Université de Montréal

Alignement de phrases parallèles dans des corpus bruités

par
Fethi Lamraoui

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

Juillet, 2013

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Alignement de phrases parallèles dans des corpus bruités

présenté par:

Fethi Lamraoui

a été évalué par un jury composé des personnes suivantes:

Guy Lapalme,	président-rapporteur
Philippe Langlais,	directeur de recherche
Derek Nowrouzezahrai,	examinateur externe

Mémoire accepté le: . . . . . .

RÉSUMÉ

La traduction statistique requiert des corpus parallèles en grande quantité. L'obtention de tels corpus passe par l'alignement automatique au niveau des phrases. L'alignement des corpus parallèles a reçu beaucoup d'attention dans les années quatre vingt et cette étape est considérée comme résolue par la communauté. Nous montrons dans notre mémoire que ce n'est pas le cas et proposons un nouvel aligneur que nous comparons à des algorithmes à l'état de l'art.

Notre aligneur est simple, rapide et permet d'aligner une très grande quantité de données. Il produit des résultats souvent meilleurs que ceux produits par les aligneurs les plus élaborés. Nous analysons la robustesse de notre aligneur en fonction du genre des textes à aligner et du bruit qu'ils contiennent. Pour cela, nos expériences se décomposent en deux grandes parties. Dans la première partie, nous travaillons sur le corpus BAF où nous mesurons la qualité d'alignement produit en fonction du bruit qui atteint les 60%. Dans la deuxième partie, nous travaillons sur le corpus EuroParl où nous revisitons la procédure d'alignement avec laquelle le corpus Europarl a été préparé et montrons que de meilleures performances au niveau des systèmes de traduction statistique peuvent être obtenues en utilisant notre aligneur.

Mots clés : Corpus parallèles, Alignement de phrases, Bruit, Corpus bruit

ABSTRACT

Current statistical machine translation systems require parallel corpora in large quantities, and typically obtain such corpora through automatic alignment at the sentence level: a text and its translation . The alignment of parallel corpora has received a lot of attention in the eighties and is largely considered to be a solved problem in the community. We show that this is not the case and propose an alignment technique that we compare to the state-of-the-art aligners.

Our technique is simple, fast and can handle large amounts of data. It often produces better results than state-of-the-art. We analyze the robustness of our alignment technique across different text genres and noise level. For this, our experiments are divided into two main parts. In the first part, we measure the alignment quality on BAF corpus with up to 60% of noise. In the second part, we use the Europarl corpus and revisit the alignment procedure with which it has been prepared; we show that better SMT performance can be obtained using our alignment technique.

Keywords : Parallel corpora, Sentence alignment, Noise, Noisy corpora.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

REMERCIEMENTS

Je tiens à remercier en premier temps mon directeur de recherche M. Philippe Langlais qui m'a soutenu, conseillé, dirigé, et m'a appris beaucoup de choses plus particulièrement sa façon de résoudre les problèmes qui a constitué un apport considérable sans lequel ce travail n'aurait pas pu être mené au bon port.

Mes remerciements s'étendent également aux professeurs du DIRO pour leur enseignement et aux membres du laboratoire RALI pour l'ambiance de travail. Un merci particulier pour M. Fabrizio Gotti qui m'a apporté toujours son aide.

Je remercie ALLAH qui nous aide et nous donne la patience et le courage. Un grand merci à mes parents, à mes frères, ma sœur, ma tente et à Asma qui m'ont soutenu tout au long ces années d'études.

CHAPTER 1

INTRODUCTION

Les investissements dans le domaine de la traduction est en croissance remarquable, notamment au sein de l'union européen où tous les documents doivent être traduits dans toutes les langues officielles . En 2006[1], l'union européenne a dépensé environ 800 millions d'euros pour faire une telle traduction, contre environ 190 millions en 2005[2]. La traduction automatique est devenue de plus en plus nécessaire avec l'utilisation excessive de la technologie et du web qui sont à la base d'une explosion de l'information. Chaque jour, des milliers de textes sont ajoutés sur le web en plusieurs langues afin de les rendre accessibles à tous les utilisateurs quel que soit leurs préférences linguistiques. Nous trouvons ces textes multilingues au niveau des entreprises commerciales ainsi qu'au niveau des grands organismes tels que l'ONU, etc.

La recherche dans le domaine de la traduction automatique est en évolution permanente. L'idée de base de la traduction automatique est de produire une interprétation automatique d'un texte écrit dans une langue naturelle à une autre. Pour faire une telle interprétation, l'approche statistique requiert des corpus parallèles (des textes et leur traduction) en grande quantité. Ces corpus sont alors alignés automatiquement au niveau des phrases. Ainsi, l'étape d'alignement des phrases est essentielle à un système de traduction statistique. Cependant, l'alignement automatique des corpus parallèles est considéré comme une tâche résolue ou verrouillée. Nous pensons que cet état de fait est inadéquat à la réalité des corpus. Beaucoup sont en effet bruités, et les performances d'alignement sont médiocres pour certains types de documents. Notamment [Langlais et al., 1998] montrent qu'aligner des traductions d'ouvrages littéraires est une tâche plus difficile que d'aligner des textes parlementaires. En particulier, les auteurs montrent que sur la nouvelle de Jules Vernes "De la terre à la lune", les systèmes obtiennent des

[1] On parle ici de plus de 20 langues officielles.

[2] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+IM-PRESS+20071017FCS11816+0+DOC+XML+V0//FR.

performances très variables. Alors que, sur d'autres textes, les aligneurs approchent la perfection.

Plusieurs méthodes d'alignement de phrases ont été proposées [Brown et al., 1991, Gale et Church, 1991, Kay et Röscheisen, 1993]. Nous pouvons regrouper ces méthodes dans deux grandes catégories, selon les indices d'alignement qu'elles utilisent. Le premier indice est la longueur des phrases [Brown et al., 1991, Gale et Church, 1991], alors que le deuxième sont les propriétés lexicales [Chen, 1993, Kay et Röscheisen, 1993, Moore, 2002, Wu, 1994]. [Chen, 1993, Wu, 1994] trouvent que les méthodes d'alignement qui se basent sur les propriétés lexicales donnent de bons résultats en les comparant avec ceux produits par les méthodes qui se basent sur la longueur des phrases.

Les méthodes d'alignement qui se basent sur les indices lexicaux diffèrent selon la manière dont les caractéristiques lexicales sont acquises. Certaines de ces méthodes nécessitent un dictionnaire bilingue initial [Li et al., 2010, Wu, 1994], tandis que d'autres apprennent les informations lexicales après une phase d'entrainement [Chen, 1993, Kay et Röscheisen, 1993, Moore, 2002].

Les méthodes d'alignement qui apprennent de l'information lexicale après une phase d'entrainement sont typiquement moins rapides, et en quelque sorte dépendent des paires de langues en considération, et de la qualité du matériel à aligner (l'alignement de mots au sein d'une petite quantité de donnée constitue un défi). Contrairement aux indices lexicaux, les méthodes qui se basent sur les indices de longueur sont indépendantes de la langue [Gale et Church, 1991].

Le but de notre mémoire est de vérifier s'il est légitime de considérer le problème de l'alignement de phrases comme résolu. A cet effet, nous proposons une approche d'alignement qui améliore l'état de l'art. Nous déployons l'aligneur résultant dans différentes situations en en montrant le bon comportement global.

1.1 Définition de l'alignement et problématique

Les systèmes de traduction automatique requièrent des corpus parallèles en grande quantité. Dans les premiers temps, l'alignement phrastique des corpus parallèles se faisait manuellement, mais avec l'explosion de l'information, les corpus parallèles sont devenus très volumineux. Dès lors, l'alignement manuel de tels corpus est devenu une tâche très fastidieuse. Ce problème a conduit les chercheurs à penser à des techniques automatiques d'alignement. Les alignements manuels produits par des experts restent une source fiable, sur laquelle nous nous appuyons pour mesurer la qualité de l'alignement automatique.
Nous définissons dans ce qui suit, en quoi consiste un alignement de phrases, et quels sont les problèmes que nous pouvons rencontrer.

1.1.1 Définition de l'alignement

Plusieurs définitions ont été données pour l'alignement, ou l'appariement, par les chercheurs dans le domaine du traitement des langues naturelles. [Langlais, 1997] définit un système d'alignement multilingue automatique comme : « un processus qui prend en entrée un corpus multilingue ; c'est-à-dire un ensemble de textes traitant d'un même sujet dans des langues différentes et qui produit une sortie constituée d'appariements mettant en correspondance les régions (ou segments) qui sont en relation de traduction dans l'ensemble des textes du corpus. Une région est une unité textuelle pouvant relever de différents niveaux comme le chapitre, la division, le paragraphe, la phrase, la proposition, le terme, le mot, ou encore le caractère. ». [Simard, 1998] définit l'alignement comme : « la relation qui existe entre un texte et sa traduction, cette relation peut être vue à différents niveaux de granularité : entre texte, paragraphes, phrases, propositions, mots ou même caractères ». [Kraif, 2001] fait la distinction entre l'alignement et l'appariement. Il définit un alignement comme des occurrences d'un segment dans une langue correspondant à d'autres occurrences d'un segment dans une autre langue. Ici, nous parlons d'une correspondance observable en contexte, tandis qu'un appariement est une correspondance sémantique telle qu'on en trouve dans les dictionnaires.

1.1.2 Problématique

Le problème général de l'alignement automatique des corpus parallèles consiste à identifier automatiquement la segmentation optimale, c'est-à-dire, repérer voire mettre en relation, les segments sources avec les segments cibles correspondants. Cette correspondance peut se faire à différents niveaux : paragraphes, phrases, mots, ou même caractères. La figure 1.1 représente un exemple d'alignement au niveau des mots.

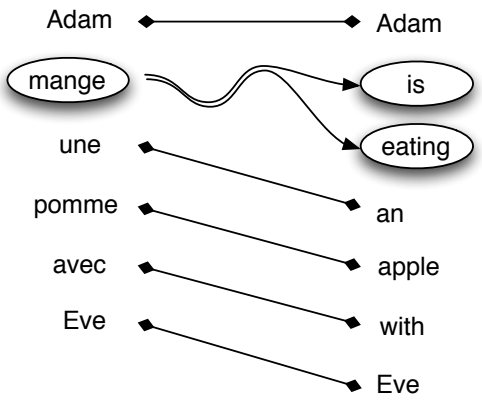


Figure 1.1: Exemple d'alignement au niveau des mots.

Pour les corpus parallèles parfaitement alignés, trouver un appariement ou une segmentation optimale, revient à mettre en relation la *i*-ème phrase du texte source avec la *i*-ème phrase correspondante du texte cible. En revanche, dans la vrai vie, la notion de textes parfaitement alignés est rare voire inexistante. Nous trouvons toujours des insertions ou des omissions dans les deux sens. Faire un alignement au niveau des corpus bruités est une opération d'autant plus difficile que le niveau de bruit augmente; ce que nous montrons dans le chapitre 5. Le tableau 1.I représente un exemple d'alignement de phrases au niveau des corpus parallèles bruités. Cet exemple s'agit d'une suppression dans la partie française, qui est elle-même considérée comme une insertion dans la partie anglaise.

Nous définissons l'alignement ou l'appariement dans notre mémoire comme la tâche de mettre en correspondance les phrases qui sont en relation de traduction dans un cor-

Français	**Anglais**
Débat L'intelligence artificielle	Artificial intelligence A debate
Depuis 35 ans, les spécialistes d'intelligence artificielle cherchent à construire des machines pensantes. Leurs avancées et leurs insuccès alternent curieusement.	Attempts to produce thinking machines have been met, during the past 35 years, with a curious mix of progress and failure.
	Two further points are important.
Les symboles et les programmes sont des notions purement abstraites.	Symbols and programs are purely abstract notions.

Table 1.I: L'alignement identifie une insertion/suppression coté anglais/français.

pus bilingue. Cette tâche se base sur des indices plutôt simples permettant d'aligner de manière satisfaisante les corpus au niveau des phrases. Nous décrivons plus en détail ces indices dans le chapitre 3

Dans ce mémoire nous décrivons notre méthode d'alignement de phrases. Dans chapitre 2 nous donnons une revue de littérature sur les corpus parallèles en décrivons les plus populaires dans le domaine de traitement de la langue. Le chapitre 3 est consacré à une revue de littérature sur les méthodes d'alignement de phrases les plus connues. Nous détaillons notre approche en chapitre 4, nous montrons pourquoi notre approche arrive à produire de bons résultats. Le chapitre 5 est consacré à des expériences menées sur deux corpus différents que nous détaillons ci-après et les résultats encourageants obtenus.

CHAPTER 2

CORPUS PARALLÈLES

De nos jours, avec l'expansion de la technologie et du web, l'accès à l'information est devenu de plus en plus facile, et ce de manière multilingue. De ce fait, le nombre de corpus parallèles a évolué d'une manière considérable. Ce chapitre est consacré à une vue d'ensemble des principaux corpus utilisés dans le domaine du traitement des langues. Nous distinguons les corpus parallèles, les corpus comparables et les corpus bruités. Nous définissons dans ce qui suit chaque type de corpus, puis, nous décrivons les principales ressources parallèles existantes.

2.1 Définition

2.1.1 Corpus parallèles

Une définition élémentaire des corpus parallèles peut se donner par un cas simple, où deux langues seulement sont impliquées : l'un des corpus est la traduction de l'autre. Nous généralisons cette définition, où plusieurs langues peuvent être impliquées : l'un des corpus dans une langue est la traduction directe des autres corpus dans d'autres langues. L'aspect du parallélisme existe dans plusieurs corpus, nous donnons l'exemple des débats parlementaires EuroParl[1], Hansard canadien[2] que nous décrivons en détail dans la section 2.2.

2.1.2 Corpus comparables

Selon [Laffling, 1992], les corpus comparables sont des textes composés dans différentes langues indépendamment les uns des autres. Ces textes expriment des idées semblables ou parlent du même sujet, mais ils ne sont pas la traduction les uns des autres. Nous trouvons les corpus comparables dans des sélections de journaux qui trait-

[1] `http://www.europarl.europa.eu/`.
[2] `http://www.parl.gc.ca`.

ent du même évènement. D'après [Déjean et al., 2002], "deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de la langue l1, respectivement l2, dont la traduction se trouve dans le corpus de la langue l2, respectivement l1". Il faut noter que, les corpus comparables peuvent également comprendre des sous-parties parallèles.

2.1.3 Corpus bruité

En partant de la définition des corpus comparables et corpus parallèles, nous représentons les corpus bruités comme des corpus parallèles contenants des segments de texte qui ne sont pas alignés. Selon [Fung et Mckeown, 1994], dans les corpus bruités, des segments dans la langue source peuvent être totalement absents dans la langue cible, ou peuvent être substitués avec des segments qui ne sont pas leur traduction correspondante. De même dans le coté inverse. Les corpus comparables peuvent être considérés comme des corpus bruités, [Gahbiche-Braham et al., 2011] considèrent les articles issus de services de presse en différentes langues comme des corpus parallèles bruités.

La figure 2.1 montre le lien de parallélisme qui se trouve entre les différents types de corpus que nous venons de discuter. Il convient de noter que la distinction de ces types de corpus revêt un coté arbitraire et que le degré de parallélisme est davantage un continuum.

2.2 Ressources

2.2.1 Historique

Selon [Véronis, 2000], le 1er corpus parallèle recensé est la pierre de Rosette. En juillet 1799, les soldats de l'armée de Napoléon ont découvert près de la ville de Rosette, sur le delta du Nil, une pierre qui s'avère d'une grande source de connaissances sur les langues presque mortes. Cette pierre est un fragment de stèle gravée de l'Égypte ancienne incluant trois versions d'un même texte. Elle a permis à Jean-François Champollion d'apporter en 1822 la clé du déchiffrement de l'écriture hiéroglyphiques. Les

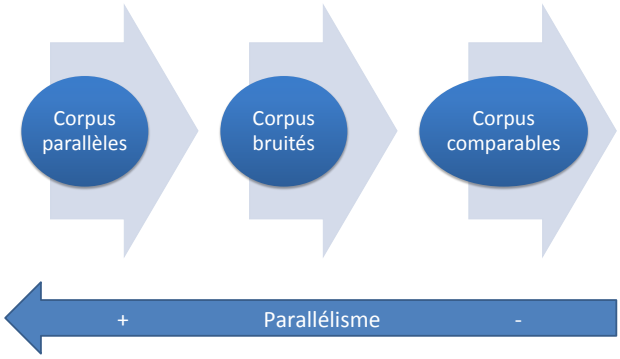


Figure 2.1: Lien entre les corpus et le niveau du parallélisme.

textes qu'elle comporte sont un décret promulgué à Memphis en 196 av. J.-C. au nom du pharaon Ptolémée V. Le décret est écrit en deux langues (égyptien ancien et grec ancien) et trois écritures : égyptien en hiéroglyphes, égyptien en écriture démotique et alphabet grec. L'apparition de cette pierre a mis fin à de nombreuses controverses et mythes qui avaient entouré les langues presque mortes.

2.2.2 Hansard

Le Hansard canadien est parmi les premiers corpus parallèles (français-anglais) utilisés dans les recherches en traduction automatique. Ce corpus est utilisé dans plusieurs recherches, [Brown et al., 1991], [Gale et Church, 1991] et [Véronis et Philippe, 2000]. Il est extrait des transcriptions officielles des débats parlementaires canadiens. Les versions distribuées en ligne ne sont pas toujours gratuites. C'est le cas avec LDC[3] qui propose une distribution payante des hansards qui a été collectée entre les années 1970 et 1988 par "IBM T. J. Watson Research Center" et "Bell Communication Research Inc.", et qui comprend 2,87 millions de paires de phrases. Une version des hansards canadiens

[3]Linguistic Data Consortium : `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T20`.

est disponible au RALI[4]. Cette version comprend 6,6 millions de phrases. Nous utiliserons ce corpus car il est très connu dans le domaine de la traduction automatique, et historiquement le plus utilisé.

2.2.3 EuroParl

Le corpus EuroParl ou (European Parliament) est extrait des débats parlementaires européens. Ces débats sont disponibles gratuitement sur internet[5]. Europarl est parmi les corpus les plus visés et les plus utilisés dans le domaine de la traduction statistique. Ce corpus est considéré comme une ressource multilingue [Koehn, 2005], car les textes sont disponibles en plusieurs langues, dont le français, l'anglais et l'allemand que nous étudions dans ce mémoire. La septième version de ce corpus comporte plus de 60 millions de mot par langue (environ 2 millions de phrases par langue). Le tableau 2.I en montre les caractéristiques principales. Nous utilisons la version la plus récente (Version 7) pour lancer une partie de nos expériences.

Langue	**Phrases**	**Vocabulaire**	**Langue**	**Phrases**	**Vocabulaire**
Bulgare	411,636	-	**Italien**	2,081,669	50,3 M
Tchèque	668,595	13,2 M	**Lithuanien**	678,665	11,5 M
Danois	2,323,099	47,8 M	**Latvien**	666,026	12,0 M
Allemand	2,176,537	47,2 M	**Hongrois**	658,824	12,6 M
Grecque	1,517,141	-	**Polonais**	387,490	7,1 M
Anglais	2,218,201	54,0 M	**Portugais**	2,121,889	52,3 M
Espagnol	2,123,835	54,8 M	**Roumain**	402,904	9,6 M
Estonien	692,210	11,4 M	**Slovaque**	674,359	13,1 M
Finnois	2,119,515	33,8 M	**Slovène**	634,488	12,7 M
Français	2,190,579	54,2 M	**Suédois**	2,241,386	45,7 M

Table 2.I: Caractéristiques principales de EuroParl.

[4]Cette version n'est pas gratuite et est disponible à la demande.

[5]Site officiel du parlement européen : `http://http://www.europarl.europa.eu/`.

2.2.4 BAF

Le corpus BAF[6] se distingue des autres corpus évoqués dans ce document. Premièrement, il ne se limite pas à un seul type de données. BAF regroupe en effet des textes de plusieurs genres. Deuxièmement, ce corpus est conçu avant tout pour être une référence pour l'évaluation des méthodes d'alignement automatique. Un alignement manuel est donc disponible, ce qui nous permet une évaluation fine. BAF a été développé par le laboratoire RALI (Recherche Appliquée en Linguistique Informatique) à l'Université de Montréal. il comporte plus de 800 000 mots (environ 400 000 mots par langue). Nous distinguons dans ce corpus 11 textes par langue qui sont répartis sous 4 types. Le tableau 2.II nous montre les caractéristiques principales de ce corpus.

N	Type	Description
1	**INST (Institutional texts)**	4 textes institutionnels pour un total de 300 000 mots par langue.
2	**SCIENCE**	5 articles scientifiques pour un total de 50 000 mots par langue.
3	**TECH (Technical Documentation)**	Manuels d'utilisation ex : (le manuel d'utilisation de Xerox). Environ 40 000 mots par langue.
4	**VERNE**	L'œuvre littéraire « De la terre à la lune » de Jules Verne qui comprend environ 50 000 mots par langue.

Table 2.II: Caractéristiques principales du corpus BAF.

Si les textes institutionnels sont réputés faciles à aligner, les textes littéraires présentent un défi. Selon [Langlais et Véronis, 1998] les traductions des textes littéraires sont plus libres. Le tableau 2.III représente quelques passages qui ne sont pas traduits dans le texte (Verne).

Nous constatons dans le tableau 2.III que, la 39ème ligne du texte français est traduite par la 32ème ligne du texte anglais. Ensuite, les lignes 40 à 59 sont sans traduction. La traduction se poursuit normalement à partir de la 60ème ligne du texte français et la 33ème ligne du texte anglais.

[6]Bi-texte Anglais Français : `http://rali.iro.umontreal.ca/rali/?q=en/node/71`.

Ligne	Français	Ligne	Anglais
39	Toutes ces inventions laissèrent loin derrière elles les timides instruments de l'artillerie européenne.	**32**	These inventions, in fact, left far in the rear the timid instruments of European artillery.
40	qu'on en juge par les chiffres suivant.		
:	:		
:	:		
59	c'était une réunion d'ange exterminateurs, au demeurant les meilleurs fils du monde		
60	il faut ajouter que ces yankees, braves à toute épreuve, ne s'en tinrent pas seulement aux formules et qu'ils payèrent de leur personne.	**33**	it is but fair to add that these Yankees, brave as they have ever proved themselves to be, did not confine themselves to theories and formulae, but that they paid heavily, in propria personal, for their inventions.

Table 2.III: Un passage non traduit du français à l'anglais dans le bi-texte VERNE.

2.2.5 JRC-Acquis

Le corpus JRC-Acquis ou (Joint Research Centre) est une ressource considérable de textes parallèles. Il s'agit des acquis communautaires[7] de l'union européenne (UE). La version 3.0 de ce corpus[8] est disponible en 22 langues officielles de l'union européenne. Le JRC-Acquis comprend 4,4 millions de documents alignés (dans toutes les langues). Nous montrons dans le tableau 2.IV, un extrait des acquis disponibles dans cette version.

[7]Ensemble des règles de droits auxquels doivent se conformer les membres de la communauté européenne

[8]`http://ipsc.jrc.ec.europa.eu/index.php?id=198`

(1) *Decision creating the "Official Journal of the European Communities"* ***(2)*** *THE COUNCIL OF THE EUROPEAN ECONOMIC COMMUNITY* ***(3)*** *Having regard to Article 191 of the treaty establishing the european economic community; having regard to the proposals from the president of the european parliament and the presidents of the high authority, the commission of the european economic community and the commission of the european atomic energy community; whereas the european economic community, the european coal and steel community and the european atomic energy community should have a joint official journal;* *Has decided:* ***(4)*** *To create, as the official journal of the community within the meaning of article 191 of the treaty establishing the european economic community, the official journal of the european communities.* ***(5)*** *Done at Brussels, 15 September 1958.*
(1) *Décision portant création du "Journal Officiel des Communautés européennes"* ***(2)*** *LE CONSEIL DE LA COMMUNAUTÉ ÉCONOMIQUE EUROPÉENNE* ***(3)*** *vu l'article 191 du Traité instituant la Communauté Économique Européenne; vu les propositions formulées par le président de l'assemblée parlementaire, les présidents de la haute autorité, de la commission de la communauté économique européenne et de la commission de la communauté européenne de l'énergie atomique; considérant qu'il est opportun que la communauté économique européenne, la communauté européenne du charbon et de l'acier et la communauté européenne de l'énergie atomique disposent d'un journal officiel commun; Décide:* ***(4)*** *De créer, en tant que journal officiel de la communauté au sens de l'article 191 du traité instituant la communauté économique européenne, le journal officiel des communautés européennes.* ***(5)*** *Fait à Bruxelles le 15 septembre 1958.*

Table 2.IV: Extrait des acquis de la version 3.0 avec l'alignement 1-1, 2-2, 3-3, 4-4, 5-5. Les numéros de ligne ont été ajouté par nos soin.

2.2.6 Wikipédia

Wikipédia[9] est une encyclopédie électronique libre créée en 2001. Elle est considérée comme une bonne ressource de corpus comparables, mais ça ne l'empêche pas

[9] `http://www.wikipedia.org`

d'avoir des pages parallèles (bruités ou pas). Le tableau 2.V représente un exemple d'un extrait de deux corpus parallèles, "Déclin de l'Empire romain d'Occident"[10] et "Decline of the Roman Empire"[11]. Ces deux corpus étaient parallèles en Octobre 2006, mais il ne le sont plus à l'heure actuelle.

(1) Le déclin de l'Empire romain, aussi appelé chute de l'Empire romain, est un terme historique de la période qui décrit l'effondrement de l'Empire romain d'Occident. **(2)** Le terme fut utilisé pour la première fois au xviiie siècle par Edward Gibbon dans sa fameuse étude Histoire de la décadence et de la chute de l'Empire romain, mais il n'était ni le premier ni le dernier à spéculer sur ce pourquoi et quand l'Empire romain a chuté. **(3)** Cela reste une des plus grandes questions historiques, et a une tradition riche de participation d'érudits. **(4)** En 1984, le professeur allemand Alexander Demandt publia une collection de 210 théories sur ce pourquoi l'Empire romain chuta.
(1) The decline of the Roman Empire, also called the fall of the Roman Empire, is a historical term of periodization that describes the collapse of the Western Roman Empire. **(2)** Gibbon in his famous study The Decline and Fall of the Roman Empire (1776) was the first to use this terminology, but he was neither the first nor the last to speculate on why and when the Empire collapsed. **(3)** It remains one of the greatest historical questions, and has a tradition rich in scholarly interest. **(4)** In 1984, German professor Alexander Demandt published a collection of 210 theories on why Rome fell.

Table 2.V: Extrait de corpus parallèles "Déclin de l'Empire romain d'Occident" et "Decline of the Roman Empire" avec l'alignement 1-1, 2-2, 3-3, 4-4. Les numéros de ligne ont été ajouté par nos soin.

Wikipédia est considérée comme une source de corpus comparables plus que paral-

[10] http://fr.wikipedia.org/w/index.php?title=D%C3%A9clin_de_l%27Empire_romain_d%27Occident&oldid=11030799

[11] http://en.wikipedia.org/w/index.php?title=Decline_of_the_Roman_Empire&oldid=83652798

lèles. C'est une des raisons pour laquelle nous ne l'avons pas choisi comme une source de données pour nos expériences.

Il existe de nombreuses autres ressources parallèles que nous ne pouvons pas toutes citer dans ce chapitre.

CHAPTER 3

ÉTAT DE L'ART DE LA RECHERCHE SUR L'ALIGNEMENT DES CORPUS PARALLÈLES

Plusieurs axes de recherche exploitant les potentialités des corpus parallèles ont été explorés. Des études ont également été réalisées sur les corpus comparables ainsi que dans le domaine de la recherche d'information multilingue. Nous décrivons en détail dans ce chapitre les indices d'alignement, ainsi que les méthodes d'alignement qui exploitent ces indices d'une manière ou d'une autre. Ensuite, nous discutons les caractéristiques de chacune de ces méthodes.

3.1 Indices d'alignement

Plusieurs indices ont été proposés pour identifier la correspondance qui se trouve entre un texte dans une langue et sa traduction. [Lecluze, 2011] classifie ces indices dans sa thèse sous deux grandes catégories : similitude de longueur et similitude de distribution. La première catégorie (similitude de longueur), englobe les indices qui se basent sur les longueurs des phrases afin de déterminer les correspondances entre les phrases et leurs traductions. La deuxième catégorie (similitude de distribution), réunit les indices qui se basent sur les propriétés lexicales pour déterminer les phrases qui sont en relation de traduction. [Shi et al., 2006] classent les indices d'alignement de phrases sous trois modèles plutôt que deux : la première catégorie se base sur la longueur des phrases, la deuxième sur le lexique et la troisième regroupe les indices hybrides. Dans ce mémoire, nous répartissons les indices d'alignement sous deux grande catégories. Indices à base de longueur de phrases et indices lexicaux. Nous dressons dans la suite un tour d'horizon des indices relevant de ces deux catégories.

3.1.1 Indices à base de longueur de phrases

Le rapport entre la longueur des phrases sources et les phrases cibles est l'information principale utilisée pour faire l'alignement.

Longueur de phrases

[Brown et al., 1991] calcule le rapport de longueur[1] entre les phrases sources et cibles pour établir des correspondances entre elles. L'intuition de base est que « les phrases les plus longues dans une langue ont tendance à être traduites par les phrases les plus longues dans d'autres langues, et inversement, les phrases les plus courtes ont tendance à être traduites par les phrases les plus courtes dans d'autres langues ». La figure 3.1 montre la forte corrélation entre la longueur des paragraphes en relation de traduction dans deux langues telle que mesurée par [Gale et Church, 1991].
.Selon [Brown et al., 1991], cet indice a permis d'atteindre un taux de réussite avoisinant les 100% (99% sur le corpus Hansard). [Gale et Church, 1991] descendent à un niveau de granularité plus bas pour calculer le rapport entre les phrases. La longueur des phrases est comptée en caractères. Ceci présente l'avantage de ne pas s'appuyer sur la notion de mot, qu'il est difficile d'appréhender dans certaines langues, notamment le chinois.

Patterns

La définition la plus simple des patterns peut se donner par l'exemple suivant : une phrase cible alignée avec deux phrases sources, nous donne le pattern 1-2.

[Gale et Church, 1991] calculent la probabilité des patterns dans un corpus donné. Après un étiquetage manuel des rapports bancaires suisses, ils proposent de considérer uniquement six patterns. Le tableau 3.I montre les patterns proposés par [Gale et Church, 1991] et leurs probabilités. Nous utilisons ces valeurs dans notre aligneur.

Nous constatons que le pattern 1-1 (une phrase source est traduite par une phrase

[1] Dans leur cas la longueur des phrases est comptée en mot.

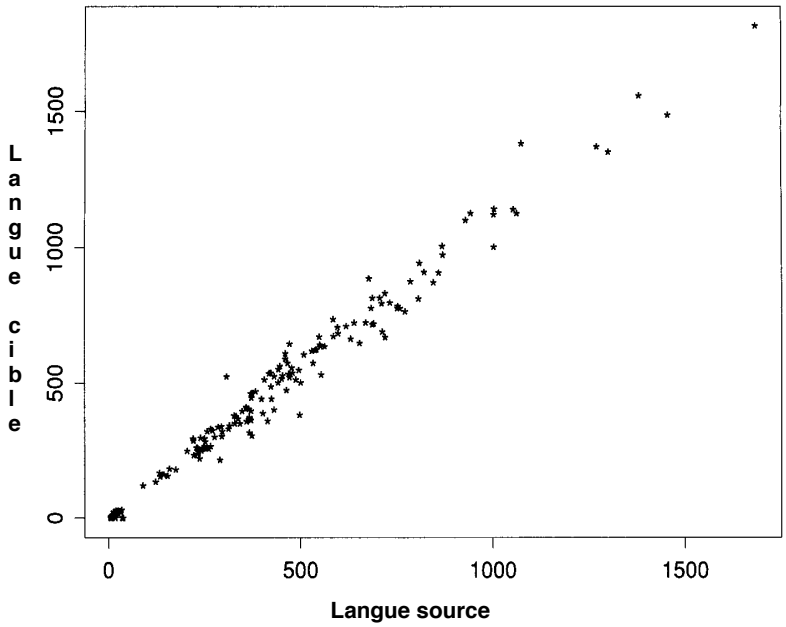


Figure 3.1: Corrélation entre les longueurs (comptées en caractère) de phrases anglaises et allemandes des rapports bancaires suisses qui sont en relation de traduction. Ce rapport est calculé par [Gale et Church, 1991].

Schéma (Pattern)	**Probabilité**
1-1	0.89 %
1-0/0-1	0.0099 %
1-2/2-1	0.089 %
2-2	0.011 %

Table 3.I: Les schémas proposés par Gale & Church et leur probabilité

cible) est le plus fréquent des patterns observés. Un étiquetage du corpus BAF (que nous décrivons plus loin) a donné approximativement les mêmes constations. Le tableau 3.II représente les valeurs obtenues à partir du texte Ilo[2].

3.1.2 Indices lexicaux

Ces indices s'appuient sur les propriétés lexicales pour faire correspondre une phrase dans une langue à une phrase dans une autre langue.

[2]Ilo: est l'un des textes qui compose le corpus BAF.

Pattern	Nb	%
1-1	6744	94.83%
2-1	203	2.85%
1-2	48	0.6%
3-1	37	0.5%
2-2	26	0.3%
4-1	15	0.2%
1-3	3	0.04%
1-0	13	0.18%
5-1	3	0.04%
0-1	6	0.08%
0-9	1	0.01%
0-2	5	0.07%
1-4	1	0.01%
2-0	5	0.07%
1-14	1	0.01%
TOTAL	7111	100%

Table 3.II: Valeurs obtenues après une évaluation manuelle du texte ilo un des textes du corpus BAF.

Les cognats

Dans ce sens, [Simard et al., 1992] proposent d'exploiter les *cognats* pour sélectionner les paires de phrases qui peuvent être traduction les unes des autres. Ils définissent les *cognats* comme : « deux mots sont dits cognats s'ils partagent des propriétés communes à un quelconque niveau (sémantique, orthographique, etc.)», plus précisément, deux mots sont considérés comme cognats s'ils partagent un préfixe de 4 caractères. Selon [Simard et al., 1992], il y a une corrélation forte entre les paires de phrases qui peuvent être traduction les unes des autres et le nombre de cognats qu'elles contiennent : "accès/access, activité/activity, parlement/parliament". La figure 3.2 est une représentation graphique de l'un des textes de BAF (VERNE). Cette représentation est générée à l'aide de l'outil "DotPlot"[3]. Chaque point dans cette figure représente la relation entre un mot source et un mot cible (en terme de *cognats*). Dans cette figure nous constatons

[3]Cet outil a été développé afin de montrer la similarité entre deux séquences. Dans notre cas, les séquences sont des paires de phrases.

qu'il y a une corrélation entre les paires de phrases qui sont traduction les unes des autres : la diagonale qui apparait est la résultante d'un plus grand nombre de cognats partagés dans les phrases qui sont en relation de traduction, et qui sont naturellement proches de la diagonale principale formée par le bi-texte.

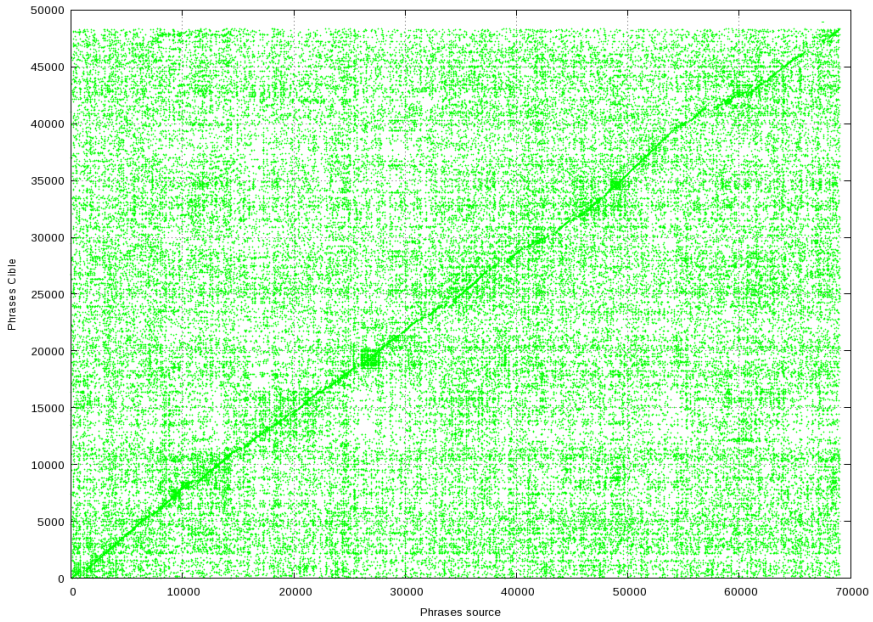


Figure 3.2: Exemple de diagonale due à la forte corrélation entre la source et la cible (VERNE) en terme de cognats. Chaque point représente une relation de cognat entre un mot source et un mot cible.

[Ribeiro et al., 2001] proposent un algorithme qui couvre davantage de cas de cognats, que la définition proposée par [Simard et al., 1992]. Un exemple de cas couvert par cet algorithme est : gouvernement/government). Ces deux mots ne partagent qu'un préfixe de trois caractères. Ils ne sont donc pas considérés par l'indice tel qu'implémenté par [Simard et al., 1992]. Toujours dans la cadre des cognats [Kondrak, 2001] propose de détecter les *cognats* au niveau phonétique. Il montre que l'utilisation des traits phonologiques des mots pour trouver les *cognats* est plus efficace. Selon lui cet indice apporte de meilleurs résultats. Notamment, s'il s'agit de deux langues qui ne partage pas le même alphabet (exemple : anglais et russe).

Les dictionnaires bilingues

Consiste à utiliser une information externe (les dictionnaires bilingues) au corpus pour faire la traduction. Dans ce sens [Varga et al., 2005] utilise un dictionnaire indépendant pour faire la traduction des mots sources.

Les modèles IBM

Le but est de trouver, comment une phrase cible $t = t_1, t_2 ... t_m$ sera traduite à partir d'une phrase source s composée de l mots, $s = s_1, s_2 ... s_l$. Dans le modèle IBM1[4] l'opération se fait comme suit.

Premièrement, la longueur m est choisie de la traduction stochastique à partir de la phrase source. Ensuite, pour chaque position de mot dans t, un mot dans s lui est associé (incluant le mot vide), dont il est supposé être la traduction. Dans les modèles IBM, seuls les alignements où chaque mot cible est associé à un mot source (et un seul) sont considérés. Ce modèle suppose que toutes les longueurs possibles de t ont la même probabilité, que nous la notons ε; que tous les positions sont possibles et équiprobables; et la probabilité $tr(t_j|s_i)$ d'un mot cible ne dépend que du mot source qu'il l'a généré:

$$P(t|s) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} tr(t_j|s_i) \tag{3.1}$$

Les paramètres $tr(t_j|s_i)$ sont appris de manière non supervisée à l'aide de l'algorithme EM[5].

3.2 Méthodes à base des indices de longueur

Plusieurs méthodes ont été proposées dans ce sens [Brown et al., 1991, Gale et Church, 1991, Kay et Röscheisen, 1993]. Ces méthodes reposent sur une idée simple

[4]Le modèle IBM1 est un modèle de traduction probabiliste.

[5]L'algorithme EM ou Expectation-Maximization : est une méthode itérative qui permet de trouver le maximum de vraisemblance des paramètres d'un modèle probabiliste. Cet algorithme peut être appliqué, lorsque le modèle dépend des variables non observables

qui est qu'il y a une relation proportionnelle entre les segments[6] qui peuvent être traduction les uns des autres, et la longueur de ces segments.

[Gale et Church, 1991] créent un modèle probabiliste qui est une approximation de la probabilité que deux patterns sont traduction l'un de l'autre. Leur méthode consiste à attribuer un score probabiliste à chaque alignement candidat, le meilleur alignement se calcul à partir des scores assignés aux candidats. Les auteurs utilisent la programmation dynamique afin de retracer le chemin des meilleurs candidats.

3.2.1 Fonction de score

La fonction de score (que nous appelons S_{gc}) permet d'assigner un score à chaque alignement. Les scores sont calculés par rapport aux longueurs des phrases (comptées en caractère).

$$\begin{aligned} S_{gc} &= -log(p(\delta|pattern) \times p(pattern)) \\ p(\delta|pattern) &= 2 \times (1 - p(|\delta|)) \\ \delta &= \frac{(l_2 - l_1 c)}{\sqrt{l_1 s^2}} \end{aligned} \tag{3.2}$$

Où $p(pattern)$: est donné par les comptes du tableau 3.I, et δ dépend des longueurs l_1 et l_2 qui sont respectivement les longueurs des phrases sources et cibles d'un pattern donné. Deux autres paramètres sont considérés. c: représente la moyenne (le nombre de caractères d'une langue source divisé par le nombre de caractères d'une langue cible), et s^2: représente la variance. Ces deux paramètres sont déterminés d'après une étude empirique faite sur les rapports bancaires suisses par: $c = 1$ et $s^2 = 6.8$. La figure 3.3 montre la forme globale de la distribution δ qui peut être approchée par une loi normale centrée réduite.

[6] Les segments peuvent être : une phrase, un paragraphe, etc.

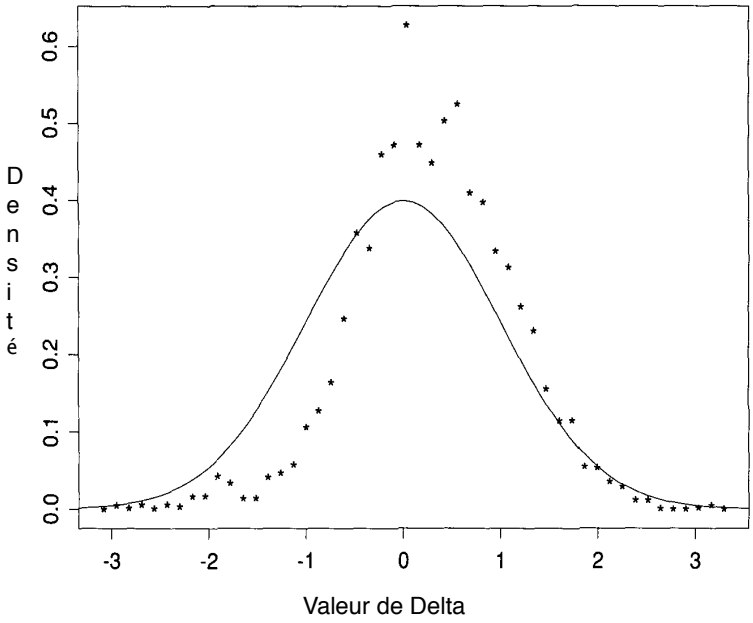


Figure 3.3: Distribution de δ sur un corpus aligné manuellement. L'axe des abscisses représente la valeur de δ, l'axe des ordonnées représente la densité de δ mesurée sur corpus aligné manuellement [Gale et Church, 1991].

3.2.2 Programmation dynamique

[Gale et Church, 1991] font appel à la programmation dynamique pour calculer le coût minimal entre les éléments alignés. Le fait d'utiliser la programmation dynamique réduit la complexité de l'alignement. Le coût d'un appariement donné est calculé par la fonction de score décrite dans la section 3.2.1. Le coût minimal se base sur les coûts des candidats, et le coût des patterns précédemment définis (0-1/1-0; 1-1; 1-2/2-1; 2-2). La figure 3.4 illustre comment le coût minimal d'un alignement est calculé par programmation dynamique.

Pour mieux illustrer l'opération. Soit $s_i, i = 1...I$ les phrases sources, et $t_j, j = 1...J$ leur traduction correspondante. Soit $D(i, j)$ le meilleur alignement entre les phrases $s_1...s_i$ et $t_1...t_j$. À l'initiale, $D(i, j)$ est nul pour tout i et j, puis on considère systématiquement tous les schémas d'appariement (0-1/1-0; 1-1; 1-2/2-1; 2-2):

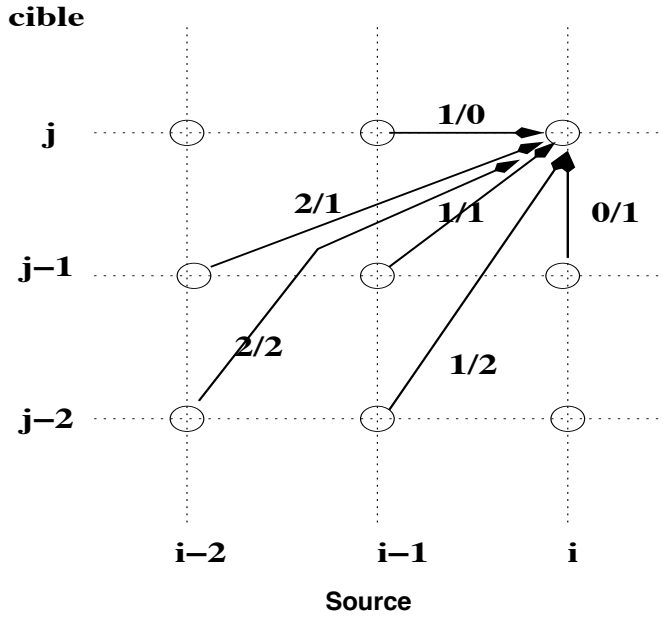


Figure 3.4: Trouver le coût minimal par programmation dynamique.

$$D(i,j) = min \begin{cases} D(i,j-1) & +d1 \\ D(i-1,j) & +d2 \\ D(i-1,j-1) & +d3 \\ D(i-1,j-2) & +d4 \\ D(i-2,j-1) & +d5 \\ D(i-2,j-2) & +d6 \end{cases} \tag{3.3}$$

Où $d_1, d_2 ... d_6$ sont les coûts respectifs de chaque pattern selon un score donné (par la fonction de score décrite dans la section 3.2.1).

Espace de recherche

[Brown et al., 1991, Chen, 1993, Gale et Church, 1991] calculent les scores des alignements d'une manière exhaustive. Un coût est assigné à chaque alignement pos-

sible. Puis, l'alignement global optimal se calcule par programmation dynamique, en prenant en considération chaque pattern. C'est-à-dire, le nombre totale des alignements possibles dans ce cas, se représente approximativement par le produit des nombres de phrases dans les deux langues. Ceci est clairement impossible avec des corpus qui contiennent plus d'un million de phrases. Par exemple EuroParl contient plus de deux millions de phrases pour le couple français-anglais.

3.3 Méthodes à base d'indices lexicaux

Plusieurs chercheurs [Chen, 1993, Simard et al., 1992] ont proposé d'utiliser l'information lexicale pour améliorer la qualité des alignements. D'autres, tentent d'enrichir la qualité d'alignement en utilisant les dictionnaires bilingues [Li et al., 2010]

3.3.1 Cognates

[Simard et al., 1992] proposent une définition pragmatique : « deux mots sont cognats s'ils ont en commun un préfixe d'au moins quatre caractères. Si l'un des mots contient au moins un chiffre, alors les deux mots sont cognats s'ils sont égaux.». Pour déterminer l'impact des *cognats* sur les bi-textes, [Simard et al., 1992] mesurent comment deux textes sont liés en terme de *cognats*. Pour cela, ils calculent γ: le nombre de *cognats* dans un segment de n mots source et m mots cible.

$$\gamma = \frac{c}{(n+m)/2} \tag{3.4}$$

3.3.2 Fonction de score

Afin de produire un alignement utilisant les cognats, [Simard et al., 1992] proposent de considérer une fonction de score, qui est la suivante:

$$S_{cog} = \frac{P_T(c|n)}{P_R(c|n)} \times p(pattern) \tag{3.5}$$

où $P_T(c|n)$ et $P_R(c|n)$ dénotent la probabilité que deux segments de longueur moyenne n (en mots) possèdent c *cognats*, sous les hypothèses respectives que les segments sont en relation de traduction $P_T(c|n)$, ou au contraire qu'ils ont été sélectionnés aléatoirement $P_R(c|n)$.
Les auteurs estiment sur une partie étiquetée du Hansard ces deux probabilités. Ils trouvent que ces deux probabilités suivent approximativement des lois binomiales. P_T et P_R ont été expérimentalement fixés à 0.3 et 0.09.

3.4 Méthodes hybrides

Les méthodes dites hybrides combinent plusieurs indices afin de produire une performance très élevée. Selon [Simard et Plamondon, 1998], deux grands problèmes se posent face à un processus d'alignement. D'une part, la robustesse de cet alignement, et de l'autre part, sa précision. Pour cela, ils proposent une méthode qui combine deux étapes pour faire l'alignement. La première étape exploite les indices à base de longueur de phrases, tandis que la deuxième, implémente un modèle statistique à bases des résultats obtenus auprès de la première étape.

Pour sa part [Moore, 2002] estime que les méthodes à base de longueur de phrases sont relativement robustes, mais moins précises, et que les méthodes à base des indices lexicaux sont généralement plus précises, mais en même temps moins rapides. À partir de cette observation, [Moore, 2002] propose une méthode qui combine les indices précédemment discutés. Sa méthode est parmi les méthodes les plus connues. Elle produit des résultats avec un faible taux d'erreur. Cette méthode passe par trois étapes essentielles. La première étape se base sur la longueur des phrases afin de les aligner. Cette étape fait appel à une version modifiée de [Brown et al., 1991]. La deuxième étape consiste à utiliser les paires de phrases déjà alignées pour entrainer un modèle de traduction statistique (une version modifiée de modèle IBM1). Le but de cette étape est de

produire un lexique bilingue de manière automatique. La dernière étape consiste à trouver l'alignement à coût minimal. Pour que sa méthode soit plus rapide, [Moore, 2002] ne prend en considération que les alignements qui ont une probabilité non négligeable.

La méthode de [Moore, 2002] est très similaire à celle de [Wu, 1994], qui utilise la longueur de phrases et l'information lexicale[7] pour trouver le meilleur alignement. Par contre, dans le cas de [Moore, 2002], l'information lexicale est dérivé automatiquement par le modèle de traduction statistique. De ce fait, aucune information externe n'est requise. Nous décrivons cette approche dans la suite.

3.4.1 Première étape : Alignement à base de longueur de phrases

[Moore, 2002] s'appuie sur l'indice de longueur tel que proposé par [Brown et al., 1991]. Dans leur cas, ils calculent le rapport entre la longueur des phrases sources l_t et la longueur des phrases cibles l_s qui varie selon une distribution gaussienne, avec la moyenne μ et la variance σ^2.

$$P(l_t|l_s) \quad = \alpha \exp(-((log(l_t/l_s) - \mu)^t/2\sigma^2)) \tag{3.6}$$

Où α est choisi de manière à ce que $P(l_t|l_s)$ somme à 1 pour les valeurs positives de l_t.

[Moore, 2002] suppose que la longueur des phrases cibles l_t varie selon une loi de Poisson dont la moyenne est simplement l_s fois le rapport r de la longueur moyenne de longueur phrases.

$$P(l_t|l_s) \quad = \exp(-l_s r)(l_s)^{l_t}/(l_t!) \tag{3.7}$$

Selon la loi de Poisson, chaque mot de la langue source est traduit en quelques mots dans la langue cible correspondante, dont la moyenne peut être simplement calculée comme le rapport entre les longueurs moyennes de phrases dans les deux langues.
Le modèle de [Moore, 2002] est simple à calculer, parce qu'il ne contient aucun paramètre

[7] L'information lexicale dans le cas de Wu est calculé sur un corpus donné avant l'alignement

caché. Alors qu'avec le modèle de [Brown et al., 1991], le calcul de la variance σ^2 se fait d'une manière itérative en utilisant EM, ce qui le rend moins rapide.

Espace de recherche

[Moore, 2002] propose une technique pour réduire l'espace de recherche. Pour cela, il considère que les alignements possibles forment une matrice, et que l'alignement optimal de tous le corpus forme en quelque sorte une diagonale. Ensuite, il considère une bande de largeur fixe autour de cette diagonale. À la fin, dans la phase de programmation dynamique, il ne tient compte que des alignements qui se trouvent dans cette bande. Cette approche a été initialement proposée par [Kay et Röscheisen, 1993].

3.4.2 Deuxième étape : modèle de traduction

Dans cette étape, [Moore, 2002] part de l'idée que, dans un bi-texte donné, 90% des patterns sont de type 1-1. Pour cela, il considère les patterns 1-1 qui ont une forte probabilité selon l'alignement précédent. Pour assurer la robustesse de sa méthode, cette probabilité est fixé par un seuil de 0.99%. Les alignements résultants sont utilisés pour l'entrainement du modèle de traduction. Le choix de n'utiliser que les patterns 1-1 va avoir un impact sur l'alignement final, car de l'ordre de 10% du corpus demeurera non aligné. Cette perte peut poser des problèmes si le bi-texte à aligner est de petite taille. Le modèle de traduction utilisé par [Moore, 2002] est une version modifiée du modèle IBM1.

3.4.3 Troisième étape : Alignement de phrase

Pour avoir le score final d'un alignement donné, [Moore, 2002] fait appel une autre fois à la programmation dynamique, mais cette fois-ci, en combinant le modèle IBM1 avec le modèle initial de longueur. Par exemple si on considère une phrase source s de longueur l, une phrase cible t de longueur m, et $P_{1-1}(l,m)$ la probabilité assignée par le modèle initial à un alignement 1-1. La combinaison se calcule alors comme suit:

$$P(s,t) \quad = \frac{P_{1-1}(l,m)}{(l+1)^m}\Big(\prod_{j=1}^{m}\sum_{i=0}^{l} tr(t_j|s_i)\Big)\Big(\prod_{i=1}^{l} f_\mu(S_i)\Big) \tag{3.8}$$

HunAlign

[Varga et al., 2005] implémentent une approche basée sur une combinaison de la longueur des phrases et la similarité lexicale (cette similarité se calcul à l'aide d'un dictionnaire). La première étape consiste à faire une traduction grossière de chaque mot source à l'aide d'un dictionnaire déjà existant. La deuxième étape est une fonction de similarité à base de longueur de phrases et les mots partagés entre la traduction résultante à partir du dictionnaire et la traduction originale du texte source. En général cette approche est similaire à celle proposée par BMA, la différence entre les deux est que HunAlign fait la traduction mot à mot à l'aide d'une ressource pré-calculée à la place d'entraîner un modèle IBM1. C'est pour cela que HunAlign est plus rapide que BMA. Dans le cas de l'absence d'un dictionnaire, un premier passage sur tout le texte est effectué sur la base des longueurs de phrases. Ce passage a pour but de créer un nouveau dictionnaire, qui va être utilisé pour faire l'alignement. Selon [Abdul-Rauf et al., 2012], bien que cette approche est optimisée pour être rapide, la consommation de mémoire est son point faible. En réalité, l'approche proposée par [Varga et al., 2005] ne peut pas gérer des corpus parallèles de plus de 20000 phrases. Ceux-ci doivent être divisées en plus petites parties.

Nous avons discuté dans ce chapitre, les principales méthodes d'alignement de phrases existantes dans le domaine de la traduction automatique. Il en existe de nombreuses variantes que nous ne décrivons pas ici.

Dans le chapitre suivant, nous présentons notre approche. Nous verrons qu'elle produit de bons résultats en comparaison avec les approches décrites précédemment.

CHAPTER 4

NOTRE APPROCHE

Les techniques à l'état de l'art (chapitre 3) produisent des résultats encourageants. Nous constatons cependant que, la performance de ces techniques se détériore notablement lors de l'alignement de grands corpus, tel qu'EuroParl. Les performances se détériorent également en présence d'un taux de bruit élevé (des segments de texte qui ne sont pas traduits), etc.
Dans ce chapitre, nous décrivons notre approche (que nous appelons **JAPA**) qui passe par trois étapes principales:

1) Le score de chaque paire.
2) La sélection des paires de phrases potentielles.
3) La sélection de l'alignement optimal selon la fonction de score.

4.1 La fonction de score

Plusieurs recherches ont été réalisées pour maximiser le score de l'alignement optimal. Chacune de ces recherches s'appuie sur des indices précis. Nous trouvons des approches qui explorent les *cognats*, d'autre la longueur des phrases, etc.
JAPA fait usage d'informations proposées dans des recherches précédentes, mais les intègre d'une manière pratique et efficace.

4.1.1 Aucune information linguistique

La comparaison de longueur de phrases a été l'un des premiers indices proposés. À cet égard, deux modèles ont été décrits: le premier [Brown et al., 1991], tient compte de la longueur des segments (comptée en mots), le second [Gale et Church, 1991], tient compte de la longueur des segments (comptée en caractère)(voir section 3.2.1).

Il est important de noter que ce modèle simple donne de bons résultats pour des corpus peu ou pas bruités. Sinon, les résultats d'alignement seront erronés depuis le point de bruit jusqu'à la fin. La figure 4.1 nous montre le comportement de l'alignement sur un corpus propre[1], tandis que la figure 4.2 nous montre ce qui peut se produire avec ce modèle lorsqu'un corpus bruité est aligné. Nous constatons que tous les alignements produits après le passage bruité[2] sont faux.

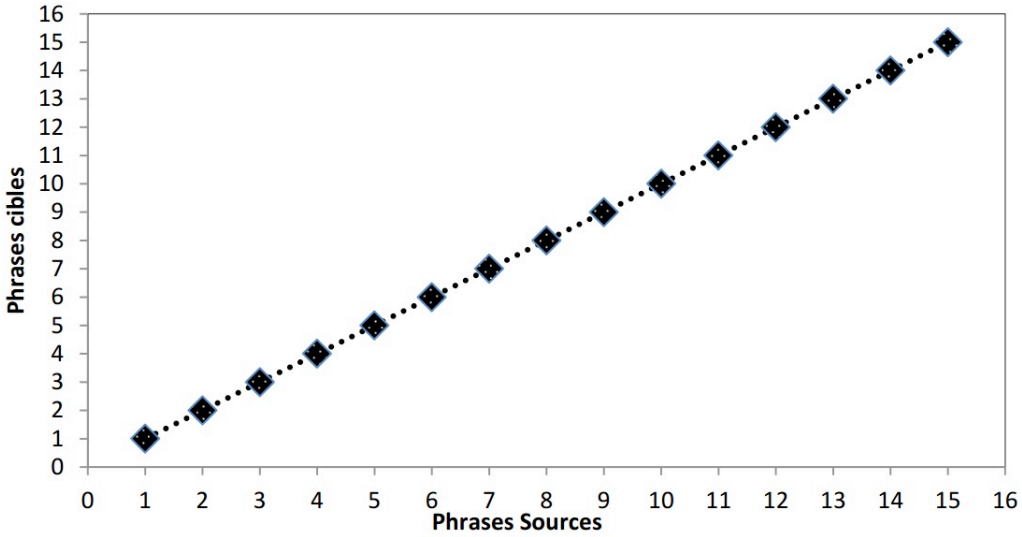


Figure 4.1: Exemple d'alignement d'un corpus propre faite par le modèle de Gale & Church (G&C).

4.1.2 L'information partagée entre langues

Quand un humain est confronté avec le problème de l'alignement d'un corpus parallèle donné, l'idée intuitive est d'utiliser des points d'ancrage[3] (même s'il ne maitrise pas parfaitement les langues en considération). Il est bien connu que, pour des raisons his-

[1]Nous voulons dire par propre que la i-ème ligne du texte source est alignée avec la i-ème ligne correspondante du texte cible exemple d'un segment de 5 phrases par langue qui seront alignés comme suit: 1-1, 2-2, 3-3, 5-5.

[2]il s'agit dans notre exemple des phrases 8,9 et 10 qui ne sont pas traduites.

[3]Les points d'ancrage à ce stade peuvent être des titres, des mots qui se ressemblent dans les deux langues, des dates, etc.

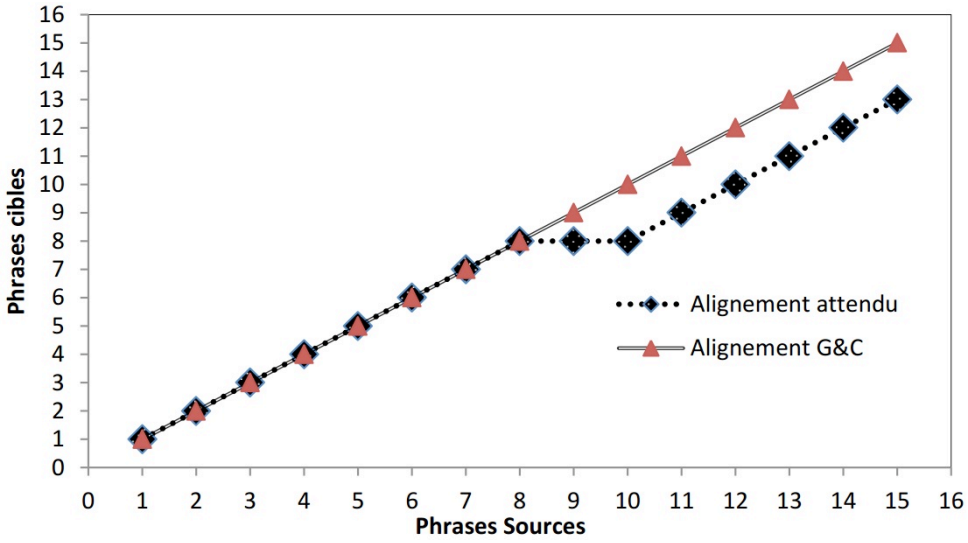


Figure 4.2: Exemple d'alignement d'un corpus bruité avec le modèle de G&C.

toriques, de nombreuses langues partagent beaucoup de mots, ou au moins des lemmes[4]. Pour remédier à ce problème automatiquement, nous explorons les *cognats* tels que définis par [Simard et al., 1992].

4.1.2.1 Lexique bilingue

Plusieurs méthodes utilisent des dictionnaires afin d'améliorer la qualité de l'alignement produit [Ma, 2006, Varga et al., 2005]. Cette information additionnelle peut être nuisible et réduire plutôt qu'améliorer la qualité d'alignement. Notre approche n'utilise aucune information extérieure pour aligner les textes.

4.1.2.2 Cognats

Les *cognats* peuvent être des entités qui ne sont pas modifiables tout au long du processus de traduction. Les noms propres, les données numériques sont les meilleurs

[4]Le lemme (ou lexie, ou item lexical) est l'unité autonome constituante du lexique d'une langue. C'est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire. Dans le vocabulaire courant, on parlera plus souvent de mot. Pour les langues qui partagent des lemmes, ceci est particulièrement vrai si les langues considérées sont des langues indo-européennes. Un exemple simple de lemme, manger est le lemme de mangera.

exemples dans ce cas. En partant de ce fait, [Simard et al., 1992] proposent quelques règles simples afin d'automatiser le processus d'alignement en utilisant les *cognats*. Voir la section 3.3.2.

4.1.3 Le score final

La fonction de score de **JAPA** (que nous appelons S_{japa}) utilise les modèles susmentionnées. Nous avons développé une extension de la fonction de score de [Simard et al., 1992] et de [Gale et Church, 1991]. Cette fonction est représentée comme suit:

$$S_{japa} = -log\left(\frac{P_T(c|n)}{P_R(c|n)} \times p(\sigma|pattern) \times p(pattern)\right) \tag{4.1}$$

Une fois cette fonction développée, elle est représentée par les trois quantités (x, y et z) suivantes:

$$\begin{aligned} S_{japa} = & -\left[c.log\frac{P_T}{P_R}\right] - \left[(n-c).log\frac{1-P_T}{1-P_R}\right] \quad (x) \\ & -logp(\sigma|pattern) \quad (y) \\ & -logp(pattern) \quad (z) \end{aligned} \tag{4.2}$$

Les trois quantités (x, y et z) représentent respectivement, le score des *cognats*, le score de longueur et le score des patterns.

Nous pouvons observer dans le tableau 4.I que ces quantités n'ont pas la même dynamique. Cette observation nous a conduit à la conclusion que le poids de chacune de ces trois quantités n'est pas égal dans le processus d'alignement. À partir de ce fait, nous introduisons trois coefficients de pondération (α_x, α_y et α_z) choisis de manière à donner un poids significatif à chaque quantité.

Nous partons de l'hypothèse que les différents scores pondérés sont statistiquement indépendants. Ensuite, Nous utilisons la technique de minimisation non dérivative appelée Simplex [Nelder et Mead, 1965] afin de trouver une combinaison qui optimise les performances (à la fois précision et rappel que nous décrivons dans le chapitre 6) sur

Score	μ	σ	Min	Max
x	0.2	9.4	-104.1	36.2
y	13	16	0.1	69.07
z	2.8	1.7	0	4.5

Table 4.I: Les trois quantités utilisées dans la fonction de score S_{japa} avec la moyenne (μ) et l'écart-type (σ) sur un corpus de 1000 paires de phrases manuellement alignées.

un corpus de développement. Les coefficients suivants sont présentement considérés par **JAPA**: $\alpha_x = 0.5,\ \alpha_y = 0.2$ *et* $\alpha_z = 1$[5]. Le score d'un pattern donné par **JAPA** est donc décrit par:

$$\begin{aligned} S_{japa} &= -0.5 \times \left(c.log\frac{P_T}{P_R} - (n-c).log\frac{1-P_T}{1-P_R} \right) \\ &\quad - 0.2 \times logp(\sigma|pattern) \\ &\quad - logp(pattern) \end{aligned} \tag{4.3}$$

4.2 Sélection des paires candidates

La sélection des paires candidates, ainsi que la réduction de l'espace de recherche, sont deux points à ne pas négliger dans la réalisation d'un aligneur de phrases.

Pour la réduction de l'espace de recherche, notre idée de base est que l'alignement de phrases peut se faire correctement en descendant à un niveau de granularité plus bas en considérant les alignements au niveau des mots. Tel que mentionné par [Debili et Sammouda, 1992], nous sommes confrontés à un cercle vicieux que nous pouvons contourner en considérant que l'alignement au niveau des phrases peut utiliser un alignement grossier au niveau des mots. Cette solution a été envisagée aussi par [Simard et al., 1992]. Les auteurs proposent un algorithme qui détermine les points par où la solution doit passer. Cette solution est efficace dans la mesure où la localisation des points est précise, ce qui est un point délicat. Pour cela, nous décrivons une méthode de réduction de l'espace de recherche, que nous trouvons plus fiable, puisqu'elle n'impose pas de

[5]Ces coefficients sont bien sûr configurables

point d'ancrage.

Nous considérons le corpus bilingue comme une matrice binaire, dénotée M. La i-ème ligne représente le i-ème mot du texte source, et la j-ème colonne représente le j-ème mot du texte cible. L'élément $M(i, j)$ est mis à 1 si et seulement si, le mot source i est en relation avec le mot cible j (dans notre cas, les deux mots sont *cognats*). Cette représentation est courante en bio-informatique, et a été utilisée en alignement par [Gale et Church, 1991]. La figure 4.3 représente un exemple d'une telle matrice calculée à partir d'un roman *"De la terre à la lune"* de Jules Verne. Chaque point indique une relation de cognats entre un mot source et un mot cible.

Nous constatons dans la figure 4.3 que, chaque mot du texte source est en relation avec plusieurs mots tout au long du texte cible. Ainsi, la diagonale est très dense où il y a une forte concentration de points par rapport aux autres endroits du texte. Cette quantité d'information peut s'avérer inutile, voire nuire à la délimitation d'un espace de recherche pertinent. Pour éliminer les paires qui constituent un bruit au processus d'alignement, plusieurs approches ont été proposées. [Chang et Chen, 1997] proposent une technique de filtrage d'image. Dans **JAPA**, nous utilisons une technique plus simple qui s'avère en pratique efficace. Nous considérons seulement les mots de basse fréquence dans le processus d'alignement au niveau des mots. La figure 4.4 représente la matrice binaire obtenue à partir du même bi-texte que celui utilisé pour générer la figure 4.3, en tenant compte uniquement des cognats vus moins de 20 fois. Nous observons que la diagonale qui caractérise l'alignement devient plus visible, et est donc à priori plus facile à identifier.

Nous faisons appel à la programmation dynamique pour calculer le coût minimal des appariements au niveau des mots. Nous définissons un coût pour l'appariement comme:

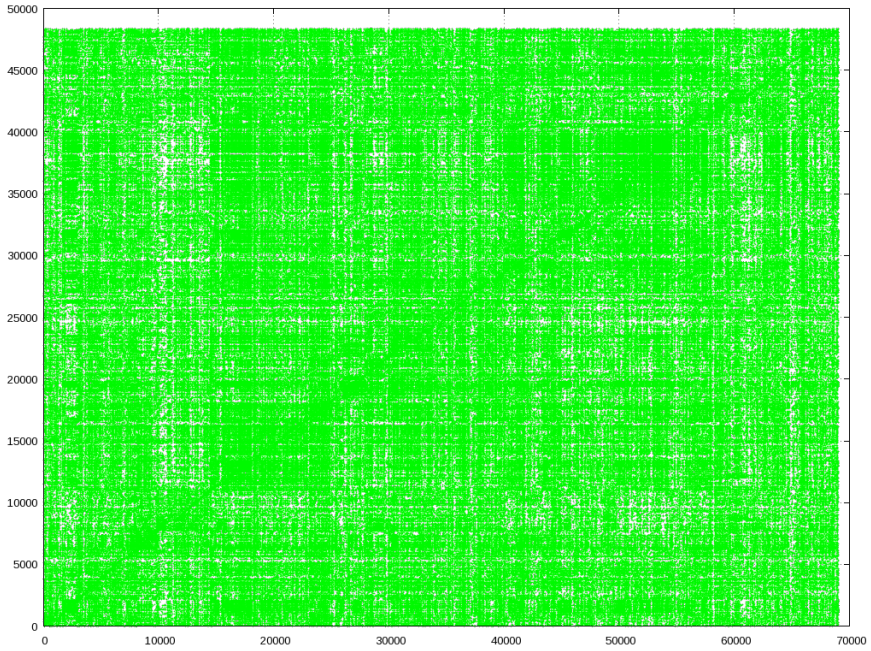


Figure 4.3: Exemple d'une matrice binaire calculée à partir de *"De la terre à la lune"* de Jules Verne. Un point indique une référence de cognats entre un mot source et un mot cible.

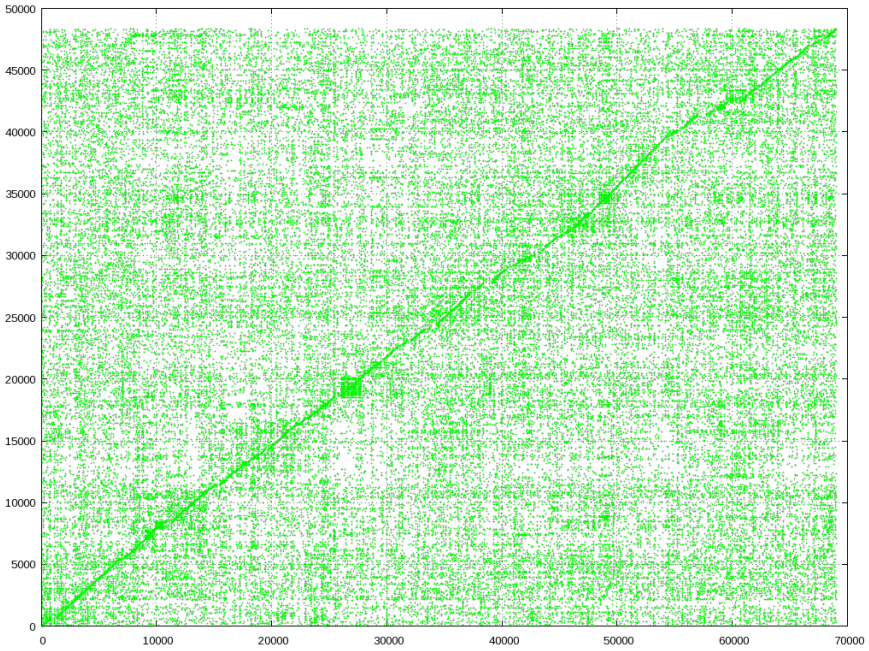


Figure 4.4: Exemple de la matrice binaire calculée à partir de *"De la terre à la lune"* de Jules Verne lorsque seuls les mots peu fréquents (<= 20) sont considérés pour le calcul de cognats.

$$\begin{aligned} S(I,J) &= \min_{i=I-R}^{\text{I-1}} \min_{j/M_{i,j}=1} (S(i,j)+F(i,j,I,J)) \\ F(i,j,I,J) &= \frac{J-1}{I-1} + (I-i-1)\times C \end{aligned} \tag{4.4}$$

où R est une constante qui définit la tolérance maximale pour la discontinuité dans l'alignement et C est une constante qui pénalise l'écart à la diagonale "naturelle". Ces constantes sont fixées empiriquement à partir d'un corpus d'entrainement[6].

4.3 Réduction de l'espace de recherche et alignement au niveau des phrases

Pour rendre **JAPA** plus rapide, nous réduisons l'espace de recherche au niveau des phrases. Pour cela, nous ne prenons en compte que les phrases qui se trouvent à l'intérieur d'un faisceau d'une taille fixe[7] centré autour de l'alignement global de mots précédemment produit. La figure 4.5 représente un extrait de la matrice calculée à partir de *"De la terre à la lune"* de Jules Verne, ainsi, le faisceau calculé à partir de l'alignement global de mots précédemment calculé.

Nous observons dans la figure 4.5 qu'une séquence de phrases (entre 2150 et 2200) n'est pas traduite dans la langue cible. Restreindre l'alignement de phrases à l'intérieur du faisceau facilite la reconnaissance d'alignement 1-0 que l'approche simple de [Gale et Church, 1991] ignore.

À la fin, nous réutilisons la programmation dynamique pour aligner les phrases sources $(S_1,...,S_I)$ avec les phrases cibles $(t_1,...,t_J)$. Pour cela, nous utilisons une implémentation similaire à celle proposée par [Gale et Church, 1991]. Cependant, seules les paires appartenant au faisceau sont considérées dans cette phase.

[6] Nous utilisons les valeurs $C = 5$ et $R = 50$

[7] Dans **JAPA** nous fixons la largeur de ce faisceau à 20 phrases.

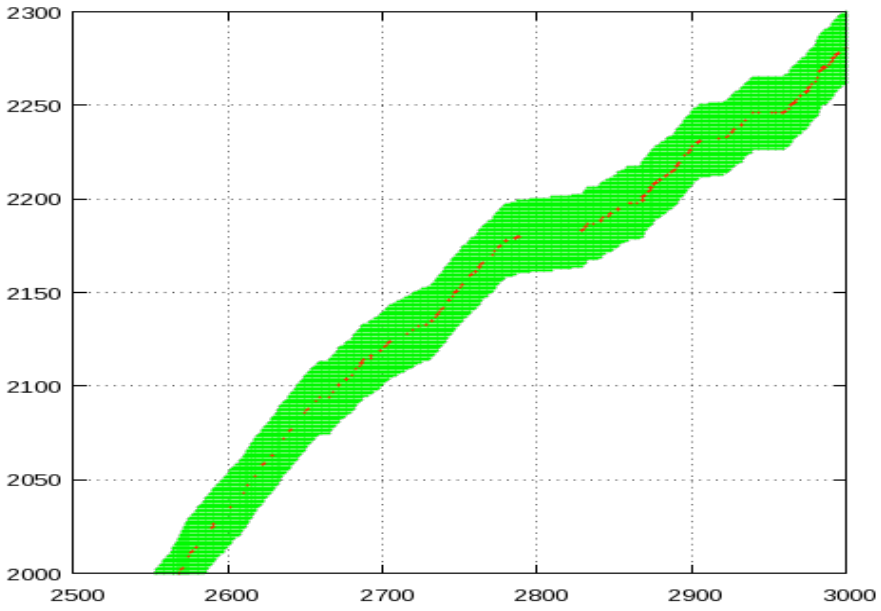


Figure 4.5: Exemple d'un extrait de matrice binaire calculée à partir de *"De la terre à la lune"* de Jules Verne avec le faisceau calculé à partir de l'alignement de mot. Les deux axes indiquent maintenant les indices de phrases dans le bi-texte.

CHAPTER 5

EXPÉRIENCES

Ce chapitre comprend plusieurs expériences menées sur différents types de données. Toutes ces expériences ont pour but la validation de la qualité d'alignement proposée par notre approche. Pour chaque expérience, nous définissons le corpus utilisé, les étapes suivies et les résultats obtenus, ainsi qu'une discussion de ces résultats. Nous classons nos expériences sous deux grandes catégories selon le corpus utilisé.

5.1 BAF

Ce corpus est décrit en détail dans le chapitre 2. Dans les expériences qui suivent, nous évaluons la qualité d'un alignement candidat par les mesures précision, rappel et F-mesure définis dans ce qui suit.

5.1.1 Évaluation

Pour mieux comprendre les mesures de rappel et précision, nous reprenons une définition formelle proposée par [Isabelle et Simard, 1996].

Soit un bi-texte dont le texte source $S = s_1, s_2, ..., s_n$ et le texte cible $T = t_1, t_2, ..., t_m$ correspond à sa traduction. Un alignement entre S et T se définit comme un sous-ensemble du produit cartésien $p(S) \times (T)$ où $p(S)$ et $p(T)$ sont respectivement l'ensemble de tous les sous-ensembles de S et T. Le triplet (S, T, A) sera appelé un bi-texte. Le rappel d'un alignement par rapport à la référence A_r est défini par:

$$Recall = |A \cap A_r| / |A_r| \tag{5.1}$$

Cela représente la proportion de bi-segments corrects dans l'alignement A par rapport à l'alignement de référence A_r. Ainsi, la précision d'un alignement par rapport à la référence A_r est défini par:

$$Precision = |A \cap A_r| / |A| \tag{5.2}$$

Nous utilisons également la F-mesure qui combine le rappel et la précision [van Rijsbergen, 1986]:

$$F = 2 \times \frac{(recall \times precision)}{(recall + precision)} \tag{5.3}$$

Le tableau 5.I représente un exemple de bi-texte que nous prendrons pour référence dans la suite. Cet alignement peut être présenté par $A_r = \left\{ \left(\{s_1\}, \{t_1\} \right), \left(\{s_2\}, \{t_2\}\{t_3\} \right) \right\}$.

Français	**Anglais**
s_1: Phrase numéro un.	t_1: The first sentence.
s_2: Phrase numéro deux qui ressemble à la première.	t_2: The second sentence. t_3: It looks like the first.

Table 5.I: Un exemple de bi-texte de référence avec l'alignement $\left(\{s_1 : t_1\}, \{s_2 : t_2, t_3\} \right)$ correspondant.

Le rappel et la précision peuvent être calculés à plusieurs niveaux de granularité: au niveau d'alignement, phrases ou à un niveau plus bas: mots ou caractères [Langlais et al., 1998]. À un niveau de granularité d'alignement F_A, un pattern obtient un point s'il correspond exactement à un pattern dans la référence. Dans l'exemple de la figure 5.1, l'alignement candidat identifie correctement deux des trois patterns de la référence $\left(\{s_3 : t_3, t_4\} et \{s_5 : t_5\} \right)$ alors sa précision P_A est 2/4 (et son rappel R_A) est donc 2/3. Considérer le niveau de granularité de la phrase F_S, permet de donner un crédit à des patterns partiellement corrects. Ainsi dans notre exemple, l'association $\{s_1 : t_1\}$ a été correctement identifiée par l'alignement candidat.

Le calcul au niveau de granularité phrase considère que dans un alignement n-m, les n phrases sources sont appariées avec les m phrases sources. Ainsi, n×m paires de phrases sont légitimes, et chacune d'elles identifiée par l'alignement candidat reçoit un crédit pour cela. Dans notre exemple, l'alignement candidat reçoit une précision P_S de 2/2 et un rappel R_S de 2/6. On notera que les taux de F-mesure au niveau des phrases

sont habituellement supérieurs à ceux mesurés au niveau des alignements.

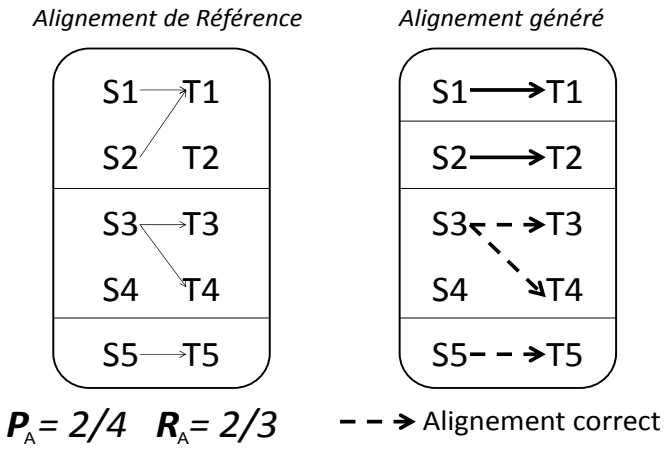


Figure 5.1: Exemple de calcul de précision et rappel au niveau de l'alignement.

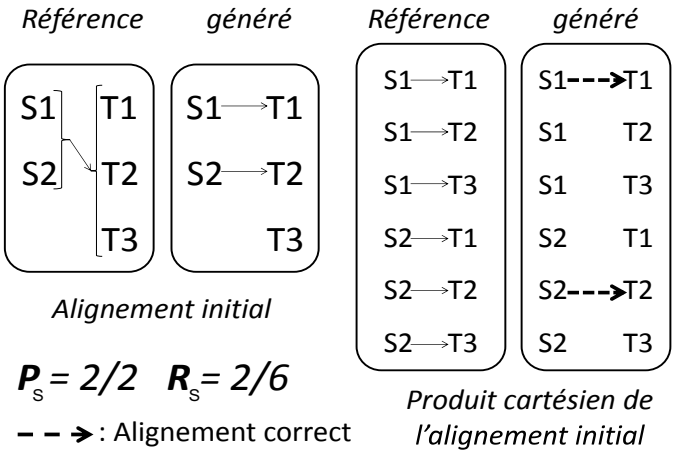


Figure 5.2: Exemple de calcul de précision et rappel au niveau des phrases

5.1.2 Performance sur BAF

Nous comparons notre méthode avec deux autres aligneurs populaires, BMA[1] et HunAlign[2] (deux parmi les meilleures méthodes d'alignement de phrases disponibles) sur les différents bi-textes du corpus BAF. Le tableau 5.II montrent les résultats obtenus à partir de chaque méthode.

	JAPA		BMA		HUNALIGN	
	F_A	F_S	F_A	F_S	F_A	F_S
INST	**94.9**	**96.4**	93.3	91.4	91.0	93.4
SCIENCE	**89.4**	**91.7**	88.9	86.4	84.8	86.5
VERNE	69.2	**86.9**	**72.3**	74.6	66.0	74.6
TECH	90.4	**10.9**	**94.2**	10.3	89.6	10.7
TOTAL	**89.6**	**85.6**	89.5	80.2	86.0	81.0

Table 5.II: F-mesure au niveau de l'alignement et au niveau des phrases des méthodes en fonction du type de texte du corpus BAF.

Nous observons qu'il y a une convergence entre les résultats obtenus par notre méthode et celle de BMA au niveau de l'alignement. Ainsi les deux méthodes ont une légère avance par rapport à HunAlign. Par contre, si nous comparons les résultats de F-mesure au niveau des phrases, notre méthode surpasse les deux autres pour tous les types de textes. Nous constatons également qu'au niveau des phrases, sur le corpus TECH, les trois méthodes ont produit beaucoup d'alignement incorrects. Une simple explication est que le bi-texte se termine par un index trié par ordre alphabétique. Ce passage n'est aligné au niveau de phrases dans la référence, ce qui affecte considérablement le rappel.

5.1.3 Bruitage de données

Les recherches menées dans le domaine de la traduction automatique montrent qu'il y a toujours un rapport entre la masse de textes multilingues traitée, et la qualité de traduction correspondante. Le bruit dans les données à aligner, ainsi que différents phénomènes apparaissent dans les bi-textes, tel que les passages non traduits qui sont

[1]BMA : se sont les initiales de Bob Moore Aligner. Cette aligneur est disponible sur `http://research.microsoft.com/apps/pubs/default.aspx?id=68886`

[2]`http://mokk.bme.hu/en/resources/hunalign/`

susceptibles de conduire à des alignements erronés. En partant de ce fait, l'utilisation d'un bon aligneur devient une étape importante pour avoir de bons résultats lors de la traduction.

Plusieurs études ont été effectuées sur le bruitage de données. Nous distinguons les recherches qui visent à filtrer et nettoyer les corpus bruités [Nie et Chen, 2002], d'autres qui s'intéressent à les aligner [Fung et Mckeown, 1994]. Nous tentons avec cette expérience de mesurer la robustesse de notre méthode d'alignement vis-à-vis du bruit. Nous présentons ici les principales caractéristiques de cette expérience qui se concentre sur le bruitage. Nous présentons également le corpus que nous choisissons pour tester la robustesse de notre aligneur, et ce même en présence de beaucoup de bruit. Ensuite, nous comparons nos résultats qui se rapprochent de ceux des meilleurs travaux effectués jusqu'à présent sur l'alignement automatique, et qui les dépassent lors de taux de bruit très élevés. Ici, nous prenons comme point de comparaison les résultats produits par la méthode de BMA décrite en détail dans le chapitre 3.

Pour cette expérimentation, et dans le but de bien étudier l'alignement sur des corpus bruités, nous avons besoin d'un corpus parallèle assez propre, pour que nous puissions contrôler le taux de bruitage. D'après les études empiriques faites sur les grands corpus disponibles sur le web [Goutte et al., 2012], la plupart des corpus disponibles contiennent un taux de bruit non négligeable. Ces auteurs estiment par exemple que les hansards contiennent de l'ordre de 0.5% de bruit, alors que le Giga corpus est bruité à 13%. Un tel taux de bruit ne permet pas de contrôler avec précision le bruit artificiel que nous souhaitons ajouter pour mettre à l'épreuve notre algorithme.

Nous choisissons donc pour réaliser cette expérience le corpus BAF[3] (version 1.1). Ce corpus ne se limite pas à un seul genre de textes, ce qui va enrichir notre expérimentation. Il est de plus aligné entièrement à la main au niveau des phrases, ce qui veut dire

[3]BAF sont les acronymes de Bi-texte Anglais Français. La version 1.1 de ce corpus est disponible sur le lien suivant : `http://www-rali.iro.umontreal.ca`

que nous avons la référence des alignements faite manuellement. Ceci nous permet de contrôler précisément le bruit introduit et son impact sur l'alignement.

5.1.3.1 Protocoles de bruitage

Dans cette partie, nous décrivons les politiques existantes pour introduire du bruit dans un corpus donné. Ainsi, nous décrivons notre politique lancée sur le corpus BAF.

Nous notons que les politiques d'introduction du bruit diffèrent d'une étude à une autre. Par exemple, dans [Goutte et al., 2012], les auteurs introduisent du bruit en décalant les phrases dans le texte source (dans leur cas: le texte anglais). Pour 10% du bruit, 10% du texte est sélectionné aléatoirement. Un décalage se fait sur les paires de phrases de cette sélection, de tel sorte que, la première phrase source de la sélection aléatoire soit alignée avec la deuxième phrase cible de la sélection, la deuxième phrase source avec la troisième phrase cible, et ainsi de suite jusqu'à la fin. Pour les autres niveaux de bruit (20%, 30%, etc.), les niveaux de bruit précédents sont inclus exemple, dans le niveau de 20%, le niveau de bruit de 10% est inclus.

Après une implémentation de cette technique, nous trouvons qu'elle est moins perturbante. Le tableau 5.III représente les résultats obtenus en lançant notre méthode sur les différents niveaux de bruit. Nous constatons dans ce tableau que ce type de bruitage n'affecte pas vraiment la qualité d'alignement produite.

	10%		20%		30%	
	Notre bruit	**Cyril**	**Notre bruit**	**Cyril**	**Notre bruit**	**Cyril**
INST	**85.9**	94.9	**81.7**	94.7	**77.8**	94.6
SCIENCE	**78.8**	89.1	**75.2**	87.2	**70.5**	86.9
VERNE	**56.1**	69.3	**54.5**	69.2	**36.7**	69.1
TECH	**72.2**	90.4	**74.7**	90.3	**69.1**	90.2

Table 5.III: Comparaison des scores F-mesure calculés au niveau d'alignement après bruitage selon notre méthode et selon la méthode de [Goutte et al., 2012].

Dans notre étude, l'introduction du bruit consiste à éliminer aléatoirement des lignes

entières dans le texte source (dans notre cas: le texte français). Nous introduisons du bruit aux différents textes par incrément de 10% pour en arriver au total à 6 niveaux allant de 10% jusqu'à 60%. Par exemple, si nous prenons un texte de 1000 phrases par langue avec une phrase par ligne, dans le premier niveau de 10%, nous supprimons aléatoirement une phrase chaque 10 phrases coté source. Nous aurons comme résultat final: 900 phrases dans le texte source avec 1000 phrases dans le texte cible. Le tableau 5.IV illustre l'opération effectuée.

Ligne	**Français**	**Anglais**	**Ancienne réf**	**Nouvelle réf**
1	1ère ligne	1st line	**1-1**	**x-1**
2	2ème ligne	2nd line	**2-2**	**1-2**
14	14ème ligne	14th line	**14-14**	**13-14**
15	15ème ligne	15th line	**15-15**	**x-15**
16	16ème ligne	16th line	**16-16**	**14-16**
1000	1000ème ligne	1000th line	**1000-1000**	**900-1000**

Table 5.IV: Niveau de 10% du bruit avec la référence correspondante.

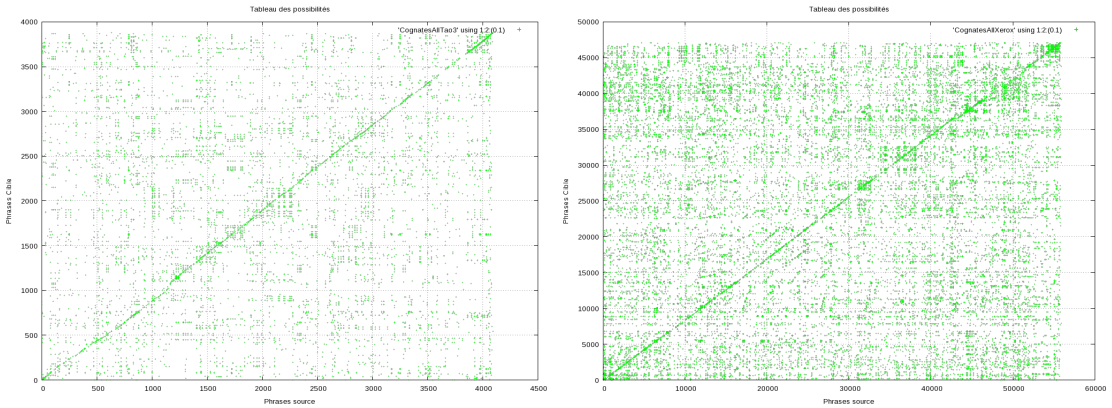
Si nous reprenons l'exemple précédent mais cette fois ci avec un taux de bruitage qui atteint les 60%, nous supprimons aléatoirement six phrases chaque dix phrases coté source. Nous aurons comme résultat final: 400 phrases dans le texte source avec 1000 phrases dans le texte cible. Ceci est illustré en tableau 5.V.

Ligne	**Français**	**Anglais**	**Ancienne réf**	**Nouvelle réf**
1	1ère-6ème ligne	1st line	**1-1**	**x-{1-6}**
7	7ème ligne	7th line	**2-2**	**1-7**
14	14ème ligne	14th line	**14-14**	**8-14**
15	15ème ligne	15th line	**15-15**	**x-15**
1000	1000ème ligne	1000th line	**1000-1000**	**400-1000**

Table 5.V: Texte de 1000 phrases avec un taux de bruit de 60% et la référence correspondante.

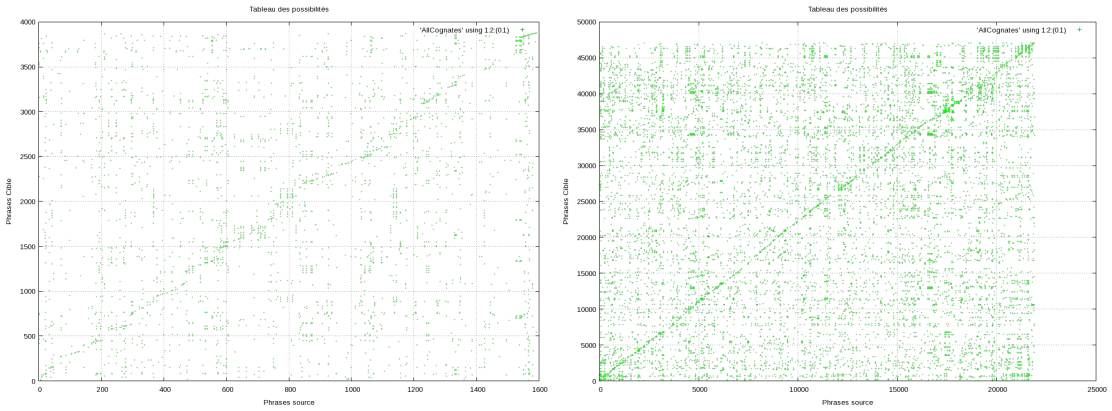
Les figures 5.3 et 5.4 montrent des Dotplots de deux bi-textes de BAF, sans et avec

bruitage à 60%.



(a) Tao3 (b) Xerox

Figure 5.3: Représentation graphique de la dispersion des *cognats* dans deux textes propres (sans bruit) de BAF. L'axe des abscisses représente la liste des mots sources (français), l'axe des ordonnées représente la liste des mots cibles (anglais). Chaque point correspond à deux mots (source,cible) qui sont *cognats*.



(a) Tao3 (b) Xerox

Figure 5.4: Représentation graphique de la dispersion des *cognats* dans les deux textes de BAF de la figure 5.3. Les textes sont bruités à 60%. L'axe des abscisses représente la liste des mots sources (français), l'axe des ordonnées représente la liste des mots cibles (anglais).

Dans la version propre (figure 5.3), nous constatons que les *cognats* forment une di-

agonale bien claire, ce qui veut dire que les meilleurs candidats pour un alignement optimal se trouvent autour de cette diagonale. Dans la figure 5.4, nous constatons qu'avec 60% de bruit, la concentration des *cognats* dans les deux textes diminue notablement. Dès lors, des passages qui ne contiennent pas de *cognats* apparaissent. Nous constatons également que, la concentration des *cognats* autour de la diagonale est encore présente, bien que 60% de bruit est ajouté. Ceci est due à la présentation des *cognats* par Dotplot (les échelles des bi-textes propres et des bi-textes bruités ne sont pas les mêmes). Ces constatations nous mènent à conclure que les textes ont été correctement bruités par notre politique.

5.1.3.2 Résultat

Nous présentons dans le tableau 5.VI les scores de F-mesure obtenus par notre méthode et celle de BMA[4] au niveau de l'alignement F_A.

%	INST	SCIENCE	VERNE	TECH
0	94.9(+1.6)	89.4(+0.5)	69.2(-3.1)	90.4(-3.8)
10	85.9(+1.1)	78.8(-1.9)	56.1(-2.5)	72.2(-7.1)
20	81.7(+2.2)	75.2(+2.9)	54.5(-3.8)	74.7(-6.5)
30	77.8(+7.0)	70.5(+8.3)	36.7(-7.6)	69.1(-3.8)
40	76.0(+16.6)	69.1(+30.5)	40.6(+3.3)	65.0(+1.8)
50	69.0(+23.0)	67.2(+41.8)	44.7(+28.8)	64.3(+14.9)
60	69.5(+44.8)	67.8(†)	49.9(+41.9)	65.8(†)

Table 5.VI: Résultat F-mesure de notre méthode en fonction du bruit. Les chiffres entre parenthèses sont les gains absolus de notre méthode envers BMA. (Une valeur négative indique que BMA surpasse notre méthode). BMA plante parfois, ceci est marqué par un symbole †.

Nous observons que la qualité d'alignement des deux méthodes se dégrade avec les différents niveaux de bruit. Avec les textes institutionnels (INST), la qualité de l'alignement produit par notre méthode diminue de 94.9 jusqu'à 69.5. Par contre, la qualité d'alignement de BMA chute d'une manière remarquable de 93.3 à 24.7. Les mêmes

[4] À ce niveau, la comparaison se fait seulement entre notre méthode et celle de BMA. La méthode HunAlign n'est pas prise en compte, parce que nous observons dans le tableau 5.II que BMA est plus meilleur que HunAlign. Cependant, nous nous concentrons seulement sur BMA.

tendances sont observées pour les autres types avec de légères différences. Cependant, en particulier pour VERNE, la qualité d'alignement produit par BMA est meilleure avec les niveaux de bruit faibles. Une représentation graphique de ces résultats est fournie avec les figures 5.5, 5.6, 5.7 et 5.8. Cette représentation a pour but, mieux observer la stabilité de la qualité d'alignement produit par notre aligneur en fonction du bruit.

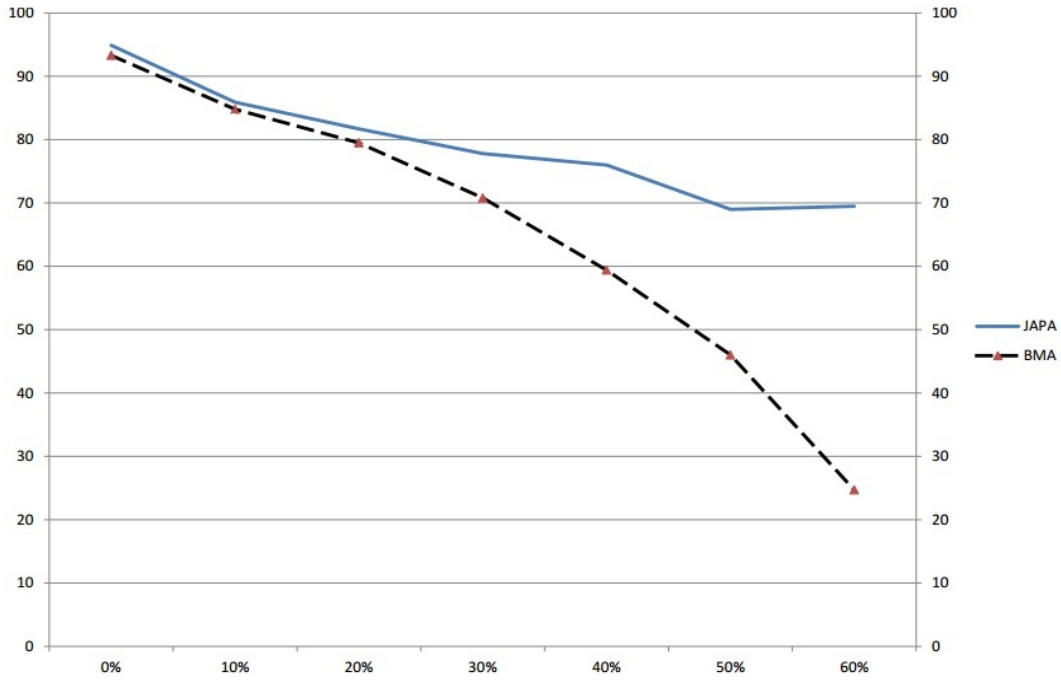


Figure 5.5: Comportement de l'alignement en fonction du bruit de notre approche contre celle de BMA sur le type INST de BAF.

Nous concluons que notre approche est plus stable que l'approche de BMA avec les taux de bruit très élevés. Nous constatons que l'approche de BMA plante parfois, ce n'est pas le cas avec notre approche.

5.2 Temps d'exécution

Un bon aligneur doit être capable de produire un alignement de qualité. Une différence notable (autre que la qualité d'alignement) entre notre méthode et celle de BMA est le temps nécessaire pour faire l'alignement. Sur le corpus BAF, nous constatons que

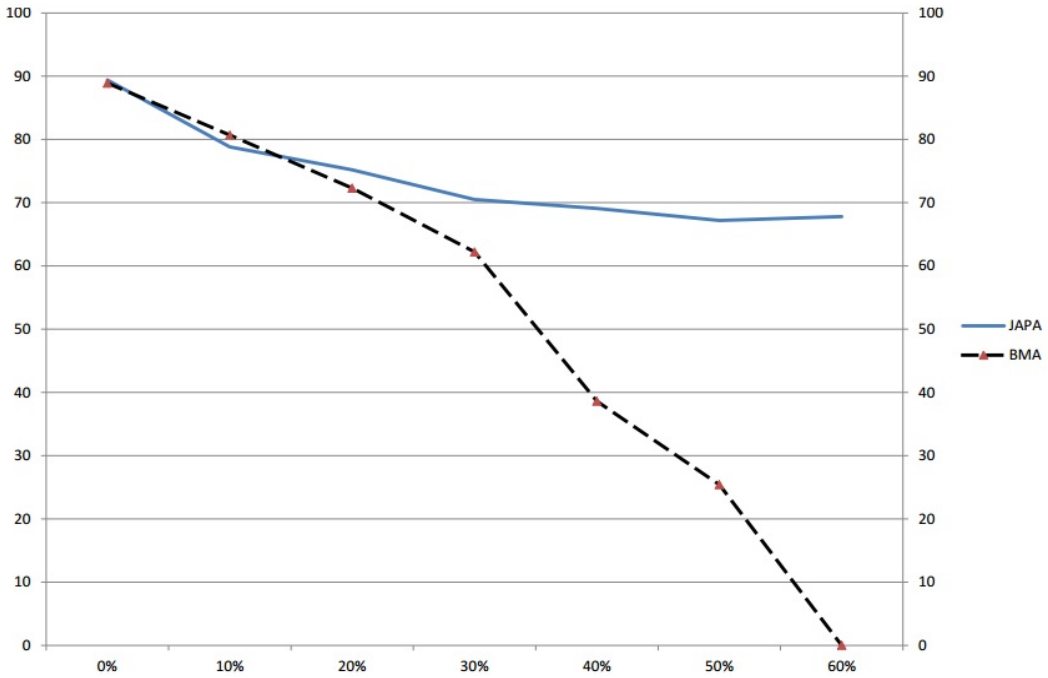


Figure 5.6: Comportement de l'alignement en fonction du bruit de notre approche contre celle de BMA sur le type SCIENCE de BAF.

notre méthode prend 27 secondes en moyenne pour aligner l'un des bi-textes, tandis que, la méthode de BMA prend en moyenne 257 secondes. Par conséquent, notre méthode est un ordre de grandeur plus rapide que BMA. La différence de vitesse augmente avec la taille des textes parallèles à aligner, ainsi que le niveau de bruit qu'ils contiennent. Pour illustrer cela, nous lançons les deux méthodes (la notre, celle de BMA) sur des passages du corpus EuroParl en faisant varier la taille de 1000 jusqu'à 1 million de paires de phrases. La vitesse des deux méthodes mesurée en secondes[5] est rapportée dans la figure 5.9. Nous observons que notre méthode est plus rapide de celle de BMA, nous constatons également que la différence de vitesse augmente considérablement avec les longs bi-textes. En particulier, pour le passage de 1 million de phrases, notre méthode prend moins de 24 minutes, tandis que BMA prend plus de 30 heures.

[5]Les deux méthodes ont été lancée sur la même machine.

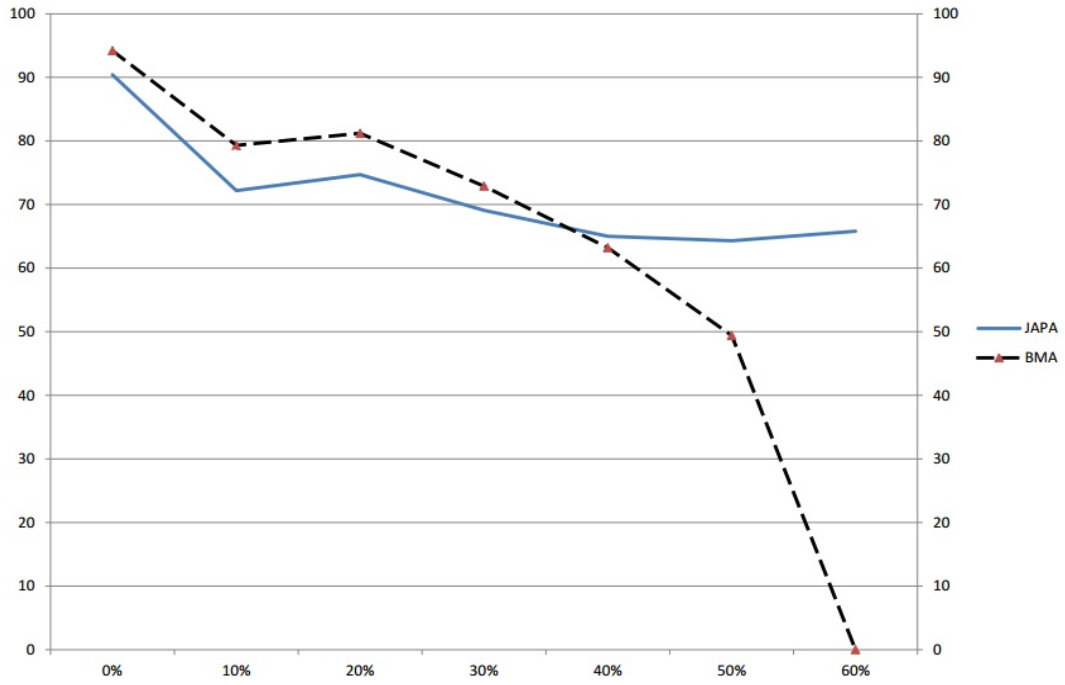


Figure 5.7: Comportement de l'alignement en fonction du bruit de notre approche contre celle de BMA sur le type TECH de BAF.

5.3 EuroParl

5.3.1 Validation de notre méthode

Nous menons cette étude dans le but de valider notre méthode sur de grands corpus. Nous présentons ici les principales caractéristiques de cette expérience, ainsi que le corpus que nous avons choisi pour tester la robustesse de notre méthode et sa qualité d'alignement au niveau des phrases. Pour cela, nous nous intéressons aux limites d'alignement rencontrées avec la méthode de P.K.[6] décrite en section 5.3.2. Nous présentons également les gains obtenus par notre méthode, en comparant nos résultats avec ceux obtenus par la méthode de P.K.[7]

Notre expérience est lancée sur le corpus Europarl version 7. Nous choisissons

[6]P.K. sont les initiales de Philippe Koehn.

[7]Les résultats obtenus par la méthode de P.K. sont disponibles sur le lien suivant : `http://www.statmt.org/wmt12/training-monolingual-europarl.tgz`

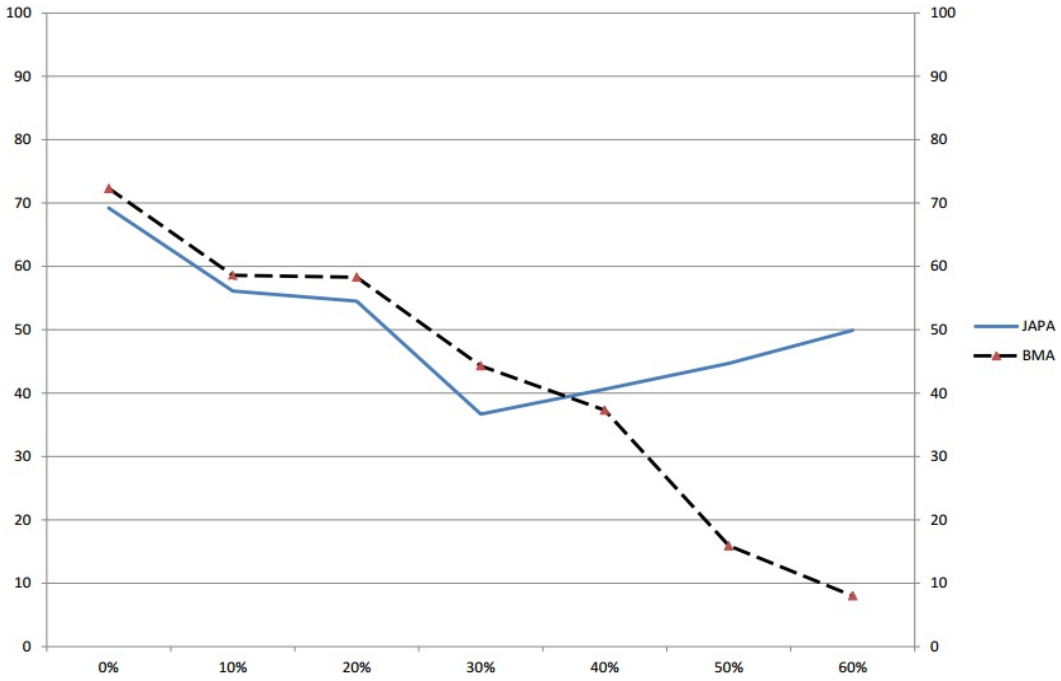


Figure 5.8: Comportement de l'alignement en fonction du bruit de notre approche contre celle de BMA sur le type VERNE de BAF.

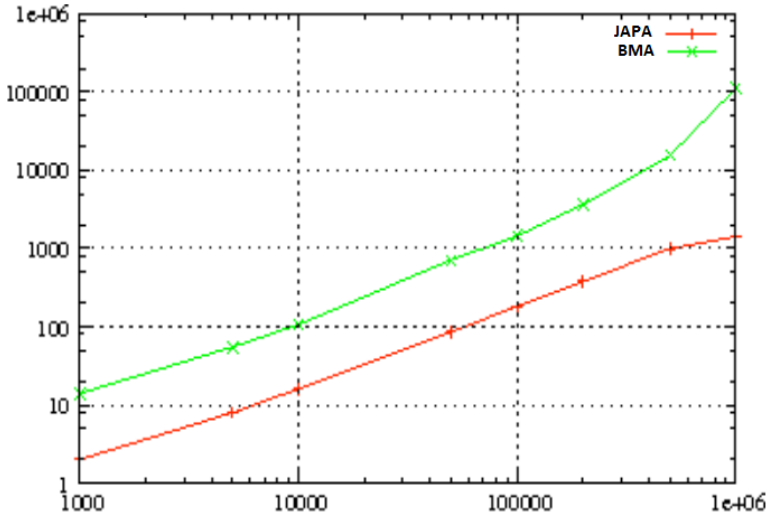


Figure 5.9: Comparaison des temps d'exécution en seconde de notre méthode par rapport à celle de BMA. L'axe des abscisses représente le nombre de phrases traitées, l'axe des ordonnées représente le temps d'exécution. L'échelle est logarithmique sur les deux axes.

délibérément ce corpus en raison de sa disponibilité et de sa popularité dans le domaine de la traduction. Le choix sélectif du corpus va nous permettre d'éprouver notre méthode sur des données représentatives.

Afin d'étudier la robustesse de notre méthode et trouver les défaillances d'alignements survenues sur EuroParl par la méthode de P.K., nous alignons avec notre méthode les textes français et anglais du corpus Europarl. Notre programme ne sait bien sûr pas que le corpus à aligner est en fait déjà aligné. Nous constatons que notre sortie est différente de celle distribuée sur internet. Nous nous intéressons seulement à la différence entre les deux sorties, en admettant que tous les alignements similaires entre les deux sorties sont corrects. Parmi les 2M de lignes du corpus, nous avons pu extraire 7151 alignements différents entre les deux approches. Nous constatons que ces différences sont dispersées tout au long du corpus.

Nous avons réalisé une analyse manuelle d'un sous ensemble de 1700 alignements à partir des alignements différents (7151). Le tableau 5.VII présente les résultats obtenus. Par exemple pour la première ligne du tableau **0-1** indique que notre sous ensemble contient 90 cas de patterns de type **0-1**. Notre méthode a aligné correctement 74 des cas et a mal aligné les 16 cas restants.

Alignement	✓	✓ **%**	×	× **%**
0-1	74	82.22	16	17.78
1-0	128	81.53	29	18.47
1-1	323	41.25	460	58.75
2-1	247	86.06	40	13.94
1-2	290	83.57	57	16.43
2-2	24	66.67	12	33.33
Total	1086	63.88	614	36.12

Table 5.VII: Évaluation manuelle du sous-ensemble de 1700 alignements produits par notre approche et leurs distributions. Les alignements corrects sont présentés par ✓ et les alignements incorrects sont présentés par ×.

En considérant tous les patterns, notre méthode est correcte dans 64% des cas. Tan-

dis que, pour les patterns 1-1, l'alignement produit par P.K. est plus précis (58.8% contre 41.2%). Si nous nous concentrons uniquement sur les patterns $n-m$[8], notre méthode a un léger avantage avec 884 cas corrects, contre 569 pour la méthode de P.K. Les tableaux 5.VIII et 5.IX représentent un extrait des alignements incorrects produit par la méthode de P.K. où la ligne *i* dans le texte source est associée à la ligne *i* du texte cible, alors que, notre méthode a pu détecté le bon alignement.

5.3.2 Impact sur les systèmes de traduction statistique

Les résultats de la dernière expérience menée sur le corpus EuroParl nous montrent qu'il y a plusieurs passages alignés incorrectement. En s'appuyant sur ces résultats, une question se pose quant aux limites et la qualité de l'alignement fourni par la chaine de traitement de P.K. que nous décrivons dans ce qui suit. Les expériences qui suivent ont pour objectif de vérifier si nous pouvons mieux aligner le corpus EuroParl. Une fois réaligné, le corpus EuroParl va nous servir comme entrée d'un système de traduction statistique que nous décrivons dans la section 5.3.2.2

Chaine de traitement de P.K. : EuroParl est le fruit d'un traitement fait par [Koehn, 2005] sur les documents source de EuroParl. La préparation de ce corpus passe par plusieurs étapes ordonnées comme suit :

1. Obtention des données brutes à partir du web.
2. Extraction des documents parallèles (alignement de documents).
3. Ajustement des textes (faire la tokenisation et mettre les textes en une phrase par ligne).
4. Mettre les phrases d'une langue en correspondance avec les phrases d'une autre langue (alignement de phrases). L'alignement se fait comme suit:

[8]Nous ignorons les patterns $0-n$ et $n-0$, qui ne sont pas importants en pratique dans la traduction statistique.

Ligne FR	Français	Ligne EN	Anglais
169540	Quand le progrès mal maîtrisé acquiert sa propre dynamique, il génère la peur et le désarroi.	**169536**	When progress that is not properly controlled acquires its own dynamic, it generates fear and disarray.
169541	Troisième idéal : l'égalité. L'égalité est un formidable levier en Europe.	**169537**	Third ideal: equality. Equality is a hugely powerful lever in Europe.
169542	C'est en son nom que les pays européens tentent d'assurer les mêmes droits, la répartition équitable des fruits de l'activité économique et la représentation démocratique des intérêts de chacun.	**169538**	It is in the name of equality that the countries of Europe try to secure equal rights, an equal share in the fruits of economic activity and the democratic representation of the interests of every individual.
169543	En Europe occidentale, elle a abouti à l'adoption d'une législation sociale étendue qui nourrit le concept d'égalité de chances et de justice.	**169539**	In Western Europe, this has led to the adoption of far-reaching social legislation to guarantee equal opportunities and equality under the law.
169544	Dans l'Europe communiste, l'égalitarisme poussé à outrance a fini par sacrifier le droit à l'identité personnelle sur l'autel du collectivisme.	**169540**	In Communist Europe, the extreme version of egalitarianism ended up sacrificing the right to a personal identity on the altar of collectivism.
169545	Liberté, progrès, égalité ne sont pas seulement des valeurs théoriques.	**169541**	Freedom, progress and equality are not just theoretical values.

Table 5.VIII: Exemple 1 d'alignements incorrects produits par la méthode de P.K. où la ligne *i* du texte source est associée à la ligne *i* du texte cible, alors que notre méthode a pu identifié le décalage {169540:169536, 169541:169537, 169542:169538, 169543:169539, 169544:169540, 169545:169541}.

Chaque document source représente un débat journalier et est disponible dans différentes langues. Chaque document contient plusieurs chapitres de la forme (<CHAPTER id>) où 'id' c'est le numéro de chapitre. Chaque chapitre contient plusieurs tours de parole de la forme (<SPEAKER id name>) où 'id' indique le numéro de speaker et 'name' son nom. Chaque speaker contient plusieurs

Ligne FR	Français	Ligne EN	Anglais
169558	Si elle veut convaincre les pays tiers du bien-fondé de ses valeurs humanistes, l'Union européenne doit donner l'exemple avant de donner des leçons.	**169553**	If it wants to convince third countries that its humanist values are sound, the European Union must teach by example before teaching lessons.
169559	Les valeurs essentielles ne figurent pas seulement dans les textes, elles sont également mises en pratique dans les politiques de l'Union européenne.	**169554**	The essential values are not just written in the texts; they are also put into practice in the European Union' s policies.
453301	L'élection du président de la Commission par le Parlement européen.	**453300**	Another point of consensus is the election of the President of the Commission by the European Parliament.
453302	De vrais progrès dans la mise en place d'un espace de liberté, de sécurité et de justice.	**453301**	Genuine progress was made in the implementation of an area of freedom, security and justice.
453303	De vraies avancées - j'espère qu'on les préservera - dans le domaine de la défense.	**453302**	Real advances were made too in the field of defence. I hope they will be retained.

Table 5.IX: Exemple 2 d'alignements incorrects produits par la méthode de P.K. où la ligne i du texte source est associée à la ligne *i* du texte cible, alors que notre méthode a pu identifié ce décalage {169558:169553, 169559:169554}, {453301:45300, 453302:45301, 453303:45302}

paragraphes délimités par des balises de type <p>. Le tableau 5.X montre un extrait d'un document journalier où la présidente de la session prend la parole. L'alignement se fait par une implémentation de [Gale et Church, 1991] seulement sur les sections qui ont le même nombre de paragraphes. Nous représentons dans la figure 5.10 un exemple de la chaine de traitement de P.K.

Afin de bien analyser les sorties de la chaine de traitement de P.K., nous lançons notre expérience avec 5 configurations. Chaque configuration est définie comme suit:

1. Configuration CPK : Cette configuration représente les textes standards produits

Français	**Anglais**
<CHAPTER ID=1> <SPEAKER ID=1 NAME="La Présidente"> Reprise de la session <P> Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vœux en espérant que vous avez passé de bonnes vacances. <P>	<CHAPTER ID=1> <SPEAKER ID=1 NAME="President"> Resumption of the session <P> I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period. <P>

Table 5.X: Extrait du document journalier 17-01-2000.txt.

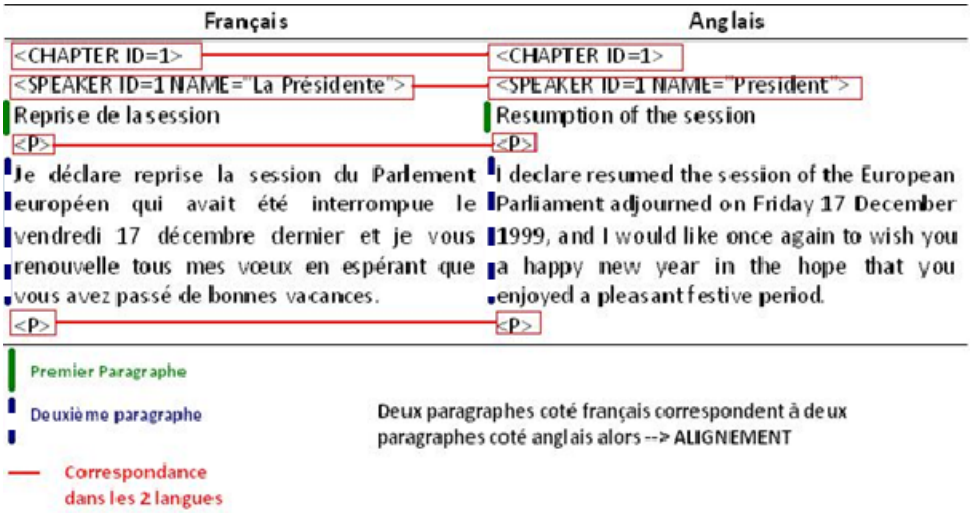


Figure 5.10: Exemple d'un alignement produit par la chaine de traitement de P.K.

par la chaine de traitement de P.K.

2. Configuration JAPA : Cette configuration est similaire à CPK à l'exception de l'aligneur utilisé qui est notre méthode au lieu de [Gale et Church, 1991]. Cette configuration se compare directement avec CPK, vu que les étapes sont les mêmes.

3. Configuration CPK++ : Le but de cette configuration est de voir la réaction de la

chaine de traitement à l'ajout des paragraphes précédemment ignorés et de tester si l'ajout de ce matériel a un impact en traduction. Cette configuration est une extension de la configuration CPK. P.K se limite à aligner les sections qui contiennent le même nombre de paragraphes dans les deux langues. Nous testons donc une variante où les paragraphes ignorés par P.K. sont également alignés.

4. Configuration CJAPA++ : Cette configuration est similaire à CPK++ à l'exception de l'aligneur utilisé qui est notre méthode au lieu de [Gale et Church, 1991]. Cette configuration se compare directement avec CPK++, vu que les étapes sont les mêmes.

5. Configuration CSOURCE : Aligner des longs textes est une opération coûteuse et plus difficile. Notamment, aligner de longs textes avec l'approche de [Gale et Church, 1991] augmente le nombre d'erreurs d'alignement. Notre méthode ne semblant pas soufrir de ce problème, nous testons une configuration dans laquelle nous alignons directement les fichiers journaliers plutôt que de passer par la chaine de traduction de P.K. Nous ne prenons pas en considération les indices fournis par la chaine de traitement (la correspondance entre les chapitres, speakers et paragraphes). Nous avons montré avec l'expérience de la section 5.3 que notre aligneur est à la fois rapide et précis, aussi nous souhaitons vérifier avec cette configuration si notre approche peut outrepasser la chaine de traitement de P.K.

5.3.2.1 Analyses et statistiques

Nous analysons au fur et à mesure et en détail les caractéristiques de chaque configuration.

Le tableau 5.XI nous donne une idée générale de ce qui est produit par les 5 configurations. Les six premières lignes représentent le type d'alignement (1-0,0-1,1-1,1-2,2-1 et 2-2) produit par chaque configuration. Comme nous pouvons le constater, dans la configuration CPK et CPK++, nous marquons les types d'alignement (1-1,1-2,2-1 et 2-2) par des '—'. Dans ces deux configurations, l'information sur les alignements produits est interne à la chaine de traitement de P.K. Les 3 lignes suivantes nous montrent si une

	CPK	JAPA	CPK++	JAPA++	CSOURCE
1-0	2,0K	9,3K	27,7K	22,7K	11,1K
0-1	2,9K	4,8K	11,1K	18,1K	7,159
1-1	—	2,7M	—	2,6M	2,7M
1-2	—	57,5K	—	62,7K	54,9K
2-1	—	63,5K	—	67,4K	61,5K
2-2	—	1,3K	—	1,4K	1,3K
<>-<>	0	619,6K	859,8K	643,2K	835,2K
<>-Texte	0	12	8,1K	2,5K	463K
Texte-<>	0	12	2,8K	2,8K	484K
NB-Lignes	2,0M	2,6M	3,0M	2,7M	2,9M
Propres	2,0M	2,0M	2,1M	2,1M	2,0M

Table 5.XI: Caractéristiques de la sortie produite par chaque configuration.

	CPK	JAPA	CPK++	JAPA++	CSOURCE
|VOC-FR|	101,2K	101,9K	103,4K	103,2K	101,9K
|VOC-EN|	91,8K	92,5K	93,8K	93,6K	92,5K
|TOKENS-FR|	60,4M	61,6M	63,6M	63,3M	61,6M
|TOKENS-EN|	50,1M	51,2M	52,9M	52,7M	51,2M

Table 5.XII: Caractéristiques de la sortie de chaque configuration après tokenisation.

balise est alignée avec un élément textuel (<>-Texte, Texte-<>) ou si elle est alignée avec une autre balise (<>-<>). Les deux dernières lignes représentent le nombre total de lignes (paires de phrases) obtenues après alignement. Ainsi que le nombre de lignes propres, c'est à dire le nombre de paires de phrases après élimination des alignements 1-0, 0-1, <>-Texte, Texte-<> et <>-<>.

Le tableau 5.XII est obtenu après la tokenisation des sorties propres. Les deux premières lignes de ce tableau représentent le vocabulaire[9] produit après la tokenisation, les deux dernières lignes représentent le nombre total de mots vus dans les sorties propres.

Nous constatons que la configuration JAPA produit plus d'alignement que CPK après la tokenisation des sorties propres. De plus, le vocabulaire obtenu par JAPA est plus riche

[9]Le vocabulaire veut dire le nombre de mots différents rencontrés.

que celui obtenu par CPK que ce soit en français ou en anglais. Pour les configurations CPK++ et CJAPA++, nous constatons que CPK++ a pu produire plus d'alignement que CJAPA++. De même, le vocabulaire obtenu par CPK++ est plus riche que celui obtenu par CJAPA++.

D'après ces analyses, nous pouvons émettre une hypothèse que JAPA est meilleur que CPK, et que, CPK++ est meilleur que CJAPA++. Pour valider cette hypothèse nous évaluons manuellement un document journalier que nous choisissons parmi les documents sources. Ce document contient des sections qui n'ont pas le même nombre de paragraphes. Pour cela nous prenons deux parties d'un document journalier « ep-06-07-05 ». La première partie contient les paragraphes du premier speaker jusqu'au 10-ème. La deuxième partie contient les paragraphes du 206-ème speaker jusqu'à la fin du document (343-ème speaker). Nous lançons les 5 configurations sur les deux parties pour évaluer leur comportement.

D'après l'évaluation manuelle des sorties de la première partie, nous trouvons que les 5 configurations ont bien aligné les paragraphes, c.à.d. qu'il n'y a pas d'erreur d'alignement. Pour la deuxième partie, vu que la chaine de traitement dans CPK et CJAPA va éliminer directement les sections qui n'ont pas un nombre de paragraphes identique, c'est inutile de les évaluer. Deux configurations parmi les 3 restantes ont échoué à aligner les paragraphes. La configuration CSOURCE quant à elle réussi à les aligner correctement. Dès lors, nous examinons la deuxième partie pour tirer plus d'information et pour comprendre comment les deux configurations (CPK++ et CJAPA++) ont réagi. Après une inspection de la deuxième partie, nous constatons qu'il y a une erreur au niveau des paragraphes liés aux speakers.

Le tableau 5.XIII nous montre le problème rencontré. Nous remarquons que les paragraphes du speaker id=206 sont décalés au speaker id=207, de ce fait, la chaine de traitement de P.K. va donner aux méthodes d'alignement des paragraphes décalés à chaque fois jusqu'à la fin du document. C'est pour cette raison que les deux configu-

Français	**Anglais**
<SPEAKER ID="**206**" LANGUAGE="" NAME="" AFFILIATION="Cf. procès-verbal."/>	
<SPEAKER ID="**207**" LANGUAGE="" NAME="Le Président." AFFILIATION="">	<SPEAKER ID="**206**" LANGUAGE="" NAME="President." AFFILIATION="">
<P>	<P>
- L'ordre du jour appelle l'heure des questions (B6-0312/2006).	The next item is Question Time (B6-0312/2006).
<P>	<P>
Les questions suivantes ont été posées au Conseil.	The following questions have been submitted to the Council.
<P>	<P>

Table 5.XIII: Exemple de problème des speakers non alignés.

rations nous donnent des alignements incorrects. En ce qui concerne la configuration CSOURCE, elle nous donne un alignement correct de tout le document, même si elle ne présuppose pas de marquage (balisage) des textes. Ceci reflète la qualité de d'alignement que nous pouvons avoir avec notre méthode.

5.3.2.2 Traduction automatique statistique

Nous souhaitons enfin vérifier si la qualité de l'alignement a un impact sur la traduction. Nous utilisons pour cela le système de traduction automatique statistique ***Moses***[10].

Ce système permet d'entrainer automatiquement (sur une grande quantité de données parallèles) des modèles de traduction pour une paire de langues donnée. L'entrée de ce système est une collection de textes parallèles. Nous voulons valider chaque sortie produite par les 5 configurations décrites dans la section 5.3.2. Pour cela, nous utilisons les bi-textes obtenus par les configurations décrites en section 5.3.2 comme des entrées pour entrainer des modèles de traduction. Pour chaque configuration nous entrainons un modèle de traduction. À la fin, nous lançons des traductions sur des corpus de test et

[10] `http://www.statmt.org/moses/`

évaluons les résultats de chaque modèle.

Le processus de traduction passe par plusieurs étapes qui sont décrites en détail dans le tutoriel de *moses* [11]:

1. **Préparation de données**: Dans notre cas, l'étape de préparation de données passe elle même par deux phases essentielles:

   Tokenisation: Il s'agit d'insérer les espaces là où il faut, par exemple entre les mots et les points.

   Cleaning: Il s'agit de nettoyer les corpus. Seules les paires non vides, dont chaque phrase contient au plus 100 mots est retenue.

2. **Entrainement de modèle de langue**: L'utilisation d'un modèle de langue nous assure une sortie fluide. Nous utilisons l'outil **SRILM** pour construire les modèles de langue[12].

3. **Entrainement du modèle de traduction**: C'est l'étape la plus longue dans le processus de traduction. L'entrainement d'un tel modèle de traduction requiert l'alignement de mot (utilisant l'outil GIZA++), un algorithme d'extraction de phrases et l'attribution de score.

4. **Tuning**: Cette étape requiert une petite quantité de données (petit bi-texte) séparée des données d'entrainement. Pour cela, nous utilisons pour nos 5 systèmes le même bi-texte "news commentary (NEWS08)"[13]. Cette étape opère un réajustement des scores obtenus par l'étape d'entrainement du modèle de traduction et du modèle de langue.

[11] `http://www.statmt.org/moses/manual/manual.pdf`
[12] `http://www.speech.sri.com/projects/srilm/`
[13] `http://www.statmt.org/wmt12/`

5.3.2.3 Évaluation et résultats

Une fois nos modèles entrainés, et les scores réajustés, nous lançons des tests de traduction sur différents corpus de test (News 2009, News 2010, News 2011 et Hansard). Le tableau 5.XIV représente les scores BLEU produits par les 5 systèmes. Le score BLEU est fondé sur la précision n-grammes entre la phrase traduite et une traduction de référence. Il s'agit de la métrique la plus utilisée pour évaluer une traduction candidate selon une ou plusieurs références. Un score proche de 100 désigne une traduction proche de la référence.

	NEWS09	**NEWS10**	**NEWS11**	**HANS**
SOURCE	**20.02**	20.46	**21.16**	23.64
PK	19.43	20.14	20.77	23.46
JAPA	19.81	20.47	20.93	23.60
PK++	19.49	20.27	20.62	23.25
JAPA++	19.96	**20.61**	21.05	**23.95**

Table 5.XIV: Résultats bleu des 5 configurations pour la paire français-anglais.

Nous constatons que JAPA++ et SOURCE sont les meilleures configurations. Cette observation nous mène à conclure que l'utilisation d'un bon aligneur a un impact sur les résultats de traduction. Nous lançons les mêmes expérimentations avec la paire allemand-anglais. Les tableaux 5.XV présente les résultats produits par les 5 systèmes. Pour la paire finnois-anglais, nous lançons les mêmes expérimentations mais avec un jeu de test obtenu à partir du corpus JRC-Acquis. Le tableau 5.XVI présente les résultats produits par les 5 systèmes.

Même si le score BLEU est moins élevé que la paire français-anglais, nous constatons que les systèmes JAPA et JAPA++ ont toujours un léger avantage (non significatif) par rapport au autres.

Pour mieux analyser les résultats obtenus par *Moses*, nous considérons les vocabulaires des bi-textes obtenus par chaque configuration. Nous nous basons que sur les deux configurations JAPA++ et PK. Nous créons une liste à partir de ces vocabulaires

	NEWS09	**NEWS10**	**NEWS11**
SOURCE	15.54	16.12	15.26
PK	16.44	17.10	16.30
JAPA	16.39	**17.13**	**16.31**
PK++	16.35	17.08	16.24
JAPA++	**16.45**	17.09	16.30

Table 5.XV: Résultats bleu des 5 configurations pour la paire allemand-anglais.

	JRC-Acquis
SOURCE	09.44
PK	09.56
JAPA	**10.42**
PK++	10.08
JAPA++	**10.48**

Table 5.XVI: Résultats bleu des 5 configurations pour la paire finnois-anglais.

dans laquelle, nous ne gardons que les mots vus au plus 3 fois par PK, et vus plus que 3 fois par JAPA++. À partir de cette nouvelle liste, nous cherchons dans les corpus de test, toutes les phrases qui contiennent des mots de cette nouvelle liste, puis, nous créons 3 jeux de test:

1. $f = 0$: Chaque phrase contient au moins un mot non vu pas PK, c'est à dire la fréquence des mots égale à zéro ($f = 0$).

2. $f \in [1,3]$: Chaque phrase contient au moins un mot de la nouvelle liste, c'est à dire la fréquence est entre un et trois ($f \in [1,3]$).

3. $f \in [0,3]$: Représente la concaténation des deux premiers, c'est à dire la fréquence est entre zéro et trois ($f \in [0,3]$).

Le tableau 5.XVII représente les résultats de traduction des trois jeux de test. Nous observons que les gains en bleu (pour les trois paires français-anglais, allemand-anglais et finnois-anglais) sont stables, et qu'ils sont en faveur de la configuration JAPA++.

	($f = 0$)		($f \in [1,3]$)		($f \in [0,3]$)	
	JAPA++	PK	JAPA++	PK	JAPA++	PK
fr-en	**30.4** (118 Phra.)	29.5	**21.5** (394 Phra.)	20.6	**23.7** (512 Phra.)	22.7
al-en	**14.2** (100 Phra.)	13.8	**16.3** (344 Phra.)	16.1	**15.9** (444 Phra.)	15.7
fi-en	**7.5** (38 Phra.)	7.0	**14.6** (227 Phra.)	14.1	**14.3** (256 Phra.)	13.9

Table 5.XVII: Comparaison de scores BLEU obtenu par JAPA++ et PK pour les phrases contenants des mots peu fréquents.

CHAPTER 6

CONCLUSION

Plusieurs approches ont été proposées pour faire l'alignement des corpus parallèles. Certaines se basent sur des indices de longueur de phrases [Gale et Church, 1991], d'autres se basent sur des indices lexicaux [Simard et al., 1992], d'autres encore se basent sur la combinaison de plusieurs indices (indices hybrides). Nous avons décrit une approche très simple qui produit de très bons résultats. Ces résultats surpassent ceux produits par les meilleurs approches disponibles.

Nous avons lancé plusieurs expériences sur plusieurs types de données. Le but de ces expériences est de valider s'il est légitime de considérer l'étape de l'alignement de phrases comme résolue, et aussi, comment notre approche va réagir face à un taux de bruit élevé. Pour la première expérience, nous avons augmenté le taux de bruit jusqu'à arriver à 60%, et nous avons observé à chaque fois la qualité de l'alignement produit. Nous avons constaté que les résultats produits par notre approche se dégradent au fur et à mesure que le bruit est introduit. En revanche l'approche de BMA se dégrade d'une manière plus importante. Nous avons ensuite réalisé des expériences sur le corpus EuroParl, dans lesquelles nous avons montré l'impact de notre approche sur des systèmes de traduction automatique. Notre méthode d'alignement permet d'obtenir des bi-textes qui contiennent davantage de mots, ce qui enrichit lors de l'entrainement, à la fois le modèle de traduction et le modèle de langue.

Nous avons constaté que notre approche est consistante face aux taux de bruit élevés. De plus, dans un cadre bien étudié, nous avons constaté que notre approche a un impact sur les systèmes de traduction automatique. Notre observation s'oppose en partie à [Goutte et al., 2012] qui concluent que les systèmes de traduction automatique peuvent traiter un taux d'erreur d'alignement de phrases atteignant les 30% sans différence notable en traduction.

À la fin, nous avons montré des situations où un bon alignement de phrases influence le processus de traduction. Nous tenons à ajouter qu'il est rare, voire impossible de trouver des corpus parallèles propres, même avec les recherches qui visent à extraire des corpus parallèles à partir des corpus comparables [Fung et Cheung, 2004, Uszkoreit et al., 2010, Utiyama et Isahara, 2003]. Ceci justifie le recourt à un algorithme d'alignement plus performant que ceux existants. Cependant, nous pensons que l'alignement de phrases mérite davantage d'être étudié.

À la fin, ce travail a donné lieu à une publication dans une conférence internationale MT Summit qui s'est tenu en Septembre 2013 à Nice (France).

BIBLIOGRAPHY

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours et Rico Sennrich. Extrinsic evaluation of sentence alignment systems. Dans *Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*, pages 6–10, May 2012. URL http://dx.doi.org/10.5167/uzh-62565.

Peter F. Brown, Jennifer C. Lai et Robert L. Mercer. Aligning sentences in parallel corpora. Dans *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 169–176, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/981344.981366.

Jason S. Chang et Mathis H. Chen. An alignment method for noisy parallel corpora based on image processing techniques. Dans *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, EACL '97, pages 297–304, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/979617.979655.

Stanley F. Chen. Aligning sentences in bilingual corpora using lexical information. Dans *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 9–16, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/981574.981576.

Fathi Debili et Elyès Sammouda. Aligning sentences in bilingual texts: French-english and french-arabic. Dans *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 517–524, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/992133.992151.

Hervé Déjean, Éric Gaussier et Fatia Sadat. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. Dans *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02,

pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1072228.1072394.

Pascale Fung et Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. Dans *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1220355.1220506.

Pascale Fung et Kathleen Mckeown. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. Dans *In Proceedings of the First Conference of the Association for Machine Translation in the Americas, 81–88*, pages 81–88, 1994.

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard et François Yvon. Two ways to use a noisy parallel news corpus for improving statistical machine translation. Dans *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 44–51, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-015. URL http://dl.acm.org/citation.cfm?id=2024236.2024246.

William A. Gale et Kenneth W. Church. A program for aligning sentences in bilingual corpora. Dans *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/981344.981367.

Cyril Goutte, Marine Carpuat et George Foster. The impact of sentence alignment errors on phrase-based machine translation performance. Dans *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012. URL http://www.mt-archive.info/AMTA-2012-Goutte.pdf.

Pierre Isabelle et Michel Simard. Propositions pour la représentation et l'évaluation des alignements de textes parallèles dans l'arc a2. *Rapport technique*, 1996.

Martin Kay et Martin Röscheisen. Text-translation alignment. *Comput. Linguist.*, 19(1):121–142, mars 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972450.972457`.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5, 2005.

Grzegorz Kondrak. Determining recurrent sound correspondences by inducing translation models. Dans *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '01, pages 1–7, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1072228.1072244`.

Olivier Kraif. Constitution et exploitation de bi-textes pour l'aide à la traduction, 2001.

John Laffling. On constructing a transfer dictionary for man and machine. *Target*, 4(1): 17–31, 1992. URL `http://www.ingentaconnect.com/content/jbp/targ/1992/00000004/00000001/art00002`.

P. Langlais, M. Simard, S Armstrong, P. Bonhomme, F. Débili, P. Isabelle, E. Soussi et P Théron. The ARCADE: A Cooperative Research Project on Bilingual Text Alignment. Dans *1st International Conference On Language Resources and Evaluation (LREC)*, Granada, Spain, 1998.

Philippe Langlais et Jean Véronis. Progress in parallel text alignment techniques for multilingual lexical acquisition: the ARCADE evaluation exercise. Dans *2nd Workshop on Lexical Semantics Systems (WLSS'98)*, Pisa, Italy, April 1998.

Charlotte Lecluze. Alignement de documents multilignues sans présupposé de parallélisme, 2011.

Peng Li, Maosong Sun et Ping Xue. Fast-champollion: a fast and robust sentence alignment algorithm. Dans *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 710–718, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944647.

Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. Dans *In Proceedings Fifth International Conference on Language Resources and Evaluation of LREC-2006*, 2006.

RobertC. Moore. Fast and accurate sentence alignment of bilingual corpora. Dans StephenD. Richardson, éditeur, *Machine Translation: From Research to Real Users*, volume 2499 de *Lecture Notes in Computer Science*, pages 135–144. Springer Berlin Heidelberg, 2002. ISBN 978-3-540-44282-0. URL http://dx.doi.org/10.1007/3-540-45820-4_14.

J. A. Nelder et R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

Jian-Yun Nie et Jiang Chen. Learning translation models from the web. Dans *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*, volume 4, pages 1999–2004 vol.4, 2002.

António Ribeiro, Gabriel Lopes et João Mexia. Extracting equivalents from aligned parallel texts: Comparison of measures of similarity. Dans MariaCarolina Monard et JaimeSimão Sichman, éditeurs, *Advances in Artificial Intelligence*, volume 1952 de *Lecture Notes in Computer Science*, pages 339–349. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-41276-2. URL http://dx.doi.org/10.1007/3-540-44399-1_35.

Lei Shi, Cheng Niu, Ming Zhou et Jianfeng Gao. A dom tree alignment model for mining parallel data from the web. Dans *In COLING/ACL-2006*, pages 489–496, 2006.

Michel Simard. The baf: A corpus of english-french bitext, 1998.

Michel Simard, George F. Foster et Pierre Isabelle. Using cognates to align sentences in bilingual corpora. Dans *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2*, CASCON '92, pages 1071–1082. IBM Press, 1992. URL http://dl.acm.org/citation.cfm?id=962367.962411.

Michel Simard et Pierre Plamondon. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80, 1998. ISSN 0922-6567. URL http://dx.doi.org/10.1023/A%3A1008010319408.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat et Moshe Dubiner. Large scale parallel document mining for machine translation. Dans *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1873781.1873905.

Masao Utiyama et Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. Dans *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1075096.1075106.

C. J. van Rijsbergen. (invited paper) a new theoretical framework for information retrieval. Dans *Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '86, pages 194–200, New York, NY, USA, 1986. ACM. ISBN 0-89791-187-3. URL http://doi.acm.org/10.1145/253168.253208.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón et V. Nagy. Parallel corpora for medium density languages. Dans *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596, 2005.

Jean Véronis. *Alignement de corpus multilingues*, chaptitre 6, pages 316–334. Ingénierie des langues. HERMES, 2000. Parallel Text Processing.

Jean Véronis et Langlais Philippe. *Evaluation of Parallel Text Alignment Systems*, chaptitre X, pages 369–388. Text Speech and Language Technology Series. Kluwer Academic Publishers, 2000. Parallel Text Processing.

Dekai Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. Dans *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 80–87, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/981732.981744`.