Conférence TALN 1999, Cargèse, 12-17 juillet 1999

Identification des cognats et alignement bi-textuel : une étude empirique

Olivier Kraif

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

Résumé

Nous nous intéressons ici aux méthodes d'alignement automatique destinées à produire des corpus bi-textuels, utiles au traducteur, au terminologue ou au linguistique. Certaines techniques ont obtenu des résultats probants en s'appuyant sur la détermination empirique des « cognats » (de l'anglais « cognate »), des mots qui se traduisent l'un par l'autre et qui présentent une ressemblance graphique. Or les cognats sont généralement captés au moyen d'une approximation abrupte, de nature opératoire : on considère tous les 4-grammes (mots possédants 4 lettres en commun) comme cognats potentiels. Aucune étude n'a été faite, à notre connaissance, à propos de la validité de cette approximation. Afin d'en démontrer les possibilités et les limites, nous avons cherché à déterminer empiriquement la qualité de cette simplification, en termes de bruit et de silence (ou de manière complémentaire, de précision et de rappel). Nous avons ensuite essayé de développer un filtrage plus efficace, basé sur l'utilisation des sous-chaînes maximales. Enfin, nous avons corrélé les améliorations du filtrage avec les résultats de l'alignement, en nous basant sur une méthode générale développée par nous : nous avons pu constater un net progrès en terme de rappel et de précision de l'alignement.

1. Introduction

Un bi-texte, noté <T1,T2,S,C>, est un corpus constitué de deux textes T1 et T2 dont l'un est traduction de l'autre, doté d'une fonction de segmentation S, permettant de découper les deux textes en unités plus petites (paragraphes, phrases, syntagmes), et d'une fonction de correspondance C permettant d'apparier les segments en relation de traduction (Isabelle, 1992). Ainsi conçu, un corpus bi-textuel constitue un matériel privilégié dont les applications sont intéressantes dans de nombreux domaines :

- dans l'apprentissage de la pratique traductionnelle, en tant que réserve d'exemples concrets, sous la forme d'un concordancier bilingue.

- dans l'enseignement des langues.

- en terminologie différentielle, afin d'observer les équivalences usitées dans la pratique effective à l'intérieur d'un domaine précis.

- en linguistique, dans l'étude contrastive des langues vivantes : étude des distributions lexicales, des correspondances morphologiques, des divergences phraséologiques, etc..

- en traduction automatique, pour la constitution de systèmes de traduction basés sur l'exemple. De même en aide à la traduction, les bi-textes peuvent être constitutifs d'une mémoire de traduction, utilisable comme un répertoire de solutions déjà trouvées en réponse à des problèmes spécifiques (de terminologie, de phraséologie).

- pour la vérification automatique de traduction (détection de faux amis, d'omissions, etc..) (Isabelle,1992), l'aide à la rédaction en langue étrangère, etc...

Les techniques d'alignement visent à la production massive de ces bi-textes : elles ont pour but de mettre en correspondance, par un traitement automatique, les portions de textes qui sont traductions les unes des autres.[1] Nous nous intéressons ici à une de ces méthodes, fondée sur l'exploitation d'indices lexicaux (les traductions des mots) et s'appuyant sur une approximation plutôt abrupte : sont considérés comme traductions potentielles tous les mots possédant plus de 4 caractères en commun (on écrira : 4-grammes). Entre des langues apparentées, comme l'anglais et le français, cette hypothèse a déjà été mise en œuvre et a fourni de bons résultats (Simard, Foster, Isabelle, 1994) (K.W. Church, 1993).

Cependant aucune étude n'a été faite, à notre connaissance, à propos de la pertinence de cette approximation. Afin de dégager les fondements théoriques de la méthode et d'en préciser les limites, nous avons cherché, à partir d'un travail empirique, à examiner cette hypothèse selon trois angles différents : dans un premier temps nous avons étudié le rappel et la précision[2] empirique de l'approximation des n-grammes vis-à-vis de l'identification des cognats ; puis nous avons tenté d'améliorer les résultats de cette approximation, moyennant quelques ajustements ; enfin nous nous sommes intéressés à l'incidence de la qualité du filtrage des cognats sur les résultats de l'alignement, dans le cadre d'une méthode originale que nous avons développée.

2. Principes de l'alignement

Nous nous plaçons dans le cadre de l'alignement au niveau des phrases. Nous ne discuterons pas des difficultés inhérentes à la notion de phrase dont les définitions peuvent varier (suivant que l'on se base sur la ponctuation, les données prosodiques, le contenu logique, etc.) et dont les limites ne sont pas toujours évidentes (problèmes de ponctuation ambiguë, phrases imbriquées, listes, etc.). La phrase, comme le mot, sera pour nous une unité opératoire.

Si l'on représente les deux textes à aligner comme étant deux ensembles de phrases S1 et S2, trouver un alignement consiste à trouver un sous ensemble du produit cartésien S1xS2. Pour que l'alignement ait un sens, les deux textes doivent respecter les deux conditions du *parallélisme* énoncés par Langé et Gaussier (1995) :

- *quasi-bijectivité* : toute phrase source a en général un correspondant dans le texte cible, et réciproquement.
- *quasi-monotonie* : la séquence des phrases sources doit suivre, en général, la séquence des phrases cibles correspondantes.

---

[1] Si l'on numérote ces portions, un alignement peut donc être représenté sous la forme d'un ensemble de points de coordonnée (x,y), où x et y représentent les numéros des deux portions appariées.

[2] Pour l'évaluation quantitative, nous utilisons les mesures de précision et de rappel mises au point dans le Projet Arcade. Pour plus de détail, cf. supra.

Nos développements sur l'alignement présupposent la vérification de ces hypothèses.

Même dans le cadre du parallélisme, l'alignement n'est pas une tâche triviale dans la mesure où une même phrase peut apparaître dans plusieurs de ces couples, ou dans aucun : il est fréquent en effet qu'une phrase soit traduite par une ou plusieurs phrases, et il arrive qu'elle soit omise.

Il existe un certain nombre de techniques d'alignement, désormais devenues classiques. Elles diffèrent par le type d'information utilisé : les longueurs de phrase (Gale et Church, 1991 ; Brown, 1991), les distributions lexicales (Kay et Röscheinsen, 1993 ; Fung, 1994), ou encore les similitudes de surface entre les mots (Simard, Isabelle et Foster ; 1992).

Ces derniers ont enrichi la méthode de Gale et Church par l'exploitation de l'identification des cognats, i.e. des équivalents traductionnels présentant une « ressemblance » tant au plan sémantique que graphique. Les cognats, dans un sens élargi, rassemblent à la fois les chaînes invariantes (comme les noms propres, les données numériques) et les mots apparentés (racines communes, emprunts). L'hypothèse de « cognacité », vérifiée empiriquement entre des langues européennes, peut être ainsi formulée : « la densité de cognats observée entre deux phrases est probablement plus élevée si elles sont traduction l'une de l'autre, que si elles sont prises au hasard ». Sans entrer dans les difficultés de la définition linguistique de cette notion de « ressemblance », les méthodes basées sur les cognats s'appuient sur une donnée très simple : la longueur de la suite maximale de caractères contigus communs. Pour une suite de longueur n, on parlera de « n-gramme ». Certains systèmes (Simard et al, 1992 ; K.W. Church, 1993), en se basant sur n=4 ont obtenu des résultats significatifs : nous pensons que ces résultats prometteurs peuvent être améliorés par un raffinement de cette approximation.

3. La notion de cognat

Nous avons travaillé sur des textes issus d'un corpus bilingue, le BAF, constitué au RALI de l'Université de Montréal, et gracieusement prêté dans le cadre du projet Arcade centré sur l'évaluation des méthodes d'alignement (Langlais, Simard, Véronis *et al*, 1998). Il s'agit de texte institutionnels issus de la Court suprême du Canada : environ 31000 mots en anglais et 33000 en français.

Afin de pouvoir évaluer le bruit généré par le recours aux n-grammes, il nous a fallu définir manuellement les couples de cognats observables au sein de notre corpus.

Etant données les difficultés inhérentes à la détermination des cognats, nous sommes parti d'une définition opératoire visant à minimiser les ambiguïtés dues à la notion de ressemblance. Deux mots M1 et M2 sont cognats si et seulement si :

1. il existe deux phrases (P1,P2) dont l'une est la traduction de l'autre, et dans lequel ils sont traduction l'un de l'autre.
2. M1 et M2 présentent un lien étymologique (emprunt, origine commune) perceptible dans leur signifiant.

On y ajoutera les transfuges, c'est-à-dire les invariants de la traduction (par exemple les nombres et certains noms propres).

Le critère 2 ainsi que la notion de transfuge n'ont pas posé de problèmes significatifs (l'existence de liens étymologiques permettant de donner une assise objective à la notion confuse de ressemblance). En revanche, décider de la *traduisibilité* d'une forme par une autre implique des difficultés :

- D'une part un mot peut être traduit par un phrasème : par exemple « because » <-> « à cause ». On retient alors le couple portant l'étymon commun : « because » <-> « cause ».
- D'autre part il est parfois difficile de déterminer si un mot *peut* en traduire un autre : la traduction mot-à-mot est un cas limite, éloigné de la pratique effective de la traduction. Comme nous l'avons souligné dans une précédente discussion sur la notion contestable « d'alignement lexical » (Kraif, 1999), il n'est pas possible d'étendre l'hypothèse de parallélisme au mot, à l'intérieur des phrases. Or des mots d'étymologie commune mais de sémantisme différent peuvent, dans un certain contexte, se retrouver en relation de traduction. Par exemple, les mots « importation »(eng.) et « export » (fra.) sont d'« assez bons » cognats : on peut leur imaginer un contexte de traduction : « Il fait de l'export vers les USA » <-> « He makes importations from France ». Mais entre « sensible » (eng.) et « sens », l'écart sémantique paraît plus grand. On peut pourtant trouver aussi un contexte de traduction : « He's sensible » <-> « Il est plein de bon sens ». Jusqu'où peut-on accepter ces distorsions sémantiques ?

Nous contournons cette difficulté par un parti pris restrictif : au sein de notre corpus, nous n'identifions comme cognats que les mots qui sont effectivement traduits l'un par l'autre, dans le corpus. Ceci peut introduire un léger biais. Par exemple « appeal » (eng.) et « appelant » (fra.) peuvent apparaître dans des phrases différentes, donc ne pas être notés comme traduction l'un de l'autre, alors que ce sont bien des cognats. Mais dans la mesure où ce sont les cognats *effectifs* du corpus qui peuvent aider à l'alignement de celui-ci, on peut supposer que ce biais n'enlèvera rien à la validité de nos résultats quant aux statistiques permettant de lier l'identification des cognats avec la qualité de l'alignement obtenu.

4. Algorithme d'alignement

Nous avons démontré dans des travaux précédents (Kraif, 1998) qu'il était possible, en se basant uniquement sur l'identification des cognats et des transfuges, d'obtenir un préalignement robuste et précis en un temps quasi-linéaire, en $O(n \log(n))$[3]. Un tel algorithme peut servir d'étape préalable à l'application de méthodes plus sophistiquées basées sur la programmation dynamique[4], en un temps qui demeure quasi-linéaire.

Notre algorithme se base sur une heuristique simple : le principe de « précision d'abord ». Concrètement, ce principe commande d'utiliser d'abord les informations les plus fiables pour en tirer un préalignement grossier mais très sûr. La méthode est ensuite appliquée de nouveau, récursivement, à l'intérieur des sections déjà alignées : le rappel peut ainsi augmenter sans décroissance de la précision.

Ainsi, l'alignement dégagé met relation des couples de phrases, mais reste fragmentaire: pour deux textes ($P_1P_{2...}P_n$) et ($P'_1P'_2\ldots P_n$), on examine tous les couples ($P_{i,}P'_j$) considérés comme alignables (situés à l'intérieur d'une bande autour de la diagonale), et on en dégage une suite de couples de phrase (ou points d'ancrage) représentant un pré-alignement (c'est à dire un sous-ensemble de l'alignement complet). Les points d'ancrages ainsi dégagés serviront ultérieurement de « points de capiton » pour un alignement exhaustif.

---

[3] à titre d'indication, nous avons obtenu une précision de 99,4% et un rappel de 77,6% en moyenne sur l'ensemble du corpus BAF. Le log(n) est du aux recherches dans nos index, sous forme d'arbres binaires.

[4] par exemple, en appliquant la méthode de Gale et Church après cet algorithme, on obtient une précision de 96% en moyenne pour un rappel de 86%, toujours sur le BAF.

Identification des cognats et alignement

Dans une première étape, on se base sur l'exploitation des transfuges seuls (considérés comme les indices les plus fiables). On implémente un processus itératif en deux temps :

1. prise en compte de tous les transfuges apparaissant le même nombre de fois dans les deux sections à aligner. Puis on apparie ces occurrences pour obtenir un ensemble de points d'alignement.
2. filtrage des points selon les critères suivants, qui traduisent l'hypothèse de parallélisme :
   - *diagonalité* : suppression des points éloignés de la diagonale.
   - *continuité* : suppression des points présentant une déviation forte par rapport aux points précédents.
   - *monotonie* : suppression des points entrant en conflit sur l'une de leur coordonnées, ainsi que les points croisés : (x,y) et (x',y') se croisent si x>x' et y<y'.

Pour maximiser la précision, on impose en outre une condition de surdétermination : on ne retient que les points générés par au moins deux transfuges différents.

A l'issue de l'étape 2, chaque point donne lieu à un découpage de la section alignée en sous-sections alignées. Puis l'on réitère les étapes 1 et 2 sur chaque sous-section, récursivement, jusqu'à stabilité.

- Dans une deuxième étape, on examine tous les couples de phrases alignables à l'intérieur des sections préalignées : on compte la fréquence des cognats f_{ij} identifiés entre les phrases P_i et P'_j de chaque section.. A partir de la matrice (f_{ij}) ainsi obtenue, on calcule une nouvelle matrice exprimant le lien statistique entre les lignes i et les colonnes j :

$$c_{ij} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

Pour obtenir des points d'alignement, on applique une condition de réciprocité : (i,j) peut donner un point si c_{ij} est la borne supérieure de la ligne i *et* de la colonne j.

Les points obtenus sont ensuite filtrés avec les mêmes critères qu'auparavant (*diagonalité*, *continuité*, *monotonie*).

Notons que chaque matrice est calculée entre les points fixés par la première étape. Si l'on peut montrer que le rappel de l'étape 1 est (pour des textes normalement parallèles) supérieur à un certain seuil, on obtient ainsi un espace de calcul en O(n). Concrètement, sur le Corpus BAF cette hypothèse est tenable : l'étape 1 obtient un rappel oscillant de 40% à 91% (sauf pour un seul texte qui ne remplit pas les conditions de parallélisme, et pour lequel le rappel a peu de signification).

5. Précision et rappel des n-grammes

Au cours de l'algorithme précédent, la présence de n-grammes a permis de filtrer un certain nombre de cognats potentiels. Nous voulons maintenant déterminer dans quelles proportions ces appariements correspondent à des cognats véritables, ou produisent du bruit.

Il existe deux façons de calculer les statistiques du filtrage des cognats : soit on ignore les fréquences des mots de chaque texte, et l'on considère tous les appariements possibles entre les deux lexiques, en comptabilisant tous les n-grammes aboutissant à des appariements justes

ou erronés ; soit on établit ces statistiques à l'intérieur des phrases qui sont effectivement comparées, dans l'espace de calcul de l'algorithme, et un même appariement interviendra autant de fois qu'il entre dans une comparaison. Nous avons opté pour cette dernière solution, car les statistiques sont ainsi directement liées à l'exploitation de l'algorithme.[5]

Pour l'évaluation du silence et du bruit on utilise les trois mesures suivantes : la précision P exprime la proportion de cognats trouvés par rapport au nombre d'appariements donnés ; le rappel R exprime la proportion de cognats trouvés par rapport nombre total de cognats existants (entre les phrases comparées) ; et la F-mesure représente la combinaison de P et R : $F=2PR/(P+R)$[6].

On obtient les statistiques du tableau 1. Nous y avons fait figurer, dans la première colonne, les statistiques de précision et rappel liées aux transfuges, afin de servir de base de comparaison. Notons qu'un certain nombre de ces transfuges sont comptabilisés dans les n-grammes.

L'augmentation de la précision avec le nombre de caractères communs indique clairement que plus un n-gramme est long, plus il est fiable dans la détermination des cognats. Malheureusement les indices qui génèrent le moins de bruit sont aussi les plus rares.

On constate que la prise en compte des transfuges d'au moins 2 caractères[7] donne une meilleure F-mesure qu'avec les n-grammes (la précision est de 100% car nous avons négligé tous les cas d'homographie). En revanche, les 50% de rappel obtenu par les transfuges indiquent clairement ce que peuvent apporter les n-grammes, ou d'autres techniques : il reste 50% de cognats à identifier. On peut sans doute améliorer les résultats globaux en combinant l'identité (les transfuges) et la ressemblance (les n-grammes ou autre). C'est ce que montre la troisième colonne du tableau 1.

6. Sous-chaînes maximales

Le filtrage par les n-grammes appelle deux remarques :

1. D'une part ils ne permettent pas de reconnaître la « ressemblance » lorsque celle-ci implique des ruptures à l'intérieur des groupes de lettres : par exemple « **do**cteur » et « **do**ttore »(it.) ne sont que des 2-grammes.
2. D'autre part, la signification d'un n-gramme dépend étroitement de la taille des mots comparés. Un 4-gramme entre « **form** » et « **form**e » paraît plus significatif qu'un 6-gramme entre « explo**ration** » et « décla**ration** ».

Pour palier le premier inconvénient, nous proposons de recourir aux sous-chaînes maximales (on notera SCM), à l'instar de Débili et Sammouda (1992) : la plus longue sous-chaîne de caractère commune aux deux mots (en autorisant les sauts). Par exemple, pour « docteur » et « dottore », la SCM est de longueur 4 : **d-o-t-r**. Mais la combinatoire des SCM

---

[5] Il faut garder à l'esprit que ces statistiques résultent de la comparaison de phrases voisines, à l'intérieur de blocs préalignés. Entre des phrases quelconques on peut supposer qu'elles seraient différentes (pour des raisons de continuité thématique), avec une précision et un rappel inférieurs.

[6] F est en quelque sorte une moyenne dynamique : elle se rapproche de la moyenne si P et R sont rapprochés, et elle décroît si P et R sont éloignés.

[7] Pour les transfuges comptant exactement 2 caractères, nous n'avons tenu compte que des nombres.

est très importante (surtout avec les mots longs), et risque de produire beaucoup de bruit : par exemple « pragmatic » est presque totalement inclus dans « paradigmatique ». Nous en avons donc implémenté une version plus contrainte : les sous-chaînes doivent être quasiment parallèles, c'est-à-dire que l'on n'autorise pas d'insertion ou de délétion de caractères en série.

Enfin, pour limiter le bruit et tenir compte de la remarque 2, nous tiendrons compte de la longueur des SCM par rapport à la taille des mots. On calcule le rapport entre la taille du mot le plus long et la longueur de la SCM : $r(M1,M2)=l(SCM)/\max(l(M1),l(M2))$. Puis on effectue un filtrage en fonction de r. Pour notre corpus nous avons testé différentes valeurs pour ce seuil : les meilleurs résultats ont été obtenus en acceptant les SCM avec $r \geq 2/3$.

Les colonnes 4 et 5 du tableau 1 contiennent les résultats avec les SCM seules, puis combinées avec les transfuges. On constate, comme on pouvait s'y attendre, une nette amélioration de la précision sans réduction du rappel (sauf pour les sous-chaînes courtes avec $n \geq 3$).

7. Corrélations avec l'alignement

Nous avons cherché à corréler les résultats de la méthode de filtrage des cognats avec les résultats de son exploitation, i.e. l'alignement. Les résultats obtenus pour différents types de paramétrages sont inscrits dans le tableau 2. Le rappel Ra et la précision Pa de l'alignement sont calculés suivant les mesures de « KR-mot » définies dans le projet Arcade. On a en outre testé deux mesures combinant n-grammes et SCM (lignes 10 et 11), et nous avons extrait un alignement en utilisant la liste des cognats de référence, déterminée manuellement (dernière ligne).

Les résultats obtenus appellent les observations suivantes :

- la précision de notre méthode d'alignement est peu sensible au bruit : même pour une précision Pc de 15%, la précision de l'alignement demeure au delà de 80%. Cette robustesse est due à la multiplicité des contraintes de filtrage des points d'alignement.
- le rappel de l'alignement, très sensible à la qualité du filtrage des cognats, est fortement corrélé au résultat global de ce filtrage : entre F_c et Ra, la corrélation linéaire est de 0,94. Cela confirme l'amélioration des résultats apportée par le recours au SCM. La densité des cognats identifiés entre les phrases des deux textes est donc déterminante.
- un point est marginal : l'alignement obtenu avec les données des cognats de référence (dernière ligne) est légèrement moins bon. Cela pourrait indiquer que trop de cognats (un rappel trop important) pourrait affecter le rappel de l'alignement. C'est évidemment lié à la nature de notre méthode : en effet la mesure du lien à tendance à favoriser les appariements avec les phrases courtes. Dès lors, deux phrases pesant trop lourd auraient moins de chance d'obtenir un bon score, et donc d'être alignées ensemble. Mais cette hypothèse demanderait une étude plus approfondie pour être confirmée. Quoiqu'il en soit il est raisonnable de supposer que l'amélioration de Fc peut conduire à des résultats meilleurs pour Fa, moyennant une exploitation différente de la cognacité.

8. Conclusion et Perspectives

Nous avons cherché à sortir de l'approximation classique cognat ≈ 4-gramme, dont la validité n'a encore jamais été précisément étudiée. Après avoir évalué, sur un extrait du corpus BAF, les mesures de rappel et de précision liées à cette approximation, nous avons indiqué une méthode permettant d'améliorer significativement les performances de la détermination des cognats : nous montrons comment combiner les n-grammes aux notions de transfuges et de sous-chaînes maximales afin d'obtenir un meilleur rendement.

Nous avons ensuite appliqué ces améliorations à l'intérieur d'un algorithme de préalignement développé par nos soins : cet algorithme, de complexité en O(n log(n)) étant destiné à obtenir rapidement un préalignement de bonne qualité (pour y appliquer ensuite des méthodes plus fines et plus coûteuses en calcul). Notre expérience montre qu'un tel algorithme est robuste du point de vue de sa précision, et tributaire de la qualité de l'identification des cognats en ce qui concerne son rappel. Appliqué à l'ensemble du corpus BAF, notre algorithme, avec utilisation des SCM, obtient une précision de 99,4% en moyenne (avec écart type de 0,78%) et un rappel de 77,6% en moyenne (85,1% si l'on ignore deux textes non parallèles), ce qui en confirme l'efficacité comme méthode de préalignement. Notons que cet algorithme constitue un cadre général pour l'utilisation d'autres sources d'information concernant les équivalences lexicales : dictionnaires, lexiques extraits automatiquement à partir de l'information mutuelle ou du *t-score*, etc.

Bien entendu ces conclusions ne peuvent être généralisées de manière trop hâtive : il est clair que les résultats obtenus dépendent étroitement des textes en question et, de manière encore plus déterminante, du couple de langues impliqué. Des questions restent ouvertes quant aux limites de la méthode d'alignement : il faudrait lier ses résultats à la densité de cognats effectivement présents dans les textes, puisque cette densité semble être le paramètre décisif.

Certes, en ce qui concerne la détermination des cognats, d'autres méthodes d'inspiration linguistique peuvent être employées, sans doute avec une meilleure précision. Mais on perd ainsi la généralité des méthodes ici décrites : dans ce cas le recours direct à un dictionnaire bilingue *ad hoc* semblerait à la fois plus simple et plus efficace. En outre, d'après les résultats obtenus avec les cognats de référence (précision maximum), l'alignement résultant n'est pas nécessairement meilleur : le recours aux cognats induit un bruit incompressible du à la dispersion des correspondances lexicales à l'extérieur des bornes des phrases alignées.

Notons enfin que l'utilisation des transfuges et des cognats peut s'étendre à des langues sans aucun lien de parenté : dans les domaines technologiques et scientifiques, une grande partie de la terminologie est normalisée et convergente. Des adaptations sont requises toutefois quant aux systèmes de transcription phonologiques, lorsque les alphabets sont différents.

Références

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 ACL*, Colombus, Ohio, pp.1-8

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

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

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

Kay M., Röscheisen M. (1993), Text-Translation Alignment, *Computational Linguistics*, Vol. 19, N°1, pp.121-142

Kraif O. (1998). Alignement de phrases basé sur les cognats. In *Actes des 5èmes rencontres de l'atelier des doctorants en linguistique*, Université Paris 7, Paris, 4-5 déc. 1998, pp. 31-33

Kraif O. (1999). Réflexions autour des concepts de correspondance et d'alignement textuel. In *Actes du colloque Linguistique contrastive et Traduction Approches Empiriques*, Louvain-la-Neuve, 5-6 février 1999, pp. 25-26

Langé J.-M., Gaussier E. (1995), Alignement de corpus multilingues au niveau des phrases, *T.A.L.*, Vol. 36, N° 1-2, pp. 67-80

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

Simard M., Foster G., Isabelle P. (1992). Using cognates to align sentences. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal, pp. 67-81

Annexe

Les valeurs de Précision, Rappel, et F-mesure sont données en pourcentage.

	transfuges			**n-grammes**			**n-grammes + transfuges**			**SCM**			**SCM + transfuges**		
n	**Pc**	**Rc**	**Fc**	**Pc**	**Rc**	**Fc**	**Pc**	**Rc**	**Fc**	**Pc**	**Rc**	**Fc**	**Pc**	**Rc**	**Fc**
≥2	100	50	66												
≥3	100	29	45	15	75	24	18	95	30	47	55	51	55	76	63
≥4	100	21	34	31	48	37	41	76	54	64	45	53	75	74	74
≥5	100	16	28	48	38	42	64	71	67	75	39	51	85	72	78
≥6	100	15	26	72	26	39	86	61	71	74	30	43	86	65	74
≥7	100	13	22	87	19	31	95	56	71	76	21	33	90	58	70
≥8	100	7	14	88	10	18	97	52	68	73	13	22	92	55	69
≥9	100	6	12	93	8	15	99	52	68	93	10	18	99	53	69

Tableau 1

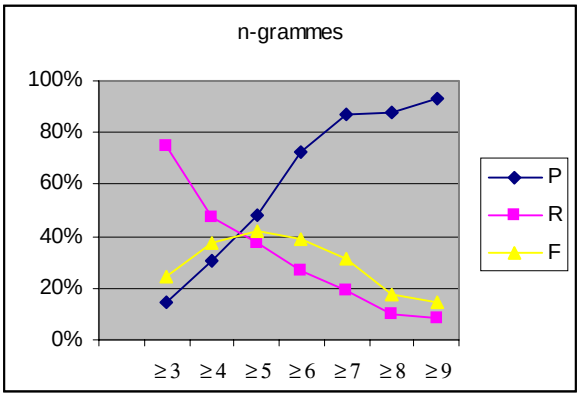

fig. 1

Précision, Rappel, F pour les n-grammes

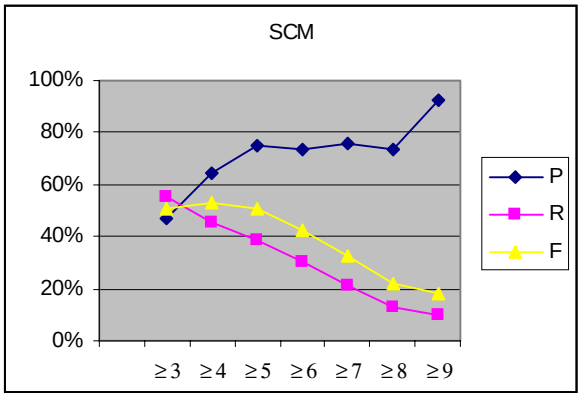

fig. 2

Précision, Rappel, F pour les SCM

	Pc	Rc	Fc	Pa	Ra	Fa
3-grammes (*)	15	75	24	82	32	45,5
4-grammes	31	48	37	97	63	76,3
CM>=3 + transfuge	55	76	63	99	76	85,9
CM>=4 + transfuge	75	74	74	100	86	92,2
CM>=5 + transfuge	85	72	78	100	85	91,9
CM>=6 + transfuge	86	65	74	100	84	91,2
CM>=7 + transfuge	90	58	70	100	82	89,9
CM>=8 + transfuge	92	55	69	100	79	88,3
CM>=9 + transfuge	99	53	69	100	76	86,3
Combinaison 1	68	73	70	99	80	88,6
Combinaison 2	75	73	74	100	86	92,5
Transfuges seuls	100	50	66	100	68	81,2
Cognats	100	100	100	100	74	85,2

Tableau 2

Coeff. de corrélation	Pa	Ra	Fa
Pc	0,76	0,74	0,75
Rc sans (*)	0,41	0,72	0,71
Fc	0,85	0,94	0,93

Tableau 3

- Combinaison 1 : transfuges, 4-grammes (mots de longueur <7), CMS avec n≥4 et r>2/3
- Combinaison 2 : transfuges, N-grammes avec N≥3 et r>2/3, CMS avec n≥5 et r=2/3
- Cognats : liste de référence obtenue manuellement.

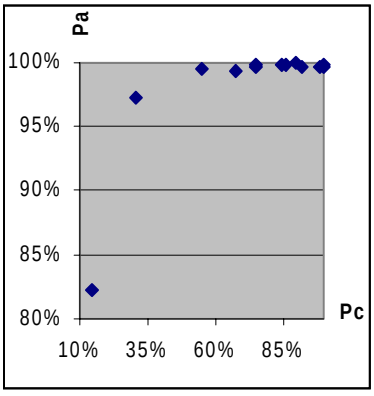

Fig. 3

Pa en fonction de Pc

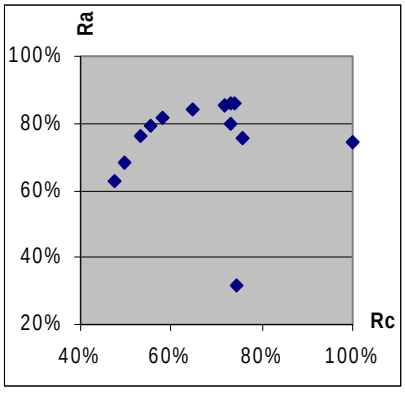

Fig. 4

Ra en fonction de Rc

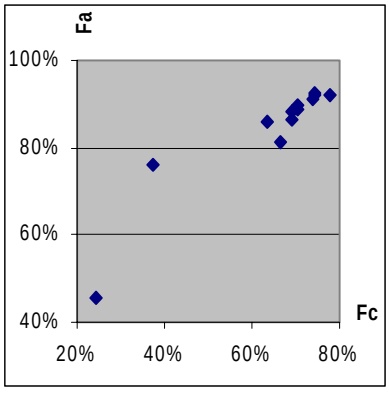

Fig. 5

Fa en fonction de Fc