Qu'attendre de l'alignement de corpus multilingues ?

Olivier Kraif

Qu'attendre de l'alignement de corpus multilingues ?

Olivier Kraif

Olivier.Kraif@u-grenoble3.fr

Laboratoire de Linguistique et didactique des langues étrangères et maternelles (LIDILEM)

Université Stendhal Grenoble 3

1 Introduction

En 1992, Pierre Isabelle notait que le volume annuel des traductions effectuées au Canada dépassait le demi-milliard de mots. Il en concluait que si les traducteurs avaient accès à cette masse gigantesque de traductions, ils y trouveraient un véritable gisement d'exemples illustrant des problèmes concrets rencontrés dans le cadre de leur pratique professionnelle. Une quinzaine d'années plus tard, avec l'explosion soudaine du Web, le souhait formulé par Isabelle est réalisé au delà de toute espérance. Avec les organismes internationaux, tels que l'ONU, l'OMS ou l'UE, qui publient rapports, comptes-rendus et décisions législatives, traduits avec précision et rigueur en plusieurs langues officielles (le corpus de l'*Acquis Communautaire* [1] compte désormais 20 langues !) ; avec les projets issus de la mouvance du Logiciel Libre, qui publient des documentations techniques et des traductions collaboratives en de très nombreuses langues[2] ; avec les collections de textes littéraires numérisés, traduits - comme dans le projet Carmel (El-Bèze, 2006)[3] - et librement diffusés[4], la quantité de traductions disponibles dépasse - et de loin - le milliard de mots, et concerne bien d'autres paires de langues que l'anglais et le français. Aujourd'hui, tout internaute peut se constituer rapidement une collection importante de tels textes - originaux et traductions - que nous appellerons désormais *corpus multilingues parallèles*.

Que faire de cette "masse" de traduction, pour reprendre le terme d'Isabelle ? Si l'on veut pouvoir en tirer parti efficacement, ne serait-ce que comme réservoir d'exemples, la première opération à effectuer consiste à *aligner* les textes, c'est-à-dire à rendre explicite les correspondances entre segments en relation d'*équivalence traductionnelle*[5]. Cette relation, initialement définie entre deux textes pris globalement[6], est généralement décomposable au niveau d'unités textuelles plus petites : sections, paragraphes, phrases, voire syntagmes ou lexies. L'alignement résulte donc de la segmentation des textes au niveau d'un certain *grain* - généralement phrastique - et de la mise en correspondance des segments - ou groupes de segments - jugés équivalents. On aboutit alors à ce qu'on appelle, pour reprendre le terme de Harris (1988), un corpus *bi-textuel* - ou *multi-textuel* dans le cas multilingue.

D'un tel corpus, on pourra par exemple extraire une concordance bilingue en sélectionnant toutes les phrases contenant une certaine construction, et en donner la traduction avec les phrases ou groupes de phrases alignées. A titre d'illustration, les occurrences suivantes de l'expression figée *mettre sur pied*, avec les phrases alignées correspondantes, ont été extraites du corpus JOC[7]:

Dans le cadre de cette « Semaine » , chaque membre des réseaux (…) **a mis sur pied** un ensemble de manifestations … .
As part of the ` Week ' , each member of the different networks (…) organized a series of events … .

(…) il n' y aura pas de programmes d' action (…) pour aider à **mettre sur pied** des industries « stratégiques » avec l' argent de « Bruxelles »
(…) there will be no action programme (…) to help launch ` strategic ' industries with money from ` Brussels ' .

Quelle action de planification de la logistique indispensable et de prévision budgétaire entreprend-elle en vue de la **mise sur pied** des autres programmes d' aide qui s' annoncent dès à présent ?
What steps is the Commission taking to plan the necessary measures and ensure the availability of the necessary budgetary appropriations for further programmes of aid which already seem likely to be necessary ?

Comme le montrent ces exemples, la notion d'alignement est assez intuitive, et cependant plus complexe qu'il n'y paraît. Dans le dernier couple aligné ci-dessus, on peut se demander si les deux phrases, ainsi isolées, sont vraiment équivalentes. Le sens de "logistique" n'apparaît pas clairement dans la version anglaise, et se réfère sans doute à des éléments co-textuels. De même, la version anglaise insiste par de multiples répétitions sur l'aspect "necessary" des mesures à prendre. La construction du sens s'effectue au niveau textuel, et se manifeste par des macrostructures sémantiques caractérisant la *cohésion* (thématique, énonciative, stylistique, anaphorique, …) interne au texte, ainsi que sa *cohérence* vis-à-vis des référents extra-linguistiques[8]. En isolant une paire de phrases de son environnement co-textuel, on la prive du réseau de coréférences sémantiques sur lequel reposait en partie l'équivalence traductionnelle. De ce fait, il apparaît que la notion d'équivalence connaît un véritable continuum de degrés. Initialement définie au niveau global, le texte traduit devant assumer de manière complète les fonctions communicatives que lui assigne le traducteur, dans le respect du sens de la source, l'équivalence traductionnelle devient de plus en plus lacunaire et morcelée quand on descend aux niveaux de granularité inférieurs. Inutile de dire qu'au niveau des mots (ou des morphèmes), elle vole en éclat, même si elle persiste ici et là pour certaines unités. Pour une discussion approfondie de ces problèmes, voir Kraif (2002).

Pour simplifier, nous considérerons ici l'alignement d'un point de vue opératoire et applicatif, afin de répondre aux questions suivantes : que sait-on faire, en terme de technologie, et en quoi les résultats de ces opérations peuvent être utiles aux utilisateurs - traducteurs, étudiants, linguistes ou lexicographes ? Après un bref historique du domaine, nous verrons comment il est possible d'obtenir de bons résultats au niveau phrastique en s'appuyant sur des indices superficiels. Nous

examinerons ensuite ce qu'on peut attendre de l'alignement au niveau lexical. Nous aborderons enfin l'alignement d'un point de vue contrastif, pour la comparaison des langues considérées en tant que systèmes morphosyntaxiques et lexicosémantiques.

2 L'alignement en quelques dates

On assiste, avec les méthodes d'extraction de corpus alignés, au développement poussé de techniques dites *knowledge-poor*, qui ne s'intéressent qu'aux phénomènes les plus superficiels concernant le passage d'une langue à une autre. Il y a comme un retour (faut-il y voir une régression ?) aux idées fondatrices de Warren Weaver, qui en 1949 comparait la traduction à une opération de décryptage d'un code vers un autre. Comment ne pas s'étonner, 40 ans après le rapport ALPAC qui dénonçait cette vision réductrice de la traduction, du développement fourmillant des indices statistiques et des modèles mathématiques issus de la théorie de l'information, qui constituent la plus grande part des travaux dans le domaine du traitement des corpus bi-textuel ?

Ce paradigme s'est pourtant affirmé dès les premiers essais d'alignement automatique, qui datent de 1987 : Kay & Röscheisen (1988, 1993) implémentent alors une méthode basée sur la distribution des mots, en n'utilisant aucune source d'information en dehors des deux textes à aligner. Les auteurs montrent qu'en observant des cooccurrences de mots à l'intérieur de zones probablement correspondantes (le début et la fin des textes, ainsi que les zones se situant au même niveau, dans chacun des textes) il est possible d'extraire des correspondances lexicales, qui peuvent servir ensuite de « points d'ancrage » pour aligner les phrases. Le grand mérite de ces premières recherches est de montrer qu'il est possible d'aligner sans passer par le sens, en se basant sur des propriétés purement formelles. Dans le même esprit, un autre type de technique se fait jour : les travaux parallèles de Brown, Lai & Mercer (1991) et Gale & Church (1991, 1993) obtiennent de bons résultats en se basant sur l'observation des longueurs de phrase. Par ailleurs, les systèmes étudiés intègrent une modélisation des probabilités empiriques des différents types de regroupements (ou *transitions*, du type 1-1, 1-0, 0-1, 1-2, 2-1, etc.).

A la suite de ces travaux fondateurs deux principales directions sont ouvertes : l'utilisation d'ancrages lexicaux d'une part, et le développement des systèmes probabilistes modélisant les variations des longueurs de phrase d'autre part, formeront le noyau dur de toutes les techniques mises en œuvre par la suite.

Etonnamment, les indices superficiels semblent très efficaces. La « philosophie » qui guide ces recherches s'appuie sur un constat de bon sens : bien souvent, un humain peut aligner deux textes sans connaître les deux langues impliquées, uniquement en se basant sur des indices formels tels

que découpages en sections, longueurs des phrases, récurrences de certains couples d'unités, graphies ressemblantes, traduction des nombres et des noms propres, etc.
Même entre des langues génétiquement éloignées, les longueurs de phrases et les ressemblances superficielles suffisent en général à déterminer les regroupements phrastiques probables. L'exemple ci-dessous montre comment deux phrases en albanais peuvent être appariées à une phrase en français, à partir de ces quelques indices.

Il faudra développer les recherches de charbon à pouvoir **calorifique** plus élevé et transformable en **coke** , s ' employer partout à substituer le charbon aux **carburants** liquides et à le consommer avec économie . (174 car.)

Të zgjerohen kërkimet për qymyre me fuqi **kalorifike** më të lartë dhe të **kok**sifikueshme .
Të punohet kudo për zëvendësimin e **karburanteve** të lëngëta me qymyr dhe për kursimin e tij . (73+76=148 car.)

Ainsi, pour tirer parti des ancrages lexicaux, Church (1993) propose de se baser non sur les distributions, mais sur les ressemblances superficielles caractérisant les chaînes de caractères : l'observation de 4-grammes (suites de 4 caractères) est appliquée au repérage de mots apparentés (ou *cognats*), et les couples d'unités lexicales en correspondance supposée forment des nuages de points, qu'on peut ensuite filtrer pour extraire l'alignement des zones de forte densité. La comparaison des chaînes de caractères aboutit au développement de systèmes mixtes, intégrant à la fois des modèles probabilistes relatifs aux longueurs et des modèles orientés vers les ressemblances superficielles (Simard, Foster & Isabelle, 1992 ; Mc Enery & Oakes, 1995). A la suite de Kay & Röscheisen, Débili & Sammouda (1992) montrent qu'il n'y a pas de cercle vicieux dans le fait d'utiliser successivement l'alignement des mots pour aligner les phrases, et l'alignement des phrases pour aligner les mots : le processus converge vers un alignement de plus en plus précis, chaque étape apportant de nouvelles informations. Par ailleurs, le repérage des cognats se raffine petit à petit : Mc Enery & Oakes (1995) en proposent une caractérisation améliorée en faisant intervenir le coefficient Dice[9] dans la comparaison de deux chaînes.
Pour l'identification des ancrages lexicaux, l'étude des distributions lexicales donne également de bons résultats. Fung & Church (1994) proposent une méthode simple basée sur un pré-découpage grossier des deux textes en zones d'égales importances : les occurrences et cooccurrences des unités dans les zones correspondantes permettent alors d'établir de manière fiable une liste d'unités équivalentes pouvant servir d'amorçage à un processus itératif du type de celui décrit par Débili & Sammouda (1992). Chen (1993) élabore aussi une méthode d'alignement en se basant sur l'appariement des mots, en s'inspirant du modèle de traduction basé sur l'exemple développé par Brown *et al*. (1993).
Enfin, Davis, Dunning & Ogden (1995) montrent comment combiner différents types d'indices pour les intégrer dans un même cadre algorithmique. Avec une approche similaire, des résultats très

satisfaisants sont obtenus par Langlais & El-Beze (1997) : divers indices, basés sur les longueurs de phrases, les chaînes identiques (transfuges), les cognats, les probabilités de transitions, sont pondérés de façon à optimiser les performances. Melamed (1997) et Kraif (2001a) combinent aussi plusieurs indices, en utilisant des heuristiques adaptées pour réduire l'espace de recherche et minimiser les chances d'erreur.

On constate que les techniques sont nombreuses : les résultats obtenus à l'issue de la campagne du projet ARCADE (Véronis, 1997) montrent en outre qu'elles sont parvenues à maturité, certains auteurs considérant le problème de l'alignement phrastique comme étant pratiquement résolu. Mais une vision trop optimiste risque de masquer la véritable nature du problème : la difficulté de la tâche ne peut être évaluée dans l'abstrait, car elle dépend étroitement du type de traduction mise en jeu, et des techniques éprouvées peuvent se révéler calamiteuses sur un corpus spécifique. Pour affirmer le problème résolu, il faudrait avoir réglé la question de l'équivalence traductionnelle en général, et nous en sommes encore loin : entre les résultats récents et cet objectif théorique encore lointain, nous sommes convaincu qu'il reste une importante marge de progression.

3 Etat de l'art

Il existe aujourd'hui de nombreux logiciels d'alignement automatique. Certains sont des produits commerciaux, comme Trados WinAlign, ou Mindo de Babeling, d'autres sont issus de la recherche, et distribués gratuitement, comme K-vec++, Giza++, Plug aligner, ou Alinea, avec pour certains des licences de type *Logiciel libre*.

Nous décrirons ici plus en détail le logiciel Alinea, développé par nous, et distribué gratuitement. Ce logiciel a été évalué lors de la campagne d'évaluation ARCADE II (Chiao *et al*., 2006), ce qui permet d'avoir une estimation rigoureuse de ses performances.

3.1 Fonctionnement d'Alinea

Pour bien comprendre le principe de l'alignement phrastique, il est utile d'avoir une représentation géométrique du processus. Un alignement étant un ensemble de paires associant des phrases, ou groupes de phrases, chacune de ces paires peut être représentée par une petite surface rectangulaire dans l'espace bidimensionnel du bi-texte (surface proportionnelle aux longueurs des deux segments alignés). Ces regroupements sont appelés *transitions*, et correspondent à différents cas de correspondance : 1-1, 1-2, 2-1, etc. L'enchaînement de ces transitions est appelé un *chemin* d'alignement, qui serpente peu ou prou autour de la diagonale du bi-texte, comme le montre la figure 1. Le but des logiciels d'alignement est de trouver le "meilleur chemin possible", à la fois *complet* (i.e. sans ignorer de zones alignables), *exact* (i.e. en respectant la relation d'équivalence

entre segments appariés) et de *grain fin* (car il est moins intéressant d'aligner au niveau des chapitres ou des paragraphes qu'au niveau des phrases).

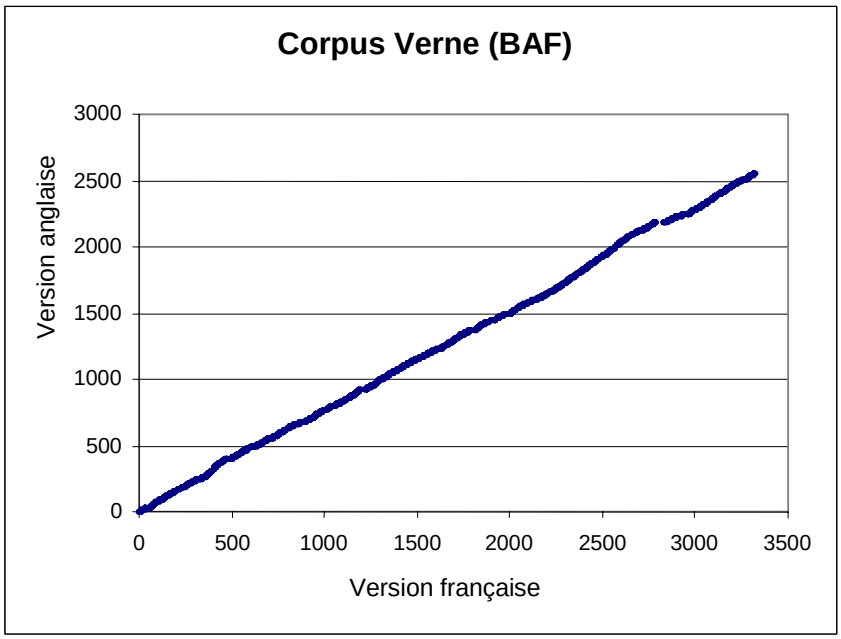

Figure 1 : chemin d'alignement extrait du corpus Verne (BAF)

Pour obtenir un tel alignement, Alinea procède en deux étapes :

- *l'extraction de points d'ancrage***,** qui vise à confiner l'espace de recherche à l'intérieur d'îlots de confiance. Ces points d'ancrage doivent être extrêmement fiables, car ils sont déterminants pour la suite. Alinea peut se baser sur des points d'ancrage explicites (balises XML), ou extraire des points d'ancrage probables en s'appuyant sur des réseaux d'appariements de *transfuges* concordants : nombres, noms propres, emprunts, etc. L'algorithme d'extraction est itératif et effectue un prédécoupage qui s'affine progressivement, en commençant par les transfuges les plus fiables (les nombres).

- *l'alignement* proprement dit, avec extraction du chemin complet optimal. On évalue la *probabilité* d'un chemin quelconque en fonction des *indices* disponibles : *rapport des longueurs*, appariement de chaînes identiques (*transfuges*), appariement de mots ressemblants (*cognats*), de mots possédant des *distributions* similaires ou de lexèmes équivalents. Un algorithme de programmation dynamique se charge alors d'extraire le chemin qui maximise cette probabilité. Pour optimiser les calculs, on se base en général sur un jeu de 8 transitions prédéfinies : 1-1, 1-0, 0-1, 2-1, 1-2, 2-2, 3-1, 1-3. Alinea permet cependant, lorsque les textes présentent des segmentations très divergentes, d'élargir la recherche à des transitions quelconques.

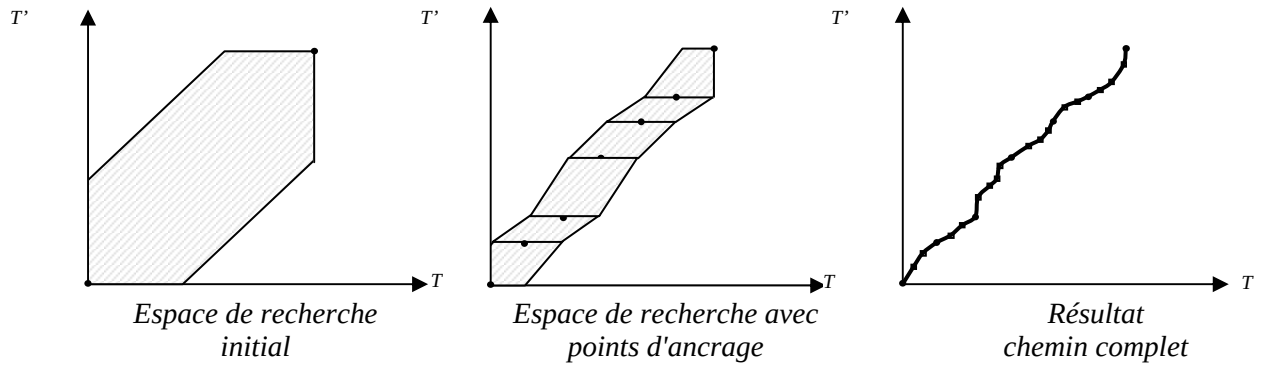

Figure 2 : réduction de l'espace de recherche avec des points d'ancrage

Notons que la pertinence relative des indices d'alignement dépend fortement de la typologie des textes et du couple de langues : entre l'anglais et le français, par exemple, la fréquence des cognats est très importantes à l'intérieur de phrases alignées. Des textes riches en nombres seront également plus faciles à aligner. Dans l'exemple ci-dessous, ces indices sont très nombreux - il s'agit donc d'un cas de figure particulièrement favorable :

Le Comité préparatoire du cinquantième anniversaire de l ' Organisation des Nations Unies , que l ' Assemblée générale a créé par sa décision 46/472 du 13 avril 1992 , s ' est réuni cinq fois et s ' est mis d ' accord sur le choix d ' un thème

The Preparatory Committee for the Fiftieth Anniversary of the United Nations , established by the General Assembly in decision 46/472 of 13 April 1992 , held five meetings.
Agreement was reached by consensus on a theme for the anniversary

Transfuges : *(Nations,Nations), (46/472,46/472), (13,13), (1992,1992)*
Cognats: *(Comité, Committee) (préparatoire, Preparatory) (anniversaire, Anniversary) (Nations, Nations) (Unies, United) (Assemblée, Assembly) (générale, General) (décision, decision) (avril, April) (thème, theme)*

Pour identifier les cognats, Alinea se base sur le calcul de la plus longue sous chaîne commune (par exemple "préparatoire" et "preparatory" partagent une sous-chaîne de longueur 9 : p-r-p-a-r-a-t-o-r), ce qui ne nécessite aucune donnée linguistique. Quand on aligne des langues à alphabets différents, on ne peut plus s'appuyer aussi simplement sur les comparaisons de caractères : mais comme le montrent les résultats de la campagne ARCADE II, la plupart des traductions contiennent suffisamment de chaînes empruntées au système graphique de la langue source (nombres, sigles ou noms propres) pour qu'on puisse tirer parti de ces indices. Des prétraitements, tels que la translittération des nombres arabo-indiens, peuvent également être requis afin d'améliorer les résultats.

Enfin Alinea permet d'ajuster les paramètres concernant le rapport des longueurs et la pondération des différents indices, afin de s'adapter aux spécificités de chaque paires de langues. Lorsque les indices de surfaces sont vraiment insuffisants, Alinea permet d'utiliser des ressources langagières, de type lexique bilingue, afin de compenser ce déficit d'information.

3.2 Résultats d'Alinea

Les premiers résultats de la campagne ARCADE II sont publiés dans Chiao *et al.* (2006). L'originalité de cette évaluation était de porter sur l'alignement du français avec d'une part des langues apparentées (anglais, allemand, espagnol, italien), et d'autre part des langues plus lointaines ou utilisant des alphabets différents (comme l'arabe, le chinois, le farsi, le grec, le japonais et le russe).

Pour le premier groupe, Alinea obtient des résultats corrects[10] à environ 98 %, à 3 dixièmes du meilleur système. Notons que les résultats sont meilleurs pour l'italien et l'espagnol que pour l'anglais et l'allemand, ce qui montre l'importance de la proximité génétique. Pour le second groupe, Alinea obtient les meilleurs résultats (mais seul un autre système était en compétition), avec une moyenne de 87,1 % : la dégradation des performances est avérée, mais pas catastrophique. Il existe tout un continuum entre les couples les plus propices (comme le français et le grec, avec 97,6 %) et les plus problématiques (comme le français et le japonais, avec seulement 78,9 % de correction). Notons que pour obtenir ces résultats, nous n'avons utilisé ni lexique bilingue, ni outil de translittération : le seul prétraitement était la segmentation en phrase.

Il existe une marge de progression réelle pour l'utilisateur qui prend le temps de régler ses paramétrages et d'enrichir Alinea d'un lexique bilingue (ce logiciel permettant aussi de constituer automatiquement ses propres ressources linguistiques). Quel que soit le couple de langues, on peut donc escompter des résultats compris dans une fourchette de 90 % à 99 % pour les techniques décrites ci-dessus, avec des traductions respectant les critères de parallélisme (sans omission ou ajout massifs).

4 Du phrase à phrase au … mot à mot ?

L'alignement lexical constitue également un défi d'importance pour les systèmes d'alignement. Comme nous l'avons montré par ailleurs (Kraif, 2002), il y a une solution de continuité de la phrase au mot : une forte proportion des unités lexicales n'ont pas, en général, d'équivalent strict dans le texte traduit. En revanche on peut parler de *correspondance lexicale* : certaines unités conservent leur stabilité référentielle lors du passage à la traduction, et comme les "raisins dans la brioche" (pour reprendre la métaphore de Seleskovitch, citée par Laplace, 1994), restent individualisables malgré la "chimie du sens" opérée par la traduction.

Or, ces correspondances peuvent être extraites de manière automatique, en se basant sur l'observation comparée des distributions des unités dans une vaste collection de textes alignés. En effet, connaissant les fréquence f_1 et f_2 de deux unités en langue source et en langue cible, il est

possible de calculer leur fréquence de cooccurrence théorique F_{12} dans l'hypothèse où ces deux unités seraient indépendantes (cas où les cooccurrences seraient dues au hasard).

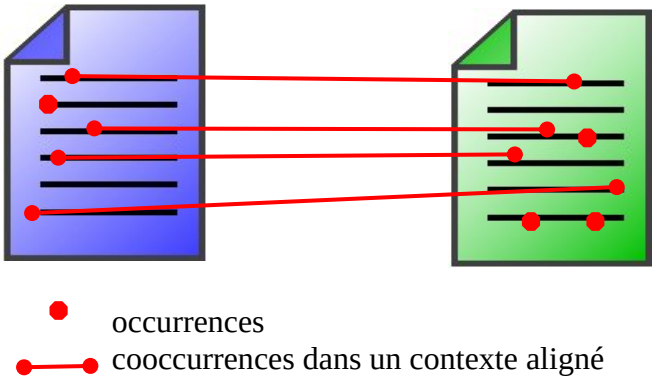

Figure 3 : Comptage des occurrences et cooccurrences

$f_1 = 5, f_2 = 7, f_{12} = 4$

Dans le cas d'unités équivalentes (p. ex. "anniversaire" et "anniversary") la fréquence observée f_{12} est en général bien supérieure à la fréquence attendue F_{12}. Différents indices statistiques, tels que l'information mutuelle, le t-score, le rapport de vraisemblance, ou la probabilité de l'hypothèse nulle, permettent de mesurer ce degré d'association entre deux unités source et cible.

Pour Alinea, nous avons construit un indice mixte pour extraire, au sein de deux phrases alignées, les couples d'unités obtenant les meilleurs scores. En se basant sur les distributions, les ressemblances formelles, les positions dans les phrases alignées et l'identité des parties du discours, nous avons montré, dans Kraif & Chen (2004), qu'on pouvait obtenir des appariements corrects à 90% pour 88% des unités possédant des correspondances (en négligeant les mots vides : articles, prépositions, etc.), sur un corpus français anglais (un extrait de *Madame Bovary*).

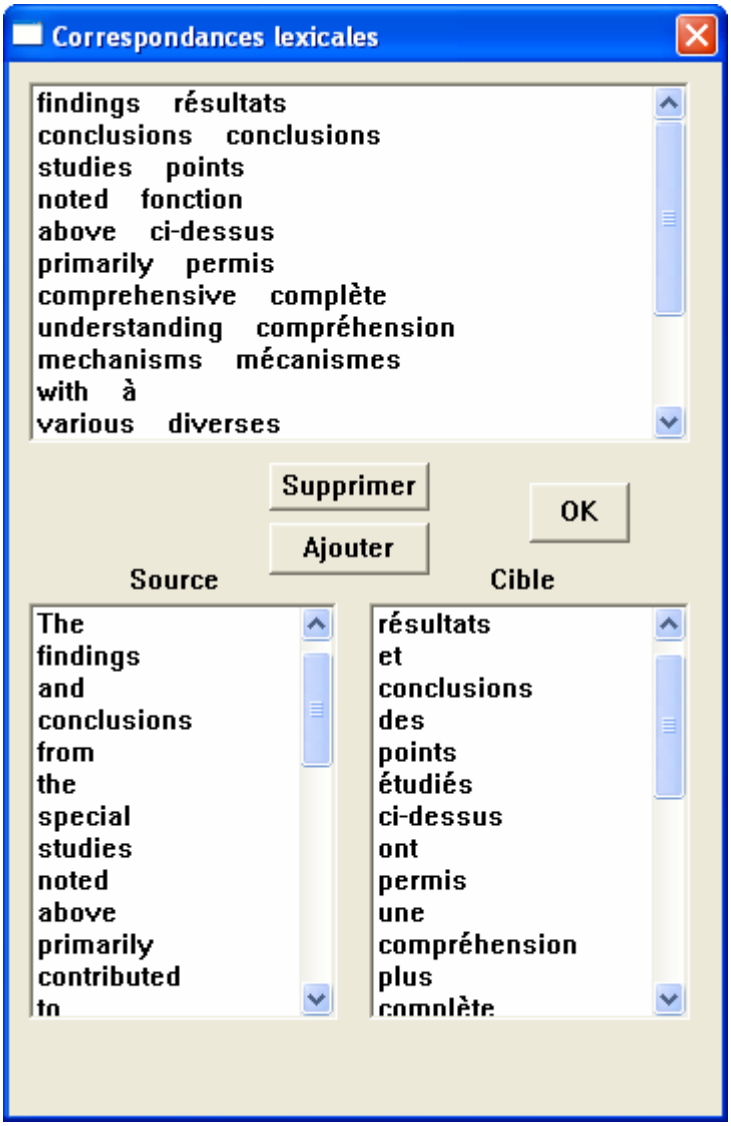

Figure 4 : Correspondances lexicales extraites par Alinea

Les résultats de telles extractions dépendent avant tout de la taille du corpus d'apprentissage, duquel les statistiques d'occurrence et de cooccurrence ont été tirées : il faut compter au moins un million de mots dans chaque langue pour des résultats corrects à 80 %. Par ailleurs, la possibilité de faire correspondre des unités polylexicales (comme "à cause de") ainsi que l'usage de corpus lemmatisés (Alinea peut traiter des corpus comportant lemmatisation et étiquetage morphosyntaxique[11]), permet d'obtenir une amélioration notable des résultats.

A partir des correspondances ainsi obtenues, il est également possible d'extraire automatiquement un lexique bilingue spécifique à un corpus : pour filtrer le bruit et éliminer les correspondances trop liées à leur co-texte, il suffit de retenir les correspondances observées avec une fréquence statistique significative. Dans l'exemple ci-dessous, seules les correspondances observées plus de 3 fois ont été retenues.

during-PRE	pendant-PRE (6)	eight-QUA	huit-QUA (4)
dust-NOM	poussière-NOM (14)	eighty-QUA	quatre-QUA (5)
dusty-ADJ	poussiéreux-ADJ (3)	elegance-NOM	élégance-NOM (3)
dwarf-NOM	rabougrir-VER (3)	eloquence-NOM	éloquence-NOM (3)
dye-PPS	teinter-PPS (3)	embarrass-PPS	embarras-NOM (5)
ear-NOM	oreille-NOM (18)	embarrasser-PPS (3)	
earth-NOM	terre-NOM (4)	employé-NOM	libre-ADJ (3)
eastern-ADJ	là-ADV (3)	empty-ADJ	vide-ADJ (6)
easy-ADJ	facile-ADJ (6)	encampment-NOM	campement-NOM (3)
eat-VER	manger-VER (6)	encumber-PPS	estimer-VER (3)
edict-NOM	le-DET (3)	engage-PPS	engager-PPS (3)
egg whisk-NOM	oeuf-NOM (3)	enough-ADV	assez-ADV (10)

Figure 5 : Extrait d'un lexique bilingue tiré d'un alignement anglais-français de with a Donkey in the Cevennes*, de Stevenson*

5 De l'alignement des textes à l'alignement des langues

Au vu du lexique ainsi obtenu, on constate qu'il reste du bruit, souvent lié à des problèmes d'identification des unités polylexicales (comme "egg whisk" <-> "œuf", ou "eighty" <-> "quatre"). Ce bruit pourrait cependant être aisément écarté pour des corpus de grande dimension : à mesure que les données deviennent statistiquement plus significatives, les régularités émergent et se distinguent des associations bruitées, plus instables par nature. Par ailleurs, les effets "textuels", liés à l'idiosyncrasie d'un texte précis, à son sujet, aux habitudes de l'auteur, aux choix du traducteur, etc., s'estompent à mesure que le corpus augmente et devient plus représentatif de la langue générale (ou d'une langue de spécialité si l'on vise un corpus spécialisé).

Comme dans toute recherche de linguistique de corpus, on peut alors partir de l'observation du texte pour viser la langue. De ce point de vue, les bi-textes ne permettent pas seulement d'étudier deux langues, prises du point de vue du code, mais de les confronter et de les éclairer réciproquement, en s'appuyant sur les structures et les régularités originales que font apparaître les contrastes.

Il est par exemple relativement aisé d'établir automatiquement des classes de synonymes, sur la base de la transitivité de la relation d'équivalence. La figure 6 montre les résultats d'une requête élaborée de manière itérative, en recherchant initialement l'expression "de temps en temps". Les couples de phrases trouvés pointent l'expression équivalente "from time to time". En recherchant cette dernière expression en anglais, de nouveaux couples de phrases sont identifiés, contenant d'autres équivalents en français "de temps à autre", "par instants". En cherchant ces nouvelles expressions, on trouve alors de nouveaux équivalents anglais "now and then", "ever and again"... On peut réitérer ce processus de l'aller-retour jusqu'à obtenir des classes stables. L'alignement contenant des appariements bruités, un filtrage est parfois nécessaire, afin de ne retenir que les équivalences les plus significatives, et constituer des classes réduites avec un noyau sémantique cohérent.

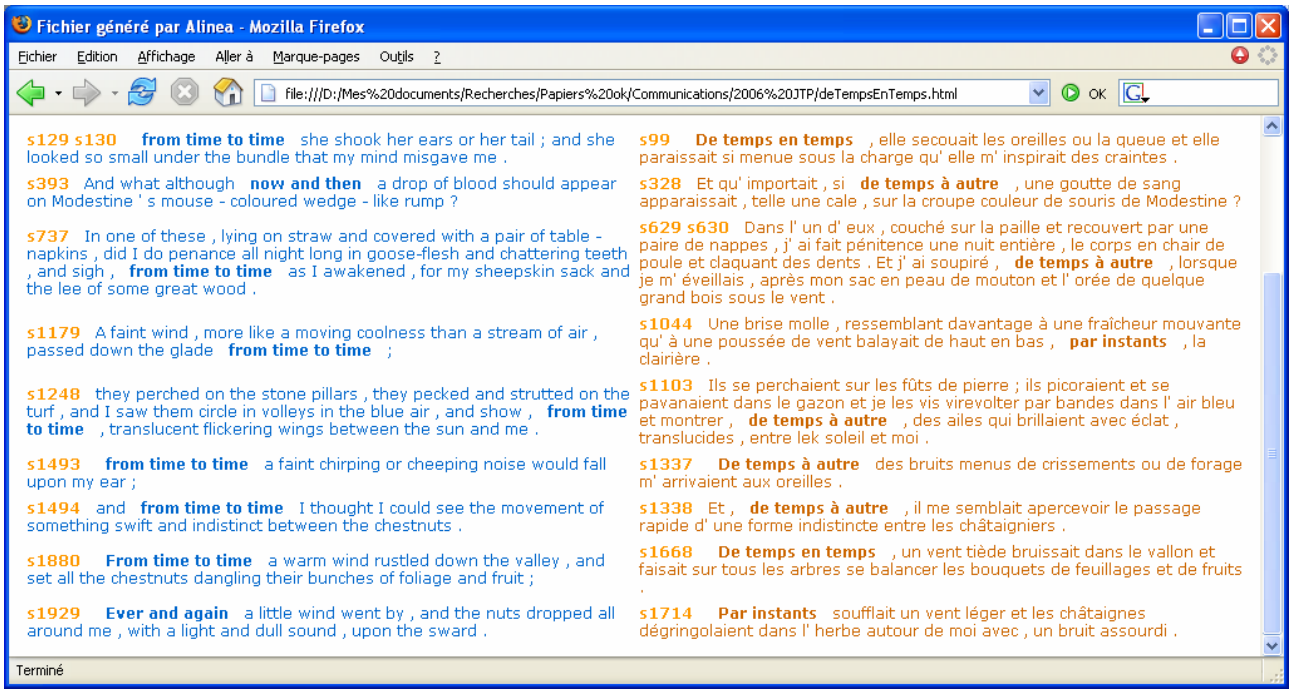

Figure 6 : Concordance extraite à partir d'une requête bilingue d'Alinea, sur la traduction de with a Donkey in the Cevennes, *de Stevenson*

Les classes ainsi obtenues à partir de l'équivalence traductionnelle représentent plus des communautés de voisinage sémantique que des relations stricte de synonymie, et elles transcendent souvent les catégories morphologiques. Par exemple, en appliquant l"aller-retour" au mot "âne" on obtient la classe suivante : *âne*, *ânesse*, *ânier*, *bourriquet*, *bourrique*, *bourricot*, *baudet*. On y trouve donc des embryons de paradigmes morphologiques qui pourraient servir d'appui à une étude sur la morphologie dérivationnelle et/ou flexionnelle.

Il est possible d'aller plus loin que la recherche de formes simples ou d'expressions figées. En appliquant ces outils de concordance à des corpus étiquetés et lemmatisés, comme Alinea le permet, il est possible de rechercher des expressions comportant un certain degré de variabilité morphosyntaxique (flexions, insertion de modifieurs, etc.). Par exemple, si l'on recherche les expressions du type "mettre en + NOM", on pourra obtenir les exemples suivants :

Requête : <base=mettre><ctag=adv>?<en><ctag=nom>
Résultats sur le corpus Stevenson :

…I buttoned myself into my coat , and walked on	Je boutonnai ma veste et me **mis en marche**
…but he flies up at you in a temper	… sans qu' il se **mette immédiatement en colère** contre vous
…and dispersed the generals of the king	… et **mettaient en fuite** les généraux du roi

Ces exemples illustrent que ces constructions sont des collocations à verbe support dont le sens est porté par la base nominale, ce qui peut bien sûr avoir des répercussions sur l'organisation des entrées dans la construction d'un dictionnaire.

6 Conclusion

Nous sommes convaincu que les outils d'alignement et de concordance bilingue sont encore largement sous-exploités. Certaines fonctionnalités, comme l'extraction automatique de lexiques bilingues, la construction de classes sémantiques, l'élaboration de requêtes tirant parti d'étiquetages morphosyntaxiques et de lemmatisation, sont encore relativement peu utilisées par la communauté des linguistes ou celle des traducteurs.

L'accès facilité à de grandes quantités de textes multilingues en ligne, ainsi que la disponibilité de certains outils qui franchissent le seuil des laboratoires, devrait conférer aux applications des techniques d'alignement un développement rapide.

Mais au-delà des prometteuses applications en aide à la traduction, en lexicographie, en terminologie ou en didactique des langues, l'usage des corpus alignés permet d'ouvrir un nouveau champ d'étude, où le traitement automatique des langues et la linguistique de corpus peuvent converger au service de la traductologie. Ces nouveaux outils d'alignement et de concordance permettront peut-être d'observer plus finement des régularités traductionnelles invisibles "à l'œil nu", mais statistiquement émergentes sur de grands corpus informatisés. Un coin du voile sera alors peut-être levé sur l'activité traduisante, si difficile à automatiser, que Martin Kay (2000) qualifie de " largement mystérieuse " :

> « *Perhaps the single most remarkable observation about machine translation is that it has attracted the attention of a vanishingly small number of researcher with some knowledge of traditional translation. And one of the most remarkable facts about translation as a field of inquiry is that it has very rarely been treated as an empirical enterprise. As a result, the literature on translation theory is replete with simplified versions of linguistic theories about morphology, syntax and semantics in the apparent belief that they have something to say about translation. But what translators actually do and how they do it remains largely mysterious.*»

7 Références

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. (1993a) The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, vol. 19, n. 2, pp. 263-311.

Brown, P., Lai, J., Mercer, R. (1991) Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91,* Morristown, NJ, pp. 169-176.

Chen, S. F. (1993) Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of ACL-93*, Columbus OH.

Church, K. W. (1993) Char align : A program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL-93*, Columbus Ohio, pp. 1-8.

Davis, M. W., Dunning T. E., Ogden W. C. (1995) Text Alignment in the Real World : Improving Alignments of Noisy Translations Using Common Lexical Features. In *Proceedings of EACL 95,* 8 p. (disp. à l'adresse : http://www.crl.nmsu.edu).

Débili, F., Sammouda, E. (1992) Appariement des phrases de textes bilingues Français - Anglais et Français - Arabe. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, 23-28 août 1992, pp. 518-524.

El-Bèze M., Richard C., Meyer R. (2006) Projet CARMEL : récits de voyages, in *Actes de TALN 2006*, Louvain-la-Neuve.

Fung, P., Church, K. W. (1994) K-vec : A New approach for Aligning Parallel Texts. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING-94*, Kyoto, pp. 1096-1102.

Gale, W., Church, K. W. (1993) A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, vol. 19, n. 1, pp. 75-91.

Harris, B. (1988) Are you Bi-Textual ? *Language Technology*, n° 7, pp. 41-41.

Isabelle, P. (1992) La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. *META*, Outremont, PQ, XXXVII, 4, pp. 721-731.

Kay M. (2000) Préface, in Jean Véronis Ed., *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers.

Kay M., Röscheisen, M. (1993) Text-Translation Alignement. *Computational Linguistics*, Morristown, NJ, vol. 19, n. 1, pp. 121-142.

Kraif O. (2001a) *Constitution et exploitation de bi-textes pour l'Aide à la traduction, Thèse de doctorat*, sous la dir. de Henri Zinglé, Université de Nice Sophia Antipolis

Kraif O. (2001b) Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation, *TAL 42 :3*, ATALA, Paris, pp. 833-867.

Kraif O. (2002) Translation alignment and lexical correspondences : a methodological reflection, *Lexis in Contrast*, eds. B. Altenberg & S. Granger, Benjamins Publisher, Amsterdam, pp. 271-290

Kraif O., Chen B. (2004) Combining clues for lexical level aligning using the Null hypothesis approach, in *Proceedings of Coling 2004*, Geneva, August 2004, pp. 1261-1264.

Langlais, P., El-Bèze, M. (1997) Alignement de corpus bilingues : algorithmes et évaluation. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 191-197.

Laplace, C. (1994) *Théorie du langage et théorie de la traduction*, Paris, Didier érudition.

McEnery, A. M., Oakes, M. P. (1995) Sentence and word alignment in the CRATER project : methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop*, Dublin.

Melamed, I. D. (1997) A Portable Algorithm for Mapping Bitext Correspondence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, 7-12 July 1997, pp. 305-312 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Simard, M., Foster, G., Isabelle, P. (1992) Using Cognates to Align Sentences in Bilingual Corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, TMI-92*, Montréal, CCRIT, pp. 67-81.

Véronis, J. (1997) Une action d'évaluation des systèmes d'alignement de textes multilingues. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 191-197.

8 Outils d'alignement

- Alinea : http://www.u-grenoble3.fr/kraif
- Giza++ : http://www.isi.edu/~och/GIZA++.html

- K-vec++ : http://www.d.umn.edu/~tpederse/parallel.html
- Mindo : http://www.babeling.com/accueil.html
- Plug aligner : http://stp.ling.uu.se/~corpora/plug/pwa/
- Trados WinAlign : http://www.translation.net/trados_winalign.html

9 Notes

[1] Le *ACQUIS COMMUNAUTAIRE Multilingual Corpus* est disponible en 20 langues à l'adresse : http://wt.jrc.it/lt/Acquis/ Il comporte environ 800 textes incluant l'ensemble des textes et des traités qui constituent le socle législatif de l'UE.

[2] Le corpus Opus, disponible à l'adresse http://logos.uio.no/opus/, contient des textes parallèles concernant jusqu'à 61 langues.

[3] Ce projet a permis de constituer un corpus de récits de voyage en quatre langues (français, anglais, italien, espagnol), alignés et comportant des annotations morphosyntaxiques, sémantiques et thématiques. Une partie du corpus est diffusé librement, à l'adresse : http://www.projetcarmel.org

[4] On trouve de nombreux textes sur le site du Projet Gutenberg (http://www.gutenberg.org/), sur le site de l'ABU (http://abu.cnam.fr/), et sur d'autres sites consacrés aux livres électroniques. *The Online Book Page* constitue un index assez complet des textes disponibles : http://onlinebooks.library.upenn.edu/ .

[5] Pour une discussion de ce concept, cf. Kraif (2001), pp. 38-63.

[6] Pour F. Israël (in Lederer & Israël, 1991 : 22) « l'unité de traduction n'est plus le mot, le syntagme ou la phrase mais le texte tout entier. »

[7] Le corpus JOC est composé de textes parallèles en neuf langues faisant partie du Journal Officiel de la Commission européenne (série C, année 1993). Les textes (au nombre de plusieurs milliers) sont constitués de questions écrites des parlementaires européens sur un large éventail de sujets, et des réponses correspondantes de la Commission européenne. La taille du corpus est d'environ 10,2 millions de mots, collectés et préparés dans le cadre des projets MLCC et MULTEXT. Le corpus est actuellement distribué sous licence par ELDA. Cf. l'adresse suivante : http://www.elda.org/catalogue/fr/text/W0017.html

[8] Ce que Rastier englobe simplement sous le terme de "textualité", cf. Rastier, F. (1987) Microsémantique et textualité. In Charolles M., Petöfi J.S., Sözer E. (Ed.), *Research in Text Connexity and Text Coherence*, Hambourg, Helmut Buske Verlag, p. 147.

[9] Pour deux chaînes de longueur n_1 et n_2, comportant n_{12} caractères communs, on a : $Dice = 2*n_{12}/(n_1+n_2)$

[10] La correction étant mesurée sur la base de la F-mesure, combinant précision et rappel.

[11] L'étiquetage morphosyntaxique et la lemmatisation sont des techniques de base du traitement automatique des langues. L'étiquetage consiste à identifier les traits morphosyntaxiques (parties du discours, flexions, etc.) de formes apparaissant dans un texte. La lemmatisation consiste à identifier la forme canonique (le lemme) d'une forme fléchie (p. ex. l'infinitif s'il s'agit d'un verbe). On arrive à obtenir automatiquement plus de 95% de correction sur des langues comme l'anglais ou le français.