Extraction automatique de correspondances lexicales : évaluation d'indices et d'algorithmes

Olivier Kraif

LILLA, Université de Nice Sophia Antipolis, 98 Bd. E. Herriot BP 369 06007 Nice Cedex
okraif@mageos.fr
http://lilla2.unice.fr

Résumé

Les bi-textes sont des corpus bilingues parallèles, généralement segmentés et alignés au niveau des phrases. Une des applications les plus directes de ces corpus consiste à en extraire automatiquement des correspondances lexicales, fournissant une information utile aux traducteurs, aux lexicographes comme aux terminologues. Comme pour l'alignement, des méthodes statistiques ont donné de bons résultats dans ce domaine. Nous pensons qu'une exploitation judicieuse d'indices statistiques adaptés et d'algorithmes de conception simple permet d'obtenir des correspondances fiables. Après avoir présenté les indices classiques, auxquels nous essayons d'apporter des améliorations, nous proposons dans cette article une étude empirique destinée à en montrer les potentialités.

1. Introduction

Depuis le début des années 90, un nouveau champ d'investigation est apparu dans le domaine du Traitement Automatique des Langues : l'exploitation des corpus bilingues parallèles, constitués d'un ensemble de textes et de leurs traductions respectives. Le problème de l'*alignement* de tels corpus, visant à la *segmentation* des textes en unités plus petites (la phrase étant l'étalon le plus courant) et à l'*appariement* des unités en relation de traduction mutuelle, a très tôt rencontré des solutions efficaces et opérationnelles, permettant d'obtenir de grandes quantités de textes parallèles alignés, ou *bi-textes*, de façon totalement automatisée. Une campagne d'évaluation sous l'égide de l'Aupelf-Uref, le projet Arcade (Langlais *et al.*, 1998), a montré que les techniques d'appariement basées sur des indices superficiels comme la longueur des phrases ou la présence de mots apparentés (les cognats) aboutissait à des résultats satisfaisant sur des textes de nature variée, allant de la traduction juridique à celle, plus " libre ", de texte littéraire.

Comme le note Isabelle (1992), les bi-textes ainsi obtenus constituent une mine d'information pour le traducteur, le linguiste ou le terminologue : de par leur taille importante, ils représentent un véritable gisement de solutions réutilisables à chaque fois qu'un problème de traduction déjà rencontré surgit à nouveau. Ainsi que le souligne Macklovitch (1992), ce type de ressource a l'avantage de mettre l'accent sur la pratique réelle de la traduction, à la différence des dictionnaires bilingues qui ne retiennent le plus souvent que les usages les plus standardisés.

Nous nous intéressons ici à une des applications les plus prometteuses des bi-textes : l'extraction automatique de correspondances lexicales (certains auteurs utilisent le terme d'" alignement lexical "[1]). Il s'agit de donner, pour des unités lexicales identifiées dans le texte source, l'équivalent traductionnel employé dans la cible[2]. Ainsi

[1] Nous préférons le terme de *correspondance* dans la mesure où nous pensons qu'il y a une solution de continuité entre l'alignement au niveau des paragraphes ou des phrases, et l'alignement au niveau des mots, qui présuppose en quelque sorte le concept de traduction mot-à-mot. Pour une discussion de ce problème, cf. Kraif (à paraître).

[2] Pour la clarté de l'énoncé, nous utilisons les termes *source* et *cible* afin de différencier les deux parties d'un bi-texte, mais sans nous soucier du sens de la traduction effectuée initialement par l'humain, puisqu'ici la relation d'équivalence traductionnelle est considérée comme étant symétrique.

automatisée, ce type d'extraction permet par exemple de constituer automatiquement des glossaires bilingues utiles aux lexicographes.

Très tôt, depuis la constitution des premiers bi-textes, des méthodes statistiques ont été développées pour obtenir automatiquement ce type de correspondance. Différents indices d'association ont été proposés, permettant de mesurer quantitativement l'importance de la corrélation statistique entre les occurrences de deux unités de part et d'autre d'un bi-texte. A ce jour, ce sont des algorithmes sophistiqués, inspirés de l'algorithme EM (Dempster *et al*., 1977), qui ont obtenu les meilleurs résultats. Nous proposons dans cette étude d'évaluer quelques uns des indices statistiques les plus utilisés (Church *et al*, 1992, Dunning 1993, Fung *et al*, 1994, Gaussier *et al*, 1995), sur la base d'algorithmes simples. Nous désirons ainsi montrer qu'une utilisation judicieuse de ces indices peut donner de bons résultats, et peut permettre, pour certaines applications, de faire l'économie des méthodes itératives[3] coûteuses en temps de calcul et en espace mémoire.

2. Constitution du corpus et mise en œuvre de l'évaluation

La première étape d'une telle évaluation consiste à se doter d'un corpus de référence contenant des correspondances établies manuellement, afin de disposer d'un étalon auquel on puisse comparer les résultats.

Or, il existe plusieurs façons de concevoir l'extraction de correspondances lexicales : en particulier, l'appariement peut être complet, si l'on cherche à apparier toutes les unités du texte source avec les unités correspondantes dans le texte cible ; ou fragmentaire, si l'on s'intéresse à une liste d'unités sources préalablement déterminées (cf. le " lexical spotting " du Projet Arcade). Dans un cas comme dans l'autre, l'appariement dépend bien sûr de ce que l'on entend par unité lexicale. En fait, quel que soit le type d'extraction considéré, il est très difficile de donner une définition rigoureuse des correspondances lexicales, qui soit linguistiquement motivée et qui permette en même temps d'énoncer des critères explicites pour la segmentation et l'appariement. Nous avons montré (Kraif, à paraître) que la notion d'alignement lexical, basé sur le concept de compositionnalité traductionnelle (Isabelle, 1992), soulevait des problèmes d'indétermination : la traduction étant d'abord une opération de transformation appliquée à des messages, et non à des unités linguistiques, la compositionnalité traductionnelle n'a pas toujours de pertinence en deçà de la phrase. La distance sémantique des unités sources et cibles est variable, et dépend étroitement du contexte. En outre cette distance est indissociable du niveau de segmentation choisi (phrase, syntagme, mot). La définition de l'alignement conduit donc à une intrication profonde entre appariement et segmentation.

Afin de lever ces indéterminations, nous préférons abandonner le concept de compositionnalité. Nous prônons une conception restreinte des correspondances lexicales, en distinguant, d'une part, l'identification des unités, et d'autre part, leur appariement. Notre définition des correspondances est donc basée sur deux conditions indépendantes :

- la détermination des unités doit être effectuée préalablement, sur des critères monolingues. Dans cette tâche, nous avons retenu toutes les unités lexicales au sens large (la lexie au sens de Mel'cuk *et al*., 1995), incluant des formes simples, des mots composés, des unités phraséologiques et des termes polylexicaux (dont l'unité est due à des critères extra-linguistiques) :

(1) mots composés : *letter of formal notice* <-> mise en demeure
(2) unités phraséologiques : *building a nuclear arsenal* <-> se doter de l'arme nucléaire
(3) termes polylexicaux : *Board of Governors* <-> Conseil des gouverneurs

L'identification de ces unités pose parfois problème : le plus simple est de caractériser les unités figées (exemple 1) qui satisfont aux critères énoncés par G. Gross (1996). Mais le figement est un phénomène scalaire par nature, et certaines unités se situent à la limite, comme par exemple " huile d'olive ". Sans compter que certaines unités échappent totalement au figement et présentent toutefois un intérêt pour le traducteur : c'est le cas des tournures idiomatiques (exemple 2), ou des termes (exemple 3). Nous avons adopté une acception large de la polylexicalité, en reconnaissant les unités dans les cas de figement, de tournure idiomatique ou d'unité terminologique définie par un concept sous-jacent.

[3] Il s'agit des algorithmes inspirés de l'algorithme EM, dont le but est d'estimer l'ensemble des paramètres maximisant la probabilité du corpus d'apprentissage. Dans la version la plus simple, les paramètres sont constitués des probabilités liées à chaque correspondance lexicale. La convergence vers les paramètres optimaux est atteinte par la répétition de deux étapes : 1. A partir des paramètres, on compte le nombre de fois qu'une lexie est susceptible d'être en correspondance avec une autre lexie, dans le corpus. 2. On réestime les paramètres (probabilités des correspondances) en fonction de ces comptes.

- dans un second temps, l'appariement doit être basé sur l'équivalence traductionnelle. Ce type d'équivalence pouvant concerner différents niveaux (équivalence dynamique, sémantique, conceptuelle, référentielle, stylistique, etc.), nous n'avons retenu qu'un seul critère : celui de généralité, l'équivalence devant être envisageable dans d'autres contextes.

Bien sûr, chacune de ces deux conditions autorise une grande variété d'interprétations. Afin de donner un maximum de cohérence à l'extraction manuelle, nous avons cherché à expliciter des critères précis permettant de trancher dans les cas litigieux. Il aurait été souhaitable, à l'instar des précédentes évaluations (Melamed, 1998, Langlais, 1998), d'assurer un consensus intersubjectif en faisant appel à plusieurs annotateurs. Malheureusement, cette tâche n'a pu être confiée qu'à une seule personne. Les critères choisis, ainsi que les correspondances manuelles sont disponibles pour consultation à l'adresse : http:\\lilla2.unice.fr.

Pour cette évaluation, nous avons réutilisé le corpus JOC fourni pour la deuxième campagne du projet Arcade. Nous avons aligné automatiquement ce corpus au niveau des phrases[4], en utilisant les techniques d'alignement décrites dans (Kraif, 1999). Nous sommes parti d'un bi-texte aligné automatiquement afin de montrer que les quelques erreurs qui subsistent (évaluées comme étant inférieures à 3%) ne compromettent pas l'utilisation des techniques ici présentées. Cet alignement nous a fourni environ 69 000 couples de phrases.

La totalité de ce corpus a été utilisée comme corpus d'apprentissage (dans le compte des cooccurrences), mais une partie seulement a été appariée manuellement, pour l'évaluation. Pour constituer cet échantillon, nous avons effectué un tirage aléatoire de 1 000 couples de phrases, parmi lesquels nous avons supprimé les phrases ne contenant qu'une seule forme, pour aboutir finalement à 767 couples de phrases. Au final, l'extraction manuelle a donné un peu plus 9 000 couples de lexies correspondantes. Notons qu'au cours de l'appariement environ 30% des mots n'ont pas trouvé de correspondant lexical satisfaisant : nous nommerons " résidu " ces unités.

	Corpus d'apprentissage		Echantillon	
	Anglais	Français	Anglais	Français
Nombre de mots	1 060 174	1 168 555	16 216	20 002
Nombre de couples de phrases	69 160		767	
Nombre de mots retenus dans les couples			11 736	13 041
Répartitions des 9079 couples appariés manuellement				
Nombre de lexies simples			7 383	6 983
Nombre de lexies polylexicales			1 996	2096

tableau 1 : constitution du corpus d'apprentissage et de l'échantillon de référence

Sur la base des couples de référence, nous nous proposons d'évaluer trois tâches différentes (un exemple de ces trois tâches est fourni à l'annexe 1) :

- *l'extraction des correspondances entre toutes les lexies* : il s'agit, étant données les lexies identifiées dans le texte source et le texte cible, d'établir automatiquement les appariements entre unités équivalentes. Ces lexies incluent mots simples et unités polylexicales, qu'il s'agisse de mots pleins ou de mots fonctionnels.

Les couples ainsi extraits sont comparés aux couples de référence par le biais des mesures classiques de précision, de rappel et de F-mesure. Si l'on note C l'ensemble des couples à évaluer, C_{ref} l'ensemble des couples de référence, on a donc :

$$P = \frac{\left|C \cap C_{ref}\right|}{\left|C\right|} \quad R = \frac{\left|C \cap C_{ref}\right|}{\left|C_{ref}\right|} \; et \; F = \frac{2 \times (P \times R)}{(P + R)} \qquad (1)$$

- *l'extraction des correspondances entre lemmes* : en supprimant les variations morphologiques superficielles des lexies (nombre pour les substantifs, genre et nombre pour les adjectifs, temps, mode et personne pour les verbes, etc.) on recalcule les cooccurrences et l'on effectue une extraction similaire à la précédente.

- *l'extraction des correspondances entre formes simples* : les unités considérées sont des mots simples (groupes de lettres séparés par des espaces) n'ayant subi aucun pré-traitement. Dans ce cas de figure, on obtient parfois des correspondances fragmentaires, lorsque des parties de lexies complètes sont appariés : dans le calcul

[4] Par soucis de simplicité, nous nous placerons toujours dans le cas d'alignement mettant en jeu des couples de phrases : en fait, il peut s'agir de couples résultant de l'appariement d'agrégats de 0, 1, 2 ou 3 phrases. Par abus de langage, nous dénommerons phrase ce type de d'agrégat.

de la précision, on considérera comme étant valide tout couple de formes inclus dans un couple de lexies appariées manuellement. Pour le calcul du rappel, le dénominateur correspond au nombre de formes simples engagées du côté source dans les couples de référence.

A la différence de l'évaluation menée au sein du Projet Arcade, nous ne nous concentrons pas sur un ensemble réduit de mots-test, choisis en fonction d'aspects sémantiques (polysémie), morphologiques (substantifs, verbes et adjectifs) et de leurs fréquences. Dans la mesure où les méthodes les plus efficaces s'appuient sur des appariements exhaustifs, il n'est pas prématuré de s'intéresser aux extractions globales qui engagent toutes les unités.

3. Indices

La mise en œuvre de la plupart des indices d'association se base sur le raisonnement suivant : si deux unités ont des distributions similaires, respectivement dans le texte source et le texte cible, i.e. si elles apparaissent fréquemment dans des zones en relation de traduction, il est probable qu'elles soient elles-mêmes en relation de traduction. Le fait d'apparaître ensemble dans des zones alignées est généralement désigné par le terme de *cooccurrence*. On devrait préciser " cooccurrence parallèle ", afin d'éviter toute ambiguïté avec la cooccurrence dans le cas monolingue. Nous préciserons à chaque fois qu'il pourrait y avoir confusion. Comme l'a montré Melamed, le dénombrement des cooccurrences dans un bi-texte peut être effectué suivant différents modèles. Dans la suite de cet exposé, nous nous limiterons au plus simple et au plus courant de ces modèles : deux unités sont cooccurrentes à chaque fois qu'elles apparaissent de part et d'autre d'un couple de phrases alignées[5].

– L'indice le plus classique est l'information mutuelle (Church et Hovy, 1992) : c'est le rapport entre le nombre de cooccurrences observées, et le nombre théorique basé sur l'hypothèse de l'indépendance des unités u_1 et u_2. Si n représente le nombre de couples de phrases alignées, n_1 et n_2 représentent le nombre des occurrences respectives de u_1 et u_2, et n_{12} le nombre total de cooccurrences de u_1 et u_2, alors l'information mutuelle se calcule de la manière suivante :

$$IM = \log\left(\frac{p_{12}}{p_1 p_2}\right) \qquad (2)$$

avec $p_1 = \frac{n_1}{n} \quad p_2 = \frac{n_2}{n} \quad p_{12} = \frac{n_{12}}{n}$

Un défaut majeur de l'information mutuelle est sa tendance à surévaluer l'association entre les unités peu fréquentes. Généralement, on estime qu'une IM élevée n'a pas de signification pour des unités cooccurrant moins de 3 fois.

– Afin de pallier ce défaut, Fung (1994), dans un article où elle développe une méthode d'alignement basée sur l'extraction préalable de correspondances lexicales, propose d'employer le *t-score*. Son calcul est voisin de celui de l'IM :

$$t \approx \frac{p_{12} - p_1 p_2}{\sqrt{\frac{p_{12}}{n}}} \qquad (3)$$

– Par ailleurs, Dunning (1993) note avec justesse que les unités peu fréquentes n'ont rien de " rare ". Ainsi les méthodes basées sur le test du Qi2 ou du z-score (supposant la normalité des distributions) seraient invalides pour la plupart des unités lexicales. En modélisant l'occurrence d'une unité par une distribution binomiale, Dunning déduit un indice évaluant la plausibilité de l'hypothèse d'indépendance des occurrences de deux unités quelconques. L'opposé du logarithme permet alors d'exprimer le degré d'association entre deux ces unités. Pour deux unités u_1 et u_2, on donne la table de contingence suivante :

	occurrence de u_2	*non occurrence de u_2*
occurrence de u_1	*a*	*b*
non occurrence de u_1	*c*	*d*

tableau 2 : occurrences et cooccurrences de u_1 et u_2

[5] En fait, le calcul est un peu plus compliqué : lorsque dans un même couple de phrases, l'unité lexicale *f* apparaît *p* fois, et l'unité *f'* apparaît *q* fois, le nombre de cooccurrences engendrées par ce couple est donné par : *cooc=min(p,q)*.

On a alors, en notant RV l'indice issu du rapport de vraisemblance :

$$RV = -2 \log \lambda = 2(S^+ - S^-)$$
$$S^+ = a \log a + b \log b + c \log c + d \log d + n \log n \quad (4)$$
$$S^- = (a+c)\log(a+c) + (b+d)\log(b+d) + (a+b)\log(a+b) + (c+d)\log(c+d)$$

Dans l'observation des cooccurrences, cet indice s'est révélé être un des plus simples et des plus fiables (Gaussier et Langé, 1995).

– Nous avons développé un autre indice basé sur le modèle binomial : nous avons tout simplement cherché à évaluer la probabilité de l'hypothèse nulle, supposant l'indépendance des occurrences de u_1 et u_2. En reprenant les notations précédentes, on a :

$$P_0(n_{12}/n, n_1, n_2) = \frac{C_n^{n_1} C_{n_1}^{n_{12}} C_{n-n_1}^{n_2 - n_{12}}}{C_n^{n_1} C_n^{n_2}} = \prod_{k=1}^{n_2 - n_{12}} \frac{(n - n_1 - n_2 + n_{12} + k)}{(n - n_2 + n_{12} + k)} \prod_{k=1}^{n_{12}} \frac{(n_1 - n_{12} + k)(n_2 - n_{12} + k)}{k(n - n_2 + k)} \quad (5)$$

En prenant l'opposé du logarithme , on obtient à nouveau un indice permettant d'évaluer le degré d'association de u_1 et u_2. On note *P0* l'indice résultant.

Enfin, nous avons mis au point un dernier type d'indice, basé non plus sur l'observation des distributions, mais sur les similitudes formelles entre les unités lexicales : il s'agit cette fois d'identifier les transfuges (lexies invariantes) et les cognats (lexies apparentées), dont la ressemblance superficielle peut être une source d'information précieuse. L'élaboration de cet indice repose sur l'identification des sous-chaînes maximales, telles que nous les avons utilisées dans les méthodes d'alignement (cf. Kraif, 1999). Pour chaque couple candidat, nous distinguons onze cas, en fonction desquels nous avons calculé empiriquement la probabilité de non-correspondance (cf. annexe).

L'opposé du logarithme de cette probabilité aboutit à l'indice noté CO. L'addition de cet indice avec l'indice P0 nous donne un dernier indice combiné, PC, mêlant information distributionnelle et ressemblances superficielles. En résumé nous étudierons 6 indices : l'information mutuelle IM, le t-score TS, le rapport de vraisemblance RV, l'invraisemblance de l'hypothèse nulle P0, l'indice basé sur les ressemblances de surface CO, et l'indice combiné PC.

4. Algorithmes

Pour les lexies comme pour les formes simples, nous avons intégré ces indices dans deux cadres algorithmiques très simples :

- *recherche de l'indice maximum* : pour chaque unité de la phrase source, on retient une correspondance obtenant la meilleure valeur de l'indice.

- *recherche de la meilleur affectation biunivoque* : on suppose que les correspondances se réalisent sous la forme d'une relation biunivoque entre les unités cibles et les unités sources. Dans ce cadre, une même unité ne peut entrer dans plusieurs correspondances à l'intérieur d'un même couple de phrases. Cet algorithme est itératif à partir de l'étape 2 :

1. *Constitution de l'ensemble des candidats Cand* : on calcule des indices pour toutes les unités (u_1,u_2) du couple de phrases (P_1,P_2). Tous ces couples sont placés dans *Cand*.
2. *Sélection* : on sélectionne un couple (U_1,U_2) de *Cand* obtenant la meilleure valeur de l'indice. On retient ce couple dans l'ensemble *Corresp* contenant le résultats des correspondances, et on élimine de *Cand* tous les couples qui mettent en jeu (U_1) ou (U_2).
3. *Retour* en 2 tant que l'ensemble *Cand* n'est pas vide.
4. *Terminaison*. *Corresp* contient le résultat.

Dans la mise en œuvre de ce dernier algorithme, nous avons procédé à quelques simplification, pour des raisons d'efficacité des calculs. D'une part, au sein de chaque phrase, une même unité ne peut rentrer que dans un seul couple, même si elle y compte plusieurs occurrences. Cette approximation affecte légèrement le rappel, puisque 9 % des couples de références concernent des unités répétées dans un même couple de phrases. D'autre part, les unités dépassant 5 000 occurrences dans la totalité du corpus d'apprentissage ne sont pas prises en

compte. Cela concerne 29 unités en anglais et 38 en français, essentiellement des mots outils et des signes de ponctuation. De par ces simplifications, 31 % des couples corrects ne pouvant être considérés, le rappel ne pourra dépasser la limite de 69 %.

5. Evaluation

L'application de ces deux algorithmes sur les 6 indices donne les valeurs de précision et de rappel représentées sur les figures 1-4. Les indices sont symbolisés de bas en haut : CO losange vide, IM carré plein, TS triangle plein, RV carré vide, P0 barre horizontale, PC astérisque. Par *FS*, *Lex* et *Lem* on désigne les trois tâches précédemment énumérées. Les résultats permettent d'établir des comparaisons selon les 3 axes suivants :

- *Comparaison des algorithmes :*

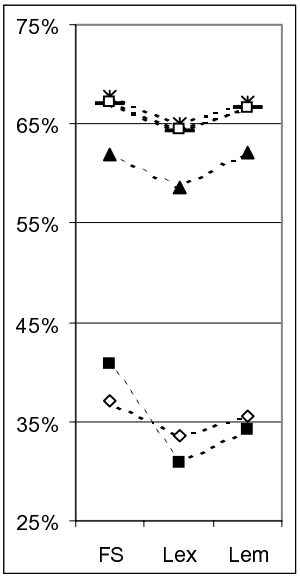
On constate une amélioration globale des résultats entre 1 et 2. C'est l'information mutuelle qui en bénéficie le plus, car dans l'algorithme 1, cet indice concentre toutes les correspondances sur les mêmes unités ; ce défaut est compensé par l'affectation biunivoque. Il est notable que CO se maintient au niveau plancher : l'information apportée par les cognats étant plutôt rare à l'intérieur d'un même couple de phrases, il est peu fréquent que deux unités cibles soient en concurrence pour être appariées avec une même unité source.

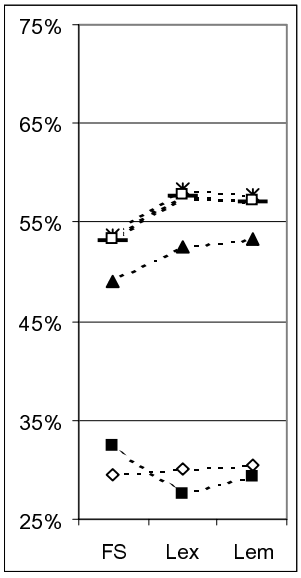
- *Comparaison des tâches :*

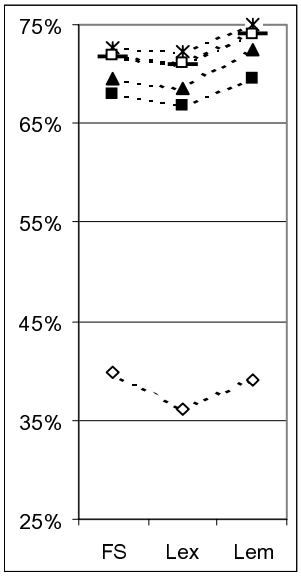
De manière étonnante, l'extraction des formes simples paraît obtenir une meilleure précision : cela peut être dû à la tolérance du mode d'évaluation choisi, dans la mesure où il suffit qu'un appariement soit inclus dans un couple de référence pour qu'il soit compté comme correct. D'ailleurs, malgré ce biais, cette extraction demeure inférieure au deux autres vis-à-vis de la F-mesure. Les meilleurs résultats globaux sont obtenus avec l'extraction *Lem*, même si celle-ci connaît un rappel inférieur à *Lex*. Cette dégradation est due à un autre artefact : l'algorithme négligeant les occurrences des unités répétées dans un même couple de phrases, la réduction morphologique amplifie l'effet de cette simplification.

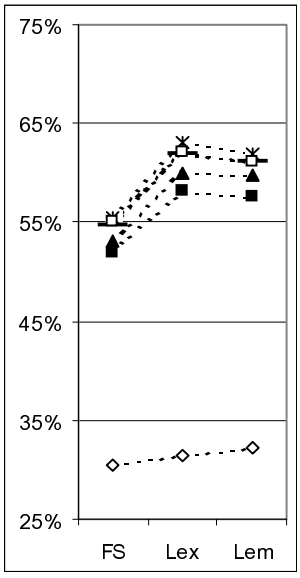
- *Comparaison des indices :*

Enfin, quel que soit l'algorithme ou la tâche, une hiérarchie se dessine au niveau des indices : les meilleurs résultats sont toujours produits par la mesure combinée PC, et les moins bons par CO. Les informations









apportées par les cognats sont donc complémentaires des cooccurrences, puisqu'elles se cumulent harmonieusement dans PC. En outre P0 semble se comporter de manière identique à RV, puisque leur deux courbes se chevauchent.

figure 1 : P algo. 1 *figure 2 : R algo. 1* *figure 3 : P algo. 2* *figure 4 : R algo. 2*

Afin de vérifier certaines des analyses précédentes, nous avons effectué une autre extraction en supprimant tous les mots outils du corpus (comme des correspondances de référence). En prenant la moyenne des six indices, avec l'algorithme 2, on constate une amélioration globale des résultats, notamment du rappel. Ceci pour deux raisons : d'une part les mots outils sont plus fréquents dans le résidu de traduction que dans les couples

appariés ; d'autre part l'effet des simplifications de l'algorithme (absence des unités répétées et/ou de fréquence supérieure à 5000) est estompé du fait de la suppression des mots outils. Ainsi, l'extraction *Lem* n'obtient plus un rappel inférieur à celui de *Lex*.

Algorithme 2	P %			R %			F %		
	FS	Lex	Lem	FS	Lex	Lem	FS	Lex	Lem
avec toutes les unités	65,6	64,3	67,4	50,1	56,1	55,6	56,8	59,9	60,9
sans mots fonctionnels	69,8	68,5	71,1	63,9	76,9	78,1	66,7	72,4	74,5

tableau 3 : comparaison des extractions avec et sans mots fonctionnels

Ainsi les écarts entre les trois tâches apparaissent plus nettement, ce qui confirme nos hypothèses. Les meilleurs résultats sont atteints avec l'indice PC : on obtient P = 78,8 %, R = 86,5 % et F = 82,5 %.

6. Filtrage des résultats

Les correspondances ainsi extraites peuvent donner lieu à différents filtrages permettant d'éliminer les appariements les plus improbables. Nous distinguons trois types de filtrage :

- *filtrage par valeur seuil* : on ne retient que les couples obtenant une valeur de l'indice supérieure à un certain seuil. Pour chaque indice, 7 seuils ont été testés : la moyenne de l'indice multipliée par 7 facteurs allant de 0,25 à 10.

- *filtrage relatif* : pour chaque phrase, on classe les couples par valeur décroissante de l'indice, et l'on ne retient que les meilleurs premiers couples dans les proportions suivantes : 80%, 60%, 40%, 20%.

- *filtrage différentiel* : pour chaque unité source, on ne retient son appariement avec une unité cible que si la valeur de l'indice obtenue avec cette unité est supérieure à toutes les valeurs obtenues avec les unités concurrentes, dans des proportions supérieures à différentes valeurs, allant de 1,05 à 4.

On trouvera en annexe les figures représentant les résultats pour l'appariement des lexies avec l'algorithme 2 (les filtrages présentent, *mutatis mutandis*, les mêmes caractéristiques pour les lemmes et les formes simples, et avec les deux algorithmes).

On constate que :

- les filtrages sont opérants puisqu'ils permettent une augmentation de la précision.

- les valeurs optimales de F sont atteintes avec le filtrage à valeur seuil, de paramètre 0,5 (PC atteint F=69 %). Pour les zones de rappel important (>50%), cette méthode obtient une meilleure précision à rappel égal que les autres filtrages (pour les trois meilleurs indices).

- pour les zones de faible rappel (<50%), le filtrage différentiel apporte une précision plus grande à rappel égal.

Ces filtrages présentent donc des profils légèrement différents : si l'on privilégie le rappel, le filtrage avec seuil convient mieux. Si l'on recherche des précisions très élevées, avec un rappel médiocre, le filtrage différentiel paraît plus adapté. Le filtrage relatif n'a d'intérêt que dans la mesure où il autorise une évolution continue et contrôlée de l'équilibre entre P et R.

8. Conclusion et Perspectives

Pour pouvoir vérifier la validité et la généralité des remarques précédentes, il faudrait certes pousser ces études empiriques vers des investigations plus approfondies, en variant le type de corpus. D'un point de vue global, elles tendent à confirmer la possibilité, moyennant des algorithmes très simples et des mesures statistiques adéquates, d'extraire à partir des bi-textes des informations de premier choix pour les traducteurs comme pour les lexicographes.

Plus spécifiquement, deux indices statistiques semblent plus adaptés à cette tâche : le rapport de vraisemblance et l'inverse de la probabilité de l'hypothèse nulle. La première est plus simple de calcul, mais la seconde présente l'avantage d'être facilement combinable avec d'autres informations sous une forme probabiliste.

En outre, nous avons montré que des sources d'information complémentaires, comme les ressemblances de surface, pouvait se révéler utiles. Dans des recherches ultérieures, il serait intéressant d'étudier ces techniques en

les combinant à des informations linguistiques, issues notamment de dictionnaires bilingues sous format électronique.

Références

Church, K., Hovy, E. (1992). Good Application for Crummy Machine Translation, *Machine Translation*, 8.

Debili F, Sammouda E. (1992). Appariements de Phrases de Textes bilingues Français-Anglais et Français-Arabes. In *Actes de COLING-92*, Nantes, pp. 528-524

Dempster, A., Laird, N., Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 34 (B).

Dunning, T. (1993). Accurate Methods for the Statistics of surprise and Coincidence. *Computational Linguistics*. Vol 19, 1, pp. 61-74

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*, Kyoto

Gale W., Church K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, CA, pp. 177-184

Gaussier, E., Langé J.-M. (1995). Modèles statistitiques pour l'extraction de lexiques bilingues, *T.A.L.*, Vol. 36, N° 1-2, pp. 133-155

Gross G. (1996). *Les expressions figées en français*, Ophrys, Paris

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

Kraif O. (1999). Identification des cognats et alignement bi-textuel :une étude empirique, *Actes de TALN'99*, Cargèse, France, pp. 205-214

Kraif O. (à paraître). Translation alignment and lexical correspondences : a methodological reflection. In B. Altenberg & S. Granger, Ed., *Lexis in contrast. Studies in Corpus Linguistics*. John Benjamins

Langlais P., Simard M., Veronis J. *et al*, (1998), ARCADE : A cooperative Research Project on Parallel Text Alignment Evaluation, disponible sur le WEB à http://www.lpl.univ-aix.fr/projects/arcade

Macklovitch, E. (1992). Corpus-Based Tools for Translators, *Proceedings of the 33rd Annual Conference of the American Translators Association*, San Diego California.

Mel'cuk, I., Clas, A., Polguere, A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Duculot, Louvain-la-neuve.

Melamed, D. (1998). Manual Annotation of Translational Equivalence : The Blinker Project. Institute for Research in Cognitive Science. Technical Report#98-06, University of Pennsylvania

Annexe

1. Exemple d'extraction :

fr. : Pour la bonne tenue de ces registres, l'évaluation des cas de mortalité constatés par les autorités apporte des informations importantes.

ang. : *The assessment of the official cause of death is a piece of information vital to these registers.*

Lexies : (Pour ;*to*) (ces ; *these*) (registres ; *registers*) (l' ; *the*) (évaluation ; *assessment*) (des ; *of the*) (cas de mortalité ; *cause of death*) (des ; *a piece of*) (informations ; information) (importantes ; *vital*)
Lemmes : (Pour ;*to*) (ce ; *this*) (registre ; *register*) (le ;*the*) (évaluation ;*assessment*) (de le ; *of the*) (cas de mortalité ;*cause of death*) (de le ; *a piece of*) (information ; *information*) (importante ;*vital*)

Formes simples : (Pour ;*to*) (ces ; *these*) (registres ;*register*) (l ; *the*) (évaluation ;*assessment*) (des ; *of*) (cas ;cause) (de ; of) (mortalité ; *death*) (des ; *piece*) (informations ;information) (importante ;vital)

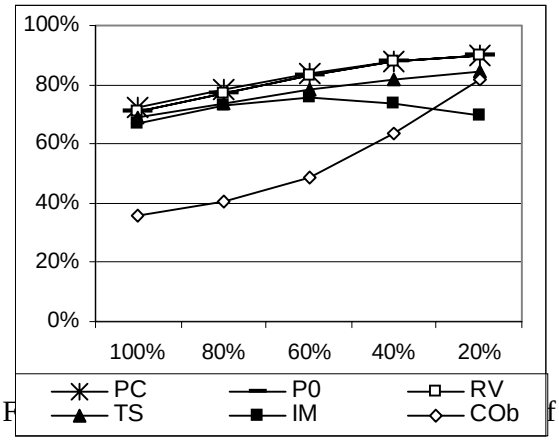
2. Paramètres pour l'indice CO :

Cas 1 = transfuge numérique, cas 2 = transfuge de long. sup. à 3, cas 3 : 4-gram de long. inf. à 7, cas 4-10 : SCM comportant 4-10 caractères, cas 0 : tout le reste.

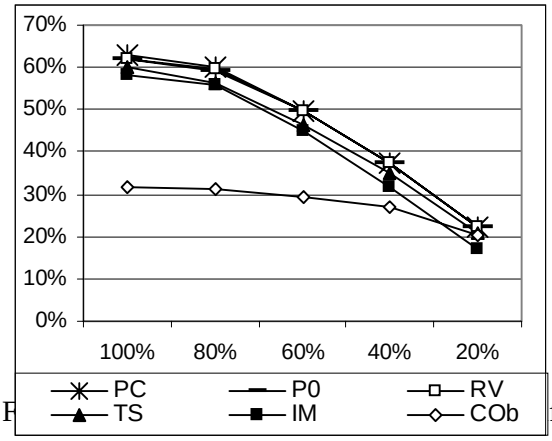
Cas	0	1	2	3	4	5	6	7	8	9	10
p(cas)	0,979	0,001	0,028	0,280	0,497	0,290	0,217	0,125	0,149	0,101	0,058

Ces probabilités ont été calculées sur l'échantillon. Nous pensons qu'elles sont réutilisables pour d'autres corpus et qu'elles dépendent surtout du couple de langues.

3. Résultats pour les différents types de filtrage :



F f



f

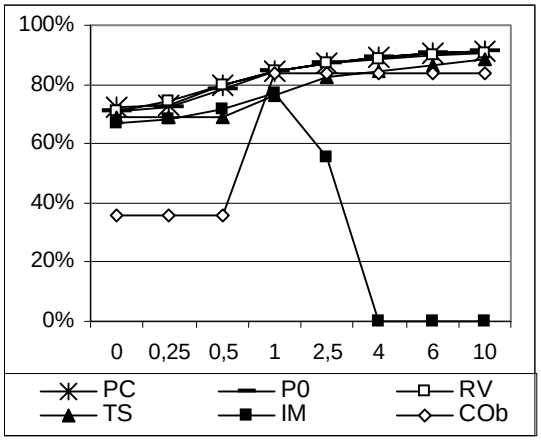
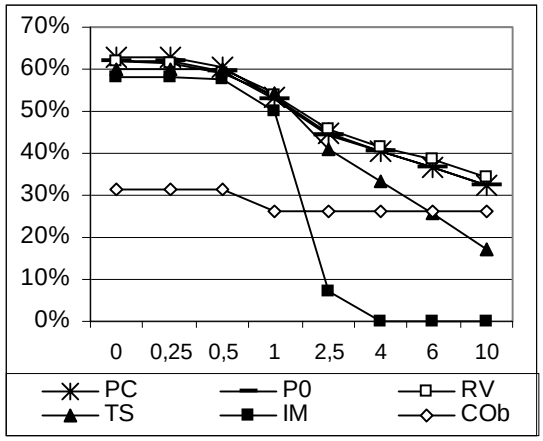
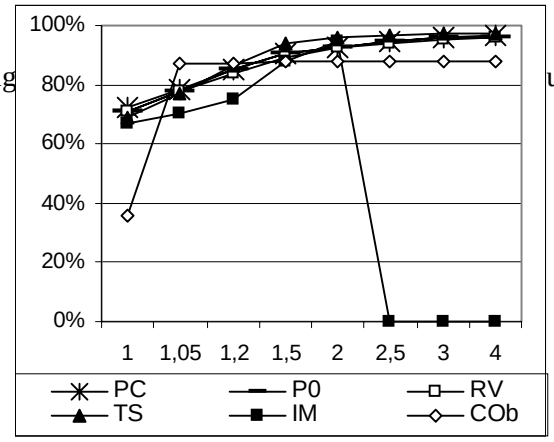


Fig uil



Figu uil



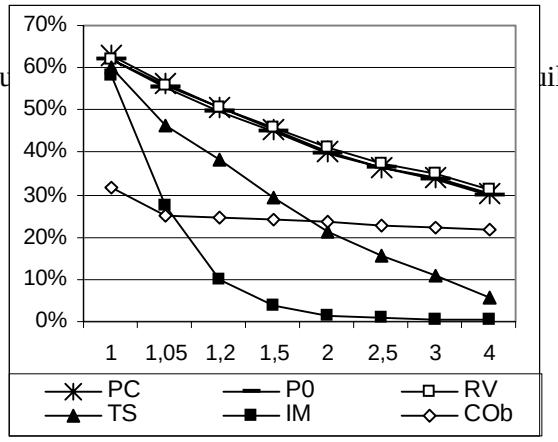


Figure 9 : évolution de P avec le filtrage diff.

Figure 10 : évolution de R avec le filtrage diff.