

Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés

Sylwia Ozdowska

ERSS – Université Toulouse le Mirail
Maison de la Recherche
5 allées Antonio Machado
31058 Toulouse Cedex 1
ozdowska@univ-tlse2.fr

Résumé – Abstract

Nous présentons une méthode d'appariement de mots, à partir de corpus français/anglais alignés, qui s'appuie sur l'analyse syntaxique en dépendance des phrases. Tout d'abord, les mots sont appariés à un niveau global grâce au calcul des fréquences de cooccurrence dans des phrases alignées. Ces mots constituent les couples amorces qui servent de point de départ à la propagation des liens d'appariement à l'aide des différentes relations de dépendance identifiées par un analyseur syntaxique dans chacune des deux langues. Pour le moment, cette méthode dite d'appariement local traite majoritairement des cas de parallélisme, c'est-à-dire des cas où les relations syntaxiques sont identiques dans les deux langues et les mots appariés de même catégorie. Elle offre un taux de réussite de 95,4% toutes relations confondues.

We present a word alignment procedure based on a syntactic dependency analysis of French/English parallel corpora. First, words are associated at a global level by comparing their co-occurrences in aligned sentences with respect to their overall occurrences in order to derive a set of anchor words. The anchor words are the starting point of the propagation process of alignment links using the different syntactic relations identified by a parser for each language. This process is called the local alignment. For the moment, it is performed basically when the syntactic relations are identical in both languages and the words aligned have the same part of speech. This method achieves a precision rate of 95,4% all syntactic relations taken into account.

Mots-clefs – Keywords

appariement syntaxique de mots, corpus parallèle, traitement automatique des langues naturelles
syntactic word alignment, parallel corpora, natural language processing

1 Introduction

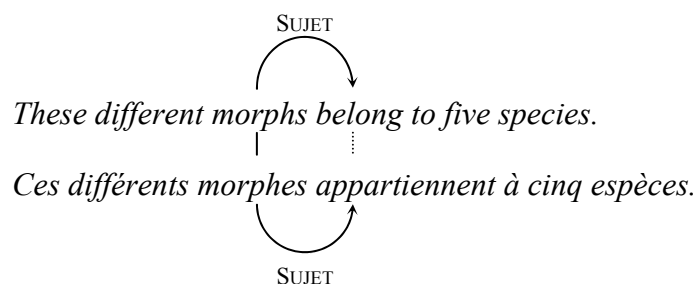
Qu'il soit destiné à la traduction automatique ou à la construction de ressources terminologiques bilingues (Daille *et al.*, 1994 ; Gaussier, 1998 ; Wu, 2000 ; Hull, 2001), l'appariement de mots s'effectue principalement à l'aide de modèles statistiques, notamment ceux, très répandus, de (Brown *et al.*, 1993). Jugée trop coûteuse (Véronis, 2000), l'utilisation de ressources linguistiques dans ce type de tâche se limite encore très souvent au recours à des dictionnaires électroniques, des étiqueteurs ou des lemmatiseurs. Cependant, des propositions sont faites depuis peu qui visent à exploiter des données issues de l'analyse syntaxique dans les systèmes de traduction probabilistes, soit pour les deux langues source et cible (Wu, 2000) soit pour la langue source uniquement (Fox, 2002 ; Lin et Cherry, 2003). Les données syntaxiques n'interviennent néanmoins que dans un second temps par rapport aux calculs statistiques, et ce pour contraindre les appariements que ces derniers peuvent produire.

La méthode d'appariement présentée dans cet article utilise les données syntaxiques non seulement pour confirmer des liens d'appariement existants mais aussi, et surtout, pour en découvrir de nouveaux. Le choix de cette méthode est motivé par l'objectif que nous poursuivons : parvenir à un appariement précis en captant aussi bien des appariements entre mots fréquents que des appariements entre mots rares et/ou spécifiques au corpus.

2 Hypothèse de départ

Le point de départ de notre étude est l'hypothèse formulée par (Debili et Zribi, 1996) selon laquelle « *les liaisons paradigmatisques peuvent aider à déterminer les relations syntagmatiques, et inversement* ». Plus particulièrement, nous reprenons l'idée que les relations de dépendance syntaxique sont susceptibles, d'une part, de confirmer ou d'infirmer des liens d'appariement et, d'autre part, de créer de nouveaux liens. Le raisonnement est le suivant :

Si deux mots $T1_i$ (*morph*, dans l'exemple) et $T2_p$ (*morphe*) sont appariés et s'il existe une relation de dépendance syntaxique entre $T1_i$ (*morph*) et $T1_j$ (*belong*), d'une part, et entre $T2_p$ (*morphe*) et $T2_q$ (*appartenir*), d'autre part, alors $T1_j$ (*belong*) et $T2_q$ (*appartenir*) peuvent être appariés.



Cet article présente une implémentation de ce mécanisme de propagation des liens d'appariement à l'aide des relations de dépendance syntaxique ainsi que son évaluation. Notre étude s'appuie sur un corpus de traduction français/anglais constitué au sein du service

linguistique de l'INRA¹ (Frérot *et al.*, 2001). Ce corpus a été aligné automatiquement au niveau des phrases et il compte au total environ 300 000 mots. Pour le traiter, nous avons choisi les outils SYNTAX (Bourigault et Fabre, 2000). Il s'agit de deux analyseurs syntaxiques, l'un pour le français, l'autre pour l'anglais, qui prennent en entrée un corpus étiqueté et effectuent une analyse en dépendance de chaque phrase, puis ils extraient du corpus un ensemble de mots et de syntagmes.

3 Appariement global

La première étape du processus d'appariement tel que nous l'avons implémenté consiste à trouver des couples susceptibles de servir de point de départ à la propagation, appelés couples amorces. Nous avons choisi de construire un ensemble de couples amorces directement à partir du corpus en associant les mots en langue 1 ($m1$), l'anglais, et les mots en langue 2 ($m2$)², le français, extraits par SYNTAX sur la base de leur fréquence d'apparition dans des phrases alignées (Gale et Church, 1991 ; Ahrenberg *et al.*, 2000). Plus précisément, il s'agit de comparer le nombre de fois où deux mots $m1$ et $m2$ apparaissent ensemble dans des phrases alignées, c'est la fréquence de cooccurrence, par rapport à la fréquence de chacun d'entre eux sur l'ensemble du corpus. Soient $f(m1)$ la fréquence de $m1$ sur l'ensemble du corpus, $f(m2)$ la fréquence de $m2$ et $f(m1, m2)$ la fréquence de cooccurrence, la mesure d'association calculée est la suivante :

$$j(m1, m2) = \frac{f(m1, m2)}{f(m1) + f(m2) - f(m1, m2)}$$

Dans le cadre de cette étude, nous calculons cette mesure uniquement pour les $m1$ et $m2$ dont la fréquence sur l'ensemble du corpus est supérieure ou égale à 5. En effet, l'appariement global concerne les mots les plus fréquents, dont les équivalents peuvent être retrouvés avec une bonne précision par des méthodes statistiques (Ozdowska et Bourigault, 2004). Ensuite, seuls sont sélectionnés les couples ($m1, m1$) tels que $j(m1, m2) \geq 0,2$. L'ensemble des couples amorces ainsi obtenus constitue le lexique global. Une fois projetés au niveau des phrases alignées, ces couples permettent d'initier le processus d'appariement local.

4 Appariement local

L'appariement local consiste à mettre en correspondance des $m1$ avec des $m2$ phrase à phrase, à partir de couples amorces et en suivant les relations de dépendance syntaxique mises en place par les analyseurs SYNTAX. Contrairement à l'appariement global, il vise à retrouver les équivalents des mots peu fréquents. Il s'agit donc de deux procédures complémentaires.

¹ Institut National de la Recherche Agronomique. Nous remercions A. Lacombe de l'INRA de nous avoir autorisée à utiliser ce corpus à des fins de recherche.

² Il s'agit de mots pleins (noms, verbes, adjectifs, adverbes) aussi bien pour l'appariement global que local.

Généralement les poissons **sont capturés** quand ils migrent de leur zone d'engraissement.

- (3) *Most of the young shad **reach** the sea.*
*La plupart des alosons **gagne** la mer.*
- (4) *The eggs are very small and **fall** to the bottom.*
*Les oeufs de très petite taille **tombent** sur le fond.*
- (5) *X is a model which **was designated** to stimulate...*
*X est un modèle qui **a été conçu** pour stimuler...*

APPARIEMENT DES VERBES REGIS. L'appariement se fait selon trois schémas de propagation : *V_Prep_V* (1), *N_Prep_V* (2) et *Adj_Prep_V* (3).

- (1) *Ploughing tends to **destroy** the soil microaggregated structure.*
*Le labour tend à **rompre** leur structure microagrégée.*
- (2) *The capacity to **colonize** the digestive mucosa...*
*L'aptitude à **coloniser** le tube digestif...*
- (3) *An established infection is impossible to **control**.*
*Toute infection en cours est impossible à **maîtriser**.*

La Figure 1 montre les résultats et les évaluations par schéma de propagation. Le nombre d'occurrences correspond au nombre de fois où le même schéma a été trouvé dans les deux langues, c'est-à-dire au nombre de cas de parallélisme. Etant donné le nombre important d'occurrences pour chaque schéma de propagation, seule une partie des occurrences a été la plupart du temps évaluée et ce suivant l'ordre de leur apparition dans le corpus.

	<i>Régi → V- Recteur</i>					<i>Recteur → V - Régi</i>		
	<i>Adv_V</i>	<i>Suj_V</i>	<i>V_Obj</i>	<i>V_Prep_N</i>	<i>V_Prep_V</i>	<i>V_Prep_V</i>	<i>N_Prep_V</i>	<i>Adj_Prep_V</i>
occurrences	177	1705	1495	934	75	52	25	50
cas évalués	95	663	692	478	57	52	25	50
réussite (%)	98,9	94,8	92,4	95,8	96,4	95,1	100	99
	94,5					98,4		

Figure 1 Appariement des verbes régis et recteurs – résultats et évaluations

4.3 Appariement des adjectifs et des noms

Les adjectifs, aussi bien que les noms, peuvent être appariés selon les deux modes présentés en 4.1. Cela implique néanmoins que les adjectifs et les noms ne soient pas traités de manière complètement indépendante, et ce pour deux raisons : a) l'ambiguïté potentielle qui existe lors de la propagation à partir des noms recteurs (4.1) et qui concerne tous leurs dépendants et b) la différence dans la formation, donc dans la structure, des syntagmes nominaux dans les deux langues qui provoque des phénomènes de non parallélisme et de transposition (Chuquet et Paillard, 1989). Par conséquent, les dépendants du nom font l'objet d'un traitement spécifique

qui autorise des appariements alors même que les relations de dépendance utilisées sont différentes dans les deux langues et/ou que les mots appariés ne sont pas de même catégorie. Les adjectifs et les noms sont donc traités de manière indépendante dans les cas suivants : a) appariement à partir des régis et b) pour les noms seulement, appariement à partir de recteurs qui ne sont pas des noms. Ils sont par contre traités ensemble lors de l'appariement à partir des noms recteurs.

APPARIEMENT DES ADJECTIFS RECTEURS. Les schémas de propagation utilisés sont : *Adv_Adj* (1), *Adj_Prep_N* (2) et *Adj_Prep_V* (3).

- (1) *The white cedar exhibits a **very common** physical defect.*
*Le Poirier-pays présente un défaut de forme **très fréquent**.*
- (2) *The area presently **devoted** to agriculture represents...*
*La surface actuellement **consacrée** à l'agriculture représenterait...*
- (3) *Only fours plots were **liable** to receive this input.*
*Seulement quatre parcelles sont **susceptibles** de recevoir ces apports.*

	<i>Adv_Adj</i>	<i>Adj_Prep_N</i>	<i>Adj_Prep_N</i>
occurrences	162	454	45
cas évalués	98	198	45
réussite (%)	98,9	97,4	97,7
	97,9		

Figure 2 Appariements des adjectifs recteurs – résultats et évaluations

APPARIEMENT DES NOMS RECTEURS. Les noms sont appariés suivant les schémas : *N_Adj* (1), *N_N/N_Prep_N* (2), *N_Prep_N* (3) et *N_Prep_V* (4).

- (1) *Allis shad remain on the continental shelf.*
*La grande alose reste sur le **plateau continental**.*
- (2) *Nature of micropolluant carriers.*
*La nature des **transporteurs** des micropolluants.*
- (3) *The **bodies** of shad are generally fusiform.*
*Le **corps** des aloses est généralement fusiforme.*
- (4) ***Ability** to react to light.*
***Capacité** à réagir à la lumière.*

APPARIEMENT NON AMBIGU DES NOMS REGIS. L'appariement des noms régis est non ambigu lorsque leur recteur n'est pas lui-même un nom. C'est le cas pour quatre schémas de propagation : *Suj_V_Obj* (1), *V_Prep_N* (2) et *Adj_Prep_N* (3).

- (1) *The **caterpillars** can inoculate the **fungus**.*
*Les **chenilles** peuvent inoculer le **champignon**.*
- (2) *The roots are placed in **tanks**.*
*Les racines sont placées en **bacs**.*

- (3) *Botrysis*, a fungus responsible for grey rot.
Botrysis, champignon responsable de la *pourriture* grise.

	Régi → Nom - Recteur				Recteur → Nom - Régi		
	N_Adj	N_N/N_Prep_N	N_Prep_N	N_Prep_V	Suj_V_Obj	V_Prep_N	Adj_Prep_N
occurrences	4706	2960	2966	18	1631	461	325
cas évalués	965	625	671	18	623	268	188
réussite (%)	93,7	87,5	91,8	100	95,6	95,5	97,8
	91,5				97,3		

Figure 3 Appariement non ambigu des noms recteurs et régis– résultats et évaluations

APPARIEMENT POTENTIELLEMENT AMBIGU DES NOMS ET ADJECTIFS REGIS. Comme nous l’avons évoqué en 4.1, l’appariement est potentiellement ambigu quand la propagation se fait à partir d’un nom recteur vers le(s) nom(s) et/ou les(s) adjectif(s) régi(s). C’est le cas non seulement parce que le nombre de régis peut être supérieur à 1 mais aussi parce que leur catégorie peut varier d’une langue à l’autre ; c’est la transposition. Ainsi, un adjectif régi peut être rendu par un nom régi et vice versa, les relations susceptibles de servir de base à la propagation sont alors différentes.

Compte tenu de l’existence d’une ambiguïté potentielle, le principe d’appariement des noms et adjectifs régis à partir des noms recteurs ($n1$, $n2$) s’articule autour des trois situations que l’on peut rencontrer :

1. $n1$ et $n2$ ont chacun un seul régi, respectivement $reg1$ et $reg2$, par la relation NN, ADJ ou PREP, on apparie $reg1$ et $reg2$;

the drained whey
le lactosérum d’égouttage
 \Rightarrow (**drained, égouttage**)

2. $n1$ a un seul régi $reg1$ et $n2$ plusieurs $\{reg2_1, reg2_2, \dots, reg2_n\}$, ou l’inverse. S’il existe un seul appariement local déjà effectué par propagation et/ou un appariement global $app(reg1, reg2_i)$, ou l’inverse, on le choisit et on élimine les autres. Sinon, on garde l’ensemble des appariements $(reg1, \{reg2_1, reg2_2, \dots, reg2_n\})$, ou l’inverse, sans prendre de décision ;

stimulant substances which are absent from...
substances solubles stimulantes absentes de...
(stimulant, {soluble, stimulant, absent})
 $app(stimulant, stimulant) = 1$
 \Rightarrow (**stimulant, stimulant**)

3. $n1$ et $n2$ ont plusieurs régis, $\{reg1_1, reg1_2, \dots, reg1_m\}$ et $\{reg2_1, reg2_2, \dots, reg2_n\}$ respectivement. Pour chaque $reg1_i$, on vérifie s'il existe un appariement local déjà effectué par propagation et/ou un appariement global $app(reg1_i, reg2_j)$. Si oui, on choisit cet appariement et on élimine tous les couples $(reg1_i, reg2_k)$ tels que $k \neq j$. Sinon, on garde tous les appariements $(reg1_i, reg2_j)$ sans prendre de décision.

<i>reference product on the market</i>	<i>a big rectangular net, which is lowered...</i>
<i>produit commercial de référence</i>	<i>un vaste filet rectangulaire immergé...</i>
$(reference, \{commercial, référence\})$	$(big, \{vaste, rectangulaire, immergé\})$
$(market, \{commercial, référence\})$	$(rectangular, \{vaste, rectangulaire, immergé\})$
$app(reference, référence) = 1$	$app(rectangular, rectangulaire) = 1$
$\Rightarrow (reference, référence)$	$\Rightarrow (rectangular, rectangulaire)$
$\Rightarrow (market, commercial)$	$\Rightarrow (big, \{vaste, immergé\})$

Ce principe de propagation présente deux avantages majeurs : il permet de lever un certain nombre d'ambiguïtés de manière locale et de gérer des cas de non parallélisme avec ou sans transposition au niveau de la catégorie grammaticale des mots appariés. Il est à noter que les appariements non désambiguïsés, comme $(big, \{vaste, immergé\})$, qui représentent près de 30% des cas, sont considérés comme non résolus à ce stade du processus et ne sont donc pas pris en compte lors de l'évaluation.

	N_{-}
occurrences	10726
cas évalués	3464
réussite (%)	97,7

Figure 4 Appariement à partir des noms recteurs – résultats et évaluation

5 Discussion et perspectives

L'appariement par propagation offre un taux de précision de 95,4% sur une base de près de 10 000 couples évalués. Il convient de noter que ce résultat est lié en partie au contexte favorable dans lequel la méthode d'appariement a été testée. Tout d'abord, nous avons travaillé à partir d'un corpus de traduction en domaine spécialisé dans lequel les structures syntaxiques présentent une forte similarité pour les deux langues. Deuxièmement, nous n'avons pas cherché à appairer tous les mots mais seulement les mots pleins régis et/ou recteurs. Les évaluations concernent donc uniquement les appariements de mots effectués dans ces conditions d'expérimentation précises. C'est pourquoi le résultat obtenu est difficilement comparable aussi bien à ceux qui sont rapportés par (Daille *et al.*, 1994), 70 à 80%, (Gaussier, 1998), 90 à 98% ainsi que (Hull, 2001), 56%, et qui concernent l'évaluation de l'appariement de termes auquel l'appariement de mots sert de support, qu'à ceux de (Lin et Cherry, 2003), précision estimée à 89,2% ou 95,7%³ dans le cadre d'une tâche d'évaluation (Mihalcea et Pedersen, 2003), qui

³ Selon la calcul statistique utilisé à la base.

concernent l'appariement de tous les mots de la phrase. (Lin et Cherry, 2003) remarquent néanmoins que l'introduction d'une contrainte syntaxique améliore de manière significative les performances de leurs algorithmes. Il s'agit dans ce cas d'une contrainte de cohésion (Fox, 2002) qui stipule que « *if two phrases are disjoint in the English sentence, the alignment must not map them to overlapping intervals in the French sentence* » (Lin et Cherry, 2003). Cette contrainte suppose une représentation de la phrase anglaise sous forme d'arbre de dépendance et vise globalement à interdire ou à désambiguïser certains appariements statistiques. Pour l'appariement par propagation, l'exploitation des données syntaxiques est au cœur même de la méthode, ces dernières ne constituent pas seulement une information additionnelle.

La propagation des liens d'appariement le long des relations syntaxiques s'avère très efficace dans les cas où il y a parallélisme dans les deux langues. Elle donne également une bonne précision dans les cas de transposition où elle conduit à mettre en correspondance des adjectifs et des noms. Le rappel reste à évaluer. En poursuivant l'analyse des cas de non parallélisme, dont on trouve une ébauche dans (Ozdowska et Bourigault, 2004), nous espérons mettre au jour des régularités liées aux phénomènes de variation interlingue et étendre ainsi la propagation à ce type de situations. Cette technique doit encore être testée et évaluée sur de nouveaux corpus parallèles, notamment ceux utilisés dans les campagnes d'évaluation Arcade (Véronis et Langlais, 2000). Enfin, notre réflexion concerne l'utilisation de cette méthode sur des corpus comparables.

Références

- AHRENBURG L., ANDERSSON M., MERKEL M. (2000), A knowledge-lite approach to word alignment, J. Véronis (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 97-138.
- BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, Université Toulouse le Mirail, pp. 131-151.
- BROWN P., DELLA PIETRA S., MERCER R. (1993), The mathematics of statistical machine translation : parameter estimation, *Computational Linguistics*, 19(2), pp. 263-311.
- DAILLE B., GAUSSIER E., LANGÉ J.-M. (1994), Towards Automatic Extraction of Monolingual and Bilingual Terminology, *Proceedings of the International Conference on Computational Linguistics (COLING'94)*, pp. 515-521.
- DEBILI F., ZRIBI A. (1996), Les dépendances syntaxiques au service de l'appariement des mots, *Actes du 10^{ème} Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*.
- FREROT C., RIGOU G., LACOMBE A. (2001), Approche phraséologique d'une extraction automatique de terminologie dans un corpus scientifique bilingue aligné, *Actes des 4èmes rencontres Terminologie et Intelligence Artificielle*, Nancy, pp 180-188.
- FOX H. J. (2002), Phrasal Cohesion and Statistical Machine Translation, *Proceedings of EMNLP-02*, pp. 304-311.
- GALE W. A., CHURCH K. W. (1991), Identifying Word Correspondences in Parallel Text, *Proceedings of the DARPA Workshop on Speech and Natural Language*.

- GAUSSIER E. (1998), Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora, *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL '98)*, pp. 444-450.
- GAUSSIER E., HULL D. A., AÏT-MOKHTAR S. (2000), Term alignment in use, J. Véronis (ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 253-274.
- HULL D. A. (2001), Software tools to support the construction of bilingual terminology lexicons, D. Bourigault, Ch. Jacquemin, M. -C. L'Homme (eds.), *Recent Advances in Computational Terminology*, John Benjamins, pp. 225-244.
- LIN D., CHERRY C. (2003), Linguistic Heuristics in Word Alignment, *Proceedings of PACLing 2003*.
- MIHALCEA R., PEDERSEN T. (2003), An Evaluation Exercise for Word Alignment, *Workshop Proceedings on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond (HLT-NAACL 2003)*, pp. 1-10.
- OZDOWSKA S., BOURIGAULT D. (2004), Détection de relations d'appariement bilingue entre termes à partir d'une analyse syntaxique de corpus, *Actes du 14^{ème} Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence artificielle*.
- VÉRONIS J. (2000), From the Rosetta stone to the information society. A survey of parallel text processing, Véronis J. (ed), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht : Kluwer Academic Publishers, pp. 1-24.
- VÉRONIS J. et LANGLAIS P. (2000), Evaluation of parallel text alignment systems. The ARCADE project, Véronis J. (ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 371-388.
- VINAY J.-P., DARBELNET J. (1958), *Stylistique comparée du français et de l'anglais*, Paris : Didier.
- WU D. (2000), Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars, J. Véronis (ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 139-167.