

Université de Montréal

**Projection d'un analyseur grammatical via alignement
bilingue de mots**

Par

Ziad Khairallah

Département d'Informatique et de Recherche Opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc)
en informatique

Avril, 2005

© Ziad Khairallah, 2005

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé :

**Projection d'un analyseur grammatical via alignement
bilingue de mots**

Présenté par :

Ziad Khairallah

a été évalué par un jury composé des personnes suivantes :

.....
(Président - rapporteur)

Philippe Langlais

.....
(Directeur de recherche)

.....
(Membre du jury)

Mémoire accepté le2005

Résumé

Des analyseurs syntaxiques (parseurs) de qualité n'existent que pour un petit nombre de langues actuellement parlées dans le monde. Ceci s'explique par le fait que jusqu'à récemment, le développement d'analyseurs syntaxiques nécessitait l'intervention massive et récurrente d'experts linguistes. Mettre au point un tel analyseur était alors une entreprise scientifique s'étalant sur plusieurs années. Depuis peu, la disponibilité de corpus annotés syntaxiquement (par des linguistes) rend possible le développement rapide d'analyseurs syntaxiques par des informaticiens. La disponibilité de ces corpus annotés est cependant limitée de nos jours à des langues très bien dotées (anglais, chinois, etc.).

Nous étudions dans ce mémoire la possibilité de dériver un analyseur du français à partir d'un analyseur existant de la langue anglaise (Link Grammar). Nous utilisons pour cela une technique consistant à projeter le dictionnaire monolingue anglais (vers le français) de cet analyseur à l'aide d'un bitexte (deux documents dont les phrases en relation de traduction sont identifiées) de grande taille et de techniques d'alignement bilingues au niveau des mots.

L'idée générale consiste premièrement à annoter le côté anglais du bitexte de projection avec des informations syntaxiques, puis à lancer l'alignement de mots entre les deux langues, et enfin à projeter l'analyse syntaxique produite par l'analyseur vers le français. La prochaine étape consiste alors à créer un dictionnaire français qui sera donné à l'analyseur projeté.

Nous étudions dans ce mémoire, outre la faisabilité de l'approche, différentes techniques de projection que nous évaluons en analyse sur un corpus (français) de référence. En effet, l'analyseur projeté est utilisé pour analyser syntaxiquement des phrases françaises, dont nous évaluons la qualité à l'aide de mesures de précision et de rappel communément utilisées.

Nous montrons que la projection d'une ressource comme un analyseur syntaxique est une idée viable. Nous nous attardons cependant à discuter les limites d'une telle approche, dressons la liste des problèmes que nous avons rencontrés et proposons des techniques pour les contourner.

Mots clé : analyseur syntaxique, alignement de mots, bitexte, corpus, dictionnaire, algorithme de projection, relations syntaxiques.

Abstract

Quality Parsers exist only for one small number of languages currently spoken in the world. This is explained by the fact why until recently, the development of parsers required the massive and recurring intervention of linguist experts. To develop such a parser was then a scientific project being spread out over several years. Recently, the availability of syntactically annotated corpora (by linguists) makes possible the fast development of parsers by computers specialists. The availability of these annotated corpora is however limited nowadays to languages equipped very well (English, Chinese, etc).

We study in this thesis the possibility of deriving a parser for French starting from an existing parser of the English language (Link Grammar). We use for that a technique consisting in projecting the English monolingual dictionary (into French) of this analyzer using a parallel corpus (two documents whose sentences in relation of translation are identified) of big size and bilingual alignment technique on the words level.

The general idea first, to annotate the English side of the parallel corpus of projection with syntactic information, then to launch the word-alignment between the two languages, and finally projecting the syntactic analysis produced by the parser into French. The next step then consists in creating a French dictionary which will be given to the projected parser.

We study in this thesis, in addition to the feasibility of the approach, various techniques of projection which we evaluate on a reference corpus (French). Indeed, the projected parser is used to annotate French sentences, of which we evaluate quality using measurements of precision and recall commonly used.

We show that the projection of a resource as a parser is a viable idea. We however discuss the limits of such an approach, draw up the list of the problems which we encountered and propose the techniques to circumvent them.

Key words: parser, syntactic annotation, word alignment, corpus, dictionary, projection algorithm.

Table des matières

Résumé.....	3
Abstract.....	5
Table des matières.....	7
Liste des figures.....	9
Liste des tableaux.....	11
Table de notations.....	12
R E M E R C I M E N T S.....	13
Chapitre 1.....	14
Introduction.....	14
1.1. Linguistique informatique.....	14
1.1.1. Grammaire.....	15
1.1.2. Analyse syntaxiques.....	15
1.2. Le concept de la projection.....	16
1.2.1. Importance applicative.....	16
1.2.2. Le concept.....	17
1.3. Travaux reliés à la projection.....	18
1.4. Notre travail.....	21
1.5. Aperçu sur la mémoire.....	22
Chapitre 2.....	24
Link Grammar et Link Parser.....	24
2.1. La logique et la notation.....	24
2.1.1. L'idée de base.....	24
2.1.2. Les règles des mots.....	25
2.1.3. Règles globales.....	28
2.2. Dictionnaire.....	28
2.3. Caractéristiques généraux du parseur.....	29
2.3.1. Indice inférieure du connecteur.....	29
2.3.2. Suffixe des mots.....	30
2.3.3. Le système des coûts.....	30
2.3.4. Exemple d'une analyse syntaxique.....	30
2.4. Conclusion.....	32
Chapitre 3.....	33
Alignement de mots.....	33
3.1. Les alignements.....	33
3.1.1. Les modèles d'alignements de IBM.....	35
3.1.1.1. Modèle de traduction par IBM ₁	36
3.1.1.2. Modèle de traduction par IBM ₂	37
3.1.2. Alignement de VITERBI.....	38
3.1.2.1. Principe.....	38
3.1.2.2. Fichier d'entrée.....	39
3.1.2.3. Fichier de sortie.....	40
3.2. Exemple.....	41
3.3. Conclusion.....	43

Chapitre 4.....	44
Corpus utilisé et évaluation.....	44
4.1 Corpus bi texte.....	44
4.2 Fichier de test.....	45
4.3 Préparation du corpus.....	45
4.4 Évaluation.....	47
Chapitre 5.....	50
Protocole général et expériences.....	50
5.1. Protocole de projection.....	50
5.2. Le mot le plus fréquent (PL-pfa).....	55
5.2.1. Description.....	55
5.2.2. Description formelle.....	56
5.2.3. Résultats et conclusion.....	58
5.3. Les mots du corpus (PL-a).....	61
5.3.1. Description.....	61
5.3.2. Description formelle.....	61
5.3.3. Résultats et conclusion.....	62
5.4. Les « n » les plus fréquemment et probablement alignés.....	65
5.4.1. Description.....	65
5.4.2. Description formelle.....	65
5.4.3. Résultats et conclusion.....	68
5.5. Projection de l'expression (PE-npfa).....	71
5.5.1. Description.....	71
5.5.2. Description formelle.....	72
5.5.3. Résultats et conclusion.....	75
5.6. Alignement bidirectionnel.....	78
5.6.1. Description.....	78
5.6.2. Description formelle.....	78
5.6.3. Résultats et conclusion.....	80
5.7. Ordre des mots.....	82
5.7.1. Description.....	82
5.7.2. Résultats et conclusion.....	84
5.8. Traduction à la main.....	87
5.8.1. Description.....	87
5.8.2. Résultats et conclusion.....	88
5.9. Bilan des expériences.....	90
5.9.1. Les limitations.....	93
Chapitre 6.....	97
Conclusion.....	97
Références.....	99

Liste de figures

Figure 1 : projection des relations syntaxiques.....	18
Figure 2 : affichage du résultat de l'analyse syntaxique du « Link-Grammar ».....	19
Figure 3 : exemple des connecteurs partant des mots anglais.....	24
Figure 4 : des connecteurs formant des liens entre les mots reliés.....	25
Figure 5 : le dictionnaire représenté par une liste de mots avec leurs formules.....	27
Figure 6 : une tranche du dictionnaire de la « Link-Grammar ».....	29
Figure 7 : un alignement dont chaque mot français est aligné à un mot anglais.....	33
Figure 8 : un autre alignement possible de ces phrases mais moins probables.....	33
Figure 9 : un alignement dont chaque mot anglais est associé à un seul mot français.....	34
Figure 10 : un alignement dont un ensemble de mots français est connecté à un ensemble de mots anglais.....	34
Figure 11 : les paramètres des paires de mots alignés.....	35
Figure 12 : la probabilité qu'une phrase f soit la traduction de e	35
Figure 13 : une tranche d'une phrase du corpus de test.....	45
Figure 14 : filtration du corpus à l'aide du vocabulaire de la référence.....	46
Figure 15 : filtration de la référence à l'aide du vocabulaire du corpus filtré.....	46
Figure 16 : filtration du corpus à l'aide du vocabulaire de la référence.....	46
Figure 17 : la fonction de projection dans l'algorithme général.....	52
Figure 18 : la liste des mots avec leurs fréquences après alignement.....	55
Figure 19 : l'ensemble $tf(f)$ formé par les mots anglais et leurs fréquences d'alignement.....	57
Figure 20 : pourcentage des mots source (anglais) en fonction de la fréquence d'alignement.....	59
Figure 21 : l'union des liens formés par les mots anglais avec « chien », et qui forme l'entrée de ce mot dans le dictionnaire.....	62
Figure 22 : pourcentage des mots cibles (français) ayant une projection.....	64
Figure 23 : sortie de l'alignement pour le mot « tradition ».....	64
Figure 24 : l'ensemble $tf(f)$ du mot « chien ».....	67
Figure 25 : comparaison des nombres de phrases analysées entre les « n » les plus fréquemment (PL-npfa) et les plus probablement(PL-nppa) alignés.....	68
Figure 26 : pourcentage de nombre de mots en fonction de la fréquence d'alignement entre PL-npfa et PL-nppa.....	69
Figure 27 : comparaison des précisions entre PL-npfa et PL-nppa.....	70
Figure 28 : comparaison de pourcentage de nombre de fois des mots anglais en fonction de leurs masses, entre PL-npfa et PL-nppa.....	71
Figure 29 : redéfinir la définition de la projection des expressions.....	73

Figure 30 : l'expression du mot «chien» dans le dictionnaire français.....	74
Figure 31 : résultat des nombres de phrases analysés par la méthode PE-npfa en fonction de « n ».....	75
Figure 32 : résultat du F-mesure de la méthode de PE-npfa en fonction de « n ».....	75
Figure 33 : comparaison de la précision des deux méthodes PL-npfa et PE-npfa.....	76
Figure 34 : comparaison de nombres de phrases analysées des deux méthodes PL-npfa et PE-npfa.....	76
Figure 35 : le nombre de phrases après analyse des PL-npfa et PL-nppa avec le nouveau fichier d'alignement.....	80
Figure 36 : pourcentage des nombres des mots anglais en fonction de la fréquence d'alignement, entre l'alignement unidirectionnel et bidirectionnel.....	81
Figure 37 : les résultats de la précision du PL-npfa et PL-nppa avec le nouveau fichier d'alignement.....	81
Figure 38 : comparaison de nombre de phrases analysées des deux méthodes PL-npfa et PE-npfa en appliquant les transformations sur les liens.....	84
Figure 39 : comparaison de la précision des deux méthodes PL-npfa et PE-npfa en appliquant les transformations sur les liens.....	85
Figure 40 : Une tranche de la liste des mots traduits à la main.....	87
Figure 41 : Une liste des mots dont chacun à un sens différent.....	88
Figure 41 : comparaison de nombre de phrases analysées des deux méthodes PL-npfa et PE-npfa après transformations des liens.....	89
Figure 42 : comparaison de la précision des deux méthodes PL-npfa et PE-npfa en appliquant les transformations sur les liens.....	90
Figure 43 : Un exemple montrant l'échéance de la projection des relations syntaxiques.....	91
Figure 44 : Un exemple du « futur simple » an anglais et sa traduction en français.....	92

Liste de tableaux

Tableau 1 : préservation des relations de l'exemple 1	18
Tableau 2 : les probabilités des alignements sur un corpus de phrase de 8 mots (anglais/français).....	37
Tableau 3 : des statistiques concernant les corpus utilisés.....	58
Tableau 4 : résultats de l'application de la méthode de projection des liens du mot le plus fréquemment aligné (PL-pfa) en fonction de la longueur des phrases..	59
Tableau 5 : les résultats de la projection de tous les mots alignés à un seul mot français (PL-a).....	63
Tableau 6 : comparaison des nombres de phrases analysés entre PE-npfa, sans croisement et avec croisement.....	85
Tableau 7 : comparaison de nombres de phrases analysées de la PL-npfa, sans et avec application des croisements des liens.....	86
Tableau 8 : résultats de la méthode qui projette l'expression du mot traduit à la main.....	88
Tableau 9 : comparaison des méthodes PL-npfa, PE-npfa et PE-main en ajoutant l'option de NULL-LINK.....	92
Tableau 10 : statistiques montrant le nombre des phrases ayant les limitations différentes.....	95

Table de notation

e_i	Le mot anglais à la position i .
e^I	Une phrase anglaise de I mots.
m	La longueur de la phrase anglaise.
I	Position en $e^I, i=0,1,\dots,m$
f_j	Le mot français de la position j .
f^l	Une phrase française de l mots.
l	La longueur de la phrase française.
j	Position en $f^l, j=0,1,\dots,l$
$t(e_i f_j)$	Probabilité de transfert
a	Alignement
$a(i j,I,J)$	Probabilité d'alignement

R E M E R C I E M E N T S

Il va sans dire que ce travail n'a pu être réalisé sans l'apport essentiel d'un grand nombre de personnes et d'organismes.

Tout d'abord je tiens à témoigner de ma plus profonde gratitude à mon directeur de recherche monsieur Philippe Langlais. Il a contribué à ce projet de recherche de façon incommensurable. Sa rigueur intellectuelle et son soutien financier ont permis au projet de suivre sa route, ses idées lumineuses ont fait sauver un temps précieux, et ses connaissances du sujet ont toujours été d'une grande utilité. Un grand merci donc, pour une formidable année, parsemée d'aide appréciée et de discussions passionnantes.

Merci au département d'informatique et de recherche opérationnelle et à la Faculté des études supérieures de m'avoir si généreusement accordé une bourse de rédaction.

Un remerciement tout particulier à mes amis qui me soutiennent toujours et particulièrement : Kamal, Ali, Hind et Youssef pour leurs aides dans ce mémoire. Finalement, une attention vers ma famille qui n'ont pas négligé les sacrifices tout au long des mes études.

Chapitre 1

Introduction¹

L'informatique est depuis plusieurs années entrée dans notre quotidien. Avec l'évolution rapide des technologies informatiques, le besoin s'est rapidement fait sentir de s'appuyer sur les techniques linguistiques pour faciliter la communication homme-machine. Des majordomes téléphoniques permettent par exemple, grâce à la reconnaissance vocale d'acheminer un appel automatiquement à son destinataire via un dialogue minimaliste. Parallèlement, la linguistique a pu profiter de la puissance des ordinateurs pour acquérir une nouvelle dimension, et ouvrir la voie à de nouveaux domaines de recherche.

1.1. Linguistique informatique

La linguistique informatique fait partie intégrante des techniques informatiques, et intervient également dans des sous-domaines de l'intelligence artificielle. Parmi les applications concrètes figurent: L'analyse (lexicale, syntaxique et sémantique) des langages informatiques ou humains, la traduction automatique, la recherche d'information, etc...Ce sont tous des applications du traitement automatique des langues (TAL) ; Une discipline à la frontière de la linguistique et de l'informatique, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain.

Enfin, le traitement automatique de langues a pour objectif de traiter, d'une façon automatique, des données linguistiques exprimées dans une langue dite « naturelle ». Ces données linguistiques peuvent être des textes écrits, qui sont des suites de phrases. Un texte doit donc pouvoir être écrit comme un ensemble de formes régi par des règles explicables : les règles de la langue.

¹ Texte extrait du site : http://fr.wikipedia.org/wiki/Linguistique_informatique

1.1.1. Grammaire

Pour pouvoir traiter automatiquement les données, il faut être capable d'explicitier ces règles de la langue, de les représenter dans des formalismes opératoires et calculables, et les implémenter à l'aide de programmes, ces règles s'appellent « la grammaire ».

Quand nous parlons de la grammaire nous pensons immédiatement aux langages ou en général, toutes autres choses associées aux langages, comme les mots dans le langage, les règles sur l'utilisation des mots en conversation, en écriture, etc...

Qu'est-elle la définition de la grammaire? La grammaire est définie comme un ensemble de règles décrivant le fonctionnement d'une langue. Une autre question se pose : Pourquoi avons-nous besoin d'étudier la grammaire d'un langage particulier?

Beaucoup de raisons, mais voici les plus importantes :

1. Un ensemble de règles fini peut décrire la structure d'un nombre infini de séquences de mots.
2. La grammaire représente notre compréhension des modèles observés. La preuve qu'on maîtrise une langue (naturelle ou pas) est l'apprentissage de sa grammaire.
3. La grammaire peut aider à fournir des informations manquantes dans une représentation linéaire. Par exemple, dans la reconnaissance de la parole, s'il y a une incertitude pour le mot qui vient juste d'être prononcé, une grammaire pourrait aider à déterminer le mot correct.

1.1.2. Analyse syntaxique².

Dans notre travail nous nous intéressons à l'analyse syntaxique qui s'intègre au début du processus.

² Prise du site : <http://french.chass.utoronto.ca/fre378/>

L'analyse syntaxique est une partie de la grammaire qui s'intéresse à l'étude des règles qui servent à expliquer d'une part, l'ordre des mots dans la phrase et, d'autre part, traite la manière dont les mots peuvent se combiner pour former des propositions (unité syntaxique construite autour d'un verbe) ainsi que l'enchaînement des propositions entre elles.

Cela consiste à associer à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités. Afin de présenter ces relations, les résultats peuvent avoir plusieurs formes qui nous aident à comprendre la syntaxe de cette chaîne (arbre syntaxique, arcs dressés entre les mots,...).

Prenons l'exemple suivant : *Jean est allé à l'école*. Le résultat de l'analyse syntaxique donc, pourra être l'arbre suivant :

```
(S (NP Jean)
  (VP est
    (VP allé
      (PP à
        (NP l'école)))))).
```

En résumé, l'analyse est le processus de structurer une représentation qui est conforme à une grammaire donnée. Ainsi par exemple : un analyseur de langage C est un programme qui examine un fichier de texte, et assigne la structure appropriée au texte selon la syntaxe et les règles du langage de programmation C.

1.2. La projection

1.2.1. L'importance applicative

En introduisant la notion de la grammaire et son importance dans les applications du traitement de la langue, on peut être amené à penser à des concepts qui à leur tour peuvent créer une grammaire d'une langue donnée sans l'effort linguistique spécialisé dans cette langue.

Ce processus vient de l'idée de la difficulté d'écrire une grammaire pour une langue donnée. Par exemple, l'analyse syntaxique d'une phrase écrite en espagnol nécessite la maîtrise de cette langue et donc sa grammaire afin de l'appliquer pour pouvoir extraire son arbre syntaxique, mais pour une personne ne connaissant pas cette langue, il est presque impossible de le faire. D'ici vient le besoin d'une application qui peut former une grammaire d'une langue quelconque « F » à l'aide d'une grammaire existante d'une langue donnée « E ». Cette application s'appelle la projection d'une grammaire de « E » vers « F », plus précisément des règles syntaxiques de chaque mot de la langue « E », qui consiste à transférer ces règles d'une langue à une autre.

1.2.2. Le concept

Jusqu'ici, nous avons présenté la grammaire et la nécessité de la créer pour une langue à partir d'une langue donnée. En présence de deux langues, il est nécessaire d'expliquer la relation entre les deux et l'importance de cette relation dans l'approche de la projection.

En fait, le besoin de communiquer rapidement dans toutes les langues devient une priorité pour les êtres humains. C'est la naissance de la traduction automatique, qui relie les mots d'une langue « E » avec les mots d'une langue « F ». Ceci constitue un but de l'informatique et de l'intelligence artificielle en particulier depuis longtemps.

Le besoin majeur de la traduction automatique dans notre approche, se concentre principalement sur la traduction des mots d'une phrase dans une langue « E » vers les mots d'une autre phrase dans une langue « F » (en considérant que les phrases sont aussi traduites l'une par rapport à l'autre). C'est la seule relation entre ces deux langues dont on a besoin pour transférer cette grammaire.

Dans sa forme la plus simple, la projection d'une grammaire d'une langue vers une autre peut s'exprimer par la projection des liens syntaxiques rencontrés dans le matériel source vers la langue cible. Si l'on fait l'hypothèse d'une projection bijective, ceci consiste à dire que deux mots sources S_i et S_j qui sont en relation syntaxique (ce que l'on notera par $R(S_i, S_j)$), voient leurs traductions T_a et T_b partager la même relation ($R(T_a, T_b)$).

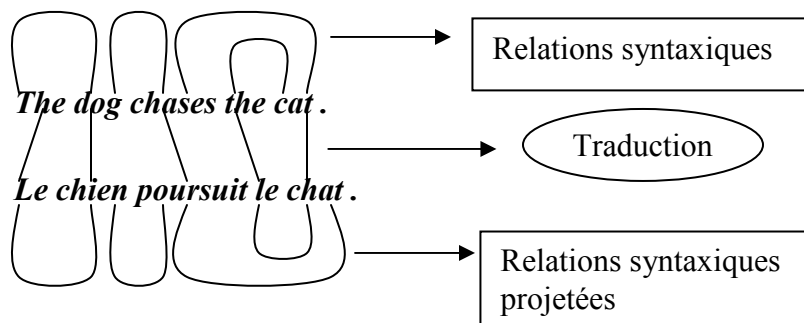


Figure 1 : projection des relations syntaxiques

Nous illustrons sur l'exemple de la figure 1 cette idée intuitive; Si le mot anglais "chase" est traduit au mot français "poursuit" et "cat" est traduit au "chat", la création d'une relation grammaticale (représentée par des arcs) entre "chase" et "cat" est accompagné d'une création d'une même relation entre "poursuit" et "chat". On peut voir dans le tableau 1 le nom de chaque relation liant les mots ensemble, dans les deux langues étudiées.

R	S _i	S _j	T _a	T _b
Det-noun	The	Dog	Le	Chien
Sub-verb	Dog	Chases	Chien	Poursuit
Det-noun	The	Cat	Le	Chat
Verb-obj	Chases	Cat	Poursuit	Chat

Tableau 1 : préservation des relations de l'exemple 1.

Comme indiqué, la projection s'élève à une prétention que la traduction à travers deux langues ressemble à un homomorphisme reliant le graphe syntaxique de la phrase source (anglais) au graphe syntaxique de la phrase cible (français).

1.3. Travaux reliés à la Projection

Il existe différents types de formalismes capables de structurer un langage donné en terme de relations syntaxiques. Une de ces applications grammaticales s'appelle la « Link-Grammar » [D. Sleator et al., 91], qui consiste à représenter une phrase comme un ensemble de mots reliés entre eux par des liens syntaxiques. Un système (Link-Parser) a été

défini et développé par Daniel Sleator, Davy Temperley et John Lafferty en 1991, qui repose sur la grammaire de « Link-Grammar »; ce système analyse syntaxiquement une phrase dans la langue anglaise en s'appuyant sur les liens créés entre les mots, dont chaque lien doit être nommé (ex : sujet-verbe, adj-nom) et répond à des conditions décrites dans un dictionnaire formé par la main. Une grammaire développée pour l'Anglais est disponible et est constituée d'un vocabulaire de 107 formes, et des liens dérivant de ces formes peuvent entretenir avec les mots anglais.

Le résultat de l'analyse est une suite de relations grammaticales entre les mots formant la phrase en question, et qui se visualise sous la forme d'un graphe d'arcs planaires et, alternativement une représentation d'arbre est offerte.

Exemple ³:

```

+-----O-----+
+-D-----S---+   +-----D-----+
|         |         |         |         |
the dog.n chases.v the cat.n

```

Constituent tree:

```

(S (NP The dog)
  (VP chases
    (NP the cat)))

```

figure 2: affichage du résultat de l'analyse syntaxique du « Link-Parser »

Cet outil d'analyse syntaxique anglais n'attaque pas la projection telle expliquée. Mais d'une façon, il est relié à notre approche par le fait qu'on utilise ses liens produits pour les projeter. En ce qui concerne l'idée de la projection, il y avait un nombre modeste d'études exploitant les corpus parallèles dans l'amorçage des outils d'analyse monolingues. Cette idée, apparemment a aidé à former des structures d'arbres pour la langue en parallèle ou encore en surmontant l'incertitude des annotations syntaxiques pour des langues autres que l'Anglais. C'est le fruit de la combinaison d'analyse syntaxique d'une langue de départ et la traduction de cette langue à une autre.

L'idée de projeter des ressources monolingues dans une autre langue n'a suscité jusqu'à maintenant que peu d'intérêt. Hwa et al. (2001) ont cependant montré qu'il était envisageable de projeter les relations syntaxiques de l'Anglais vers le Chinois. Leur but

³ Résultat pris du site : <http://www.link.cs.cmu.edu/link/submit-sentence-4.html>

était d'étudier la performance de la projection de l'analyse grammaticale vers le Chinois, en utilisant un alignement de mots comme un pont de cette projection, une technique que nous décrivons dans le chapitre 3.

Cette idée se formalise en considérant deux paires de phrases, E (source) et F (cible), telle qu'elles sont la traduction l'une de l'autre ayant les structures syntaxiques $Tree_E$ et $Tree_F$. Donc, si les nœuds X_E et Y_E du $Tree_E$, sont alignés avec les nœuds X_F et Y_F du $Tree_F$, respectivement, et si la relation syntaxique $R(X_E, Y_E)$ est vraie dans $Tree_E$, alors $R(X_F, Y_F)$ sera valide dans $Tree_F$.

Dans leur cadre expérimental, l'analyse syntaxique pour l'Anglais est fournie par les modèles de Collins [Collins, 1997] et corrigée manuellement. Un corpus bilingue de 124 phrases est utilisé pour la projection, construit aussi par des linguistes professionnels. L'alignement de mots entre ces deux langues est fait manuellement, et même l'analyse du corpus de test est construite par des humains (88 phrases).

Ils montrent, dans leur travail, que la projection directe des relations syntaxiques ne peut pas être assurée. Cette projection offre de très faibles résultats (précision :30% et rappel :39%). Une inspection de ces résultats indique que leur corpus parallèle aligné manuellement contient plusieurs instances des alignements multiples ou des mots non alignés. Un ensemble de règles syntaxiques est appliqué alors durant la projection afin de confronter ces problèmes. Ceci, leur permet d'obtenir 67% pour le F-mesure (expliquée dans le chapitre 3) qui représente 76% de gain relatif sur la projection directe.

Dans la même année, 2001, David Yarowski et Grace Ngai ont également décrit un travail où ils projettent avec succès des étiquettes **POS** (catégorie syntaxique d'un mot) de l'Anglais vers le Français et le Chinois. Ils utilisent pour cela l'alignement de mots aussi.

Exemple :	Anglais :	DT	NNS
		The	laws
	Français :	Les	lois
		DT	NNS

L'idée, illustrée dans l'exemple précédent, peut être vue très clairement. Après l'alignement de mots entre ces deux phrases, les tags de chaque mot anglais sont transférés (projetés) vers les mots français qui lui sont associés.

Deux limitations pour ce paradigme, le premier est la faible précision des alignements de mots, due aux limitations courantes des algorithmes d'alignement de mots. Par exemple, si « The » n'est pas aligné à aucun mot français, et « laws » est aligné vers « Les » et « lois », alors « Les » va avoir le tag « NNS », ce qui n'est pas correcte dans la grammaire française. La deuxième limitation est la disparité potentielle dans les besoins d'annotation de deux langues : toutes les distinctions faites dans une langue ne sont pas nécessairement pertinentes dans une autre langue.

Ils ont montré que la simple projection directe des annotations est très bruitée, même lorsque les alignements de mots ont été manuellement corrigés. Des filtres robustes des données et des procédures sont alors décrits pour s'exécuter efficacement sur ces données de mauvaise qualité. Les taggers résultants surpassent de manière significative les projections directes sur lesquelles ils ont été formés, en donnant une précision de 96%.

Nous avons observé que même s'ils ont atteint un niveau élevé dans la précision, leur approche semble plus facile à appliquer avec succès. Puisque c'est seulement le tag de chaque mot qui est projeté, il n'existe pas de relations liant les mots. Ainsi, il n'y aurait pas de confusion en ce qui concerne l'ordre des mots dans la phrase (une limite étudiée au chapitre 5). Notons que l'ordre de mots peut constituer un problème pour les approches qui projettent la relation entre les deux mots (tel que [Hwa et al., 2001] et notre travail, expliqué dans la section 1.4). Donc il suffit d'avoir un aligneur de mot efficace pour la réussite de leur approche. Par contre, les mots alignés incorrectement obtiennent des probabilités trop petites permettant de les éliminer durant l'entraînement sur les tags projetés.

1.4. Contribution

La projection des relations syntaxiques semble être un principe raisonnable, particulièrement quand elle est exprimée en termes de dépendances syntaxiques qui sont

largement préservées à travers l'alignement, comme illustré dans le tableau 1. De plus elle peut nous permettre d'utiliser l'analyse syntaxique d'un langage pour construire des annotations pour la phrase correspondante dans l'autre langage.

Nous nous intéressons dans cette étude à la projection de l'Anglais vers le Français d'une grammaire lexicalisée disponible pour l'anglais dont nous avons parlé précédemment et que nous détaillons dans le chapitre 2. Nous utilisons pour cela l'idée proposée par Rebecca Hwa [Hwa et al.,2001], et faisons usage d'un alignement bilingue de mots que nous décrivons au chapitre 3. Nous construisons un dictionnaire français à partir des mots du vocabulaire du corpus avec leurs liens grammaticaux projetés (qui forment les entrées de ce dictionnaire). Ce dictionnaire va être appliqué sur le système d'analyse de l'anglais (le « Link-Parser » décrit dans le chapitre 2). Enfin, les performances de notre grammaire projetée sont étudiées sur un corpus de phrases arborées de phrases françaises que nous présentons dans le chapitre 4.

Rappelons que l'idée de Yarowski [Yarowski et al., 2001] était de projeter seulement des tags appartenant à chaque mot de la langue source. D'autre part, notre travail est inspiré de l'idée de [Hwa et al., 2001] sur la projection des relations grammaticales entre les mots, aidant aussi à réduire le travail et le coût pour la création des arbres syntaxiques dans les nouveaux langages. Notons qu'un arbre syntaxique joue un rôle important dans l'analyse syntaxique. Pratiquement, un arbre peut servir d'interface entre les différentes phases de traduction d'un analyseur, et permet encore de spécifier de meilleures stratégies d'analyse syntaxique.

Contrairement à l'approche de Hwa, nous avons relevé le défi avec l'utilisation de données beaucoup plus grandes que celles de Hwa. Nous avons appliqué la projection sur à peu près 32 500 paires de phrases (contre 124) du corpus bitexte, dont l'analyse du coté anglais est fournie par un système sans correction manuelle. Même l'alignement de mots, construit manuellement dans Hwa, est implémenté automatiquement par une technique des modèles IBM dans notre approche (expliqué au Chapitre 3). Notre travail est donc un vrai test d'une application complètement automatique de la projection des relations grammaticales.

Notre travail se démarque également de celui de Hwa et al. [2001] de par la nature même de la grammaire que nous projetons. Dans notre cas, nous projetons le dictionnaire sous-jacent à la grammaire. Ce dictionnaire consigne l'ensemble des relations grammaticales que peuvent entretenir chaque mot de la langue étudiée. Cette diversité des relations que peut y avoir un mot donné, facilite la construction d'une large grammaire. Ainsi, les liens appartenant à chaque mot de notre dictionnaire peuvent être modifiés de façon qu'ils s'adaptent avec la situation du mot utilisé dans une phrase. Par exemple, la relation « adjective-nom » appartenant à une définition d'un mot quelconque, peut être modifiée sans affecter les autres relations du même genre. Cet avantage peut être une solution au problème de l'ordre des mots rencontré et expliqué au chapitre 5.

Notre résultat montre que, bien que la projection est souvent trop restrictive, un petit ensemble de transformations linguistiques élémentaires peut améliorer la qualité de la projection.

1.5. Aperçu sur le contenu du mémoire

Dans ce mémoire, nous présentons une étude empirique qui mesure le degré de préservation des relations syntaxiques lors de la projection des annotations de l'anglais vers le français. Dans cette partie, nous allons présenter un bref aperçu sur le contenu des différents chapitres de notre mémoire. Nous décrivons tout d'abord dans le chapitre 2 le système d'analyse syntaxique que nous avons utilisé dans ce travail (Link-Grammar).

Nous présentons dans le chapitre 3 la technique d'alignement que nous avons implémenté; une technique initialement proposée par la traduction statistique par [Brown et al., 1993].

Nous présentons alors dans le chapitre 4 les corpus que nous avons réunis et traités pour ce travail, et précisons le système d'évaluation que nous avons mis en place pour mesurer nos progrès.

Le chapitre 5 constitue le cœur de notre approche et décrit les différentes techniques de projection que nous avons étudiées et implémentées. Toutes les expérimentations ainsi

que les résultats de ces projections obtenus à l'aide de PARSEVAL sont montrés dans ce chapitre.

Nous concluons et résumons notre travail dans le chapitre 6, par les points les plus intéressants dans cette recherche ainsi que les perspectives qui peuvent étendre notre projet dans le futur.

Chapitre 2

Link Grammar et Link Parser

Dans le chapitre précédent, on a expliqué le principe de notre travail et qui comprend une analyse syntaxique au début du processus. Dans ce chapitre, nous expliquons les fondations des “Link-Grammar” et de l’analyseur que nous avons utilisé le “Link Parser”. La “Link Grammar” a été développée par Daniel D. Sleator et Danny Temperly, et leur article original est inclus dans la référence.

2.1. La logique et la notation des link grammars

2.1.1. L’idée de base.

La « Link grammar » un formalisme qui permet de relier des mots entre eux à l’aide d’un ensemble de liens. Dans ce formalisme, les mots sont représentés par des ensembles de connecteurs sortant d’eux.

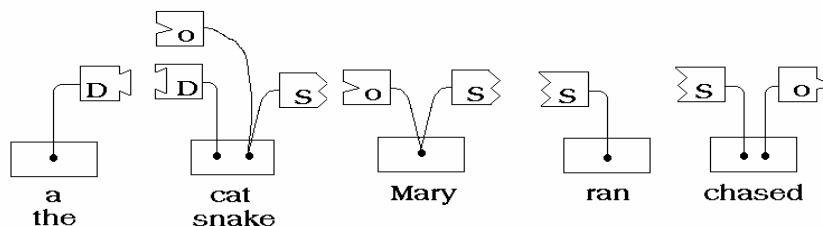


Figure 3 : exemple des connecteurs partant des mots anglais.

Un connecteur caractérise un lien syntaxique que le mot décrit peut avoir avec un autre dans la même phrase. Par exemple, les mots « a » et « the » de la figure 3 sont à la recherche de mots pour former une relation de « déterminant » (D). De la même manière, « Mary » pour se réaliser doit entretenir une relation « objet(O) » et « sujet(S) » avec d’autres mots.

Ecrire une Link-Grammar consiste à préciser l’ensemble des liens (connecteurs) qu’un mot peut prendre dans une langue.

Il y a différents types de connecteurs, et ces connecteurs peuvent aussi se pointer à la droite ou à la gauche du mot. Les connecteurs qui pointent à la droite sont marqués par “+”, et ceux à la gauche sont marqués par “-“. Un connecteur qui pointe à la gauche se connecte avec celui qui pointe à la droite pour un même type sur un différent mot. Les deux connecteurs liés ensemble forment un lien (link). Par exemple, pour la phrase “The cat chased a snake”, les liens suivants peuvent être formés:

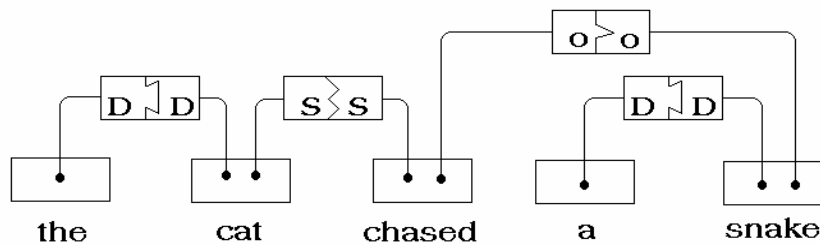


figure 4⁴ : des connecteur formant des liens entre les mots reliés ensemble.

Une phrase valide est une phrase dans laquelle tous les mots sont liés entre eux de manière cohérente et selon certaines règles globales que nous décrivons dans la prochaine section.

2.1.2. Les règles des mots.

Tous ces mots avec leurs règles doivent être arrangés dans un dictionnaire. Une simple entrée du dictionnaire ressemble à ceci:

blah: A+;

Ceci signifie que si le mot “blah” est utilisé dans une phrase, il doit former un lien “A” avec un autre mot à sa droite; c.à-d. il doit y avoir un autre mot à la droite de lui avec un connecteur “A-“. Autrement la phrase est inadmissible. L’expression après les deux points est le “linking requirement” pour le mot, également appelé sa formule.

⁴ exemple extrait du papier : Daniel Sleator and Davy Temperley, *Parsing English with a Link Grammar*, Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.

Un mot peut avoir plus d'un connecteur qu'ils doivent se connectés. On utilisera alors le symbole & pour indiquer que les différents connecteurs doivent être satisfaits. Ça serait noté comme suit:

blah: A+ & B+;

Un mot peut avoir une règle dans laquelle un des deux (ou plusieurs) connecteurs peut être utilisé mais exactement un seul doit être utilisé. Dans le dictionnaire, on la note comme ceci:

blah: A+ or B-;

Ceci signifie que si le mot peut faire un lien à la droite, ou un lien vers la gauche, leurs utilisations dans la phrase sont valides; mais elle doit faire l'un ou l'autre, et elle ne peut pas faire les deux ensembles.

Ces règles peuvent être combinées. Par exemple, considérons la notation suivante:

blah: A+ or (B- & C+);

Ceci signifie que le mot doit faire un lien "A" vers la droite, ou un lien "B" vers la gauche et un lien "C" vers la droite. Aucune autre combinaison ne peut être valide.

De telles expressions peuvent être combinées sans limite, comme:

blah: (A+ or B-) & ((C- & A+ & (D- or E-)) or F+);

Quelques connecteurs sont optionnels et sont notés avec des accolades, par exemple:

blah: A+ & {B+};

Ceci signifie que le mot doit faire un lien "A" vers la droite et peut faire un lien "B" vers la droite aussi, mais pas nécessairement.

Un mot peut encore faire plusieurs liens du même type à d'autres mots. Pour ceci, nous employons le symbole de "multi-connecteur" le "@". Par exemple, le mot ci-dessous pourrait faire tout nombre des liens de "F" aux mots vers la droite (mais n'est pas obligé d'en faire aucun).

blah: (A+ or B+) & {C- & (D+ or E-)} & {@F+};

L'ordre des éléments dans l'expression est important. Et ceci indique la proximité relative des mots auxquels ils sont reliés. Plus loin vers la gauche du nom du connecteur, plus la connexion doit être proche. Par exemple:

`blah: A+ & B+;`

Signifie que "blah" doit faire un "A" lien vers la droite et un "B" vers la droite avec, mais également que le mot fait un "A" lien avec un autre mot plus proche que le mot qu'il se connecte avec "B". Prenons l'exemple de la figure 3 :

`Snake : D- & O-;`

Donc « snake » se doit connecter avec « a » par le lien « D » en premier et puis avec « chased » par « O ».

Ceci concerne seulement, cependant, des connections dans la même direction. Pour des connecteurs qui pointent dans des directions opposées, l'ordre est non pertinent. De plus, dans les "or" expressions, comme "A+ or B+", l'ordre des éléments est non pertinent.

Nous devrions mentionner le « disjoint », qui est un concept important de l'analyseur. C'est l'ensemble des connecteurs qui constitue une utilisation légale d'un mot. Un mot dans un dictionnaire de la Link-Grammar est représenté par un ensemble de disjoints. Si un mot a l'expression suivante:

`blah: {C-} & (A+ or B+);`

Alors, il a les quatre disjoints suivants:

D1 : C- A+

D2 : A+

D3 : C- B+

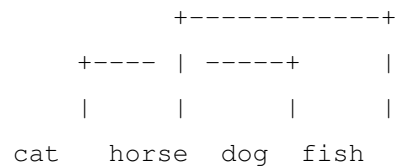
D4 : B+

Ces disjoints représentent toutes les utilisations légales du mot "blah". Employer le C- et A+ est une utilisation légale du mot, mais ce n'est pas le cas pour A+ et B+. Les disjoints jouent un rôle important pour faciliter la recherche des liens dans l'algorithme de l'analyseur.

2.1.3. Règles Globales.

En plus d'avoir chaque mot spécifié par un ensemble de disjoints, une Link-Grammar fait usage de deux règles globales qui contrôlent comment les mots peuvent se lier.

Tout d'abord, les liens ne peuvent pas se croiser. Par exemple, la manière suivante de relier ces quatre mots ("cat" se reliant au "dog", et "horse" au "fish") serait illégale. L'analyseur simplement ne trouvera pas de tels liens.



C'est la règle du "crossing-link" (ou "**planarity**").

Deuxièmement, tous les mots dans une phrase doivent être indirectement reliés entre eux. Par conséquent la manière suivante de relier ces quatre mots serait illégale.



C'est la règle de "**connectivity**".

Une phrase valide est donc; qui peut être liée d'une manière que : a) tous les mots sont employés dans une façon qui répond à leurs exigences des liens, et b) le croisement des liens et la connectivité ne sont pas violées.

2.2. Dictionnaire

Une liste des mots avec les formules est appelée un dictionnaire;

words	formula
a the	D+
snake cat	D- & (O- or S+)
Mary	O- or S+
ran	S-
chased	S- & O+

figure 5 : le dictionnaire représenté par une liste de mots avec leur formule.

Une entrée dans le dictionnaire se compose ainsi d'un mot, suivis des deux points, suivis d'une expression de connecteurs, et dans la plupart des cas sont regroupés dans des rubriques indiquant leurs fonctions, un exemple de ces groupes illustré dans la figure 6:

```
% NOUNS
places.n: ({@AN-} & {@A- & {[[@AN-]]}} & (({Dmc-} & {@M+} & {TON+ or TH+ or Ce+ or (R+ & Bp+)}) & {@MXp+} &
(<noun-main-p> or Bpm+)) or Up- or (YP+ & {Dmc-}) or (GN+ & (DD- or ()))) or [[AN+]];
times.n: ({@AN-} & {@A- & {[[@AN-]]}} & (({Dmc-} & {@M+} & {TON+ or VN+ or TH+ or Ce+ or (R+ & Bp+)}) & {@MXp+}
& (<noun-main-p> or Bpm+)) or Up- or (YP+ & {Dmc-}) or (GN+ & (DD- or ()))) or [[AN+]];

%PRONOUNS
she he: ({[[R+ & Bs+]]) & ((Ss+ & <CLAUSE>) or Sls-);
me him them us: J- or Ox-;
myself yourself himself herself itself themselves ourselves yourselves: J- or O- or E+ or MVa-;
each_other: J- or O- or YS+;
```

Figure 6 : une tranche du dictionnaire de la Link-Grammar.

Le dictionnaire se compose d'une série de telles entrées. Tous les mots possédant la même condition d'enchaînement, sont mis dans une liste, séparée par les espaces. Par exemple :

```
me him them us : J- or Ox-;
```

Notre travail consiste à projeter ce dictionnaire qui est une ressource qui n'existe que pour l'anglais et qui représente de nombreuses heures de spécialistes, qui l'ont construite à la main.

2.3. Caractéristiques générales du parseur

2.3.1. Indice inférieur du connecteur.

En général, un connecteur peut seulement se lier à un autre avec le même nom, c'est-à-dire les mêmes lettres majuscules. Cependant, il y a une autre manière de contrôler comment les connecteurs peuvent se lier entre eux, en utilisant des indices inférieurs du connecteur. Un indice inférieur est une lettre minuscule qui

suit le nom du connecteur, comme “Ss+”. Un “Ss+” peut se relier à un “S-“ ou a un “Ss-“, mais pas avec un connecteur “Sp-“.

2.3.2. Suffixe des mots.

Un mot pouvant avoir plusieurs entrées différentes dans le dictionnaire, les entrées sont distinguées avec des suffixes. Des mots peuvent être suivis d’un indice inférieur tel que “.n”. Par exemple:

run.n: A+ or B+...

run.v: C+ or D+...

En cherchant les liens, le parseur considèrera chaque entrée pour le mot comme un mot différent, et produira tous les liens trouvés pour toutes les entrées. Le suffixe est affiché, indiquant quelle entrée le parseur a choisi, pour un lien particulier.

2.3.3. Le système des coûts.

C’est un système pour assigner un coût à un lien. Ceci permet au parseur d’exprimer des préférences parmi les liens qu’il trouve. Le système des coûts utilise les crochets (“[“,”]”). Si un connecteur ou des séries de connecteurs, est entouré par les crochets, il est assigné un coût. Le montant du coût est égal au nombre de crochets de chaque coté: [A+] recevra un coût de 1; [[A+]] recevra un coût de 2, etc.... le parseur utilise ce coût comme critère pour décider quel lien choisir le premier, ce sont produit par ordre de coût (c’est à dire le plus faible coût en premier).

2.3.4. Exemple d’une analyse syntaxique.

Afin de bien comprendre les différents contextes introduits, nous proposons ici un exemple d’analyse utilisant le dictionnaire fourni, pour la phrase suivante:

It is a powerful grammar we can use to analyse English sentences

Après le lancement du parseur :

Une autre option du système d'analyse qui permet des liens vides de s'établir dans une phrase. Donc une phrase peut être analysée, même si des mots ne peuvent pas être liés. Cette fonction s'appelle la « NULL-LINK » qui peut prendre des valeurs entre 1 et le nombre de liens établis dans la phrase. Donc si NULL-LINK est égale à 1, le système analyse la phrase et permet d'avoir un seul lien non établi dans les résultats produits. Ainsi de suite pour chaque valeur de NULL-LINK.

2.4. Conclusion

Nous avons expliqué dans ce chapitre le fonctionnement d'un système d'analyse syntaxique, que nous avons utilisé afin de fournir les liens de chaque mot Anglais. Notons que ce sont les mots qui jouent le rôle du mot source dans notre projection, et leurs liens doivent être projetés sur le mot cible (le français dans notre cas). C'était la première étape de notre approche. L'application qui relie ces deux langages, plus précisément les deux mots de chaque langage va être le sujet du prochain chapitre.

Chapitre 3

Alignement de mots

Comme nous l'avons mentionné dans notre introduction, l'idée de notre travail est de construire ou de former un analyseur syntaxique pour une langue où un dictionnaire n'est pas déjà disponible. Ceci nous permet de penser à un pont pouvant relier ces deux langues, nous utilisons pour cela un alignement bilingue de mots que nous expliquons maintenant.

3.1. Les alignements

L'alignement de mots dans un corpus parallèle, c'est à dire un corpus de deux textes dont la i ème phrase source est traduite à la i ème phrase cible, a été proposé par l'équipe d'IBM [Brown et al., 93] dans le cadre de la traduction statistique. Un alignement de mots consiste simplement à relier entre eux les mots d'une paire de phrases qui sont en relation de traduction.

Les figures⁵ 7, 8, 9 et 10 nous montrent plusieurs alignements qui sont tous acceptables avec des probabilités différentes (l'alignement de la figure 8 est moins probable que celui de la figure 7).

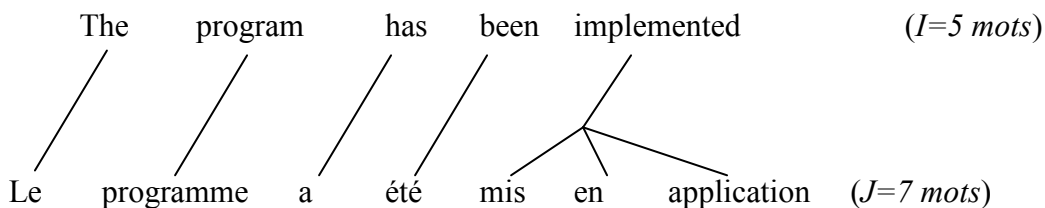


Figure 7: Un alignement dont chaque mot français est aligné à un seul mot anglais.

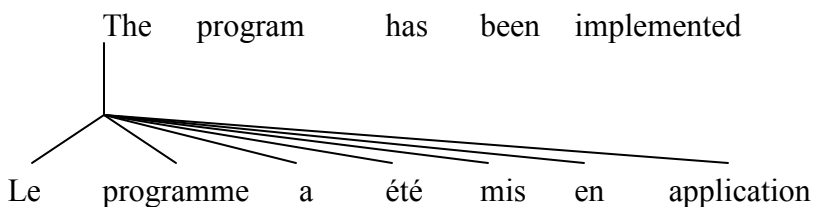


Figure 8: Un autre alignement possible de ces phrases mais moins probable.

⁵Exemples pris de [Brown et al., 1993].

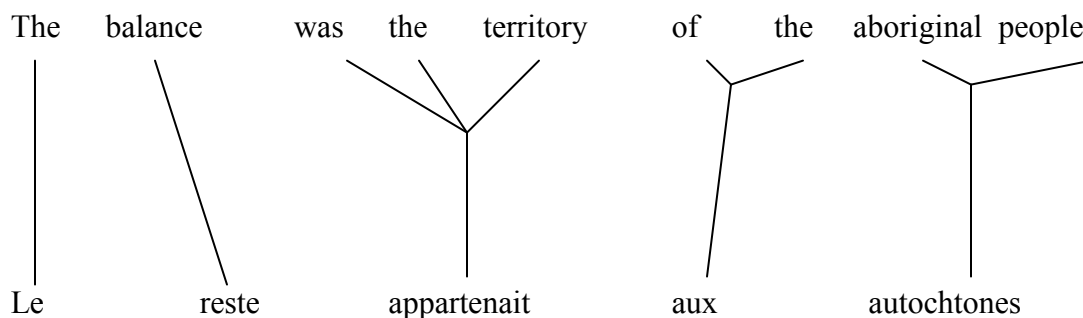


Figure 9: Un alignement dont chaque mot anglais est associé à un seul mot français.

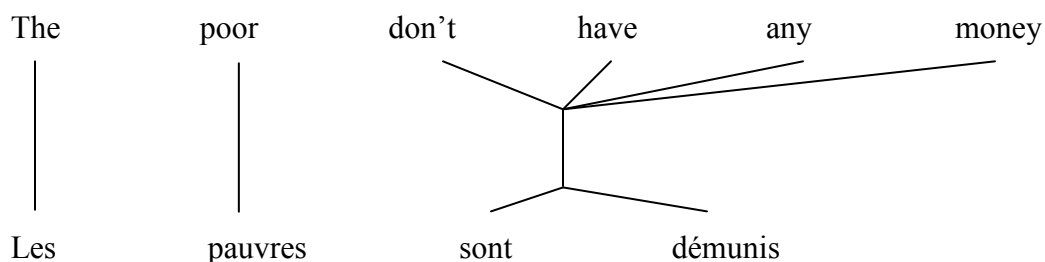


Figure 10: Un alignement dont un ensemble de mots français est connecté à un ensemble de mots anglais.

Nous montrons en figure 7 et 9 et 10 trois types d'alignements. Dans celui de la figure 7, chaque mot français est connecté à un et un seul mot anglais; Cependant un mot anglais peut-être associé à un ou plusieurs mots français. La figure 9 montre un alignement où un mot anglais n'est connecté qu'à un seul mot français. Enfin, la figure 10 présente un alignement général où plusieurs mots français peuvent se connecter à plusieurs mots anglais. Dans cet exemple, les quatre derniers mots anglais ensemble (*don't have any money*) sont alignés aux deux mots français (*sont démunis*).

Les modèles que nous utilisons ne reconnaissent que les alignements pour lesquels un mot français est aligné à au plus un mot anglais (figure 7 et 8). Un alignement peut donc être représenté comme un vecteur affectant, à chacune des positions françaises, une position anglaise. Par exemple, (*Le programme a été mis en application | the (1) program (2) has (3) been (4) implemented (5,6,7)*) représente l'alignement de la figure 6.

L'ensemble des mots français associés à un mot anglais est dénommé par "cept", et si un mot anglais n'est connecté à aucun mot français on dit que le "cept" est vide. Ce cept vide qui n'a pas de position, est connecté par convention à la position 0 et on le note e_0 .

Exemple: (*J'applaudis à la décision* | $e_0(3)$ *I(1) applaud(2) the(4) decision(5)*), on voit que "à" est connecté au mot NULL qui est par convention à la position zéro de la phrase anglaise.

3.1.1. Les modèles d'alignements proposés par IBM

[Brown et al, 1993] propose cinq modèles de traduction 1, 2, 3, 4 et 5. Chaque modèle a sa propre prescription pour aligner les mots entre les deux langages ainsi que pour modéliser la probabilité conditionnelle $P(f|e)$, la probabilité qu'une phrase f soit la traduction d'une phrase e . Toute paire de mots reliés (e_i, f_j) du corpus d'entraînement est un paramètre du modèle, dont la probabilité est apprise automatiquement par un algorithme décrit dans l'article. Nous montrons dans la figure 11 un exemple de paramètres obtenu après l'application de l'alignement à partir d'un corpus parallèle.

The (*le*, 0, 18) (*la*, 0.15) (*de*, 0.12)
 Les paires (e_i, f_j) avec leurs probabilités seront alors :
 (*the*, *le*, 0.18)
 (*the*, *la*, 0.15)
 (*the*, *de*, 0.12)

figure 11 : les paramètres des paires de mots alignés.

Notation: Soit E l'ensemble de phrases anglaises, F l'ensemble de phrases françaises. $e^I = e_1, \dots, e_I$ et $f^J = f_1, \dots, f_J$ sont deux phrases particulières de E et F . $A(e^I, f^J)$ est l'ensemble des alignements liant une phrase anglaise donnée à une phrase française. On note par $P(F=f^J, A=a | E=e^I)$ la probabilité de jointe de f^J et d'un alignement particulier a .

Alors :

$$P(f_j^J | e_i^I) = \sum_{a \in A} P(f_j^J, a | e_i^I).$$

Figure 12 : la probabilité qu'une phrase f soit la traduction de e.

Bien que nous ne nous intéressons pas directement dans ce travail à la tâche de traduction, nous décrivons dans des grandes lignes le principe sous jacent proposé par [Brown et al., 93], ce qui permettra de mieux comprendre la notion d'alignement que nous utilisons dans cette étude.

Le point de départ de la traduction statistique peut être exprimé par l'équation suivante :

$$\hat{e} = \underset{e}{\operatorname{Argmax}} P(e^1 | f^1) = \underset{e}{\operatorname{Argmax}} P(e^1) \times P(f^1 | e^1).$$

Où $P(e^1)$ est un modèle de langue qui contrôle la bonne formation d'une phrase du langage cible, et $P(f^1 | e^1)$ est un modèle de traduction (notons qu'il s'agit d'un modèle inversé; L'information conditionnée étant inconnue). Argmax représente l'opération de décodage et consiste à extraire, parmi toutes les séquences possibles de la langue cible F , celle qui maximise le produit des deux distributions. Nous concentrons notre exposé sur la distribution $P(f^1 | e^1)$ qui capture le modèle d'alignement que nous utilisons ici.

La distribution dans la figure 12 peut être obtenue par une marginalisation sur toutes les alignements « a » possibles entre deux phrases e et f . Rappelons que dans notre cas, un alignement peut être représenté par un vecteur a_j^1 dont la coordonnée a_i indique une position source j telle que f_j est la traduction de e_{a_i} (dans le cas d'un mot français non aligné à un mot anglais particulier, il est associé à un mot fictif e_0).

- **Modèle de traduction probabiliste IBM1**

Dans IBM1, la probabilité d'alignement d'un mot anglais en position i avec un mot français en position j est indépendante des positions i et j . Toutes les positions sont possibles et équiprobables.

Donc,

$$P(f^J, a | e^1) = \prod_{j=1}^J t(f_j, e_{a_j})$$

Où $t(f_j, e_{a_j})$ est appelée la probabilité de transfert et peut être vue comme une version probabilisée d'une entrée de dictionnaire.

- **Modèle de traduction probabiliste IBM2**

Cependant l'expérience sur un corpus bilingue (*français/ anglais telles que les phrases sources et cibles soient de 8 mots*) montre que dans 70% de cas (Voir tableau 2), les mots correspondants ont les mêmes positions, dans 10% des cas les mots diffèrent d'une position et dans 5% ils diffèrent de deux positions.

Source Cible	1	2	3	4	5	6	7	8
1	0.85	0.1	0.5					
2	0.15	0.7	0.1	0.05				
3	0.05	0.1	0.7	0.1	0.05			
4		0.5	0.1	0.7	0.1	0.05		
5			0.05	0.1	0.7	0.1	0.05	
6				0.05	0.1	0.7	0.1	0.05
7					0.05	0.1	0.7	0.15
8						0.05	0.1	0.85

Tableau 2: Les probabilités des alignements sur un corpus de phrases de 8 mots (anglais /français).

IBM2 remédie à cette simplification à outrance en introduisant des probabilités d'alignement $a(i|j, J, I)$: la probabilité qu'un mot anglais en position i soit associé à un mot français en position j , sachant les longueurs respectives (I et J comptées en mot) des deux phrases considérées. On élabore alors la probabilité d'alignement $a(a_j|j, J, I)$ comme un nouveau paramètre pour le modèle 2 sous la contrainte stochastique.

$$\sum_{i=0}^I a(i|j, J, I) = 1$$

Dans le cas des modèles IBM2 que nous utilisons ici, une prescription simple à calculé pour $P(e^I | f^J)$ peut être faite :

$$P(f^J, a | e^I) = \prod_{j=1}^J t(f_j | e_{a_j}) a(i | j, J, I).$$

Ces deux distributions sont apprises par un algorithme décrit dans [Brown et al., 93] à partir d'un bitexte. Nous utilisons dans notre travail ces distributions pour calculer l'algorithme de viterbi.

3.2. Alignement de VITERBI

a. Principe

Quelques soit le modèle, il existe un alignement d'une paire (e, f) qui obtient le plus grand score selon le modèle considéré. Nous appelons cet alignement, **l'alignement de viterbi**: $V(f | e) = \underset{a}{\operatorname{argmax}} P(a | e, f) = \underset{a}{\operatorname{argmax}} P(a, f | e) / P(f | e) = \underset{a}{\operatorname{argmax}} P(f, a | e)$.

Pour les modèles IBM1&2, on trouve l'alignement de viterbi très facilement, chaque prédiction étant indépendante entre elles. L'opération de maximisation dans **IBM1** s'exprime par :

$$V = \underset{a}{\operatorname{argmax}} \prod_{j=1}^J t(f_j | e_j) \quad \text{et donc :}$$

$$V_j = \underset{a_j}{\operatorname{argmax}} t(f_j | e_j)$$

Dans le cas du modèle **IBM2** :

$$V = \underset{a}{\operatorname{argmax}} \prod_{j=1}^J t(f_j | e_{a_j}) a(a_j | j, J, I)$$

$$V_{a_j} = \underset{a_j}{\operatorname{argmax}} t(f_j | e_{a_j}) a(a_j | j, J, I)$$

Alors pour trouver le meilleur alignement du modèle 1, nous choisissons pour chaque mot français f_j (indépendamment) de la position anglaise i qui maximise le facteur de transfert, alors que dans le modèle 2 il suffit de maximiser en plus le facteur d'alignement (distorsion). Ceci se fait donc en une complexité $J*(I+1)$. Rappelons que la position e_0 est ajoutée à chaque phrase source pour expliquer les mots cibles non traduits.

Nous décrivons maintenant les types des fichiers qu'on utilise pour l'application de l'algorithme de viterbi et le fichier de sortie qu'on veut l'utiliser pour accomplir la tâche de la projection.

b. Fichiers d'entrée

a- Table de transfert (T-table)

Les paramètres de transfert sont représentés par une grande matrice creuse à deux dimensions et initialisés uniformément en donnant une probabilité de transfert à chaque paire de mots croisés au moins une fois dans une paire de phrases où les paramètres sont exprimés selon le format:

Source_id cible_id P(cible_id/source_id).

b- Vocabulaires

Puisque les paramètres sont exprimés en fonction de l'identité du mot, ce qui facilite la tâche de programmation en temps d'exécution et la place en mémoire. Alors, on a besoin d'un vocabulaire qui liste tous les mots en assignant des indices pour chacun d'eux. Deux vocabulaires existent pour la langue française et la langue anglaise, de taille respectivement, x et y mots différents :

mot-id mot

Où,

mot-id = L'identité du mot exprimé en entier

mot = Le mot lui-même

c- Tables d'alignement

Les deux derniers modèles de traduction 2 et 3 possèdent des tables d'alignement.

-Dans le modèle 2 : A-tables est exprimé selon ce format : $i j I J P(i | j, J, I)$.

Où,

i = La position dans la phrase source.

j = La position dans la phrase cible.

I = La longueur de la phrase source.

J = La longueur de la phrase cible.

Et $P(i | j, J, I)$ est la probabilité que le mot anglais de la position i sera à la position j dans une paire de phrases de longueur I et J .

c. Fichier de sortie

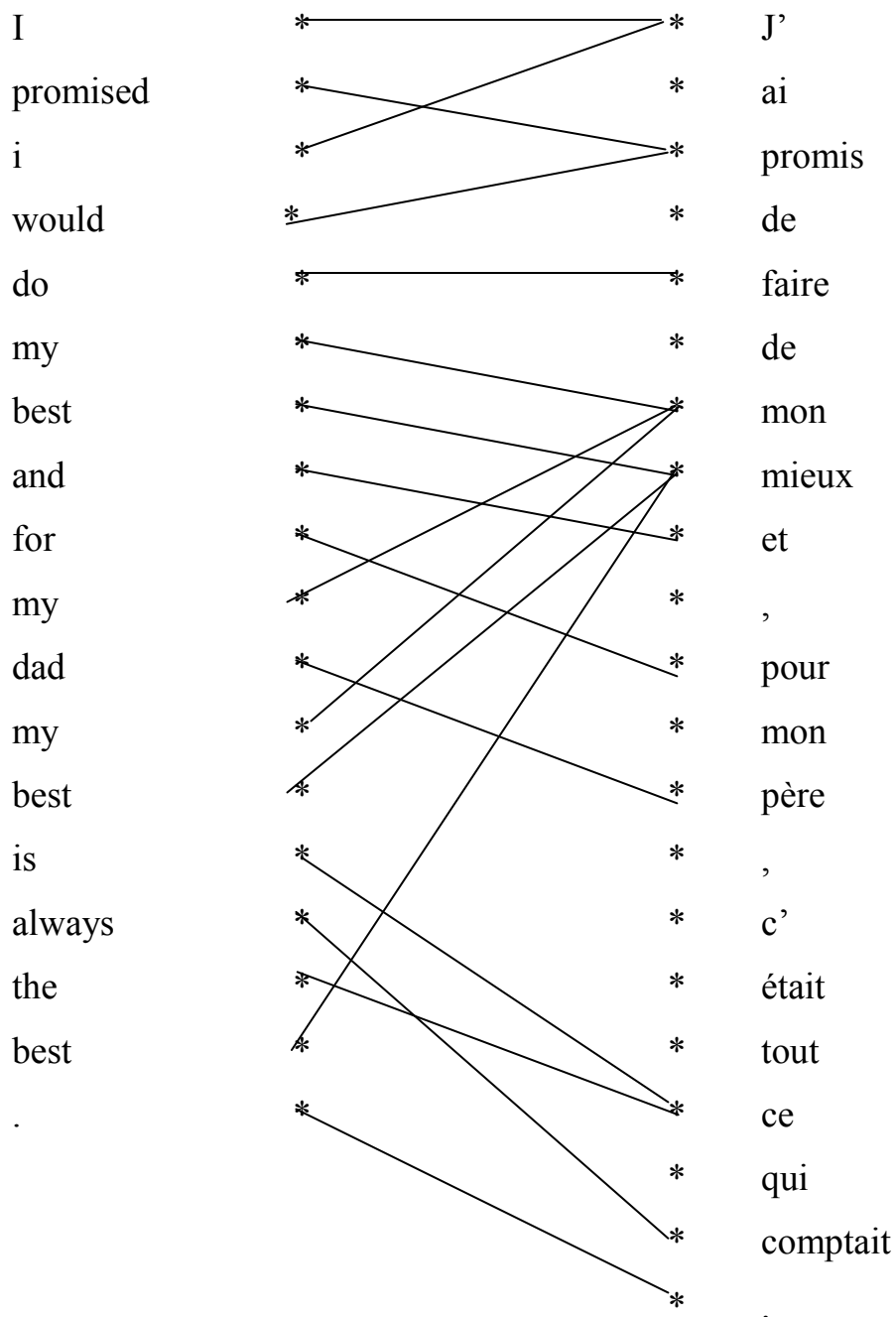
Le résultat est stocké sous forme d'un fichier qu'on utilise comme une entrée pour l'algorithme de projection. Ce sont les informations pour chaque phrase qui nous indique du quel mot anglais on doit extraire les liens pour le projeter suivant la masse calculée durant l'exécution de **Viterbi**. Le fichier est sous le format : *phrase-id source-phrase-id cible-phrase-id masse*

Où,

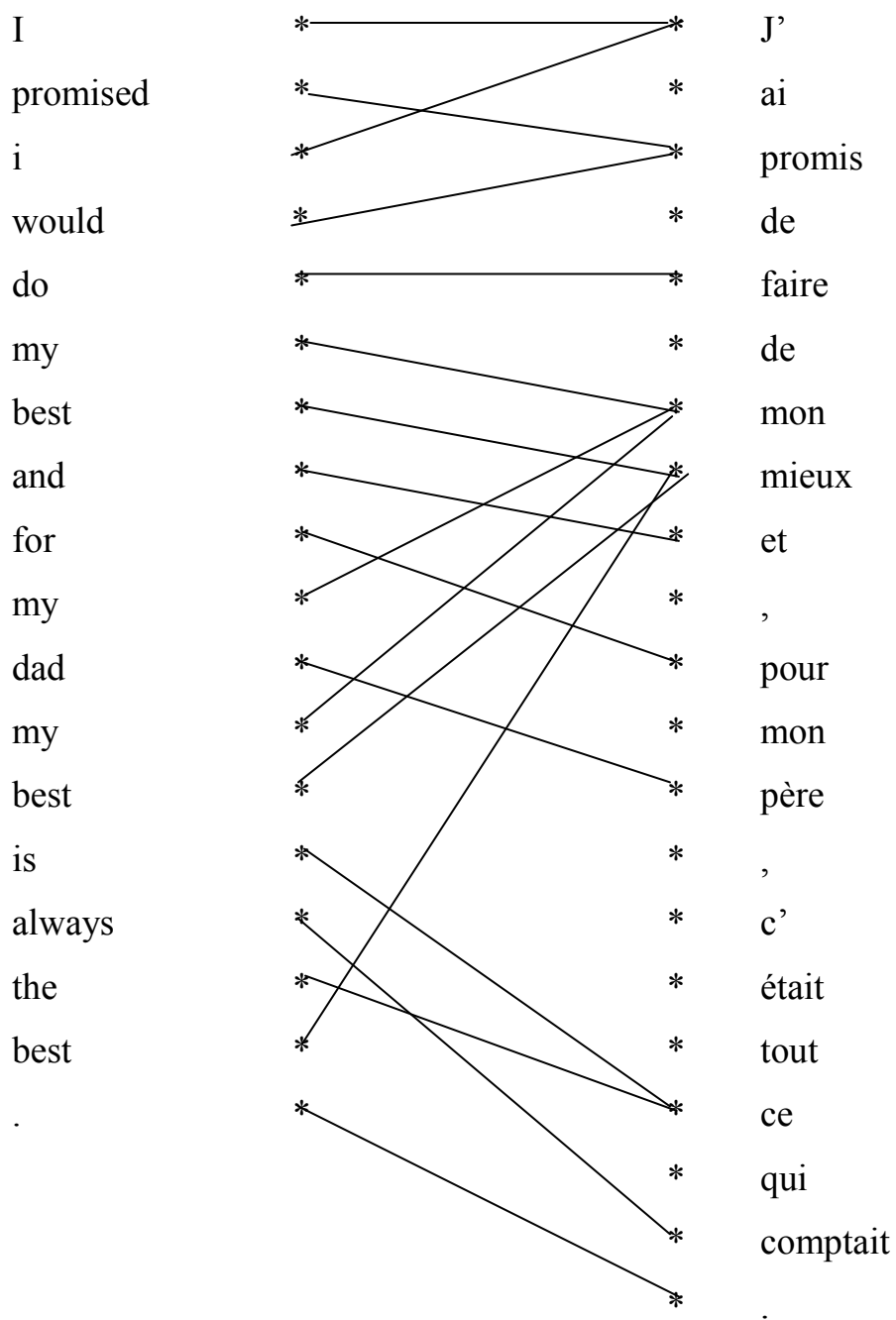
<i>phrase-id</i>	= Le numéro de la phrase
<i>source-phrase-id</i>	= La position de la phrase source
<i>cible-phrase-id</i>	= La position de la phrase cible
<i>masse</i>	= Le produit de la probabilité de transfert et la probabilité d'alignement du modèle 2.

3.3. Exemple⁶

Alignement de Viterbi obtenus par IBM1



⁶ Exemple pris du site : <http://www.iro.umontreal.ca/~felipe/IFT6010-Automne2004>

Alignement de Viterbi obtenu par IBM2

On peut observer que “my” est toujours aligné par le modèle *IBM1* au même mot “mon” (c’est assez normal car chaque position source (ici française) est équiprobable selon le modèle 1). Contrairement au modèle 2 qui prend en considération l’emplacement des mots dans la phrase en ajoutant la distribution $a(i / j, m, l)$. Nous constatons que le premier “my” sera associé au premier “mon” et les autres avec le “mon” qui est proche dans son emplacement à celui des autres “my”.

3.4. Conclusion

Dans ce chapitre, on a expliqué brièvement le principe de la traduction statistique et nous avons étudié d’avantage les modèles de traduction probabiliste et en particulier les deux premiers modèles proposés par [Brown et al. 1993].

Nous avons présenté aussi la différence entre les deux premiers modèles *IBM1* et *IBM2*, et le principe de la méthode d’alignement Viterbi appliqué sur *IBM2* et utilisé par notre étude. D’autre part, nous mettrons à l’épreuve un algorithme de projection directe qui utilise ce modèle, et fera l’objet du chapitre 5.

Chapitre 4

Corpus utilisé et évaluation

Nous avons décrit au chapitre 2 comment une Link-Grammar pouvait à l'aide d'un dictionnaire anglais identifier les liens de dépendance entre les mots d'une phrase à analyser. Nous venons également de décrire comment un corpus bilingue parallèle pouvait être aligné au niveau des mots.

Nous nous intéressons donc de savoir si la projection d'un dictionnaire anglais de la Link-Grammar peut être faite via alignement vers un dictionnaire français et fournir de bonnes analyses de phrases françaises. Nous avons pour cela utilisé différents corpus et implémenté une mesure de la qualité d'une analyse grammaticale que nous décrivons dans ce chapitre.

4.1. Corpus – Bitexte

Nous avons préparé et utilisé deux corpus pour notre étude. Le premier, qui est le point de départ de l'entraînement qu'on désigne par « bitexte ». Un bitexte est un corpus bilingue parallèle (un texte dans une langue de départ et sa traduction) ou les liens de traduction entre les phrases ou groupes de phrases sont explicites.

Le corpus des débats parlementaires canadiens (connu sous le nom de Hansard) a été utilisé dans notre étude pour entraîner les paramètres des modèles de traduction. Ce corpus est constitué de 1 639 250 paires de phrases, 29 547 933 mots anglais et 31 826 112 mots français. Les tailles des vocabulaires anglais et français sont respectivement de 103 830 et de 83 106 mots différents.

Nous avons utilisé ce corpus pour entraîner les distributions nécessaires pour le calcul des alignements de mots. Le corpus a été aligné par la suite au niveau des mots selon le procédé décrit au chapitre 3.

4.2. Fichier de Test

Nous avons également réuni un corpus de test, différent du corpus d'entraînement pour tester nos différentes expériences de projection. Il nous fallait pour cela disposer d'un corpus dont les analyses syntaxiques sont connues.

Nous avons pour cela travaillé sur un extrait du corpus Corfran [Abeille, 1999] auquel le RALI⁷ a un droit d'accès. Le corpus contient 2 000 phrases et un vocabulaire de 9 394 mots différents. Les phrases ont été annotées, sous forme d'arbre, manuellement par des experts linguistiques. Un exemple d'un tel arbre est proposé en figure 13.

[SENT [NP Un [PP des [NP éléments [AP essentiels AP].....
Un des éléments essentiels.....

Figure 13 : une tranche d'une phrase de la corpus de test.

4.3. Filtrage

Pour rendre la projection efficace, nous avons effectué des pré-traitements sur notre bitexte d'entraînement et le fichier de test, mentionnés, afin d'isoler deux sous corpus.

1. Nous avons éliminé les paires de phrases du corpus parallèle, qui ne peuvent pas être analysées par le système d'analyse que nous avons utilisé, le « Link-Parser ». Nous pensons en effet qu'il est prématuré à ce stade de nous intéresser à des phrases dont l'analyse complète ne peut être faite.
2. Pour des raisons de fiabilité et de performance, nous avons appliqué plusieurs filtres sur les phrases du corpus bitexte pour ne garder que les phrases ayant leurs mots inclus dans le vocabulaire du corpus de test. L'importance de cette phase vient de l'inutilité de projeter des mots qui ne seront pas utilisés dans le corpus de test.

Illustrons par des schémas ces traitements, et soit « filtrage » une fonction prenant un ensemble de phrases et un vocabulaire de mots en entrée, afin d'extraire de cet

⁷ Recherche Appliquée en Linguistique Informatique à l'Université de Montréal.

ensemble les phrases dont les mots sont éléments de ce vocabulaire. Le résultat de cette fonction est un fichier construit de ces phrases, qui va constituer le nouveau corpus de test.

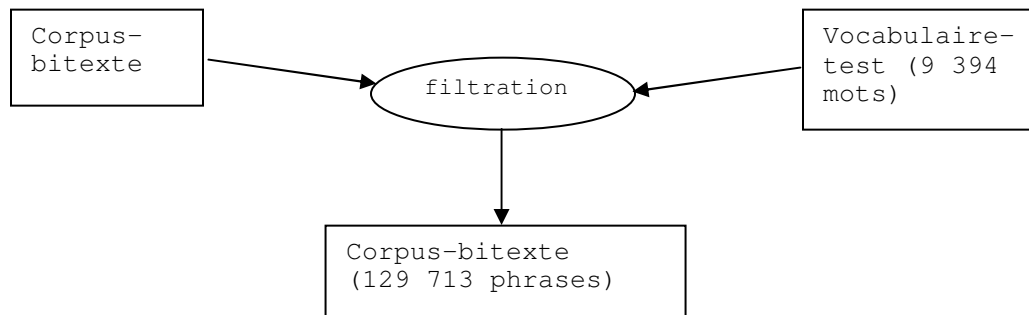


Figure 14 : filtration du corpus à l'aide du vocabulaire de la référence.

Nous avons diminué ainsi le corpus bitexte jusqu'à 129 713 phrases avec un vocabulaire de 6 745 mots différents (figure 14).

Nous appliquons maintenant la « filtration » sur le corpus de test, mais en utilisant le vocabulaire du corpus bitexte (figure 15). Alors, des 2 000 phrases de test, il en reste seulement celles ayant leurs mots éléments au vocabulaire du corpus bitexte (6 745 mots).

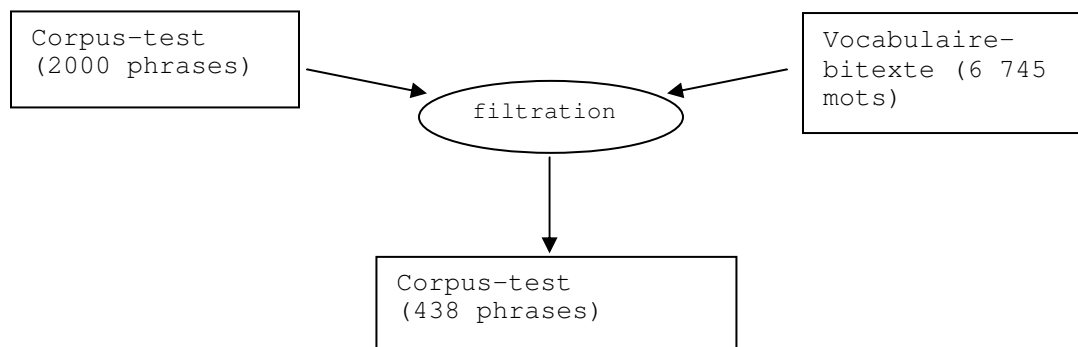


Figure 15: filtration de la référence à l'aide du vocabulaire du corpus filtré.

Un nouveau corpus de test est ainsi formé de 438 phrases et un vocabulaire de 2 370 mots (figure 15). De même on filtre de nouveau les 129 713 phrases du corpus, mais maintenant sur les 2 370 mots du vocabulaire du corpus de test. Arrivant à un corpus bitexte de 32 570 phrases (figure 16), que nous utiliserons dans notre approche.

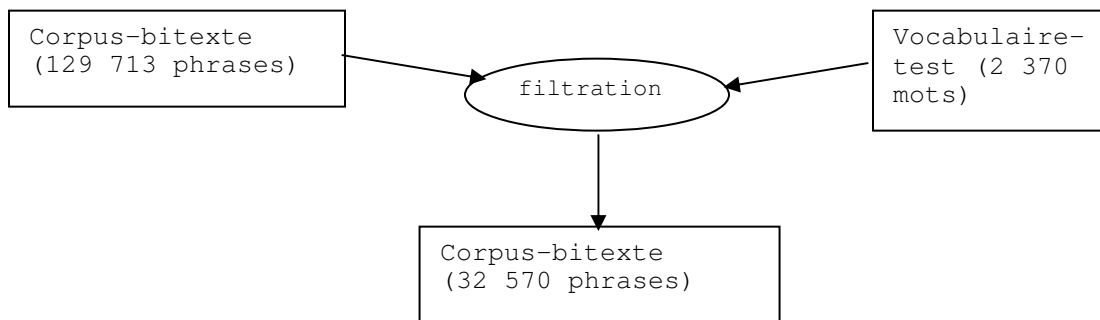


Figure 16: filtration du corpus à l'aide du vocabulaire de la référence.

Pour bien illustrer la procédure de filtration, nous considérons les deux phrases suivantes : Le chien poursuit le chat // La voiture poursuit le chien, du corpus bitexte du coté français, et soit V le vocabulaire du corpus du test, constitué par les mots français suivants : $V = \{\text{Le, chien, poursuit, le, chat, pomme, arbre, voiture}\}$.

En appliquant la filtration sur ces deux phrases, on aura un nouveau corpus, formé dans ce cas de la phrase « Le chien poursuit le chat » (tous ses mots sont inclus dans V). De l'autre coté, la deuxième phrase sera rejetée parce que « La » n'est pas incluse dans le vocabulaire.

Et ainsi de suite afin de réduire le nombre de phrases de bitexte de 32 570 phrases et environ 2 073 mots français différents et 200 phrases du corpus de test.

De plus, pour réduire les temps de calculs, nous analysons les phrases du corpus bitexte ne dépassant pas une longueur donnée (70 mots dans nos expériences). Enfin, pour les mêmes raisons citées, nous avons constaté que des phrases banales du corpus de test n'ont pas de sens sémantique pourraient affecter les résultats de l'analyse, et alors elles seront éliminées. Les phrases trop longues (plus grandes que 40 mots) ne sont pas considérées non plus, nous laissant 135 phrases pour les tester.

4.4. Évaluation⁸

Nous évaluons nos différentes techniques de projection de dictionnaire par leur aptitude à produire les relations annotées manuellement dans le corpus de test.

Nous utilisons pour cela notre implémentation de l'évaluateur d'analyse, le « PARSEVAL » qui utilise une métrique, appelée le « GEIG »-« Grammar Evaluation Interest Group »- ou simplement la métrique PARSEVAL. Ces métriques ont été proposées pour évaluer la qualité d'analyseurs syntaxiques [Black et al., 1991]. La nature de ces métriques a été fortement influencée par le besoin de permettre des sorties des différents analyseurs des groupes de recherche, basée sur des schémas d'analyse très divers, pour être comparé sur un pied d'égalité.

On se sert, pour l'appliquer, d'une référence, c'est à dire d'un corpus arboré manuellement (expliqué dans la partie Corpus de ce chapitre), qui représente le but à atteindre (par les concepteurs de grammaire).

Les métriques de PARSEVAL se basent sur trois mesures : précision (*precision*), rappel (*recall*) et parenthésage croisé (*crossing bracket*).

$$\text{Précision} = \frac{|\text{brackets candidats corrects}|}{|\text{brackets candidats}|}$$

$$\text{Rappel} = \frac{|\text{brackets candidats corrects}|}{|\text{brackets référence}|}$$

Crossing = moyenne des parenthésages croisant ceux de la référence.

$$\text{Croisement} : \exists [i, j] \text{ et } [i', j'] / i < i' \leq j < j'$$

À l'origine ces mesures ne tenaient pas compte de l'étiquette attachée à un constituant. Lorsque l'on tient compte de cette étiquette, on parle alors de *labeled precision* et de *labeled recall*. À cause de l'incompatibilité des étiquettes du système du *Link-Parser*

⁸ Une grande partie de cette explication est prise du cours de IFT6010-Automne 2004, enseigné par Philippe Langlais

et celles de notre corpus de test, nous avons utilisé des mesures sans tenir compte des étiquettes.

Nous pouvons illustrer ces mesures sur l'exemple suivant:

Arbre de référence : [Ils [[mangent] le dîner]]

[1, 4] [2, 2] [2, 4]

Arbre de candidat : [[Ils [mangent]] [le dîner]]

[1, 4] [1, 2] [2, 2] [3, 4]

précision : $2/4 = 50\%$

rappel : $2/3 = 66.7\%$

crossing = 1 ([2,3] dans la référence et [1,2] dans le candidat)

Nous utilisons également la F-mesure, moyenne harmonique de la précision et du rappel [Van Rijsbergen, 1979].

$$\begin{aligned} \text{F - mesure} &= \frac{2 * \text{précision} * \text{rappel}}{\text{précision} + \text{rappel}} \\ &= 57.15\% \end{aligned}$$

A la fin de tout résultat, et indépendamment de l'application lancée, notre but est de maximiser les précisions et rappels et de minimiser le crossing-bracket. Ce sont les métriques qui doivent être étudiées et analysées dans toutes les expérimentations décrites au chapitre 5. Notons que dans cette étude, nous calculons ces taux (précision, rappel, etc.) sur les seules phrases que notre système a réussi à analyser. Ceci peut laisser à tort penser que notre système est meilleur qu'il ne l'est en pratique. Pour cette raison, nous rapportons dans toutes nos expériences le pourcentage de phrases analysées. L'ensemble de ces métriques constitue donc un moyen efficace de mesurer la qualité de nos techniques de projection.

Chapitre 5

Protocole général et expérience

Nous décrivons dans ce chapitre plusieurs méthodes de projection que nous avons implémentées, et testées. Plus particulièrement, nous nous intéressons à montrer les limites de chaque hypothèse de projection étudiée et à proposer des façons de palier à ces limites.

Nous commençons par décrire notre protocole expérimental et décrivons dans la suite, différentes approches à la projection que nous avons implémentée. Nous proposons que les bonnes questions soient : dans quelle mesure la projection est-elle vraie? Et comment est-elle utile quand elle est préservée?

Dans le reste du chapitre, nous nous contentons de répondre à la première question empiriquement en considérant les relations syntaxiques et les alignements entre des paires de phrases dans les deux langues (anglais et français).

5.1. Protocole de Projection.

Dans notre cadre expérimental, plusieurs approches à la projection ont été étudiées afin de trouver laquelle parmi elles, nous donnaient les meilleurs résultats. Toutes ces approches ont été testées dans le même cadre expérimental que nous décrivons maintenant.

Notre point de départ dans ces expériences est un bitexte « B » de $|B|$ paires de phrases :

$$B = \{ S^t = \langle E^t, F^t \rangle / t \in [1, |B|] \}$$

Où, $F^t = f_1 \dots f_l$ est une phrase française de l mots qui constitue avec $E^t = e_1 \dots e_m$ une phrase anglaise de m mots, une paire de notre bitexte.

Nous avons également à notre disposition un alignement au niveau des mots. Dans notre cas, un alignement de deux phrases F^t et E^t est un vecteur « a », tel que :

$$a = \{(a_k) / \exists A_t(e_{ak}^t, f_k^t), k \in [1, l], a_k \in [0, m]\}$$

Où, $A_t(e_{ak}^t, f_k^t)$ indique un lien entre e_{ak}^t et f_k^t tel qu'identifié par l'algorithme d'alignement décrit dans le chapitre 3. On rappelle que la phrase source est étendue d'un mot (d'indice 0) rendant compte des mots cibles (français ici) non alignés à un mot source particulier.

En plus de l'aligneur, un système d'analyse syntaxique anglais est fourni qui crée des relations grammaticales pour chaque mot d'une phrase anglaise (voir chapitre 2). Une relation syntaxique dans notre travail est un lien établi entre deux mots différents d'une phrase anglaise. Tout lien formé est un élément d'un ensemble de « *Tag* » utilisé par la « *Link-Grammar* » dont le nombre peut arriver jusqu'à 610 « *Tags* » différents. On peut représenter cet ensemble comme suit :

$G(e_{i1}^t, e_{i2}^t) = \{\Phi \mid tg \mid tg \in T\}$, T étant l'ensemble des tags. Cet ensemble est égal à Φ s'il n'existe pas une relation reliant les deux mots e_{i1}^t et e_{i2}^t ensembles.

Avant de formaliser la projection, nous définissons maintenant la notion des liens :

Soit $L(e_i^t) = \{ \cup G(e_{i1}^t, e_{i2}^t) / i1 \in [1, m], t \in [1, |B|] \text{ et } \exists i2 \in [1, m] / G(e_{i2}^t, e_{i2}^t) = tg \}$, l'ensemble des liens (tags) du mot e_i^t de la phrase E^t , formés durant l'analyse de cette phrase.

On nomme $L(f_v) = \{ \cup L(e_i^t), t \in [1, |B|], e_i^t \in E^t, v \in [1, |V|] / \exists A_t(f_v, e_i^t) \}$, l'ensemble de tous les liens que le mot f_v entretient sur notre corpus de projection. V étant le vocabulaire français de $|V|$ mots différents.

Nous proposons en figure 17 une description algorithmique de la notion de projection :

```

Projection ( $S^t$ )
{
  pour chaque  $S^t = \langle E^t, F^t \rangle$ 

    analyser ( $E^t$ );
    => formation de  $L(e^t_i) \forall i \in [1, m]$ 

    aligner ( $E^t, F^t$ );
    => formation de  $a = \{a_k\}$ 

    pour chaque  $f^t_j \in F^t, \forall j \in [1, l]$ 

      si ( $a_j \neq 0$ ) (c.à.d.  $f_j$  est aligné à un mot anglais)
        association-liens ( $f^t_j, e^t_{a_j}, L(e^t_{a_j})$ );
        => formation de  $L(f_v)$ 
}

```

Figure 17 : la fonction de projection dans l'algorithme général.

Les deux premières fonctions « **analyser** » et « **aligner** » sont des appels à des algorithmes déjà implémentés. Le premier défini dans le système utilisé « *Link-Parser* » et le deuxième est notre implémentation de l'algorithme de *Viterbi* (décrit dans le chapitre 2). Quant à la troisième fonction, « **association-liens** », elle constitue le cœur de ce chapitre et une implémentation possible de cette fonction sera décrite pour chacune de nos tentatives.

Après avoir associé tous les liens à un mot français, ce dernier constitue une entrée dans le dictionnaire projeté. Ce dictionnaire peut être représenté par l'ensemble suivant :

$$D(f_v) = \{(f_v, L(f_v)) / v \in [1, V]\}$$

Nous nous retrouvons maintenant dans une situation telle qu'un dictionnaire français est formé à l'aide de la projection, et est prêt à être utilisé par le système de *Link-Parser*. Nous pouvons alors mesurer la performance de l'analyseur projeté à l'aide des métriques PARSEVAL décrites au chapitre 4.

Nous illustrons le mécanisme général de ce protocole sur un exemple que nous compléterons dans nos différents essais.

Considérons la paire de phrases suivante :

$\langle E_t : \textit{The dog chases the cat} , F_t : \textit{Le chien poursuit le chat} \rangle$

Nous considérons toujours que $t = 1$, c'est-à-dire, nous appliquons la projection pour la première paire de phrases dans le corpus parallèle.

Analyse syntaxique de E^t

$$\begin{array}{cccccc} & & & +---O_s---+ & & \\ +---D_s---+ & +---S_s---+ & & & +---D_s---+ & \\ | & | & | & | & | & | \\ \textit{The} & \textit{dog.n} & \textit{chases.v} & \textit{the} & \textit{cat.n} & \end{array}$$

$$L(e^t_{i_1}) = \{ \cup G(e^t_{i_1}, e^t_{i_2}) / t \in [1, |B|], i_1 \in [1, m] \text{ et } i_2 \in [1, m] / G(e^t_{i_1}, e^t_{i_2}) = tg \}$$

Pour $t=1$,

$$L(e^1_1) = L(\textit{The}) = \{D_s+\}$$

$$L(e^1_2) = L(\textit{dog.n}) = \{D_s-, S_s+\}$$

$$L(e^1_3) = L(\textit{chases.v}) = \{S_s-, O_s+\}$$

$$L(e^1_4) = L(\textit{the}) = \{D_s+\}$$

$$L(e^1_5) = L(\textit{cat}) = \{D_s-, O_s-\}$$

Aligner (E^t, F^t)

Source :	The	dog	chases	the	cat
	↓	↓	↓	↓	↓
Cible :	Le	chien	poursuit	le	chat

$$a = \{ (a_k) / \exists A_t(e^t_{a_k}, f^t_k) , k \in [1, 5] , a_k \in [0, 5] \}$$

$$a = \{ 1 , 2 , 3 , 4 , 5 \}$$

Après avoir formé l'ensemble $L(e_i^t)$ et les associés anglais « e_{ak} » de chaque mot français sont formés aussi, l'information suffisante pour une projection est construite. La fonction « **association-liens** » donc, transfère les liens $L(e_i^t)$ du mot anglais « e_{ak} » vers le mot français « f_v » qui lui est aligné. l'ensemble $L(f_v)$ de chaque mot français serait alors :

$$L_v(f_v) = \left\{ \bigcup_{ak \in [1,m]} L(e_{ak}) \mid \exists A_t(f_v, e_{ak}), v \in [1, V] \right\}$$

$ak = 1 \Rightarrow L_1(\text{Le}) = \cup L(e_1) = \{L(\text{The})\} = \{Ds+\}$	
$ak = 2 \Rightarrow L_2(\text{chien})$	$= \{Ds-, Ss+\}$
$ak = 3 \Rightarrow L_3(\text{poursuit})$	$= \{Ss-, Os+\}$
$ak = 4 \Rightarrow L_4(\text{le})$	$= \{Ds+\}$
$ak = 5 \Rightarrow L_5(\text{chat})$	$= \{Ds-, Os-\}$

Nous avons formé ainsi les liens de chaque mot français. Nous pouvons construire alors un dictionnaire français qui est constitué, dans notre exemple, des entrées suivantes :

```

Le      : Ds+;
chien   : Ds- & Ss+ ;
poursuit : Ss- & Os+ ;
le      : Ds+ ;
chat    : Ds- & Os- ;
mange   : Ss- & Os+ ;

```

Nous admettons que notre corpus de test contient la phrase : ***Le chien mange le chat***. Si le système d'analyse de la Link-Grammar utilise le dictionnaire projeté, il peut alors produire l'analyse suivante :

```

          +----O_S-----+
+-D_S-+----S_S---+      +-D_S-+
|      |           |      |      |
Le  chien  mange  le  chat

```

Avec une représentation arborée équivalente:

```

(S (NP Le chien)
  (VP mange
    (NP le chat)))

```

Enfin, en analysant toutes les phrases du corpus de test, chacune est évaluée à l'aide de la métrique décrite dans le chapitre 4. Des valeurs de la précision et le rappel sont ainsi fournies, afin de nous aider de mieux comprendre les erreurs et d'améliorer la performance de notre projection.

Notre corpus de projection contient de nombreuses paires de phrases. De nombreux mots apparaissent donc plusieurs fois. Un certain nombre de questions se posent alors. Est-il pertinent de projeter tous les mots? Peut-on projeter les liens anglais sur les mots français directement? Est-ce que la présence de bruit dans l'alignement nuit à la projection? Autant de questions auxquelles nous apportons des réponses dans la suite de ce chapitre.

5.2. Le mot le plus fréquent

Description:

Rappelons que les alignements *IBM* tels que nous les utilisons n'autorisent à une occurrence d'un mot cible (français) qu'un seul mot source (anglais). Il est cependant tout à fait possible que dans deux paires de phrases, deux occurrences d'un même mot français se retrouvent alignées à deux mots anglais différents. Donc en général, pour un mot français, il existe un ou plusieurs mots anglais qui lui sont associés dans notre Bitexte.

Plus le nombre de paires de phrases est grand et plus la liste des mots anglais associés à un mot français donné augmente. Une entrée française risque donc de se trouver polluée de liens nuisant à l'analyse. Dans cette variante, nous décidons de ne projeter que les liens du mot anglais le plus fréquemment associé à un mot français donné.

Dans cette approche nous profitons du nombre de fois qu'un mot anglais est aligné au même mot français, ce que nous appelons la « fréquence d'alignement » d'un mot anglais. Nous obtenons ainsi un dictionnaire français qui sera utilisé par le système « *link-parser* », afin d'analyser les phrases de test.

```
rester:(stay.v , 10) , (remain.v, 4) , (leave.v, 3) ...
maintenir:(maintain.v , 6), (sustain.v , 3);
```

figure 18 : liste des mots avec leurs fréquences après alignement.

Nous montrons en figure 18 les fréquences d'alignement, dans notre corpus de projection, des mots « *rester* » et « *maintenir* ». Les entrées respectives de ces deux mots sont celles de « *stay.v* » et « *maintain.v* » qui sont les mots anglais qui leur sont les plus fréquemment alignés.

Décrivons maintenant les ensembles utilisés dans cette variante :

Soit $L(f_v, e) = \{ \cup L(e^t_i) / \exists e, \exists i \in [1, m] / e = e^t_i \text{ et } A(e^t_i, f_v), \forall t \in [1, |B|], \}$, l'ensemble des liens de chaque mot anglais « *e* » alignés à un seul mot français « *f_v* ».

Notons ici, que les liens dont nous parlons ne sont que les liens qui partent de chaque mot anglais « *e* » lors de l'analyse de la phrase contenant ce mot. C'est-à-dire, les liens qu'il utilise seulement pour se connecter avec les autres mots d'une même phrase.

Soit $tf(f) = \{(e, c) / \exists t, \exists a_k \in [1, m] / A(e^t_{a_k}, f_k) \text{ et } e^t_{a_k} = e\}$, l'ensemble des mots anglais « *e* » qui sont alignés à un mot français donné « *f* », avec « *c* » la fréquence d'alignement de « *e* » dans le corpus.

Cette fréquence « *c* » est le nombre de fois qu'un mot anglais est aligné à un mot français, et elle est incrémentée à chaque fois qu'il existe un alignement entre ces deux mots. Cette fréquence est représentée par la formule suivante :

$$c = \sum_{t=1}^{|B|} \sum_{ak=1}^m \delta(A(e^t_{ak}, f)) , \quad \delta(x) = 1 \text{ si } x \text{ est vrai}$$

$$\delta(x) = 0 \text{ si non}$$

Dans cette approche, la fonction « **association-liens** » projette les liens $L(f_v, e)$ du mot anglais *e*, le plus fréquemment de l'ensemble $tf(f)$, vers le mot français f_v qui lui est associé.

Prenons un exemple et montrons comment se construisent les entrées de dictionnaire sur la paire de phrases suivante :

Exemple 2 :

< $E^t = I \text{ bring the dog with me}$ $F^t = J' \text{ amène le chien avec moi}$ >

après
analyse =>
de E^t

$L(e^1_1) = L(I.p)$	=	{Sp*i+}
$L(e^1_2) = L(bring.v)$	=	{Sp*i-, Os+, MVp+}
$L(e^1_3) = L(the)$	=	{Ds+}
$L(e^1_4) = L(dog.n)$	=	{Ds-, Os-}
$L(e^1_5) = L(with)$	=	{MVp-, J+}
$I.(e^1_6) = I.(me)$	=	{.T-}

L'ensemble des mots anglais e associés au mot français f avec leur fréquence d'alignement, est illustré dans la figure 19 :

$tf(J')$	$= \{(I.p, 1)\}$
$tf(amène)$	$= \{(bring.v, 1)\}$
$tf(le)$	$= \{(the, 1)\}$
$tf(chien)$	$= \{(dog.n, 1)\}$
$tf(avec)$	$= \{(with, 1)\}$
$tf(moi)$	$= \{(me, 1)\}$

Figure 19 : l'ensemble $tf(f)$ formé par les mots anglais et leur fréquence d'alignement.

Le dictionnaire sera constitué alors, des mots de la phrase française de l'exemple 2 avec les liens des mots anglais appartenant à $tf(f)$:

J'	: (Sp*i+);
amène	: (Sp*i- & MVp+);
le	: (Ds+);
chien	: (Ds- & Os-);
avec	: (MVp- & J+);
moi	: (J-);

Si maintenant notre corpus est étendu de la paire de phrase de l'exemple-1, alors l'ensemble $tf(f)$ pour le mot «chien», qui se forme suivant les mêmes étapes devient :

$$tf(chien) = \{(dog.n, 2)\}$$

Nous constatons deux occurrences du mot «dog.n» aligné au mot «chien». Mais puisqu'il n'y a qu'un seul mot dans l'ensemble $tf(chien)$, alors par évidence, il est le mot qui a la plus grande fréquence. L'ensemble des liens du mot «chien» sera l'union des liens du mot «dog.n» utilisés dans les deux occurrences.

Donc,

$$L(chien, dog.n) = \{\cup L(dog.n)\}$$

$$L(chien, dog.n) = \{(Os- \& Ds-) \text{ or } (Ds- \& Ss+)\}$$

L'entrée dans le dictionnaire projeté pour le mot « chien » devient alors:

chien : (Os- & Ds-) or (Ds- & Ss+);

Nous ajoutons maintenant dans notre corpus de projection la paire de phrases de l'exemple 3 :

Exemple 3 :

< E^t = *He killed the monster* , F^t = *Il a tué le chien* >

Où, « chien » est aligné avec « monster ».

et,

$tf(\text{chien}) = \{(\text{dog.n}, 2) , (\text{monster.n}, 1)\}$
--

Ici, encore les liens du mot «dog.n», ayant la plus grande fréquence d'alignement, vont être ceux qui constituent l'entrée du mot « chien » dans le dictionnaire.

Résultats et Conclusion :

Avant de discuter les résultats obtenus par cette méthode, nous montrons en tableau 3 les statistiques principales issues de notre corpus de projection, des deux cotés Anglais et Français, et du corpus du test.

	Corpus anglais	Corpus français	Corpus de test (référence)
Nombre de phrases	32 571	32 571	134
Nombre de mots différents	6 568	2 073	733
Nombre de mots différents projetés	4 663	2 030	

Tableau 3 : des statistiques concernant les corpus utilisés.

Du tableau 3, nous déduisons quelques points qui nous aideront quantifier les erreurs. Du côté anglais, il y a 71 % des mots du vocabulaire qui sont projetés sur les différents mots du vocabulaire français, donc 29% n'ont pas été alignés et sont donc non projetés. En revanche, le pourcentage des mots français cible d'une projection est de 90%.

Rappelons que l'alignement entre deux mots signifie qu'un pont est établi entre ces deux mots et par la suite des relations syntaxiques vont être projetées. Donc à chaque fois qu'un mot anglais est aligné, ses relations vont être projetées sur le mot français qui lui est associé.

La fréquence d'alignement indique alors la quantité de liens projetés, intéressante à connaître parmi les mots (71%) qui sont alignés.

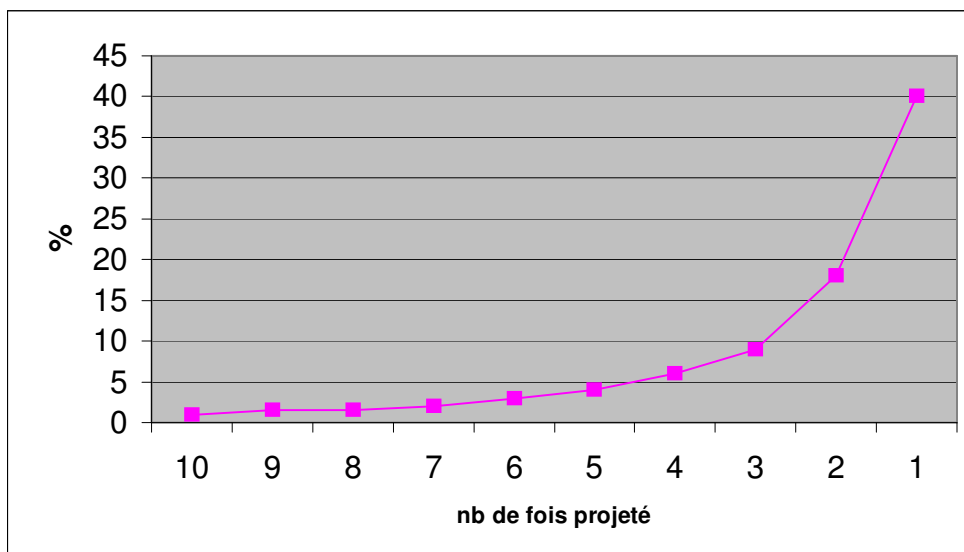


figure 20: pourcentage des mots source (anglais) en fonction de la fréquence d'alignement.

Nous montrons à la figure 20 la fréquence de projection des 71% des mots anglais. Sans grande surprise, on observe que 40% de ces mots anglais ne se projettent qu'une seule fois. Notons que les alignements de faible fréquence sont probablement plus affectés au bruit de l'alignement que les autres. Un mot est aligné une seule fois, va diminuer la performance de la projection, s'il est aligné incorrectement.

	Précision	Rappel	F-mesure	Crosing-bracket	Nb de phrase
Nombre de mots ≤ 10	71.9 %	78.5 %	75 %	0	7
10 < Nombre de mots ≤ 20	-	-	-	-	0
20 < Nombre de mot ≤ 40	-	-	-	-	0

Tableau 4 : résultats de la méthode de projection des liens du mot le plus fréquemment aligné (PL-pfa) en fonction de la longueur des phrases.

Nous montrons maintenant dans le tableau 4 les résultats de cette méthode. La précision et le rappel qu'on peut les constater sont acceptables au niveau de l'analyse. Mais le nombre insuffisant de phrases nous laisse penser à une ou plusieurs limites qui ont empêché une analyse plus étendue des phrases du corpus de test.

Le nombre de phrases analysées est affecté, le plus souvent, par le nombre de fois les mots peuvent se connecter. La seule cause qui limite un mot français de se connecter dans une phrase donnée est des liens inappropriés qui lui sont projetés ou le manque de lien lors de la projection.

Rappelons que chaque mot anglais dans le dictionnaire de la Link-Grammar est représenté par une longue expression de liens. Plusieurs combinaisons de liens possibles vont dériver de cette expression (voir la définition du disjoint dans le chapitre 2). Dans cette méthode, on ne prend qu'une seule combinaison pour la projeter, celle utilisée par les mots les plus fréquemment alignés. L'idée alors de cette approche a une tendance à omettre des liens utiles pouvant aider un mot à se connecter. Revenons à la figure 20, il y a 40% des mots anglais se projettent une seule fois, ayant alors une seule combinaison de liens. Ça revient à dire que ces mots anglais n'ont pas suffisamment de liens qui peuvent couvrir tous les cas possibles pour un mot français donné.

Exemple :

```
intacts => intact.a (freq=1)
intacts : (Pa-);
```

Par exemple, le mot anglais « `intact.a` » est parmi les 40% des mots qui sont projetés une seule fois. Ce mot à un seul lien qui est sa seule combinaison, et donc une seule façon de se connecter dans une phrase. Or, « `intact.a` » avec son expression extraite du dictionnaire de la Link-Grammar : `{EA- or EF+} & (({[[@Ec-]]) & {Xc+} & A+) or ((Pa- or AF+ or Ma-) & {@MV+}) or AA+ or [[DD- & <noun-main-p>]] or <adj-op>`, a plusieurs combinaisons de liens possibles.

En conclusion, cette méthode ne permet pas de projeter suffisamment de liens vers les mots français, limitant ainsi les analyses des phrases françaises du corpus de test. Pour cette raison, on a pensé à une méthode qui va prendre plus de liens à projeter.

5.3. Les mots du corpus (PL-a)

Description :

Rappelons que la projection des liens des mots anglais vers les mots français se fait en passant par toutes les paires de phrases du corpus. A chaque fois qu'un mot français est rencontré, il est aligné à un et un seul mot anglais qui peut être le même ou différent, des mots associés à lui. Ces mots anglais vont former une liste à la fin de l'alignement du corpus bitexte. Dans cette approche, on utilise tous les mots de cette liste afin de projeter leurs liens sur le mot français qui est aligné à eux.

La différence de la première méthode (*PL-pfa*) est que pour chaque mot français, il existe une ou plusieurs combinaisons de liens appartenant aux différents mots anglais qui lui sont associés. Un mot français donné est donc présenté par l'union des liens de chaque mot anglais qui lui sont alignés dans le corpus.

Formellement :

Puisque cette méthode est exécutée sans modification de l'alignement de mots et de l'analyseur syntaxique offert par le système « *Link-Parser* », alors tous les ensembles décrits dans la première approche demeurent les mêmes.

Rappelons que l'ensemble $tf(f)$ est constitué d'une liste de mots anglais qui lui sont alignés, avec leurs fréquences. Mais, nous ne tenons pas compte dans cette variante de la fréquence d'alignement. Notre implémentation de la fonction « **association-liens** » consiste alors, de projeter les liens de tous les mots anglais « e » de $tf(f)$ alignés à un mot français f_v donné.

Donc pour chaque « f_v » dans le vocabulaire français, il existe un ensemble de liens $L(f_v, e)$ (voir section 5.2) réunissant tous les différents liens que « e » contient.

Le changement alors sera dans la définition du dictionnaire.

Donc $D(f_v) = \{(f_v, \cup L(f_v, e)) / v \in [1, V] \text{ et } A(e, f_v)\}$ sera l'ensemble des mots français avec l'union des liens de tous les « e » alignés à f_v .

Et ainsi de suite pour chaque mot français dans le corpus, on lui ajoute les liens utilisés par le mot source qui lui est aligné.

Résultats et conclusion :

Nous comparons cette méthode avec la première et nous présentons les résultats, dans le tableau 5, en fonction de la longueur des phrases (mesuré en nombre de mots).

	Précision	Rappel	F-mesure	Crosing-bracket	Nb de phrase
PL-pfa	71.9 %	78.5 %	75 %	0	7
<= 10	48 %	65 %	55 %	0	19
<= 20	31.6 %	45 %	37 %	2.7	37
<= 40	31.2 %	45 %	37 %	5	2

Tableau 5 : les résultats de la projection de tous les mots alignés à un mot français (PL-a).

Nous observons une baisse dans la précision et rappel par rapport au « PL-pfa » tel que mesurés sur les phrases analysées. Mais cette variante permet d'analyser beaucoup plus de phrases (58) que la première, ce qui est meilleur dans la performance en générale (nous considérons le rapport F-mesure et le nombre de phrases ensembles). Rappelons que Hwa [Hwa et al., 2001] ont reçu, après une projection directe similaire à cette méthode, un F-mesure proche de 38% contre 43%, que notre première tentative nous a offert. Une comparaison des résultats finals sera détaillée dans la section 5.9 de ce même chapitre.

Comme déjà expliqué, chaque mot français subit plusieurs projections. Cette approche donc, consiste à les prendre tous en considération. Nous serons intéressés alors de savoir la quantité de liens projetés vers un mot français. Nous montrons pour cela en figure 22 le pourcentage de nombre de fois des mots français ayant une projection.

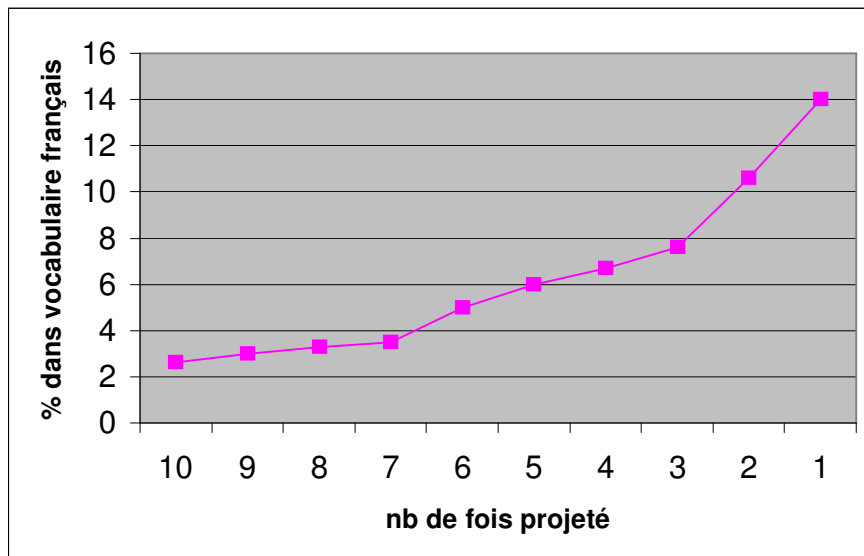


figure 22 : pourcentage des mots cibles (français) ayant une projection.

On peut constater directement que 14% seulement, des mots français ont subi une seule projection, et puis la courbe descend pour atteindre un niveau à peu près constant dès qu'on dépasse les 6 alignements. La plupart des mots ont plusieurs projections, et donc, plus de couverture des liens pour se connecter, mais plus de tendance à utiliser des liens inutiles transférés avec le bruit d'alignement.

L'avantage de cette approche est de pouvoir se connecter de plusieurs façons sur les mots cibles. Mais d'autre part, on perd la précision en utilisant des liens non convenables au mot français. C'est-à-dire, quand un mot cible n'est pas aligné correctement à un mot source. Nous pouvons la constater dans l'exemple de la figure 23 :

Exemple :

```
tradition: (tradition.n, liens)
           (traditional.a, liens)
           (is.v, liens)
```

figure 23 : sortie de l'alignement pour le mot français « tradition »

Le mot « tradition », est aligné avec trois mots anglais différents. La projection consiste à projeter les liens des mots alignés vers le mot « tradition ». Ainsi ce mot va pouvoir utiliser les liens projetés du mot « is.v », ce qui n'est pas correct grammaticalement, et conduit à une baisse dans la précision. C'est l'inconvénient de cette méthode.

Pour mettre une fin à cette expérience, on en déduit que des liens inutiles sont transférés lors de la projection de plusieurs mots sources. Donc une limitation du nombre de mots projetés sera importante pour augmenter la précision, et elle sera le sujet de notre prochaine expérience.

5.4. Les plus fréquents ET les plus probables

Description :

En introduisant les deux premières méthodes ; l'une prend le mot anglais le plus fréquemment aligné pour projeter ses liens, tandis que l'autre prend en considération tous les mots possibles en projetant leurs liens sur le mot français en question. Si on considère ces méthodes comme étant deux fonctions, leur variable serait alors « n » le nombre de mots anglais utilisés pour la projection vers un mot français : il est égal à « 1 » dans la première méthode et à un « ∞ » dans la deuxième.

Une nouvelle variante est intégrée dans cette approche : les mots les plus probablement alignés. Rappelons que l'alignement de *Viterbi* s'appuie sur une masse calculée, durant son exécution, pour la sélection du meilleur alignement. Cette masse peut être considérée encore comme un critère pour trouver les mots les plus probablement alignés.

Enfin chaque mot français va avoir une liste de mots anglais alignés avec lui, dont chacun à sa fréquence et sa probabilité d'alignement. Dans cette approche, on va essayer de faire varier « n » sur un ensemble fini, en fonction de la fréquence d'alignement du mot d'une part et la masse probabilisée d'une autre part. Donc notre courante expérience sera une comparaison entre les « n » les plus fréquemment alignés et les « n » les plus probablement alignés.

Formellement :

L'ajout de la masse probabilisée dans cette approche implique un changement dans l'ensemble $tf(f)$ qui peut se présenter de la façon suivante :

$tf(f) = \{(e,c,pb) / \exists k \in [1,l], \exists a_k \in [1,m] / \exists A(e_{ak}^t, f) \text{ et } e_{ak}^t = e\}$, L'ensemble des mots anglais « e » qui sont alignés à un mot français donné « f », avec « c » la fréquence d'alignement de « e » et « pb » sa probabilité d'alignement.

Le changement va affecter la fonction « **association-liens** » qui va prendre comme arguments, le nombre « n » et la variante qui indique sur quel critère on s'appuie pour choisir les mots anglais. Cette variante peut prendre la valeur « *fréquence* » ou « *masse* ». Donc deux étapes sont exécutées afin de projeter les liens sur un mot français. Premièrement nous indiquons sur quelle variante notre projection s'appuie, puis nous choisissons le nombre de mots anglais qu'on doit projeter pour chaque mot français.

- Si variante = « *fréquence* »
 - Pour $n=1$,
 - Appliquons la première méthode qui maximise sur la fréquence « c » de l'ensemble $tf(f)$.
 - Pour $n=2$,
 - Prenons les deux premiers « e » les plus fréquemment alignés de $tf(f)$.
 - Projetons les liens de ces deux mots sur le mot français associé à eux.

Et ainsi de suite pour toutes les valeurs de « n » qu'on choisit.

Si variante = « *masse* », la même procédure est suivie, mais la maximisation serait sur la masse probabilisées.

$$tf(f) = \underset{pb}{\operatorname{argmax}} (e, c, pb)$$

Dans les deux cas, la définition du dictionnaire est la même et dépend de la valeur de « n ». On aura donc deux nouveaux dictionnaires pour chaque valeur de « n », appartenant aux différentes variantes choisies.

Soit $D(f_v) = \{ (e, L(f_v, e)) / v \in [1, V] \}$ l'ensemble des mots « f_v » du vocabulaire français suivi de leurs liens, formés par l'union des liens des « n » mots anglais « e » alignés à f_v .

Illustrons cette idée sur les exemples 1, 2 et 3, et indiquons seulement le changement qui a eu lieu lors de l'application de cette méthode.

$$\begin{aligned} \text{tf}(f_v) &= \{ (e, c, \text{pb}) \} \\ \text{tf}(\text{chien}) &= \{ (\text{dog.n}, 2, -3), (\text{monster.n}, 1, -2) \} \end{aligned}$$

Figure 24 : l'ensemble $\text{tf}(f)$ du mot « chien ».

Par exemple, pour le mot « chien », l'ensemble $\text{tf}(f)$ sera constitué des mots anglais qui lui sont alignés, et dont chacun est suivi de deux nombres, qui sont respectivement, la fréquence d'alignement et la masse probabilisée (on calcule le Log de la probabilité, ce qui rend le résultat négatif), comme illustré dans la figure 24. Donc deux cas se présentent suivant la variante nous choisie comme critère de sélection :

1er cas, si variante = fréquence

$$\begin{aligned} D(f_v) &= \{ (e, L(f_v, e)) \} \\ \text{Pour } n=1, \\ D(\text{chien}) &= \{ (\text{dog.n}, (\text{Os-} \& \text{Ds-}) \text{ or } (\text{Ds-} \& \text{Ss+})) \} \\ \text{Pour } n=2, \\ D(\text{chien}) &= \{ (\text{dog.n}, (\text{Os-} \& \text{Ds-}) \text{ or } (\text{Ds-} \& \text{Ss+})) \cup \\ &\quad (\text{monster.n}, (\text{Os-} \& \text{Ds-})) \} \end{aligned}$$

Donc deux dictionnaires seront formés pour cette variante. Nous les illustrons avec l'entrée du mot « chien » :

Pour n=1, chien : ((Os- & Ds-) or (Ds- & Ss+));

pour n=2, chien : ((Os- & Ds-) or (Ds- & Os-)) or
(Os- & Ds-);

2ième cas, si variante = masse

$D(f_v) = \{ (e, L(f_v, e)) \}$
 Pour n=1,
 $D(\text{chien}) = \{ (\text{monster.n}, (Os- \& Ds-)) \}$
 Pour n=2,
 $D(\text{chien}) = \{ (\text{monster.n}, (Os- \& Ds-)) \cup$
 $(\text{dog.n}, (Os- \& Ds-) \text{ or } (Ds- \& Ss+)) \}$

L'entrée « chien » dans les deux dictionnaires sera alors dans ce cas :

pour n=1, chien : (Os- & Ds-);

pour n=2, chien : (Os- & Ds-) or
((Os- & Ds-) or (Ds- & Os-));

Et ainsi de suite, nous créons pour chaque variante plusieurs dictionnaires en fonction de « n ». Dans notre approche, « n » se situe entre 1 et 15. Donc pour la variante « fréquence », on aura 15 dictionnaires à tester, et de même pour la variante « masse ».

Résultats et conclusion :

Les résultats des deux variantes pour chaque valeur de « n », sont comparés en fonction du nombre de phrases analysées, et de la précision et rappel qui sont présentées par le F-mesure. La représentation graphique des figures 25 et 27 présente cette comparaison.

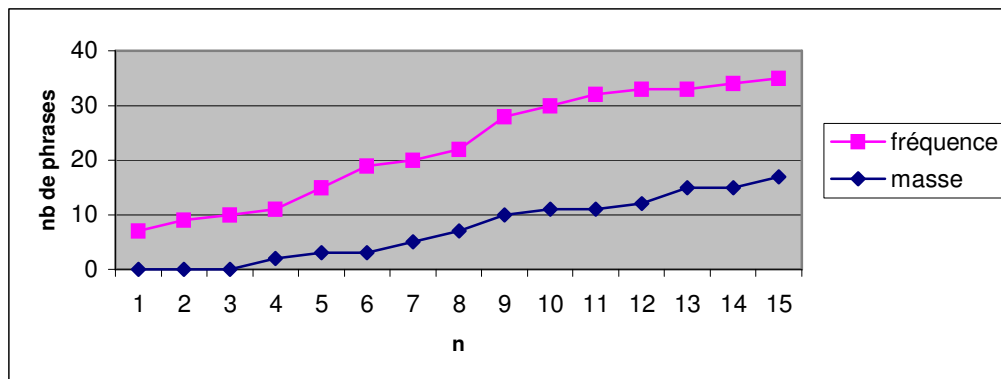


figure 25 : comparaison du nombre de phrases analysées entre les n les plus fréquemment alignés (PL-npfa), et les n les plus probablement alignés (PL-nppa).

Pour les deux méthodes le nombre de phrases augmente en fonction de « n ». Ce qui est normal, puisque « n » est le nombre de mots projetés. Nous constatons que plus de phrases sont analysées par la méthode *PL-npfa*. La méthode *PL-nppa* quant à elle ne retourne aucune analyse pour des valeurs de n inférieures à 4.

Notons que plus il y a de phrases analysées, plus les mots de cette phrase ont plus d'options pour se connecter. La seule explication qu'un mot contient plusieurs combinaisons de liens, est sa fréquence d'alignement. Comparons alors ces deux variantes en montrant, en figure 26, le pourcentage du nombre de mots dans le corpus (les plus fréquemment et les plus probablement alignés) en fonction de la fréquence d'alignement :

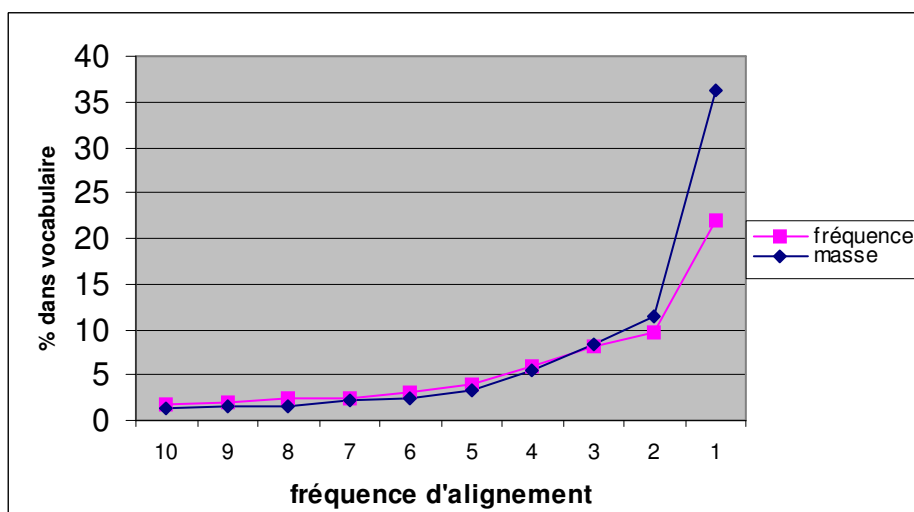


figure 26: pourcentage de nombre de mot en fonction de la fréquence d'alignement entre PL-npfa et PL-nppa.

On constate que les mots les plus probablement alignés sont plus nombreux que les mots les plus fréquemment alignés, ayant la fréquence d'alignement égale à 1. Donc, dans la première méthode (*PL-nppa*), il y a plus de mots qui utilisent une seule combinaison de liens, c'est pour cette raison cette méthode à moins analysées de phrases.

Nous mesurons la qualité de ces deux variantes à l'aide de la F-mesure présenté en figure 27:

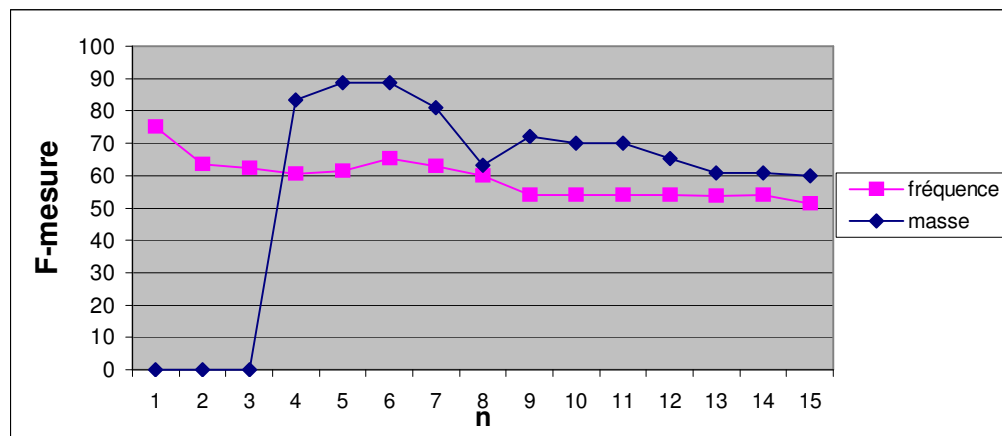


figure 27 : comparaison des F-mesure entre PL-npfa et PL-nppa.

Pour des valeurs de « n » petites (>3), la F-mesure des n -mots les plus probablement alignés est plus grande que celle des n -mots les plus fréquemment alignés. Mais en augmentant « n », on constate que les deux méthodes sont proches (des valeurs entre 50 et 60%). Même si la courbe de la « *masse* » dépasse un peu celle de la « *fréquence* », il est important de savoir la raison pour laquelle les mots les plus probablement alignés ont un meilleur F-mesure. Les valeurs nulles de la f-mesure pour les valeurs de n inférieures à 4 s'expliquent par le fait que notre grammaire ne produit aucune analyse complète, tel qu'expliqué précédemment.

Montrons pour cela sur un graphe, dans la figure 28, le pourcentage de nombre de mots les plus fréquemment et probablement alignés, en fonction de la masse probabilisée maintenant.

Rappelons que cette masse indique le critère de sélection qu'utilise l'aligneur pour aligner les mots anglais aux mots français. Puisque cette masse est négative, alors plus elle est proche de 0, plus le mot sera le plus convenable à aligner. Elle est découpée en 4 parties, indiquant 4 niveaux de précision d'alignement. La valeur (-5) indique les mots

ayant la masse située entre (0) et (-5), qui est une valeur la plus acceptable d'après l'alignement de Viterbi (voir chapitre 3). La valeur (-10) dont la masse entre (-5) et (-10) est moins acceptable, et de même pour les mots situant dans la valeur (-20) sont les moins acceptables.

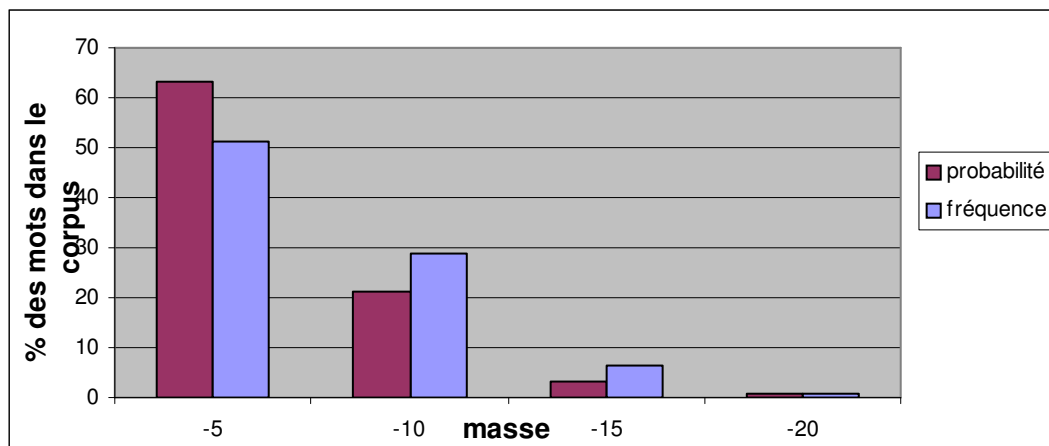


figure 28 : comparaison de pourcentage de nombre de fois des mots du (PL-npfa) et du (PL-nppa) en fonction de leurs masses.

En diminuant la masse (de -5 vers -20), on constate que le nombre de mots les plus fréquemment alignés dépasse celui des mots les plus probablement alignés. Ceci indique qu'on a plus de mots dans la méthode PL-npfa, avec leurs probabilités d'alignement plus grandes que (-5). Alors il existe plus de mots dans cette méthode qui sont moins précis au niveau d'alignement. Ce qui explique la différence de la F-mesure entre ces deux méthodes.

En prenant en considération les nombres de phrases pour chaque variante, on déduit qu'une meilleure performance en général de la méthode PL-npfa sur la PL-nppa. En fin de compte le défi se situe sur la maximisation du nombre de phrases analysées et du F-mesure ensemble. Un compromis classique à de nombreuses applications.

5.5. Projection de l'expression

Description :

Les mots anglais du dictionnaire de la « *Link-Grammar* » sont exprimés à l'aide d'une expression contenant tous les liens possibles qu'un mot donné puisse prendre. À chaque fois ce mot est utilisé dans une phrase pour l'analyser, la combinaison de liens qui

le laisse se connecter avec les autres mots de la phrase, est extraite de son expression dans le dictionnaire. Voici par exemple l'entrée dans le dictionnaire pour le mot « dog.n » :

Exemple:

```
dog.n : ({@AN-} & {@A- & {[[@AN-]]}} & ((Ds- & <noun-sub-s> &
<noun-main-s> or Bsm+)) or (YS+ & Ds-) or (GN+ & (DD- or [( ])) or
Us-)) or AN+;
```

Dans cette approche, on projette toute l'expression du mot anglais vers le mot français qui lui est associé par l'alignement. Le fait que l'expression soit décrite seulement dans le dictionnaire, on doit alors, à chaque fois qu'on rencontre un mot anglais, le chercher dans le dictionnaire pour extraire son expression afin de la projeter sur le mot français.

Formellement :

Si on veut comparer cette approche au protocole général, qui dans sa description indique que la projection se fait au niveau des liens partant de chaque mot, on peut étendre alors ces liens sur toutes les combinaisons possibles d'un mot, et alors projeter toute son expression.

Pour cela, il nous faudra redéfinir tous les ensembles qui ont rapport avec les liens. Par contre, il n'y aura aucun changement pour l'ensemble d'alignement et tout ce qui est en rapport avec la « *Link-Grammar* ».

Premièrement, commençons à définir l'expression d'un mot par $Exp(e)$, qui se formalise par l'expression des liens du mot « e » dans le dictionnaire de la *Link-Grammar*.

Soit $D(f_v) = \{(e, Exp(e_d)), \exists v \in [1, |V|] / \exists A_t(f_v, e)\}$, l'ensemble du dictionnaire français projeté.

Nous proposons en figure 29 une redéfinition formelle de l'opération de la projection :

```

Projection ( $S^t$ )
{
  pour chaque  $S^t = \langle E^t, F^t \rangle$ 

    analyser ( $E^t$ );
    => extraire expression de chaque e dans  $E^t$ 

    aligner ( $E^t, F^t$ );
    => formation de  $a = \{(a_j)\}$ 

  pour chaque  $f_j^t \in F^t$ ,  $\forall j \in [1, l]$ 

    si ( $a_j \neq 0$ )
      association-expression ( $f_j^t$ ,  $e_{a_j}^t$ ,  $\text{Exp}(e_{a_j}^t)$ );
      => formation de  $D(f_j^t)$ 
}

```

Figure 29 : redéfinir la définition de la projection (des expressions).

Nous illustrons ce formalisme sur l'exemple suivant :

$\langle E^t : \textit{The dog chases the cat}$, $F^t : \textit{Le chien poursuit le chat} \rangle$

Analyse syntaxique de E^t

```

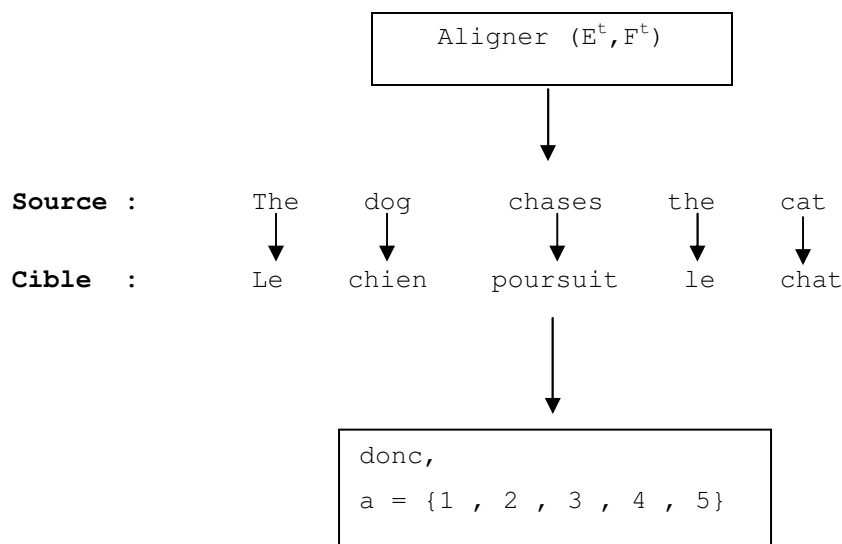
      +-----Os-----+
+-Ds-+-----Ss---+   +-Ds-+
|   |   |   |   |   |   |   |
The dog.n chases.v the cat.n

```

Pour $t=1$ (considérons toujours la première phrase), $i=2$ (formons l'exemple sur un seul mot, « dog.n »)

$\text{Exp}(e_i^t) = \text{Exp}(e_2^1) = \text{Exp}(\text{dog.n}) =$

$(\{ @AN- \} \& \{ @A- \& \{ [[@AN-]] \} \} \& ((Ds- \& \langle \text{noun-sub-s} \rangle \& (\langle \text{noun-main-s} \rangle \text{ or Bsm+})) \text{ or } (YS+ \& Ds-) \text{ or } (GN+ \& (DD- \text{ or } [()])) \text{ or } Us-)) \text{ or } AN+$



Après qu'on a extrait l'expression de chaque mot anglais du dictionnaire, et on a formé l'ensemble « a » des alignements ; il devient facile de trouver pour chaque mot source (anglais), son associé, le mot cible (français) et de projeter l'expression des liens directement.

Alors, l'ensemble du dictionnaire français du mot « chien » après la projection sera formé de la façon suivante :

$$\begin{aligned}
 D(\text{chien}) &= \{(\text{dog.n}, \text{Exp}(\text{dog.n}) / \exists A_t(\text{chien}, \text{dog.n})\} \\
 &= \{(\text{dog.n}, \{ \{ @AN- \} \ \& \ \{ @A- \ \& \ \{ [[@AN-]] \} \} \ \& \ ((Ds- \ \& \\
 &\langle \text{noun-sub-s} \rangle \ \& \ (\langle \text{noun-main-s} \rangle \ \text{or} \ \text{Bsm+})) \ \text{or} \ (\text{YS+} \ \& \ \text{Ds-}) \ \text{or} \\
 &(\text{GN+} \ \& \ (\text{DD-} \ \text{or} \ [()])) \ \text{or} \ \text{Us-}) \ \text{or} \ \text{AN+}) \}
 \end{aligned}$$

Et enfin, les entrées du dictionnaire français seront formées avec les mêmes expressions de celles d'anglais associées aux mots français. Par exemple, la figure 30 montre typiquement l'entrée dans le dictionnaire pour le mot « chien ».

chien : ({ @AN- } & { @A- & { [[@AN-]] } } & ((Ds- & <noun-sub-s> & <noun-main-s> or Bsm+)) or (YS+ & Ds-) or (GN+ & (DD- or [()])) or Us-) or AN+;

Figure 30: l'expression du mot "chien" dans le dictionnaire français.

Résultats et conclusion

Avant de présenter les résultats sur les graphes des figures 31 et 32, rappelons que pour chaque occurrence du mot français dans le corpus, il existe un ou plusieurs alignements dans la partie anglaise. Chaque mot anglais a une fréquence d'alignement, qui représente le nombre de fois que ce mot anglais est aligné au mot français, ce qui nous aide à choisir lequel on doit projeter son expression.

Nous allons dans cette expérience projeter les « n » entrées des mots anglais les plus fréquemment alignés à un mot français donné.

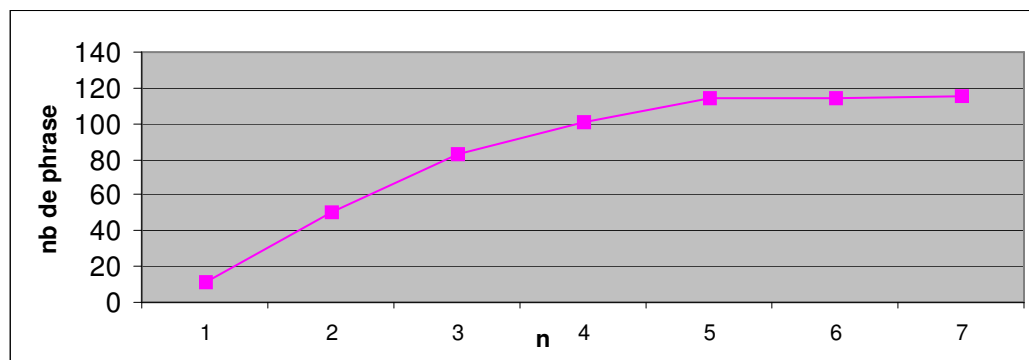


figure 31 : résultat des nombres de phrases analysées par la méthode PE-npfa en fonction de « n » (le nombre de mots anglais ayant servi à la projection)

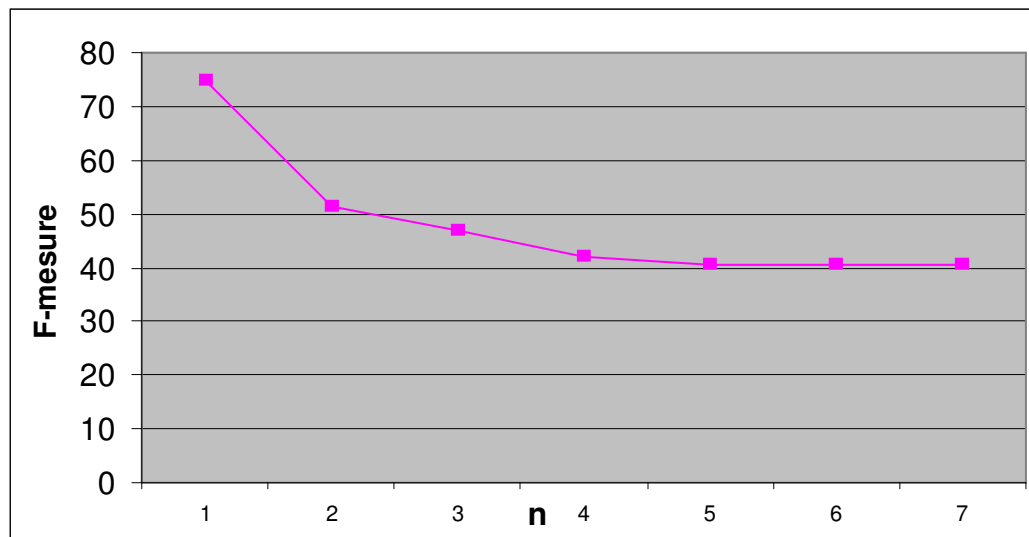


figure 32 : résultat du F-mesure par la méthode PE-npfa en fonction de « n ».

La F-mesure, comme constatée dans la figure 32, diminue à chaque fois « n » augmente, et c'est à cause de nombre des expressions projetées, qui vont augmenter aussi. Mais, pour la même raison, on peut voir une augmentation du nombre des phrases

analysées en fonction de « n », parce qu'on peut avoir plus d'option de se connecter pour un mot français donné.

Revenons à l'approche qui projette juste les liens utilisés par les mots anglais, et comparons ces deux méthodes par deux graphes, illustré en figure 33 et 34. Nous montrons ainsi le nombre de phrases analysées et la F-mesure de chacune :

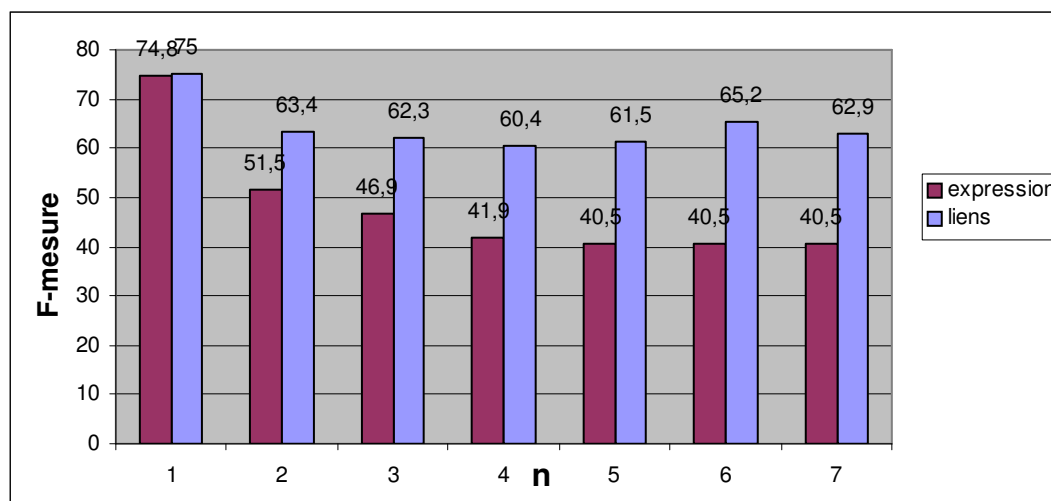


figure 33 : comparaison de la précision des deux méthodes PL-npfa et PE-npfa.

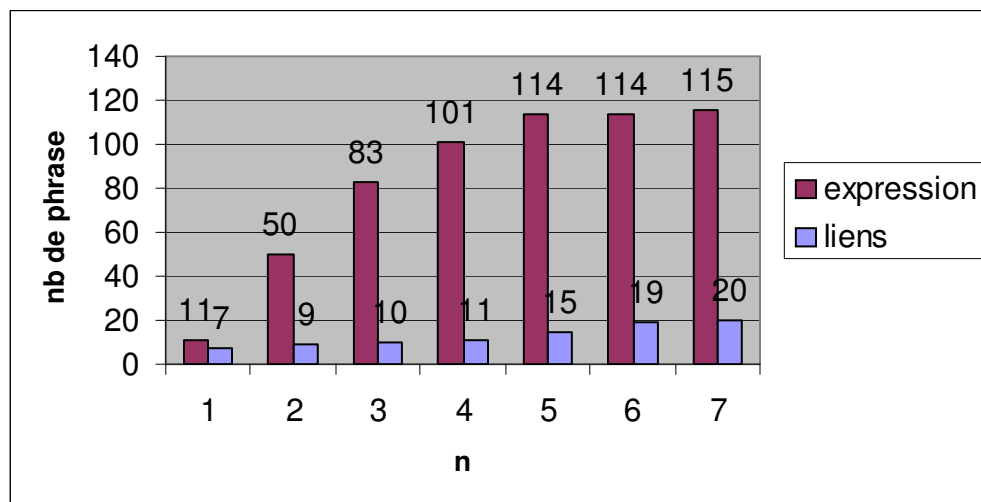


figure 34 : comparaison du nombre de phrases analysées des deux méthodes PL-npfa et PE-npfa.

Avant de faire une conclusion, rappelons un peu comment une expression d'un mot donné peut être interprétée par l'analyseur de la « *Link-Grammar* ». Par exemple, le mot : `car : {@A-} & D- & {B+} & (O- or S+);` peut avoir 8 combinaisons différentes de

liens. Alors il peut se connecter de plusieurs façons différentes dans une phrase donnée (chaque combinaison (disjoint) permet au mot de se connecter de plusieurs façons).

Combinaison de liens : ((@A or D) & (S or B))

((D) & (S or B))

((@A or D or O) & (B))

((D or O) & (B))

((@A or D) & (S))

((D) & (S))

((@A or D or O) & ())

((D or O) & ())

Donc, tout mot aligné avec « car », va avoir ces mêmes combinaisons dans une phrase française. Dans la projection de toute l'expression, il y aura parfois des liens qui connectent un mot anglais dans une phrase et qui sont valides en anglais, mais ils ne le sont pas en français. Ces liens, dans la méthode *PE-npfa*, vont être transférés avec la projection de l'expression d'un mot. C'est pourquoi la précision de la méthode *PL-npfa* est plus élevée que la méthode *PE-npfa*.

Par contre, dans la méthode *PL-npfa*, les liens utilisés par un mot anglais sont les seuls qui subissent la projection à travers l'alignement. Donc les liens les plus convenables sont transférés vers le mot français. C'est pour cette raison, on voit que la méthode *PL-npfa* est plus précise, et qui converge vers une valeur de 62%, au contraire de la méthode *PE-npfa* qui se stabilise vers 40%.

D'autre part, le nombre de phrases analysées est beaucoup plus grand dans la projection de l'expression. Ceci s'explique encore par la multitude de combinaisons de liens existant dans une seule expression, qui laisse pour un mot français des options plus vaste de se connecter.

Une limite est constatée durant cette approche. L'imprécision de l'alignement qui laisse la projection dans un niveau insuffisant par rapport à la F-mesure et le nombre de phrases analysées. Dans une nouvelle approche, nous voulons essayer d'améliorer la projection en diminuant le bruit de l'alignement.

5.6. Alignement bidirectionnel

Description :

Dans les expériences présentées jusqu'à maintenant, nous avons fait usage d'un modèle d'alignement IBM 2 donnant la probabilité d'un mot français étant donnée la probabilité d'un mot anglais.

Nous avons expliqué au chapitre 3 que ce modèle est par nature non symétrique : chaque mot cible (français jusqu'à maintenant) reçoit un et un seul mot source (qui peut être le mot NULL). Ceci n'implique pas l'inverse : tous les mots sources ne sont pas nécessairement alignés à un mot cible particulier.

Nous avons donc étudié la pertinence d'aligner de manière bidirectionnelle notre corpus de projection : une fois à l'aide d'un modèle $P(f/e)$ et une fois avec un modèle $P(e/f)$, puis de prendre l'union des deux alignements ainsi produits. Cette approche est souvent appliquée pour renforcer la qualité d'un alignement [Och et al.,2004], et contourner la limitation des modèles IBM.

Formellement :

Cette approche va être l'application de la méthode comparant les « n » les plus fréquemment alignés avec les « n » les plus probablement alignés. Donc, la projection va utiliser les mêmes ensembles, mais en fonction d'un nouveau fichier qui unit celles des deux sens d'alignement. Le changement est alors au niveau de la construction de ce nouveau fichier.

Nous présentons alors les deux ensembles représentant les deux fichiers d'alignement (de chaque direction);

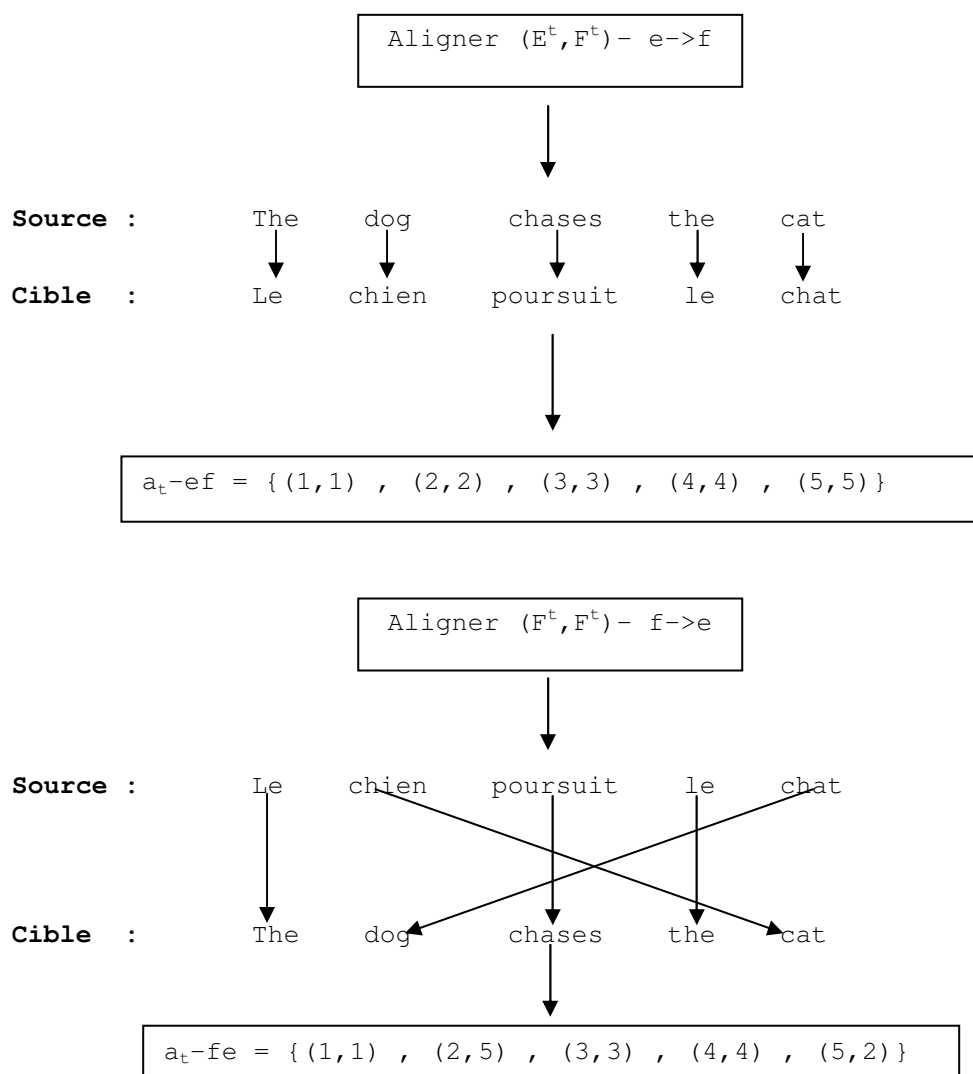
Soit $a_{r-ef} = \{(e_{ak}^t, f_k) / \exists t \in [1, |B|], \exists a_k \in [1, m], \exists k \in [1, l] / \exists A(e_{ak}^t, f_k)\}$, l'ensemble de paires de mots anglais et français alignés de e vers f d'une paire de phrase donnée t , qui constitue le fichier résultant de l'application de la méthode $P(f/e)$.

Et soit $a_{rfe} = \{(f_{ak}, e_{ak}^t) / \exists t \in [1, |B|], \exists a_k \in [1, m], \exists k \in [1, l] / \exists A(f_{ak}, e_{ak}^t)\}$, l'ensemble de paires de mots français et anglais, alignés de f vers e d'une paire de phrase donnée t , qui constitue le fichier résultant de l'application de la méthode $P(e/f)$.

Finalement le nouveau fichier présenté par $F = \{(a_{r-ef} \cup a_{r-fe})\}$, et construit à partir de l'union des deux fichiers d'alignement de $e \rightarrow f$ et de $f \rightarrow e$.

Prenons l'exemple 1 pour bien illustrer l'ensemble F du nouveau fichier :

$\langle E^t : \text{The dog chases the cat} \quad , \quad F^t : \text{Le chien poursuit le chat} \rangle$



Le fichier résultant sera formé de la façon suivante :

```

F = { (at-ef ∪ at-fe) }
F = { (1,1), (2,2), (2,5), (3,3), (4,4), (5,5), (5,2) }
et, l'ensemble d'alignement devient:
a = {1, 2, 2, 3, 4, 5, 5}

```

On constate qu'il existe pour une seule occurrence d'un mot français, deux mots anglais qui lui sont associés. Par exemple, les paires d'indices (2,2) et (5,2) qui appartiennent aux paires des mots (dog , chien) et (cat , chien), sont une indication de deux possibilités d'alignement du mot « chien ».

Contrairement à l'idée de IBM, qui permet seulement un seul associé anglais pour chaque occurrence d'un mot français de la même phrase. Cette méthode dépasse cette limite en considérant plus qu'un seul mot anglais pour un mot français donné.

Résultats et conclusion :

Revenons maintenant à la représentation graphique, et montrons en figure 35, une comparaison des deux variantes de l'approche **PL-npfa** et **PL-nppa** dans l'alignement bidirectionnel.

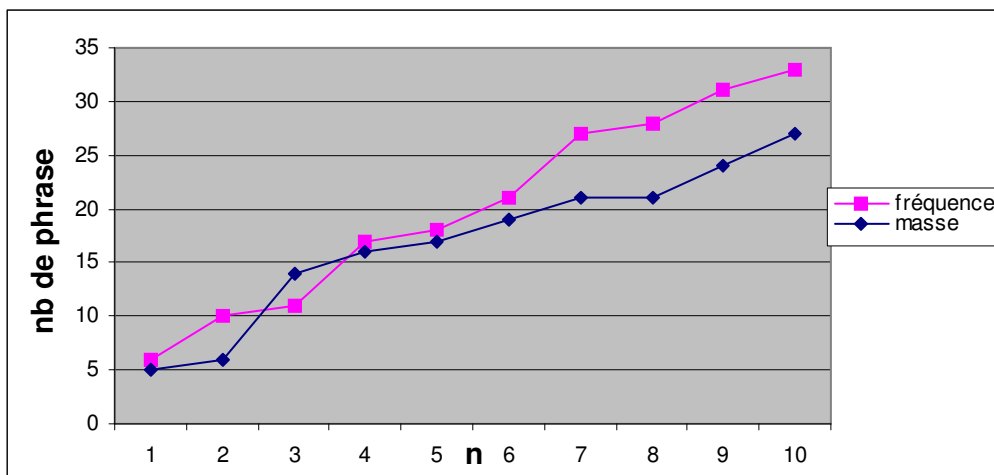


figure 35 : le nombre de phrases après analyse des PL-npfa et PL-nppa avec l'alignement bilingue.

Nous observons dans la méthode *PL-npfa* un nombre de phrases analysées plus élevé que celui de la méthode *PL-nppa*. Mais cette fois, les « n » les plus probablement alignés ont été améliorés de 50% de celle appliquée avec le modèle *P(f/e)*. Cette augmentation est due parce que les mots alignés une seule fois sont moins de 10% environ que les mots dans la méthode de l'alignement unidirectionnel. Nous montrons en figure 36 qu'il y a plus de mots ont été alignés plusieurs fois (>1). Par la suite, on aura plus de liens à projeter, ce qui explique l'augmentation de nombre de phrases analysées dans la méthode *PL-nppa* par rapport à l'alignement unidirectionnel.

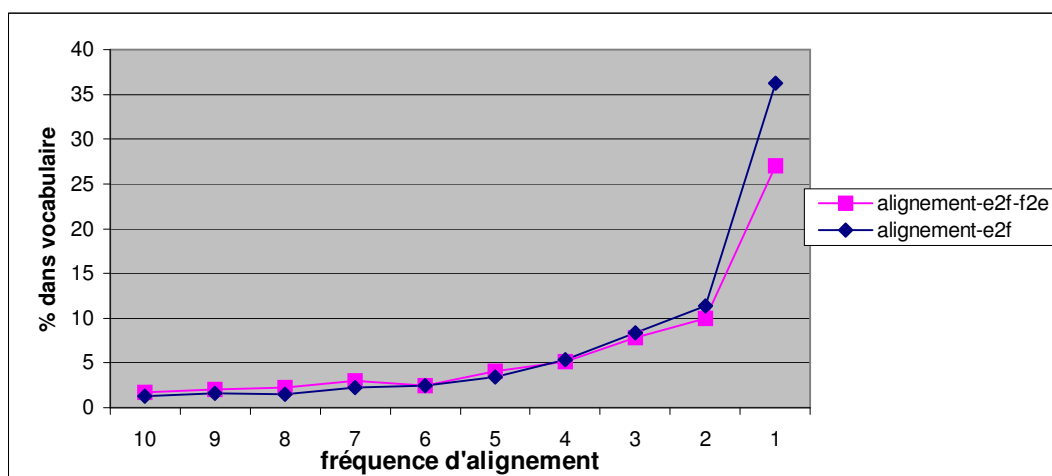


figure 36: pourcentage de nombre de mots en fonction de la fréquence d'alignement entre l'alignement unidirectionnel et bidirectionnel.

D'autre part la F-mesure converge, avec les deux variantes, vers 40 et 50% après la 7^{ème} valeur de n , ce que l'on observe sur le graphe de la figure 37.

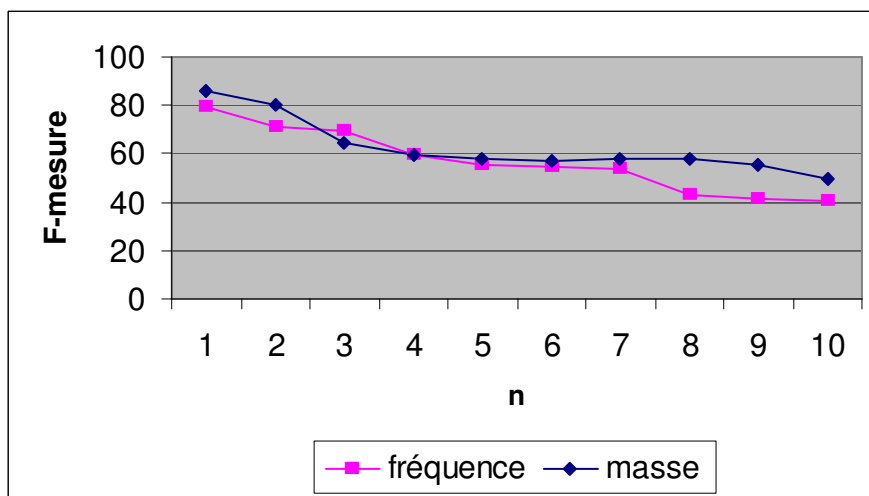


figure 37 : les résultats de la précision du PL-npfa et PL-nppa avec le nouveau fichier d'alignement.

Nous constatons aussi, une convergence des deux méthodes vers 60% (dans la F-mesure). Un taux acceptable au niveau de la performance d'une projection directe. La cause de la diminution de cette précision au 8^{ième} mot projeté, revient aux phrases de grande longueur (>20 mots) qui commence à apparaître à ce niveau. Beaucoup d'erreurs se situent dans ces genres de phrases à cause de la différence d'ordre des mots entre ces deux langues qui affecte la précision de l'analyse.

5.7. Ordre des mots

Description :

Le français et l'anglais sont deux langues reliées ensemble dans un sens, parce que le français est une langue latine avec l'influence allemande et de l'anglais, alors que l'anglais est une langue germanique avec l'influence du latin et du français. Ainsi il y a quelques similitudes entre ces langues qui partagent le même alphabet et un certain nombre de cognates. Si ces langues partagent de nombreux éléments, il n'en reste pas moins qu'elles divergent sur de nombreux points. Notamment, il est connu que l'ordre des adjectifs et des noms dans une phrase anglaise est habituellement inversé en français. On parle par exemple d'une « montagne bleue », alors qu'en anglais, on parlera d'une « blue montain ». D'autres modifications dans l'ordre des mots apparaissent également, qui vont être expliquées dans cette section.

Ces différences peuvent affecter les performances de notre algorithme de projection. Dans cette section, on s'intéresse à l'ordre des mots entre ces deux langages, et l'effet qu'il a sur les liens projetés.

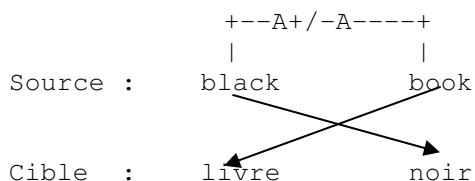
Pour comprendre cette idée, on va essayer d'expliquer une des plus importantes de ces différences : la construction des adjectifs. L'utilisation des adjectifs en français peut être difficile, parce qu'ils peuvent être placés avant ou après le nom, selon leur type et signification. Cela peut être différent ou inversé en anglais.

Prenons un exemple qui illustre ces deux types d'emplacement :

Une	table	ronde	-	round	table
un	livre	noir	-	black	book
une	jolie	fille	-	pretty	girl
un	jeune	homme	-	young	man

On constate que les adjectifs en anglais se placent avant le nom qu'il qualifie, ce qui diffère dans la plupart des cas en français. Puisque dans notre système d'analyse, les relations entre les mots se font à l'aide des liens avec des sens opposés, la projection du lien tel quel sera inefficace.

Reprenons la phrase de l'exemple précédent : *un livre noir* avec sa traduction *black book* :



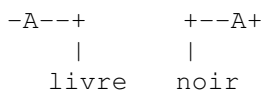
Dans le dictionnaire de la Link-Grammar, les deux mots anglais vont avoir les connecteurs suivants :

black : (A+).....
 book : (A-).....

Après projection des liens, les entrées du dictionnaire projeté pour ces deux mots seront :

noir : (A+).....
 livre : (A-).....

Et enfin, si on veut analyser la phrase : *un livre noir* ça serait impossible de lier ces deux mots ensemble:



Pour palier cette limitation, vient notre idée d'appliquer des règles de transformations sur les liens appartenant aux mots dont l'emplacement est différent dans la langue cible. Les règles consistent à ajouter des liens de sens opposés sur ceux concernés, afin d'avoir l'option de se connecter dans les deux sens. Pour permettre ceci, il faut à toute rencontre de ces genres de liens durant l'analyse anglaise, ajouter leurs homologues de sens opposés, afin qu'ils soient projetés sur les mots français.

Revenons au dictionnaire français et présentons ses entrées pour les deux mots de l'exemple précédent après la transformation subie par les liens :

noir : (A+ or A-).....
livre : (A- or A+).....

Maintenant alors, la relation entre ces deux mots sera possible en ajoutant les deux connecteurs convenables pour chacun d'eux. Dans cette approche, on veut essayer d'ajouter des règles similaires à celles expliquées dans le cas des adjectifs, pour tous les connecteurs appartenant aux mots qui ont un ordre différent dans la langue cible (français).

Résultats et conclusion :

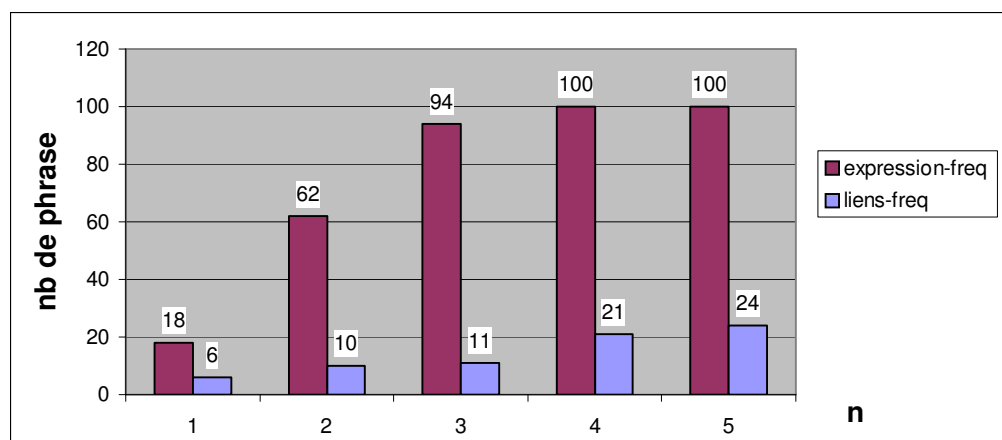


figure 38 : comparaison du nombre de phrases analysées des deux méthodes PL-npfa et PE-npfa en appliquant les transformations sur les liens.

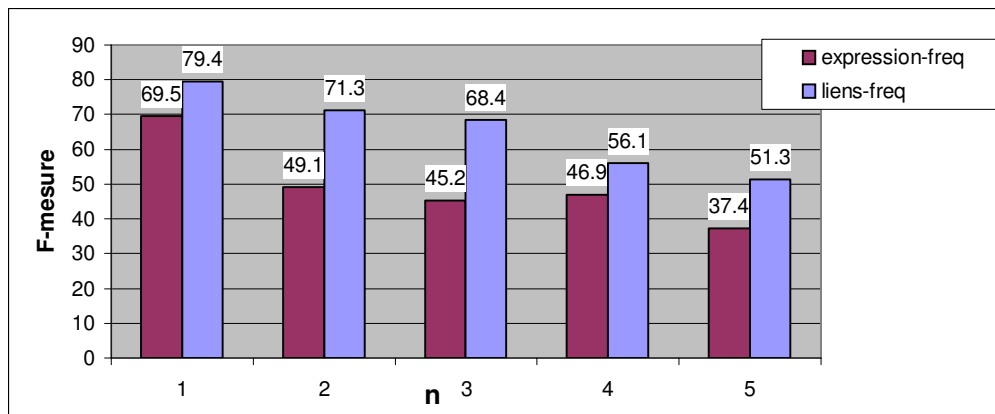


figure 39 : comparaison de la F-mesure des deux méthodes *PL-npfa* et *PE-npfa* en appliquant les transformations sur les liens.

Les résultats présentés dans les deux figures, 38 et 39, nous montrent la performance de la méthode qui projette toute l'expression d'un mot (*PE-npfa*), et la méthode qui projette juste les liens formés lors de l'analyse d'un mot (*PL-npfa*), avec « n », le nombre de mots anglais projetés, variant de 1 à 5.

Pour les mêmes raisons citées dans les méthodes précédentes, on voit toujours un nombre plus élevé de phrases analysées dans la méthode qui projette toute l'expression (*PE-npfa*). En plus, cette fois ce nombre est augmenté dans les deux méthodes utilisées.

Montrons donc combien on a pu analyser de phrases avec le croisement des connecteurs, par rapport aux méthodes *PE-npfa* (tableau 6) et *PL-npfa* (tableau 7).

<i>PE-npfa</i>	1	2	3	4	5
Sans règles de transformations	11	50	83	101	114
Règles de transformations	18	62	94	100	100
Différence	7	12	11		

Tableau 6 : comparaison des nombres de phrases analysées entre *PE-npfa* sans croisement et avec croisement.

Notons que l'ensemble des phrases du corpus de test, dont l'ordre des mots (au plus deux mots pour chaque phrase) est différent dans l'Anglais, est constitué de 28 phrases. En appliquant les transformations, nous avons montré dans le tableau 6 qu'il y avait 12 phrases analysées appartenant à cet ensemble, et donc une réussite de 42%.

Pour des valeurs de « n » égale à 4 et 5, on voit qu'on analyse moins de phrases en ajoutant ces règles. Et ceci à cause de l'augmentation du nombre des expressions, parfois qui n'appartient pas aux mots convenables ou alignés convenablement.

Projection des liens <i>PL-npfa</i>	1	2	3	4	5
Sans règles de transformations	7	9	10	11	15
Règles de transformations	7	10	11	21	24
Différence	0	1	1	10	9

Tableau 7 : comparaison des nombres de phrases analysées entre *PL-npfa* sans croisement et avec croisement.

Au contraire, le tableau 7 nous montre que la méthode qui projette les liens n'était pas trop affectée par les transformations des connecteurs. Dans cette méthode les liens projetés sont ceux utilisés par le mot anglais durant l'analyse. Alors si ce mot préserve le même ordre qu'en français, les liens qui se croisent ne sont pas alors utilisés par ce mot. Par la suite ces liens et leurs transformations ne seront pas projetés sur le mot cible (français). C'est pourquoi, dans cette méthode on ne voit pas une grande amélioration au niveau du nombre de phrases analysées.

Dans cette approche, on a essayé d'introduire des règles sur les connecteurs, afin d'augmenter le nombre de phrases analysées. Comme c'est illustré dans les tableaux 7 et 8, la méthode *PE-npfa* a pu analyser plus de phrases, mais n'a pas atteint un niveau satisfaisant. Enfin, il nous paraît qu'il reste des limitations, que nous essayons de les décrire, empêchant de faire une analyse complète des phrases du corpus de test.

5.8. Traduction à la main (PE-main)

Description :

Pour une application, telle que la projection via alignement de mot, il est très important d'avoir un alignement fait manuellement. Le résultat sera une liste de mots du vocabulaire français du bitexte, où chacun est associé à un ou plusieurs mots anglais (le plus approprié). Mais 30 000 phrases à aligner est une tâche très longue à réaliser.

Une méthode consiste alors à traduire les mots du vocabulaire français, sans passer par le corpus bilingue, qui sera beaucoup plus facile à le faire. Le but de cette expérience est de simuler le travail d'un alignement de mot idéal, et ainsi d'éliminer toutes les possibilités des bruits d'alignement et de se concentrer sur les problèmes, en général, grammaticaux rencontrés durant la projection.

Nous présentons ainsi, en figure 40, un exemple d'une tranche de cette liste des mots français avec leurs mots anglais traduits à la main :

```

éviter avoid.v
syndicaux union.n
réagir react.v
rester stay.v
condamnation conviction.n
difficulté difficulty.n
camp camp.n
spectaculaire spectacular.n
maintenir maintain.v
réparties distributed
dépassent exceed.v
limité limited.a
limité limited.v

```

Figure 40 : Une tranche de la liste des mots traduits à la main.

Alors, pour chaque mot nous essayons de trouver sa traduction en anglais comme elle est présente dans le dictionnaire de la Link-Grammar. Par exemple, le mot « *dépassent* » est traduit vers « *exceed.v* » et non vers « *exceed.n* », c'est pourquoi on doit identifier le fonctionnement du mot par les sous-groupes (.v, .n,...). Et parfois, un mot français peut être utilisé dans plusieurs fonctions grammaticales. Par exemple, le cas du mot « *limité* » qui peut être un adjectif comme il peut prendre le rôle d'un verbe aussi. Nous

montrons, dans la figure 41, une liste des quelques mots dont chacun peut avoir plusieurs sens différent.

```
Est is.v
Est East.a

fait fact.n
fait make.v

pas not
pas step.n
```

Figure 41 : une liste des mots dont chacun à un sens différent.

Nous avons donc essayé de couvrir autant que possible les cas que peut prendre un mot français pour pouvoir remplacer un bon aligneur de mots et un grand corpus d'entraînement.

Résultats et conclusion

Cette méthode peut être comparée à celle qui projette l'expression du mot le plus fréquemment aligné, puisque ces deux méthodes projettent l'expression entière du mot. On peut voir les résultats dans le tableau 8, qui indique une amélioration de la performance, au niveau du nombre des phrases analysées (50) et la F-mesure (63%).

<i>PE-main</i>	Précision	Rappel	F-mesure	Crosing-bracket	Nombre phrase
<= 10	78.1 %	87 %	82.3 %	0	18
<= 20	49.5 %	62.2 %	55.1 %	2.5	28
<= 40	30.6 %	39 %	34.3 %	5.2	4
La moyenne	57.7 %	68.7 %	62.7 %	1.9	50

Tableau 8 : résultats de la méthode qui projète l'expression du mot par traduction à la main.

En tant que nombre de phrases, le système a pu analyser 32 de plus (77% de gain) avec cette méthode (18 avec la *PE-pfa*), ce qui confirme l'importance de l'alignement de mot dans la performance. La précision (F-mesure) étant située à un niveau acceptable et très proche par rapport à la méthode *PE-pfa*. En appliquant aussi les transformations sur les

liens qui se croisent, nous avons pu analyser 17 phrases des 28 (le nombre des phrases dont le croisement des liens apparaisse) au total. Donc 60% de réussite par rapport aux phrases dans lesquelles le croisement des liens apparaît.

Les problèmes rencontrés qui nous ont empêché d'analyser plus de phrases, sont nombreux. Nous décrivons les plus importantes dans la section suivante.

5.9. Bilan des expériences

Après plusieurs expériences, il est très important de voir enfin, les résultats que notre travail nous a apportés. Nous comparons les méthodes décrites dans ce chapitre, et nous choisissons les trois les plus importantes. La projection de l'**expression** des « n » les plus fréquemment alignés, et surtout après l'application de l'alignement bidirectionnel et les règles de transformations sur les liens. De l'autre côté, et du même esprit, la projection des **liens** des « n » les plus fréquemment alignés. En ce qui concerne la troisième méthode, on a choisi celle qui projette l'**expression** du mot traduit à la main. On peut voir leurs résultats sur les deux graphes des figures 41 et 42 :

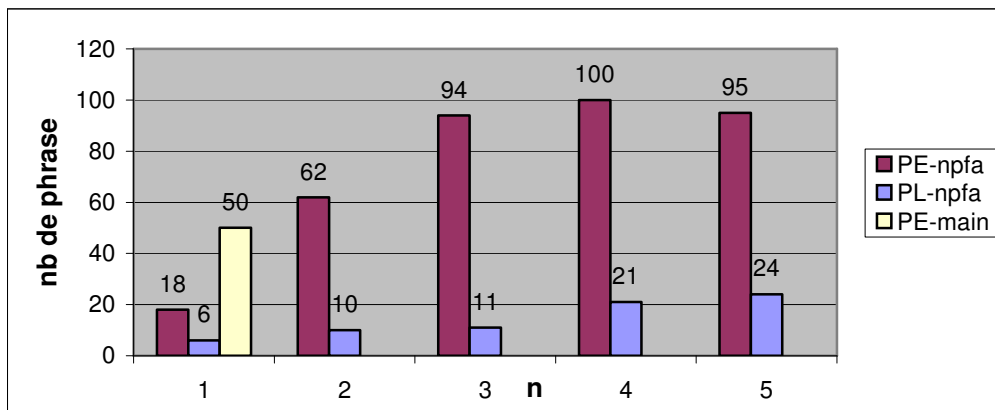


Figure 41 : Comparaison du nombre de phrases analysées des méthodes PL-npfa, PE-npfa et PE-main.

On constate, qu'il y a plus de phrases analysées par la méthode **PE-npfa** que par la méthode **PL-npfa**, pour toute valeur de n . Sans surprise, la méthode **PE-main** a pu analyser 50 phrases dépassant ainsi les deux autres méthodes. Ce qui est normal à cause du bruit d'alignement diminué dans cette méthode, qui joue un rôle important dans la projection. Ces résultats nous laissent conclure qu'il y a des limitations, autres que le croisement des mots et le bruit d'alignement, empêchant cette approche de bien couvrir autant des phrases

que possible. Ces problèmes envisagés et non résolus encore vont être le sujet de la prochaine section.

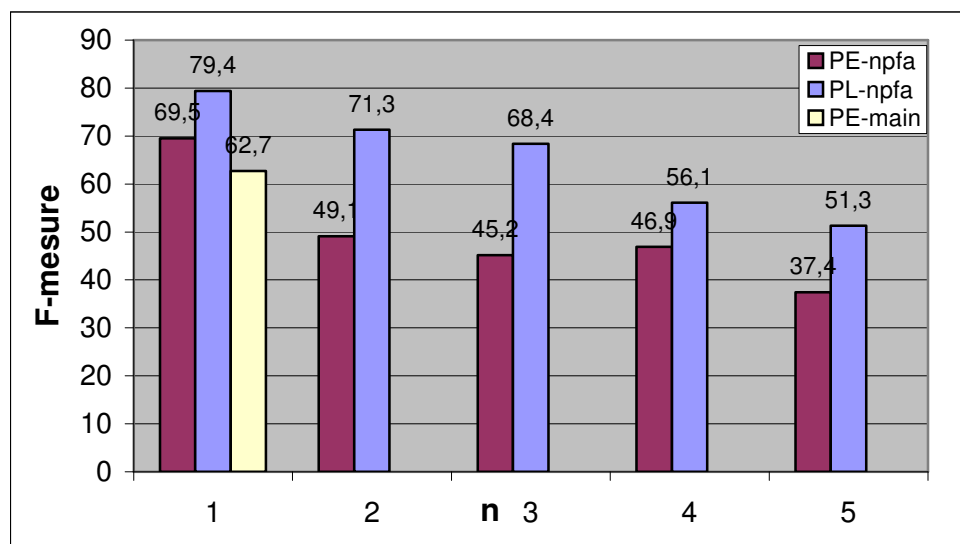


Figure 42 : Comparaison de la F-mesure des méthodes PL-npfa, PE-npfa et PE-main.

Dans la figure 42, nous montrons une comparaison entre ces trois méthodes en ce qui concerne leurs précisions (F-mesure). Nous constatons que le résultat de la méthode *PE-main* dépend d'une seule valeur de n parce que les mots sont traduits à la main. Dans ce contexte, il est préférable de comparer ces trois méthodes pour « n » égale à 1. La F-mesure, pour les trois dépasse un niveau acceptable du point de vue précision dont le résultat est proche l'un de l'autre.

Rappelons maintenant que l'option NULL-LINK décrite dans le chapitre 2 permet au système d'analyser des phrases en laissant des liens non établis (c'est ce qu'on va l'appeler lien vide). Dans une comparaison avec le travail de Hwa [Hwa et al., 2001], nous aimerions prendre cette option en considération. Puisque dans toutes nos expériences sa valeur n'était que nulle (interdit d'établir des liens vides). Tandis que le modèle de Collins [Collins, 1997] utilisé par Hwa analyse toutes les phrases, même si des relations n'ont pas pu être établies. Nous montrons pour cela, au tableau 9, les résultats des trois méthodes en prenant la valeur 2 pour NULL-LINK (toujours pour $n=1$). Ainsi le système de la Link-Grammar va permettre deux liens vides de s'établir dans chaque phrase du corpus de test, dont son résultat est traduit sur la ligne où « Null=2 ». Tandis que « No Null » montre les résultats où les liens vides sont interdits de s'établir.

performance		Nombre de phrase	F-mesure
PL-npfa	Null = 2	13	71.1%
	No Null	6	79.4%
PE-npfa	Null = 2	37	62.6%
	No Null	18	69.5%
PE-main	Null = 2	89	51.6%
	No Null	50	62.7%

Tableau 9 : comparaison des méthodes PL-npfa, PE-npfa et PE-main en ajoutant l'option de NULL-LINK.

Nous avons pris une valeur faible de « NULL-LINK » pour montrer que le système peut analyser autant de phrases, qui dépend du nombre de liens vides dans une phrase donnée. Nous constatons que le nombre de phrases analysées est augmenté et que la F-mesure diminue. Rappelons, que le critère de jugement, quelle que soit l'approche, est le rapport entre la précision et le nombre de phrases analysées qui doivent tous les deux être à un niveau élevé. Pour cela, on constate encore une meilleure performance de la méthode *PE-main*, dont le nombre de phrases analysées est amélioré par rapport aux autres (66% des phrases du corpus de test et un gain de 78% de la méthode de base). Ce nombre est encourageant pour les autres méthodes qui seront améliorées si un alignement de mot non bruité est appliqué.

Revenons à l'idée de Hwa [Hwa et al., 2001], ils obtiennent après l'application des règles de transformations linguistiques environ 67% pour la F-mesure. Par contre, notre meilleure performance dont le résultat du F-mesure atteint environ les 63% est offerte par la méthode *PE-main*. Un résultat proche d'eux, même avec les limites trouvées (que nous décrierons dans la section 5.9.1) qui ont empêché cette approche d'atteindre une meilleure performance. Notons que la méthode *PE-main*, qui compte sur la traduction manuelle, n'est pas pratique et même difficile de faire une traduction à la main pour un corpus bitexte qui peut s'étendre jusqu'à la centaine des milliers de phrases. Sa description et son

implémentation alors, permettent d'éliminer les doutes de l'alignement des mots bruités. Cette conclusion fait ressortir les limites de l'approche de la projection des relations syntaxiques.

5.9.1. Les limitations.

Nous allons maintenant décrire les limitations de cette approche suite à cette analyse qualitative et quantitative.

- Un pour plusieurs (**one-to-many**) : C'est le cas d'un mot anglais aligné à un ou plusieurs mots français. On sait que *IBM* permet un tel alignement. Parfois dans une même phrase, on observe plusieurs mots français alignés à un seul mot anglais. Pour la mieux comprendre, nous illustrons cette limitation sur un exemple en figure 43. Nous considérons pour cela la phrase anglaise, *it is not sure*, analysée avec *Link-Parser*. Sa traduction française, *ce n'est pas sûr*, est analysée par le même système mais en utilisant le dictionnaire formé par la projection :

Phrase anglaise:

```

          +-----Pa-----+
    +---Wd---+---Ss+EBm+   |
    |         |   |   |   |
LEFT-WALL it is.v not sure.a

```

Phrase française:

```

          +---Pa-----+
    +---Wd---+---Ss+EBm+   |
    |         | +--?--+   | |
    |         | |   |   | |
LEFT-WALL ce n' est pas sûr

```

Figure 43 : Un exemple montrant l'échec de la projection des relations syntaxiques.

Dans cet exemple, il n'est pas trivial d'établir l'entrée dictionnaire du mot « *n'* » car il n'existe pas de relation dans la phrase anglaise liant « *not* » au mot « *is* » dans ce sens. Rappelons que dans le formalisme des links grammars, les liens sont directionnels (nous avons une relation de *is* à *not* – *R(is,not)*, mais pas l'inverse). Il serait bien sûr possible d'appliquer des heuristiques pour ces cas comme celui-ci. Nous préférons à ce stade de notre travail considérer ce problème (one-to-many) comme une limitation des techniques

de projections que nous étudions. La négation, rencontrée et expliquée dans cet exemple, est la plus souvent envisagée parmi les limitations. D'autres ont été rencontrés dans le corpus de test, par exemple :

d'autres => others

d'entre => between

les rumeurs les plus alarmistes => the most alarmist rumours

Ce sont tous des exemples où un seul mot anglais est associé à plusieurs mots français. Dans la dernière phrase : le mot « *rumeurs* » ne peut pas être relié avec « *les* ». Le Link-Parser ne permettant pas une relation d'un nom avec un déterminant si ce nom est en relation avec l'adjectif qu'il quantifie (*R(the, rumours) ne peut pas être établie dans cette phrase*). Alors, même si « *les* » est aligné à « *the* », on aura pas le droit de lier « *les* » (un déterminant) avec « *rumeurs* » (un nom qu'on décrit).

- Plusieurs pour un (**many-to-one**) : C'est le cas de plusieurs mots anglais alignés à un seul mot français. Même si ce cas n'existe pas dans les alignements de *IBM*, mais en appliquant l'alignement bidirectionnel, cette limitation sera encore présente. Nous montrons donc comment elle affecte notre résultat sur un exemple illustré en figure 44, du *futur simple* en anglais et son équivalent en français.

Exemple:

```

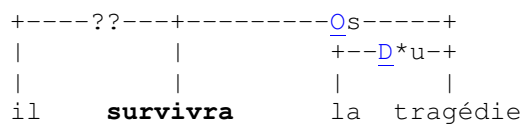
+-----O-----+
+-S-+---I---+   +---D*u-+
|   |   |   |   |   |
he will.v survive.v the tragedy.n

il      survivra      la tragédie

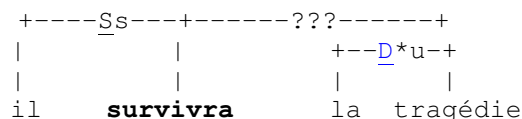
```

Figure 44 : Un exemple du futur simple en anglais et sa traduction en français.

Dans *IBM*, « *survivra* » va être aligné à un seul mot anglais de la phrase en parallèle. S'il est aligné avec « *survive.v* », alors il ne peut pas être lié vers sa gauche avec « *il* ». Parce que la relation *R(he, survive.v)* n'existe pas dans le système de la Link-Grammar, et donc la relation *R(il, survivra)* ne peut pas être établie par la suite.



Et si «*survivra*» est aligné avec «*will.v*», alors de même, le mot français ne peut pas être lié avec les autres mots de la phrase, à sa droite.



Donc dans les deux cas il ne peut pas être lié, et par la suite telle phrase ne peut pas être analysée.

Maintenant, en appliquant l'alignement bidirectionnel, et si «*survivra*» est aligné aux mots «*will*» et «*survive*», l'entrée du dictionnaire projeté du mot «*survivra*» va utiliser les liens de l'un de ces mots, et non pas tous les deux ensembles. Alors, on rencontrera le même problème envisagé avec le modèle de **IBM**. Nous comptons environ 22 phrases du corpus de test, dans lesquelles existe cette limitation. Toutes les statistiques sont donc montrées dans le tableau 10.

Limitations trouvées	One-to-many	Many-to-one	Croisement
Limitations traitées			Croisement
Nombre de phrases dans le corpus de test	30	22	28
Pourcentage du nombre de phrases analysées			42%

Tableau 10 : Statistiques montrant le nombre des phrases ayant les limitations.

Nous pouvons en déduire que ces limitations dépendent de la différence linguistique entre ces deux langues (Français/Anglais). Cette différence, à un certain niveau, nous a empêché de construire une grammaire française efficace à l'aide de la projection directe de celle de l'Anglais.

Un autre problème lié à la diversité des liens qu'un mot peut avoir dans la grammaire de la « *Link-Grammar* ». Il y aura des conditions plus délicates afin d'utiliser ses liens dans une phrase, mais ces conditions qui sont valables pour l'Anglais, ne le sont pas, parfois pour le français. Par exemple, en anglais, la relation d'un déterminant avec son nom n'est pas toujours valide, ce qui n'est pas le cas pour le français, qui est toujours vraie. Contrairement à l'idée utilisé par Hwa [Hwa et al., 2001], dont la relation syntaxique projetée n'est pas soumise à des conditions : ***R(det, nom)*** est toujours valide en anglais.

Ces limites pourraient nous amener à combiner des informations statistiques et linguistiques afin d'améliorer le résultat.

Chapitre 6

Conclusion

Le problème de la projection directe des relations syntaxiques se divise en deux sous-problèmes que nous avons décrits dans le chapitre 5 (le bruit de l'alignement et les limitations liées à la linguistique). Dans ce contexte, peu de recherches sont publiées durant ces dernières années. On peut citer [Hwa et al.,2001] qui ont créé un algorithme adressant ces limitations en donnant des résultats insuffisants, mais encourageants.

Nos expérimentations confirment l'intuition linguistique, indiquant qu'on ne peut pas, sans risque, transférer directement des relations syntaxiques d'un langage vers un autre. Les analyses syntaxiques projetées de l'anglais vers le français, en principe rendent l'analyse française vers un niveau presque précis après l'application d'un ensemble des règles linguistiques.

Nous avons concentré nos efforts sur la réalisation d'un algorithme de projection suivant deux méthodes. La première consiste à prendre les liens syntaxiques partant de chaque mot source, qui sont formés lors de l'analyse du système Link-Parser, en les projetant sur les mots cibles qui lui sont alignés. Cette méthode paraît acceptable en offrant une précision proche des 80%. La deuxième implémente une technique qui utilise la définition du mot dans le dictionnaire de la Link-Grammar et la projette telle quelle. La précision de cette méthode atteint les 68%. Une valeur, qui même si elle est plus petite que la précédente, reste prometteuse. Plusieurs méthodes se dérivent ainsi de ces deux approches. Nous avons dans ce mémoire présenté et testé plusieurs de ces méthodes et discuté de leurs limites.

Quelle que soit la méthode utilisée dans cette approche, elle n'a pas pu couvrir toutes les phrases du corpus de test. C'est-à-dire que le nombre de phrases françaises analysées est insuffisant pour pouvoir se confier à cette idée de projection bruitée.

L'avantage qui nous a permis de projeter une Link-Grammar est la diversité des liens que peut avoir un mot donné. Cet avantage est transformé en une limitation (beaucoup de conditions d'utilisation pour un lien donné) qui nous a empêché d'analyser plus de phrases (si un lien dans une phrase donnée n'est pas établi alors toute la phrase ne serait pas acceptée par l'analyseur). Même encore, le système Link-Parser lancé sur un corpus anglais uniforme, ne peut pas analyser toutes les phrases du corpus. Par contre, nous avons montré, par la traduction manuelle des mots du vocabulaire du corpus de projection, que l'utilisation d'un alignement de mots non bruité peut améliorer la performance de la projection après l'application des règles de transformations sur les liens.

Dans une perspective future, le travail peut adresser les limites de notre projection de la grammaire (décrites dans la section Bilan du chapitre 5), qui méritent un traitement linguistique uniforme. Ces limites représentent les différences dans les expressions dans les deux langues. Par exemple, l'expression nominale se traduit par un verbe ou le contraire. Nous croyons aussi que le contexte peut être employé pour effectuer les transformations correctes des expressions. La correction peut être également faite par l'intermédiaire des techniques d'entraînement statistique.

Références

1. Daniel Sleator and Davy Temperley, *Parsing English with a Link Grammar*, Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.
2. Daniel Sleator and Davy Temperley, *Parsing English with a Link Grammar, Third International Workshop on Parsing Technologies*, August 1993. This is a shorter but more up-to-date version of the technical report above. It contains a better introduction to link grammars, and gives a more detailed description of the relationship between link grammar and other formalisms.
3. John Lafferty, Daniel Sleator, and Davy Temperley, *Grammatical Trigrams: A Probabilistic Model of Link Grammar*, *Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language*, October, 1992. This paper introduces a statistical language model based on link grammars.
4. Dennis Grinberg, John Lafferty and Daniel Sleator, *A robust parsing algorithm for link grammars*, Carnegie Mellon University Computer Science technical report CMU-CS-95-125, and *Proceedings of the Fourth International Workshop on Parsing Technologies*, Prague, September, 1995. This paper describes the modifications of the parsing algorithm used to allow null links.
5. E. Fong and D. Wu, *Learning Restricted Probabilistic Link Grammars*, *IJCAII Workshop on New Approaches to Learning for Natural Language Processing*, August, 1995, Montreal, Canada, pp 49-56.
6. Carol Liu, *Towards A Link Grammar for Chinese*, Submitted for publication in *Computer Processing of Chinese and Oriental Languages - the Journal of the Chinese Language Computer Society*. Full text available on request from carol@csvax1.ucc.ie or gordon@csvax1.ucc.ie.
7. Carl de Marcken, *Lexical Heads, Phrase Structure, and the Induction of Grammar*, In *1996 Workshop on Very Large Corpora*.
8. Richard Sutcliffe, Tom Brehony, and Annette McElligott, *The grammatical analysis of technical texts using a link parser*, in *PACLING-II*, April 1995. Full text available from richard.sutcliffe@ul.ie.
9. Yarowsky, D. and G. Ngai, "Inducing Multilingual POS Taggers and NP Brackets via Robust Projection Across Aligned Corpora." In *Proceedings of NAACL-2001* (ISBN: 1-55860-775-7), pp. 200-207, 2001.
10. Mann, G. and D. Yarowsky, "Multipath Translation Lexicon Induction via Bridge Languages." In *Proceedings of NAACL-2001* (ISBN: 1-55860-775-7), pp. 151-158, 2001.
11. Yarowsky, D., G. Ngai and R. Wicentowski, "[Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora](#)." In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research* (ISBN: 1-55860-786-2), 2001.
12. Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. [Bootstrapping Parsers via Syntactic Projection across Parallel Texts](#). *Special Issue of the*

Journal of Natural Language Engineering on Parallel Texts, Eds. Rada Mihalcea and Michel Simard. To appear.

13. Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak, "Evaluating Translational Correspondence using Annotation Projection," to appear in the Proceedings of the 40th Annual Meeting of the ACL 2002.
14. Kuhn, Jonas. 2004b. Exploiting parallel corpora for monolingual grammar induction--a pilot study. In Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora, LREC 2004.
15. Knight, K. 1999. *A Statistical Machine Translation Tutorial Workbook*. Tech. Rep., USC/ISI. (available at <http://www.clsp.jhu.edu/ws/projects/mt/wkbk.rtf>).
16. Brown P. F., Cocke J., Pietra S. A. D., Pietra V. J. D. Jelinek F., Lafferty J. Roosin P. S., MERCER R. L. (1993). *A Statistical Approach to Machine Translation*.
17. E. Black and et al. *A procedure for quantitatively comparing the syntactic coverage of english grammars*. In Proc. of the 1991 DARPA Speech and Natural Language Workshop, pages 306-- 311, 1991.
18. C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth.