Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné

Cet article présente une expérience de construction d'un lexique français/anglais des droits de l'homme à partir de l'analyse automatique d'un corpus bilingue constitué d'arrêts rendus par la Cour européenne des droits de l'homme de Strasbourg. Ce corpus a été aligné automatiquement au niveau des phrases à l'aide d'heuristiques simples exploitant la structure logique des arrêts, identique dans les deux langues. Le logiciel *Lexter* a extrait des candidats termes de la partie française du corpus. Les juristes terminologues ont construit le lexique en repérant dans les phrases anglaises les équivalents des candidats termes français jugés pertinents.

Termes-clés :
extraction de terminologie, alignement terminologique, lexique bilingue, droits de l'homme

1 Introduction

Les systèmes de mémoire de traduction rencontrent un succès grandissant. Il est admis que leur utilité n'est avérée que dans les situations où la quantité de documents à traduire sur un même domaine est très importante. La présence de tels systèmes ne condamne donc pas le recours aux terminologies bilingues. Mémoire de traduction et terminologies multilingues sont des outils complémentaires dans l'usage que le traducteur peut en faire. Ces outils peuvent aussi être complémentaires dans leur mode d'élaboration réciproque. Ainsi, une terminologie bilingue peut être exploitée pour aligner un corpus parallèle, de même que, réciproquement, un corpus aligné peut être exploité pour (re)construire une terminologie bilingue. Dans cet article, nous exposons comment nous élaborons un lexique bilingue des droits de l'homme, à partir de l'analyse (semi-automatique) d'un corpus bilingue français/anglais, aligné au niveau des phrases, constitué d'un ensemble d'arrêts rendus par la Cour européenne des droits de l'homme de Strasbourg. Nous présentons dans la section 2 la problématique générale de l'alignement terminologique dans les recherches en traitement automatique des langues. Dans la section 3, nous posons le cadre générale du projet «lexique multilingue des droits de l'homme», auquel collaborent juristes, terminologues et linguistes informaticiens. Nous décrivons dans la section 4 les traitements informatiques effectués, concernant l'alignement du corpus et l'extraction terminologique automatique sur la partie française du corpus, et nous présentons une analyse quantitative des premiers résultats obtenus.

2 Alignement de terminologie

2.1 Alignement de phrases, de mots, de termes

Dans le domaine de la recherche en traitement automatique des langues (TAL), le thème de l'alignement multilingue suscite un grand nombre de travaux depuis plusieurs années. On s'est intéressé d'abord à l'alignement de phrases dans un corpus parallèle, c'est-à-dire un corpus bilingue dont l'une des parties est une traduction de l'autre. L'objectif est de construire des couples de phrases, extraites de chacune des parties, qui soient les traductions l'une de l'autre. On élabore ainsi un corpus aligné. Différents algorithmes et différentes techniques de type statistique ou linguistique, s'appuyant sur le niveau lexical ou sur celui des caractères, sont mis en œuvre (Brown *et al.* 1995, Church 1993, Gale et Church 1993). À partir d'un corpus aligné au niveau des phrases, on peut chercher à aligner des mots, et donc à construire des lexiques bilingues (Dagan *et al.* 1993). Les opérations d'alignement de phrases et d'alignement de mots sont souvent

interdépendantes, puisqu'on peut s'appuyer sur des couples de mots pour identifier des associations de phrases, et réciproquement, sur des couples de phrases pour identifier des associations de mots.

Depuis quelques années, les efforts portent sur l'alignement de termes. L'objectif est, à partir d'un corpus aligné au niveau des phrases, de construire non seulement des couples de mots simples, mais aussi des couples de séquences de mots, qu'elles soient désignées sous le nom de (candidats) termes, de collocations ou de syntagmes nominaux (Smadja et McKeown 1994, Kupiec 1993). Pour l'alignement de termes, parmi tous les choix méthodologiques à opérer pour élaborer un système d'alignement, l'un concerne l'extraction monolingue: dans certains travaux, par exemple Hull (1998), l'extraction de candidats termes est effectuée indépendamment sur chacun des deux corpus, et les candidats termes automatiquement extraits sont alors appariés à l'aide d'algorithmes basés sur des principes analogues à ceux adoptés pour l'alignement de mots; dans d'autres travaux, par exemple Gaussier (1998), l'extraction automatique est effectuée uniquement sur l'un des deux corpus, et les algorithmes mis en œuvre réalisent de façon conjointe les tâches d'extraction des termes dans l'autre corpus et d'appariement avec les termes extraits automatiquement du premier corpus.

2.2 Une expérience d'alignement de terminologie sans appariement statistique

Dans l'expérience que nous décrivons ci-dessous, notre démarche a été la suivante: à partir d'un corpus bilingue français/anglais, aligné au niveau des phrases, nous avons d'abord effectué une extraction automatique de candidats termes sur la partie française du corpus; ensuite, ce sont les juristes terminologues eux-mêmes qui sont allé chercher les équivalents anglais dans les contextes. Cette démarche est rendue possible, et efficace, grâce à une interface de validation spécialement conçue pour cette tâche, dans laquelle le juriste terminologue accède directement à l'affichage des couples de phrases, pour lesquels le candidat terme en cours d'analyse est présent dans la phrase française (*cf.* figure 1).

Sur le plan technique et algorithmique, notre démarche est donc «pauvre», puisqu'aucun appariement statistique n'est réalisé. Cependant, rien ne prouve qu'elle ne supporte pas la concurrence avec des techniques plus sophistiquées, qui seraient en mesure de proposer des appariements pour les candidats termes les plus fréquents. Se pose ici le problème de l'évaluation en ingénierie linguistique. Toutes les techniques d'alignement de terminologie sont présentées par leurs auteurs comme étant susceptibles d'apporter une aide à un utilisateur humain chargé de construire une terminologie bilingue. Les modes d'évaluation systématiquement évoqués par ces auteurs font appel aux notions de taux de précision (le plus souvent) et de taux de rappel (parfois). Pour mesurer le taux de précision, on évalue la proportion de couples corrects dans une liste de couples extraits automatiquement. Nous estimons que ce paramètre, certainement utile à un moment donné de l'élaboration du système, n'est pas le paramètre à prendre en compte de façon prioritaire pour évaluer les systèmes d'extraction multilingue. Le problème est identique à celui de l'extraction monolingue (*cf.* Bourigault et Habert 1998): puisque l'objectif n'est pas une extraction automatique, mais une aide efficace à l'utilisateur, ce sont les gains en temps et les gains en qualité, apportés par l'utilisation d'outils d'extraction et d'alignement, qu'il convient de mesurer. Une évaluation sérieuse des techniques d'extraction et d'alignement passe donc de façon incontournable par une phase d'observation de leur utilisation dans des expériences en grandeur réelle de constitution de terminologie. C'est ainsi que l'on peut espérer pouvoir mesurer leur apport effectif et donc leur intérêt, et identifier leurs lacunes et donc déterminer les directions de recherche futures en extraction automatique de terminologie.

Dans la suite de cet article, nous relatons une expérience au cours de laquelle des juristes terminologues ont construit «à la main» un lexique bilingue des droits de l'homme, à partir des résultats fournis par un outil informatique d'aide au dépouillement terminologique. Notre propos dans cet article est en particulier de fournir des indications quantitatives concernant le nombre de couples construits et le temps passé par les juristes terminologues. Ces données, avec le lexique lui-même, pourraient servir de base pour appréhender quel pourrait être l'intérêt de fournir aux juristes terminologues les résultats d'un appariement statistique de candidats termes.

3 Le projet «lexique des droits de l'homme»

Le travail a été effectué dans le cadre d'un projet financé par le Ministère français de la Recherche et de l'Enseignement supérieur(1), et s'est déroulé sous la forme d'une collaboration entre linguistes

(1) Nous remercions Anne Guyon, de la Direction de l'information scientifique et technique et des bibliothèques, pour son soutien constant et amical.

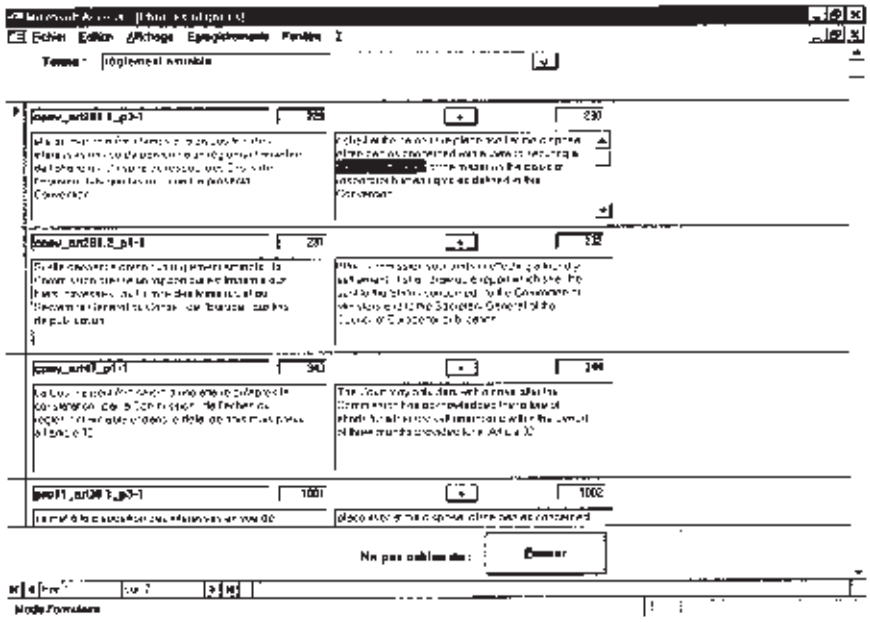

Figure 1 :
Illustration de la démarche. Pour chaque candidat terme (français) extrait par *Lexter*, l'interface HTL affiche les phrases du corpus français dans lesquelles il a été détecté avec en regard les phrases du corpus anglais qui en sont les traductions.

informaticiens et juristes terminologues. Ce lexique a été constitué à partir de l'analyse d'un corpus textuel bilingue composé de la Convention de sauvegarde des droits de l'homme et des libertés fondamentales, et de sa douzaine de protocoles, et de 36 arrêts rendus par la Cour européenne des droits de l'homme (CDEH) de Strasbourg en 1995.

La Cour européenne des droits de l'homme règle les litiges relatifs à l'interprétation et à l'application de la Convention de sauvegarde des droits de l'homme et des libertés fondamentales signée le 4 novembre 1950 et entrée en vigueur le 3 septembre 1953. Le texte de cette convention a été révisé depuis cette date par des protocoles (une dizaine) qui font partie intégrante de celle-ci. La compétence de la Cour s'exerce à l'égard des États qui l'ont reconnue de plein droit ou ont donné leur agrément à la saisine de celle-ci dans une affaire déterminée. À ce jour, une quarantaine d'états ont accepté la juridiction obligatoire de la Cour. Il existe deux versions officielles des textes précités : l'une en français, l'autre en anglais. Du 20 avril 1959 – date de son entrée en fonction – à 1997, la Cour a rendu plus de huit cents arrêts (dont environ 600 au cours des sept dernières années). Ces arrêts sont rédigés – comme la Convention et des protocoles – en anglais et en français. Les deux langues font également foi et il est impossible de distinguer une langue source et une langue cible. La Convention, les protocoles et les arrêts constituent ainsi un corpus juridique bilingue délimité. À une phrase dans une langue correspond, en effet, presque toujours exactement une phrase dans l'autre langue.

Une lecture attentive de ces textes fait apparaître certaines disparités de terminologie : ainsi pour un terme ou une expression français correspond parfois – dans les différents textes – plusieurs termes ou expressions anglais... et réciproquement. Les noms des institutions (*e.g.* Cour suprême autrichienne), en particulier, sont traduits de façon très variable d'un arrêt à l'autre... ce qui préoccupe vivement les traducteurs de la Cour. C'est pour mieux mettre en lumière la spécificité et la richesse du vocabulaire employé par la Cour dans les deux langues et, à terme et en accord avec la volonté de celle-ci de « normaliser » – dans une certaine mesure – la terminologie dans le domaine que nous avons entrepris ce travail. Celui-ci correspond en outre à un besoin : il n'existe aucun lexique fiable et récent en la matière et la CEDH elle-même ainsi que d'autres institutions (le Conseil constitutionnel en France, notamment) réclament un tel document.

Dès le départ du projet, et devant l'ampleur de la tâche, les juristes terminologues ont souhaité utiliser un outil informatique d'aide au dépouillement terminologique. C'est ainsi que s'est mise en place une collaboration entre les juristes terminologues du Centre de terminologie et de néologie (CTN) du Laboratoire de linguistique informatique de Villetaneuse, et les linguistes informaticiens de l'Équipe de recherche en syntaxe et sémantique de Toulouse.

4 Les traitements informatiques

4.1 Préparation des corpus : balisage et alignement

Le corpus de travail nous a été fourni sous forme électronique par la Cour européenne des droits de

l'homme de Strasbourg. Rappelons qu'il s'agit, pour chacune des deux langues, de la Convention accompagnée de ses protocoles et de 36 arrêts rendus par la Cour pendant l'année 1995. Il s'agissait de fichier Ascii pauvres, c'est-à-dire sans balisage logique des sections et encore moins de phrases. Dans cet état, plutôt que d'un corpus aligné, nous disposions d'un corpus «à aligner». En effet, la structure des arrêts est extrêmement bien marquée, et de façon équivalente pour les deux langues, sur le plan physique: parties, sections, alinéas, paragraphes, etc (*cf.* figure 2). De ce fait, pour n'importe quel lecteur humain, maîtrisant le français et l'anglais, il est aisé, à la lecture conjointe des deux corpus, d'associer à chaque phrase de l'un des corpus sa traduction dans l'autre. Mais pour permettre les traitements informatiques ultérieurs, il convenait de transformer cette structuration physique (visuelle), en une structuration logique manipulable par l'ordinateur: la tâche a consisté à identifier les phrases de chacun des deux corpus, et à associer le même identifiant aux couples de phrases équivalentes.

Dans le cas présent, étant donné la forte structuration du corpus, l'alignement a été effectué sur des bases uniquement formelles, il n'a pas été nécessaire de recourir aux techniques statistiques d'alignement de phrases, telles que celles évoquées dans la section 1. Les tâches de balisage et de segmentation ont été réalisées par une chaîne de programmes, qui ont été réalisés au fur et à mesure en s'assurant que l'application des règles s'effectuaient de façon identique sur chacun des deux corpus. La chaîne se décompose en 7 étapes:

1) Repérage du numéro de l'arrêt. Ce numéro (450 dans l'exemple ci-après) apparaît toujours dans le même contexte: *« TITLE: Affaire ALLENET de RIBEMONT c. France, CASE: 3/1994/450/529»*.

2) Repérage des grandes parties. Chaque arrêt se décompose en un certain nombre de grandes parties, chacune étant marquée par un titre normalisé: «*PROCÉDURE ET FAITS (*ang. *PROCEDURE AND FACTS)»,* dans laquelle est décrite la procédure qui a été suivie dans le pays d'origine avant que la Cour ne soit saisie, «*EN FAIT (ang. AS TO THE FACTS»*, qui présente les faits, *«EN DROIT (ang. AS TO THE LAW)»*, où sont détaillés les éléments du droit pertinents pour le cas, *«PAR CES MOTIFS, LA COUR (ang. FOR THESE REASONS, THE COURT)»*, qui expose la décision des juges de la Cour, *« OPINION DISSIDENTE (ang. DISSENTING OPINION)»*, dans le cas où certains des juges ne se sont pas rangés à l'avis de la majorité.

3) Repérage des numéros de section. Chaque arrêt est découpé en sections numérotées. C'est essentiellement ce découpage rigoureux en section, très largement répandu dans les textes de droit, qui rend possible l'approche formelle adoptée pour le découpage et l'alignement des corpus.

4) Repérage des paragraphes. Au sein de chaque section, le texte peut être organisé en paragraphes (portion entre deux retours chariots). Dans une grande majorité des cas, le découpage en paragraphes est identique dans les deux corpus. Dans les cas contraires, nous nous sommes autorisé l'insertion de marques de paragraphe pour rétablir un parallèle exact.

5) Repérage des citations. Dans le texte d'un arrêt peuvent apparaître des citations, soit des propos rapportés d'un des antagonistes du cas traité, soit des extraits de textes de loi des pays concernés. Pour distinguer ces passages du discours de la Cour lui-même, il a été jugé indispensable de repérer ces situations, heureusement marquées de façon régulière par un décalage du texte sur la droite et par des guillemets.

6) Élimination d'éléments divers non textuels. Les arrêts fourmillent d'élements textuels, peu intéressants sur le plan terminologique, que nous avons jugés bons d'éliminer pour simplifier la tâche, et de l'extracteur terminologique, et des juristes. Il s'agit en particulier des références à des cas déjà jugés, donnés sous la forme de leur titre ou de leur numéro (ex. «*Minelli c. Suisse, n° 266-A, p. 13* »); des références à des articles de loi (ex.: «*article 6 par. 2 (art. 6-2)»*); des dates; et enfin des noms d'individus.

7) Segmentation en phrases. Le programme de découpage en phrases s'appuie de façon très classique sur le repérage des ponctuations fortes, qui, hélas, diffèrent légèrement entre le français et l'anglais. Une évaluation de la qualité de l'alignement au niveau du paragraphe était fournie par le comptage du nombre de phrases pour chaque paragraphe. En cas de distorsion, nous nous sommes autorisé, là aussi, à insérer un certain nombre de modifications mineures dans les corpus (ajout, élimination ou changement de signes de ponctuations, ajout de retours chariots) de façon à ce que pour chaque paragraphe les nombres de phrases en français et en anglais soient les mêmes. Un exemple du résultat obtenu est présenté sur la figure 3.

Chaque corpus, qui compte environ 300 000 mots, a été segmenté en 12 131 phrases. La mise au point de ces programmes, qui ont été écrit en *Flex* sous Linux et la vérification du balisage ont pris une vingtaine d'heures. C'est très peu en regard du temps passé ensuite par les juristes à analyser le corpus pour construire le lexique. Cette phase de balisage et d'alignement était nécessaire pour la mise en place de la suite des opérations. À l'issue de ces

(...)
EN DROIT
I. SUR LA VIOLATION ALLÉGUÉE DE L'ARTICLE 6 PAR. 2 DE LA CONVENTION
31. M. Allenet de Ribemont dénonce les propos tenus lors de la conférence de presse du 29 décembre 1976 par le ministre de l'Intérieur et les hauts fonctionnaires de police qui l'accompagnaient. Il invoque l'article 6 par. 2 (art. 6-2) de la Convention, ainsi libellé:
«Toute personne accusée d'une infraction est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie.»
32. Le Gouvernement conteste en substance l'applicabilité de l'article 6 par. 2 (art. 6-2), en se fondant sur l'arrêt Minelli c. Suisse du 25 mars 1983 (série A n° 62). D'après lui, une atteinte à la présomption d'innocence ne peut provenir que d'une autorité judiciaire et ne se révéler qu'à l'issue de la procédure en cas de condamnation si la motivation du juge permet de supposer que celui-ci considérait a priori l'intéressé comme coupable.

(...)
AS TO THE LAW
I. ALLEGED VIOLATION OF ARTICLE 6 PARA. 2 OF THE CONVENTION
31. Mr Allenet de Ribemont complained of the remarks made by the Minister of the Interior and the senior police officers accompanying him at the press conference of 29 December 1976. He relied on Article 6 para. 2 (art. 6-2) of the Convention, which provides:
«Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law.»
32. The Government contested, in substance, the applicability of Article 6 para. 2 (art. 6-2), relying on the Minelli v. Switzerland judgment of 25 March 1983 (Series A no. 62). They maintained that the presumption of innocence could be infringed only by a judicial authority, and could be shown to have been infringed only where, at the conclusion of proceedings ending in a conviction, the court's reasoning suggested that it regarded the defendant as guilty in advance.

Figure 2:
Extraits des corpus français et anglais avant balisage et alignement

#A450_DR_1-p2-1
SUR LA VIOLATION ALLÉGUÉE DE L'ARTICLE 6 PAR. 2 DE LA CONVENTION
#A450_DR_a31_1-p1-1
<elim nompr>M._Allenet_de_Ribemont</elim> dénonce les propos tenus lors de la conférence de presse <elim date>du 29 décembre 1976</elim> par le ministre de l'Intérieur et les hauts fonctionnaires de police qui l'accompagnaient.
#A450_DR_a31_1-p1-2
Il invoque l'<elim article>article 6 par. 2</elim> <elim art>(art. 6-2)</elim> de la Convention, ainsi libellé:
#A450_DR_a31_1_CIT1-p1-1
Toute personne accusée d'une infraction est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie.
#A450_DR_a32_1-p1-1
Le Gouvernement conteste en substance l'applicabilité de l'<elim article>article 6 par. 2</elim> <elim art>(art. 6-2)</elim>, en se fondant sur l'arrêt <elim cas> Minelli c. Suisse</elim> <elim date>du 25 mars 1983</elim> (série A <elim numero>n° 62)</elim>.
#A450_DR_a32_1-p1-2
D'après lui , une atteinte à la présomption d'innocence ne peut provenir que d'une autorité judiciaire et ne se révéler qu'à l'issue de la procédure en cas de condamnation si la motivation du juge permet de supposer que celui-ci considérait a priori l'intéressé comme coupable.

#A450_DR_1-p2-1
ALLEGED VIOLATION OF ARTICLE 6 PARA. 2 (art. 6-2) OF THE CONVENTION
#A450_DR_a31_1-p1-1
<elim nompr>Mr Allenet de Ribemont</elim> complained of the remarks made by the Minister of the Interior and the senior police officers accompanying him at the press conference of <elim date>29 December 1976</elim>.
#A450_DR_a31_1-p1-2
He relied on <elim article>Article 6 para. 2 </elim><elim art>(art. 6-2)</elim> of the Convention, which provides:
#A450_DR_a31_1_CIT1-p1-1
Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law .
#A450_DR_a32_1-p1-1
The Government contested, in substance, the applicability of <elim article>Article 6 para. 2. 2</elim> <elim art>(art. 6-2) </elim>, relying on the l'arrêt <elim cas>Minelli v. Switzerland</elim> judgment of <elim date>25 March 1983</elim>. (Series A <elim numero>no. 62</elim>).
#A450_DR_a32_1-p1-2
They maintained that the presumption of innocence could be infringed only by a judicial authority, and could be shown to have been infringed only where, at the conclusion of proceedings ending in a conviction, the court's reasoning suggested that it regarded the defendant as guilty in advance.

Figure 3:
Extraits des corpus français et anglais après balisage et alignement. À l'issue de la phase de balisage et d'alignement, les deux corpus (300 000 mots chacun) sont découpés en 12 131 phrases, alignées.

sept étapes, le découpage ainsi effectué est exact jusqu'au niveau du paragraphe. En ce qui concerne le découpage en phrases au sein des paragraphes, une évaluation au moment de la recherche des équivalents dans l'interface de validation conduit à une estimation du taux d'alignement correct d'environ 90%. Les erreurs d'alignement ne sont pas préjudiciables au moment de la construction du lexique, puisque que dans tous les cas la phrase équivalente précède ou suit la phrase alignée, et l'utilisateur y accède donc rapidement.

4.2 Extraction terminologique sur le corpus français

Après le balisage et l'alignement, la seconde phase du travail informatique a consisté à faire traiter le corpus français par l'extracteur de terminologie *Lexter* (Bourigault 1993, Bourigault *et al.* 1996). *Lexter* reçoit en entrée un corpus de texte, en français, portant sur un domaine quelconque, effectue une analyse morpho-syntaxique de ce corpus et extrait une liste de candidats termes, c'est-à-dire de mots ou de séquences de mots susceptibles d'être retenus comme termes du domaine. Ces candidats termes sont soumis au terminologue par l'intermédiaire d'une interface hypertextuelle de validation, dite « Hypertexte terminologique *Lexter* » (HTL). Pour chacun des candidats termes, le terminologue accède, entre autres, à l'ensemble des phrases du corpus dans lesquelles le logiciel a détecté le candidat.

Lexter est habituellement utilisé dans des contextes monolingues. Dans ce projet, il a pu être utilisé pour élaborer une terminologie bilingue parce que nous disposions d'un corpus aligné. La démarche a donc été la suivante. *Lexter* a traité le sous-corpus français et a extrait des candidats termes français qui ont été validés par les juristes. L'interface de validation HTL a été légèrement modifiée de façon à ce qu'en regard les phrases du sous-corpus français dans lesquelles a été détecté un terme soient affichées les phrases du sous-corpus anglais qui en sont les traductions. Grâce à cela, les juristes retrouvent très facilement dans ces phrases, dans la grande majorité des cas, les équivalents anglais des candidats termes français (*cf.* figure 1).

4.3 Premiers résultats quantitatifs

L'expérience n'a pas encore atteint son terme. Nous estimons avoir accomplir les trois quarts du travail[2] (avant la soumission de nos résultats aux traducteurs experts de la Cour). Le lexique comporte actuellement un peu plus de 4 000 couples de termes français/anglais. Nous ne présentons ici que quelques indications de type quantitatif. Une analyse des résultats d'un point de vue théorie de la traduction et théorie du droit est exposée dans Humbley *et al.* (1999).

Conformément à ce que nous avons annoncé dans la section 2.2, c'est aux paramètres de la qualité du lexique et du temps d'élaboration qu'il convient de s'intéresser. La qualité est assurée par le fait que tous les couples de termes sont construits « manuellement » par les juristes terminologues par observation des attestations en corpus. Le lexique est donc une image fidèle, une photographie, de la façon dont les traducteurs de la Cour travaillent. Bien entendu, cela ne signifie pas que toutes les traductions sont bonnes, et l'un des résultats du travail pourrait être une tentative de normalisation, qui devrait être entreprise par les traducteurs eux-mêmes sur la base des résultats que nous leur fournissons.

Sur le plan du temps d'élaboration, les termes peu fréquents dans le corpus sont très rapidement analysés, puisque le juriste terminologue peut observer d'un coup d'œil l'ensemble des couples de phrases, pour choisir le ou les équivalents. Bien entendu, pour les termes très fréquents l'analyse est plus longue. Nous estimons à une vingtaine de jours pleins la durée d'élaboration du lexique dans son état actuel. C'est sur cette base que devrait être mesuré l'intérêt de l'introduction de techniques statistiques d'appariement de termes, dont les résultats ne sont précis que pour les termes extrêmement fréquents.

Dans le tableau 1, nous présentons les résultats chiffrés concernant les termes complexes (syntagmes nominaux). Environ 60% des candidats termes fournis par le logiciel ont été retenus par les juristes terminologues. Ce chiffre se situe dans la fourchette des taux habituellement observés en extraction automatique de terminologie. Précisons d'une part que l'élimination des candidats termes jugés non pertinents est une opération très simple et très rapide, et d'autre part que parmi les 40% éliminés, seuls quelques pour-cents doivent être considérés à proprement dit comme du bruit, consécutif à des erreurs d'analyse du logiciel, soit au moment de l'étiquetage, soit au moment du repérage des syntagmes nominaux. Une bonne partie des syntagmes non retenus apparaissent dans des parties du corpus décrivant les faits jugés (parties «EN FAIT», *cf.* section 4.1), et ne présentent pas de pertinence d'un point de vue du droit.

Il convient de constater que les hapax – les candidats termes extraits

(2) Nous remercions Delphine Bailly, stagiaire en traductrion juridique, pour sa contribution efficace au projet.

	Extraits	Vus	Non retenus	Retenus	analysés
fréquence = 1	12 193	6 375 *43%*	2 720 *57%*	3 655	1 183
fréquence > 1	4 283	3 185 *33%*	1 058 *66%*	2 127	2 127
TOTAL	16 476	9 560 *40%*	3 778 *60%*	5 483	3 310

Tableau 1.
Nombre de candidats termes extraits par *Lexter*, vus, retenus et analysés par les juristes terminologues.

une seule fois dans le corpus – présentent un intérêt certain. 57% des hapax ont été retenus (3 655/6 375), 36% des termes retenus sont des hapax (1183/3310). Ceci laisse entrevoir d'éventuelles limites des outils d'appariement statistique, qui se basent sur la récurrence des associations.

Le travail de constitution du lexique a très vite fait apparaître que les cas de traductions multiples étaient très fréquents, et ce dans les deux sens. C'est ainsi que sur les 2 127 termes français dont le nombre d'occurrence est supérieur à 2, 553 (soit 26%) ont au moins deux équivalents différents! 17% (168/981) des termes français qui n'apparaissent que deux fois dans le corpus se voient associés à deux équivalents anglais différents. Les chiffres sont du même ordre de grandeur de l'anglais vers le français. Une analyse des cas de traduction multiple est présentée dans Humbley *et al.* (1999). Un exemple pour conclure: le terme français *détention provisoire* est traduit par: *detention on remand, pre-trail detention, detention pending trial, imprisonment in default, remand custody, remand detention*.

5 Perspectives

En ce qui concerne le projet de lexique, notre objectif est d'achever la validation et l'analyse des candidats termes, de façon à atteindre une couverture maximale du corpus. Ce lexique sera ensuite soumis aux traducteurs de la Cour. Par ailleurs, le lexique, dans son état actuel, a servi de base pour l'intégration de deux nouvelles langues: le polonais et le roumain. Pour ces deux langues, seule la convention et certains protocoles ont été analysés. L'élaboration de lexiques multilingues incluant d'autres langues que les langues officielles de la Cour est indispensable pour les pays adhérents du Conseil de l'Europe qui souhaiteraient faire traduire les arrêts de la Cour dans leur(s) langue(s) nationale(s). Nous souhaitons étendre le lexique aux verbes et syntagmes verbaux. Le travail a été entièrement réalisé, à la main, sur la Convention et ses protocoles. Un travail à grande échelle, sur l'ensemble du corpus, est envisageable dans un avenir proche, dès que le logiciel *Lexter* aura été étendu à l'extraction des syntagmes verbaux. Sur le plan des recherches en extraction terminologique bilingue, nous avons entamé une collaboration avec les chercheurs du Centre de recherche de Xerox à Grenoble pour une confrontation et une évaluation comparative de nos approches.

Didier Bourigault,
Équipe de recherche en syntaxe et sémantique,
CNRS et Université Toulouse 2.

Christine Chodkiewicz,
Centre de terminologie et néologie,
Laboratoire de linguistique informatique,
Université Paris XIII.

John Humbley,
Centre de terminologie et néologie,
Laboratoire de linguistique informatique,
Université Paris XIII.

Bibliographie

Bourigault (D.), 1993: «Analyse syntaxique locale pour le repérage de termes complexes dans un texte», dans la *Revue Tal*, volume 34, n°2.

Bourigault (D.), Gonzalez-Mulliez (I.) et Gros (C.), 1996: «Lexter, a Natural Language Tool for Terminology Extraction», dans *Actes du 7e congrès international Euralex*, Göteborg, Suède.

Bourigault (D.) et Habert (B.), 1998: «Evaluation of Terminology Extractors: Principles and Experiments», dans Rubio (A.), Gallardo (N.), Castro (R.) et Tejada (A.), Éditeurs, dans *Actes de la première conférence internationale sur les ressources linguistiques et l'évaluation*, volume I, p. 299-305, Grenade, Espagne.

Brown (P.), Lai (J.) et Mercer (R.), 1991: «Aligning sentences in parallel corpora», dans *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics (ACL'91).*

Church (K. W.), 1993: «Char_align: A program for aligning parallel texts at the character level», dans *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics (ACL'93)*, Columbus.

Dagan (I.), Church (K. W.) et Gale (W. A.), 1993: «Robust bilingual word alignment for machine aided translation»,

dans *Proceedings of the workshop on Very Large Corpora* (*VLC'93*), Columbus.

Gale (W. A.) et Church (K. W.), 1993, «A program for aligning sentences in bilingual corpora», dans *Computational Linguistics*, 19(1).

Gaussier (E.), 1998: «Flow network models for word alignment and terminology extraction from bilingual corpora», dans *Actes de la 17e conférence internationale de linguistique informatique (COLING-ACL'99)*, volume I, pp. 444-450, Montréal, Canada.

Hull (D.) 1998: «A pratical approach to terminology alignment», dans Bourigault (D.), Jacquemin (C.) et L'Homme (M.-C.) éditeurs, dans *Proceedings of the first workshop on Computational Terminology (COMPUTERM'98)*, Montréal.

Humbley (J.), Chodkiewicz (C.) et Bourigault(D.), 1999: «Using *Lexter* to establish a glossary of Human Rights», (à paraître) dans *Actes de la conférence Terminology and Knowledge Engineering (TKE'99)*, Vienne.

Kupiec (J.), 1993: «An algorithm for finding noun phrase correspondences in bilingual corpora», dans *Proceedings of the 31st Annual Conference of the Association for Computational Lingustics (ACL'93)*, Columbus.

Smadja (F.) et McKeown (K.), 1994: «Translating collocations for use in bilingual lexicons», dans *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, New Jersay.