

Détection de relations d'appariement bilingue entre termes à partir d'une analyse syntaxique de corpus

Detection of bilingual alignment relations based on corpus syntactic analysis

Sylvia Ozdowska

Didier Bourigault

Equipe de Recherche en Syntaxe et Sémantique
Université Toulouse le Mirail
Maison de la Recherche
5, allées Antonio Machado
31058 Toulouse Cedex 1
{[@univ-tlse2.fr">ozdowska;didier.bourigault](mailto:ozdowska;didier.bourigault)}@univ-tlse2.fr

Résumé

Nous présentons une procédure d'appariement bilingue de mots (français/anglais) basée sur la propagation des liens d'appariement le long des relations de dépendance syntaxique. Nous procédons, dans un premier temps, à un appariement global qui repose sur le calcul de la fréquence de cooccurrence des termes sources et termes cibles dans des paragraphes alignés. Puis, à partir des résultats de l'appariement global, nous mettons en oeuvre un appariement local, de phrase à phrase, qui consiste à propager des liens d'appariement le long des relations syntaxiques identifiées par l'analyseur SYNTAX. Nous illustrons et évaluons la procédure sur la relation SUJET. Le taux de précision obtenu pour cette relation est de 91,7%.

Mots Clef

analyse syntaxique automatique, appariement de mots, corpus, traitement des langues naturelles.

Abstract

We present a word alignment procedure (French/English) based on the propagation of alignment links using syntactic dependency relations. First, we perform a "global" alignment by comparing the distribution of source and target terms in aligned paragraphs, i.e. by counting co-occurrences in these segments in respect with overall occurrences. Second, we rely on the results of the "global" alignment and on those provided by a parser, SYNTAX, to make a "local" alignment, from the source to the target sentence, which consists in propagating alignment links from nouns to verbs with the syntactic relation SUBJECT. This approach achieves a precision rate of 91,7%.

Keywords

parsing, word alignment, corpus, natural language processing.

1 Introduction

L'appariement est la mise en correspondance de mots ou expressions équivalentes dans des textes qui sont la traduction l'un de l'autre. Cette procédure représente un enjeu important, notamment pour la construction de ressources terminologiques multilingues. Divers systèmes visant à automatiser cette tâche ont vu le jour. Nombreux sont ceux qui privilégient l'utilisation de données statistiques dans leur stratégie de sélection, pour un terme source, du/des terme(s) cibles correspondants, plus rares ceux qui exploitent des données linguistiques.

L'objectif de cet article est de proposer une méthode d'appariement de mots et de syntagmes, appelés termes¹, qui s'appuie principalement sur des connaissances linguistiques et, plus précisément, sur les relations de dépendance syntaxique identifiées, pour les deux langues source et cible, par l'analyseur syntaxique de corpus SYNTAX.

2 Etat de l'art

Il existe globalement deux manières de procéder pour apparier des termes. Elles ont pour point commun de s'appuyer sur un alignement statistique au niveau des mots et d'opérer sur des corpus alignés au niveau des phrases ou bi-textes [16]. Dans un cas, il s'agit de repérer des termes simples et complexes dans les deux langues, source et cible, puis à les mettre en correspondance [1] [5]. Dans l'autre, seuls les termes de la langue source sont extraits, ceux de la langue cible étant mis en évidence par le processus d'appariement même [13] [18]. Le repérage des termes sources, et cibles le cas échéant, repose sur la reconnaissance de patrons syntaxiques plus ou moins contraints.

La sélection du bon appariement se fait soit par calcul d'une mesure d'association [12] [13], soit par

¹ Il s'agit plutôt, dans notre cas, de candidats termes, c'est-à-dire de mots et de syntagmes susceptibles d'acquérir le statut de termes spécifiques à un domaine donné.

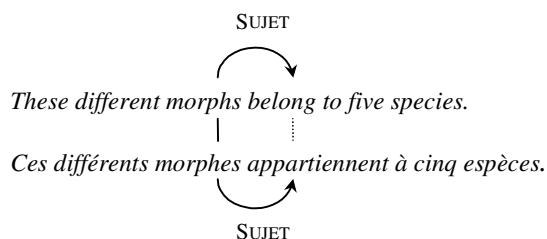
identification, en langue cible, d'une séquence de mots susceptibles de contenir la traduction du terme source [18].

A ces deux méthodes, il convient d'ajouter celle fondée sur une analyse parallèle des phrases source et cibles alignées et sur une identification simultanée du terme source et de son équivalent [24].

L'utilisation des connaissances linguistiques n'intervient, la plupart du temps, que lors de l'extraction des termes, à l'exception de [18] (dans une moindre mesure) et de [24]. Avec l'appariement syntaxique, nous proposons d'étendre l'utilisation de ce type de connaissances à la phase d'appariement à proprement parler. Ce choix est motivé par les deux principaux objectifs que nous poursuivons : a) parvenir à un appariement précis à un niveau de granularité fin, c'est-à-dire celui des termes simples et complexes, b) capter des appariements entre termes peu fréquents et/ou spécifiques au corpus.

3 Appariement syntaxique

Nous reprenons à notre compte l'hypothèse formulée par Debili et Zribi [6] selon laquelle « *les liaisons paradigmatisées peuvent aider à déterminer les relations syntagmatiques, et inversement* » et, plus particulièrement, 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 Ts_j et Tc_q sont appariés et s'il existe une relation de dépendance syntaxique entre Ts_j et Ts_i , d'une part, et entre Tc_q et Tc_p , d'autre part, alors Ts_i et Tc_p peuvent être appariés.



Dans cet article, nous présentons une implémentation de ce mécanisme de propagation des liens d'appariement suivant les relations de dépendance syntaxique, et évaluons ses résultats dans le cas de la relation syntaxique SUJET.

4 Environnement de travail

4.1 Corpus

Le corpus de travail sur lequel s'appuie notre étude a été constitué dans le cadre d'une expérience, menée à l'INRA², sur l'enrichissement de la base de données terminologiques alimentée et exploitée par les traducteurs

du service linguistique [9]³. Il s'agit d'un corpus bilingue de traduction, avec le français pour langue source et l'anglais pour langue cible, qui a été aligné automatiquement au niveau des paragraphes. Il compte environ 300 000 mots et comprend, par ordre décroissant d'importance, des articles de recherche, des articles de vulgarisation, un manuel d'utilisation de logiciel, des plaquettes de présentation, un contrat de licence, des résumés de monographies. Par ailleurs, il couvre différents domaines dont les plus représentatifs sont : l'agronomie, les sciences du sol, l'hydrobiologie, l'environnement, la biométrie et la modélisation, la génétique et l'amélioration des plantes, la pathologie végétale et la malherbiologie. Ce corpus présente donc une relative hétérogénéité tant du point de vue des thèmes abordés que des types de textes qui le composent.

4.2 Outil d'analyse

L'outil choisi pour le traitement du corpus est SYNTAX [2]. Il s'agit d'un analyseur syntaxique qui prend en entrée un corpus étiqueté et effectue une analyse en dépendances de chaque phrase. Il existe deux versions de cet outil : une pour le français et une pour l'anglais. SYNTAX prend en charge le repérage du/des sujet(s) et objet(s) des verbes (SUJET, OBJET)⁴, avec une distinction de la relation lorsque le recteur est un verbe d'état (ATTRIBUT, VERBE ATTRIBUT), le repérage des antécédents des pronoms relatifs (PROREL), ainsi que le rattachement des prépositions (PREP) et de leurs régis (NPREP), celui des adjectifs (ADJ), des noms épithètes (EPI), des adverbes (ADV) ou encore des déterminants (DET).

A la différence de [24], l'analyse est faite de manière indépendante pour chacune des deux langues, le traitement reste néanmoins homogène étant donné que les relations de dépendance syntaxique identifiées ainsi que leur représentation dans l'une et l'autre langue sont les mêmes. C'est là le principal avantage de cet outil pour une étude comme la nôtre qui s'appuie sur l'exploitation de corpus alignés.

A partir des résultats de l'analyse syntaxique, SYNTAX extrait un ensemble de mots et de syntagmes. Ces derniers sont organisés au sein d'un réseau où chacun d'entre eux est lié à sa tête et à son expansion. Comme dans [5], les mots et syntagmes sont extraits dans chacune des deux langues et servent de point de départ au processus d'appariement. Cependant, l'extraction n'est pas restreinte aux unités qui relèveraient de patrons morphosyntaxiques prédéfinis et encore moins à un type d'unités particulier, celui des syntagmes nominaux. Elle concerne aussi bien des termes simples tels que les noms, les verbes, les adjectifs, que des termes complexes comme les syntagmes nominaux, verbaux ou encore adjectivaux. Ainsi, nous

² Institut National de la Recherche Agronomique

³ Nous remercions A. Lacombe de l'INRA de nous avoir autorisés à utiliser ce corpus à des fins de recherche.

⁴ Nous indiquons entre parenthèses les noms que nous utiliserons pour faire référence à ces différentes relations dans la suite de l'article.

disposons de deux ensembles d'unités, l'un en langue source, l'autre en langue cible, suffisamment larges pour que l'appariement puisse avoir lieu même si les unités concernées n'ont pas la même catégorie grammaticale.

5 Processus d'appariement et architecture du système

Le processus d'appariement tel que nous le concevons comprend deux étapes (figure 1). La première consiste à appairer les termes sources (Ts) et termes cibles (Tc) extraits par SYNTAX en se basant sur leur fréquence d'apparition dans des paragraphes alignés ; on parlera dans ce cas d'appariement global au niveau du corpus. Cet appariement concerne les termes les plus fréquents, dont les équivalents peuvent être retrouvés avec une bonne précision par des méthodes statistiques. La seconde étape vise à retrouver les équivalents des termes peu fréquents. Elle s'appuie sur les résultats obtenus à l'étape précédente, et elle consiste à mettre en correspondance des Ts avec des Tc à un niveau local, c'est-à-dire phrase à phrase, à partir des relations de dépendance syntaxique identifiées par l'analyseur SYNTAX.

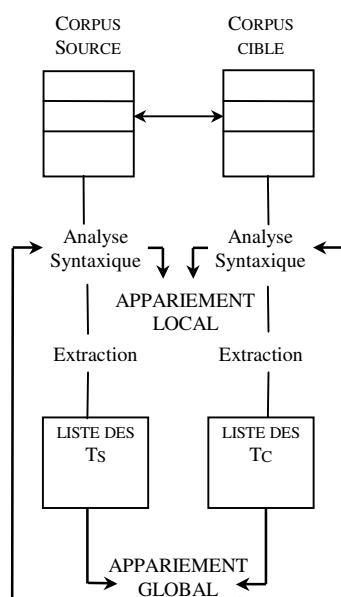


Figure 1 – Architecture du système

A terme, il s'agira de mettre en place un processus itératif où l'appariement global servira de base à l'appariement local qui permettra à son tour, d'une part, d'améliorer l'appariement global en éliminant les couples mis en correspondance par erreur et, d'autre part, de poursuivre la propagation des appariements à partir des unités mises en correspondance de manière locale.

6 Appariement global

6.1 Procédure d'appariement global

L'appariement global met en oeuvre une méthode largement utilisée, qu'il s'agisse de systèmes d'alignement de textes au niveau des phrases [19] ou au niveau des mots [1] [10] [20]. Cette méthode est celle du calcul des fréquences de cooccurrence des Ts et Tc dans des segments alignés d'un bi-texte. L'idée sous-jacente à ce type de calcul est la suivante : si un Ts et un Tc apparaissent souvent dans des segments alignés, alors ils ont de fortes chances d'être la traduction l'un de l'autre. Dans notre cas, puisque nous disposons d'un corpus aligné de manière fiable au niveau des paragraphes, il s'agit de comparer le nombre de fois où un Ts et un Tc donnés apparaissent ensemble dans des paragraphes alignés, c'est la fréquence de cooccurrence, par rapport à la fréquence, ou nombre d'occurrences, de chacun d'eux sur l'ensemble du corpus. Soient $\text{freq}(\text{Ts})$, la fréquence du Ts, $\text{freq}(\text{Tc})$, la fréquence du Tc et $\text{freq}(\text{Ts}, \text{Tc})$, la fréquence de cooccurrence. Pour chaque Ts et chaque Tc, nous calculons la mesure d'association suivante :

$$j(\text{Ts}, \text{Tc}) = \frac{\text{freq}(\text{Ts}, \text{Tc})}{\text{freq}(\text{Ts}) + \text{freq}(\text{Tc}) - \text{freq}(\text{Ts}, \text{Tc})}$$

Cette mesure (Jaccard) n'est pertinente que pour les Ts et Tc qui ont une fréquence suffisamment élevée, c'est pourquoi nous ne la calculons que lorsque cette dernière est supérieure ou égale à 5 sur l'ensemble du corpus.

Par ailleurs, pour un Ts donné, nous ne retenons dans cette expérience que les Tc pour lesquels $j(\text{Ts}, \text{Tc}) \geq 0,2$. L'ensemble des couples (Ts, Tc) ainsi obtenus forme ce que l'on appellera le lexique global.

Enfin, cet appariement est fait dans les deux sens, c'est-à-dire d'une part avec l'anglais comme langue source et le français comme langue cible et, d'autre part, avec le français comme langue source et l'anglais comme langue cible. En effet, si la mesure d'association est symétrique, il n'en est pas de même pour l'appariement, les Tc étant sélectionnés en fonction de la valeur de $j(\text{Ts}, \text{Tc})$. Par ailleurs, nous pensons que l'existence d'une « réciprocité » au niveau de l'appariement, c'est-à-dire d'un lien du Ts vers le Tc, d'une part, et, du Tc vers le Ts, d'autre part, peut constituer un indice qui confirme le lien d'équivalence entre les deux termes.

6.2 Résultats et évaluation

L'évaluation porte sur l'appariement global avec filtrage sur la seule valeur de $j(\text{Ts}, \text{Tc})$. Nous relevons tout d'abord les informations d'ordre général suivantes :

nombre total de Ts extraits	47230
nombre de Ts avec $\text{freq}(\text{Ts}) \geq 5$	3864
nombre de Tc avec $\text{freq}(\text{Tc}) \geq 5$	3951
nombre de Ts appariés	3219
nombre moyen de Tc par Ts	3,29
valeur moyenne de $j(\text{Ts}, \text{Tc})$	0,37

Tableau 1 - Appariement global

Le tableau 1 montre que seuls 3864 Ts ont une fréquence supérieure ou égale à 5. Par conséquent, seuls 8,20% des Ts extrait par SYNTEX font l'objet d'un appariement global. Sur ces 3864 Ts, 3219 sont appariés, soit 83,30%. Pour ce qui est de l'évaluation de l'appariement global à proprement parler (tableau 2), elle prend en compte le critère du rang des Tc, ces derniers étant classés par ordre décroissant de la valeur de $j(Ts, Tc)$.

nombre de cas validés 1151							
rang du Tc	1	2	3	4	5	10	15
%							
rappel	64,4	76,4	78,4	79,2	79,8	79,9	80
précision	73,3	44,5	35,1	30,7	28,5	24,4	23,4

Tableau 2 – Evaluation de l'appariement global

Les taux de rappel et de précision varient de manière inverse, l'un de 73,30% à 91,80% et l'autre de 23,40% à 73,30% selon le rang maximal n où on s'attend à trouver le bon Tc. Comme le montre le tableau 3, plus de 95% des Tc valides occupent le premier ou le second rang par ordre décroissant de la valeur de $j(Ts, Tc)$. Le rappel augmente donc considérablement quand l'on passe du rang 1 au rang 2 et, inversement, la précision baisse de manière importante.

rang du Tc	nombre de Tc	%
1	741	80,4
2	138	15
3	23	2,5
4	10	1,1
5	6	0,7
> 5	3	0,3

Tableau 3 – Rang des Tc corrects

Les résultats de l'appariement global servent de base à l'appariement local. Même si ces résultats sont satisfaisants, nous envisageons dans la section suivante un certain nombre de pistes pour améliorer leur précision.

6.3 Filtrage de l'appariement global

Comme nous venons de le voir, le filtrage du lexique global ne fait intervenir, pour le moment, qu'un seul paramètre, à savoir la valeur de $j(Ts, Tc)$. D'autres critères, aussi bien de type statistique que de nature linguistique, sont cependant susceptibles de rentrer en ligne de compte. Nous n'en citerons à titre d'exemple, que deux : celui de la catégorie grammaticale et celui de la « réciprocité » de l'appariement.

6.3.1 Catégorie grammaticale

Les termes sont actuellement appariés indépendamment de leur catégorie grammaticale. L'avantage d'un appariement non restrictif du point de vue de la catégorie est qu'il autorise la mise en correspondance, d'une part, de termes simples et de termes complexes, comme par

exemple *databases/bases de données* ou encore *pollution removal/dépollution*. D'autre part, l'appariement peut avoir lieu même lorsque les deux termes appartiennent à des catégories distinctes, on parle alors de transposition, phénomène relativement fréquent en traduction [4] [23]⁵. C'est le cas par exemple pour *calculate/calcul* :

Plant water requirements are calculated in several steps.
Le calcul des besoins en eau au niveau des plantes se fait en plusieurs temps.

Ou encore de *winter/hivernal* :

The hypothesis is that a cold winter climate favours the cycle involving sexual reproduction.
L'hypothèse est qu'un climat hivernal froid favorise le cycle faisant intervenir la production sexuée.

L'inconvénient d'un tel appariement est qu'il est susceptible de générer du bruit. Il nous semble par conséquent nécessaire de travailler à une solution intermédiaire. En effet, lors du passage d'un terme simple à un terme complexe, et vice versa, il y a maintien, pour le noyau du terme complexe, de la catégorie du terme simple. Par ailleurs, quand il y a transposition, elle ne peut se faire de n'importe quelle catégorie de départ vers n'importe quelle catégorie d'arrivée. Une transposition telle qu'un syntagme verbal est rendu par un adverbe semble très peu fréquente. C'est pourquoi il serait intéressant d'identifier, pour chaque catégorie de départ, quelles sont les catégories d'arrivée possibles.

6.3.2 Réciprocité de l'appariement

L'idée est de vérifier, pour chaque Tc proposé, qui devient dès lors le terme source (Ts'), quels sont ses équivalents (Tc'), et de ne garder que les Tc qui ont comme premier équivalent le Ts de départ (Tc'1 est identique à Ts). Prenons un exemple : pour le Ts *surface strength*, les quatre Tc proposés sont, dans l'ordre décroissant de la valeur de $j(Ts, Tc)$, : *mesurer la résistance*, *mesure de la résistance*, *résistance d'un sol* et *pénétrmètre*. Si l'on regarde, pour chacun de ces termes Ts', quels sont les Tc' qui leur sont associés, on s'aperçoit que le Ts' *résistance d'un sol* a pour seul et unique Tc' *surface strength* (Tc' identique à Ts). Les autres Ts' ont, respectivement, 5, 4 et 7 Tc' équivalents et *surface strength* occupe, respectivement, la troisième, quatrième et troisième rang dans la liste (Tc' différent de Ts). Le Tc *résistance d'un sol* sera donc le seul à être retenu. Grâce à cet « aller-retour » de la langue source vers la langue cible puis de la langue cible vers la langue source, nous sommes en mesure de proposer, pour le Ts *surface strength*, le seul Tc *résistance du sol*, alors que ce dernier n'occupe que la troisième position dans la liste des Tc. De même, pour le Ts *market gardening* quatre Tc sont proposés, toujours dans l'ordre décroissant de $j(Ts, Tc)$:

⁵ Voir notamment le cas des nominalisations en 7.3.

marâcher, haies, marâchages et paysages. Chacun de ces Ts' se voit attribuer, respectivement, 2, 4, 2 et 3 Tc'. Cependant, seuls les Ts' *marâcher* et *marâchage*, qui ont comme première traduction *market gardening*, seront retenus. Un retour aux phrases permet de confirmer, dans ce cas, la justesse du filtrage puisque *market gardening* est rendu par *marâcher* dans des constructions telles que *to ameliorate market gardening soils/pour l'amélioration des sols marâchers* et par *marâchage* dans des constructions du type *crops for market gardening/cultures tournées vers le marâchage*. Le critère de la « réciprocité » de l'appariement demande à être testé et évalué, notamment en observant les effets de son application sur les taux de rappel et de précision.

7 Appariement local

7.1 Algorithme d'appariement

La première opération à effectuer avant de procéder à l'appariement local consiste à projeter le lexique global au niveau local, c'est-à-dire phrase à phrase. Pour ce faire on vérifie, pour chaque couple de phrases appariées, quels sont les mots appartenant à ces phrases qui ont été appariés au niveau global. Dans la mesure où le découpage en phrases n'est pas tout à fait le même en anglais et en français⁶, nous ne disposons pas à ce niveau d'un alignement aussi fiable que l'alignement des paragraphes. C'est pourquoi, dès qu'un décalage apparaît, les phrases non alignées sont ignorées. Les résultats de la projection du lexique global aux phrases sont présentés dans le tableau 4.

nombre de phrases appariées	7055
nombre de mots appariés à partir du lexique	39852
nombre moyen de mots par phrase (anglais)	20
nombre moyen de mots par phrase (français)	22
nombre moyen de mots appariés par phrase	5,8

Tableau 1 – Projection de l'appariement global au niveau local

Afin de tester les performances d'un appariement local de mots par propagation des liens d'appariement le long des relations de dépendance syntaxique, nous avons choisi de travailler à partir de la relation SUJET et, plus précisément, à partir des noms qui sont régis par cette relation. Il en résulte que, dans le cas de la relation SUJET, le sens de la propagation des appariements est celui qui va du régi, le nom, vers le recteur, le verbe.

Soient Ns, un nom anglais dans une phrase source, et Nc, le nom français apparié à Ns dans la phrase cible. L'algorithme d'appariement local traite pour le moment deux types de cas : ceux où les deux noms Ns et Nc (souligné) sont en relation SUJET respectivement avec un

verbe Vs et un verbe Vc (gras). L'appariement des verbes se fait indépendamment de leur forme, active ou passive. :

*The fish **are** generally **caught** when they migrate from their feeding areas towards their spawning grounds.*

*Généralement les poissons **sont capturés** lorsqu'ils migrent de leur zone d'engraissement vers celles de reproduction.*

Ceux où Ns est en relation SUJET avec le verbe Vs et Nc en relation OBJET avec Vc. L'appariement se fait alors en fonction de la forme du Vs, avec la condition que ce dernier soit au passif :

*The predictor (1) **can then be constructed**.*

*On peut alors **construire** le prédicteur (1).*

Dans l'un comme dans l'autre de ces deux cas, nous parlerons de propagation « directe » dans la mesure où elle implique des relations de dépendance syntaxique qui sont les mêmes en langue source et en langue cible⁷.

7.2 Résultats et évaluation

La propagation des liens d'appariement basée sur la relation syntaxique SUJET/OBJET a donné lieu à l'appariement de 1591 couples de verbes⁸. Nous avons validé manuellement 649 cas afin de constituer une base de référence pour l'évaluation de l'appariement local. Ont été évalués comme corrects aussi bien des appariements où un verbe simple en langue source correspond à un verbe simple en langue cible, que des appariements où un verbe simple en langue source correspond à une locution verbale en langue cible, et inversement, l'appariement ne concernant que le seul noyau verbal de la locution :

[improve]
[conduire] à l'amélioration

[have] an influence
[influer]

[calculate]
[permettre] le calcul

En effet, nous considérons que, même s'il n'y a pas d'équivalence du point de vue du sens entre les verbes

⁷ Si l'on admet toutefois que les relation SUJET et OBJET sont équivalentes lorsque le verbe est passif dans un cas et actif dans l'autre.

⁸ Etant donné que nous cherchons à évaluer l'appariement par propagation, nous ne faisons, pour le moment, aucune distinction entre les verbes qui ont été appariés uniquement au niveau local et ceux qui ont par ailleurs été appariés entre eux au niveau global. Les chiffres suivants donnent une idée de la répartition : 60% des verbes ont été appariés uniquement au niveau local, grâce à la propagation des liens d'appariement à partir des régis, alors que 40% d'entre eux l'ont également été au niveau global.

⁶ Les signes de ponctuation ne reçoivent pas toujours la même interprétation dans l'une ou l'autre des deux langues. Il arrive donc que le découpage intervienne à des endroits différents.

ainsi appariés, il n'en reste pas moins que l'appariement est juste du point de vue de la propagation. Le tableau 5 présente les résultats de l'évaluation.

cas validés	649
succès	595
échecs	54
Précision (%)	91,70%

Tableau 2 – Evaluation de l'appariement local

Il est à noter que sur les 54 échecs relevés, 35 ont pour origine une erreur dans l'analyse syntaxique, 19 seulement viennent d'une erreur de l'algorithme d'appariement même. L'objectif de ce travail est d'une part d'obtenir une première évaluation de la précision de la technique d'appariement syntaxique, sur le cas de la relation SUJET, et d'autre part de repérer et d'analyser soigneusement les cas de non correspondance entre structure syntaxique, pour lesquels cette technique est silencieuse, pour mieux déterminer comment la généraliser. C'est pourquoi nous nous intéressons dans la section suivante à ces cas de non correspondance.

7.3 Typologie des cas de non correspondance

Le tableau 6 présente la répartition des Nc en termes de relations syntaxiques lorsque le Ns est régi par la relation SUJET. Il montre que dans 64,30% des cas le Nc est, lui aussi, régi par la relation SUJET (OBJET, lorsque le verbe source est au passif). Parmi les autres cas possibles, les plus représentées sont la relation NPREP⁹ (16,93%), aucune relation de dépendance (9,82%) et la relation ATTRIBUT (5,05%). Les relations OBJET (verbe source à l'actif), PROREL, et EPI représentent au total un peu moins de 4% des cas.

Ns régis par SUJET	2474
Cas de correspondance (traités par l'algorithme d'appariement local)	
Nc régis par SUJET	1512
Nc régis par OBJET (verbe source passif)	79
Cas de non correspondance	
Nc régis par NPREP	419
Nc non régis	243
Nc régis par ATTRIBUT	125
Nc régis par OBJET (verbe source actif)	52
Nc régis par PROREL	30
Nc régis par EPI	21
Nc régis par VERBE ATTRIBUT	2

Tableau 3 – Relations syntaxiques des noms cibles

Nous avons passé en revue un certain nombre de cas où le Nc n'est pas SUJET/OBJET (avec passif) et ce avec l'idée d'établir une typologie de ces cas afin de mettre en évidence des équivalences entre la langue source et la langue cible, cette fois-ci en termes de relations

syntactiques. En effet, nous pensons que la propagation des liens d'appariement peut se faire non seulement de manière directe, comme nous venons de le voir ci-dessus, mais aussi de manière indirecte, c'est-à-dire suivant des relations de dépendance syntaxique qui ne sont pas les mêmes en langue source et en langue cible.

i) RELATION NPREP. Nous avons observé 232 cas, soit environ 55% de l'ensemble. Dans près de 57% des cas observés, la présence de la relation NPREP est due à une différence de formulation entre la langue source et la langue cible. Cette différence relève de l'un des quatre types suivants :

- nominalisation (37% des cas) : le verbe source est rendu par un nom, recteur de la préposition qui régit à son tour le nom cible ;

genes are regulated
la régulation des gènes

- ajout d'un élément nominal en français (23% des cas), recteur de la préposition qui régit le nom cible. Il peut s'agir soit d'un ajout pur et simple, soit d'une transposition. Dans le premier cas, l'élément nominal n'a pas de correspondant en langue source :

a team has shown
les travaux d'une équipe ont montré

Dans le deuxième cas, l'élément nominal a un équivalent en langue source mais qui est de nature et/ou de fonction différente :

pollution arises partly
une partie de cette pollution provient

Qu'il soit question d'ajout ou de transposition, l'introduction de l'élément nominal ne modifie pas fondamentalement la structure de la proposition. De plus, grâce à la relation PREP, il est possible de remonter jusqu'au recteur de la préposition dont dépend le Nc qui est, par ailleurs, sujet du verbe que l'on cherche à aligner.

- passage d'une construction active à une construction passive, le sujet de la langue source devient complément d'agent en langue cible ; passage d'une construction passive à une construction active en *on*, le sujet de la langue source devient complément de verbe en langue cible (18% des cas) ;

the researchers used a scanner
un scanner a été utilisé par les chercheurs

⁹ Les noms des types de relations sont définis en 4.2.

*threshold is referred to
on se réfère au seuil*

- transformations qui affectent fondamentalement la structure de la proposition et/ou de la phrase et qui ne présentent aucune régularité (22% des cas) ;

*our results showed
à partir de nos résultats, nous avons montré*

*the characteristic originates from
il faut considérer deux origines pour le caractère*

Les reformulations relevant du dernier type mises à part, toutes les autres présentent un caractère suffisamment régulier pour que l'on puisse envisager une propagation indirecte des liens d'appariement.

Dans les 43% des cas où il n'y a pas de reformulation, la présence de la relation NPREP vient soit d'une erreur de l'analyse syntaxique (26%), soit d'une erreur d'appariement des noms source et cible (10%) ou encore d'une erreur d'alignement au niveau des phrases (6%).

ii) AUCUNE RELATION DE DEPENDANCE SYNTAXIQUE. Nous avons observé 139 cas, soit près de 57% de l'ensemble. Dans 70% des cas observés, l'absence de relation syntaxique vient d'une erreur de l'analyse syntaxique en langue source (29%) ou en langue cible (41%). Le premier type d'erreur a deux sortes de répercussions sur l'appariement local. Soit l'absence d'une quelconque relation de dépendance syntaxique en langue cible empêche uniquement de produire des appariements bruyants. Soit elle crée en plus du silence, les bons recteurs n'étant identifiés ni dans l'un ni dans l'autre des deux langues.

Le deuxième type d'erreur, quant à lui, est à l'origine du silence. En effet, l'appariement ne peut avoir lieu puisque le Nc, qui est bien sujet du verbe cible que l'on cherche à appairer, n'a pas été reconnu comme tel.

Pour les 30% restants, 23% des cas relèvent de la reformulation, 6% d'une erreur d'appariement des noms source et cible et 1% d'une erreur d'alignement des phrases.

iii) RELATION ATTRIBUT : nous avons observé 80 cas sur 125, ce qui représente un taux d'environ 70%. La relation ATTRIBUT ne lie pas directement le Nc à un verbe mais à son attribut qui, lui, est rattaché au verbe d'état par la relation VERBE ATTRIBUT. Combinées, ces deux relations peuvent être assimilées, dans près de 87% des cas, à une relation SUJET et donner lieu à la propagation des liens d'appariement.

*the intensity varies
l'intensité est variable*

Dans les autres cas, la présence de la relation ATTRIBUT a pour origine une erreur dans l'analyse syntaxique (7%), dans l'appariement des noms source et cible (1%), dans l'alignement des phrases (2%) ou bien une reformulation (3%).

iv) RELATION OBJET (verbe source actif) : nous avons observé l'ensemble des cas. Dans 49% d'entre eux, les verbes source et cible sont actifs. On trouve parmi ces cas des constructions du type suivant : *il existe un plateau*. Le Nc reconnu comme objet du verbe est en fait un sujet.

Dans 49% des cas, il s'agit d'une erreur dans l'analyse syntaxique et, dans les 8% restants, d'une erreur d'appariement au niveau des phrases.

v) RELATION PROREL : nous avons observé 22 cas, soit près de 73% de l'ensemble. Dans 32% d'entre eux, la relation PROREL sert d'intermédiaire entre le nom cible sujet et le verbe recteur et peut par conséquent permettre de propager les liens d'appariement de manière indirecte.

*this society comprises
cette société qui regroupe*

Dans 32% des cas restants il y a erreur d'analyse, 5% des cas correspondent à une erreur dans l'alignement des phrases et enfin 31% à une reformulation.

vi) RELATION EPI : sur l'ensemble des cas, 47% viennent d'une erreur d'analyse, 43% d'une erreur dans l'appariement des mots et 10% d'une erreur dans l'appariement des phrases. On peut donc en conclure que la relation EPI ne peut déboucher sur une propagation indirecte des liens d'appariement.

vii) RELATION VERBE ATTRIBUT : dans les deux cas cette relation apparaît comme un cas particulier de la relation OBJET lorsque le verbe source est à la forme passive et le verbe source est un verbe d'état. Cette relation, comme la relation ATTRIBUT, permet donc de propager les appariements des noms vers leurs verbes recteurs.

8 Discussion

Avec un taux de précision de 91,7%, l'appariement local par propagation offre un résultat comparable, voire supérieur, à ceux obtenus par Daille *et al.* [5], précision variant entre 70 et 80% selon la taille de la liste des candidats, Gaussier [13], qui fait état d'une précision allant de 90 à 98% suivant que le nombre de meilleures associations prises en compte est 500 ou 100. Wu [24], quant à lui, estime ce taux à 81,5% et enfin Hull [18] qui, privilégiant le taux de rappel, se contente d'une précision ne dépassant pas les 56%.

Les premiers résultats obtenus par une technique de propagation de lien d'appariement le long de relations syntaxiques sont prometteurs, et nous encouragent à poursuivre dans cette voie. L'analyse détaillée des cas où ce principe est mis en défaut est aussi extrêmement

enrichissante, à la fois d'un point de vue linguistique, sur la problématique de la variation syntaxique interlingue, et du point de vue de l'implémentation, pour identifier comment étendre le principe de la propagation syntaxique de l'appariement pour diminuer le silence. Le travail présenté ici se poursuit par l'affinement de l'algorithme de propagation le long de la relation SUJET, et par son extension à d'autres relations syntaxiques. Par ailleurs, notre réflexion porte sur l'optimisation du partage des tâches entre l'alignement global, efficace et relativement précis sur les termes fréquents dans le corpus, et l'alignement local qui s'appuie sur une projection dans les phrases du résultat de l'alignement global pour détecter des appariements entre termes peu fréquents. Elle porte aussi sur l'utilisation éventuelle de ressources exogènes, notamment de dictionnaires électroniques, susceptibles de compléter le lexique global construit lors de la phase d'alignement global.

Bibliographie

- [1] L. Ahrenberg, M. Andersson, M. Merkel, 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, 2000.
- [2] D. Bourigault, C. Fabre, Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, pp. 131-151, Université Toulouse le Mirail, 2000.
- [3] P. Brown, S. Della Pietra, R. Mercer, The mathematics of statistical machine translation : parameter estimation, *Computational Linguistics*, 19(2), pp. 263-311, 1993.
- [4] H. Chuquet, M. Paillard, *Approche linguistique des problèmes de traduction anglais/français*, Ophrys, 1989.
- [5] B. Daille, E. Gaussier, J.-M. Langé, Towards Automatic Extraction of Monolingual and Bilingual Terminology, *Proceedings of the International Conference on Computational Linguistics (COLING'94)*, pp. 515-521, 1994.
- [6] F. Debili, A. Zribi, Les dépendances syntaxiques au service de l'appariement des mots, *Actes du 10^{ème} Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*, 1996.
- [7] F. Debili, L'appariement : quels problèmes ?, *Actes des 1^{ères} JST 1997 FRANCIL de L'AUPELF-UREF*, 1997.
- [8] H. Déjean, E. Gaussier, Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables, *Lexicometrica*, numéro spécial *Alignement lexical dans les corpus multilingues*, 2002.
- [9] C. Frérot, G. Rigou, A. Lacombe, Approche phraséologique d'une extraction automatique de terminologie dans un corpus scientifique bilingue aligné, *Actes des 4^{èmes} rencontres Terminologie et Intelligence Artificielle*, Nancy, pp 180-188, 2001.
- [10] W. A. Gale, K. W. Church, Identifying Word Correspondences in Parallel Text, *Proceedings of the DARPA Workshop on Speech and Natural Language*, 1991.
- [11] W. A. Gale, K. W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19(3), pp. 75-102, 1993.
- [12] E. Gaussier, *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues*, Thèse de doctorat, Paris VII, 1995.
- [13] E. Gaussier, 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, 1998.
- [14] E. Gaussier, General considerations on bilingual terminology extraction, D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, John Benjamins, pp. 167-183, 2001.
- [15] E. Gaussier, D. Hull, S. Aït-Mokhtar, Term alignment in use, J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 253-274, 2000.
- [16] B. Harris, Bi-text, A New Concept in Translation Theory, *Language Monthly*, 54, pp.8-10, 1988.
- [17] D. A. Hull, A Practical Approach to Terminology Alignment, *Proceedings of the First Workshop on Computational Terminology, COLING-ACL'98*, pp. 1-7, 1998.
- [18] D. A. Hull, 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, 2001.
- [19] M. Kay, M. Röscheisen, Text-Translation Alignment, *Computational Linguistics*, 19(1), pp. 121-142, 1993.
- [20] P. van der Eijk, Automating the Acquisition of Bilingual Terminology, *Proceedings of the EACL'93*, pp. 113-119, 1993.
- [21] J. Véronis, Alignement de corpus multilingues, J.-M. Pierrel (Ed.), *Ingénierie des langues*, Paris, Editions Hermès, pp. 115-150, 2000.
- [22] J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Parallel Corpora*, Dordrecht : Kluwer Academic Publishers, 2000.
- [23] J.-P. Vinay, J. Darbelnet, *Stylistique comparée du français et de l'anglais*, Paris, Didier, 1958.
- [24] D. Wu, 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, 2000.