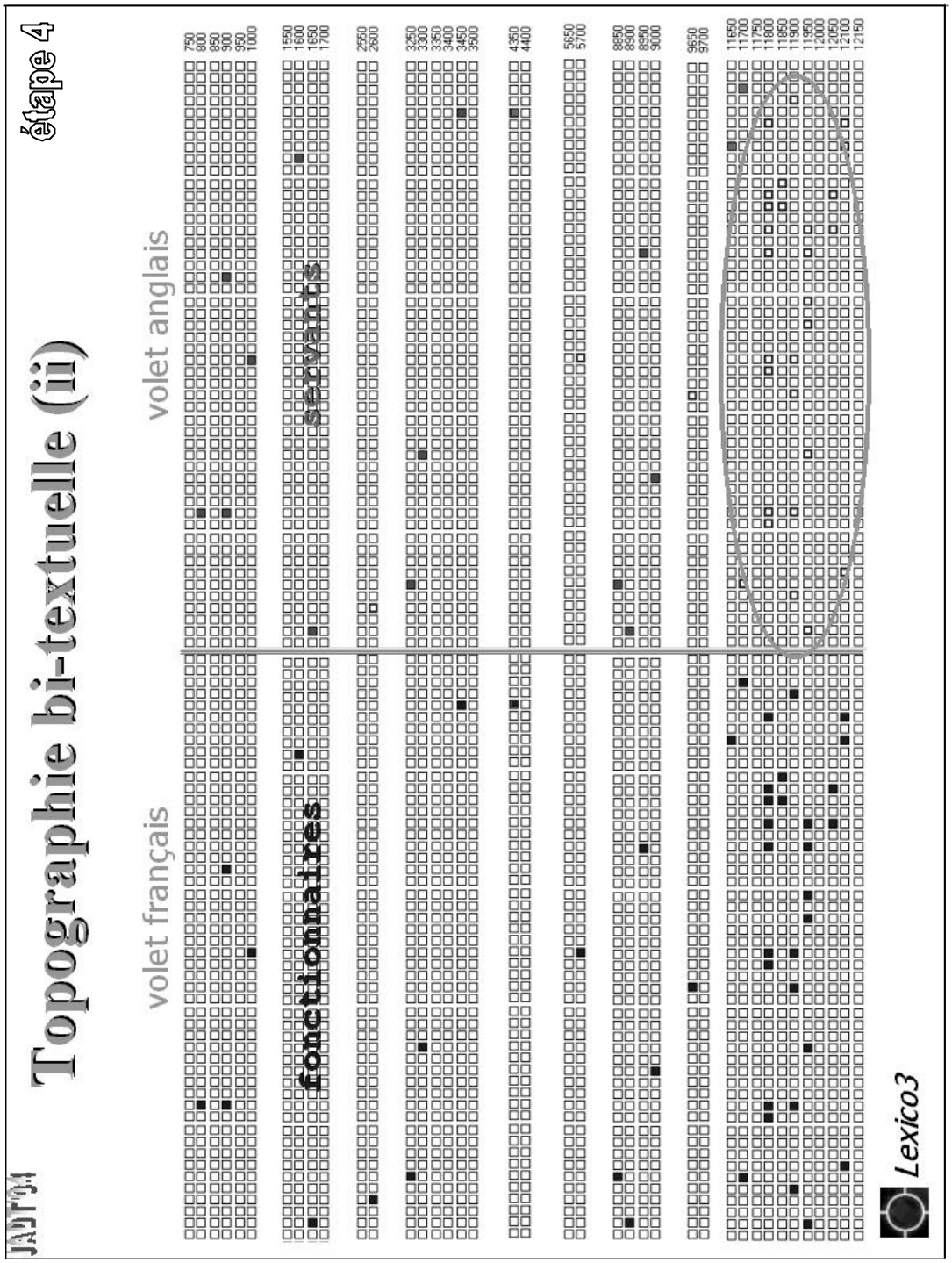
Maquette de visualisation bi-textuelle : étape 3

/fichierCD/stmz/page5_fichiers/JADT_2004F.ppt



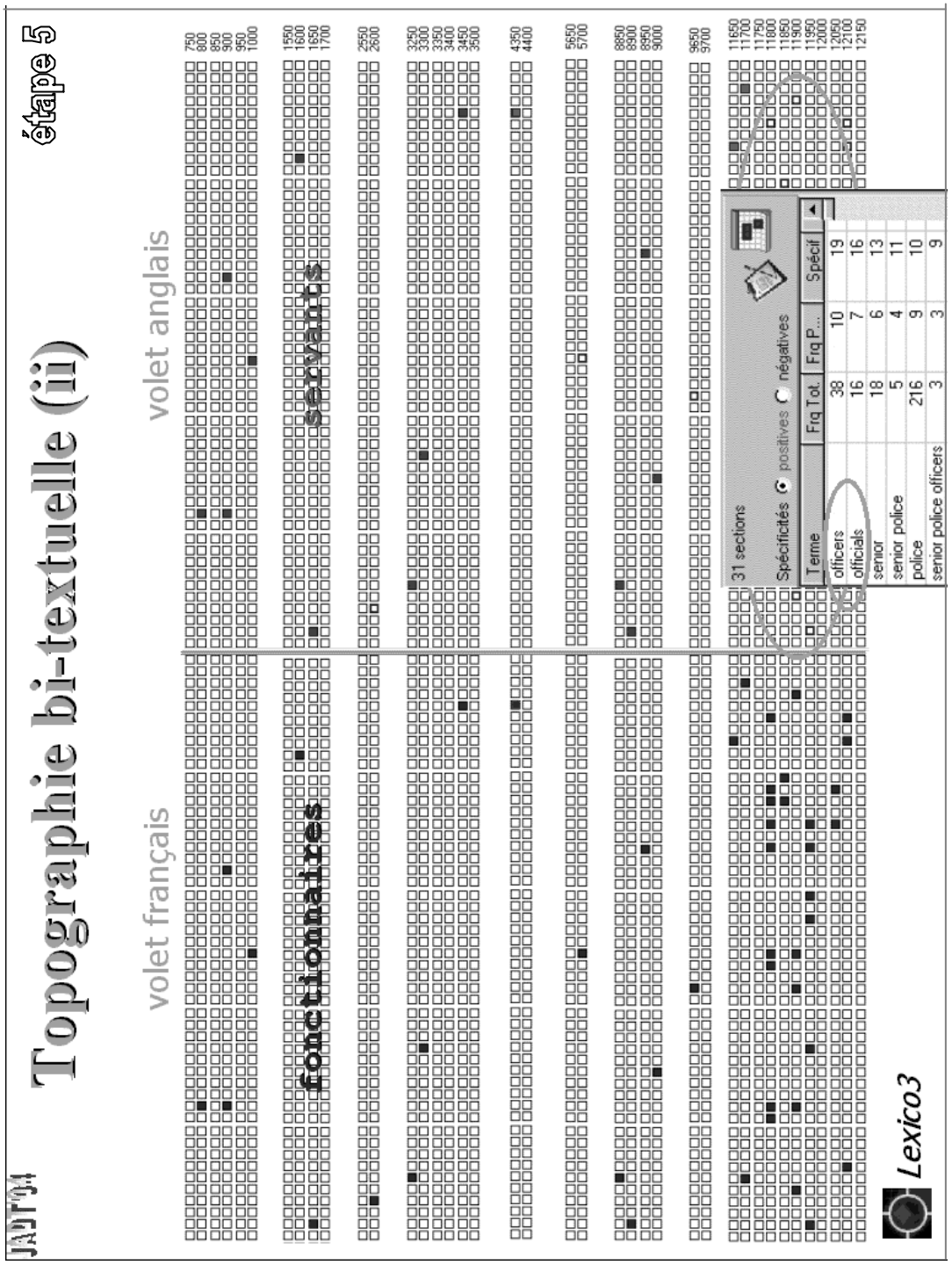
Maquette de visualisation bi-textuelle : étape 4

/fichierCD/stmz/page5_fichiers/JADT_2004F.ppt



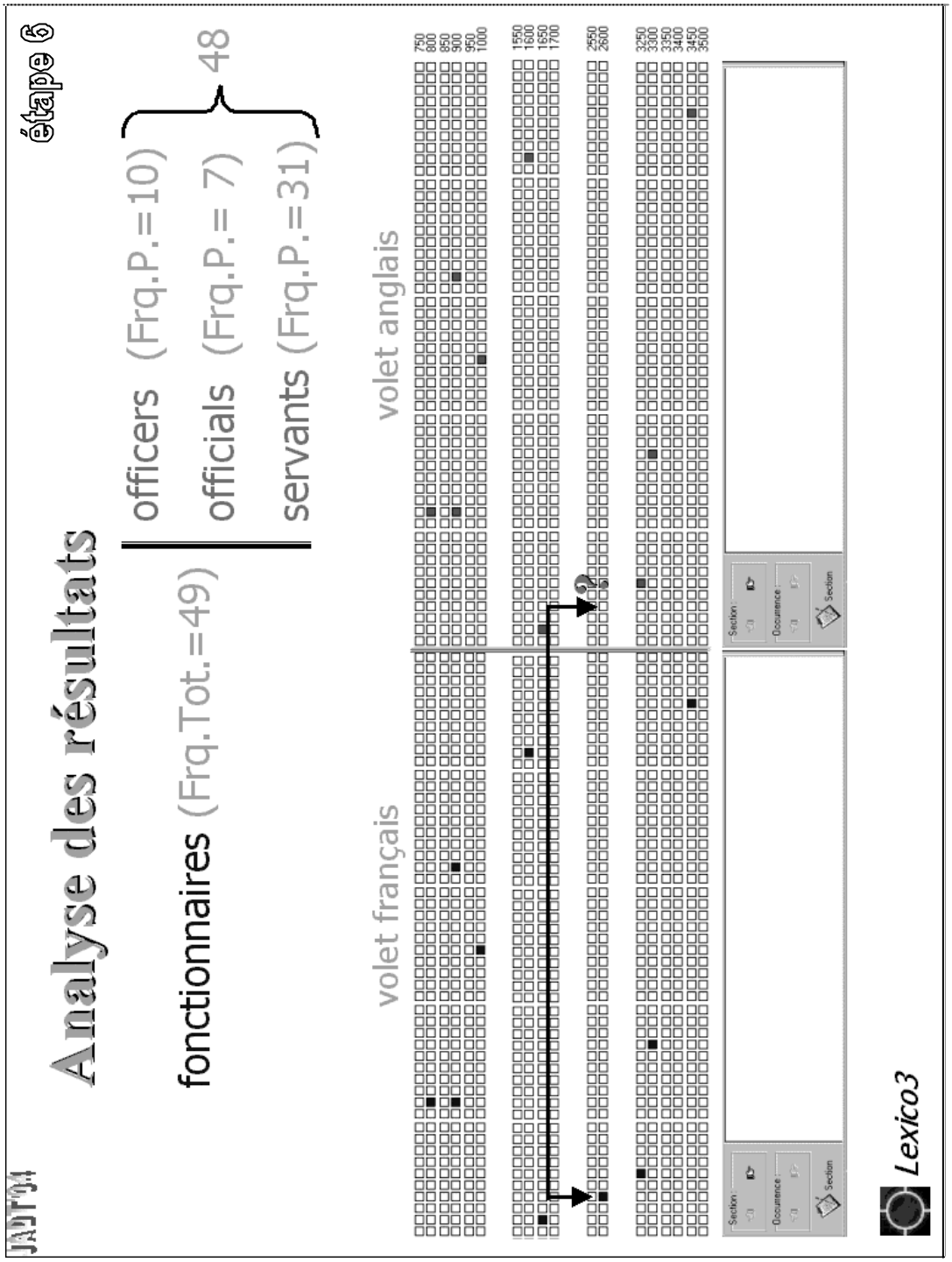
Maquette de visualisation bi-textuelle : étape 5

/fichierCD/stmz/page5_fichiers/JADT_2004F.ppt



Analyse des résultats : étape 6

/fichierCD/stmz/page5_fichiers/JADT_2004F.ppt



JADT'04

étape 7

Analyse des résultats

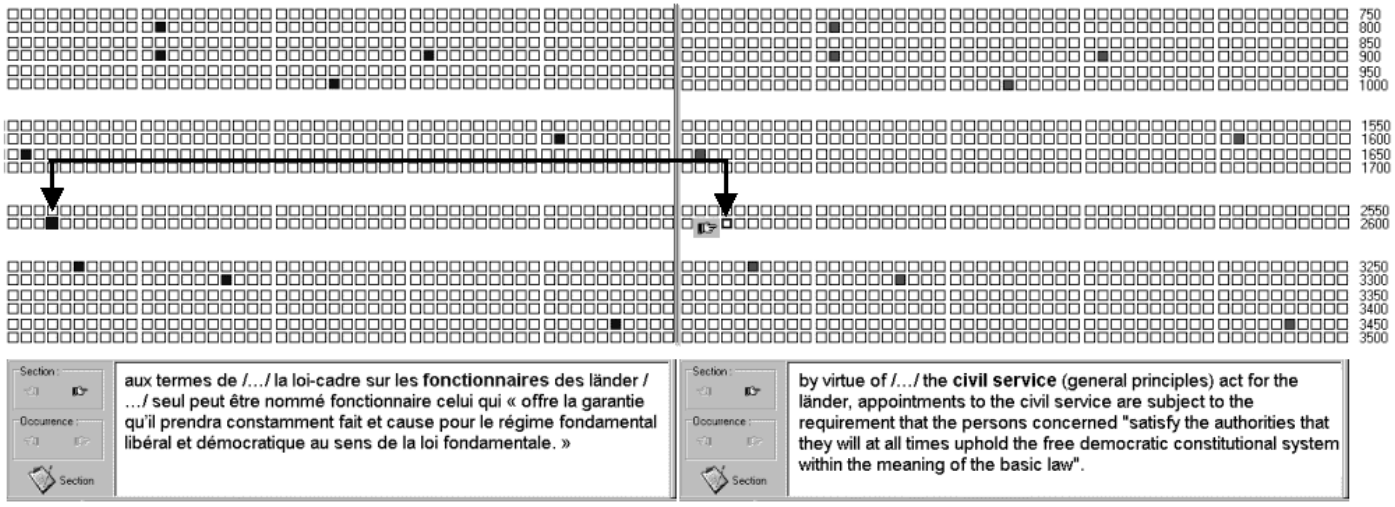
fonctionnaires (Frq.Tot.=49)

- officers (Frq.P.=10)
- officials (Frq.P.= 7)
- servants (Frq.P.=31)

} 48

volet français

volet anglais



Lexico3

Les captures d'écran que nous venons de présenter concernent les maquettes de logiciels d'exploration textométrique intertextuelle que nous avons proposées. La réalisation informatique de ces fonctionnalités est actuellement en cours au sein du Centre de Lexicométrie et d'Analyse Automatique des Textes (CLA2T – SYLED) de l'Université de la Sorbonne nouvelle – Paris 3.

Etablissement :

Université de la Sorbonne nouvelle – Paris 3

Domaine :

Sciences du langage

Etudiant :

Maria ZIMINA-POIROT

Directeur :

André SALEM, Professeur

Titre :

Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles

Résumé :

Ce travail présente les résultats d'une série de recherches consacrées au développement d'une nouvelle famille d'outils d'exploration textométrique intertextuelle. De nombreuses méthodes de statistique textuelle ont été articulées et adaptées au cadre multilingue : la méthode des *segments répétés*, les *spécificités*, la *topographie bi-textuelle*, les *cooccurrences multiples*, l'*analyse factorielle des correspondances,* la *classification automatique*, etc. L'utilisation de chaque méthode dans le contexte multilingue est illustrée par des exemples d'applications concrètes, accompagnés d'échantillons de ressources traductionnelles obtenues à partir du corpus parallèle français/anglais de la *Convention de sauvegarde des Droits de l'Homme*. Les perspectives ouvertes par cette approche offrent aux traducteurs, enseignants en langues étrangères, terminologues, lexicographes, etc., des moyens automatisés pour explorer la structure des équivalences lexicales dans les corpus de traduction.

Mots-clés :

alignement, bi-texte, corpus parallèles, correspondances traductionnelles, statistique textuelle, textométrie, topographie textuelle.

Title :

Quantitative approaches of extracting translation resources from parallel corpora

Summary :

This research work presents the results of a series of experiments devoted to the development of new tools for intertextual textometric exploration of translation corpora. Various methods of textual statistics have been adapted for use in a multilingual context and put into practice for parallel text processing, such as: *repeated segments extraction*, *characteristic elements computation*, *bi-textual topography*, *multiple co-occurrences*, *factorial analysis*, *automatic classification*, etc. Examples of concrete applications illustrate the use of each of these methods in a multilingual context. These examples are accompanied by sample translation resources obtained on quantitative bases from the parallel French/English corpus of the *Convention for the Protection of Human Rights*. The suggested approach opens up new horizons for automatic exploration of lexical equivalences of translation corpora by a variety of users: translators, foreign language teachers, terminologists, lexicographers, etc.

Key words :

alignment, bi-text, parallel corpora, textometrics, textual statistics, textual topography, translation correspondences.