

Chapitre 6

Alignement de corpus multilingues

6. 1. Introduction

En juillet 1799, les soldats de l'armée de Napoléon découvraient près de la ville de Rosette, sur le delta du Nil, une pierre qui allait devenir l'une des plus célèbres de l'Antiquité. Cette pierre, datant de 196 av. J.-C., relatait les honneurs rendus au roi Ptolémée V par les temples d'Égypte sous forme d'un "texte 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). Son étude permit à Jean-François Champollion d'apporter en 1822 la clé du déchiffrement de l'écriture hiéroglyphique, découverte qui eut un retentissement considérable car elle mettait fin aux nombreuses controverses et mythes qui avaient entouré cette écriture.

En fait, la pierre de Rosette est relativement récente : on a trouvé de très nombreuses inscriptions multilingues qui jalonnent à peu près toutes les périodes de l'Antiquité depuis l'invention de l'écriture. Les inscriptions en deux langues des tombes des princes d'Éléphantine en Égypte, qui datent du III^{ème} millénaire avant J.-C., témoignent par exemple d'une activité de traduction extrêmement ancienne. Pas toujours autant médiatisées que la pierre de Rosette, les inscriptions antiques ont été d'une utilité fondamentale pour le déchiffrement des langues et écritures anciennes : on dispose ainsi de nombreuses inscriptions dans des combinaisons de langues variées : sumérien/akkadien, hittite/babylonien, vieux-perse/babylonien/élamite, phénicien/étrusque, etc. Jusqu'à nos jours, l'Histoire est constellée de textes parallèles (contrats, traités, œuvres sacrées, littérature, etc.), datant de toutes les époques et concernant presque tous les couples de langues en contact, même si ce parallélisme est parfois seulement virtuel, les textes et leur traductions n'étant pas toujours sur le même support physique comme dans le cas de la pierre de Rosette.

2 Ingénierie des langues

C'est seulement à partir des années 80, cependant, que les textes parallèles ont commencé à être exploités de façon systématique dans le cadre du traitement automatique des langues. Quelques tentatives semblent avoir été faites en traduction automatique à la fin des années 50 [KOUTSOUDAS 1957 ; LEON 1998, p.292], mais les capacités de stockage et de calcul des ordinateurs, ainsi que les difficultés de saisie de quantités importantes de textes, ne permettaient probablement pas alors une exploitation pertinente. Selon Alan Melby, l'idée du stockage d'exemples de traductions en vue de leur réutilisation future semble avoir germé de façon indépendante dans divers centres de recherche à la fin des années 70, notamment à Brigham Young et Xerox PARC [MELBY 1981 ; KAY 1980]. La première méthode automatique d'alignement de textes parallèles a, quant à elle, été développée par Martin Kay à partir de 1984 [KAY 1988], et rapidement, de nombreuses méthodes ont été proposées pour l'*alignement* de différents niveaux d'unités entre les textes (c'est-à-dire la mise en correspondance des unités qui sont la traduction l'une de l'autre) : paragraphes, phrases, mots et expressions¹.

Les applications des *textes parallèles alignés* (que l'on appelle parfois *bitextes* [HARRIS 1998] ou *multitextes*) sont extrêmement diverses : constitution 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 ou la linguistique contrastive, etc. Étant donnée l'importance croissante du multilinguisme dans les industries de la langue, provoquée par la globalisation des marchés et de l'information, l'exploitation des corpus de textes parallèles semble être une technologie promise à un avenir florissant. Les textes parallèles eux-mêmes sont disponibles de façon de plus en plus massive, grâce à l'archivage électronique croissant des documents dans les entreprises et grâce au World Wide Web, qui fournit une source de plus en plus riche de documents multilingues.

6.2. Applications

Nous détaillons dans cette section les applications principales des textes parallèles alignés : lexicographie et terminologie, traduction, recherche d'informations, ainsi que quelques applications (par exemple en enseignement des langues assisté par ordinateur) qui ont fait l'objet de peu de publications pour l'instant, mais sont en essor constant.

¹ Cet essor coïncide bien sûr avec le renouveau de la linguistique de corpus, stimulé à la fois par la disponibilité de grandes masses de textes électroniques, la capacité accrue des ordinateurs et le relatif désarroi des programmes linguistiques refusant le recours aux données empiriques.

6.2.1. Lexicographie et terminologie

La lexicographie moderne utilise de façon de plus en plus systématique des corpus de textes informatisés. Le projet pionnier du *Trésor de la Langue Française* [IMBS 1971] a montré l'utilité de tels corpus dès la fin des années 50, mais leur utilisation par des maisons d'édition est beaucoup plus récente : elle remonte au projet COBUILD [SINCLAIR 1987]. L'utilité des corpus pour la lexicographie bilingue a également été mentionnée depuis un certain temps [HARTMANN 1980 ; ATKINS 1990] et les projets de dictionnaires commencent à faire appel à des corpus. La compilation du *Oxford-Hachette French Dictionary* a reposé sur un corpus anglais et un corpus français de plus de 10 millions de mots chacun [GRUNDY 1996]. Bien qu'il s'agisse dans ce cas de corpus indépendants (de textes *comparables*, c'est-à-dire de genre, domaine, époque, etc. similaires), il est évident que l'utilisation de corpus parallèles constituerait une ressource additionnelle précieuse pour le lexicographe. L'alignement préalable des textes, à des unités grossières telles que le paragraphe ou la phrase, permet en effet d'utiliser des outils de concordance bilingue² à l'aide desquels le lexicographe peut examiner rapidement un grand nombre de traductions d'un mot donné, et repérer des éléments importants de contexte tels que les collocations (WARWICK 1990 ; LANGLOIS 1996). Le projet interuniversitaire du *Dictionnaire Canadien Bilingue* fait ainsi appel à un corpus parallèle aligné de près de 50 millions de mots (le *Hansard*, voir section 6.3), en complément à un corpus plus important de textes comparables [ROBERTS 1996].

Si l'utilisation des corpus parallèles n'a pas encore atteint massivement l'édition de dictionnaires classiques³, elle est déjà très importante dans le domaine de la terminologie et de la conception de lexiques computationnels. Il n'est pas possible de citer toutes les études, et nous nous bornerons à mentionner quelques jalons importants.

² Pour des exemples de concordanciers parallèles, voir [SIMARD 1993] (<http://www-rali.iro.umontreal.ca/TransSearch/TS-project.en.html>), [BARLOW 1995] (<http://www.ruf.rice.edu/~barlow/parac.html>), [EBELING 1998a].

³ Outre la nouveauté des techniques d'alignement et la taille pour l'instant modeste des corpus parallèles par rapport aux corpus monolingues, on peut noter deux phénomènes qui freinent leur utilisation dans le cadre de la lexicographie classique. D'une part, une certaine réticence des lexicographes traditionnels à utiliser des traductions, perçues souvent comme n'étant pas de vrais actes de discours mais des artefacts— mais le but des dictionnaires bilingues étant justement, au moins en partie, la production de tels artefacts, cette réticence devrait aller en s'amenuisant. D'autre part, les corpus parallèles sont nécessairement biaisés dans leur représentativité : les textes traduits disponibles relèvent généralement de domaines particuliers (textes légaux, techniques, etc.). Certains genres y sont sous-représentés, en particulier conversations, émissions radiophoniques, etc.).

4 Ingénierie des langues

Dès 1989, Catizone, Russel et Warwick utilisent des textes parallèles en anglais et allemand pour extraire des lexiques bilingues à l'aide de méthodes statistiques et en conjonction avec des dictionnaires en ligne (*machine-readable dictionaries*). [SADLER 1989] utilise des corpus parallèles analysés sous forme d'arbres de dépendances pour proposer des entrées lexicales bilingues à un utilisateur. Klavans et Tzoukerman [1990] proposent un système (BICORD) qui combine l'information dérivée d'un dictionnaire bilingue avec celle extraite d'un corpus parallèle et montrent son application à l'étude de verbes de mouvement.

Beaucoup d'études se sont attachées à l'extraction de dictionnaires de mots simples, le plus souvent par des méthodes statistiques [DAGAN 1993 ; WU 1994b ; DAGAN 1994 ; RESNIK 1997]. Très rapidement, les travaux se sont toutefois orientés vers l'extraction d'unités plus longues que le mot graphique : collocations, terminologie et phraséologie. En effet, ces unités composées sont de toute évidence le domaine dans lesquels les dictionnaires classiques (surtout s'ils n'ont pas été basés sur l'analyse de corpus) sont le moins satisfaisants. De plus, surtout dans le domaine terminologique, les besoins peuvent être temporaires et/ou très spécialisés, et il est courant que les ressources correspondantes n'existent pas, alors que les traductions peuvent, elles, exister en abondance. De nombreux auteurs se sont attachés à extraire des unités complexes de textes alignés [KUPIEC 1993 ; VAN DER EIJK 1993 ; DAGAN 1994 ; DAILLE 1994 ; GAUSSIER 1995], et les études les plus récentes [SMADJA 1996 ; MELAMED 1997 ; HIEMSTRA 1998] semblent indiquer des avancées importantes dans le domaine.

6.2.2. Traduction

L'idée d'une traduction entièrement automatique de haute qualité (*fully automated high quality translation*, ou *FAHQT*) a été certainement abandonnée par la plupart des chercheurs, du moins comme objectif de recherche à court et moyen terme. La plupart des directions de recherches actuelles se situent dans un continuum entre deux pôles : d'une part, la traduction humaine assistée par différents outils informatisés, d'autre part la traduction automatique assistée par l'homme. Tout au long de ce continuum, les bases de textes parallèles peuvent fournir des outils et des ressources utiles. [ISABELLE 1992] faisait remarquer que le volume colossal de traductions (au moins un demi milliard de mots par an pour le seul Canada), si l'on pouvait l'exploiter systématiquement, fournirait au traducteur la solution à bien plus de problèmes que tous les dictionnaires possibles. Sans doute la solution de nombreux problèmes de traduction automatique est-elle aussi présente dans ces masses de traductions existantes. Il y a de bonnes raisons, en fait, de compléter les outils classiques (dictionnaires, listes terminologiques), qu'ils soient à destination du traducteur humain ou des systèmes de traduction automatique. D'une part, bien des domaines spécialisés, surtout s'ils sont récents ou transitoires, n'ont pas fait l'objet

d'une compilation systématique par les lexicographes et terminologues. D'autre part, les dictionnaires bilingues classiques souffrent d'un certain nombre de défauts, et sont souvent assez pauvres (bien que la lexicographie moderne essaie de corriger cette tendance) dans le domaine des restrictions de sélections lexicales, des collocations, de la phraséologie, etc.

En 1980, Kay proposait une approche graduelle de l'automatisation des tâches de traduction, dans laquelle l'une des premières étapes pourrait consister à fournir au traducteur humain des exemples de textes contenant un matériau similaire à celui qu'il doit traduire. Cette idée a été implémentée par différentes équipes, qui proposent, souvent dans le cadre de "stations de travail du traducteur" (*translator's workstation*) des outils de concordances bilingues et de gestion d'une mémoire de traduction [KJAERGAARD 1987 ; ISABELLE 1992 ; MACKLOVITCH 1992 ; PICCHI 1992].

L'idée de réutiliser des fragments de traduction pour la traduction automatique semble, comme nous l'avons mentionné dans l'introduction, remonter à la fin des années 70. Un courant de recherche appelé "traduction automatique basée sur la mémoire" (*memory-based machine translation*) ou "traduction automatique basée sur les exemples" (*exemple-based machine translation* ou *EBMT*) s'est constitué à partir du milieu des années 80 [NAGAO 1984 ; SADLER, 1989 ; SATO 1990 ; SUMITA 1990]. Le principe général de ce type de traduction consiste à rechercher dans une base d'exemples de traduction des fragments similaires à certaines portions du texte à traduire et à les recombinaison de façon adéquate – ce qui peut demander l'établissement un ensemble de règles d'une grande complexité. Un autre courant de recherches s'est développé à peu près simultanément, en particulier chez IBM, visant à s'affranchir de cette complexité en la laissant "apprendre" automatiquement par la machine à partir de modèles statistiques. [BROWN 1990], renouant ainsi d'une certaine façon avec l'approche initiale de [WEAVER 1949], propose un modèle de traduction dont les paramètres sont estimés sur 40000 paires de phrases tirées du corpus *Hansard* (français-anglais), et qui obtient des résultats étonnamment bons par rapport à la simplicité du modèle mis en œuvre. Les systèmes basés sur l'exemple, qui à l'origine dans la conception de [NAGAO 1984] utilisaient des règles explicites, utilisent de plus en plus des composants statistiques, par exemple [BROWN, 1996].

Enfin, mentionnons que les techniques d'alignement peuvent intervenir lors de la création même de textes parallèles, et peuvent ainsi fournir un support à la création et à la maintenance de documents multilingues. [ISABELLE 1993 ; MACKLOVITCH 1995] proposent ainsi un système (*TransCheck*) destiné à vérifier automatiquement les traductions, à l'aide de l'alignement automatique du texte et de sa traduction. Les erreurs potentielles qui peuvent être repérées sont de diverses natures : faux amis, omissions, cohérence terminologique, etc. (voir aussi [MELAMED 1996]). Ce type d'outil peut être particulièrement important au niveau de la gestion des documents

multilingues : la cohérence des versions et révisions successives est très difficile à assurer, et la détection de différences structurelles (parties manquantes ou ajoutées, etc.) peut être extrêmement utile. Foster, Isabelle et Plamondon [1997] montrent par ailleurs comment le travail de frappe du traducteur peut être minimisé à l'aide d'un modèle de langage qui anticipe les mots possibles à chaque point du texte traduit à partir du texte source. Enfin, [ISABELLE 1993] décrivent un projet de dictée automatique pour le traducteur (*TransTalk*), dans lequel la reconnaissance des mots de la traduction dictée peut être améliorée grâce à l'alignement avec le texte source.

6.2.3. Recherche d'information multilingue

Le domaine de la recherche d'information multilingue (*cross-language information retrieval*) est en progression constante depuis une dizaine d'années, et vit actuellement une véritable explosion grâce au Web. La recherche d'information multilingue consiste à poser des requêtes dans une langue, alors que les documents (ou une partie d'entre eux) sont dans une ou plusieurs autres langues : cette situation est rendue nécessaire lorsque de nombreux utilisateurs sont suffisamment bilingues pour comprendre des documents présentés dans une autre langue, mais ne seraient pas forcément capables de bâtir des requêtes avec les mots-clés appropriés. Diverses techniques ont été proposées (voir les états de l'art de [FLUHR 1995 ; OARD 1996]). En particulier, la traduction (entièrement) automatique des requêtes ou simplement la reformulation mot à mot dans l'autre langue à l'aide de dictionnaires bilingues. Ces techniques ne sont cependant pas totalement satisfaisantes, à cause de la difficulté de la traduction automatique et des imperfections des dictionnaires bilingues, qui ne couvrent pas forcément tout le champ de la base. Lorsque des textes parallèles existent pour au moins une partie de la base de textes, ils peuvent être exploités de façon utile. D'une part, des dictionnaires bilingues peuvent être extraits de la directement de la base, selon des techniques analogues à celles mentionnées précédemment. Ces dictionnaires peuvent remplacer ou compléter de façon utile les dictionnaires bilingues et thesaurus classiques [EVANS 1991 ; OARD 1994]. D'autres techniques peuvent cependant mettre à contribution la partie parallèle de la base de façon plus directe. Supposons par exemple une requête en français. Celle-ci, appliquée par les moyens traditionnels à la partie française de la base parallèle (par exemple français-anglais), ramènera des documents français qui peuvent être classés par ordre de pertinence. Les documents anglais correspondant aux documents français classés en tête peuvent à leur tour être utilisés comme requête sur le reste de la base anglaise, non parallèle. Diverses variantes de cette idée générale ont été proposées utilisant des techniques d'indexage sémantique latent (*semantic latent indexing*), *relevance feedback*, etc., combinées parfois avec l'extraction de dictionnaires bilingues des corpus pour la reformulation des requêtes [LANDAUER 1990 ; DAVIS 1995 ; YANG, 1998]

6.2.4. Autres applications

Diverses autres applications des textes parallèles ont été proposées, en dehors des grands courants mentionnés ci-dessus. L'utilisation de textes parallèles a ainsi été proposée pour la résolution d'un problème monolingue classique, celui de la désambiguïsation du sens des mots dans des textes, problème qui intervient sous une forme ou sous une autre dans la quasi totalité des applications de traitement automatique des langues [IDE 1998]. [BROWN 1991a ; GALE 1993b] proposent l'utilisation de corpus parallèles pour constituer un corpus d'entraînement d'amorçage (*bootstrap*) pour les systèmes de désambiguïsation automatique : les auteurs utilisent le fait que dans de nombreux cas, l'ambiguïté lexicale présente dans une langue est levée par des choix lexicaux différents dans l'autre (par exemple, l'anglais *pen* se traduit en français soit par *stylo* soit par *enclos* selon son sens). Les textes alignés peuvent également fournir des banques d'exemples extrêmement précieuses pour l'enseignement des langues, où l'observation de la véritable utilisation des mots et expressions en contexte peut compléter de façon très efficace les outils classiques que sont les dictionnaires et les grammaires [PIENEMANN 1992 ; JAGTMAN, 1994 ; ZANETTIN, 1994 ; BONHOMME ; 1995 ; ROMARY, 1995 ; BARLOW, 1996]. Enfin, les corpus alignés peuvent constituer des ressources de base pour les recherches linguistiques comparatives et l'étude théorique de la traduction [MELBY 1981 ; SALKIE 1995 ; EBELING 1998b ; JOHANSSON 1998 ; KENNING 1999].

6.3. Corpus

De nombreux projets ont émergé à travers le monde, visant à la compilation de corpus de textes parallèles, et nous nous bornerons ici à mentionner les principaux. Le corpus *Hansard* (français-anglais) est sans doute l'un des premiers, et en tous cas le plus célèbre des corpus parallèles. Collecté dans les années 80 par l'*IBM T. J. Watson Research Center* et *Bell Communications Research*, il contient plusieurs dizaines de millions de mots, tirés des transcriptions des débats du parlement Canadien sur une période allant du milieu des années 1970 à 1988. Ce corpus a été utilisé dans de multiples études, et a constitué au fil des années une sorte d'étalon de fait pour l'évaluation et la mise au point des systèmes. Toutefois, sa limitation à genre de texte et à un couple de langue, ont rendu nécessaire le recueil d'autres données. Ainsi, en 1992-93, l'*European Corpus Initiative* (ECI) a collecté une masse importante de données parallèles, dans des langues européennes diverses et en particulier les textes trilingues (français, anglais, espagnol) de l'*International Telecommunications Union CCITT handbook* (13,5 M de mots) et l'*International Labour Organisation* (5M). En 1994-95, les projets MULTEXT-MLCC ont collaboré pour la collecte et la préparation d'un volume important de données dans les neuf langues de la Communauté Européenne de l'époque : questions écrites des parlementaires (10 M de mots), débats du parlement (environ 60 M). De nombreux

autres projets sont en cours, tel que l'*English-Norwegian Parallel Corpus* (ENPC) [JOHANSSON 1994], le projet LINGUA [ROMARY 1995], le projet TRIPTIC [PAULUSSEN 1995], le projet PEDANT [DANIELSSON 1996], etc. De nombreux efforts ont été faits, également, pour la constitution de corpus parallèles pour les langues asiatiques (par exemple le projet JEIDA [ISAHARA 2000]).

Étant donnés les coûts, les textes alignés (et vérifiés) sont beaucoup moins nombreux. IBM et Bellcore, puis le projet ARCADE ont aligné des parties du *Hansard* [SIMARD 1998 ; VERONIS 2000] au niveau des phrases. Le projet CRATER a aligné 3 millions de mots du *CCITT handbook* [GARSIDE 1994]. Le projet MULTEXT a aligné environ 1 million de mots au niveau des phrases, en cinq langues [IDE 1994]. A part quelques expériences telles que les projets BLINKER [MELAMED 1998a] ou ARCADE [VERONIS 2000], on se semble pas disposer pour l'instant de corpus significatifs alignés au niveau des mots ou expressions.

Enfin, des efforts sont en cours pour constituer des corpus parallèles de langues rares ou minoritaires, ou de pays moins technologiques. L'ECI proposait déjà quelques textes parallèles (non alignés) entre l'anglais, le serbe, le slovène, le croate et l'uzbek. Les projets MULTEXT-EAST [ERJAVEC 1995] et TELRI ont réuni un corpus de langues de pays européens de l'Est (partiellement aligné en phrases).

6.4. Techniques d'alignement

6.4.1. Alignement de phrases

La plupart des méthodes publiées à ce jour dérivent de deux groupes d'études initiales : [KAY 1988] d'une part, et [GALE 1993a] ainsi que [BROWNb 1991] d'autre part. Ces deux groupes d'études sont basés sur des principes différents, bien que, comme nous le verrons plus bas, ils reposent sur un certain nombre d'hypothèses simplificatrices communes.

La méthode proposée par [KAY 1988] fait l'hypothèse que pour que des phrases soient en correspondance de traduction, il faut que les mots qui les composent soient également en correspondance. Elle n'utilise qu'une information *interne*, c'est-à-dire que toute l'information nécessaire (et en particulier les correspondances lexicales) est dérivée des textes à aligner eux-mêmes. Kay et Röscheisen soulignent bien sûr la difficulté d'aligner les mots entre deux textes, mais utilisent le fait qu'un tel alignement, même très grossier et très imparfait, peut conduire à un alignement satisfaisant au niveau des phrases. Le point de départ de l'algorithme est un ensemble initial de phrases raisonnablement candidates à l'alignement : la première phrase et la dernière ont de bonnes chances de se correspondre dans chaque texte, et les phrases intermédiaires sont certainement en correspondance dans un couloir

diagonal relativement étroit. L'algorithme compare ensuite la distribution des mots de cet ensemble de phrases dans chacun des textes et fait l'hypothèse que si ces distributions sont similaires au-delà d'un certain seuil pour un couple de mot donné, ces mots ont de bonnes chances d'être en relation de traduction. Les mots en question fournissent alors un ensemble de points d'ancrage qui permettent de réduire le couloir diagonal des alignements de phrases candidats. La procédure est itérée jusqu'à convergence vers une solution minimale.

Gale et Church [1993] proposent une méthode qui n'utilise également qu'une information interne, mais ne fait aucune hypothèse directe sur le contenu lexical des phrases. Les auteurs partent de la constatation que la longueur des phrases dans le texte source et de leurs traductions dans le texte cible sont fortement corrélées : des phrases courtes ont tendance à être traduites par des phrases courtes, et des phrases longues par des phrases longues. De plus, il semble exister un rapport assez constant entre les longueurs de phrases d'une langue à l'autre en termes de nombre de caractères (ainsi il est connu que les textes français sont plus longs que leurs équivalents anglais : une évaluation sur le corpus du projet ARCADE montre par exemple que ce rapport est de l'ordre de 1.1 entre le français et l'anglais et qu'il varie peu selon le genre des textes [VERONIS 2000]). Cette observation permet de construire un modèle probabiliste et une mesure de dissimilarité entre phrases des deux textes à aligner, qui prennent en compte la proportion des types d'alignements attendus : le cas de loin le plus courant est celui où une phrase du texte source correspond exactement à une phrase du texte cible (1:1), mais d'autres cas sont possibles, correspondant à des omissions (1:0), des additions (0:1) ou des fusions plus ou moins complexes ($m:n$) (avec $m, n > 1$). L'alignement optimal est celui qui minimise la mesure de dissimilarité accumulée sur l'ensemble du texte. Pour des raisons de calculabilité, Gale & Church sont amenés à faire des hypothèses simplificatrices, et en particulier à réduire le cas ($m:n$) à $m, n \leq 2$. L'alignement optimal peut alors être calculé de façon efficace par un algorithme classique de programmation dynamique. [BROWN 1991b] utilisent également la même idée de corrélation entre les longueurs de phrases, mais ils formulent le problème à l'aide de modèles de Markov cachés.

Comme nous l'avons mentionné, la plupart des méthodes utilisent ces deux idées, ancrage lexical et corrélation des longueurs de phrases, la plupart du temps en les combinant entre elles. [DEBILI 1992] propose un alignement de phrases basé sur un alignement préalable des mots à l'aide d'un dictionnaire bilingue classique. [SIMARD 1992 ; CHURCH 1993a ; JOHANSSON 1993 ; MCENERY 1995] proposent quant à eux une méthode d'ancrage lexical très simple, qui donne cependant de bons résultats combinée avec une méthode de type Gale & Church. Il s'agit de repérer dans les textes à aligner, la présence de *cognates*, c'est-à-dire d'occurrences qui sont identiques ou se ressemblent graphiquement. Il peut s'agir de dates ou de symboles divers (incluant certaines punctuations), ou bien de mots graphiquement apparentés,

tels que *language* en anglais et *langue* en français. [SIMARD 1992] propose de considérer comme cognates des mots qui ont les mêmes quatre lettres initiales, ce qui exclut *government/gouvernement*. [MCENERY 1995] améliore la définition des cognates en calculant la similarité des deux mots à l'aide du coefficient de Dice, correspondant au double du rapport du nombre de lettres communes à la somme des lettres des deux mots.

La méthode des cognates trouve toutefois ses limites lors de l'alignement de couples de langues non apparentés. [CHURCH 1993b] faisait part de résultats plutôt encourageants sur l'alignement anglais-japonais, mais il s'est avéré que ceux-ci étaient dus au fait que le texte utilisé était un manuel technique contenant de nombreux exemples et termes techniques identiques dans les deux langues. [WU 1994a] observe que la corrélation entre les longueurs de phrases est bien moins bonne entre l'anglais et le chinois qu'entre l'anglais et le français, et il propose l'incorporation d'indices lexicaux propres au domaine. [FUNG 1994a] propose une méthode simple d'estimation d'un petit lexique bilingue d'amorçage à partir des textes à aligner (*K-Vec* amélioré en *DK-Vec* [FUNG 1994b]). [CHEN 1996] propose l'addition d'un modèle de traduction probabiliste à la technique de [BROWN 1991a]. [DAGAN 1993] utilise également un modèle de traduction, proposé par [BROWN 1993] pour un alignement en mots qui améliore notablement l'alignement en phrases.

[LANGLAIS 1997] montrent l'importance de la combinaison des différentes sources d'information (lexique, cognates, longueur des phrases, fréquence des types d'appariement⁴) et surtout l'importance d'un modèle adéquat pour cette combinaison. [MELAMED 2000] propose un algorithme qui combine également ces différentes sources d'information avec des méthodes efficaces de filtrage et de réduction de l'espace de recherche. Les deux dernières études ci-dessus ont montré leur efficacité dans l'évaluation du projet ARCADE, et semblent représenter l'état de l'art à ce point dans le temps : les systèmes atteignent désormais plus de 98.5% d'efficacité sur des textes "propres" (voir section 6.4.4 ci-dessous, et [VERONIS 2000]). Toutefois, bien que la robustesse des systèmes se soit incontestablement améliorée, les performances se dégradent encore très fortement en présence de textes "bruités" qui présentent des différences structurelles importantes (fragments manquants, inversions, etc.) qui sont malheureusement courantes dans les applications réelles.

Comme nous l'avons mentionné la quasi totalité des méthodes actuelles se fondent sur des hypothèses communes. En particulier, elles supposent que :

⁴ L'utilisation d'autres sources d'information a été proposée : Papageorgiou, Craniias & Piperidis [1994] utilisent un étiquetage grammatical, en faisant l'hypothèse que les parties du discours tendent à être respectées dans la traduction, mais cette voie a été peu explorée pour l'instant.

- l’ordre des phrases dans les deux textes est identique ou très proche ;
- les textes contiennent peu de suppressions ou d’adjonctions ;
- les alignements 1:1 sont très largement prépondérants et que les rares alignements m:n sont limités à de petites valeurs de m et n (typiquement 2).

Ces hypothèses sont rendues nécessaires pour des raisons d’efficacité, mais elles rendent les systèmes vulnérables aux différences structurelles importantes. [FLUHR 2000] propose une approche originale qui permet de s’affranchir des hypothèses mentionnées ci-dessus. Les textes ne sont plus traités séquentiellement, mais sont transformés en bases de données, qui sont alors traitées comme un système de recherche d’information : les auteurs ramènent le problème de l’alignement de phrases à celui d’une interrogation documentaire multilingue, dont le but est de ramener la phrase la plus similaire dans le texte cible à partir de la “ requête ” que constitue la phrase du texte source. Si certains détails restent à régler dans cette approche, elle semble constituer une direction de recherche très prometteuse.

On notera enfin que d’autres directions de recherches restent ouvertes ou ont été à peine explorées. Ainsi, la plupart des études portent sur l’alignement de paires de textes, alors que de nombreuses sources (par exemple, de façon typique, les traductions dans le cadre de l’Union Européenne) font intervenir plus de deux langues. [SIMARD 2000] montre que l’alignement par paires n’est pas optimal, et que l’alignement simultané de plusieurs langages peut améliorer le résultat global.

6.4.2. Alignement de mots et expressions

La section précédente a montré que bien des méthodes d’alignement de phrases utilisaient comme point d’ancrage un alignement (souvent partiel) des mots [KAY 1988 ; DEBILI 1992 ; DAGAN 1993 ; FUNG 1994a]. A l’inverse, l’alignement de phrases peut être un point de départ à l’alignement en mots. La circularité, qui a été notée par divers auteurs, tels que [DEBILI 1992], n’est qu’apparente : diverses techniques, d’amorçage, de relaxation, etc. permettent en fait un calcul concomitant des deux types d’alignements. Cependant, dans le cadre de l’alignement de phrases, l’alignement des mots n’est pas le but premier, et n’est au mieux qu’un produit dérivé qui n’a pas forcément une importance en tant que tel. Lorsque l’alignement de mots est le but premier⁵, on ne peut plus se satisfaire d’alignements grossiers et partiellement erronés. Divers auteurs se sont donc intéressés directement au problème du filtrage du bruit dans les alignements et extractions [DAGAN 1994 ; RESNIK 1997 ; JONES 1997 ; CHOUKA 2000 ; FUNG 2000].

⁵ Ou l’extraction de lexiques bilingues. Du point de vue technique, les deux problèmes ne sont pas strictement identiques, mais suffisamment apparentés pour que nous ne fassions pas la distinction dans le cadre de cette introduction.

Cependant, les techniques d'ancrage lexical propres à l'alignement de phrases mettent le plus souvent en jeu des occurrences isolées qui sont loin d'être satisfaisantes : les textes sont fortement constitués d'occurrences en rapport complexe : mots composés, locutions, phraséologie, et aucun alignement ou extraction ne peut sérieusement être fait au niveau lexical sans prendre en compte ces phénomènes. L'alignement des unités complexes est d'ailleurs très souvent l'un des buts recherchés, notamment en terminologie (voir section 6.2.1). L'alignement de certains couples de langues est d'ailleurs rendu particulièrement difficile par la présence très importante de mots composés, comme le montrent [AHRENBURG 2000] à propos du suédois, et [BLANK 2000] à propos de l'allemand (voir aussi [VAN DER EIJK 1993 ; JONES 1994]). Par ailleurs, le problème se complexifie encore par la présence des nombreux mots grammaticaux (à peu près 50% des occurrences de n'importe quel texte), dont la traduction est encore moins biunivoque (*one-to-one*) que celle des mots pleins : ceux-ci se traduisent souvent par un affixe dans la langue cible, par une information positionnelle, voire ne se traduisent pas du tout. Ignorer les mots grammaticaux n'est pas totalement possible : ceux-ci sont souvent partie intégrante des expressions terminologiques ou phraséologies à repérer.

L'alignement ou l'extraction de lexiques peut se décomposer, au moins conceptuellement, en deux aspects : il s'agit de repérer les mots et expressions du texte source et du texte cible, puis de les mettre en correspondance. Du point de vue pratique, ces deux tâches ne peuvent pas être totalement modularisées : la détermination des unités dans la langue source est dépendante de la langue cible (par exemple, il faut aligner d'un bloc *demande de brevet* et *Patentanmeldung* [BLANK 2000], alors que l'alignement peut se fractionner avec *domanda di brevetto*). Diverses méthodes statistiques ont été proposées pour la sélection d'expressions complexes dans une langue [LAFON 1984 ; CHURCH 1990 ; SMADJA 1990]. Cependant, les méthodes purement statistiques se heurtent à des difficultés importantes. D'une part, la rareté et la non-normalité des "événements" lexicaux observés rend le choix de statistiques extrêmement délicat. D'autre part, la plupart des expressions sont seulement "semi-figées", et supportent un certain nombre d'opérations linguistiques (flexion, insertion d'adjectifs et d'adverbes, passivisation, etc.), qui mettent en défaut au moins les modèles statistiques simples. Divers auteurs ont donc proposé soit des approches purement linguistiques, soit la combinaison des méthodes statistiques avec de telles approches, généralement basées sur la reconnaissance de patrons et modèles (*patterns, templates*) à l'aide d'expressions régulières ou de grammaires LOCALES [JACQUEMIN 1991 ; BOURIGAULT 1992 ; SMADJA 1993 ; DAILLE 1994]. Ce type de technique a été appliqué avec un certain succès à l'alignement ou l'extraction bilingues [DAILLE 1994 ; SMADJA 1996 ; MCENERY 1997 ; BLANK 2000 ; PIPERIDIS 2000].

L'introduction de connaissances linguistiques est toutefois relativement coûteuse, et n'est pas indépendante des langues. Divers auteurs continuent donc à rechercher

une amélioration des méthodes purement ou principalement statistiques, et des progrès remarquables ont été récemment faits dans ce domaine [MELAMED, 1998b ; WU 1997 ; HIEMSTRA 1998 ; AHRENBURG 2000 ; GAUSSIER 2000].

6.4.3. Autres types d'alignement

A la suite de Church [1993a], de nombreux auteurs ont essayé de s'affranchir du "bruit" que peuvent contenir les corpus et qui dégrade fortement les performances. Toutefois, dans le cas de Church [1993a] et de plusieurs autres études que nous avons citées précédemment, les divergences entre les deux parties du bitexte sont effectivement un véritable bruit (par exemple dû à la mauvaise détection des ponctuations par des techniques d'OCR), mais dans de nombreux autres cas, les différences sont de nature structurelle, à cause par exemple d'une présentation différente des documents, qui peuvent être parfois explicitement marquées lorsque les documents sont codés dans des langages de représentations tels que SGML ou XML. Peu d'études (publiées) se sont intéressées à ce problème, pourtant fondamental pour l'alignement de "vrais" documents (voir [ROMARY 2000]).

Enfin, un domaine semble prendre une importance grandissante, celui de l'alignement de segments linguistiques supérieurs aux mots ou termes, et inférieurs à la phrase : clauses, fragments d'arbres syntaxiques, squelettes de phrases (skeleton sentences). Un tel type d'alignement serait très intéressant pour diverses applications, et notamment pour la traduction basée sur l'exemple, l'enseignement des langues et la linguistique comparative. Le problème est extrêmement difficile, à cause de la difficulté de détecter des frontières de clauses dans chaque langue, de la difficulté de l'analyse syntaxique, même partielle, et des grandes divergences de structures entre langues, mêmes apparentées. Diverses études ont cependant commencé à aborder ce problème (qui forme en fait un continuum avec le problème de l'alignement d'expressions décrit dans la section précédente) [KAJI 1992 ; MATSUMOTO 1993 ; GRISHMAN 1994 ; PAPAGEORGIOU 1997]. [PIPERIDIS 2000 ; WU 2000] donnent une idée de l'état d'avancement de ce champ de recherches.

6.4.4. Evaluation

Jusqu'à une période récente, l'évaluation des performances des systèmes d'alignement était relativement imprécise. Il faut pourtant rendre justice à ce domaine de recherche, car il a une tendance sans doute plus prononcée que d'autres dans le traitement automatique des langues à fournir des évaluations quantitatives. Cependant à part quelques rares exceptions (comme [BLANK 1995]), les évaluations fournies sont principalement des auto-évaluations, difficilement comparables entre

elles. D'une part, les corpus utilisés sont différents d'une étude à l'autre⁶ et souvent peu d'éléments sont connus sur leur composition exacte et leurs particularités. D'autre part, ni le mode de constitution des alignements de référence, ni le mode de jugement par rapport à la référence sont communs, et comme dans toutes les applications linguistiques, leur variabilité a des chances d'être élevée. C'est ce qui a motivé la mise en place à partir de 1995 du projet ARCADE [VERONIS 2000]. Ce projet est destiné à évaluer les performances des systèmes d'alignement. Dans un premier temps, le projet s'est consacré à l'alignement en phrases, puis s'est ouvert au problème de l'alignement en mots. Les buts du projet sont d'établir des corpus-étalons pour l'évaluation de l'alignement, de développer des protocoles et mesures adaptés, et bien sûr de fournir un "instantané" sur les performances des systèmes actuels. Malgré sa jeunesse et ses limitations inévitables, le projet a déjà produit des avancées méthodologiques et des résultats appréciables [VERONIS 2000].

6.5. Conclusion

Le domaine du traitement des textes multilingues parallèles est relativement jeune, mais la présente introduction (et la taille de sa bibliographie, pourtant partielle) aura probablement convaincu le lecteur qu'il est en plein essor. L'explosion actuelle de l'information multilingue, à travers le Web, la globalisation des marchés, etc. fournit des sources de plus en plus considérables de ces "pierres de Rosette" électroniques que sont les textes parallèles, et constitue un moteur extraordinaire pour la recherche et le développement des technologies qui les exploitent.

Nous avons essayé de faire un tour d'horizon des nombreuses applications (constitution 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 ou la linguistique contrastive, etc.), ainsi que des techniques d'alignement qui rendent les textes parallèles utiles et exploitables (alignement de phrases, de mots et expressions complexes, etc.). Le grand nombre de publications consacrées aux textes parallèles montrent le chemin parcouru en à peine plus d'une dizaine d'années. Elles montrent aussi que de nombreuses difficultés subsistent, et que le champ de recherche dans ce domaine est encore largement ouvert.

⁶ Le *Hansard* qui a été utilisé dans beaucoup d'études est bien sûr devenu une sorte d'étalon *de facto*, mais il ne représente qu'un seul type de texte et un seul couple de langues

Remerciements

Je remercie chaleureusement tous les collègues qui ont bien voulu lire ce chapitre et me faire bénéficier de leurs remarques et en particulier David Hull et Elliott Macklovitch ; Je suis également redevable à René Fourneau et Serge Rosmorduc pour nos échanges sur les inscriptions antiques. Je remercie enfin mon étudiante Marie-Dominique Mahimon pour sa précieuse aide bibliographique.

6.6. Références

- [AHRENBURG 2000] AHRENBURG L., ANDERSSON M. & MERKEL M. A knowledge-lite approach to word alignment. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [ATKINS 1990] ATKINS B.T.S. Corpus Lexicography: The Bilingual Dimension. In Cignoni, L., & Peters, C. *Computational Lexicology and Lexicography* (Special issue dedicated to Bernard Quemada). *Linguistica Computazionale*, Vol. VI.
- [BARLOW 1995] BARLOW M. *ParaConc: A concordancer for parallel texts*. *Computers and Texts*, 10.
- [BARLOW 1996] BARLOW M. Parallel texts in language teaching. In McEnery A. M., Botley S. P., Glass J., Wilson A. (Eds.), *Corpora and language research : A selection of papers from Talc96*. UCREL Technical Papers Special Issue, Lancaster University.
- [BLANK 1995] BLANK I. Sentence alignment: methods and implementation. *Traitement automatique des langues*, 36 (1-2), 81-99.
- [BLANK 2000] BLANK I. Terminology extraction from parallel technical texts. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [BONHOMME 1995] BONHOMME P., ROMARY L. Projet de Concordances Parallèles Lingua : gestion de textes multilingues pour l'apprentissage des langues, *Actes des Quinzièmes Journées Internationales IA 95*, Montpellier.
- [BOURIGAULT 1992] BOURIGAULT D. Surface grammatical analysis for the extraction of terminological noun phrases. *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, France, 977-981.
- [BROWN 1990] BROWN P. F., COCKE J., DELLA PIETRA S., DELLA PIETRA V. J., JELINEK F., LAFFERTY J., MERCER R. L., ROOSIN P. A statistical approach to machine translation. *Computational Linguistics*, 16 (2), 79-85.
- [BROWN 1991a] BROWN P. F., DELLA PIETRA S., DELLA PIETRA V. J., & MERCER R. L. Word sense disambiguation using statistical methods. *Proceedings of the 29th Annual Meeting of Association for Computational Linguistics*, Berkeley, California, 264-270.
- [BROWN 1991b] BROWN P. F., LAI J. C., & MERCER R. L. Aligning Sentences in Parallel Corpora, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, 169-176.
- [BROWN 1993] BROWN P. F., DELLA PIETRA S., DELLA PIETRA V. J., & MERCER R. L. The mathematics of statistical machine translation : parameter estimation. *Computational Linguistics*, 19 (2), 263-311.

- [BROWN 1996] BROWN R. D. Example-Based Machine Translation in the Pangloss System. *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, 169-174.
- [CATIZONE 1993] CATIZONE R., RUSSELL G., WARWICK S. Deriving Translation Data from Bilingual Texts, *Proceedings of the First International Lexical Acquisition Workshop*. Detroit, 1-7.
- [CHEN 1996] CHEN S. *Building Probabilistic Models for Natural Language*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.
- [CHOUÉKA 2000] CHOUÉKA Y., CONLEY E.S, DAGAN I. A comprehensive bilingual word alignment system: Accommodating disparate languages: Hebrew and English. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [CHURCH 1990] CHURCH K. W., HANKS P. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16 (2), 22-29.
- [CHURCH 1993a] CHURCH K. Char_align: a program for aligning parallel texts at the character level. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio.
- [CHURCH 1993b] CHURCH K., DAGAN I., GALE W, FUNG P., HELFMAN J., SATISH B. Aligning parallel texts : do methods developed for English-French generalize to Asian languages ? *Proceedings of the Pacific Asia Conference on Formal and Computational Linguistics*, Taipei:Academica Sinica.
- [DAGAN 1993] DAGAN I., CHURCH K. W., GALE W. Robust Bilingual Word Alignment for Machine-Aided Translation. *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, 1-8.
- [DAGAN 1994] DAGAN I., CHURCH K. W. Termight : identifying and translating technical terminology. *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*, University of Stuttgart, Germany.
- [DAILLEb 1994] DAILLE B., GAUSSIER E., LANGE J.-M. Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan, 712-716.
- [DANIELSSON 1996] DANIELSSON P., RIDINGS D. *PEDANT. Parallel texts in Göteborg*. Språkbanken, Institutionen för svenska språket, Göteborgs universitet.
- [DAVIS 1995] DAVIS M. W., DUNNING T. E. A TREC evaluation of query translation methods for multi-lingual text retrieval. In Harman D. K. (Ed.), *The Fourth Text Retrieval Conference (TREC-4)*. NIST, <http://crl.nmsu.edu/ANG/MWD/Book2/trec4.ps>.
- [DEBILI 1992] DEBILI F., SAMMOUDA E. Appariement des Phrases de Textes Bilingues. *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, France, 517-538.
- [EBELING 1998A] EBELING J. The Translation Corpus Explorer: A browser for parallel texts. In Johansson S. Oksefjell S. (Eds.), *Corpora and Cross-linguistic Research: Theory, Method and Case Studies* (pp. 101-112). Amsterdam: Rodopi.
- [EBELING 1998b] Ebeling, J. Contrastive linguistics, translation, and parallel corpora. *Meta*, 43 (4), 602-615.

- [ERJAVEC 1995] ERJAVEC T., IDE N., PETKEVIC V., VERONIS J. Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. *TELRI , Proceedings of the First European Seminar, "Language resources for Language Technologies"*, Tihany, Hungary, 87-97.
- [EVANS 1991] EVANS D. A., HANDERSON S. K., MONARCH I. A., PEREIRO J., DELON L., HERSH W. R. *Mapping vocabularies using "latent semantics"*. Technical Report CMU-LCL-91-1, Carnegie Mellon University, Laboratory for Computational Linguistics.
- [FLUHR 1995] FLUHR C. Multilingual information retrieval. In Cole R. A., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Eds.) *Survey of the State of the Art in Human Language Technology* (pp. 391-405). Center for Spoken Language Understanding, Oregon Graduate Institute, <http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html>.
- [FLUHR 2000] FLUHR C., BISSON F., ELKATEB F. Mutual benefit of sentence/word alignment and crosslingual information retrieval. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [FOSTER 1997] FOSTER G., ISABELLE P., PLAMONDON P. Target-text mediated interactive machine translation. *Machine Translation*, 12 (1-2), 175-194.
- [FUNG 1994a] FUNG P., CHURCH K. K-vec: A new approach for aligning parallel texts, *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, 1096-1102.
- [FUNG 1994b] FUNG P., MCKEOWN K. Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping, *Proceedings of the Conference of the Association for Machine Translation in the Americas*. Columbia, MD, 81-88.
- [FUNG 2000] FUNG, P. (2000). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [GALE 1993a] GALE W. A., CHURCH K. W. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19 (3), 75-102.
- [GALE 1993b] GALE W. A., CHURCH K. W. , YAROWSKY D. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26, 415-439.
- [GARSIDE 1994] GARSIDE R., HUTCHINSON J., LEECH G.N., MCENERY A.M., OAKES M.P. The exploitation of parallel corpora in projects ET10/63 and CRATER. In Jones, D. (Ed.) *New Methods in Language Processing* (pp. 108-115), UMIST.
- [GAUSSIER 1995] GAUSSIER E., LANGE J.-M. Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique des Langues*, 36(1-2), 133-155.
- [GAUSSIER 2000] GAUSSIER E., HULL D., AIT-MOKHTAR S. Word alignment in use: Translation memory and cross-language retrieval. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [GRISHMAN 1994] GRISHMAN R. Iterative alignment of syntactic structures for a bilingual corpus. *Proceedings of the Second Annual Workshop on Very Large Corpora*, Kyoto, Japan, 57-68.

- [GRUNDY 1996] GRUNDY V. L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue. In Béjoint H., Thoiron Ph. *Les dictionnaires bilingues* (pp. 127-149). Louvain-la-Neuve : Duculot.
- [HARRIS 1988] HARRIS B. Are you bi-textual? *Language Technology*, 7, 41-41.
- [HARTMANN 1980] HARTMANN R.R.K. *Contrastive Textology. Comparative Discourse Analysis in Applied Linguistics* (Studies in Descriptive Linguistics 5). Heidelberg: J. Gross.
- [HIEMSTRA 1998] HIEMSTRA D. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. *Proceedings of the eighth CLIN meeting*, 41-58.
- [IDE 1994] IDE N., VERONIS J. MULTEXT (Multilingual Tools and Corpora). *Proceedings of the 14th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan.
- [IDE 1998] IDE N., VERONIS J. Introduction to the Special Issue on Word Sense Disambiguation: the State of the Art. *Computational Linguistic*, 24 (1), 1-40.
- [IMBS 1971] IMBS, P. *Trésor de la Langue Française. Dictionnaire de la langue du XIXè et du XXè siècles (1789-1960)*. Paris : Editions du CNRS
- [ISABELLE 1992] ISABELLE P. La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. *META*, 37 (4), 721-737.
- [ISABELLE 1993] ISABELLE P., DYMETMAN M., FOSTER G., JUTRAS J-M., MACKLOVITCH E., PERRAULT F., REN X., SIMARD M. Translation analysis and translation automation. *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'93)*, Kyoto, Japon.
- [ISAHARA 2000] ISAHARA H., HIRUNO, M. Japanese-English aligned bilingual corpora.
- [JACQUEMIN 1991] JACQUEMIN C. *Transformation des noms composés*. Unpublished doctoral dissertation, Université de Paris VII.
- [JAGTMAN 1994] JAGTMAN M. COMOLA: A computer system for the analysis of interlanguage data. *Second Language Research*, 10, 49-83.
- [JOHANSSON 1993] JOHANSSON S., EBELING J. , HOFLAND K. 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, 4-5 March 1994, pp. 85-112). Lund : Lund University Press.
- [JOHANSSON 1994] JOHANSSON S., HOFLAND K. Towards an English-Norwegian parallel corpus. In Fries U., Tottie G., Schneider P. (Eds.), *Creating and Using English Language Corpora* (pp. 25-37). Amsterdam: Rodopi.
- [JOHANSSON 1998] JOHANSSON S. On the role of corpora in cross-linguistic research. In Johansson S., Oksefjell S. (Eds.), *Corpora and Cross-linguistic Research: Theory, Method and Case Studies* (pp. 3-24). Amsterdam: Rodopi.
- [JONES 1994] JONES D., ALEXA M. Towards Automatically Aligning German Compounds with English Word Groups in an Example-Based Translation System. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, 66-7. Reprinted in Jones D. B., Somers H. L. (Eds) (1997), *New Methods in Language Processing* (pp. 199-206), London: UCL Press.

- [JONES 1997] JONES D. B., SOMERS H. L. Bilingual vocabulary estimation from noisy parallel corpora using variable bag estimation. In Mitkov R., Nicolov N. (Eds.) (1997). *Recent advances in natural language processing* (pp. 427-437). Amsterdam : John Benjamins.
- [KAJI 1992] KAJI H., KIDA Y., MORIMOTO, Y. Learning translation templates from bilingual text. *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, France, 672-678.
- [KAY 1980] KAY M. *The proper place of men and machines in translation*. Technical Report CSL-80-11, Xerox Palo Alto Research Center.
- [KAY 1988] KAY M., RÖSCHEISEN M. *Text-translation alignment*. Technical Report. Xerox Palo Alto Research Center.
- [KENNING 1999] KENNING M.-M. Parallel Concordancing and French Personal Pronouns, *Languages in Contrast*, 1 (1).
- [KJAERGAARD 1987] KJAERGAARD P. REFTEX. A context-based translation aid. *Proceedings of the 3rd conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, 109-112.
- [KLAVANS 1990] KLAVANS J., TZOUKERMAN E. The BICORD system: combining lexical information from bilingual corpora and machine-readable dictionaries. *Proceedings of the 12th International Conference on Computational Linguistics (COLING'90)*, Helsinki, Finland, 174-179.
- [KOUTSOUDAS 1957] KOUTSOUDAS A., HUMECKY A. Ambiguity of syntactic function resolved by linear context. *Word*, 13 (3), 403-414.
- [KUPIEC 1993] KUPIEC J. An algorithm for finding noun phrase correspondences in bilingual corpora. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 17-22.
- [LAFON 1984] LAFON P. *Dépouillements et statistiques en lexicométrie*. Genève : Slatkine-Champion.
- [LANDAUER 1990] LANDAUER T., LITTMAN M. Fully-automatic cross-language document retrieval using latent semantic indexing. *Proceedings of the 6th Conference of the UW Centre for the New OED*, Waterloo, Canada, 31-38.
- [LANGLAIS 1997] LANGLAIS PH., EL-BEZE M. Alignement de corpus bilingues : algorithmes et évaluation. *Actes de 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la langue de l'AUPELF-UREF (JST)*, Avignon, Avril 1997.
- [LANGLOIS 1996] LANGLOIS L. Bilingual Concordancers: A New Tool for Bilingual Lexicographers, *Second international conference of the Association for Machine Translation in the Americas (AMTA'96)*. Montréal, Canada.
- [LEON 1998] LEON J. Les premiers outils pour la traduction automatique. Demande sociale, technologie et linguistique (1948-1960). *Bulag*, 23, 273-295.
- [MACKLOVITCH 1992] MACKLOVITCH E. Corpus-based tools for translators. *Proceedings of the 33rd Annual Conference of the American Translators Association*, San Diego, California.
- [MACKLOVITCH 1995] MACKLOVITCH E. TransCheck - or the automatic validation of human translations. *Proceedings of the MT Summit V*, Luxembourg.

- [MATSUMOTO 1993] MATSUMOTO Y, ISHIMOTO H., UTSURO T., NAGAO M. Structural matching of parallel text. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 23-30.
- [MCENERY 1995] MCENERY A.M., OAKES M.P. Sentence and word alignment in the CRATER project : methods and assessment. *Proceedings of the EACL-SIGDAT Workshop*, Dublin.
- [MCENERY 1997] MCENERY A., LANGE J.-M., OAKES M., VERONIS J. The exploitation of multilingual annotated corpora for term extraction. In Garside R., Leech G., McEnery A. (Eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 220-230). London : Addison Wesley Longman.
- [MELAMED 1996] MELAMED I. D. Automatic detection of omissions in translations. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, Copenhagen, 764-769.
- [MELAMED 1997] MELAMED I. D. Automatic discovery of non-compositional compounds in parallel data. *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, Providence, RI, 97-108.
- [MELAMED 1998a] MELAMED I. D. *Manual Annotation of Translational Equivalence: The Blinker Project*, University of Pennsylvania (IRCS Technical Report #98-07).
- [MELAMED 1998b] MELAMED I. D. Word-to-word models of translational equivalence. *Computational Linguistics*,
- [MELAMED 2000] MELAMED I. D. Bitext maps and alignments via pattern recognition. In Véronis, J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [MELBY 1981] MELBY A. A bilingual concordance system and its use in linguistic studies. *Proceedings of the Eighth LACUS Forum*, Columbia, SC, 541-549.
- [NAGAO 1984] NAGAO M. A framework of mechanical translation between Japanese and English by analogy principle. In Elithorn A., Banerji R. (Eds.) *Artificial and human intelligence*. Elsevier Science Publishers, 173-180.
- [OARD 1994] OARD D. W., DECLARIS N., DORR B. J., FALOUTSOS C. (1994). On automatic filtering of multilingual texts. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, (volume 2, pp. 1645-1650). Available : <http://www.ee.umd.edu/medlab/fillter/papers/smc.ps>.
- [OARD 1996] OARD D. W., DORR B. J. *A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19*, University of Maryland, Institute for Advanced Computer Studies, April 1996. Available : <http://www.glue.umd.edu/~oard/research.html>.
- [PAPAGEORGIOU 1994] PAPAGEORGIOU H., CRANIAS L., PIPERIDIS S. Automatic Alignment in Parallel Corpora, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Student Session)*. Las Cruces, NM, 334-336.
- [PAPAGEORGIOU 1997] PAPAGEORGIOU H. Clause recognition in the framework of alignment. In Mitkov R., Nicolov N. (Eds.) (1997). *Recent advances in natural language processing* (pp. 417-425). Amsterdam : John Benjamins.
- [PAULUSSEN 1995] PAULUSSEN H. Compiling a trilingual parallel corpus. *Quarterly Newsletter of the Contrastive Grammar Research Group of the University of Gent*, 3. Available : <http://bank.rug.ac.be/contragram/newslet3.html>.

- [PICCHI 1992] PICCHI E., PETERS C., MARINAI E. A translator's workstation. *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, France, 972-976.
- [PIENEMANN 1992] PIENEMANN M. COALA - A computational system for interlanguage analysis. *Second Language Research*, 8, 59-92.
- [PIPERIDIS 2000] PIPERIDIS S., BOUTSIS S., PAPAGEORGIOU H. From sentences to words and clauses. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [RESNIK 1997] RESNIK PH., MELAMED I. D. Semi-automatic acquisition of domain-specific translation lexicons. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, Washington, DC, 340-347.
- [ROBERTS 1996] ROBERTS R.P., MONTGOMERY C. The Use of Corpora in Bilingual Lexicography. *Proceedings of EURALEX '96*.
- [ROMARY 1995] ROMARY L., MEHL N., WOOLLS D. The Lingua Parallel Concordancing Project: Managing Multilingual Texts for Educational Purpose, *Text Technology*, 5(3), 206-220.
- [ROMARY 2000] ROMARY L., BONHOMME P. Parallel alignment of structured documents. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [SADLER 1989] SADLER V. *The bilingual knowledge bank : a new conceptual basis for MT*. Technical report. Utrecht : BSO/Research.
- [SALKIE 1995] SALKIE R. Parallel Corpora, Translation Equivalence and Contrastive Linguistics. *Conference Abstracts: ACH/ALLC '95*, University of California, Santa Barbara, 106-108.
- [SATO 1990] SATO S., NAGAO M. Toward memory-based translation. *Proceedings of the 12th International Conference on Computational Linguistics, COLING'90*, Helsinki, Finland, 247-252.
- [SIMARD 1992] SIMARD M., FOSTER G., ISABELLE P. Using cognates to align sentences in bilingual corpora. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Montréal, Canada, 25-27 June 1992, 67-81.
- [SIMARD 1993] SIMARD M., FOSTER G. F., PERRAULT F. *TransSearch: a bilingual concordance tool*. Technical Report. Laval, Canada : Centre d'innovation en technologies de l'information.
- [SIMARD 1998] SIMARD M. The BAF: a corpus of English-French bitext. *Proceedings of First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 489-496.
- [SIMARD 2000] SIMARD M. Multilingual text alignment: Three languages are better than two. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [SINCLAIR 1987] SINCLAIR John (Ed.) *Looking up: An account of the COBUILD project in lexical computing*. London : Collins, 182pp.
- [SMADJA 1990] SMADJA F., MCKEOWN K. Automatically extracting and representing collocations for language generation. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania, 252-259.

- [SMADJA 1993] SMADJA F. Retrieving collocations from text : Xtract. *Computational Linguistics*, 19 (1), 143-177.
- [SMADJA 1996] SMADJA F., MCKEOWN K., HATZIVASSILOGLU V. Translation Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22 (1), 1-38.
- [SUMITA 1990] SUMITA E., IIDA H., KOHYAMA H. Translating with examples : a new approach to machine translation. *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI'90)*. Austin, Texas, 203-212.
- [VAN DER EIJK 1993] VAN DER EIJK P. Automating the acquisition of bilingual terminology. *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, Utrecht, 113-119.
- [VERONIS 2000] VERONIS J. & LANGLAIS Ph. Evaluation of parallel text alignment systems: ARCADE. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [WARWICK 1990] WARWICK S., G. RUSSELL Bilingual concordancing and bilingual lexicography. *Proceedings of the Fourth International EURALEX Conference*, Malaga,.
- [WEAVER 1949] WEAVER W. *Translation*. Mimeographed, 12 pp., July 15, 1949. Reprinted in Locke W. N., Booth, A. D. (1955) (Eds.), *Machine translation of languages* (pp. 15-23). New York : John Wiley & Sons.
- [WU 1994a] WU D. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, 80-87.
- [WU 1994b] WU D., XIA X. Learning an English-Chinese Lexicon from a Parallel Corpus. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland.
- [WU 1997] WU D. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23 (3), 377-404.
- [WU 2000] WU D. Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammars. In Véronis J. (Ed.). *Parallel Text Processing*. Dordrecht: Kluwer.
- [YANG 1998] YANG Y., CARBONELL J. G., BROWN R. D., FREDERKING R. E. Translingual Information Retrieval: Learning from Bilingual Corpora. *Artificial Intelligence Journal (Special issue: Best of IJCAI-97)*.
- [ZANETTIN 1994] ZANETTIN F. Parallel words: Designing a bilingual database for translation activities. In Wilson A., McEnery A. M. (Eds.), *Corpora and language research : A selection of papers from Talc94*. UCREL Technical Papers Special Issue, Lancaster University.

Résumé : Ce chapitre présente les diverses applications des textes parallèles, c'est-à-dire de textes accompagnés de leur traduction : lexicographie et terminologie, traduction humaine ou par machine, recherche d'information multilingue, enseignement des langues, études linguistiques contrastives, etc. Les principaux corpus de textes parallèles sont ensuite brièvement présentés. Le chapitre fait enfin un tour d'horizon des principales techniques d'alignement de textes parallèles (c'est-à-dire la mise en correspondance de différents niveaux d'unités : phrases, mots ou expressions, clauses, etc.), et de leur évaluation.

Mots clés : Textes parallèles, corpus multilingues, techniques d'alignement, évaluation.

- [AARTS 1990] AARTS J. Corpus linguistics: An appraisal. In Hammesse J., Zampolli A. (Eds.), *Computers in literary and linguistic research* (pp. 13-28). Paris-Genève: Champion Slatkine.
- Arad, I. (1991). *A quasi-statistical approach to automatic generation of linguistic knowledge*. Unpublished dissertation. UMIST, Manchester.
- Andrews, C. (1981). *The Rosetta Stone*. London: British Museum Publications. [traduction française: *La Pierre de Rosette*, London: British Museum Publications, 1993].
- Bonfante, L., Chadwick, J., Cook, B.F., Davies, W.V., Healey, J.F., Hooker, J.T., & Walker, C.B.F. (1990). *Reading the past: Ancient Writing from Cuneiform to the Alphabet*. London: British Museum Publications. [traduction française: *La naissance des écritures: du cunéiforme à l'alphabet*, Paris: Seuil, 1994].
- Brown, P. F., Cocke, J., Della Pietra, S., Della Pietra, V. J., Jelinek, F., Mercer, R. L., & Roosin, P. (1988). A statistical approach to machine translation. *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88)*, Budapest, 71-76.
- Brown, P. F., Della Pietra, S., Della Pietra, V. J., Lafferty, J., & Mercer, R. L. (1992). Analysis, statistical transfer, and synthesis in machine translation. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92)*, Montréal.
- [BROWN à paraître] Brown R., Carbonell J. & Yang Y. Automatic dictionary extraction for cross-language information retrieval. *Parallel Text Processing*. Dordrecht: Kluwer.
- [DAILLE 1994a] DAILLE B. Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *Unpublished doctoral dissertation*, Université de Paris VII.
- Davis, M. W., & Dunning, T. E. (1995a). Query translation using evolutionary programming for multi-lingual information retrieval. *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, Available : <http://crl.nmsu.edu/ANG/MWD/Book2/evolmltr1.ps.gz>.
- Davis, M. W., & Dunning, T. E. (1996). Query translation using evolutionary programming for multilingual information retrieval II. *Proceedings of the Fifth Conference on Evolutionary Programming*. Available : <http://crl.nmsu.edu/ANG/MWD/Book2/ep96.ps>.
- [DECLARIS 1994] DECLARIS N., HARMAN D., FALOUTSOS C. DUMAIS S., OARD D. W. Information filtering and retrieval: Overview, issues and directions. *Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (Vol 1, pages 42-49). IEEE, <http://www.ee.umd.edu/medlab/filter/papers>.
- [Dunning 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1), 61-74.
- Fung, P., & McKeown, K. (1996). A Technical Word and Term Translation Aid using Noisy Parallel Corpora Across Language Groups. *The Machine Translation Journal*, Special Issue on New Tools for Human Translators, xxx, 53-87.
- Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, 177-184.
- Gale, W. A., Church, K. W. , & Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92)*, Montréal, 101-112.

- Gaussier E. (1995). Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues. *Thèse de Doctorat de l'Université Paris VII*.
- Harris, B. (1988b). Bi-texts: A new concept in translation theory. *Language Monthly*, 54, 8-10.
- Hartmann, R.R.K. 1994. The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography. *Proceedings of EURALEX '94*, Amsterdam: Vrije Universiteit, 291-297.
- Isabelle P. (1992b). Bi-textual aids for translators. *Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Waterloo, Canada.
- Kay, M., & Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19 (1), 121-142.
- [KNOWLES 1996] KNOWLES F. L'informatisation de la fabrication des dictionnaires bilingues. In Béjoint H., Thoiron Ph. *Les dictionnaires bilingues* (pp. 151-168). Louvain-la-Neuve : Duculot.
- Leech, G. (1991). The state of the art in corpus linguistics. In Aijmer, K., & Altenberg, B. (Eds.), *English corpus linguistics* (pp. 8-29). London: Longman.
- Léon, J. (1996-1997). Les premières machines à traduire (1948-1960) et la filiation cybernétique. *Bulag*, 22, 9-33.
- Macklovitch, E. (1991). The Translators's Workstation ... in plain prose. *Proceedings of the 32nd Annual Conference of the American Translators Association*, Salt Lake City, Utah.
- Macklovitch, E. (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.
- Macklovitch, E. (1995a). *Can terminological consistency be validated automatically?* Technical report. Laval, Canada : Centre d'innovation en technologies de l'information. 15 pages.
- [MACKLOVITCH 1996] MACKLOVITCH E., HANNAN M.L Line'Em Up: advances in alignment technology and their impact on translation support tools. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*, Montréal, Québec.
- Melamed, I. D. (1996b). Automatic construction of clean broad-coverage translation lexicons. *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA'96)*, Montreal, 125-134.
- [MELAMED 1997a] MELAMED I.D. *A scalable architecture for bilingual lexicography*. Dept. of Computer and Information Science Technical Report #MS-CIS-91-01, University of Pennsylvania.
- Melamed, I. D. (1997c). A word-to-word model of translational equivalence. *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL'97)*, Madrid, 490-497.
- Melby, A.K. (à paraître). Sharing of translation memory databases derived from parallel text. *Parallel Text Processing*. Dordrecht: Kluwer.

- Nerbonne, J. (à paraître). Parallel texts in computer-assisted language learning. *Parallel Text Processing*. Dordrecht: Kluwer.
- Resnik, P. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA-98)*, Langhorne, PA, October, 1998.
- Sadler, V. (1989b). *Translating with a simulated bilingual knowledge bank*. Technical report. Utrecht : BSO/Research.
- Santos, D. (à paraître). The translation network: A model for the fine-grained description of translations.v
- Singh, S., Mcenery, T. & Baker, P. (à paraître). Building a parallel corpus of Punjabi-English: a feasibility study. *Parallel Text Processing*. Dordrecht: Kluwer.
- [SUMITA 1988] SUMITA E., TSUTSUMI Y. *A translation aid system using flexible text-retrieval based on syntax matching*. TRL Research report TR-87-1019. Tokyo Research Laboratory, IBM.
- Yang, Y., Brown, R. D., Frederking, R. E., Carbonell, J. G., Geng, Y., & Lee, D. (1997). Bilingual-corpus Based Approaches to Translingual Information Retrieval. *Proceedings of MULSAIC'97*.

[voir par exemple les bilans de Aarts (1990) et Leech (1991)]