Inférence de règles de propagation syntaxique pour l'alignement de mots

Sylwia Ozdowska* — **Vincent Claveau****

* *CLLE/ERSS - CNRS & Université de Toulouse Le Mirail*
5, allées Antonio Machado
31058 Toulouse cedex 1
Ozdowska@univ-tlse2.fr

** *IRISA - CNRS*
Campus de Beaulieu
35042 Rennes cedex
Vincent.Claveau@irisa.fr

RÉSUMÉ. *Cet article présente et évalue une approche originale d'alignement automatique de bitexte au niveau des mots. Pour cela, elle tire parti d'une analyse syntaxique en dépendances et utilise une technique d'apprentissage artificiel, la programmation logique inductive, pour apprendre des règles dites de propagation. Celles-ci s'appuient sur les informations syntaxiques connues pour aligner les mots avec grande précision. La méthode est entièrement automatique et ne requiert que peu de données d'entraînement ; les résultats présentés montrent qu'elle se compare aux meilleures techniques existantes. Enfin, l'examen des règles inférées permet d'identifier facilement les cas d'isomorphismes syntaxiques entre les deux langues traitées.*

ABSTRACT. *This paper presents and evaluates an original approach to automatically align bitexts at the word level. It relies on a syntactic dependency analysis of the texts and uses a machine-learning technique, namely inductive logic programming, to automatically infer rules called propagation rules. These rules make the most of the syntactic information to precisely align words. This approach is entirely automatic, uses very few training data, and its results rival the ones of the best existing alignment systems. Moreover, syntactic isomorphisms between the two spotted languages are easily identified through the inferred rules.*

MOTS-CLÉS : *alignement de mots, apprentissage artificiel, programmation logique inductive, analyse syntaxique*

KEYWORDS: *word alignment, machine learning, inductive logic programming, syntactic analysis*

TAL. Volume 47 – n° 1/2006, pages 167 à 186

1. Introduction

L'enjeu que représente l'alignement des corpus parallèles au niveau des mots n'est plus à démontrer : ce dernier trouve ses applications dans des tâches telles que la traduction automatique ou encore la construction de ressources lexicales bi- ou multilingues (Véronis, 2000a). Il existe principalement deux types d'approches pour aligner des mots : celles à dominante statistique qui s'appuient notamment sur les modèles IBM (Brown *et al.*, 1993), et celles qui tendent à combiner calculs statistiques simples et différentes sources d'information linguistique (Ahrenberg *et al.*, 2000 ; Barbu, 2004).

Pivots de l'approche statistique, les travaux de Brown *et al.* (1993) ont mené à l'établissement d'une famille de modèles de traduction statistiques — les modèles IBM. Ceux-ci sont de différentes complexités et procèdent sans connaissances linguistiques. Les probabilités de traduction entre phrases sont estimées en fonction des probabilités de correspondance entre mots obtenues via leur alignement dans les bi-phrases d'un corpus d'entraînement. L'outil d'alignement GIZA++[1] (Och *et al.*, 2003) implémente notamment ce type d'approche. Au fil des modèles IBM, de nouveaux paramètres sont introduits qui rendent compte de certains phénomènes linguistiques :

– modèle 1 : tous les alignements sont considérés comme équiprobables ;

– modèle 2 : les alignements dépendent de la position des mots sources et cibles ;

– modèle 3 : un mot source peut être rendu par plusieurs mots cibles ;

– modèles 4 et 5 : il existe une cohésion à l'intérieur de certains groupes de mots ; les alignements ne doivent pas être choisis indépendemment les uns des autres.

Afin de mieux traiter ce type de phénomènes et, plus généralement, d'autres variations systématiques entre les langues impliquées dans le processus de traduction (Dorr, 1994 ; Fox, 2002), les systèmes statistiques destinés à la traduction automatique se sont progressivement enrichis en incorporant des données structurelles. Des approches statistiques à composante syntaxique ont ainsi vu le jour. L'objectif est de contraindre les alignements possibles à l'aide de l'analyse syntaxique, qu'il s'agisse de constituants ou de dépendances. Pour ce faire, deux techniques sont utilisées : *tree-to-string* ou *tree-to-tree*.

Dans le premier cas, l'analyse est effectuée uniquement dans la langue source. Ainsi, Lin *et al.* (2003) définissent une contrainte de cohésion d'un alignement A au regard d'une analyse en dépendances de la phrase source T_E. Il y a cohésion si les dépendances induites en langue cible à partir de A et de T_E ne se croisent pas. Les alignements qui ne respectent pas cette condition sont éliminés. Dans le second cas, des analyses syntaxiques sont menées dans chacune des langues ; c'est le cas de l'approche que nous présentons dans cet article. Par exemple, Ding *et al.* (2003) partent de l'idée qu'un alignement partiel entre un arbre de dépendances source et un arbre de dépendances cible limite les connexions possibles entre les nœuds qui restent à

1. GIZA++ est disponible à l'adresse http://www.jfoch.com/GIZA++.html.

aligner. Les auteurs mettent ainsi au point une méthode basée sur le partitionnement itératif des arbres qui consiste à trouver les alignements les plus sûrs entre nœuds et à découper les arbres en fonction de ces alignements.

À cela vient s'ajouter l'approche proposée par Wu (2000) où l'alignement n'est pas effectué en fonction d'une analyse syntaxique préalablement établie mais y est incorporé. L'auteur utilise le formalisme ITG (*Inversion Transduction Grammars*) qui sert à décrire des grammaires bilingues permettant de produire une analyse simultanée de paires de phrases alignées par application récursive de règles syntaxiques et lexicales. Les règles syntaxiques sont de la forme $X \longrightarrow [X_1 X_2]$ ou $X \longrightarrow < X_1 X_2 >$. Elles précisent l'ordre des constituants dans les deux langues : les crochets ([]) indiquent que le même ordre est conservé d'une langue à l'autre tandis que les chevrons (<>) signalent qu'il est inversé, ce qui permet de mettre en correspondance des structures telles que Adj Nom en anglais et Nom Adj en français, par exemple. Les règles lexicales sont de la forme $X \longrightarrow x /\!/ y$ où $x /\!/ y$ est une paire mot source/mot cible. Dans la version stochastique des ITG (SITG pour *Stochastic Inversion Transduction Grammars*), une probabilité qui peut être estimée automatiquement est associée à chaque règle. Dans ce cadre, l'alignement au niveau des mots et des syntagmes est contraint par les possibilités de réordonnancement des constituants représentées dans la grammaire. Il faut noter qu'il ne s'agit pas d'une grammaire motivée d'un point de vue linguistique. Les « syntagmes » qui font l'objet d'un alignement ne correspondent pas nécessairement à la définition traditionnelle pour ce type d'unités ; il s'agit davantage de segments de phrase comme « *are about 3.5 million* » ou « *job opportunities ; and last month* ».

En dehors du paradigme statistique, la démarche proposée par Ozdowska (2004), dont nous reprenons le cadre expérimental, a consisté à définir manuellement des règles d'alignement, dites de propagation, qui exploitent les relations de dépendance mises en évidence dans chaque partie d'un corpus parallèle. Cet article présente une technique d'alignement proche de cette dernière. Cependant, l'originalité de notre démarche réside dans le fait que les règles de propagation sont acquises de manière automatique en corpus par une technique d'apprentissage artificiel, la programmation logique inductive. Ces règles de propagation, exploitant des informations syntaxiques issues des analyseurs SYNTEX, sont automatiquement inférées à partir d'exemples d'alignements valides. L'objectif de cet article est d'une part de montrer que, contrairement aux approches statistiques, notre technique ne nécessite que très peu de données d'apprentissage. D'autre part, on se propose de vérifier si les règles obtenues et les alignements qu'elles produisent varient en fonction du type de corpus d'apprentissage.

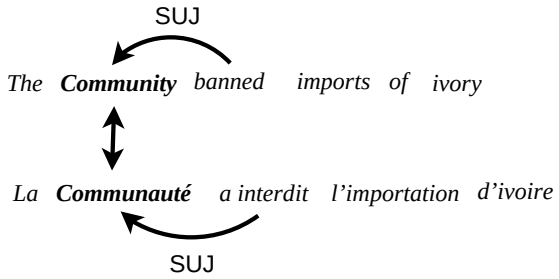
Pour ce faire, nous exposons tout d'abord le cadre méthodologique dans lequel nous avons mené nos travaux. Puis, nous décrivons la technique d'apprentissage automatique des règles de propagation en section 3. Enfin, nous présentons et discutons les résultats obtenus en section 4 avant d'indiquer les perspectives de poursuite de ce travail.

2. Contexte d'expérimentation

Dans cette section, nous présentons tout d'abord le principe de l'alignement par propagation syntaxique. Nous décrivons ensuite en sous-section 2.2 les différentes données d'apprentissage et d'évaluation utilisées dans nos expériences. Nous terminons en présentant brièvement les analyseurs syntaxiques SYNTEX utilisés pour étiqueter les textes français et anglais.

2.1. *Alignement de mots par propagation syntaxique*

L'utilisation de règles de propagation pour aligner des bitextes au niveau des mots a déjà fait l'objet de plusieurs travaux. Ainsi, Ozdowska (2004) exploite les relations de dépendance syntaxique dans le processus d'alignement. Elle utilise des règles de propagation syntaxique définies à la main qui, étant donnés deux mots en relation d'équivalence dans un couple de phrases alignées, appelés *couple amorce*, permettent de propager le lien d'alignement à d'autres mots en suivant les relations de dépendance syntaxique connues. Dans l'exemple suivant, il est ainsi possible d'aligner *ban* et *interdire* en exploitant la relation sujet portant sur le couple amorce *Community // Communauté*. Dans cet exemple et les suivants, les couples amorces sont notés en gras et les dépendances sont fléchées du recteur vers le régi et surmontées du nom de la relation syntaxique.



Chaque règle de propagation est donc décrite en fonction de la relation syntaxique qui sert de base à la propagation et de la direction dans laquelle s'effectue la propagation (et éventuellement des restrictions portant sur les parties du discours des mots concernés). La propagation peut donc se faire dans le sens de la dépendance (du recteur au régi) ou dans le sens contraire (du régi au recteur). Si nous reprenons l'exemple précédent, la règle de propagation anglais/français utilisée est :

$$\mathrm{V} \xrightarrow{\text{suj}} \textbf{Nom} \;//\; \mathrm{V} \xrightarrow{\text{suj}} \textbf{Nom}$$

Elle indique que la propagation se fait à partir d'un couple amorce de noms régis (*Community // Communauté)* vers un couple de verbes recteurs (*ban // interdire*) par

la relation sujet. Une autre règle de propagation possible est celle qui va du couple amorce de noms régis (*ivory // ivoire*) au couple de recteurs (*imports // importation*) par la relation de préposition :

$$\text{Nom} \xrightarrow{\text{prep}} \textbf{Nom} \;/\!/\; \text{Nom} \xrightarrow{\text{prep}} \textbf{Nom}$$

La plupart des règles utilisées dans ce type d'approche ont été définies en accord avec l'hypothèse d'isomorphisme direct entre les langues selon laquelle les structures syntaxiques seraient conservées lors de la traduction, comme dans l'exemple précédent (Hwa *et al.*, 2002). Cependant quelques-unes traitent des cas de non-isomorphisme, comme l'alignement de *tax* et *fiscales* dans la biphrase : *tax* ***expenditures*** *have been [...] // les* ***dépenses*** *fiscales demeurent [...]*. Si l'on part du couple amorce de noms recteurs (*expenditures // dépenses*), les structures syntaxiques qui se correspondent dans les deux langues sont (nn représente la dépendance entre deux noms et adj la dépendance entre un adjectif et un nom) :

$$\textbf{Nom} \xrightarrow{\text{nn}} \text{Nom} \;/\!/\; \textbf{Nom} \xrightarrow{\text{adj}} \text{Adj}$$

Ce type d'approche permet d'obtenir des alignements qui offrent en général une bonne précision, le rappel se révélant cependant de moins bonne qualité (voir en section 4.1 pour une définition formelle du rappel et de la précision dans ce cadre). Le principe d'isomorphisme permet en effet de générer des alignements corrects dans la plupart des cas où il s'applique mais il semble, dans certains cas, trop contraignant. Par ailleurs, ces approches nécessitent une expertise humaine pour écrire les règles de propagation, ce qui peut se révéler coûteux. C'est ce dernier point que nous proposons de contourner en utilisant une technique d'apprentissage artificiel pour inférer automatiquement des règles de propagation.

2.2. *Données d'apprentissage et d'évaluation*

Nous avons choisi comme données de référence celles mises à disposition dans le cadre d'une campagne d'évaluation des systèmes d'alignement au niveau des mots notamment pour la paire de langues anglais/français (Mihalcea *et al.*, 2003). En voici la description (Och *et al.*, 2000) :

– corpus d'entraînement anglais/français, issu du HANSARD (débats parlementaires canadiens), comptant 1,3 million de biphrases. Pour les expériences que nous reportons en section 4, nous n'avons utilisé qu'une portion variant de 10 à 1 000 couples de phrases alignées de ce corpus ;

– corpus de test constitué de 447 phrases alignées extraites d'une partie différente du HANSARD ;

– le jeu de référence contient les alignements effectués par deux annotateurs sur le corpus de test. Chaque lien d'appariement établi s'est vu attribuer la valeur S, s'il

s'agissait d'un lien considéré comme non ambigu, ou P dans le cas contraire. La valeur P est choisie en présence d'expressions figées ou de traductions libres. Dans le cas où celles-ci comprennent plusieurs mots sources et/ou cibles, leur mise en correspondance s'effectue par énumération de l'ensemble des liens individuels entre chaque mot source et chaque mot cible. Dans le jeu de référence final, la valeur S est conservée pour les alignements pour lesquels il y a accord inter-annotateurs sur S ; la valeur P est attribuée dans tous les autres cas. La figure 1 présente un exemple de phrase annotée ; les alignements S sont en traits pleins et les P en pointillés. Les expérimentations décrites en section 4 présentent les résultats pour ces deux types d'alignements, à l'exception de ceux concernant les signes de ponctuation.

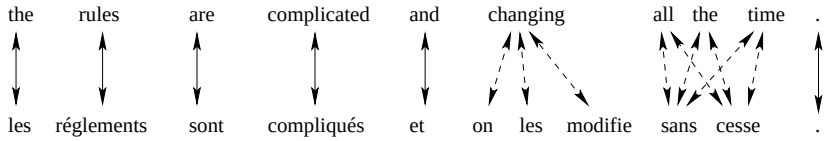


Figure 1. *Annotation pour la campagne d'alignement HLT*

En plus du HANSARD, les expériences d'inférence de règles de propagation que nous présentons en section 4 sont effectuées sur deux autres corpus. Le premier est un extrait du corpus INRA[2]. Il s'agit d'un corpus spécialisé anglais/français du domaine de la recherche agronomique dont nous avons prélevé 1 000 biphrases. Le second est un corpus fourni dans le cadre de la campagne d'évaluation ARCADE I (Véronis *et al.*, 2000). Il est constitué de questions-réponses traitées à la Commission Européenne. Là encore, nous n'avons retenu que 1 000 biphrases.

2.3. *Étiquetage syntaxique*

Le repérage des relations de dépendance syntaxique dans les trois corpus d'entraînement est effectué indépendamment pour chacune des deux langues par les analyseurs SYNTEX français et anglais (Bourigault *et al.*, 2000). Ces derniers prennent en entrée un texte étiqueté et identifient, pour chaque phrase, des relations telles que sujet, objet direct et indirect, modifieur (adjectival, adverbial), etc. Le tableau 1 résume quelques-unes des principales relations mises en évidence. Les deux outils sont conçus suivant la même architecture et mettent en œuvre les mêmes procédures de repérage des relations de dépendance. Par ailleurs, les relations identifiées ainsi que leur représentation sont globalement les mêmes d'une langue à l'autre, à l'exception de la relation nn de nature quelque peu différente en anglais et en français.

Le tableau 2 montre la précision et le rappel des deux analyseurs en fonction des principales relations syntaxiques identifiées (Ozdowska, 2006). Les performances ont

2. Nous remercions A. Lacombe, INRA, pour nous avoir permis d'utiliser ce corpus.

Relation :	sujet (suj)
Recteur/Régi :	verbe/nom, pronom
Exemple :	(the) Commission $\xleftarrow{\text{suj}}$ decides // (la) Commission $\xleftarrow{\text{suj}}$ détermine
Relation :	attribut du sujet (att)
Recteur/Régi :	verbe/nom, adjectif
Exemple :	(the Commission) is $\xrightarrow{\text{att}}$ aware // (la Commission) est $\xrightarrow{\text{att}}$ consciente
Relation :	objet direct (obj)
Recteur/Régi :	verbe/nom, pronom, verbe
Exemple :	minimize $\xrightarrow{\text{obj}}$ (the) losses // minimiser $\xrightarrow{\text{obj}}$ (les) pertes
Relation :	préposition (prep)
Recteur/Régi :	nom, adjectif, verbe/préposition
Exemple :	(the) effects $\xrightarrow{\text{prep}}$ of (this stress) // (les) effets $\xrightarrow{\text{prep}}$ de (ces tensions)
Relation :	complément prépositionnel (cprep)
Recteur/Régi :	préposition/nom, verbe
Exemple :	(the effects) of $\xrightarrow{\text{cprep}}$ (this) stress // (les effets) de $\xrightarrow{\text{cprep}}$ (ces) tensions
Relation :	adjectif épithète (adj)
Recteur/Régi :	nom/adjectif
Exemple :	(the) financial $\xleftarrow{\text{adj}}$ losses // (les) pertes $\xrightarrow{\text{adj}}$ financières
Relation :	nom épithète (nn)
Recteur/Régi :	nom/nom
Exemple :	(the) Member $\xleftarrow{\text{nn}}$ States // (les) États $\xrightarrow{\text{nn}}$ Membres
Relation :	adverbe (adv)
Recteur/Régi :	verbe, adjectif, adverbe/adverbe
Exemple :	(the debates) clearly $\xleftarrow{\text{adv}}$ illustrate // (les débats) montrent $\xrightarrow{\text{adv}}$ clairement

Tableau 1. *Exemples de dépendances syntaxiques repérées par* SYNTEX

été mesurées sur des échantillons du corpus HANSARD. Les différences sont globalement de l'ordre de 5 à 10 points, à l'exception de la relation suj pour laquelle elles sont plus importantes[3], aussi bien en précision qu'en rappel, et ce en général à l'avantage de l'analyseur pour le français. On peut s'attendre à ce que les taux d'erreur non négligeables aient des répercussions sur les performances de l'alignement syntaxique.

3. Alignement par apprentissage artificiel

Comme nous l'avons déjà dit, l'originalité de notre approche tient au fait que contrairement aux travaux précédemment exposés (Ozdowska, 2004), les règles de propagation syntaxique ne sont pas données manuellement mais inférées automati-

3. Si on ne tient pas compte de la relation nn dont le fonctionnement et la fréquence sont très différents en anglais et en français.

	Précision		Rappel	
	fr	en	fr	en
suj	94 %	88 %	91 %	76 %
att	87 %	94 %	98 %	94 %
obj	92 %	83 %	91 %	84 %
prep	91 %	86 %	90 %	81 %
adj	96 %	88 %	97 %	87 %
nn	85 %	68 %	77 %	86 %

Tableau 2. *Performances de* SYNTEX *sur le corpus* HANSARD

quement. Les deux sous-sections suivantes présentent la technique d'apprentissage artificiel et son utilisation pour inférer ces règles. La méthode d'amorçage fournissant automatiquement les exemples nécessaires à cette technique supervisée est décrite en sous-section 3.3.

3.1. *Programmation logique inductive*

Le principe de notre approche est le suivant : à partir d'exemples de propagations valides au sein de deux phrases alignées, on tente d'apprendre des règles qui définissent ces propagations. Pour ce faire, nous nous plaçons dans un cadre d'apprentissage artificiel supervisé (se reporter à l'ouvrage de Cornuéjols *et al.* (2002) pour une présentation détaillée de l'apprentissage artificiel). Parmi les différentes techniques existantes, c'est la programmation logique inductive ou PLI que nous utilisons pour inférer les règles de propagation. Une présentation approfondie de la PLI est proposée par Muggleton *et al.* (1994) ; nous n'en donnons ci-dessous que les grandes lignes. La PLI permet d'inférer des règles générales (des clauses de Horn généralement exprimées en Prolog) décrivant un concept à partir d'un jeu d'exemples de ce concept E^+ (avec éventuellement des contre-exemples E^-) et un ensemble d'informations externes B, appelées *Background Knowledge*. L'ensemble de règles inférées, appelé classifieur et noté H par la suite, est obtenu en généralisant les exemples en utilisant les informations contenues dans B.

Quelques conditions imposées à cette tâche d'apprentissage forment le cadre logique de la PLI. Les deux suivantes portent sur les données ($\Box$ signifie faux et $\models$ représente l'implication logique) :

– la consistance *a priori* assure que les exemples négatifs n'entrent pas en contradiction avec le *Background Knowledge*, soit $B \wedge E^- \not\models \Box$;

– la nécessité *a priori* traduit le besoin de connaissances additionnelles, de règles à inférer, pour expliquer les exemples, soit $B \not\models E^+$.

À cela s'ajoutent les deux conditions portant sur l'ensemble de règles H que l'on

recherche :

– la consistance *a posteriori* impose qu'aucune contradiction n'existe entre B, H et E^- : $B \wedge H \wedge E^- \not\models \square$;

– la complétude assure que tous les exemples positifs sont expliqués avec H et les informations du *Background Knowledge*, soit $B \wedge H \models E^+$.

En pratique, les règles composant H sont recherchées à travers un espace d'hypothèses regroupant toutes les règles possibles. Cet espace est organisé hiérarchiquement, ce qui permet de le parcourir efficacement. Une règle de cet espace est retenue si elle maximise un score, généralement défini en fonction du nombre d'exemples (et éventuellement de contre-exemples) qu'elle couvre.

La PLI, de par son expressivité (exemples et règles sont exprimés en logique des prédicats), a été utilisée pour de nombreuses tâches d'apprentissage, et notamment en TAL (Cussens *et al.*, 2000 ; Claveau, 2003, *inter alia*). Cette technique d'apprentissage artificiel a été choisie pour notre tâche pour deux raisons liées à ce pouvoir expressif. Tout d'abord, l'encodage des relations syntaxiques (les dépendances) et des relations de traduction (les amorces) se fait très naturellement avec des prédicats (voir ci-après pour des exemples). Par ailleurs, le classifieur inféré, c'est-à-dire l'ensemble de règles H, est parfaitement interprétable et se prête donc aisément à une analyse linguistique, conférant à cette approche une plus-value certaine par rapport à d'autres techniques d'apprentissage.

3.2. *Apprentissage de règles de propagation*

Dans notre cas, les règles recherchées sont des propagations syntaxiques et les exemples que nous utilisons sont des phrases alignées analysées syntaxiquement comportant des alignements valides ; nous n'utilisons pas de contre-exemples. L'algorithme de PLI que nous utilisons est ALEPH[4]. Dans B sont stockées en Prolog toutes les informations concernant les dépendances syntaxiques entre les mots des phrases exemples et les couples amorces connus. Comme nous l'avons précisé auparavant, le formalisme logique de la PLI permet d'encoder facilement ces informations relationnelles. Ainsi, si l'on sait que *companies // entreprises* peuvent être alignés dans l'extrait de biphrase suivant (l'identifiant de chaque mot est noté après les barres obliques) :

... private/id_1_en ***sector****/id_2_en companies/id_3_en*
... les/id_1_fr entreprises/id_2_fr du/id_3_fr ***secteur****/id_4_fr privé/id_5_fr*

4. ALEPH est développé par A. Srinivasan et disponible à http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph.

On ajoute à E^+ :

alignement(id_3_en,id_2_fr).

et à B (le nom du prédicat représente le nom de la relation syntaxique tel que défini dans le tableau 1, section 2.3 ; les deux arguments sont ici des constantes : le premier argument représente l'identifiant du recteur et le second celui du régi) :

det(id_2_fr,id_1_fr).
prep(id_2_fr,id_3_fr).
adj(id_2_en,id_1_en).
cprep(id_3_fr,id_4_fr).
adj(id_4_fr,id_5_fr).
nn(id_3_en,id_2_en).
amorce(id_2_en,id_4_fr).

Une règle qui peut être inférée à partir de cet exemple est (les arguments sont cette fois-ci des variables pouvant représenter n'importe quel identifiant de mot) :

alignement(M_en,M_fr) :- nn(M_en,A1), prep(M_fr,F1), cprep(F1,F2), amorce(A1,F2).

Avec les notations précédentes, cette règle s'écrit :

$$\text{M\_en} \xrightarrow{\text{nn}} \mathbf{A1} \;/\!/\; \text{M\_fr} \xrightarrow{\text{prep}} \text{F1} \xrightarrow{\text{cprep}} \mathbf{F2}.$$

Elle souligne l'équivalence bien connue des structures Nom Nom en anglais avec Nom de Nom en français. Tout couple apparaissant dans une biphrase avec cette structure peut ainsi être aligné.

En l'absence de contre-exemples, il est important de s'assurer que cette technique ne surgénéralise pas les exemples, c'est-à-dire que les règles inférées ne couvrent que des cas d'alignement corrects. Pour ce faire, pour qu'une règle inférée soit retenue dans H, on impose qu'elle couvre un nombre minimum d'exemples présents dans E^+ ; on parle de seuil de couverture. Par ailleurs, un autre mécanisme est utilisé pour rejeter éventuellement des règles qui pourraient être inappropriées ; il s'agit des biais de langage. Ces biais, qui font partie intégrante de tout système d'apprentissage (Mitchell, 1980), s'assurent de la bonne formation des règles. Dans notre cas, une règle est dite bien formée si elle contient un couple amorce et qu'un chemin de relations syntaxiques relie bien ce couple amorce au couple à aligner. Seules les règles obéissant à ces critères de couverture et de bonne formation sont donc finalement proposées par notre algorithme de PLI et utilisées par la suite pour l'alignement.

3.3. *Amorçage*

Des exemples d'alignements valides sont nécessaires à notre technique d'apprentissage. Cette phase de supervision, pénible si elle était conduite manuellement, est

dans notre cas automatisée par une technique dite d'amorçage. Notre approche d'inférence de règles de propagation ne requiert donc finalement aucune intervention humaine ; elle est dite semi-supervisée.

Pour générer ces alignements exemples, ou couples amorces, nous utilisons deux approches complémentaires. Il s'agit d'une part d'une technique statistique classique et d'autre part d'une recherche de cognats. En ce qui concerne la méthode statistique, nous considérons comme couples amorces les paires de mots (anglais/français) apparaissant ensemble dans des phrases alignées de manière statistiquement significative (Ahrenberg *et al.*, 2000) ; la force du lien entre deux mots est calculée par un Jaccard sur les fréquences d'apparition conjointe des deux mots (Ozdowska, 2004). Pour le repérage de cognats, c'est-à-dire de chaînes de caractères identiques ou proches dans les deux langues, la méthode mise au point est similaire à celle décrite par Fluhr *et al.* (2000). Elle consiste à identifier la sous-chaîne maximale commune à deux mots qui cooccurrent dans un couple de phrases alignées.

Par la conjonction de ces deux méthodes, ce sont ainsi en moyenne entre 4 et 6 couples amorces par phrase qui sont détectés, selon les corpus. À titre d'exemple, environ 15 % des couples amorces se révèlent erronés (c'est-à-dire des mots ne devant pas être alignés) en ce qui concerne le corpus HANSARD. Chaque couple amorce, allié aux deux phrases alignées analysées syntaxiquement dont il est tiré, peut ainsi servir d'exemple pour notre technique d'apprentissage de règles de propagation. Une phrase permet donc de produire autant d'exemples qu'elle comporte de couples amorces.

3.4. *Inférence et utilisation des règles de propagation*

À partir des exemples obtenus par la technique décrite ci-avant, il nous est donc possible d'inférer des règles de propagation. Ces règles peuvent ensuite être appliquées à de nouvelles données dans lesquelles on aura préalablement repéré des couples amorces.

Deux questions nous intéressent particulièrement. Il s'agit d'une part de savoir quelle est la taille minimale du corpus d'entraînement qui nous est nécessaire pour mener une inférence de règles efficace. Pour estimer celle-ci, on mène plusieurs phases d'apprentissage en faisant varier le nombre de phrases de ce corpus, et donc le nombre d'exemples.

Le second point que nous souhaitons explorer est celui de la généricité des règles inférées. En effet, on peut se demander si ces règles de propagation sont propres au corpus traité, ou au contraire valables quel que soit le domaine du bitexte. Pour répondre à cette question, nous proposons trois expériences :

1) les règles sont inférées à partir du corpus HANSARD[5] ;

2) les règles sont inférées à partir du corpus ARCADE ;

5. Il s'agit bien entendu d'une partie distincte de celle du jeu de test présenté en section 4.2.

3) les règles sont inférées à partir du corpus INRA.

Pour ces différents types d'expériences, les règles sont ensuite évaluées grâce au jeu de test HLT qui nous permet de nous comparer aux systèmes existants.

4. Résultats

Cette section présente tout d'abord la méthodologie d'évaluation et les résultats obtenus par notre approche sur le jeu d'évaluation HLT. Nous examinons ensuite en section 4.3 l'influence de la taille du corpus d'apprentissage. Nous décrivons en section 4.4 quelques causes d'erreurs récurrentes, puis en section 4.5 nous analysons certaines des règles inférées. Nous nous intéressons enfin à la généricité des règles inférées en comparant les résultats obtenus à partir des trois corpus d'entraînement.

4.1. *Méthodologie d'évaluation*

Dans un premier temps, nous évaluons notre approche à l'aide des données de la campagne HLT. Les règles de propagation utilisées ont été inférées sur le corpus HANSARD. Les mesures de performances utilisées sont le taux de rappel, le taux de précision et la f-mesure qui se définissent respectivement dans notre cadre par :

$$R = \frac{\text{nb d'alignements corrects trouvés}}{\text{nb d'alignements attendus}}$$

$$P = \frac{\text{nb d'alignements corrects trouvés}}{\text{nb d'alignements trouvés}}$$

$$F = \frac{2PR}{P+R}$$

À des fins de comparaison, nous indiquons ci-après les résultats obtenus par les meilleurs systèmes d'alignement — en terme de f-mesure — ayant participé à la compétition HLT : il s'agit de Ralign (Simard *et al.*, 2003), XRCE (Déjean *et al.*, 2003), BiBr (Zhao *et al.*, 2003) et ProAlign (Lin *et al.*, 2003). D'autres systèmes très performants, notamment celui présenté par Taskar *et al.* (2005) ne sont pas présentés ici. Ces derniers utilisent en effet une mesure de performances unique différente des nôtres (l'AER - *Alignment Error Rate*) qui ne permet pas d'analyser finement le comportement des systèmes d'alignement. Dans notre cas, en particulier, l'AER apparaît comme une mesure trop globale pour permettre un examen détaillé de l'efficacité des règles inférées par notre méthode.

Les systèmes Ralign, XRCE, BiBr et ProAlign auxquels nous nous comparons utilisent principalement des approches statistiques, notamment celles dérivées des modèles IBM qui peuvent être entraînées grâce à GIZA++. Ils exploitent éventuellement différents types de contraintes : contrainte de compositionnalité pour Ralign et

contraintes syntaxiques pour BiBr et ProAlign. Les trois premiers concouraient dans la catégorie « limitée », à savoir celle des systèmes n'utilisant pas de ressources autres que le corpus d'entraînement fourni dans le cadre de la campagne[6]. Les contraintes syntaxiques de BiBr étant de type ITG, elles ne requièrent pas d'analyse préalable à l'alignement comme nous l'avons vu en introduction. Le quatrième système, ProAlign, concourait dans la catégorie « illimitée », à savoir celle des systèmes utilisant des ressources externes complémentaires. En effet, comme nous l'avons également mentionné en introduction, la contrainte de cohésion syntaxique que ce système met en œuvre requiert que la partie source du bitexte soit analysée syntaxiquement. Nous indiquons aussi les résultats du système ALIBI (Ozdowska, 2006) dans lequel les règles de propagation sont définies manuellement. Si on s'en réfère à la campagne HLT, les deux systèmes PLI et ALIBI relèveraient de la catégorie « illimitée » eu égard à l'utilisation de SYNTEX.

4.2. *Performances d'alignement*

Le tableau 3 présente les résultats des systèmes cités ci-dessus et de notre approche, référée en tant que système PLI, pour la tâche d'alignement S (alignements non ambigus). Pour ces expériences, la phase d'apprentissage par PLI a été menée sur 1 000 biphrases du corpus HANSARD (non comprises dans le jeu de test).

Système	PLI	ALIBI	Ralign	XRCE	BiBr	ProAlign
Précision	82,11 %	**88,78 %**	72,54 %	55,54 %	63,03 %	71,94 %
Rappel	74,09 %	66,86 %	80,61 %	**93,46 %**	74,59 %	91,48 %
F-mesure	77,89 %	76,28 %	76,36 %	69,68 %	68,32 %	**80,54** %

Tableau 3. *Performances du système d'alignement par PLI sur les données HLT (alignements S)*

À l'examen de ce tableau, on remarque que les performances globales de notre système sont de niveau comparable à celles des autres. Ce dernier se classerait en effet deuxième en terme de f-mesure, derrière le système ProAlign. Le système PLI jouit d'une précision supérieure aux autres systèmes, à l'exception d'ALIBI. On peut supposer que cette précision est imputable au fait que l'analyse syntaxique est effectuée sur les deux parties du bitexte, source et cible, ce qui renforce probablement la contrainte de cohésion et limite d'autant plus les alignements à ceux qui sont pertinents syntaxiquement parlant. Le rappel est, quant à lui, relativement plus bas par rapport à celui des participants à la campagne HLT, qu'il s'agisse de ceux de la catégorie « illimitée » ou de ceux de la catégorie « limitée ». Ce rappel s'explique en partie par l'insuffisance de couples amorces et la couverture imparfaite de l'étiquetage syntaxique, ce qui rend

6. BiBr et XRCE concouraient dans les catégories « limitée » et « illimitée », les résultats que l'on montre sont ceux de la catégorie « limitée » où ils obtiennent les meilleures performances.

certains couples inaccessibles à nos règles de propagation puisqu'il n'existe pas de chemin de relations syntaxiques pour aller d'un couple amorce à ces couples.

Bien que les systèmes PLI et ALIBI soient plus spécifiquement conçus pour effectuer des alignements non ambigus (alignements S), nous présentons dans le tableau 4 les résultats obtenus sur la tâche d'alignement P. Comme précédemment, nous indiquons à titre de comparaison les résultats des autres systèmes.

Système	PLI	ALIBI	Ralign	XRCE	BiBr	ProAlign
Précision	80,65 %	90,81 %	77,56 %	89,65 %	66,11 %	**96,49 %**
Rappel	28,44 %	22,49 %	**38,19 %**	34,92 %	30,06 %	28,41 %
F-mesure	42,05 %	36,05 %	**51,18 %**	50,27 %	41,33 %	43,89 %

Tableau 4. *Performances du système d'alignement par PLI sur les données HLT (alignements P)*

Plusieurs choses peuvent être notées à partir de ces chiffres. Tout d'abord, les résultats du système PLI sont meilleurs en terme de f-mesure que ceux de son pendant manuel, ALIBI, ce dernier étant handicapé par un faible rappel. Cela montre que les règles inférées couvrent plus de cas que les règles manuelles utilisées dans ALIBI mais au détriment de quelques points de précision. Le rappel du système PLI est néanmoins trop faible pour concurrencer les autres systèmes. Cela n'a rien de surprenant puisque l'emploi de contraintes syntaxiques n'est pas adapté pour produire des alignements ambigus de type P où des déterminants peuvent se trouver alignés avec des verbes, par exemple. Le système ProAlign, utilisant lui aussi des contraintes syntaxiques, souffre d'ailleurs du même problème comparativement aux trois autres systèmes à fonctionnement plus purement statistique. Par ailleurs, la décomposition des correspondances impliquant des segments de plusieurs mots en alignements individuels de type P pose problème. En effet, il n'est pas certain qu'il s'agisse d'une solution adéquate pour rendre compte des performances, plus particulièrement pour fournir un rappel qui soit représentatif (Ozdowska, 2006). En effet, cette décomposition introduit une surgénération importante des alignements de référence — une correspondance telle que *all the time // sans cesse* est représentée par 6 alignements — par rapport aux alignements qui peuvent être effectivement produits par les systèmes.

4.3. *Variation de la taille du corpus d'entraînement*

On s'intéresse dans un second temps à l'évolution des performances selon la taille des corpus d'entraînement. Pour cela, on fait varier le nombre de phrases servant à produire les exemples pour l'apprentissage. La figure 2 présente les taux de rappel, de précision et la f-mesure obtenus selon le nombre de phrases à partir du corpus HANSARD. Les résultats sont très éloquents : il n'y a quasiment aucune variation de rappel et de précision de 300 à 1 000 phrases. En dessous de 300 phrases, la précision augmente sensiblement alors que le rappel décroît. Cela s'explique par le fait que seules

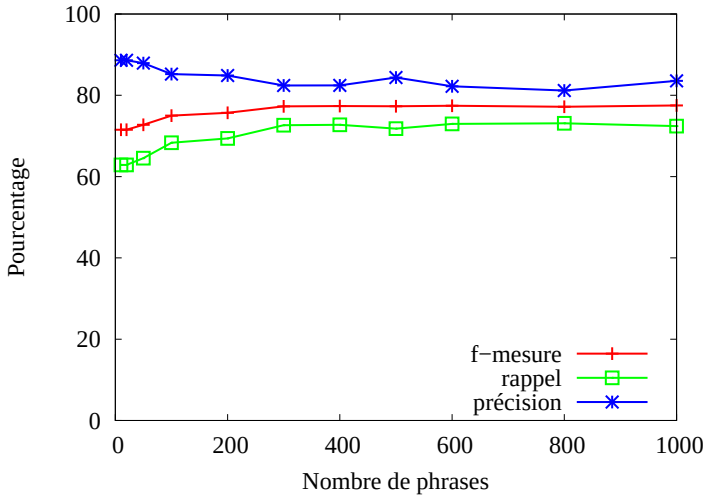


Figure 2. *Variation des performances selon le nombre de phrases utilisées à l'apprentissage*

quelques règles de propagation, parmi les plus sûres, sont trouvées. On remarque enfin qu'avec 10 phrases seulement, notre algorithme d'apprentissage est capable d'inférer des règles suffisamment pertinentes pour mener à une f-mesure de 70 %. Ces résultats sont donc très positifs, notamment en regard des tailles très restreintes de nos corpus d'entraînement. À titre de comparaison, les systèmes Ralign, XRCE, BiBr et ProAlign utilisent 1,3 million de phrases pour s'entraîner.

4.4. *Examen des résultats*

Les erreurs d'alignement les plus courantes faites par nos systèmes peuvent se classer en plusieurs grandes catégories. Comme nous l'avons dit précédemment, une grande part des faux négatifs (*i.e.* des alignements non détectés) est due à une trop faible densité de couples amorces en plus d'absences de dépendances au sein de certaines phrases.

En ce qui concerne les faux positifs (*i.e.* des alignements détectés à tort), certains viennent simplement d'erreurs d'étiquetage de SYNTEX (elles-mêmes parfois causées par des erreurs de l'étiqueteur catégoriel utilisé en amont). Par exemple, dans la biphrase *federal government carpenters get $ 6.42 // Les menuisiers du gouvernement fédéral touchent $ 6.42*, l'adjectif *federal* a incorrectement été rattaché à *carpenters*, ce qui a provoqué l'alignement incorrect de *carpenter // gouvernement*, tous deux notés recteurs du couple amorce *federal // fédéral*.

D'autres erreurs de ce type sont causées par certaines des règles inférées qui ne sont pas assez spécifiques pour éviter de ramener du bruit. C'est notamment le cas

des règles manipulant les dépendances objet ou sujet qui, à cause du manque d'informations dont dispose l'algorithme d'apprentissage, ne font pas de différence entre les voix actives et passives. Ainsi, à partir du couple amorce *bring ‖ apporter* dans la biphrase *good legislation has been brought in by Liberal governments ‖ les gouvernements libéraux ont apporté de bonnes mesures législatives*, *gouvernement* et *legislation* ont été alignés à tort.

Enfin, des phénomènes de reformulations plus ou moins fidèles lors de la traduction perturbent parfois nos tentatives d'alignement. Ainsi, dans la phrase *the Government must implement the recommendations of the Commissioner of Official Languages ‖ le gouvernement se doit de respecter les recommandations du Commissaire aux langues officielles*, *implement* et *respecter* ont été alignés alors que ce couple n'est pas noté valide dans le jeu de test HLT.

4.5. *Règles obtenues*

À partir de 1 000 phrases d'entraînement, ce sont 81 règles qui sont inférées sur le corpus INRA, 82 sur le corpus JOC et 65 sur le corpus HANSARD pour un seuil de couverture égal ou supérieur à 2 (*cf.* section 3.2). Le nombre de règles est de 67, 62 et 54 respectivement pour chacun des trois corpus si on prend en compte uniquement celles dont la couverture est égale ou supérieure à 10, qui sont celles effectivement utilisées dans nos expériences d'alignement. Dans les deux cas, le nombre de règles communes aux trois corpus est d'environ 40. Par conséquent, pratiquement toutes les règles communes satisfont au seuil de couverture requis pour être appliquées aux données à aligner. Par ailleurs, pour ce seuil de couverture et pour un corpus donné, le nombre de règles spécifiques, à savoir celles trouvées uniquement sur le corpus en question, est de 21 pour INRA et de 13 pour JOC ainsi que pour HANSARD. Parmi les règles inférées, on retrouve une bonne partie de celles définies manuellement par Ozdowska (2006).

La plupart des règles mettent en exergue des isomorphismes connus entre la syntaxe anglaise et française, comme l'alignement des adjectifs modifiant deux noms alignés, ou l'alignement des compléments d'objet direct de deux verbes alignés :

```
alignement(M_en,M_fr) :- adj(C,M_en), adj(D,M_fr), amorce(C,D).
alignement(M_en,M_fr) :- obj(C,M_en), obj(D,M_fr), amorce(C,D).
```

Ces cas d'isomorphismes parfaits représentent près de 50 % des règles de propagation. Certains cas de non-isomorphisme syntaxique sont également trouvés, comme par exemple la construction standard des syntagmes nominaux Nom Nom en anglais et Nom de Nom en français (*cf.* section 3.2). D'autres types de non-isomorphismes peuvent même mener à l'alignement de mots ayant des parties du discours différentes, comme par exemple des noms et des adjectifs :

```
alignement(M_en,M_fr) :- nn(C,M_en), adj(D,M_fr), amorce(C,D).
```

D'une manière générale, il ressort de l'examen de ces règles que la plupart d'entre elles sont des règles de propagation que l'on peut qualifier de génériques. Leur similarité avec celles trouvées manuellement par Ozdowska (2006) confirme la validité de notre processus d'apprentissage. Cependant quelques règles inférées sont plus inattendues – et leur validité peut être discutée – comme par exemple :

alignement(M_en,M_fr) :- adj(M_fr,C), nn(D,M_en), adj(D,E), amorce(E,C).

Cette règle permet d'aligner *bargaining* et *négociation* dans la biphrase *[...] to have some hang-up with regard to the **collective** bargaining process // [...] éprouver certains complexes à l'égard de la négociation **collective***.

4.6. *Généricité des règles de propagation*

Comme nous l'indiquions précédemment, il est intéressant d'examiner si les règles de propagation que nous inférons sont propres à un corpus d'apprentissage ou non. Pour avoir une vue quantitative de cette généricité, nous comparons tout d'abord les résultats d'alignement des trois jeux de règles inférés à partir des corpus HANSARD, ARCADE et INRA sur le même jeu de données que précédemment, celui de la campagne HLT. Ici encore, la phase d'apprentissage a été menée sur 1 000 phrases tirées de chacun des 3 corpus. Le tableau 5 présente les résultats des trois systèmes.

Corpus d'entraînement	HANSARD	ARCADE	INRA
Précision	82,08 %	80,65 %	**83,16** %
Rappel	74,09 %	**74,10 %**	66,90 %
F-mesure	**77,88 %**	77,20 %	74,15 %

Tableau 5. *Performances du système PLI selon le corpus d'entraînement*

Il est intéressant de noter que les trois jeux de règles de propagation obtiennent des résultats relativement similaires. Bien que les taux de précision soient très proches dans les trois cas, ce sont les règles issues du corpus INRA qui donnent les alignements les plus précis. À l'inverse, ce sont aussi ces mêmes règles qui offrent le rappel le plus bas, le plus élevé étant obtenu par les règles inférées sur le corpus JOC. Sans surprise, les meilleures performances en terme de f-mesure sont obtenues avec les règles issues du corpus similaire à celui du jeu de test (mais correspondant à une partie différente de celui-ci), à savoir le corpus HANSARD.

D'un point de vue plus qualitatif, on observe que le nombre de règles inférées à partir de chaque corpus est relativement proche (soit environ une soixantaine pour 1 000 biphrases d'entraînement et pour une couverture supérieure ou égale à 10). Il y a peu de différences entre ces règles dans les trois corpus : la proportion de règles communes représente entre 2/3 et 3/4 de l'ensemble des règles inférées à partir de chaque corpus. Par conséquent, seules quelques règles de propagation diffèrent d'un

corpus à l'autre, dont certaines sont erronées à cause d'erreurs d'étiquetage syntaxique présentes dans les données d'apprentissage, ce qui explique la proximité des performances observées dans le tableau précédent[7].

5. Conclusion et perspectives

Nous avons présenté une méthode originale d'alignement de mots basée sur la syntaxe et sur une technique d'apprentissage semi-supervisée. Celle-ci permet d'apprendre automatiquement des règles de propagation à partir d'exemples de couples de mots alignés. Ces exemples sont par ailleurs fournis à l'aide d'une procédure d'amorçage qui confère à notre approche une complète autonomie. Les résultats d'alignement obtenus par cette approche entièrement symbolique sont bons et comparables aux meilleurs systèmes d'alignement actuels. De plus, et c'est l'originalité de ce travail, contrairement aux systèmes existants, très peu de données sont nécessaires pour entraîner notre système qui reste dans le même temps entièrement automatique. Une plus-value intéressante de cette approche réside dans le fait que les règles d'alignement inférées permettent une analyse des cas d'isomorphismes et de non-isomorphismes entre les deux langues traitées. Enfin, on a montré également que les règles de propagation sont relativement génériques et changent peu d'un bitexte à un autre pour l'alignement du français et de l'anglais.

Plusieurs perspectives sont ouvertes par ce travail. Étant donnée l'entière autonomie de notre approche, il pourrait être intéressant de mener des expérimentations avec d'autres analyseurs syntaxiques que SYNTEX afin de vérifier dans quelle mesure son efficacité dépend de l'outillage mis en œuvre. Concernant la technique d'apprentissage, nous prévoyons d'intégrer les informations catégorielles pour permettre d'inférer des règles ne portant plus seulement sur les dépendances syntaxiques mais aussi sur les parties du discours. Cela permettra d'éviter certaines fausses détections reportées précédemment. Également dans le but d'améliorer la qualité des règles inférées, nous envisageons d'utiliser des exemples négatifs lors de la phase d'apprentissage. Ces exemples négatifs doivent être des couples de mots dont on sait qu'ils ne sont pas traduction l'un de l'autre dans des biphrases. Ils devraient permettre d'empêcher des généralisations excessives, et donc des règles de propagation pas assez précises. Néanmoins, il est nécessaire que la génération de ces exemples négatifs, à l'instar des exemples positifs, se fasse sans intervention humaine pour préserver à notre approche son entière autonomie.

D'un point de vue applicatif, notre méthode étant entièrement automatique, elle peut aisément être adaptée à d'autres paires de langues. Il faut bien sûr que celles-ci soient suffisamment proches d'un point de vue syntaxique, ou plus précisément que les correspondances entre les formulations syntaxiques soient suffisament régulières pour pouvoir être apprises par apprentissage artificiel. Une autre condition qui

7. Ozdowska (2006) propose une analyse détaillée des différences entre les règles inférées sur chaque corpus à laquelle le lecteur intéressé peut se reporter.

peut se révéler déterminante pour certaines langues est la disponibilité d'analyseurs en dépendances pour chacune des deux langues traitées. Des expériences dans ce sens permettraient pour ces langues d'intéressantes études des cas d'isomorphismes et de non-isomorphismes syntaxiques dans les phrases alignées à travers l'étude des règles de propagation inférées.

6. Bibliographie

Ahrenberg L., Andersson M., Merkel M., « A knowledge-lite approach to word alignment », *in* Véronis (2000b), chapter 5, 2000.

Barbu A. M., « Simple linguistic methods for improving a word alignment algorithm », *Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data, JADT'04*, Louvain-la-Neuve, Belgique, 2004.

Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaire*, vol. 25, p. 131-151, 2000. Université Toulouse le Mirail.

Brown P. F., Pietra S. A. D., Pietra V. J. D., Mercer R. L., « The mathematics of statistical machine translation : parameter estimation », *Computational Linguistics*, vol. 19, n° 2, p. 263-311, 1993.

Claveau V., Acquisition automatique de lexiques sémantiques pour la recherhe d'information, Thèse de doctorat, Université de Rennes 1, Rennes, France, décembre, 2003.

Cornuéjols A., Miclet L., *Apprentissage artificiel*, Eyrolles, 2002.

Cussens J., Džeroski S., (eds), *Learning Language in Logic*, vol. 1925 of *Lecture Notes in Artificial Intelligence*, Sringer Verlag, June, 2000.

Ding Y., Gildea D., Palmer M., « An Algorithm for Word-Level Alignment of Parallel Dependency Trees », *Proceedings of the 9th Machine Translation Summit of the International Association of Machine Translation*, New Orleans, LA, États-Unis, 2003.

Dorr B., « Machine translation divergences : A formal description and proposed solution », *Computational Linguistics*, vol. 20, n° 4, p. 597-633, 1994.

Déjean H., Gaussier E., Goutte C., Yamada K., « Reducing parameter space for word alignment », *Proceedings of the HLT-NAACL 2003 Workshop Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada, p. 23-26, mai, 2003.

Fluhr C., Bisson F., Elkateb F., « Parallel Text Alignment Using Crosslingual Information Retrieval Techniques », *in* Véronis (2000b), chapter 9, 2000.

Fox H. J., « Phrasal cohesion and statistical machine translation », *Proceedings of Empirical Methods in Natural Language Processing, EMNLP'02*, Philadelphia, PA, États-Unis, p. 304-311, 2002.

Hwa R., Resnik P., Weinberg A., Kolak O., « Evaluating Translational Correspondence Using Annotation Projection », *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics*, Philadelphia, PA, États-Unis, 2002.

Lin D., Cherry C., « ProAlign : Shared Task System Description », *Proceedings of the HLT-NAACL 2003 Workshop Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada, 2003.

Mihalcea R., Pederson T., « An evaluation exercise for word alignment », *Proceedings of the HLT-NAACL 2003 Workshop Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada, p. 1-10, 2003.

Mitchell T. M., « The Need for Biases in Learning Generalizations », *Readings in Machine Learning*, Morgan Kaufmann, p. 184-191, 1980. Publié en 1991.

Muggleton S., De Raedt L., « Inductive Logic Programming : Theory and Methods », *Journal of Logic Programming*, vol. 19-20, p. 629-679, 1994.

Och F. J., Ney H., « Improved statistical alignment models », *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, Hong Kong, p. 440-447, octobre, 2000.

Och F. J., Ney H., « A Systematic Comparison of Various Statisical Alignment Models », *Computational Linguistics*, vol. 1, n° 29, p. 19-51, 2003.

Ozdowska S., « Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés », *Actes de la conférence RECITAL'04*, Fès, Maroc, 2004.

Ozdowska S., ALIBI, un système d'ALIgnement BIlingue à base de règles de propagation syntaxique, Thèse de doctorat, Université de Toulouse 2, Toulouse, France, décembre, 2006.

Simard M., Langlais P., « Statistical translation alignment with compositionality constraints », *Proceedings of the HLT-NAACL 2003 Workshop Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada, p. 19-22, mai, 2003.

Taskar B., Lacoste-Julien S., Klein D., « A Discriminative Matching Approach to Word Alignment », *Proceedings of Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada, p. 73-80, 2005.

Véronis J., « Alignement de corpus multilingues », *Ingénierie des langues*, Jean-Marie Pierrel (ed.), Hermès, Paris, chapter 6, p. 151-171, 2000a.

Véronis J., *Parallel Text Processing*, Kluwer Academic Publishers, Dordrecht, 2000b.

Véronis J., Langlais P., *Evaluation of Parallel Text Alignment Systems. The ARCADE Project*, in Véronis (2000b), chapter 19, 2000.

Wu D., *Bracketing and Aligning Words and Constituents in Parallel Text using Stochastic Inversion Transduction Grammars*, in Véronis (2000b), chapter 7, 2000.

Zhao B., Vogel S., « Word alignment based on bilingual bracketing », *Proceedings of the HLT-NAACL 2003 Workshop Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, Edmonton, Alberta, Canada, p. 15-18, mai, 2003.