# Méthodes de filtrage pour l'extraction d'un lexique bilingue à partir d'un corpus aligné

**Olivier Kraif**

*Olivier.Kraif@u-grenoble3.fr*

ABSTRACT.

*This paper addresses the problem of lexical aligning and correspondences extraction from multilingual aligned corpora. We first draw a line between lexical aligning, which concerns segments that are translational equivalents in a particular context, and lexical correspondences extraction, which aims at spotting lexical units that are equivalent at a general level, through their respective language. Then we explain the principle of common methods, which are based on distributional criteria. The results of our own experiments are presented, and we show that it is possible to automatically extract very reliable correspondences by applying an appropriate filter, without an important loss of recall.*

KEYWORDS: *lexical aligning, lexical correspondence, bilingual lexicon, bi-text.*

RESUME.

*Le présent article est centré sur les méthodes dédiées à l'extraction de correspondances lexicales à partir de corpus multilingues alignés. Après avoir établi une distinction entre l'alignement lexical, concernant des segments variables en relation d'équivalence traductionnelle, et l'extraction de correspondances lexicales limitée à des couples de lexies équivalentes au niveau des codes linguistiques, nous dégageons le principe des méthodes basées sur l'observation des occurrences et des cooccurrences des unités. Nous exposons ensuite les résultats de nos expérimentations, en montrant qu'il est possible d'extraire automatiquement des correspondances très fiables au moyen de filtrages adéquats, tout en limitant la dégradation du rappel.*

MOTS-CLES : *alignement lexical, correspondance lexicale, lexique bilingue, bi-texte.*

# Introduction

Comme disait Jean Cocteau dans *Le Potomak*, « le plus grand chef-d'œuvre de la littérature n'est jamais qu'un dictionnaire en désordre. » Avec le même humour, on pourrait ajouter : un texte et sa traduction ne forment alors qu'un dictionnaire bilingue en vrac. Mais au-delà du paradoxe, la question mérite d'être soulevée : du texte au dictionnaire, y aurait-il un chemin qu'on pourrait retrouver ?

Au risque d'être provocateur on pourrait noter avec Saint-Jérôme, grand précurseur parmi les traducteurs, que la traduction n'est pas une affaire de *mots* : « Non verbum de verbo reddere sed sensum », écrivait-il, il ne faut pas rendre un mot pour un mot, mais respecter le *sens*. Mais gardons nous de prendre au mot l'ancêtre des traducteurs. Certes, la traduction est affaire de *sens*, et l'équivalence traductionnelle, située au niveau pragmatique du *message* [PER 93], entretient avec l'équivalence des mots une relation hiérarchique, de même que la fin prime sur les moyens. Ainsi, dans le passage à la traduction, de nombreux mots perdent leur identité formelle, sur les deux plans de l'expression et du contenu, et se dissolvent dans ce que Danica Seleskovitch appelle la « chimie » du sens [SEL 75]. Mais certaines associations lexicales résistent, et des couples se dessinent d'une langue à l'autre : ce que Seleskovitch nomme malicieusement les « raisins dans la brioche », qui restent identifiables après la « cuisson » traductionnelle.

Comme la Bête de Cocteau qui enfile ses gants à l'envers, mais sans violer le second principe de la thermodynamique, nous nous proposons d'étudier ce pari audacieux : reconstituer un dictionnaire bilingue à partir d'un texte et de sa traduction. Cette tâche est devenue possible avec le développement croissant des corpus textuels sous format numérique. L'accès à de vastes corpus traduits en plusieurs langues a permis, en dix ans, le développement des *multi-textes*, ces corpus multilingues alignés, dont des segments équivalents (en général des paragraphes ou des phrases) sont appariés grâce à des méthodes automatiques. Lors de la campagne d'évaluation Arcade [VER 00], les meilleurs systèmes ont obtenus des résultats très satisfaisants sur la tâche de l'alignement des phrases. Pour des traductions respectant des conditions relatives de parallélisme, c'est-à-dire ne comportant qu'une faible proportion d'ajout ou d'omission, précision et rappel dépassaient les 95 %. La deuxième phase d'Arcade, portant sur l'appariement d'unités lexicales, s'est avérée plus problématique. Pour la soixantaine d'unités (adjectifs, noms et verbes) faisant partie du corpus de référence, les meilleurs résultats ne dépassaient pas 75 % de précision et rappel.

De nombreux travaux ont été consacrés à ce type d'appariement [DAG 93] [FUN 94] [GAU 95] [CHA 96] [MEL 97]. La plupart des techniques employées s'appuient sur des critères distributionnels et sur les ressemblances formelles entre les mots. La présente étude se propose de faire le point sur ces méthodes, dans la perspective d'extraire automatiquement, à partir d'un multi-texte, des listes de correspondances lexicales suffisamment fiables et significatives pour être dignes de figurer dans un dictionnaire bilingue. Nous espérons ainsi montrer que, de même qu'il peut être utile d'avoir un dictionnaire bilingue pour traduire, il peut être utile de disposer d'une traduction pour constituer un dictionnaire bilingue.

## La notion de correspondance lexicale

Pour reprendre une distinction établie par Seleskovitch [SEL 80], notons que la *traduction*, en tant qu'activité communicative, ne peut se réduire au *transcodage*, qui concerne le simple transfert d'une unité linguistique vers une unité équivalente dans le système d'arrivée. Tandis que la traduction prend son sens par référence à une constellation de paramètres extra-linguistiques instantanés (participants, intention du message, présupposés, situation, contexte culturel, etc.), le transcodage concerne seulement les deux codes en présence, avec ce qu'ils véhiculent comme valeurs stables, conventionnelles, partagées par l'ensemble des locuteurs. Les unités linguistiques, considérées du point de vue du code, représentent en quelque sorte la synthèse de leurs réalisations possibles, et ne sont pas liées à une actualisation particulière.

Prenons l'exemple suivant, cité par Kay [KAY 00] :

*angl. : Gravity is a pervasive force in the world… (Scientific American)*
*fr. : La pesanteur s'exerce partout sur la terre… (Pour la science)*

Dans ces deux versions, on peut lier sémantiquement l'anglais *pervasive* avec *partout*. Mais la question se pose : ces deux unités sont-elles équivalentes au niveau des codes, y a-t-il entre elles l'« identité formelle » des raisins de Seleskovitch ? Kay pose très clairement le problème en ces termes (nous traduisons) : « Pour un chercheur intéressé par la traduction de grande qualité, un programme d'alignement qui apparierait *pervasive force*, ou seulement *pervasive*, avec *partout*, pourrait ouvrir d'intéressantes perspectives, mais comme source d'entrée potentielle dans un dictionnaire bilingue, cela pourrait constituer une source de frustration. » [KAY 00 : xiv]

Ce type de distorsion, généralement évité dans l'exercice scolaire de la version (qui se cantonne pour des raisons didactiques au transcodage), est beaucoup plus fréquent dans l'activité réelle de la traduction, même pour des textes où l'exigence de littéralité est très forte. Par exemple, dans les questions écrites du Journal de la Communauté européenne, on trouve de nombreuses solutions de ce type :

> *fr. : (…) sur l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite.*
> *angl. : (…) on the marking of banknote for the benefit of the blind and partially sighted.*

Que dire du couple (*émission*, *marking*) d'un point de vue lexicographique ?

Il y a donc un hiatus : d'un côté, les corpus de traductions recèlent un grand nombre d'équivalences, intéressantes à la fois d'un point de vue contrastif et traductionnel, mais diffuses et difficiles à isoler de leur contexte ; de l'autre côté, dans une perspective dictionnairique, on voudrait des unités bien délimitées, transcodables et équivalentes sur le plan de leur *valeur* (au sens saussurien).

Ce hiatus n'est insurmontable que si l'on *oppose* artificiellement le transcodage et la traduction, au lieu de les *articuler*. C'est ce que révèlent les études sur la prise de note en traduction consécutive, exemple type de traduction où la globalité du sens du message l'emporte sur les signifiés locaux. En analysant les notes griffonnées à la va-vite au cours de l'interprétation, Seleskovitch remarque qu'à travers la chimie du sens, des couples d'unités se détachent et transitent directement du *code* vers le *message*, de la *langue* vers la *parole* : « Ce n'est donc pas l'existence d'une équivalence dans une autre langue qui amène l'interprète à prendre le mot en notes, mais la conscience que le mot entendu a une personnalité propre, indépendante du message ; sa signification pertinente se dégage grâce au contexte mais, à part cela, il passe du niveau de la langue à celui de la parole, sans prendre de sens autre que celui que lui confère le code. » (citée par [LAP 94 : 239]).

Ces couples forment en quelque sorte le pivot, la charpente stable de la relation d'équivalence en cours de construction. Dans l'exemple précédent, on peut extraire (*billet de banque*, *banknote*) (*aveugle*, *blind*) (*personne à vision réduite*, *partially sighted*) qui forment la structure thématique de la phrase. Du fait de leur autonomie, notons que ces équivalences intéressent le

lexicographe autant que le terminographe, le premier se situant sur le plan de la langue en général, et le second sur celui des codes pertinents dans un domaine particulier.

Extraire des correspondances lexicales valides au niveau des codes, à partir d'un corpus issu de la pratique concrète de la traduction, ne consiste donc pas à relier chaque mot de la cible avec le (ou les) mot(s) de la source qui entretiennent un rapport traductionnel avec lui, mais à *filtrer* les associations susceptibles de s'extraire de leur contexte. C'est la différence que nous opérons entre *alignement lexical* et *correspondance lexicale* : *émission* et *marking* peuvent être alignés, car ils existe un lien génératif entre l'un et l'autre, lien que le test de commutation interlingue discuté par Mahimon [MAH 99] peut révéler ; en revanche, *émission* et *marking* ne forment pas de correspondance lexicale, telle qu'on pourrait la trouver dans un dictionnaire bilingue.

Notons que cette distinction répond aussi à des contraintes d'ordre épistémologiques et pratiques. En effet, l'alignement lexical reste selon nous une tâche problématique, difficile à circonscrire, car elle entremêle deux continuums, celui de la granularité et celui du degré d'équivalence retenu : faut-il préférer (*émission*, *marking*), ou (*émission de billet de banques*, *marking of banknote*), ou encore (*émission de billet de banques identifiables par les aveugles*, *marking of banknote for the benefit of the blind*) ? Plus la granularité est grossière et meilleure est l'équivalence : le degré d'équivalence traductionnelle et la granularité sont donc en relation antagoniste, et aucun critère simple ne nous permet de fixer le bon dosage. En outre, les associations étudiées sont singulières, étroitement liées à leur contexte, et donc peu fréquentes.

A l'inverse, l'extraction de correspondances lexicales nous ramène à des problèmes mieux connus : d'une part, l'identification des lexies dans leur contexte monolingue, et d'autre part le filtrage de régularités dans l'association des unes avec les autres.

## Méthodes

Nous ne nous intéresserons ici qu'aux méthodes dédiées à ce filtrage, en considérant les lexies sources et cibles comme déjà données, notamment en ce qui concerne les unités polylexicales.

Des travaux divers ([FUN 94] [GAU 95] [CHA 96] [MAC 96] [MEL 97] [KRA 00]) ont montré comment extraire des lexiques bilingues « bruts » à

partir de l'observation des occurrences et des cooccurrences des unités au sein d'un bi-texte. La plupart des méthodes ainsi développées se basent sur une idée simple : des unités sources et cibles qui apparaissent très fréquemment (c'est-à-dire plus souvent que le hasard ne le laisserait escompter) dans des segments équivalents, sont vraisemblablement équivalentes. Considérons l'exemple de la figure 1 :

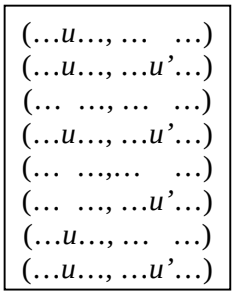



Figure 1. *Occurrences et cooccurrences de deux unités ($n_1$= 5, $n_2$=4, $n_{12}$=3)*

On compte 5 occurrences de l'unité *u*, 4 occurrences de l'unité *u'* et 3 cooccurrences des deux unités. En fonction des seules occurrences, on peut estimer le nombre de cooccurrences qu'on obtiendrait dans le cas d'une distribution aléatoire (8*(5/8)*(4/8) = 2,5). Si le nombre de cooccurrences observées dépasse de manière significative cette estimation, on peut alors faire l'hypothèse que les unités sont équivalentes. En outre, plus on dénombre de cooccurrences dans des contextes variés, plus il est vraisemblable de supposer que l'équivalence ainsi repérée est générale, indépendante d'un co-texte particulier, et donc valide au niveau des codes.

Plusieurs indices statistiques permettent de chiffrer la vraisemblance de cette hypothèse : l'information mutuelle [CHU 90], le t-score [FUN 94], le rapport de vraisemblance [DUN 93], et la log-probabilité de l'hypothèse nulle [KRA 00] (cf. les formules données en annexe II).

Divers algorithmes ont été proposés pour l'extraction des correspondances :

- en calculant une matrice d'association entre unités sources et cibles, à l'intérieur de segments textuels plus ou moins larges. Le nuage de points obtenus est ensuite filtré en fonction des zones de forte densité [FUN 94].

- en calculant l'indice d'association pour tous les couples d'unités à l'intérieur des couples de phrases préalablement alignées. Plusieurs configurations sont possibles : retenir, pour chaque unité source, l'unité cible qui obtient le meilleur score [KRA 00] ; retenir, pour chaque unité source, l'unité cible qui obtient le meilleur score, à condition que la réciproque soit vraie [GAU 95] ; sélectionner le couple d'unités obtenant le meilleur score à l'intérieur des phrases source et cible, et éliminer tous les couples concurrents (i.e. impliquant une des deux unités), puis réitérer l'opération jusqu'à épuisement des couples candidats. Cette dernière méthode, réputée plus efficace, correspond au « *competitive linking algorithm* » [MEL 97] ou à l'« algorithme de meilleure affectation biunivoque » [KRA 00].

Par ailleurs les indices d'associations peuvent s'intégrer dans des modèles probabilistes plus complexes, en faisant alterner itérativement l'estimation des paramètres d'un modèle de traduction, et l'extraction des associations les plus probables en fonction de ces paramètres [BRO 93] [KUP 93] [MEL 97]. Ces méthodes, plus coûteuses en calcul, sont en général basées sur l'algorithme EM [DEM 77].

# Expérimentation

## Extraction des correspondances

Dans des travaux précédents [KRA 00], nous avons implanté différents indices d'association sur des unités lexicales pré-segmentées, en utilisant l'algorithme de meilleure affectation biunivoque, sur le corpus JOC[1], aligné automatiquement au niveau des phrases avec les techniques décrites dans [KRA 01a]. Nous avons évalué les résultats obtenus par rapport à un corpus de référence d'environ 700 couples de phrases, extraits par tirage aléatoire, et dont les unités ont été appariées manuellement. Les valeurs de F-mesure[2],

[1] Le corpus JOC, utilisé dans le projet Arcade dans ses versions anglaise et française, est constitué de questions écrites soumises à la Commission européennes en 1993, dans les Séries C du Journal officiel de la Communauté européenne, et collectées dans le cadre du projet MLCC-MULTEXT. Il compte environ un million de mots dans chaque langue. Cf. http://www.lpl.univ-aix.fr/projects/multext/CORP/JOC.html

[2] La précision P est la proportion de couples corrects parmi les couples extraits automatiquement, le rappel R est la proportion couples extraits automatiquement parmi les couples de référence, et la F-mesure, correspondant au coefficient Dice, est la moyenne harmonique de P et R (cf. annexe I pour les formules).

pour les extractions réalisées avec ces différents indices sont représentées figure 2.

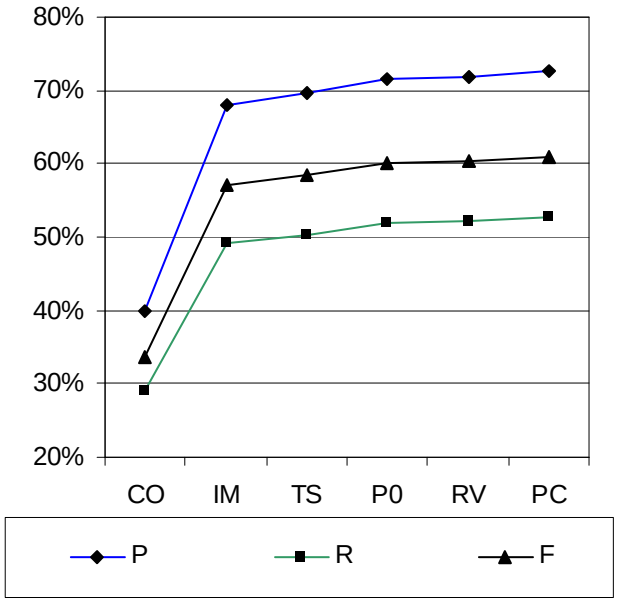



Figure 2. *F-mesure des extractions de correspondances lexicales. CO désigne un indice basé sur la cognation (mots apparentés ressemblants), IM l'information mutuelle, TS le T-score, RV le rapport de vraisemblance, P0 la log-probabilité de l'hypothèse nulle, et PC la combinaison de CO et P0.*

A la différence des autres indices, CO n'est pas construit sur l'observation des occurrences et des cooccurrences, mais sur la comparaison formelle des chaînes de caractère, afin de détecter les mots apparentés ou cognats. L'élaboration de cet indice repose sur l'identification des sous-chaînes maximales, telles que nous les avons utilisées dans les méthodes d'alignement [KRA 01a]. Pour chaque couple candidat, nous distinguons onze cas, en fonction desquels nous avons calculé empiriquement la probabilité de non-correspondance (cf. annexe III). Le meilleur indice, PC, est l'addition de CO et P0 : il mesure l'invraisemblance de l'hypothèse selon laquelle les ressemblances superficielles et distributionnelles seraient dues au jeu du hasard.

Avec PC, indice basé sur les cooccurrences et l'identification des cognats, nous avons obtenu des résultats relativement bons, avec une

précision de 72,2 %. Le rappel, de 58,8 % est plus modeste du fait de certaines approximations effectuées pour des raisons d'efficacité, qui ont conduit à ignorer environ 30 % des couples de référence. Lorsque les extractions sont lancées sur des formes simples, sans regroupement des unités polylexicales, la F-mesure tombe à 61 %. En revanche, lorsqu'on ignore les mots outils du corpus (articles, conjonctions, prépositions, auxiliaires, etc.), et lorsqu'en outre les unités lexicales sont lemmatisées avant l'appariement, on obtient une précision et un rappel voisin de 78,5 %.

Cependant, même dans le cas le plus favorable, une précision de 78 % peut sembler insuffisante. Dans l'idée de constituer des couples candidats à l'entrée dans un dictionnaire, obtenir un couple erroné sur cinq paraît rédhibitoire. Il faut donc se doter d'outils visant à l'élimination de ces « mauvais » couples.

## Filtrage

Pour chaque indice, nous avons testé trois méthodes de filtrage :

- *Filtrage relatif*. Etant donné la nature de nos algorithmes, pour chaque couple de phrases, l'extraction fournit une série de correspondances ordonnées de façon décroissante en fonction des valeurs de l'indice. Le filtrage relatif consiste à conserver les meilleurs couples, suivant différentes proportions. Nous avons testé les proportions suivantes : 80 %, 60 %, 40 %, 20 %.

- *Filtrage absolu*. Cette fois le seuil de rejet n'est plus relatif, lié au classement des candidats à l'intérieur de chaque couple de phrases, mais fixé en valeur absolue. Pour chaque indice, nous avons calculé la moyenne des valeurs obtenues pour des couples quelconques. Nous avons ensuite éliminé tous les couples obtenant une valeur inférieure à $x$ fois la moyenne, avec différentes valeurs de $x$ :

$x = 0{,}25$ $x = 0{,}5$ $x = 1$ $x = 2{,}5$ $x = 4$ $x = 6$ $x = 10$

- *Filtrage différentiel*. Outre la valeur absolue, il existe un autre indicateur de la fiabilité d'un couple de correspondances : il s'agit de l'existence d'associations concurrentes, mettant en jeu une des deux unités du couple. Ainsi, pour chaque couple extrait, on calcule le rapport entre la valeur obtenue et la deuxième meilleure valeur atteinte par tout couple concurrent (i.e. qui met en jeu une des deux unités du couple). Plus ce rapport se rapproche de 1, et plus le couple sera considéré comme suspect. On élimine donc tous les couples pour lesquels le rapport est inférieur à un certain seuil $s$. Les seuils suivants ont été testés :

*s = 1,05  s = 1,2  s = 1,5  s = 2  s = 2,5  s = 3  s = 4*

Nous avons implanté les trois méthodes de filtrage, avec ces différents paramètres, sur tous les résultats obtenus précédemment.

Les graphiques de la figure 3 montrent l'évolution conjuguée de la précision et du rappel avec les différents seuils envisagés, pour les 4 indices : CO, TS, P0 et PC.

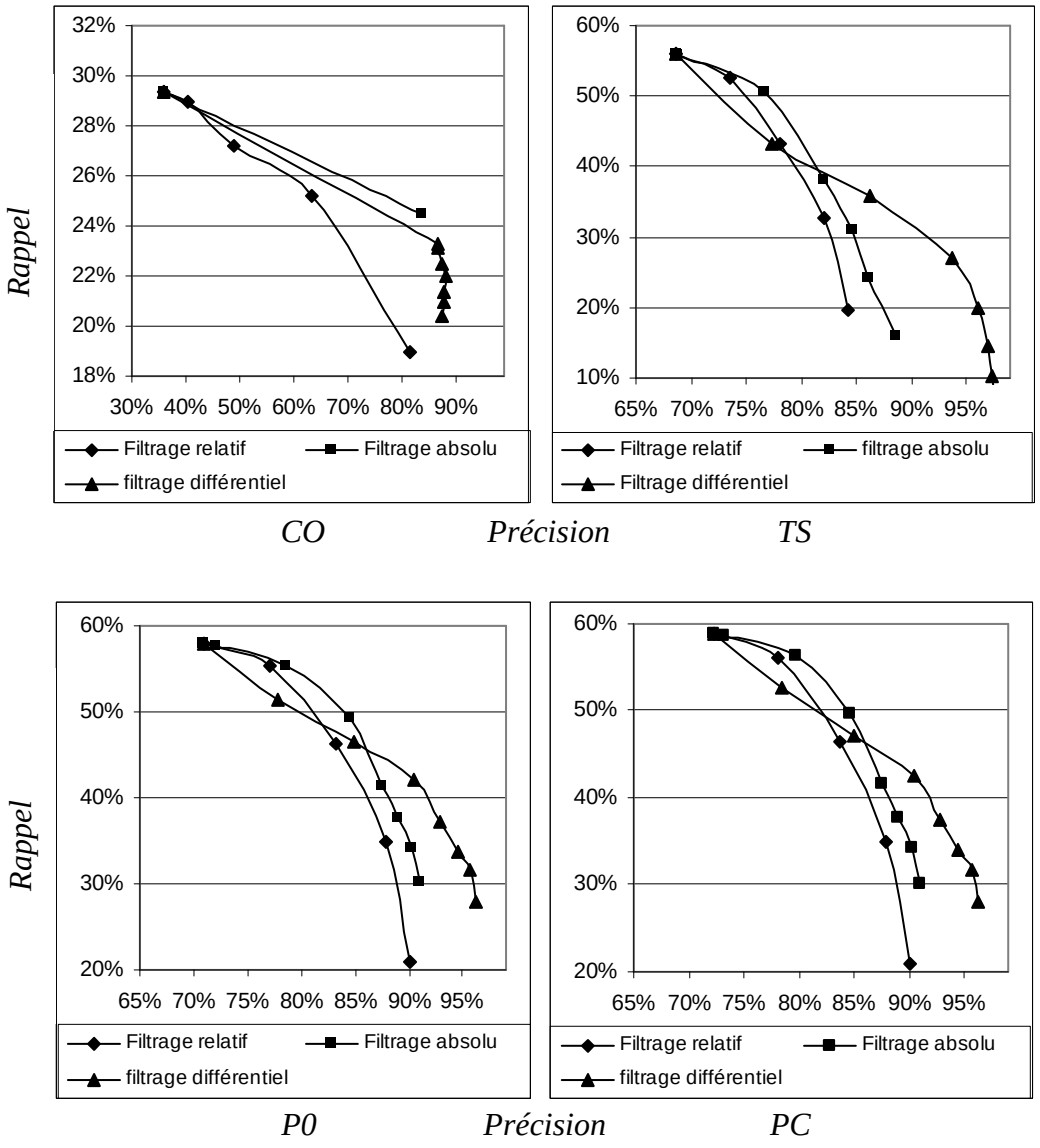



Figure 3. *Profil des différentes méthodes de filtrage pour 4 indices.*

Il apparaît que ces méthodes permettent effectivement d'augmenter la précision, au détriment du rappel, mais avec des profils différents. Le filtrage relatif implique une évolution graduelle et continue de *P* et *R*, tandis que les deux autres méthodes réalisent des variations beaucoup plus brusques. Le filtrage relatif, globalement, est moins bon, mais il permet un meilleur contrôle du taux d'écrémage, puisque l'on sait *a priori* combien de couples il va éliminer. Même si les courbes sont incomplètes du côté des valeurs basses du rappel, on constate que pour tous les indices, le filtrage absolu est supérieur si l'on cherche à maintenir un rappel élevé. En revanche, pour TS, P0 et PC, il apparaît clairement que si l'on accepte un rappel inférieur à 50 %, le filtrage différentiel est meilleur.

Si l'on se contente d'environ 30 % des couples de référence, on obtient des candidats au dictionnaire beaucoup plus fiables, avec plus de 95 % de couples corrects.

## Nouvel indice

En partant d'une liste de correspondances filtrées, on constitue en quelque sorte un mini-dictionnaire *ad hoc*, propre au corpus étudié. On peut alors renverser les rôles, et examiner si un tel « dictionnaire » peut servir à élaborer un nouvel indice. Puisque ce type de dictionnaire enregistre des fréquences d'association, l'élaboration d'un tel indice est simple : on se contentera de calculer la probabilité, étant donnée une unité *u* de la phrase source, qu'elle soit en correspondance avec une unité cible u'.

$$d = p(u'/u) = \frac{n_{12}}{n_1}$$

Nous avons effectué une extraction des correspondances lexicales sur l'intégralité du corpus. Pour des raisons de calcul, cette extraction a porté sur l'indice P0 (sans les cognats) et sur les formes simples seulement (sans grouper les unités polylexicales). A partir du jeu de correspondances obtenu, différents « dictionnaires » ont été constitués, correspondant à deux types de filtrage, et différents paramètres :

- le filtrage absolu avec les 8 seuils suivants :
  x = 0,25 x = 0,5 x = 1 x = 2,5 x = 4 x = 6 x = 10 x = 20

- le filtrage différentiel avec 11 seuils suivants :
  s = 1 s = 1,05 s = 1,2 s = 1,5 s = 2 s = 2,5
  s = 3 s = 4 s = 5 s = 6 s = 7 s = 10

Pour chacun des « dictionnaires » ainsi construits, nous avons calculé les valeurs de précision et rappel *P* et *R* correspondantes, puis nous avons effectué une nouvelle extraction en utilisant l'indice *d*. Les figures 3 et 4 montrent les valeurs de précision $P_d$ et rappel $R_d$ pour chacune de ces extractions, comparées aux valeurs de précision et rappel *P* et *R* liées aux « dictionnaires ».

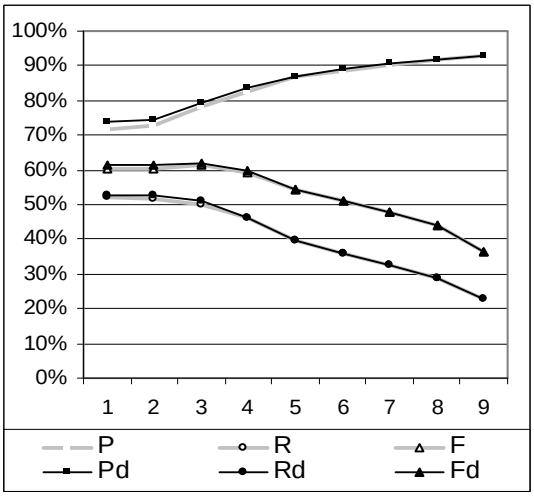



Figure 4. *Précisions, rappels et F-mesures des « dictionnaires » obtenus par filtrage absolu, et des extractions déduites de l'indice d.*

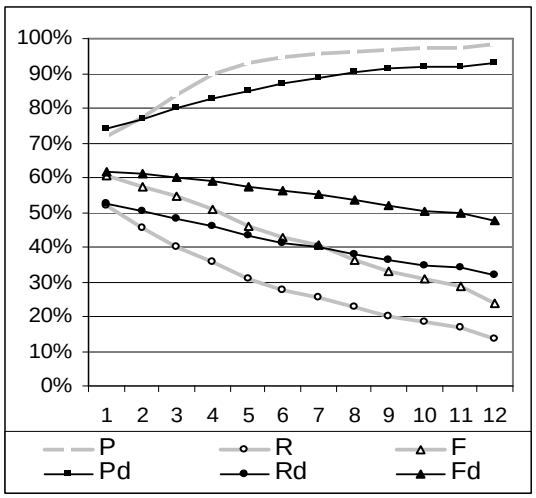



Figure 5. *Précisions, rappels et F-mesures des « dictionnaires » obtenus par filtrage différentiel, et des extractions déduites de l'indice d.*

En ce qui concerne le filtrage absolu, on constate sans étonnement que $P_d$ et $R_d$ sont calquées sur $R_d$ : l'indice *d* renvoie les mêmes correspondances que celles qui ont permis de le fabriquer.

Pour le filtrage différentiel, le résultat est moins trivial : en fait, on assiste a un accroissement notable du rappel surtout pour les seuils élevés. La précision diminue légèrement, mais dans une moindre mesure, si bien que l'on constate une différence de presque 25 % de F-mesure pour le seuil le plus élevé (s = 10). L'interprétation de ce phénomène est simple : le filtrage différentiel est basé sur la concurrence des associations. Ainsi, un couple qui est éliminé dans un certain contexte, pour cause de forte concurrence, peut néanmoins être retenu dans un autre contexte, où aucune autre association ne vient lui faire de l'ombre. Ainsi, les couples retenus sont plus variés qu'avec le filtrage absolu, qui filtre toujours les mêmes couples quel que soit le contexte. C'est pourquoi certains couples éliminés lors de l'extraction du dictionnaire peuvent être récupérés dans l'extraction déduite de *d*, si globalement leur indice *d* est bon. D'où l'augmentation du rappel, sans baisse notable de la précision, car les couples sont presque tous corrects.

En d'autres termes, le fait de réinjecter les correspondances dans l'indice *d* après un filtrage différentiel permet d'obtenir les mêmes correspondances, mais dans un plus grand nombre de contextes. Ce qui peut être intéressant pour un lexicographe, car on obtient ainsi une liste de correspondances fiables pour des contextes variés : par exemple avec s = 7, on obtient des couples corrects à 92 % représentant 33,9 % des couples de référence.

## Evaluation du « repérage de traduction » (Arcade)

Afin de situer ces résultats par rapport à ceux des systèmes ayant participé à la campagne d'évaluation Arcade [VER 00], nous avons testé les précédentes techniques sur la tâche du « repérage de traduction » (en anglais « *translation spotting* »), c'est-à-dire sur l'appariement des 3 722 occurrences d'une soixantaine d'unités (20 adjectifs, 20 noms et 20 verbes) du corpus JOC, choisies pour leur caractère polysémique. Les appariements de référence, manuellement extraits par deux annotateurs, intègrent les unités sélectionnées en y agrégeant parfois des unités voisines lorsque l'équivalence traductionnelle l'exige : par exemple, pour l'appariement de l'adjectif *sensible*, on trouve des couples tels que : (*est très sensible* ; *shares the concern*) ou (*plus en plus sensible* : *greater*). En fait, ces appariements représentent une forme d'alignement différente par rapport aux

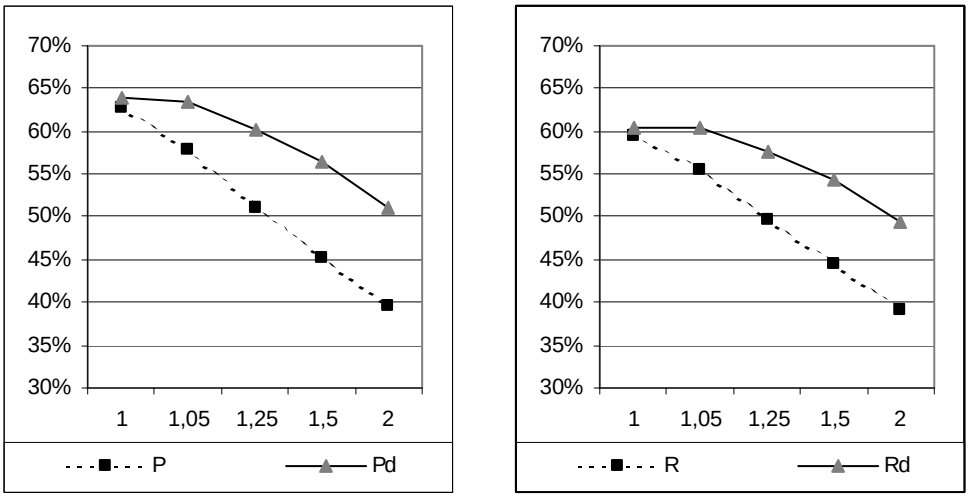
correspondances lexicales telles que nous les avons définies. Les unités concernées n'ont pas exactement le même statut : l'alignement concerne des segments textuels, tandis que les correspondances concernent des lexies. Mais vis-à-vis des expérimentations modestes que nous avons menées, cela ne change rien : nous n'avons apparié que des formes simples, sans chercher à effectuer les regroupements nécessaires pour obtenir des lexies à part entière ou des segments pleinement équivalents. Cette simplification est bien sûr pénalisée par le mode d'évaluation : si notre système renvoie (*populaire* ; *people*) au lieu de (*populaire* ; *of the people*), la précision sera de 1 avec un rappel de seulement 1/3.

En outre, dans le cas d'une unité qui n'a pas de contrepartie dans le corpus de référence, le système d'évaluation d'Arcade considère qu'elle est appariée avec le mot vide (NULL). Ainsi, lors de l'évaluation d'un jeu de correspondances, toute correspondance absente est prise comme un appariement avec le mot vide. Ceci pose problème vis-à-vis de nos méthodes de filtrage : lorsque des couples sont éliminés, ils sont considérés par le système d'évaluation d'Arcade comme des appariement avec NULL, et sont donc comptés comme des appariements erronés. Le filtrage engendre donc une baisse de la précision, au lieu d'affecter seulement le rappel. C'est pourquoi nous avons utilisé deux métriques : la première étant identique à celle du système d'Arcade, et la seconde limitant le calcul de la précision et du rappel au seuls couples extraits[3].

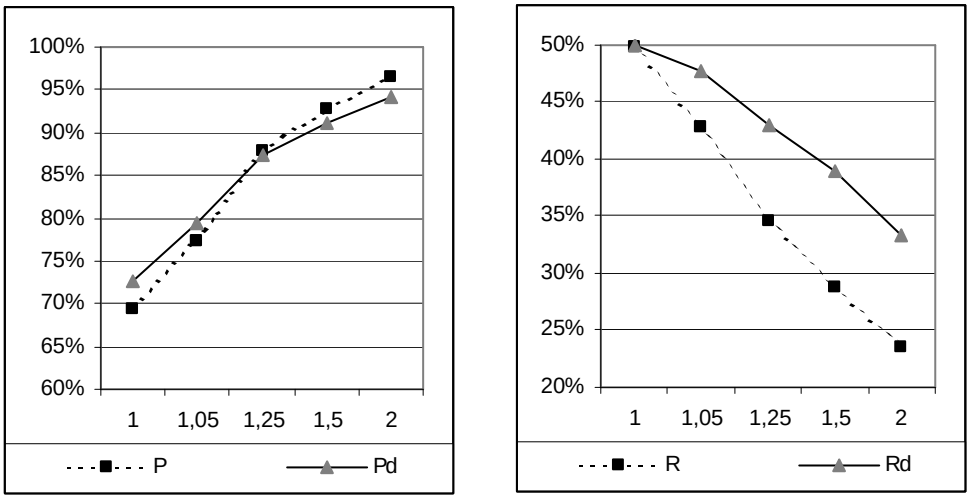
Nous avons d'abord effectué 5 extractions en utilisant l'indice P0, avec 5 seuils de filtrage différentiel ($s = 1$ ; $s = 1{,}05$ ; $s = 1{,}25$ ; $s = 1{,}5$ et $s = 2$). Ensuite, pour chacune de ces extractions, nous avons réitéré une nouvelle extraction à partir de l'indice *d*.

Les résultats de ces 10 extractions sont présentés figure 6.

[3] Toutefois, le rappel est calculé par rapport à l'ensemble des appariements de référence.




*Métrique d'Arcade (pas de correspondance = mot vide)*




*Evaluation des seuls couples extraits*

Figure 6. *Précisions et rappels* P *et* R *pour* P0 *en fonction des filtrages de seuil s = 1 ; s = 1,05 ; s = 1,25 ; s = 1,5 et s = 2.* $P_d$ *et* $R_d$ *correspondent aux extractions résultant de l'indice* d *dans une deuxième itération.*

On constate que, là encore, l'itération avec l'indice *d* améliore notablement le rappel après un premier filtrage. Même lorsque aucun filtre n'a été appliqué, les résultats progressent légèrement : on a, avec la métrique

d'Arcade, P = 62,8 % et R = 59,4 % puis $P_d$ = 63,8 % et $R_d$ = 60,5 %. L'élargissement des couples extraits à de nouveaux contextes est donc un phénomène assez général lors de la deuxième itération.

Lors de la campagne de 1998, le meilleur système avait obtenu de bons résultats : 77 % de précision et 73 % de rappel. Nos propres résultats sont honorables étant donnée la simplicité, et la « pauvreté » des techniques employées : nous n'avons utilisé ni dictionnaire, ni lemmatiseur, ni tokeniseur. Dans leur nudité, les outils présentés sont cependant intéressants, comme en atteste le haut degré de précision atteint après filtrage.

# Conclusion

Avec Saint-Jérôme, nous avons admis que, dans la traduction, le transcodage des mots n'a rien d'automatique, car il n'y a pas de règles générales pour le transfert des unités lexicales. Cependant, les résultats précédents confirment que si la traduction ne peut se réduire au simple transcodage des mots, elle n'est pas exclusive de celui-ci. L'absence de *règles* n'empêche en rien l'émergence de *régularités* : les unités équivalentes au niveau de leur valeurs, définies abstraitement au sein de leurs codes respectifs, ont tendance à apparaître fréquemment ensemble, malgré des contextes variés. Ainsi, nous avons montré qu'en appliquant des outils statistiques et des algorithmes simples, essentiellement basés sur l'observation des occurrences et des cooccurrences, il est possible d'extraire automatiquement des listes de couples susceptibles de figurer dans un dictionnaire bilingue. L'intérêt de ces méthodes est évident, puisque le « dictionnaire brut » ainsi constitué permet de donner pour chaque unité, en même temps qu'une liste d'équivalents potentiels, des contextes pour illustrer ces équivalences, ainsi que les fréquences de celles-ci. Bien sûr, il ne s'agit pas de supplanter le lexicographe : le « dictionnaire » ainsi extrait est tout au plus un lexique bilingue, qui peut servir d'outil pour la constitution d'un véritable dictionnaire, ou comme adjuvant à celui-ci, en offrant un mode consultation différent, puisqu'il permet de naviguer à l'intérieur d'exemples issus de la pratique concrète de la traduction.

Notre parti pris, pour que ces lexiques bilingues soient utilisables, est de rechercher la précision maximale, au détriment du rappel : certes, de la sorte, on court le risque de ne repérer que les correspondances les plus fréquentes, qui sont capables de se détacher au dessus de certaines, plus rares, qui ne présentent pas moins d'intérêt. Mais cet écueil peut être

contourné simplement : nous avons montré ailleurs [KRA 01b] que la taille du corpus est déterminante, et que plus le corpus est vaste, meilleure est la précision des résultats. De ce fait, en augmentant la quantité de textes traités, on peut améliorer la qualité des correspondances tout en élargissant la couverture à une plus grande variété de couples. Le principal enjeu est de trouver un corpus suffisamment grand, traduit, et cohérent avec la variété d'idiome que l'on se propose d'explorer.

Enfin, notons que les résultats des techniques présentées sont tributaires d'une dimension très importante, qui a été négligée dans la présente étude : l'identification des unités polylexicales qui forment un tout dans le passage à la traduction. Appliqué à des corpus de grandes dimensions, l'identification automatique de telles unités apparaît comme une nécessité. Le recours à des dictionnaires monolingues, couplé à des ressources telles que lemmatiseurs et outils de traitement statistiques, doit permettre d'aller un peu plus loin dans le filtrage des régularités. Du « désordre » des textes traduits, on pourra alors faire émerger l'ordre des équivalences interlingues qui s'y dissimule.

## Remerciements

Merci à Jean Véronis pour ses encouragements, et pour son aide dans la fourniture des corpus et des outils nécessaires à nos expérimentations.

## Références

[BRO 93] BROWN P., DELLA PIETRA S., DELLA PIETRA V., MERCER R., The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, vol. 19, n. 2, 1993, pp. 263-311.

[CHA 96] CHANG J. J. S., KER S. J., Aligning More Words with High Precision for Small Bilingual Corpora, *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996.

[CHU 90] CHURCH K. W., HANKS P., Word Association Norms, Mutual Information, and Lexicography. *Machine Translation*, vol. 16, n. 1, 1990, pp. 22-29.

[DAG 93] DAGAN I., K.W. CHURCH, W. GALE, Robust Bilingual Word Alignment for Machine Aided Translation, *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, 1993, p. 1-8.

[DEM 77] DEMPSTER A., LAIRD N., RUBIN D., Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 34 (B), 1977.

[DUN 93] DUNNING T., Accurate Methods for the Statistics of surprise and Coincidence. *Computational Linguistics*. Vol 19, 1, 1993, pp. 61-74.

[FUN 94] FUNG P., CHURCH K.W., K-vec : A New Approach for Aligning Parallel Texts, *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, 1994, p. 1096-1102.

[GAU 95] GAUSSIER E., LANGE J.-M., Modèles statistiques pour l'extraction de lexiques bilingues, *T.A.L.*, Vol. 36, N° 1-2, 1995, p. 133-155.

[KAY 00] KAY Martin, Preface, In Véronis, J. (Ed.), *Parallel Text Processing*, Kluwer Academic Publishers, 2000, p. xi-xviii.

[KRA 00] KRAIF Olivier, Extraction automatique de correspondances lexicales évaluation d'indices et d'algorithmes, *Actes de TALN 2000*, 16-18 octobre 2000, Lausanne, pp. 225-236.

[KRA 01a] KRAIF Olivier, Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation, *TAL*, ATALA, Paris, N°. 42 vol. 3, 2001.

[KRA 01b] KRAIF Olivier, *Constitution et exploitation de bi-textes pour l'aide à la traduction, Thèse de Doctorat*, Université de Nice Sophia-Antipolis, 2001.

[KUP 93] KUPIEC J., An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics, ACL-93*, Colombus, Ohio, 1993, pp. 17-22.

[LAP 94] LAPLACE Colette, *Théorie du langage et théorie de la tradu*ction, Paris, Didier, 1994.

[MAC 96] MC ENERY A. M., OAKES M. P., Sentence and word alignment in the CRATER project, In Thomas J., Short M. (Ed.), *Using Corpora for Language Research*, London, Longman, 1996, p. 211-231.

[MAH 99] MAHIMON, Marie-Dominique, *Identification des équivalences traductionnelles sur un corpus Français / Anglais*, Mémoire de DEA sous la dir. de Jean Véronis, Université de Provence Aix-Marseille 1, URL: http ://m.mahimon.free.fr/align_a.htm, 1999.

[MEL 97] MELAMED I. D., A Word-to-Word Model of Translational Equivalence, *35th Conference of the Association for Computational Linguistics (ACL'97)*, Madrid, 1997.

[PER 93] PERGNIER Maurice, *Les fondements sociolinguistiques de la traduction*, Lille, Presses Universitaires de Lille, 1993.

[SEL 75] SELESKOVITCH Danica, *Langue, Langage et Mémoire*, 1975.

[SEL 80] SELESKOVITCH Danica, Pour une théorie de la traduction inspirée de sa pratique, *META*, Vol. 45, N°4, déc. 1980, Outremont, Canada.

[VER 00] VERONIS Jean (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, 2000.

[VER 00] VERONIS Jean, LANGLAIS Philippe, *Evaluation of parallel text alignment systems – The ARCADE project*. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, 2000.

## Annexe I

Calcul de la précision et du rappel pour une extraction de correspondances lexicales.
Si l'on note *C* l'ensemble des couples à évaluer, $C_{ref}$ l'ensemble des couples de référence, on a :

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

## Annexe II

- *Information mutuelle* : 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 :

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

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

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 qui co-occurrent moins de 3 fois.

- *T-score* : son calcul est voisin de celui de l'IM, et il corrige le défaut mentionné ci-dessus.

$$t \approx \frac{p_{12} - p_1 p_2}{\sqrt{\frac{p_{12}}{n}}}$$

- *Rapport de vraisemblance* (ou « log-like ») : Dunning [DUN 93] 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 ces deux 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* |

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

$$RV = -2\log\lambda = 2(S_+ - S_-)$$

$$S_+ = a\log a + b\log b + c\log c + d\log d + n\log n$$

$$S_- = (a+c)\log(a+c) + (b+d)\log(b+d) + (a+b)\log(a+b) + (c+d)\log(c+d)$$

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

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

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

## Annexe III

*Paramètres pour l'indice CO :* Dans la comparaison superficielle des caractères de deux chaînes, nous avons identifié 11 cas de figures, pour lesquels les probabilités empiriques sont distinctes dans le cas de l'hypothèse nulle (chaînes prises au hasard) :

Cas 1 = chaînes numériques identiques ;
Cas 2 = transfuge (chaîne identique) de longueur supérieure à 3 ;
Cas 3 : 4-grammes – 7-grammes (sous chaîne commune contiguë de longueur comprise entre 4 et 7) ;
Cas 4-9 : Sous-chaînes maximales (plus longue sous-chaîne commune) comportant entre 4 et 9 caractères.

Cas 10 : Sous-chaînes maximales (plus longue sous-chaîne commune) comportant au moins 10 caractères.
Cas 0 : complémentaire des cas 1 à 10.

Nous avons relevé les probabilités empiriques suivantes, pour tous les couples possibles à l'intérieur des phrases alignées du corpus COUR (sous-partie du corpus BAF employé dans la campagne d'évaluation ARCADE).

| 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 empiriques, considérées comme représentatives du couple français- anglais ont été réutilisées avec succès pour d'autres corpus.

L'indice CO est calculé comme $-log(p(cas))$.