Exploitation des cognats pour l'alignement

Architecture et évaluation

Olivier Kraif

Laboratoire d'informatique d'Avignon
339, chemin des Meinajariès
Agroparc BP 1228
84911 Avignon Cedex 9 - France
Olivier.Kraif@lia.univ-avignon.fr

RÉSUMÉ. *Nous nous intéressons dans cet article aux méthodes d'alignement automatique destinées à produire des corpus bi-textuels. Certaines techniques ont obtenu des résultats probants en s'appuyant sur la détermination empirique des mots étymologiquement apparentés, les " cognats ". Or les cognats sont généralement captés au moyen d'une approximation abrupte : on considère tous les mots ayant 4 lettres consécutives communes comme cognats potentiels. Nous avons cherché à déterminer empiriquement la qualité de cette simplification, en termes de précision et de rappel, afin d'en démontrer les possibilités et les limites, et d'y apporter quelques améliorations. Enfin, nous corrélons les améliorations obtenues avec l'alignement résultant, en utilisant une méthode générale de préalignement.*

ABSTRACT. *In this paper, we focus on automatic aligning methods, which aim at producing massive bi-textual corpora. Some systems have yielded good results by taking advantage of the "cognateness", i.e. the amount of related words that occur in both parts of translation. Usually, cognates are identified by a very rough approximation : if two words share the same four letter string, they are considered as potential cognates. No empirical study, as far as we know, has been conducted to validate this simplification. In order to show precisely the scope and the limits of the cognate based approach, we evaluate, on a French-English corpus, the precision and recall of n-gram simplification, and we give some possible improvements. We finally implement a cognate-based aligning system, in order to correlate the results of cognate identification with the results of the subsequent alignment.*

MOTS-CLÉS : *alignement, corpus bilingue, corpus parallèles, bi-texte, cognat.*

KEYWORDS: *Aligning, Bilingual Corpora, Parallel Corpora, Bi-text, Cognate.*

1. Introduction

Un bi-texte, noté $<T_1,T_2,S,C>$, est un corpus constitué de deux textes parallèles T_1 et T_2 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 [ISA 92]. Le concept sous-jacent au bi-texte est la *compositionnalité* de la traduction (*Ibid.*) : on entend par ce terme que la relation d'équivalence, pragmatiquement définie au niveau de la globalité des deux textes, peut s'observer entre des parties plus petites de ces textes. Généralement, l'hypothèse de compositionnalité est assez bien vérifiée jusqu'à l'échelle des phrases, et devient peu exploitable en-deçà.

Applications de l'alignement

Ainsi conçu, un corpus bi-textuel constitue un matériau privilégié dont les applications sont intéressantes dans de nombreux domaines :

– *Didactique* : En didactique des langues ou de la traduction, le bi-texte se situe dans le prolongement naturel des éditions bilingues classiques.

– *Recherches empiriques à base de corpus (linguistique contrastive, terminologie) :* Dans l'étude empirique des phénomènes contrastifs, le bi-texte permet d'accéder à un nouvelle catégorie de faits, observables seulement sur le plan des régularités statistiques. En effet, de nombreux outils statistiques permettent de corréler de façon significative, sur une grande quantité de segments alignés, les occurrences d'" événements " linguistiques de tout type : traduction de lexies[1], correspondances sur le plan de la morphosyntaxe. Le linguiste peut alors observer, au delà de la variété des choix de traduction locaux, s'il existe des régularités fortes entre deux langues, dans le passage d'une construction syntaxique à une autre comme dans les équivalences lexicales, phraséologiques ou terminologiques.

– *TA et TAO :* En ce qui concerne la famille des systèmes de Traduction Automatique Basée sur l'Exemple (cf. [NAG 84] et [SAT 90]), la constitution d'un bi-texte est la première étape de la construction d'une *mémoire de traduction*. Mais son application la plus immédiate, et pour le moment la plus réaliste, se situe véritablement dans le domaine de la TAO. Le bi-texte est en effet l'instrument idéal pour mettre en œuvre ce que Bar-Hillel [BAR 64] appelait " a judicious and modest use of mechanical aids ".

[1] A l'instar de Mel'cuk *et al*. [MEL 95] nous employons le terme de *lexie* au sens général d'unité lexicale, incluant les formes simples (*lexèmes*), les mots composés et les locutions (*phrasèmes*).

– *Le bi-texte comme Mémoire d'entreprise :* Isabelle [ISA 92] constate que " la masse des traductions produites chaque année contient infiniment plus de solutions à plus de problèmes que tous les outils de référence existants et imaginables ! ". Autrement dit, il est parfois plus intéressant pour un traducteur de savoir chercher ses informations dans un corpus de traduction adapté à son sujet que d'utiliser un dictionnaire souvent trop généraliste. La possibilité d'établir des " concordanciers à usages divers " permettant d'extraire des occurrences au sein de leur contexte immédiat est la condition *sine qua non* pour faire une recherche efficace. Macklovitch [MAC 92] montre comment, en partant d'une base de données contenant des extraits du Journal des débats de la Chambre des communes du Canada, un tel concordancier peut donner tout un éventail de solutions pour la traduction d'une expression argotique telle que " to be out of lunch ", sans véritable équivalent en français[2]. Comme le suggère [MAC 92], " en alignant automatiquement les textes sources et leur traduction, nous pouvons transformer les archives d'un service de traduction en une mémoire d'entreprise interactive qui, à la différence de nos collègues humains, n'oublie jamais ".

– *Vérification des traductions :* D'autre part, l'alignement d'un texte et de sa traduction permet d'appliquer des outils de vérification automatique signalant au réviseur toutes les éventuelles anomalies. Comme le note Isabelle[ISA 92], les meilleurs traducteurs ne sont pas à l'abri d' " erreurs " élémentaires comme l'emploi de faux amis tels que *library* (angl.) pour *librairie* (fr.), *physician* (angl.) pour *physicien* (fr.): " Even though these translations are the work of some of the best translators in Canada, it appears that the time pressure under which they are produced makes linguistic interference harder to control. ". Pour pallier ces inconvénients relativement superficiels le système Transcheck [ISA 93] détecte la présence de faux amis, d'emprunts et de calques jugés illicites. En outre, les portions de texte " oubliées " par le traducteur sont automatiquement repérées par le système d'alignement.

Etude empirique

L'utilité des corpus bi-textuels est fonction de leur volume : pour remplir leur fonction de base d'exemples et permettre des traitements statistiques, ils doivent être d'une taille importante. C'est pourquoi on s'intéresse aux techniques d'*alignement* qui visent à automatiser la production de bi-textes, en appariant les portions de texte qui sont traductions les unes des autres. Nous examinons ici 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 consécutifs en commun (on écrira : 4-grammes), généralement en début de mots. Entre des langues apparentées,

[2] L'auteur insiste sur la variété des solutions tirées de son corpus, telle que :" être tout à fait dans l'erreur ", " se fourvoyer " ou encore " être dans la choucroute ", etc.

comme l'anglais et le français, cette hypothèse a déjà été mise en œuvre et a fourni de bons résultats [SIM 92] [CHU 93].

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 sous plusieurs angles (cf. figure 1) :

– Dans une première phase (section 3), nous avons étudié le rappel et la précision empirique de l'identification des mots apparentés ou *cognats* (de l'anglais " *cognate* ") en recourant aux *n*-grammes, sur un corpus français-anglais. Pour cette évaluation, on compare les cognats candidats à une liste de couples de cognats de référence, établie manuellement.

– Puis nous avons cherché à améliorer les résultats de l'approximation des *n*-grammes en en corrigeant certaines insuffisances : nous exposons ainsi une autre méthode de comparaison superficielle des chaînes de caractères, par l'extraction d'une sous-chaîne maximale commune respectant des contraintes de longueur et de parallélisme.

– dans une seconde phase (section 4), nous nous intéressons à l'utilisation des cognats comme indices d'alignement. Nous discutons d'abord de la place de cet indice au sein de l'architecture globale d'un système d'alignement. En partant d'une heuristique simple, nous en tirons une méthode permettant d'obtenir un préalignement privilégiant l'exactitude au détriment de l'exhaustivité.

– nous avons enfin (section 5) cherché à corréler la qualité de l'identification des cognats avec l'évaluation du préalignement résultant, afin de déterminer les conséquences des améliorations que nous avons apportées à cette identification sur les résultats finaux. Par ailleurs, en vue de déterminer plus précisément l'applicabilité des méthodes exposées sur d'autres corpus, nous avons rapporté les résultats du préalignement à des caractéristiques formelles des textes parallèles, calculables en amont de l'alignement.

Pour la mise en œuvre de cette expérimentation, nous avons travaillé sur des textes issus de corpus bilingues alignés manuellement, le corpus *BAF*[3] (pour Bi-texte anglais français, cf. [SIM 98], initialement constitué au CITI[4], puis confié au RALI[5] de l'Université de Montréal, dans le cadre d'un financement de l'AUPELF-UREF[6].

[3] Ce corpus nous a été gracieusement fourni dans le cadre du projet ARCADE centré sur l'évaluation des méthodes d'alignement [LAN 98].

[4] CITI : Centre d'innovation en technologies de l'information (Laval, Canada).

[5] RALI : Laboratoire de recherche appliquée en linguistique informatique. URL : http://www-rali.iro.umontreal.ca

[6] AUPELF : Association des universités partiellement ou entièrement de langue française, UREF : Université des réseaux d'expression française. URL : http://www.aupelf-uref.org/

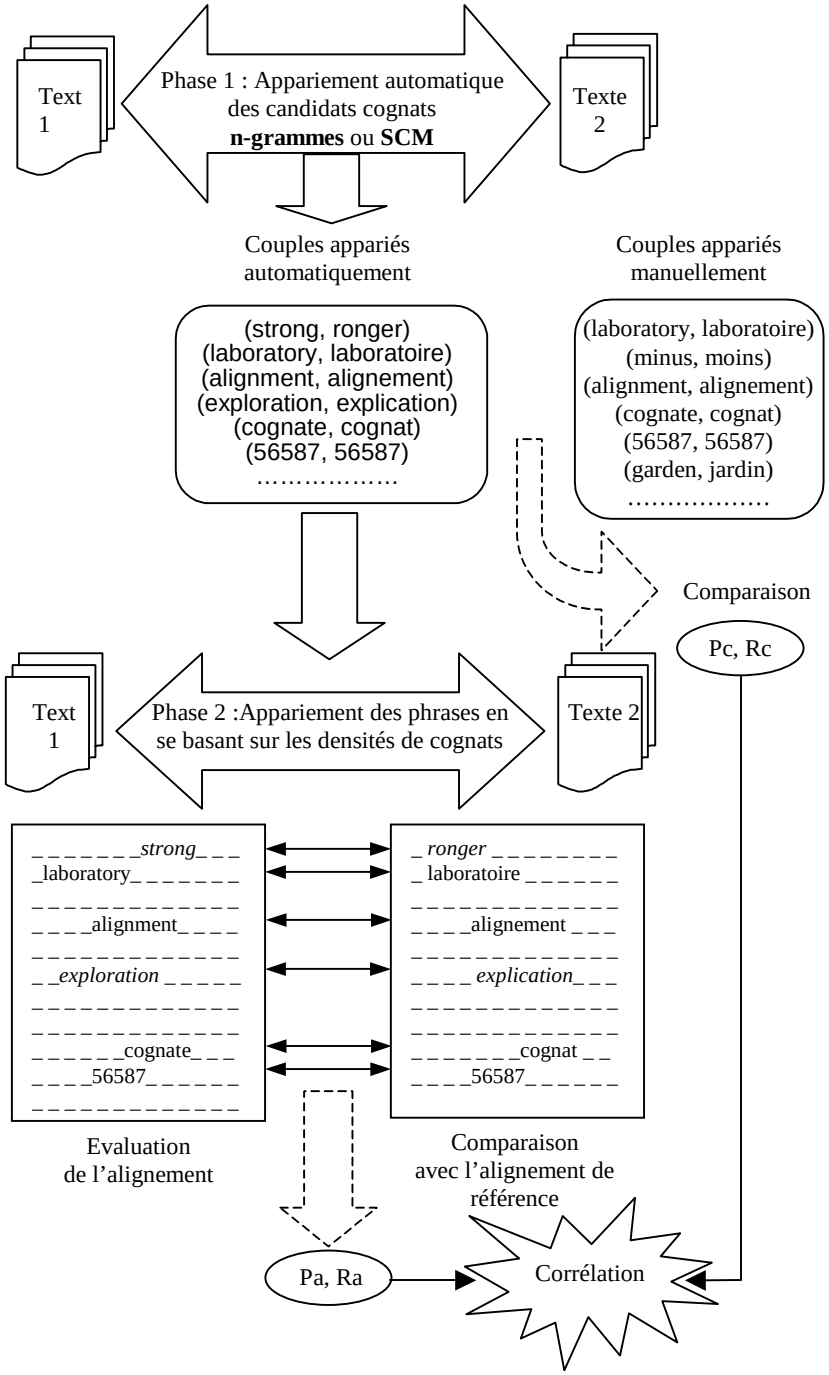


Figure 1. Organisation des phases de l'évaluation.

Ce corpus est constitué de textes scientifiques (fr. 3 134 + angl. 3 118 phr.), institutionnels (fr. 14 725 + angl. 14 460 phr.) et littéraires (fr. 3 871 + angl. 3 766 phr.). Nous avons délibérément écarté une quatrième composante du corpus, une documentation technique de *Xerox*, qui ne satisfait pas à nos hypothèses de parallélisme (cf. section 2), car constituée d'un index alphabétique ordonné différemment dans chaque langue.

2. Principe de l'alignement

Comme le remarquent Simard et Plamondon [SIM 96], il existe deux types d'alignement : l'extraction de points d'ancrage (" bi-text mapping "), qui dégage une suite de points *(i,j)* représentant des emplacements respectifs des deux textes censés se correspondre dans la traduction mutuelle ; l'alignement proprement dit, qui désigne la production effective d'un bi-texte, avec segmentation des deux textes et appariement des segments correspondants.

2.1. Points d'ancrage et alignement

Dans le premier cas, aucune segmentation préalable n'est nécessaire, les coordonnées des points pouvant être exprimées en numéros de caractère. Mais ce genre d'alignement n'est pas complet : il permet de déterminer *grosso modo* par où passe l'alignement, mais ne permet pas de façon directe de mettre en correspondance des unités cohérentes suivant le principe de compositionnalité.

Dans le second cas l'alignement produit un résultat final directement exploitable comme bi-texte. Mais cette fois, on s'appuie nécessairement sur une segmentation préalable : généralement en paragraphes, en phrases ou groupes de phrases.

Notons que la segmentation en phrases pose des problèmes épineux, et la qualité du bi-texte résultant en dépend [SIM 96]. D'une part, la définition linguistique de la phrase est au plus haut point équivoque : la phrase peut être définie selon des critères prosodiques, logiques (proposition), morphosyntaxiques (p. ex. autour de la rection verbale), ou encore selon des conventions typographiques (ponctuation, majuscules). Même dans ce dernier cas il n'est pas toujours facile de déterminer manuellement les limites d'une phrase : les phénomènes d'imbrication (parenthèses, tirets longs) posent problème, ainsi que les énumérations rapportées à un même noyau verbal, fréquentes en langue juridique.

En outre, les algorithmes de segmentation automatique ajoutent encore du bruit à ces indéterminations, car la ponctuation est ambiguë : la suite point-espace-majuscule apparaît aussi bien en fin de phrase que dans des abréviations.

Malgré ces difficultés, la phrase demeure un étalon privilégié et la plupart des systèmes d'alignement travaillent à ce niveau : on s'en sert comme d'une unité

opératoire, et la donnée d'un certain nombre de critères superficiels permet aux annotateurs humains de parvenir à un consensus raisonnable, ainsi qu'aux systèmes de segmentation automatique d'atteindre une précision suffisante [SIM 98]. Par exemple, pour la constitution du corpus *BAF*, Simard [SIM 98] énumère les critères de segmentation suivants : " Une phrase est une séquence syntaxiquement autonome de mots, terminée par un point final. (…) Les titres sont des phrases. (…) Les énumérateurs [par exemple III, N.B. etc…] sont des phrases. (…) Les éléments d'une énumération sont des phrases (…) Chaque case d'un tableau est une phrase. " (nous traduisons). Dans la suite de notre évaluation, nous nous situons dans le cadre de l'alignement phrastique, bien que les méthodes présentées ne dépendent pas spécifiquement de tel ou tel niveau de granularité.

Une fois la segmentation des deux textes obtenue, l'alignement consiste à déterminer *A*, un ensemble d'appariements entre les segments[7] de $S(T_1)$ et $S(T_2)$:

$A \subset S(T_1) \times S(T_2)$

Le calcul automatique de *A*, i.e. l'alignement, n'est pas une tâche triviale, car la traduction ne conserve pas rigoureusement la segmentation. On peut ainsi rencontrer différents cas de figure :

1. Conservation :		1:1		
2. Insertion :	simple :	0:1	multiple :	0:2, 0:3, etc.
3. Suppression :	simple :	1:0	multiple :	2:0, 3:0, etc.
4. Fusion :	simple :	2:1	multiple :	3:1, 4:1, etc.
5. Scission :	simple :	1:2	multiple :	1:3, 1:4, etc.
6. Coalescence[8] :	simple :	2:2	multiple :	2:3, 3:2, 3:3, 2:4, 4:2, 3:4, etc.

Par exemple :

$S(T_1) = P_1\ P_2\ P_3\ P_4\ P_5\ P_6$
$S(T_2) = P'_1\ P'_2\ P'_3\ P'_4\ P'_5\ P'_6\ P'_7$
$A = \{(P_1;P'_1\ P'_2)(P_2;\ )(P_3;P'_3)(P_4\ P_5;P'_4)(P_6;P'_5\ P'_6\ P'_7)\}$

L'alignement A peut aussi se noter sous la forme d'une suite de transitions, ou *chemin d'alignement* : A=[(1:2), (1:0), (1:1), (2:1), (1:3)]

L'objectif des algorithmes d'alignement[9] est de déterminer de façon automatique un tel chemin.

---

[7] On se place dans le cadre d'un alignement complet, où chaque segment apparaît une fois et une seule, en autorisant les appariements avec la " phrase vide " pour les cas d'insertion ou de suppression.

[8] Par *coalescence*, nous désignons une fusion qui s'opère de part et d'autre sur plusieurs segments à la fois.

[9] L'emploi du terme d'*alignement*, généralement usité dans la littérature, peut paraître impropre, dans la mesure où tel que nous l'avons défini, l'alignement autorise les correspondances croisées. Dans le cas général, on peut lui préférer le terme d'*appariement*

2.2. Parallélisme

Pour que l'alignement soit envisageable, les deux traductions mutuelles doivent respecter les conditions du *parallélisme* énoncées par Langé et Gaussier [LAN 95] :

– *quasi-bijectivité* : toute phrase source a en général un correspondant dans le texte cible, et réciproquement.

– *quasi-monotonie*[10] : la séquence des phrases sources doit suivre, en général, la séquence des phrases cibles correspondantes.

Le premier critère est inhérent à la notion même de *compositionnalité* : s'il n'est pas possible de faire correspondre de façon pratiquement bi-univoque des sous-parties de T_1 avec des sous-parties de T_2, aligner n'a pas de sens.

Le second critère est lié à une limitation d'ordre pratique et méthodologique : d'une part, la plupart des traductions conservent l'ordre des unités textuelles (sauf dans certains cas, comme par exemple le corpus *Xerox* déjà cité) ; d'autre part, les stratégies d'alignement deviennent beaucoup plus complexes et lourdes à mettre en œuvre si l'on admet la possibilité de réagencement des unités. On préfère alors se limiter d'abord au cas plus simple du parallélisme, quitte à étudier dans un second temps les heuristiques permettant d'aligner des textes non strictement parallèles.

Enfin, tout l'intérêt des méthodes d'alignement réside dans cette notion floue de " quasi " : le quasi-parallélisme autorise des infractions locales au principe de bijection et de monotonie (que l'on observe dans toute traduction), des suppressions, insertions et interversions mineures étant permises. Ce sont ces infractions mêmes qui justifient l'élaboration d'algorithmes et de stratégies sophistiquées, et c'est la limitation de leur portée qui rend possible l'application de ces algorithmes. La notion de " quasi " n'a jamais été quantifiée, sans doute parce qu'elle dépend en grande partie de la robustesse des méthodes : tandis que les résultats de certaines se dégradent très rapidement, passé un certain seuil de non-respect du parallélisme, d'autres suivent un processus d'altération progressive beaucoup plus harmonieux (" *gracious degradation* ").

2.3. Indices et algorithmes

Il existe un certain nombre de techniques d'alignement, désormais devenues classiques. Elles diffèrent essentiellement par le type d'information utilisé :

– *Les longueurs de segment* [GAL 91] [BRO 91]. Ces méthodes s'appuient sur une constatation empirique très simple : les longueurs d'un segment et de sa

---

employé par Débili et Sammouda. Mais nous nous plaçons ici dans un cadre restreint où nous négligeons les interversions.

[10] Nous préférons le terme *monotonie*, introduit par Isabelle et Simard [ISA 96] dans leurs *Propositions*, au terme *synchronisation* utilisé par les auteurs.

traduction sont fortement corrélées. Gale et Church ont ainsi montré que le rapport des longueurs, exprimé en nombre de caractères, suit une distribution normale centrée sur le rapport moyen des longueurs, estimé globalement constant pour un même couple de langues. Cette modélisation permet alors d'évaluer la probabilité *a priori* de n'importe quel chemin d'alignement en fonction de la coïncidence des longueurs et de la probabilité présumée des transitions. Le logarithme de l'inverse de cette probabilité donne une mesure de distance pour chaque chemin différent, et l'on extrait le chemin le plus court (i.e. le plus probable) à l'aide d'un algorithme de type Viterbi (cf. annexe 3).

– *Des dictionnaires de traduction préétablis.* Par exemple, dans un couple de phrases, la proportion d'unités lexicales ne trouvant pas de contrepartie (en se basant sur un dictionnaire) dans la phrase correspondante permet d'exprimer un *résidu de traduction* [DAV 95], que l'on peut intégrer, de façon similaire aux longueurs de phrase, à un modèle probabiliste.

– *Les distributions lexicales.* Cette fois, on se base sur la distribution des unités lexicales pour les apparier : on fait l'hypothèse que des lexies qui sont traductions mutuelles doivent être distribuées de façon similaire de part et d'autre du corpus parallèle. En convertissant les occurrences en vecteurs [FUN 94], et en comparant ces vecteurs au moyen de mesures statistiques (information mutuelle, *t-score*), on peut établir des correspondances lexicales dont les appariements donnent un nuage de points dans le produit cartésien des deux textes. Un filtrage des points permettant de dégager les zones de forte densité aboutit à une suite de points d'ancrages situés sur le chemin d'alignement. Dans d'autres architectures, on déduit l'appariement des mots d'un alignement manuel des premières phrases [CHE 93], et l'on aligne le reste du texte à partir des paramètres initiaux. La consolidation mutuelle de l'appariement des mots et des phrases permet aussi de partir d'un préalignement grossier, d'en extraire des appariements lexicaux, et de recalculer un alignement des phrases plus fin en fonction des mots nouvellement appariés [KAY 93] [DEB 92]. Un processus itératif aboutit ainsi à la convergence vers un alignement de plus en plus fin avec l'extraction d'un glossaire de correspondances lexicales de plus en plus riche.

– *Les transfuges et cognats.* On a pensé très tôt [GAL 91] à utiliser des indices superficiels (découpage en paragraphes, titres) pour obtenir un préalignement grossier mais fiable. Les *transfuges* [GAU 95], c'est-à-dire les chaînes de caractères invariantes à la traduction (nombres, sigles, noms propres, etc.) font partie de cette famille d'indices. L'idée de s'appuyer sur des ressemblances formelles a conduit notamment à l'exploitation des *cognats*, ces mots apparentés qui portent encore dans leur graphie une ressemblance de surface. L'hypothèse de " cognation " (en anglais " *cognateness* ", cf. [SIM 92], 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 traductions l'une de l'autre, que si elles sont prises au hasard. Le calcul d'une distance basée sur la " cognation " à permis d'améliorer notablement les résultats des méthodes basées sur les longueurs (*ibid.*, [DAV 95], [LAN 97b]). D'autres en ont tiré des procédés fiables pour obtenir

des points d'ancrage et réduire l'espace de recherche ([CHU 93], [SIM 96], [LAN 97a], [MEL 96]).

Toutes les méthodes basées sur les cognats présentent un point commun : les couples de cognats sont identifiés par la prise en compte de *n*-grammes. Il s'agit la plupart du temps de couples de mots ayant 4 caractères consécutifs en communs (4-grammes), généralement situés en début de mot ([SIM 92], [SIM 96], [LAN 97b]). La section suivante est consacrée à l'étude empirique de cette approximation.

3. Identification des cognats

Un parti pris méthodologique oriente nos recherches : nous nous limitons ici au parallélisme en tant que propriété formelle d'un couple de textes qui sont traductions mutuelles, indépendamment de toute information linguistique complémentaire. Nous voulons déterminer jusqu'où cette hypothèse peut être vérifiée, et à quel moment elle cesse d'être pertinente pour la constitution d'un bi-texte. En effet, pour des raisons de coût de mise en œuvre, de robustesse et de généralité, les méthodes linguistiques étant plus spécifiquement liées à un couple de langues, nous préconisons d'y recourir dans un second temps, avec une pleine conscience des possibilités et des limites des systèmes « alinguistiques », qui n'utilisent pas de ressources telles que grammaires et dictionnaires.

Ainsi, pour la détermination des cognats, nous n'étudions pas la possibilité de tirer parti de systèmes de conversion graphémique *ad hoc*, permettant d'obtenir des équivalences (telles que : *administration* → *amministrazione* en italien), avec un minimum de règles (*-dm-* → *-mm-*, *-tion* → *-zione*). Des modules linguistiques de ce genre peuvent néanmoins s'articuler, de manière complémentaire, avec les méthodes présentées ici.

3.1. Détermination manuelle des cognats

Dans cette première phase de l'évaluation, nous nous sommes restreint (pour des raisons pratiques) au corpus *Cour*, une portion du *BAF* : il s'agit de textes institutionnels issus de la Cour suprême du Canada, représentant environ 31 000 mots en anglais et 33 000 en français.

La mise en place du protocole expérimental a nécessité une détermination manuelle des couples de cognats observables au sein de notre corpus : ces couples n'entrent pas en jeu dans l'alignement, mais dans l'évaluation, afin de déterminer la validité des cognats extraits automatiquement. Etant données les difficultés inhérentes à la notion de ressemblance, tant sur le plan du contenu que sur celui de l'expression, nous sommes parti d'une définition opératoire visant à en contourner l'aspect subjectif. Deux lexies L_1 et L_2 sont des cognats si et seulement si :

– L_1 et L_2 ont un lien étymologique (emprunt, origine commune) perceptible dans leur signifiant.

– On peut trouver deux phrases (P_1,P_2) dont l'une est la traduction de l'autre, et dans lesquelles L_1 et L_2 sont traductions mutuelles.

Les *transfuges*, invariants dans la traduction, peuvent être considérés comme des cognats particuliers. Le critère 1 ainsi que la notion de transfuge ne nous 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, pour le critère 2, décider de la possibilité de traduire une lexie par une autre implique des difficultés :

– D'une part, un lexème peut être traduit par un phrasème : par exemple *because* (angl.) ↔ *à cause* (fr.). Nous avons décidé de ne prendre en compte que des formes simples, en nous limitant à celles qui portent 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, rare dans la pratique de la traduction. Comme nous l'avons souligné, ailleurs, dans une précédente discussion sur la notion contestable " d'alignement lexical " [KRA 99b], il n'est pas possible d'étendre l'hypothèse de parallélisme (ni même de compositionnalité) au mot, à l'intérieur des phrases. Or, des mots d'étymologie commune mais de sens différents peuvent, dans un certain contexte, se retrouver en relation de traduction :

Par exemple, les mots *importation* (angl.) et *export* (fr.) peuvent être considérés comme des cognats bien que leur contenu sémantique présente une différence d'orientation. On peut leur imaginer le contexte de traduction suivant :

Il fait de l'export vers les USA ↔ He makes importations from France

En revanche, entre *translation* (angl.) et *transfert* (fr.), l'écart sémantique est plus grand. Et pourtant, ces deux unités sont apparentées (*translatum* est le participe passé latin de *transferre*), et il est possible de trouver des contextes à l'intérieur desquels ils sont en relation de traduction :

J'ai effectué un transfert du français vers l'anglais ↔ I did a translation from French to English

Ici *transfert* est employé en tant que terme *générique* pouvant subsumer le terme anglais plus précis. Ce genre de saut du générique au spécifique n'est pas rare lors de la traduction. Il faut en outre détailler la notion d'étymologie commune : tous les étymons doivent-ils entrer en ligne de compte ou bien seulement les radicaux ? Après tout, si deux lexies sont des équivalents possibles, le fait qu'elles soient apparentées par le biais de préfixes (ou de suffixes) n'enlève rien à leur fonction heuristique d'indice : nous les considérerons donc comme des cognats. Le problème se situe donc sur le plan sémantique : à partir de quel niveau d'écart sémantique doit-on rejeter un cognat ? Où situer le seuil sans tomber dans l'arbitraire ?

Ne pouvant donner de réponse satisfaisante à ces questions, nous avons contourné la difficulté par un parti pris restrictif : nous ne retenons comme couple de cognats de référence (pour l'évaluation) que les lexèmes qui sont effectivement traduits l'un par l'autre, dans notre corpus[11]. Ceci peut introduire un léger biais dans nos résultats, avec une précision parfois sous-évaluée. Par exemple, *appeal* (angl.) et *appelant* (fra.) n'apparaissent pas comme traductions mutuelles dans notre corpus, et ne sont donc pas retenus comme couple de cognats, alors que d'après notre définition ce sont bien des cognats. Par ailleurs, lorsque l'on confronte les candidats avec les couples de référence, toutes les chaînes homographes sont assimilées, qu'il s'agisse d'homonymes ou d'unités polysémiques utilisées dans des acceptions différentes. Cette assimilation peut conduire à accepter des couples d'unités non-équivalentes, comme *because* et *cause* (dans le sens de " *parle* "), et augmenter artificiellement la précision. Cependant, ces deux biais antagonistes n'affectent le calcul de la précision que de façon marginale et non significative, car celle-ci ne nous intéresse pas pour sa valeur absolue mais pour ses variations, qui sont statistiquement indépendantes de ces biais.

3.2. Evaluation des n*-grammes*

La prise en compte des chaînes de *n* caractères consécutifs communs, les *n*-grammes, permet de déterminer automatiquement un ensemble de couples de cognats candidats, en comparant les formes des deux textes. On évalue ensuite la pertinence de ces candidats en dénombrant la proportion de candidats erronés (comme *ronger* (fr.) et *strong* (angl.)), i.e. le bruit, et la proportion de couples de référence qui ont été ignorés, i.e. le silence.

Pour la quantification du silence et du bruit on utilise les trois mesures suivantes :

– La précision *Pc* exprime la proportion de couples de cognats corrects par rapport au nombre de candidats produits (complémentaire du bruit). En utilisant la notation entre deux barres pour exprimer le cardinal d'un ensemble, on écrit :

$$Pc = \frac{\left| CognatsCandidats \cap CognatsDeRéférence \right|}{\left| CognatsCandidats \right|}$$

– Le rappel *Rc* exprime la proportion de couples de cognats corrects par rapport nombre de couples cognats de référence (complémentaire du silence).

[11] En d'autres termes, nous avons extrait toutes les correspondances lexicales qui impliquaient des lexies apparentées. Ces correspondances ont été établies manuellement, sur la base de l'identité de *désignation* (au sens de [PER 93] des lexies. On a ainsi relevé une liste de 944 couples différents, sans compter les tranfuges.

$$Pc = \frac{|CognatsCandidats \cap CognatsDeRéférence|}{|CognatsDeRéférence|}$$

– La F-mesure est une valeur combinée[12] représentant la moyenne harmonique de P et R :

$$Fc = \frac{2 \times (P \times R)}{(P + R)}$$

Notons qu'il existe deux façons de calculer les valeurs de *Pc* et *Rc* : soit on compare tous les *mots-types*[13] de T_1 avec tous les mots-types de T_2 ; soit on compare tous les *mots-occurrences* de T_1 avec tous les mots-occurrences de T_2, un même couple pouvant intervenir plusieurs fois dans l'évaluation. Nous avons opté pour cette dernière solution, en limitant les comparaisons à l'ensemble des couples de phrases qui font partie de l'espace de recherche de l'algorithme d'alignement : les statistiques sont ainsi directement liées aux cognats réellement utilisés par cet algorithme[14]. Pour chaque comparaison de deux mots-occurrences (M_1,M_2), quatre cas de figure peuvent se présenter, que l'on dénombre séparément, comme le montre le tableau 1 :

a+b+c+d = nombre total de (M_1,M_2) comparés	*(M_1,M_2) de la liste des Cognats de réf.*	*(M_1,M_2) hors de la liste des Cognats de réf.*
(M_1,M_2) retenus comme candidats	a	b
Couples (M_1,M_2) non retenus comme candidats	c	d

Tableau 1. *Dénombrement des cas de figure pour les couples de mots comparés*

On a donc : $Pc = a/(a+b)$ et $Pc = a/(a+c)$

On obtient les données du tableau 10 de l'annexe 1. Nous y avons fait figurer, dans la première colonne, les valeurs de précision et rappel liées aux transfuges,

[12] F a la propriété de se rapprocher de la moyenne si P et R sont proches, et de décroître si P et R sont éloignés.

[13] Par *mots-types* on entend l'ensemble des mots apparaissant dans chaque texte, chaque mot étant compté une fois indépendamment de sa fréquence. Par *mots-occurrences* on désigne toutes les occurrences particulières des *mots-types*.

[14] On verra que ces statistiques résultent de la comparaison de phrases voisines, à l'intérieur des sections préalignées automatiquement. Entre des phrases quelconques on peut supposer qu'elles seraient différentes (pour des raisons de continuité thématique), avec une précision et un rappel inférieurs.

comme base de comparaison. Notons qu'un certain nombre de ces transfuges sont également filtrés comme *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 pour la détermination des cognats. Malheureusement les indices qui génèrent le moins de bruit sont aussi les plus rares : comme le montre la figure 2, lorsque *n* augmente le rappel chute aussi rapidement que la précision croît.

On constate que la prise en compte des transfuges d'au moins 3 caractères[15] donne une meilleure F-mesure qu'avec les *n*-grammes (la précision est de 100 % car nous avons négligé toutes les erreurs dues à l'homographie comme (angl.) *case* et (fr.) *case*). 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 couples cognats à identifier. On peut sans doute améliorer les résultats globaux en combinant l'identité (les transfuges) et la ressemblance (les *n*-grammes ou autres). C'est ce que montre la troisième colonne du tableau 10 de l'annexe 1.

3.3. Sous-chaînes maximales

L'identification des cognats par les *n*-grammes appelle deux remarques :

– 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 *doc**tor*** (angl.) et *dot**tor**e* (it.) n'ont au plus que 3 caractères consécutifs communs.

– D'autre part, la signification d'un n-gramme dépend étroitement de la taille des mots comparés. Il est clair que 4 caractères consécutifs communs entre ***form*** (angl.) et ***form**e* (fr.) sont plus significatifs qu'un 6-grammes entre *explo**ration*** (angl.) et *décla**ration*** (fr.).

Pour pallier le premier inconvénient, nous proposons de recourir aux sous-chaînes maximales (on notera *SCM*), à l'instar de Debili et Sammouda [DEB 92] : on s'intéresse à la taille de la plus longue sous-chaîne de caractères commune aux deux mots en autorisant les sauts.

Ainsi, pour *doctor* et *dottore*, la *SCM* est de longueur 5 : **d-o-t-o-r**. Mais la combinatoire des *SCM* est très importante (surtout avec les mots longs), et risque de produire beaucoup de bruit : par exemple *pragmatic* (angl.) est presque totalement inclus dans *paradigmatique* (fr.). 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 plusieurs décrochements du parallélisme (insertion ou suppression) en série, et les décrochements ne sont tolérés que lorsqu'ils sont encadrés de caractères identiques.

[15] Pour les transfuges comptant 1 ou 2 caractères, nous n'avons tenu compte que des nombres.

Ainsi, **p-r-a-g-m-a-t-i** n'est pas une sous-chaîne quasi-parallèle de *pragmatic* et *paradigmatique*, car les caractères **d-i** représentent deux décrochements consécutifs.

Enfin, pour limiter le bruit et tenir compte de la remarque 2, nous rapporterons la longueur des *SCM* à la taille des mots. En calculant le rapport entre la taille du plus long mot du couple et la longueur de la *SCM*. Pour deux lexies L_1 et L_2 on calcule :

$$r(L_1,L_2)=\frac{l(SCM\,(L_1,L_2))}{\max(\,l(L_1),l(L_2))}$$

on peut ensuite filtrer les candidats en fonction de r. Dans notre expérimentation nous avons testé différentes valeurs pour ce seuil : les meilleurs résultats ont été obtenus en acceptant les *SCM* avec r ≥ 2/3.

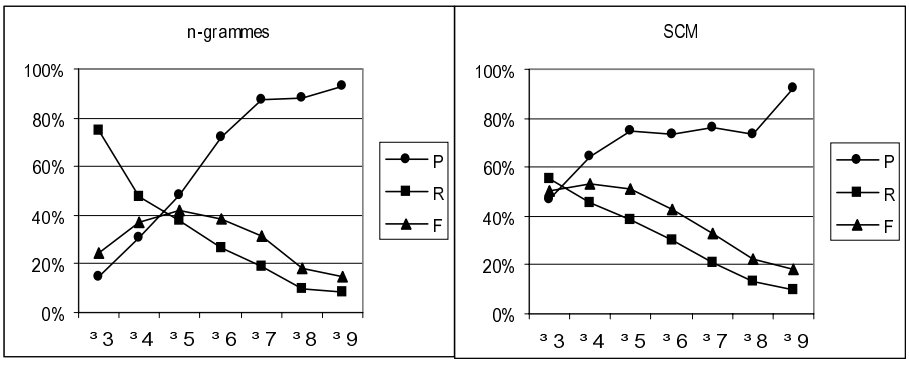


Figure 2. *Pc Rc Fc pour les n-grammes et les SCM*

La figure 2 montre les résultats liés aux *SCM* en fonction de leur longueur. On constate que l'évolution est parallèle à celle des *n*-grammes, avec cependant un niveau de précision supérieur pour un rappel similaire dès $n \geq 4$. La valeur maximum de F_c est de 53 %, soit 11 points de plus qu'avec les *n*-grammes.

4. Architecture du système d'alignement

Il reste à déterminer si la légère progression observée en fin de section précédente entraîne des améliorations significatives de l'alignement. Nous décrivons ci-dessous un algorithme basé sur les cognats et destiné à fournir d'abord un préalignement, c'est-à-dire une suite de points d'ancrage très sûrs pour établir des îlots de confiance et réduire l'espace de recherche des algorithmes plus coûteux.

4.1 Principe heuristique

Les développements récents ([DAV 95], [SIM 96], [LAN 97a]) ont montré que les meilleurs résultats pour l'alignement sont obtenus par des combinaisons d'indices, où l'on essaye de tenir compte de toutes les informations disponibles (longueurs, cognats, distributions lexicales) pour gagner en robustesse. Dès lors, un aspect déterminant de l'architecture d'un système, tant sur le plan de la complexité des calculs que sur celui des résultats, est l'*ordre* dans lequel on intègre ces indices. Pour mieux circonscrire cet aspect, nous proposons une heuristique très simple, le principe de *précision d'abord* :

L'alignement final pouvant s'obtenir de façon itérative, par l'appariement de segments de plus en plus petits, il convient d'exploiter les indices par ordre de précision décroissante. Les indices les plus sûrs sont donc utilisés en priorité.

La délimitation des possibilités d'erreur et la réduction de l'espace de recherche sont les principaux atouts d'une telle démarche. Ce principe rejoint les considérations de Simard et Plamondon [SIM 96] dans leur recherche d'équilibre entre *robustesse* (" *robustness* ") et *résolution* (" *accuracy* ") de l'alignement, lorsqu'ils préconisent un alignement en deux temps : d'abord la délimitation de l'espace de recherche par la donnée de points d'ancrage extraits à partir de critères superficiels comme les 4-grammes ; ensuite la mise en œuvre de " l'artillerie lourde " basée sur une architecture dynamique et des modèles stochastiques sophistiqués.

En appliquant le principe de précision d'abord, nous avons démontré dans des travaux précédents [KRA 99a] qu'il était profitable de se baser, dès les premières étapes du préalignement, sur l'exploitation des cognats et des transfuges.

4.2 Première étape : exploitation des transfuges

Dans une première étape, on se base sur l'exploitation des transfuges seuls, plus fiables que les cognats car produisant moins de bruit. On met en œuvre un processus itératif en deux temps :

1. Equi-occurrence : 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'ancrage candidats, sous la forme de couples *(i,j)* exprimant les coordonnées des phrases contenant les transfuges appariés.

2. Filtrage des points d'ancrage candidats, selon les critères suivants, qui traduisent l'hypothèse de parallélisme :

- *diagonalité* : on ne retient que les points situés à l'intérieur d'un couloir centré sur la diagonale de l'espace à aligner[16].

- *continuité* : suppression des points présentant une déviation forte par rapport aux points précédents[17].

- *monotonie* : suppression des points entrant en conflit sur l'une de leur coordonnées, ainsi que des points croisés (*(i,j)* et *(i',j')* se croisent si *i>i'* et *j<j'*). On ne considère ici que les possibilités d'alignement monotone.

Pour maximiser la précision des résultats, 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 des îlots de confiance dégagés. En fait, l'application du principe de précision d'abord nous commande de hiérarchiser les transfuges en fonction de leur nature, car il semble qu'ils ne soient pas tous aussi fiables. Ainsi, en distinguant entre les alphanumériques (contenant au moins un chiffre), les formes avec une majuscule et les transfuges quelconques, nous avons obtenu les résultats[18] suivants sur l'ensemble du *BAF* (hors texte *Xerox*) :

	P_a	R_a	F_a
Alphanumériques	99,9 %	5,1 %	9,3 %
Majuscules	99,7 %	17,1 %	28,2 %
Transfuges quelconques	99,5 %	53,0 %	67,7 %

[16] Si *(X0,Y0)* et *(X1,Y1)* sont respectivement les coordonnées du premier et du dernier point de l'espace de recherche, on peut calculer ainsi la distance de *(x,y)* à la diagonale : *d(x,y)=|(x-X0)/(X1-X0)-(y-Y0)/(Y1-Y0)|*. Le couloir est défini comme l'ensemble des points vérifiant : *d(x,y)<seuil*. Nous avons pris un seuil égal à 0.2, aboutissant à un couloir occupant une surface égale à environ 36 % de l'espace de recherche.

[17] On calcule la déviation entre deux points *(Xi,Yi)* et *(Xj,Yj)* comme le *log* du coefficient directeur de la droite passant par ces points, rapporté à *l1* et *l2*, tailles respectives de *T1* et *T2* : *dév=log(((Xj-Xi)*l2)/((Yj-Yi)*l1))*. On ne retient que les points vérifiant : *-log 2 < dév < log 2*.

[18] Les mesures de précision, rappel, et F-mesure sont inspirées de la métrique *KR-mot* développée dans le projet ARCADE [LAN 98]. Pour évaluer un préalignement, constitué de points *(i,j)* désignant des coordonnées de phrases, on considère simplement le préalignement comme un alignement fragmentaire, constitué de l'appariement des couples de phrases *(i,j)*. C'est pourquoi le rappel est nécessairement faible.

Tableau 2. *Résultats des préalignements issus de différents types de transfuge*[19]

La multiplication des contraintes de filtrage des points (diagonalité, continuité, monotonie et surdétermination) implique que plus les points sont denses, plus il est improbable qu'un point erroné persiste. Or, comme le montre le tableau 2, ce n'est pas le cas : nous faisons l'hypothèse que ce phénomène est dû à la plus ou moins grande fiabilité des différentes classes de transfuges (que l'on pourrait classer par ordre de fiabilité : *alphanumériques* > *majuscules* > *quelconques*).

Dès lors, si l'on applique les trois indices successivement dans l'ordre de la précision décroissante, en reprenant à chaque étape les points donnés par l'étape précédente, on obtient les résultats suivants, avec un gain de près de 4 points de F-mesure.

	P_a	R_a	F_a
Alphanumériques + majuscules + transfuges quelconques	99,5 %	57,1 %	71,4 %

Tableau 3. *Résultats après l'utilisation successive des différents types de transfuge*

L'heuristique de précision d'abord semble donc porter ces fruits dès cette étape initiale. Notons que la complexité en temps de cet algorithme est presque linéaire, puisque en $O(n \log(n))$, où n représente la taille du bi-texte[20].

4.3 Deuxième étape : exploitation des cognats

Dans un premier temps, nous avons effectué des observations empiriques sur deux textes tirés d'un rapport de la Commission européenne[21]. Notre objectif était d'étudier les propriétés formelles d'un bi-texte sur le plan de la densité des cognats observés. Ce bi-texte est constitué de 490 x 490 segments (phrases ou groupes de phrases) alignés manuellement par nous. Pour des raisons pratiques (limitation d'espace mémoire), nous avons regroupé ces segments en 49 blocs de 10 segments. Un comptage des 4-grammes relevés entre chaque bloc anglais/français nous a permis d'extraire une matrice des densités. Pour les deux textes T_1 et T_2, on note : $T_1 = (P_1 P_2 \ldots P_n)$, $T_2 = (P'_1 P'_2 \ldots P'_m)$, n_{ij} le nombre de couples de mots comportant au moins un 4-grammes entre P_i et P'_j Si l_i et l_j sont les longueurs respectives (en

[19] Pour éviter les ambiguïtés avec *Pc*, *Rc* et *Fc*, on notera *Pa*, *Ra* et *Fa* les valeurs de précision, rappel et F-mesure liées à l'alignement.

[20] Le terme *log(n)* est dû aux recherches dans nos index, sous forme d'arbres binaires.

[21] *http://www.europarl.eu.int.* La référence du rapport est *A4-0391/96*.

nombre de mots) des phrases P_i et P_j, le produit $l_i l_j$ exprime le nombre de couples de mots comparés entre P_i et P_j. Par conséquent, la densité de 4-grammes entre *i* et *j* s'écrit : $d_{ij} = n_{ij}/(l_i l_j)$.

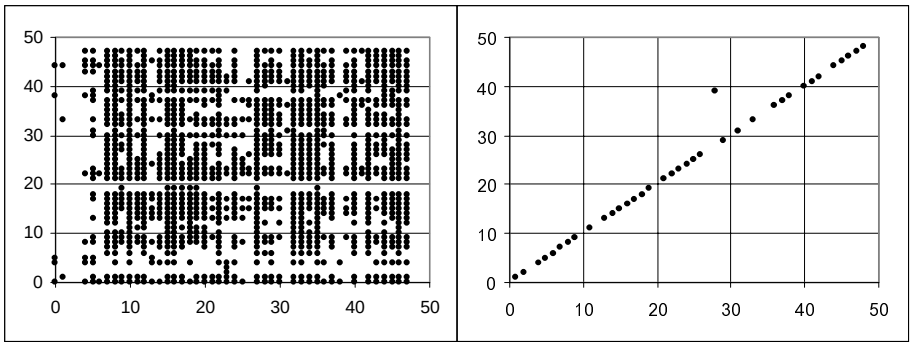
On obtient une première matrice $D = (d_{ij})_{i=1..n,j=1..m}$ contenant, pour chaque couple de phrases P_i et P'_j la densité de cognats associée. Si l'on représente sur un graphique l'ensemble des points dont la densité est supérieure à la moyenne, on aboutit à la figure 3(a). Les lignes verticales et horizontales indiquent un effet marginal important : certains segments sont très productifs en 4-grammes communs, et " pèsent " donc plus lourd que d'autres. Ceci peut s'expliquer par la présence de groupes de lettres récurrents entre l'anglais et le français : par exemple *t-i-o-n*, ou *m-e-n-t*. La densité ne peut donc être utilisée telle quelle comme mesure d'association entre les lignes et les colonnes. Pour neutraliser le bruit engendré par ces chaînes récurrentes, nous proposons d'utiliser la mesure du lien c_{ij} exprimant " la contribution de f_{ij} à l'information apportée par la matrice des fréquences " [VOL 97] :

$$c_{ij} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} \quad n = \sum_i \sum_j n_{ij} \quad n_i = \sum_j n_{ij} \quad n_j = \sum_i n_{ij} \quad f_{ij} = \frac{n_{ij}}{n} \quad f_i = \frac{n_i}{n} \quad f_j = \frac{n_j}{n}$$

On obtient alors la matrice d'association $C=(c_{ij})_{i=1..n,\ j=1..m}$

Si l'on applique une contrainte de réciprocité, en retenant tous les points *(i,j)* tels que P_i atteint son association maximum avec P'_j, et P'_j atteint son association maximum avec P_i, i.e. *(i,j) tels que* $i = argmax_{k=1..n}(c_{kj})$ *et* $j = argmax_{k=1..m}(c_{ik})$, on obtient la figure 3(b) (en cas de conflit entre deux points sur une même ligne ou une même colonne, on ne retient aucun des deux). Il est notable que cette fois, les points retenus (sauf un) correspondent à l'alignement réel des blocs.

En appliquant par la suite les trois critères de filtrage traduisant leu parallélisme - *diagonalité*, *continuité*, et *monotonie* - on arrive à 36 points correctement alignés sur 49, soit une précision de 100 % pour un rappel de 73 %.



(a) (b)

Figure 3. (a) Couples de phrases excédant la densité moyenne ; (b) couples de phrases obtenant un lien maximum pour *i* et *j*.

4.3.1 Exploitation

Dans cette deuxième étape de notre algorithme, on examine tous les couples de phrases alignables à l'intérieur des îlots de confiance obtenus, i.e. tous les points situés dans un couloir autour de la diagonale de chaque section. On aurait pu opter pour un couloir de largeur proportionnelle à la taille de la section, comme dans l'étape 1 : mais vu l'importance du rappel précédemment obtenu, on se contente d'un couloir de largeur constante (10 phrases), afin de limiter les calculs. Comme le montre la figure 4, dès cette étape, l'espace de recherche peut être considérablement réduit, si le rappel consécutif à la première étape est suffisant (en l'occurrence, c'est le cas).

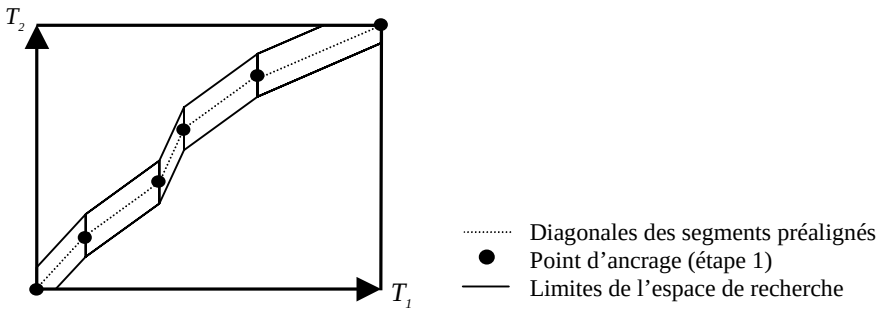


Figure 4. *Espace de recherche guidé par les points d'ancrage de l'étape 1.*

A l'intérieur de chaque îlot de confiance on comptabilise les cognats pour chaque couple de phrases alignables, et l'on extrait une matrice *C* comme précédemment. On applique alors la contrainte de réciprocité des maxima d'association pour obtenir une suite de points *(i,j)*. Les points obtenus sont ensuite filtrés par les critères de parallélisme : *continuité*, *monotonie* (la *diagonalité* a déjà été appliquée dans la définition de l'espace de recherche). Notons que chaque matrice est calculée entre des points fixés par la première étape, à condition qu'ils soient au moins distants de 10 phrases. En effet, si les points sont trop rapprochés, le calcul de la précédente matrice perd de sa pertinence. On ignore donc certains points issus de l'étape 1.

Sur le plan de la complexité des calculs, celle-ci est bornée par *O(n)*, car la largeur du couloir est constante quelle que soit la taille du texte. Pour le corpus

étudié, la constance de la largeur du couloir (10 phrases) est sans risque de " perte " du chemin d'alignement, car les îlots de confiance sont très petits (on obtient un rappel oscillant entre 40 % pour *Verne*, une traduction légèrement contractée de *De la terre à la lune*, de Jules Verne, et 91 % pour *TAO2*, bi-texte d'un article scientifique dans le domaine de la traduction assistée). Un corpus un peu moins riche en transfuges demanderait probablement un élargissement du couloir, avec une simple augmentation linéaire de la complexité. La largeur du couloir est donc liée au type du corpus, et non à sa taille totale.

4.3.2 Résultats

Pour l'identification des cognats, nous avons utilisé la seconde combinaison d'indices du tableau 11 en annexe 1 (transfuges + *n*-grammes pour $n \geq 3$ et $r > 2/3$, *SCM* pour $n \geq 5$ et $r > 2/3$), qui a donné les meilleurs résultats sur le corpus *Cour*. Appliquée à l'ensemble du corpus *BAF*, cette combinaison aboutit une F-mesure moyenne globale de 88,4 % (hors *Xerox*).

Dans le décompte des cognats, pour le calcul de (f_{ij}), nous avons cherché à pondérer les différentes classes de cognats candidats en fonction de leur fiabilité respective. Nous avons testé deux jeux de pondération : la première globalement proportionnelle à la précision de l'indice dans la détermination des cognats ; et la seconde, homogène à la première, tendant à corriger les irrégularités[22] de celle-ci en affectant des poids croissant avec la longueur des *SCM* :

Cas	trans-fuges	4grams l<7	*SCM* *n*=4	*SCM* *n*=5	*SCM* *n*=6	*SCM* *n*=7	*SCM* *n*=8	*SCM* *n*=9	*SCM* ≥10
Précision	100 %	48 %	15 %	76 %	62 %	58 %	33 %	83 %	100 %
Pondération 1	10	5	1	8	6	6	3	8	10
Pondération 2	10	6	2	6	7	8	8	9	10

Tableau 4 : *Précisions et pondérations des différentes catégories d'indice.*

Modèle	*Pa*		*Ra*		*Fa*	
	Moy.	Min.	Moy.	Min.	Moy.	Min.
Sans pondération	99,4 %	97,9 %	80,0 %	60,5 %	88,4 %	74,8 %
Pondération 1	99,6 %	98,6 %	84,5 %	62,0 %	91,2 %	76,1 %
Pondération 2	99,5 %	97,2 %	83,6 %	59,7 %	90,6 %	74,0 %

Tableau 5 : *Résultats des différentes pondérations.*

[22] Etrangement, on constate que les sous-chaînes de longueur 8 sont peu fiables, et obtiennent une précision plus faible que les sous-chaînes de longueur 4.

Les résultats avec et sans pondération sont inscrits dans le tableau 5. On constate une légère amélioration avec le recours aux pondérations, le premier jeu étant légèrement meilleur sur ce corpus. Notons que tous les minima sont atteints pour le corpus *Verne*. La précision se maintient malgré tout au delà de 97 %, ce qui montre la robustesse de ce genre de méthode.

4.4 Troisième étape : alignement final

Un algorithme de programmation dynamique peut désormais être appliqué entre les points d'ancrage afin de produire un alignement complet, en imposant au chemin de passer à une distance inférieure à trois phrases de ceux-ci. Notons que la densité importante des points d'ancrage a permis de borner la largeur du couloir délimitant l'espace de recherche, pour réduire les calculs en $O(n)$ (au lieu de $O(n^2)$).

Dans la mise en œuvre de l'algorithme de Viterbi (cf. détail en annexe 3), nous considérons 8 transitions possibles, correspondant aux transitions les plus fréquentes dans notre corpus. Pour l'estimation des probabilités de transition *a priori*, nous avons repris les valeurs des six transitions de la méthode de Gale et Church (cf. deuxième ligne du tableau 6). En partant de l'alignement de référence du corpus Cour, nous avons estimé les probabilités de T7 et T8 à environ 1/10 des probabilités des transitions T2 et T3. En conservant les mêmes proportions que dans l'estimation de Gale et Church et en normalisant (pour obtenir une somme égale à 1), on obtient les valeurs estimées de la troisième ligne du tableau 6.

	T1 (1:1)	T2 (2:1)	T3 (1:2)	T4 (0:1)	T5 (1:0)	T6 (2:2)	T7 (3:1)	T8 (1:3)
GC	0,89	0,0445	0,0445	0,00495	0,00495	0,011		
GC étendue	0,883	0,0442	0,0442	0,0049	0,0049	0,01	0,0044	0,0044

Tableau 6. *Probabilités a priori des transitions.*

Nous noterons *DGC* la mesure de distance basée sur la méthode de Gale et Church (pour le détail de cette distance, voir en annexe 2).

4.4.1 Distance basée sur la densité de cognats

Comme le suggèrent Dunning *et al*. [DUN 95], il est possible d'élaborer une distance basée sur la densité des cognats. Nous reprenons la précédente mesure de la densité, en tenant compte des pondérations précédemment définies. Si n_{ij} représente la somme *pondérée* (cf. pondération 2) des cognats observés entre P_i et P_j, on a :

$$d_{ij} = \frac{n_{ij}}{l_i \, l_j}$$

En calculant cette densité pour tous les couples alignés, d'une part, et pour des phrases quelconques, d'autre part, on détermine les probabilités empiriques d'obtenir d_{ij} inférieure à une valeur donnée. Nous avons effectué ces relevés sur les corpus *Cour* et *Verne*, dont les densités de cognats moyennes (par couple de phrases comparées) sont assez différentes : respectivement $d_c = 0{,}093$ et $d_c = 0{,}057$. On observe les distributions suivantes (à un facteur multiplicatif près : c'est d_{ij} x100 qui est représentée) sur la figure 5 :

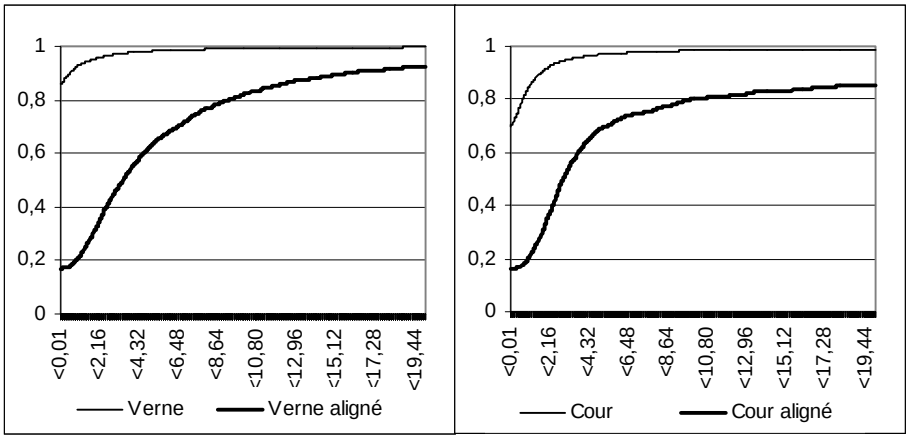


Figure 5 : *$p(d_c \leq x)$ pour les couples alignés et quelconques.*

Pour chaque corpus, on peut mesurer la valeur discriminante de notre indice à l'importante différence de surface entre les deux : la courbe en gras correspondant aux couples alignés, et l'autre à des couples de phrases quelconques. Par exemple, la probabilité d'obtenir $d_c<0{,}01$ est $p_a = 0{,}2$ pour deux phrases alignées et $p = 0{,}85$ entre deux phrases quelconques.

On constate en outre que les structures de ces distributions, entre *Verne* et *Cour*, sont similaires. Ceci nous permet de généraliser ces distributions empiriques et de nous en servir d'estimation *a priori* pour l'ensemble du corpus. Nous avons choisi d'utiliser les distributions issues de *Cour* correspondant à une traduction plus proche du texte de départ[23].

[23] On pourrait ici déceler un raisonnement circulaire : on utilise des données issues d'un corpus aligné manuellement pour en tirer par la suite un alignement automatique. Nous parlons cependant d'estimation *a priori* car nous faisons l'hypothèse que la courbe obtenue empiriquement peut être employée pour n'importe quel corpus, et c'est précisément ce que nous faisons : il faut noter que le corpus *Cour* ne représente que 8 % de la masse totale des corpus que nous traitons. Un éventuel biais ne peut donc concerner que les résultats de *Cour* : or, si l'on néglige ceux-ci, nos résultats globaux restent pratiquement inchangés.

Si l'on considère la probabilité d'aligner deux groupes de phrases (0,1,2 ou 3 phrases) dans un binôme *B* connaissant la densité de cognats d_c, on a :

$$P(B/d_c)=\frac{p(d_c/B)p(B)}{p(d_c)}$$

d'où l'on déduit la mesure de distance suivante, en assimilant *p(B)* à la probabilité *a priori* des transitions (à l'instar de [GAL 91], cf. annexe 2) :

$$\text{Dcognat1}=-\log P(B/d_c)=-(\log p(d_c/B)-\log p(d_c)+\log p(\textit{transition}))$$

où $p(d_c / B)$ et $p(d_c)$ correspondent aux deux distributions empiriques de la figure 5 (pour le corpus *Cour*).

En effectuant la même approximation que Gale et Church, on peut estimer que les variations de $p(d_C)$ sont négligeables. Ainsi, $-\log p(d_c)$ étant assimilé à une constante positive, on peut ignorer ce terme dans le calcul. On aboutit alors à une distance simplifiée :

$$\text{Dcognat2}=-(\log p(d_c/B)+\log p(\textit{transition}))$$

4.4.2 Résultats

Nous avons étudié *Dcognat1* et *Dcognat2* combinées à la distance *DGC*, avec les transitions *T1-T8*. Différentes pondérations ont été évaluées, en appliquant la formule :

$$\textit{Distance} = (1- k_{co})\cdot \textit{DGC} + k_{co}\cdot \textit{Dcognat}$$

La figure 6 montre l'évolution la F-mesure globale pour les différentes valeurs de k_{co}. Il apparaît que la distance *Dcognat2* donne des résultats légèrement meilleurs.

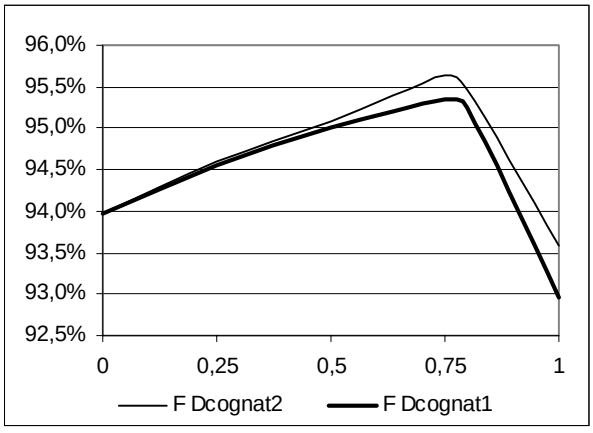


Figure 6. *Evolution de F en fonction de k_{co} (après lissage de la courbe établie d'après 5 points).*

L'interprétation de ce phénomène est délicate. *Dcognat2* étant le résultat d'une approximation supplémentaire, cette mesure aurait dû en toute logique fonctionner moins bien. Par rapport à *Dcognat1*, *Dcognat2* accorde une importance accrue au terme *p(d/cognat)*, qui n'est pas compensé par la soustraction de *p(d)*. Le troisième terme *p(transition)* y intervient donc dans une moindre mesure. Or, ce terme est lui-même issu d'une approximation générale, puisque les probabilités de transition sont estimées *a priori*, identiques pour tous les textes du corpus. On peut alors faire l'hypothèse que le bruit engendré par ces valeurs approchées est minimisé dans *Dcognat2*.

Par ailleurs, la structure des deux courbes montre une amélioration constante jusqu'à $k_{co} = 0{,}75$ suivie d'une chute assez rapide au delà. Par rapport aux deux points extrêmes, pour $k_{co} = 0$ et $k_{co} = 1$, les courbes se situent toujours au dessus (courbes convexes) : cela confirme le caractère additif des deux sources d'information. Chaque indice pallie les insuffisances de l'autre indice : lorsque les longueurs de phrase ne peuvent permettre de discriminer, les cognats sont susceptibles de prendre le relais, et réciproquement. L'hypothèse de [SIM 92] est donc confirmée : les deux indices ne se contrarient pas, mais fonctionnent en corrélation et de façon complémentaire.

Ainsi, les meilleurs résultats sont obtenus pour $k_{co} = 0{,}75$:

	P_a		R_a		F_a	
	Moy.	Min.	Moy.	Min.	Moy.	Min.
BAF\Xerox	96,5 %	70,6 %	94,8 %	72,7 %	95,6 %	71,7 %

Tableau 7. *Résultats pour k_{co}=0,75.*

Là encore les minima sont liés au corpus *Verne* : alors que pour toutes les autres parties du corpus l'alignement final aboutit à une augmentation notable de *F*, les résultats pour *Verne* sont moins bons que ceux du préalignement. Nous attribuons cette dégradation à l'inadéquation des paramètres, et notamment des probabilités de transition *a priori* (cette traduction comportant de nombreuses omissions). Cependant, même dans ce cas, on n'observe pas une dégradation catastrophique des

performances : les points d'ancrages issus du préalignement permettent de limiter les déviations trop fortes par rapport au chemin optimal.

5. Corrélations entre les résultats

5.1 Corrélation entre l'identification des cognats et l'alignement résultant

Nous avons cherché à corréler les résultats de la méthode d'identification des cognats avec les résultats de son exploitation, i.e. l'alignement. En fait, pour établir ces corrélations, nous n'avons pris en compte que les résultats du préalignement sur *Cour*, car l'alignement final ne dépend pas que des transfuges et cognats, mais d'un ensemble d'indices plus large (comme les probabilités de transition et les longueurs des phrases). Les résultats obtenus pour les différents types de paramétrages sont inscrits dans le tableau 11 de l'annexe 1, et sont représentés par les nuages de points de la figure 7.

On a en outre testé deux mesures combinant *n*-grammes et *SCM* (lignes 10 et 11 du tableau 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 :

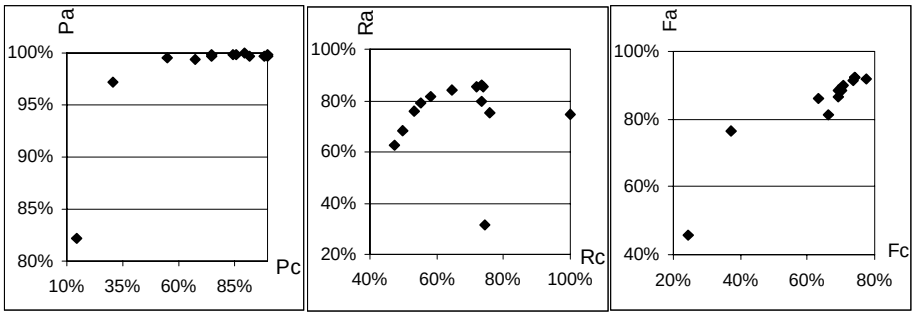
– La précision de notre méthode de préalignement est peu sensible au bruit : même pour une précision P_c de 15 %, la précision de l'alignement demeure au delà de 80 %. Cette robustesse est due à la multiplicité des contraintes de filtrage des points d'alignement.

– Le rappel du préalignement est fortement corrélé au résultat global de l'identification des cognats : entre F_c et R_a, la corrélation linéaire est de 0,94 (cf. tableau 12 de l'annexe 1). 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 le fait qu'il y ait trop de cognats (un rappel trop important) pourrait affecter le rappel de l'alignement. C'est sans doute lié à la nature de notre méthode : en effet la mesure du lien c_{li} a tendance à favoriser les appariements avec les phrases courtes. Dès lors, deux phrases contenant beaucoup de cognats 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. Quoi qu'il en soit, il est raisonnable de supposer que l'amélioration de F_c peut conduire à des résultats meilleurs pour F_a, moyennant une exploitation différente de la cognation.

5.2. Corrélations des résultats de l'alignement avec les caractéristiques du corpus

Dans la mesure où les méthodes développées ici ne dépendent que des caractéristiques formelles des textes parallèles, il paraît assez naturel de confronter les résultats à des paramètres quantitatifs synthétisant ces caractéristiques. Il serait intéressant en effet de pouvoir préjuger de l'applicabilité des méthodes sur n'importe quel corpus, voire d'en prédire grossièrement les résultats *a priori*, à partir de caractères significatifs facilement calculables.



P_a *en fonction de* P_c R_a *en fonction de* R_c F_a *en fonction de* F_c

Figure 7. *Résultats de l'alignement en fonction de l'identification des cognats.*

5.2.1 Corrélations des résultats de l'alignement avec la densité a priori des transfuges

Ainsi, dans un premier temps, nous avons cherché à corréler les résultats obtenus avec la densité des transfuges. Nous avons calculé la densité *Dtransfuge* comme le nombre moyen de transfuges *a priori* par phrase : pour chaque mot apparaissant à l'identique dans les deux textes, et comptant respectivement *nbOcc1* et *nbOcc2* occurrences dans T_1 et dans T_2, on somme *min(nbOcc1, nbOcc2)*, et l'on divise le total par la moyenne des tailles de T_1 et T_2 en nombre de phrases :

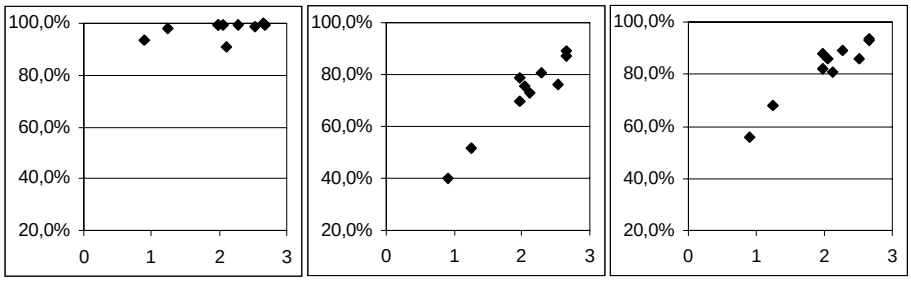
$$Dtransfuge = \frac{\sum_{mot1=mot2} \min(nbOcc(mot1), nbOcc(mot2))}{moyenne(nbPhrases(T_1), nbPhrases(T_2))}$$

Pour donner un ordre d'idée, cette densité oscille entre 1 et 2,5 pour les textes du corpus *BAF*. Le tableau 8 montre les corrélations linéaires entre ces densités et les résultats obtenus à chaque stade de l'étape 1 de l'algorithme :

	P_a	R_a	F_a
Dtransfuge	42,9 %	95,9 %	94,9 %

Tableau 8. *Corrélations entre Dtransfuge et les résultats.*

On constate que la corrélation est très forte entre la densité et le rappel. Ce résultat était prévisible : plus il y a de transfuges par phrase et plus on peut en tirer de points d'ancrage. Par ailleurs, la corrélation négative entre densité et précision n'est pas significative : l'augmentation du rappel au cours de l'algorithme n'engendre pas une dégradation importante de la précision. Ces conclusions sont illustrées par la figure 8 où sont représentées (pour les dix textes du corpus *BAF* hormis *Xerox*), les valeurs de R_a, P_a et F_a en fonction de *Dtransfuge*.



(Dtransfuge, P_a) (*Dtransfuge, R_a)* (*Dtransfuge, F_a)*

Figure 8. *Résultats de l'alignement en fonction de Dtransfuge.*

Il est notable que pour tous les points on peut minorer *F* par une fonction linéaire de la densité.

$F \geq Dtransfuge/3$

cette minoration n'étant définie que pour $Dtransfuge \leq 3$. Il faudrait conduire une étude empirique plus approfondie pour mieux déterminer la portée d'une telle minoration. En effet, les résultats de cette méthode dépendent de propriétés formelles des textes concernés, comme *Dtransfuge*, mais aussi de relations traductionnelles (comme le parallélisme ou la proportion d'homographes qui ne sont pas traductions mutuelles) dont on ne peut présumer sans connaissances linguistiques. Il nous semble néanmoins qu'étant donnée la multiplicité des contraintes de filtrages que nous imposons, dans le but de limiter le bruit, le facteur déterminant reste *Dtransfuge*, cette densité permettant d'estimer *a priori* les résultats du préalignement.

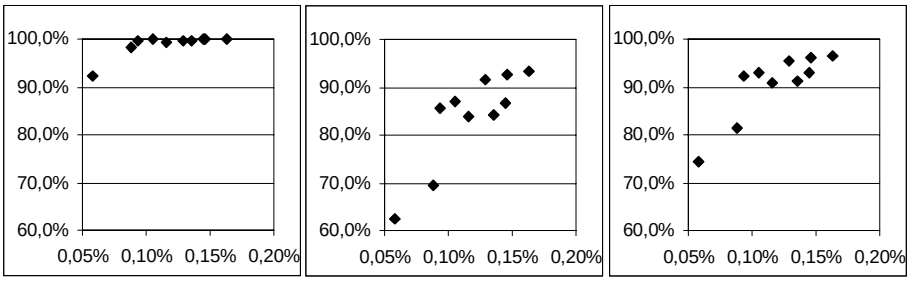
5.2.2 Corrélations des résultats de l'alignement avec la densité a priori *de cognat*

De même on peut raisonnablement supposer que la réussite du préalignement basé sur les cognats est conditionnée par la fréquence des cognats observés au sein du corpus. Plus cet indice est important, et plus il est probable que l'alignement résultat sera correct et complet.

Pour valider cette hypothèse, nous avons dénombré, pour chaque texte le nombre de *cognats candidats* identifiés. Nous avons ensuite rapporté cette quantité au nombre de couples de mots comparés, pour obtenir la densité de cognats candidats *Dcognat*. On peut alors étudier les corrélations entre *Dcognat* et les résultats du préalignement à la sortie de l'étape 2 de l'algorithme.

	P_a	R_a	F_a
Dcognat	0,76	0,85	0,85

Tableau 9. *Corrélations entre Dcognat et les résultats.*



(*Dcognat*, P_a) (*Dcognat*, R_a) (*Dcognat*, F_a)

Figure 9. *Résultats de l'alignement en fonction de Dcognat.*

Les nuages de points de la figure 9 représentent les résultats en fonction de *Dcognat* (chaque point correspond à un texte du corpus). On constate le même type de corrélation linéaire que précédemment : faible sensibilité de la précision, et corrélation forte du rappel de l'alignement avec la densité de cognats potentiels. Cette corrélation donne un caractère de prévisibilité aux résultats obtenus : pour deux textes dont on sait *a priori* qu'ils contiennent, entre eux, de nombreux cognats potentiels (on peut estimer cette densité à partir d'un échantillon), on peut prévoir que le rappel de l'alignement basé sur les cognats (au sens large, incluant les transfuges) sera important.

Comme avec les transfuges seuls, on peut se risquer à une minoration par une fonction linéaire, empiriquement vérifiée pour notre corpus :

$F \geq Dcognat*500$

6. Conclusion

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, sur ce corpus, les performances de la détermination des cognats : nous montrons comment combiner les *n*-grammes aux notions de transfuges et de sous-chaînes maximales quasi-parallèles afin d'obtenir de meilleurs résultats.

Nous avons par ailleurs tenté de déterminer la meilleure façon d'intégrer la cognation dans un système d'alignement : l'heuristique de précision d'abord nous a confirmé que l'indice, tout comme les transfuges, est applicable dès la phase de préalignement visant à réduire la taille de l'espace de recherche. La grande redondance des informations fournies par transfuges et cognats permet en effet d'appliquer un système de contraintes très serrées qui assure une précision importante. Par suite, la cognation peut être réinjectée au sein d'une mesure de distance applicable à la totalité du chemin d'alignement, et produire une nouvelle amélioration des résultats (par rapport à la méthode basée sur les seules longueurs de phrases).

Bien que ces résultats fussent connus ([SIM 96], [MEL 96], [LAN 98]), nous avons présenté une méthode d'intégration originale fondée sur des résultats empiriques précis et une heuristique efficace permettant de réduire les calculs (en *O(n log(n))* : la bonne tenue des résultats en est la preuve (96,5 % de précision et 94,8 % de rappel en moyenne pour le *BAF*, hors *Xerox*).

Notons que le cadre algorithmique présenté est indépendant des techniques permettant d'établir des correspondances lexicales : celles-ci pourraient être captées par des dictionnaires bilingues *ad hoc*, des mesures d'association statistiques (telles que rapport de vraisemblance, *t-score*, information mutuelle), ou des modules visant à capter plus finement les cognats à partir de connaissances linguistiques. L'architecture du système est par conséquent indépendante du couple de langues impliqué, et ses résultats ne sont pas liés à la nature des correspondances lexicales (cognats, transfuges ou autres) : il nous est en effet apparu que les résultats sont conditionnés essentiellement par la *densité* de ces correspondances ainsi que par l'application prioritaire des correspondances les plus fiables.

Quant à la méthode de détermination des cognats, elle dépend bien entendu des caractéristiques des textes parallèles et surtout du couple de langues considéré. Il serait intéressant d'en évaluer les résultats sur des langues non apparentées, pour des textes de spécialité : le plus souvent en effet le vocabulaire spécialisé est contraint à une forte convergence sous la pression des standards internationaux, basés

essentiellement sur le fonds gréco-latin. Nous n'avons pas encore, à ce stade de nos recherches, effectué une telle vérification.

Enfin, nous pensons que la méthode de préalignement est adaptée au développement d'heuristiques pour la détection d'omission ou d'interversion de sections importantes, dans la mesure où la forte densité des points d'ancrage permet de faire apparaître clairement les ruptures dans le parcours du chemin. Les techniques ici étudiées pourraient alors être étendues à des corpus parallèles ne respectant pas rigoureusement le critère de monotonie.

Remerciements

L'auteur remercie Jean Véronis pour ses encouragements à participer au projet ARCADE, ainsi que les relecteurs anonymes, pour leurs remarques pertinentes et leurs critiques constructives concernant la première version de cet article.

7. Bibliographie

[BAR 64] BAR-HILLEL, Y., "The future of Machine Translation", In *Language and Information : Selected Essays on their Theory and Application*, Addison-Wesley Publishing Company, Inc., p. 180-184, 1964.

[BRO 91] BROWN P., LAI J., MERCER R., "Aligning Sentences in Parallel Corpora", *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, Morristown NJ, p. 169-176, 1991.

[CHE 93] CHEN S. (1993), "Aligning sentences in bilingual corpora using lexical information", *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL-93*, Columbus Ohio, p. 9-16, 1993.

[CHU 93] CHURCH K.W., "Char-align : A Program for Aligning Parallel Texts at the Character Level", *Proceedings of the 31st Annual Meeting of the ACL*, Colombus, Ohio, pp.1-8, 1993.

[DAV 95] DAVIS M. W., DUNNING T. E., OGDEN W. C., "Text Alignment in the Real World : Improving Alignments of Noisy Translations...", *Proceedings of EACL 95*, 1995.

[DEB 92] DEBILI F, SAMMOUDA E., "Appariements de Phrases de Textes bilingues Français-Anglais et Français-Arabes", *Actes de COLING-92*, Nantes, p. 528-524, 1992.

[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.

[GAL 91] GALE W., CHURCH K. W., "A program for aligning sentences in bilingual corpora", *Proceedings of the 29th Annual Meeting of the ACL*, Berkeley, CA, p. 177-184, 1991.

[ISA 92] ISABELLE P., "La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie", *Meta*, XXXVII, 4, p.721-731, 1992.

[ISA 93] ISABELLE P., DYMETMAN M., FOSTER G., JUTRAS J.-M., MACKLOVITCH E., "Translation Analysis and Translation Automation", *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in MT*, Kyoto, 1993.

[ISA 96] ISABELLE P., SIMARD M., "Propositions pour la représentation et l'évaluation des alignements de textes parallèles", URL : http://www-rali.iro.umontreal.ca/arc-a2/PropEval, 1996.

[KAY 93] KAY M., RÖSCHEISEN M., "Text-Translation Alignment", *Computational Linguistics*, Vol. 19, N°1, p.121-142, 1993.

[KRA 99a] KRAIF O., "Architecture d'un système d'alignement : étude pour une intégration optimale des indices d'alignement", *Actes des Journées internationales de linguistique appliquée, JILA '99*, Nice, p. 161-164, 1999.

[KRA 99b] KRAIF O., "Réflexions autour des concepts de correspondance lexicale et d'alignement textuel", *Actes du colloque Linguistique contrastive et Traduction Approches Empiriques*, Louvain-la-Neuve, p. 25-26, 1999.

[LAN 95] LANGÉ J.-M., GAUSSIER E., "Alignement de corpus multilingues au niveau des phrases", *T.A.L.*, Vol. 36, N° 1-2, p. 67-80, 1995.

[LAN 97a] LANGLAIS P., "Alignement de corpus multilingues : intérêts, algorithmes et évaluations", *FRACTAL 1997*, p.245-254, 1997.

[LAN 97b] LANGLAIS P., EL-BÈZE M., "Alignement de corpus bilingues : algorithmes et évaluation", *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, p. 191-197, 1997.

[LAN 98] LANGLAIS P., SIMARD M., VERONIS J., ARMSTRONG S., BONHOMME P., DEBILI F., ISABELLE P., SOUISSI E., THÉRON P., "ARCADE : A cooperative Research Project on Parallel Text Alignment Evaluation". *Proceedings of First International Conference on Language Resources and Evaluation (LREC)*, Grenade, Expagne, 28-30 Mai, p. 289-292, 1998.

[MAC 92] MACKLOVITCH E., "Corpus-Based Tools for Translators", *Proceedings of the 33rd Annual Conference of the American Tranlators Association*, San Diego California, 1992.

[MEL 95] MEL'CUK I., CLAS A., POLGUERE A., *Introduction à la lexicologie explicative et combinatoire*, Duculot, Louvain-la-neuve, 1995.

[MEL 96] MELAMED I. D., "A geometric Approach to Mapping Bitext Correspondence", *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, p. 1-12, 1996.

[NAG 84] NAGAO M., "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle", *Artificial and Human Intelligence*, Editions A. Elithorn et R. Banerji, Elsevier Science Publishers, Amsterdam, 1984.

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

[SAT 90] SATO S., NAGAO M., "Toward Memory-based Translation", *Proceedings of the 13th International Conference on Computational Linguistics, COLING-90*, Helsinky, p. 247-252, 1990.

[SIM 92] SIMARD M., FOSTER G., ISABELLE P., "Using cognates to align sentences", *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montréal Canada, p. 67-81, 1992.

[SIM 96] SIMARD M., PLAMONDON P., "Bilingual Sentence Alignment : balancing robustness and accuracy", *Proceedings of AMTA-96*, Montréal Canada, p. 135-144, 1996.

[SIM 98] SIMARD M., "The BAF : A Corpus of English-French Bitext", *Proceedings of the First International Conference on Language Resources and Evaluation*, Grenade Espagne, p. 489-494, 1998.

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

[VOL 97] VOLLE M., *Analyse des données*, Economica, Paris, 1997.

8. Annexes

Annexe 1 : Résultat des méthodes d'identification des cognats

	transfuges			**n-grammes**			**n-grammes + transfuges**			***SCM***			***SCM* + transfuges**		
n	P_c	R_c	F_c	P_c	R_c	F_c	P_c	R_c	F_c	P_c	R_c	F_c	P_c	R_c	F_c
≥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 10. *Les valeurs de Précision, Rappel, et F-mesure sont données en pourcentage. Par SCM, on désigne la méthode basée sur les sous-chaînes maximales.*

	P_c	R_c	F_c	P_a	R_a	F_a
3-grammes (*)	15	75	24	82	32	45,5
4-grammes	31	48	37	97	63	76,3
SCM>=3 + transfuge	55	76	63	99	76	85,9
SCM>=4 + transfuge	75	74	74	100	86	92,2
SCM>=5 + transfuge	85	72	78	100	85	91,9
SCM>=6 + transfuge	86	65	74	100	84	91,2
SCM>=7 + transfuge	90	58	70	100	82	89,9
SCM>=8 + transfuge	92	55	69	100	79	88,3
SCM>=9 + transfuge	99	53	69	100	76	86,3
Combinaison 1 [1]	68	73	70	99	80	88,6
Combinaison 2 [2]	75	73	74	100	86	92,5
Transfuges seuls	100	50	66	100	68	81,2
Cognats [3]	100	100	100	100	74	85,2

Tableau 11. *[1] transfuges, 4-grammes (mots de longueur <7), SCM avec n≥4 et r>2/3, [2] transfuges, n-grammes avec n≥3 et r>2/3, SCM avec n≥5 et r=2/3, [3] cognats obtenus manuellement.*

Coeff. de corrélation	P_a	R_a	F_a
P_c	0,76	0,74	0,75
R_c sans (*)	0,41	0,72	0,71
F_c	0,85	0,94	0,93

Tableau 12.

Annexe 2 : Calcul de DGC (Distance basée sur la méthode de Gale et Church, [GAL 92])

Les auteurs ont constaté, à partir d'observations empiriques, que le rapport des longueurs entre des unités textuelles alignées, exprimées en nombre de caractère, suit approximativement une distribution normale. Considérant que 1 caractère de la source donne *X* caractères de la cible, les auteurs supposent que *X* est une variable aléatoire avec une

distribution normale, de moyenne c et de variance S^2. Pour une phrase source de n_1 caractères, on peut considérer la longueur n_2 de la phrase cible comme résultant de la somme de n_1 variables aléatoires X_i. D'après le théorème de la limite centrale, la variable aléatoire δ définie au niveau de la phrase :

$$\delta = \frac{(X_1 + X_2 + .. + X_{n1}) - n_1 * c}{\sqrt{S^2 n_1}} = \frac{(n_2 - n_1 * c)}{\sqrt{S^2 n_1}}$$

suit donc une loi normale centrée réduite (moyenne nulle, variance 1). De la probabilité de l'alignement sachant une valeur donnée de δ, on peut tirer une mesure de distance entre deux phrases P et P' : distance(P,P')=-log Prob(alignement/δ)

D'après le théorème de Bayes, on a :
prob(alignement/δ) = Prob(δ/alignement)*prob(alignement)/ prob(δ)

Les auteurs supposent que *prob(δ)* correspond à une constante, qui peut donc être négligée car elle intervient de manière identique pour tous les chemins possibles. Le deuxième facteur, *prob(alignement)* est assimilé à la probabilité générale d'observer telle ou telle transition. Les auteurs lui affectent des valeurs moyennes issues de l'observation (cf. la première ligne du tableau 6). Enfin, comme δ suit une distribution normale on obtient l'estimation :

$$prob(\delta / alignement) = 2(1 - prob(|\delta|)) = 2 * (1 - \int_{-\infty}^{\delta} \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}} dt)$$

Annexe 3 : Algorithme de Viterbi

Le principe en est le suivant : pour calculer un chemin optimal, on calcule d'abord tous les sous-chemins optimaux menant à ce chemin. Et pour chaque sous-chemin, on réitère l'opération récursivement.

Ainsi, pour un point de coordonnées *(i,j)*, la distance du meilleur sous-chemin menant à *(i,j)* est calculée en fonction des sous-chemins y menant. Dans le cas suivant, si l'on ne tient compte que des transitions 1:1, 1:0, 0:1, 2:1, 1:2, 2:2, on a :

$$D(i,j) = min \begin{cases} D(i-1;j-1) + d(P_{i-1};P'_{j-1}) \\ D(i;j-1) + d(; P'_{j-1}) \\ D(i-1;j) + d(P_{i-1};) \\ D(i-2;j-1) + d(P_{i-2}P_{i-1};P'_{j-1}) \\ D(i-1;j-2) + d(P_{i-1};P'_{j-2}P'_{j-1}) \\ D(i-2;j-2) + d(P_{i-2}P_{i-1};P'_{j-2}P'_{j-1}) \end{cases}$$

Où *D(x,y)* est la distance du sous-chemin optimal menant à *(x,y)* et d($P_1P_2..P_n$;$P'_1P'_2..P'_m$) est la distance liée au binôme ($P_1P_2..P_n$;$P'_1P'_2..P'_m$).