Généralisation de l'alignement sous-phrastique par échantillonnage

Adrien Lardilleux, François Yvon, Yves Lepage

TALN 2011, Montpellier, 27 juin –1er juillet 2011

Généralisation de l'alignement sous-phrastique par échantillonnage

Adrien Lardilleux[1] François Yvon[1,2] Yves Lepage[3]
(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex
(2) Université Paris-Sud
(3) IPS, université Waseda, Japon
Adrien.Lardilleux@limsi.fr, Francois.Yvon@limsi.fr, Yves.Lepage@aoni.waseda.jp

Résumé. L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de textes multilingues parallèles alignés au niveau de la phrase. Un tel alignement est nécessaire, par exemple, pour entraîner des systèmes de traduction statistique. L'approche standard pour réaliser cette tâche implique l'estimation successive de plusieurs modèles probabilistes de complexité croissante et l'utilisation d'heuristiques qui permettent d'aligner des mots isolés, puis, par extension, des groupes de mots. Dans cet article, nous considérons une approche alternative, initialement proposée dans (Lardilleux & Lepage, 2008), qui repose sur un principe beaucoup plus simple, à savoir la comparaison des profils d'occurrences dans des sous-corpus obtenus par échantillonnage. Après avoir analysé les forces et faiblesses de cette approche, nous montrons comment améliorer la détection d'unités de traduction longues, et évaluons ces améliorations sur des tâches de traduction automatique.

Abstract. Sub-sentential alignment is the process by which multi-word translation units are extracted from sentence-aligned multilingual parallel texts. Such alignment is necessary, for instance, to train statistical machine translation systems. Standard approaches typically rely on the estimation of several probabilistic models of increasing complexity and on the use of various heuristics that make it possible to align, first isolated words, then, by extension, groups of words. In this paper, we explore an alternative approach, originally proposed in (Lardilleux & Lepage, 2008), that relies on a much simpler principle, which is the comparison of occurrence profiles in sub-corpora obtained by sampling. After analyzing the strengths and weaknesses of this approach, we show how to improve the detection of long translation units, and evaluate these improvements on machine translation tasks.

Mots-clés : alignement sous-phrastique, traduction automatique par fragments.

Keywords: sub-sentential alignment, phrase-based machine translation.

1 Introduction

L'alignement sous-phrastique consiste à extraire des traductions d'unités textuelles de grain inférieur à la phrase à partir de corpus multilingues parallèles, c'est-à-dire dont les phrases ont préalablement été mises en correspondance. Cette tâche constitue la première étape de la plupart des systèmes de traduction automatique fondés sur les données (traduction statistique et traduction par l'exemple). Les systèmes qui concentrent aujourd'hui les efforts de recherche sont majoritairement des systèmes statistiques par fragments (*phrases* en anglais), qui utilisent comme principale ressource une table de traductions, dérivée d'alignements sous-phrastiques. Un telle table consiste en une liste pré-calculée de couples de traductions associant à chaque couple de fragments (*source*, *cible*) un certain nombre de scores reflétant la probabilité que *source* se traduise par *cible*.

On peut globalement inscrire les méthodes d'alignement sous-phrastique dans l'un des deux courants suivants : l'approche estimative, introduite par Brown *et al.* (1988), et l'approche associative, introduite par Gale & Church (1991). La première est la plus utilisée à ce jour, principalement parce qu'elle est parfaitement intégrée à la traduction automatique statistique, dont elle constitue un pilier depuis l'apparition des modèles IBM (Brown *et al.*, 1993). Cette approche consiste à définir un modèle probabiliste du corpus parallèle dont les paramètres sont estimés selon un processus de maximisation globale sur l'ensemble des couples de phrases disponibles. Pratiquement, le but est de déterminer les meilleurs appariements possibles entre les mots sources et cibles dans chacun des couples de phrases parallèles. Dans la seconde approche, on établit une liste de traductions candidates soumises à un test d'indépendance statistique, tels que l'information mutuelle (Fung & Church, 1994) ou le rapport de vraisemblance

(Dunning, 1993) — voir (Melamed, 2000; Moore, 2005) pour des travaux récents dans cette lignée. Il s'agit ici d'un processus de maximisation *locale* : chaque segment est traité indépendamment des autres. Cette approche est plus souvent utilisée pour extraire directement des couples de traductions, tandis que la première cherche avant tout à établir des *liens* de traduction entre les mots sources et cibles de chacun des couples de phrases du corpus d'entrée. Ces liens permettent, dans un deuxième temps, d'extraire des couples de traductions.

Nous avons récemment proposé une méthode d'alignement sous-phrastique (Lardilleux & Lepage, 2008, 2009; Lardilleux, 2010), apparentée aux méthodes associatives, s'attaquant à un certain nombre de problèmes souvent négligés dans le domaine : traitement simultané de multiples langues, parallélisme massif, passage à l'échelle au cœur de la méthode, et simplicité de mise en œuvre. En moyenne, cette méthode s'est révélée meilleure que l'état de l'art sur des tâches de constitution de lexiques bilingues, mais en retrait sur des tâches de traduction automatique par fragments (Lardilleux *et al.*, 2009). Nous n'avions émis jusqu'alors que des *hypothèses* pour expliquer ces résultats a priori contradictoires. Dans cet article, nous proposons une analyse fine du comportement de notre méthode afin de déterminer l'origine de ces différences, ainsi qu'une généralisation destinée à améliorer ses performances en traduction automatique par fragments.

Cet article est organisé de la façon suivante : la section 2 présente une vue d'ensemble de la méthode d'alignement d'origine ; la section 3 présente des expériences mettant en évidence l'origine de ses faiblesses ; nous décrivons dans la section 4 une généralisation, et évaluons ses performances ; et la section 5 conclut ces travaux.

2 Vue d'ensemble de la méthode d'alignement d'origine

2.1 Principes de base

Notre méthode d'alignement peut être vue comme une émulation des méthodes associatives, à la différence (majeure) près qu'elle ne se restreint pas à aligner des *couples de mots*[1] (*source*, *cible*). Elle permet, en effet, de considérer des *séquences de mots* de taille variable, éventuellement discontinues, qui partagent strictement la même distribution (répartition) dans les phrases du corpus parallèle d'entrée, indépendamment de leur langue. Ces séquences constituent en fait un sous-ensemble des candidats de traduction qui obtiendraient un score maximal par des tests d'association statistiques. Le nombre de séquences de mots ayant exactement la même distribution étant réduit, nous ne recherchons pas ces séquences dans le corpus d'entrée même, mais dans des sous-corpus de celui-ci, l'idée étant que plus un sous-corpus est petit, plus les mots qu'il contient ont de chances de partager la même distribution, et que par conséquent plus le nombre de mots alignés dans ce sous-corpus est élevée.

Le cœur de la méthode consiste donc à extraire des alignements à partir de multiples sous-corpus indépendants construits par échantillonnage. En pratique, nous privilégions les sous-corpus de petite taille car ils sont plus rapides à traiter et semblent donner de meilleurs résultats (Lardilleux, 2010). Pour chaque séquence de mots de même distribution dans un sous-corpus, deux alignements sont extraits : la séquence elle-même, d'une part, et son complémentaire, d'autre part. Le nombre de sous-corpus à traiter n'étant pas défini à l'avance, le processus est *anytime*, c'est-à-dire qu'il peut être interrompu à tout moment par l'utilisateur, ou selon des critères tels que le temps écoulé ou le taux de couverture du corpus de départ. Plus le nombre de sous-corpus traités est élevé, plus la couverture du corpus de départ est grande et plus les mesures d'association sont précises. Les alignements extraits sont collectés à partir de l'ensemble des sous-corpus traités, et sont évalués par divers scores (probabilité de traduction et poids lexicaux (Koehn *et al.*, 2003)) à proportion du nombre de fois qu'ils ont été extraits. Le résultat est une table de traductions directement utilisable, par exemple, pour des tâches de traduction automatique.

2.2 Algorithme complet

L'algorithme d'extraction complet est schématisé dans le tableau 1.

La figure 1 illustre les principales étapes de l'algorithme sur un exemple d'alignement d'un texte trilingue. Dans la suite de cet article consacré aux applications de l'alignement en traduction automatique, nous nous limiterons à une application bilingue de la méthode, bien que son caractère multilingue en constitue un atout majeur.

[1] Nous employons le terme « mot » pour désigner toute forme graphique identifiée par un programme de *tokenisation*.

Entrée : un corpus multilingue, ici arabe-français-anglais.

1	قهوة ، من فضلك .	↔ Un café , s'il vous plaît .	↔ One coffee , please .
2	هذه قهوة ممتازة .	↔ Ce café est excellent .	↔ This coffee is excellent .
3	شاي ثقيل .	↔ Un thé fort .	↔ One strong tea .
4	قهوة ثقيلة .	↔ Un café fort .	↔ One strong coffee .

⇓

Transformation en corpus alingue *(= monolingue) en concaténant les traductions d'une même phrase et distinguant les mots en fonction de leur langue d'origine. Sélection d'un sous-corpus aléatoire (ici, les trois premières lignes du corpus d'origine).*

1	$_{1}$. $_{1}$فضلك $_{1}$من $_{1}$، $_{1}$قهوة	Un$_2$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$	One$_3$ coffee$_3$,$_3$ please$_3$.$_3$
2	$_{1}$. $_{1}$ممتازة $_{1}$قهوة $_{1}$هذه	Ce$_2$ café$_2$ est$_2$ excellent$_2$.$_2$	This$_3$ coffee$_3$ is$_3$ excellent$_3$.$_3$
3	$_{1}$. $_{1}$ثقيل $_{1}$شاي	Un$_2$ thé$_2$ fort$_2$.$_2$	One$_3$ strong$_3$ tea$_3$.$_3$

⇓

Indexation des mots (calcul des vecteurs de présence). Les mots ayant même distribution sont regroupés.

	$_{1}$.	.$_2$	.$_3$	$_{1}$قهوة	café$_2$	coffee$_3$	One$_3$	Un$_2$	$_{1}$،	$_{1}$فضلك	$_{1}$من	,$_3$	,$_2$	plaît$_2$	please$_3$	s'il$_2$	vous$_2$	$_{1}$هذه	$_{1}$ممتازة	Ce$_2$	This$_3$	est$_2$	excellent$_2$	…
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	…
2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	…
3	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…

⇓

Chaque groupe de mots permet d'extraire deux alignements par phrase où il apparaît.

Les mots :	apparaissent dans les phrases :	d'où sont extraits :
$_{1}$قهوة café$_2$ coffee$_3$	1	$_{1}$قهوة café$_2$ coffee$_3$ $_{1}$. $_{1}$فضلك $_{1}$من $_{1}$، Un$_2$ _ ,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ One$_3$ _ ,$_3$ please$_3$.$_3$
	2	$_{1}$قهوة café$_2$ coffee$_3$ $_{1}$. $_{1}$ممتازة _ $_{1}$هذه Ce$_2$ _ est$_2$ excellent$_2$.$_2$ This$_3$ _ is$_3$ excellent$_3$.$_3$

⋮

⇓

Décompte des alignements et rétablissement des limites entre langues.

Arabe		Français		Anglais	Décompte
قهوة	↔	café	↔	coffee	2
. ، من فضلك	↔	Un _ , s'il vous plaît .	↔	One _ , please .	1
. هذه _ ممتازة	↔	Ce _ est excellent .	↔	This _ is excellent .	1

⋮

FIG. 1 – Vue d'ensemble de la méthode d'alignement. C'est la phase d'indexation et de constitution des groupes de mots (troisième étape sur la figure) que nous généraliserons dans la suite de l'article.

Transformer le corpus parallèle d'entrée, multilingue, en corpus *alingue* (= monolingue)
Initialiser un tableau associatif *CompteurAlignements*
Faire
    Sélectionner un sous-corpus par échantillonnage
    Indexer les mots par leur vecteur de présence dans les *phrase*s du sous-corpus
    Les mots de même distribution sont rassemblés dans un même *groupe*
    Pour chaque *groupe* de mots :
        Pour chaque *phrase* où le *groupe* apparaît :
            Rétablir l'ordre des mots du *groupe*
            CompteurAlignements[groupe] ++
            CompteurAlignements[phrase - groupe] ++
Jusqu'à interruption par l'utilisateur **ou** temps imparti écoulé **ou** plus aucun alignement obtenu **ou** tout autre critère
Calculer les scores des alignements

TAB. 1 – Les étapes de la méthode d'alignement.

2.3 Résultats

Dans cette section, nous résumons les principaux résultats et conclusions de (Lardilleux, 2010). Nous avons évalué cette méthode d'alignement sur deux tâches : en traduction automatique statistique par fragments et en constitution de lexiques bilingues. L'implémentation de notre méthode, *Anymalign*[2], est comparée à MGIZA++[3] (Gao & Vogel, 2008), l'implantation la plus récente des modèles IBM. Anymalign étant *anytime*, nous commençons en pratique par exécuter MGIZA++ avec ses paramètres par défaut (5 itérations de chacun des modèles IBM1, HMM, IBM3 et IBM4), mesurons son temps d'exécution, et exécutons Anymalign pendant la même durée. Les corpus parallèles utilisés dans les expériences sont principalement Europarl (Koehn, 2005) et des extraits du BTEC (Takezawa *et al.*, 2002), distribués lors des campagnes d'évaluation de traduction automatique IWSLT (Fordyce, 2007). Les extraits du BTEC sont constitués de 20 000 à 40 000 couples de phrases courtes alignées (10 mots anglais en moyenne) et ceux d'Europarl de 100 000 couples de phrases longues (30 mots anglais).

Dans la tâche de traduction automatique statistique par fragments, nous comparons les scores obtenus par Moses (Koehn *et al.*, 2007) avec sa table de traductions par défaut, construite à partir des alignements de MGIZA++, et celle produite par Anymalign. En moyenne, Anymalign est en retrait de deux points BLEU (Papineni *et al.*, 2002) sur l'ensemble des expériences que nous avons menées. Dans le meilleur des cas, nous avons obtenu un gain d'un point par rapport à MGIZA++ (BTEC, japonais-anglais) ; dans le pire, une perte de huit points (Europarl, finnois-anglais). Dans l'ensemble, les écarts sont plus prononcés sur Europarl que sur le BTEC.

Dans la tâche de constitution de lexiques bilingues, nous comparons les tables de traductions produites par les deux aligneurs avec un lexique bilingue de référence[4]. Dans un premier temps, ce lexique est filtré de façon qu'il ne contienne que des couples de traductions qui peuvent effectivement être extraits par les aligneurs à partir du corpus parallèle d'entrée. En pratique, un couple de traductions du lexique de référence est conservé s'il s'agit d'une sous-séquence d'un couple de phrases du corpus parallèle. Nous définissons alors le score d'une table de traductions relativement à ce lexique de référence filtré comme la somme des probabilités de traduction source → cible des alignements de la table de traductions présents dans la référence, divisée par le nombre d'entrées distinctes dans la référence. Le résultat s'interprète comme un score de rappel, entre 0 et 1. En moyenne, Anymalign est meilleur de 7 % relativement à MGIZA++ sur l'ensemble des expériences que nous avons menées. Dans le meilleur des cas, nous avons obtenu un gain relatif de 70 % (Europarl, finnois-français) ; dans le pire une perte de 18 % (Europarl, suédois-finnois). Le genre de textes constituant le corpus ne semble pas avoir d'influence majeure sur ces scores.

En résumé, notre méthode est en retrait sur les tâches de traduction automatique par fragments, mais produit de meilleurs alignements de mots, comme l'attestent les résultats de comparaison avec lexiques de référence, dont les entrées sont majoritairement des mots simples (le nombre moyen de mots par entrée est 1,2). Nous avons montré (Lardilleux *et al.*, 2009) que cela est en fait principalement dû à la faible capacité de cette méthode à produire des alignements de n-grammes de mots avec $n \geqslant 2$, comme l'illustre la figure 2. Le but de la section suivante est de mettre en évidence l'origine de ces différences.

[2] `http://users.info.unicaen.fr/~alardill/anymalign`
[3] `http://geek.kyloo.net/software/doku.php/mgiza:overview`
[4] Nos lexiques proviennent principalement du site XDXF : `http://xdxf.sourceforge.net`

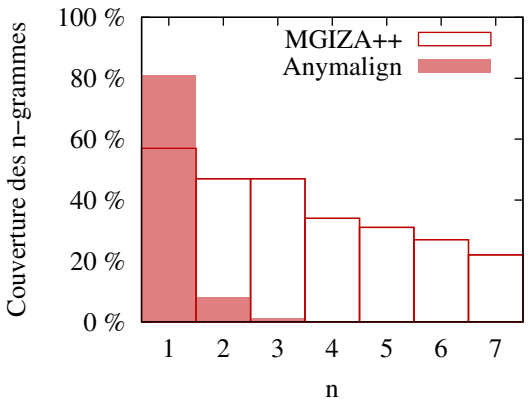

FIG. 2 – Couverture de la partie source d'un échantillon d'Europarl français-anglais par les tables de traductions de MGIZA++ et d'Anymalign. Anymalign aligne plus d'unigrammes, mais peu de n-grammes plus longs.

3 Une analyse du comportement de la méthode

Dans cette section, nous présentons des expériences montrant que deux causes principales sont à l'origine des résultats apparemment contradictoires présentés ci-dessus : les différences de fréquences des mots qui composent les séquences à aligner (cause propre à la méthode), et les fréquences de mots utiles à ces tâches (cause propre à la tâche). Les expériences présentées ici sont réalisées sur un extrait d'environ 320 000 phrases d'Europarl, avec les couples de langues portugais-espagnol (cas extrêmes de langues proches dans nos expériences) et finnois-anglais (cas extrême de langues éloignées : le finnois est une langue ouralienne agglutinante, l'anglais une langue germanique d'influence romane isolante, ce qui s'exprime par une grande différence de taille des vocabulaires). Le tableau 2 présente le nombre de mots de chaque partie de nos corpus.

Langue	Nombre de mots *(tokens)*	Taille du vocabulaire
portugais	9 249 177	87 341
espagnol	9 330 199	85 366
finnois	6 472 649	274 958
anglais	8 955 995	53 704

TAB. 2 – Caractéristiques des corpus utilisés pour nos analyses.

3.1 Différences de fréquences

Nous avons précédemment montré (Lardilleux *et al.*, 2009) qu'en pratique, la contrainte d'identité des distributions qui est au cœur de la méthode empêche d'extraire des séquences composées de mots de fréquences différentes. Par exemple, un bigramme constitué d'un mot hapax suivi du point de fin de phrase (assimilé à un mot typographique) ne peut être produit, car en supposant que le point apparaisse dans toutes les phrases du corpus d'entrée, la seule configuration dans laquelle ces deux mots partageraient la même distribution serait un sous-corpus constitué d'une seule phrase. Dans une telle configuration, presque tous les mots seraient hapax, et la séquence extraite consisterait donc en l'unique phrase de ce sous-corpus. Le bigramme attendu serait donc « masqué » et ne pourrait pas être extrait isolément.

Nous faisons un pas supplémentaire en étudiant la taille des sous-corpus d'où les mots sont extraits en fonction de la fréquence de ces mots. Étant donné un mot source m_s à aligner isolément, trois cas peuvent se produire :

1. dans un sous-corpus « trop petit », d'autres mots sources ont la même distribution que m_s. Il n'est donc pas possible d'aligner m_s isolément.
2. dans un sous-corpus de taille « idéale », aucun autre mot source n'a la même distribution que m_s, et au moins un mot cible a cette distribution. m_s peut donc être aligné isolément.

3. dans un sous-corpus « trop grand », aucun autre mot source n'a la même distribution que m_s, mais aucun mot cible non plus. m_s ne peut donc pas être aligné du tout.

Il existe ainsi une plage de tailles de sous-corpus qui permet d'extraire un mot isolément. Cette plage dépend bien entendu du mot à extraire et plus particulièrement de sa fréquence. Ces plages sont déterminées empiriquement en mesurant, pour chaque mot source d'un corpus parallèle, la taille moyenne des sous-corpus à partir de laquelle il peut être aligné isolément, ainsi que celle à partir de laquelle il ne peut plus être aligné du tout. Pour cela, nous commençons par tirer aléatoirement un sous-corpus d'une seule phrase contenant ce mot, testons si le mot peut y être aligné, puis recommençons ce test en augmentant le sous-corpus d'une nouvelle phrase tirée aléatoirement. Le processus s'arrête lorsque plus aucun mot cible n'a la même distribution que le mot source testé.

Chaque expérience produit deux nombres : la taille à partir de laquelle le mot peut être aligné isolément (passage du cas 1 au cas 2 ci-dessus), et celle à partir de laquelle le mot ne peut plus être aligné du tout (du cas 2 au cas 3). Ce test est répété 1 000 fois pour chaque mot source, et nous effectuons la moyenne des mesures recueillies sur l'ensemble des 1 000 tirages. Les résultats sont présentés à la figure 3, par classes de mots de fréquences proches.

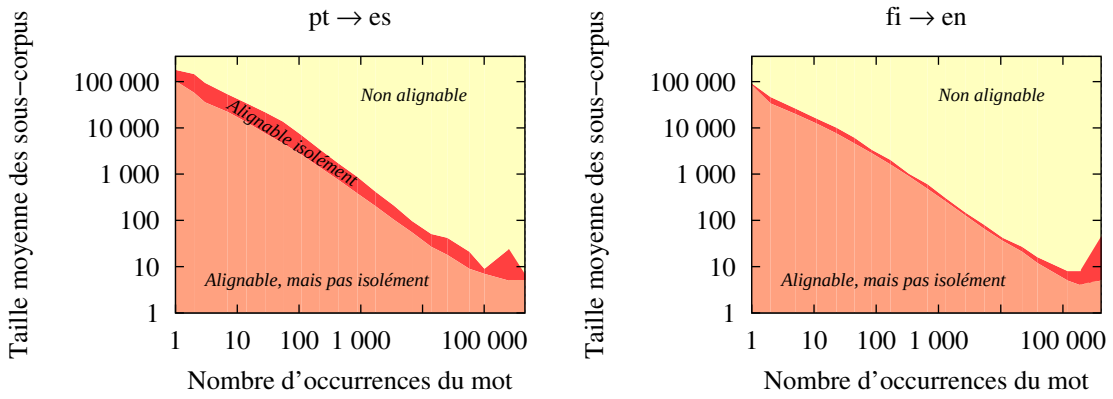

FIG. 3 – Tailles moyennes des sous-corpus à partir desquelles un mot source peut être extrait en fonction de la fréquence de ce mot. Dans la zone inférieure, le mot ne peut pas être aligné isolément (cas 1). Dans la zone du milieu, le mot peut être aligné isolément (cas 2). Dans la zone supérieure, le mot ne peut pas être aligné du tout (cas 3). Le petit sursaut de la limite supérieure à l'extrémité droite des deux graphiques est dû au point de fin de phrase, qui s'aligne plus facilement que les autres mots fréquents : il peut être aligné isolément dans des sous-corpus de 5 à 80 phrases environ.

Ces graphiques nous permettent de faire deux remarques. D'abord, la plage des tailles « idéales » des sous-corpus, autrement dit la largeur de la zone du milieu, varie grandement d'un couple de langues à l'autre. Notons que l'échelle logarithmique fait paraître cette plage plus étroite qu'elle ne l'est en réalité : le rapport moyen entre sa limite supérieure et sa limite inférieure est de 2,2 pour le couple espagnol-portugais et 1,2 pour le couple finnois-anglais. Cette différence de rapport s'explique aisément par les différences de morphologie des langues dans chacun de ces couples. Nous pouvons donc nous attendre à ce que l'alignement d'un mot donné par Anymalign nécessite le traitement de davantage de sous-corpus avec le couple finnois-anglais qu'avec le couple portugais-espagnol, puisqu'il est alors plus difficile de tirer aléatoirement un sous-corpus de la « bonne » taille.

La seconde remarque nous intéresse tout particulièrement dans le cadre de cet article : plus un mot est fréquent, plus les sous-corpus à partir desquels il est extrait sont petits, et réciproquement. Les mots rares (partie gauche des graphiques) sont donc alignés à partir de grands sous-corpus, tandis que les mots fréquents (partie droite des graphiques) sont alignés à partir de petits sous-corpus, constitués par exemple de 5 à 9 phrases pour la virgule. Ces résultats valident nos premières hypothèses : s'il est difficile de tirer un sous-corpus dans lequel deux mots source de fréquences différentes partagent la même distribution, c'est avant tout parce que ces mots ne peuvent pas être alignés à partir du même sous-corpus. Pour aligner des mots de fréquences différentes, il est nécessaire de les extraire à partir de sous-corpus de tailles différentes. Nous proposerons une alternative dans la section suivante.

3.2 Fréquences utiles

La seconde explication des différences de résultats d'Anymalign sur les deux tâches sur lesquelles il a été évalué provient en fait de la tâche elle-même, ou pour être plus précis du couple (aligneur, tâche).

Notre méthode et les modèles IBM reposent sur des intuitions opposées : la première tire parti de la rareté des mots pour les aligner (on réduit artificiellement et temporairement la fréquence de tous les mots en se plaçant dans un sous-corpus), tandis que les seconds sont estimés à partir des observations mesurées sur l'ensemble du corpus. En conséquence, Anymalign aligne mieux les mots rares, tandis que MGIZA++ aligne mieux les mots fréquents, comme l'illustre la figure 4.

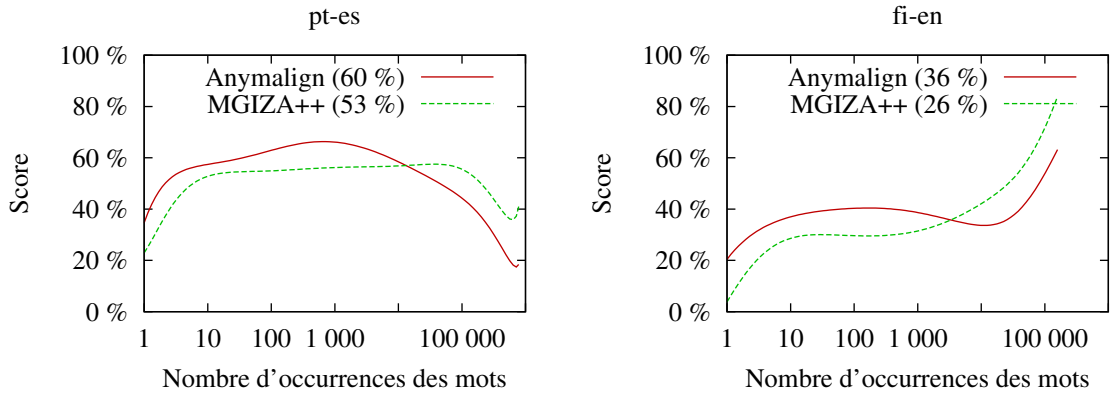

FIG. 4 – Scores obtenus par les tables de traductions produites par Anymalign et MGIZA++ sur la tâche de constitution de lexiques bilingues. Les scores entre parenthèses sont les scores globaux, calculés comme décrits au 3e paragraphe de la section 2.3. Les courbes présentent le détail de ces scores, en fonction du nombre d'occurrences du mot source de chacun des alignements : un score a été calculé localement pour chaque effectif de mot. Les courbes ont été lissées pour améliorer leur lisibilité.

Ce qui nous intéresse ici n'est pas tant l'allure générale des courbes que leur position relative : la courbe correspondant à Anymalign est au-dessus de celle de MGIZA++ pour les mots d'effectif 1 à 5 000 environ, et en-dessous pour les effectifs supérieurs. Cela montre qu'Anymalign aligne mieux non seulement les mots rares, mais également les mots de fréquence intermédiaire. Cette observation a été corroborée sur d'autres couples de langues (de-en, es-en, fr-en).

Or, les mots rares étant beaucoup plus nombreux dans tout texte — cf. loi d'Estoup-Zipf (Zipf, 1965; Mandelbrot, 1954; Montemurro, 2004) —, *a fortiori* dans notre corpus parallèle ainsi que dans les tables de traductions produites, et notre protocole d'évaluation par comparaison avec lexiques de référence traitant les mots indépendemment de leur fréquence, il est attendu que notre méthode obtienne de meilleurs scores en constitution de lexiques bilingues, puisque les mots qu'elle aligne le mieux sont au total les plus nombreux. À l'opposé, les mots fréquents sont beaucoup moins nombreux, mais autrement plus importants en traduction automatique car ils y sont beaucoup plus sollicités : un mot fréquent a plus de chances d'apparaître dans un jeu de test qu'un mot rare. Cela peut expliquer, au moins pour partie, les scores plus faibles d'Anymalign en traduction automatique. Idéalement, nous aimerions pouvoir utiliser les alignements de tel ou tel aligneur en fonction de la fréquence des mots, par exemple en combinant les tables de traductions produites par les aligneurs. Des expériences préliminaires utilisant les probabilités de traduction d'Anymalign comme fonction de trait supplémentaire dans la table de traduction par défaut de Moses ont donné des résultats prometteurs. Cela sort cependant du cadre de cet article, et nous nous consacrons par la suite à l'alignement de mots de fréquences différentes. Nous garderons néanmoins à l'esprit que, pour bien faire en traduction automatique, notre méthode devra également aligner plus efficacement les mots fréquents, ce que nous gardons pour des recherches futures.

4 Généralisation de la méthode à toutes les chaînes de mots

Dans cette section, nous présentons une généralisation de la méthode destinée à améliorer ses performances en traduction automatique statistique par fragments. En conformité avec la méthode d'origine, nous travaillerons

toujours sur les formes surfaciques des mots et sans ressource autre que le corpus d'entrée (traitement endogène). Notre but est d'extraire davantage d'alignements de n-grammes (chaînes de mots) avec $n \geqslant 2$ (cf. figure 2), tout en contournant le problème de l'extraction des mots de fréquences différentes (section 3.1).

4.1 Phase d'indexation

Nous introduisons le traitement à un grain variable en indexant des n-grammes plutôt que des mots. Nous ne chercherons pas à effectuer une segmentation particulière des phrases, par exemple en chunks, dont Vergne (2009) a montré qu'ils pouvaient être déterminés de façon endogène, mais traiterons plus simplement tous les n-grammes de mots se chevauchant. Considérons le (sous-)corpus d'entrée alingue[5] suivant, constitué de trois phrases :

1 a b c
2 a b d e
3 a c

L'indexation sur l'ensemble des n-grammes de ce corpus, *avant* recensement des groupes de même distribution servant de base à l'extraction des alignements, produit le résultat suivant :

	n = 1					n = 2					n = 3			n = 4
	a	b	c	d	e	ab	ac	bc	bd	de	abc	abd	bde	abde
1	1	1	1	0	0	1	0	1	0	0	1	0	0	0
2	1	1	0	1	1	1	0	0	1	1	0	1	1	1
3	1	0	1	0	0	0	1	0	0	0	0	0	0	0

Dans l'étape suivante, le recensement des groupes de même distribution, nous introduisons un changement majeur : si des n-grammes de même distribution se chevauchent, le groupe de mots résultant est constitué de l'union de ces n-grammes. Par exemple, les bigrammes de même distribution *bd* et *de* formeront le groupe de mots *bde*. Autrement dit, les groupes ne sont plus constitués de mots de même distribution, mais de mots issus de n-grammes de même distribution. Un même mot peut désormais apparaître dans plusieurs groupes, ce qui n'était pas le cas dans la méthode d'origine.

Ce changement soulève un problème qui ne pouvait pas se produire avec la méthode d'origine : des n-grammes peuvent masquer des (n−1)-grammes, et ce récursivement. L'unigramme *b* est par exemple masqué par le bigramme de même distribution *ab*, car l'union de *b* et *ab* donne *ab*, et *b* ne peut plus être aligné isolément. Il est donc nécessaire de traiter l'introduction de chaque longueur de n-gramme de façon spécifique.

4.2 Stratégie de constitution des groupes de mots

Nous avons testé trois stratégies :

1. traiter séparément les n-grammes en fonction de leur longueur. Ainsi, les groupes de mots ne sont construits qu'à partir de n-grammes de même longueur en source *et* en cible. Cela est bien entendu d'efficacité limitée sur des couples de langues tels que finnois-anglais : il serait préférable d'autoriser l'extraction d'un seul mot d'une langue agglutinante avec plusieurs mots d'une langue isolante.
2. permettre le mélange de toutes les longueurs de n-grammes, mais en ajoutant progressivement chaque longueur. L'ensemble initial ne contient que des unigrammes (méthode d'origine). Dans un deuxième temps, nous ajoutons les bigrammes et recréons tous les groupes de mots : certains sont identiques (les décomptes des alignements correspondants sont renforcés), d'autres sont nouveaux, d'autres enfin sont masqués mais cela n'a pas d'importance car ils ont déjà été extraits à partir des unigrammes. On ajoute ensuite les trigrammes, etc. Les alignements sont extraits à chaque fois que des n-grammes sont ajoutés.
3. forcer l'alignement de n-grammes de longueurs différentes, à contrepied de la première stratégie, en traitant séquentiellement tous les couples de longueurs (*source*, *cible*) possibles (produit cartésien). Cela permet l'alignement de n-grammes de longueurs très différentes en source et en cible, voire *trop* : puisque nous n'avons recours à aucune connaissance extérieure, Anymalign ne sait pas *a priori* quelle langue est traitée, et rien ne l'empêche par exemple de vouloir aligner des unigrammes en anglais avec de longs n-grammes en finnois, quand bien même il est peu probable que le moindre alignement puisse être produit à partir d'une telle configuration. En outre, la complexité de cette approche est bien plus importante que celle des deux précédentes, et ne passe pas à l'échelle lorsque nous traitons plus de deux langues simultanément.

[5]Comme décrit à la section 2, notre principal algorithme ne fait pas de différence entre corpus multilingues et corpus monolingues.

Pour comparer ces trois stratégies, nous préparons un ensemble de 100 000 sous-corpus aléatoires issus d'Europarl (français-anglais) et en extrayons les alignements selon chacune de ces stratégies. Nous réalisons l'expérience pour des longueurs maximales de n-grammes allant de 1 à 5. Les tables de traductions ($3 \times 5 = 15$ tables au total), obtenues à partir de ce *même* ensemble de sous-corpus, sont évaluées sur les mêmes tâches que précédemment : en traduction automatique statistique par fragments (les critères d'évaluation sont BLEU et TER (Snover *et al.*, 2006)) et en constitution de lexiques bilingues. Les résultats sont présentés dans le tableau 3.

Stratégie	*n* max.	Score en lexique (%)	BLEU (%)	TER (%)	Nombre d'entrées	Long. moy. des entrées
1.	1	36,19	21,12	63,57	83 967	1,92
	2	**36,71**	22,62	61,93	277 858	2,79
	3	36,66	23,08	62,06	366 971	3,13
	4	36,60	**23,23**	**61,43**	393 453	3,24
	5	36,58	22,92	62,14	399 810	3,27
2.	1	36,19	21,12	63,57	83 967	1,92
	2	37,08	23,63	60,68	290 631	2,78
	3	37,35	**24,72**	**59,86**	398 880	3,12
	4	37,45	24,47	60,69	436 760	3,25
	5	**37,56**	24,25	59,94	448 212	3,31
3.	1	**36,19**	21,12	63,57	83 967	1,92
	2	31,71	23,85	60,41	312 273	2,86
	3	30,90	**24,50**	60,68	453 429	3,24
	4	30,48	24,47	**59,96**	507 359	3,39
	5	30,25	24,26	60,03	524 091	3,45

TAB. 3 – Qualité et caractéristiques des tables de traductions produites selon chacune des trois stratégies de constitution de groupes de mots, pour différente longueurs maximales de n-grammes indexés. Les lignes où *n* max. = 1 sont identiques pour les trois stratégies et correspondent à la méthode d'origine.

Comme il était attendu, plus la longueur maximale des n-grammes indexés est grande, plus le nombre d'entrées dans la table de traductions et la longueur de ces entrées sont également élevés, car les alignements produits avec un *n* max. donné contiennent ceux produits avec un *n* max. inférieur (inclusion des tables). Les scores en constitution de lexiques augmentent de façon négligeable lorsque *n* max. augmente avec les deux premières approches, mais se dégradent de façon significative avec la troisième. Le gain en traduction automatique est significatif avec les trois approches. La seconde semble néanmoins fournir des résultats très légèrement meilleurs selon les trois critères d'évaluation. Son temps d'exécution est légèrement supérieur à celui de la première (au pire deux fois plus lent avec les 5-grammes), mais bien en-deçà de celui de la troisième (de l'ordre de l'heure à celui de la journée avec les 5-grammes).

La stratégie que nous utiliserons par la suite sera donc la deuxième. Elle constitue sur le fond un bon compromis entre les deux autres. La figure 5 présente le détail de la colonne « Nombre d'entrées » du tableau 3 pour cette deuxième stratégie, et est à confronter avec la figure 2.

Dans l'ensemble, l'ajout d'une longueur de n-grammes indexés, autrement dit le passage d'une courbe à celle immédiatement au-dessus, augmente considérablement la quantité de l'ensemble des n-grammes produits (y compris, de façon marginale, les n-grammes de taille inférieure, mais cela n'est dû qu'à l'extraction des complémentaires des groupes de mots). Le cas le plus significatif est celui de l'indexation des bigrammes (*n* max. = 2), qui fait exploser la quantité de bigrammes en sortie, et dans une moindre mesure de toutes les tailles de n-grammes supérieures. Le phénomène se produit également en indexant des n-grammes encore plus longs, mais cela est de moins en moins significatif à mesure que *n* max. augmente. Le graphique semble montrer qu'il n'est pas utile d'indexer des n-grammes de plus de 3 ou 4 mots, car cela se révèle peu productif. Les n-grammes qui nous intéressent le plus sont de toute façon ceux de longueur 1 à 3, parce que ce sont généralement les plus utiles en traduction automatique par fragments.

4.3 Expériences et nouveaux résultats

Nous comparons à présent notre méthode généralisée (indexation des n-grammes + constitution des groupes de mots selon la deuxième stratégie testée) à MGIZA++ sur des tâches de traduction automatique statistique par

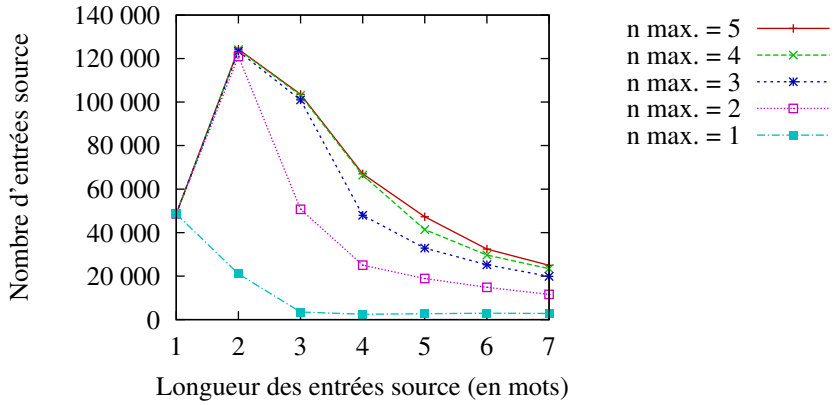

FIG. 5 – Distribution des n-grammes dans les cinq tables de traductions obtenues par la deuxième stratégie de constitution de groupes de mots. Chaque courbe correspond à une ligne du tableau 3, et la somme des ordonnées de ses points reportés est égale à la valeur indiquée dans la colonne « Nombre d'entrées » du tableau. La courbe la plus basse (*n* max. = 1) correspond à la méthode d'origine (cf. figure 2).

Tâche	Entraînement	Développement	Test	Références par phrase de test
BTEC : ar-en	19 972	1 512	489	7
BTEC : zh-en	19 972	1 512	989	7
Europarl : fi-en, fr-en, pt-es	318 804	500	1 000	1

TAB. 4 – Caractéristiques des corpus utilisés pour notre évaluation.

	Aligneur	*n* max.	BLEU (%)	TER (%)	Nombre d'entrées
ar-en	*MGIZA++*		*33,68*	*46,17*	*217 512*
	Anymalign	1	26,33	51,17	170 521
	-	2	30,88	49,70	269 454
	-	3	31,81	51,48	273 197
	-	4	**33,75**	**48,80**	258 141
zh-en	*MGIZA++*		*15,46*	*70,49*	*141 773*
	Anymalign	1	14,77	**68,97**	158 904
	-	2	16,35	71,70	263 315
	-	3	16,54	70,62	250 292
	-	4	**16,84**	69,45	269 353

TAB. 5 – Résultats des tâches de traduction sur le BTEC.

			Même temps de traitement que MGIZA++			*Temps théorique = 20 × MGIZA++*		
	Aligneur	*n* max.	BLEU (%)	TER (%)	Nombre d'entrées	BLEU (%)	TER (%)	Nombre d'entrées
fi-en	*MGIZA++*		*21,68*	*65,50*	*5 241 325*			
	Anymalign	1	13,73	77,57	1 871 639	13,54	74,34	5 178 683
	-	2	14,39	**76,59**	890 644	16,21	**71,18**	5 948 094
	-	3	**14,64**	77,15	696 420	**17,44**	72,63	4 001 816
	-	4	12,79	78,46	279 437	16,80	71,34	2 266 448
fr-en	*MGIZA++*		*29,39*	*54,37*	*10 783 083*			
	Anymalign	1	22,74	61,85	1 755 334	23,58	61,09	7 882 822
	-	2	**24,68**	60,22	1 805 297	24,55	58,42	8 317 221
	-	3	24,40	**59,77**	1 074 258	**25,29**	**57,66**	6 943 421
	-	4	23,01	61,86	492 530	24,78	58,11	5 121 617
pt-es	*MGIZA++*		*38,22*	*47,47*	*17 828 592*			
	Anymalign	1	34,63	50,25	1 532 520	34,84	50,35	6 730 554
	-	2	**36,03**	**49,63**	987 884	36,72	49,10	7 295 581
	-	3	35,72	49,95	744 947	35,98	49,02	6 126 896
	-	4	35,18	50,34	342 168	**37,01**	**48,71**	3 926 578

TAB. 6 – Résultats des tâches de traduction sur Europarl.

fragments. Le tableau 4 présente les caractéristiques des données utilisées pour chacune de ces expériences, et les tableaux 5 et 6 présentent les résultats.

Les lignes où *n* max. = 1 correspondent à la version d'origine d'Anymalign. Comme décrit précédemment (section 2.3), Anymalign étant *anytime*, la condition d'arrêt que nous lui imposons dépend du temps d'exécution de MGIZA++. Ce temps est constant quelle que soit la valeur de *n* max. Le temps de traitement augmentant avec ce paramètre, plus ce paramètre est élevé et plus le nombre de sous-corpus traités est *faible*, contrairement aux expériences présentées dans la section 4.2 où l'ensemble des sous-corpus à traiter était fixé à l'avance, impliquant un temps de traitement dépendant de *n* max. Théoriquement, les tables produites pour un *n* max. donné sont plus grandes que pour un *n* max. inférieur, à condition que l'aligneur soit exécuté suffisamment longtemps. Cela explique pourquoi les tables de traductions des tableaux 5 et 6 peuvent contenir moins d'entrées pour de plus grandes valeurs de *n* max. En pratique, ces tables contiennent tout de même davantage de longs n-grammes, ce qui permet une amélioration très significative des scores, malgré une table de traductions plus petite.

Sur les tâches impliquant le BTEC, les lignes où *n* max. = 1 montrent que la version d'origine d'Anymalign obtient des scores BLEU comparables à MGIZA++ en chinois-anglais, et est loin derrière en arabe-anglais. La généralisation aux n-grammes lui permet de devancer MGIZA++ de plus d'un point BLEU en chinois-anglais, et de l'égaliser en arabe-anglais, soit un gain spectaculaire de 7 points BLEU.

Sur les tâches impliquant Europarl, les scores de la version d'origine d'Anymalign sont en retrait de façon significative par rapport à MGIZA++, ce qui est conforme aux expériences que nous avions menées précédemment. Cela dit, la différence n'était pas aussi prononcée dans nos anciennes expériences : nous observions une différence de 2 à 3 points BLEU en moyenne, alors qu'elle est ici de 6 points. Nous pensons que ce changement est dû à la taille de notre corpus qui est désormais beaucoup plus élevé : 320 000 couples de phrases contre 100 000 précédemment. La taille des tables de traductions d'Anymalign, très petites par rapport à celles de MGIZA++, semble indiquer que le temps d'exécution de notre méthode n'est pas suffisant. Pour cette raison, le tableau 6 contient dans sa partie droite une deuxième série de résultats, qui correspondent à l'exécution d'Anymalign pendant une durée totale égale à 20 fois le temps d'exécution de MGIZA++. En pratique, Anymalign étant massivement parallélisable, nous avons découpé les traitements en 140 processus et les avons exécutés sur un cluster, pour finalement profiter d'un temps de traitement 7 fois plus rapide qu'avec les résultats présentés dans la partie gauche du tableau. Les tailles des tables de traductions dans la partie droite du tableau sont plus proches de celles de MGIZA++, ce qui confirme que le temps d'exécution n'était pas suffisant[6], mais le gain en BLEU de la version d'origine d'Anymalign n'est pas significatif pour autant. Il l'est par contre lorsque nous augmentons *n* max. : nous gagnons jusqu'à 3 points BLEU en finnois-anglais (*n* max. = 3) simplement en exécutant Anymalign plus longtemps. Dans tous les cas de la partie droite du tableau, l'indexation des n-grammes permet un gain en BLEU allant d'1,7 point en français-anglais à près de 4 points en finnois-anglais. En moyenne, les meilleurs scores d'Anymalign sont désormais en retrait de 3,5 points BLEU par rapport à MGIZA++, divisant pratiquement par deux son retard initial.

5 Conclusion

Cet article a présenté une généralisation de notre méthode d'alignement sous-phrastique afin d'améliorer ses résultats en traduction automatique. La méthode d'origine obtient de meilleurs résultats que l'état de l'art sur des tâches de constitution de lexiques bilingues, mais des résultats inférieurs en traduction automatique statistique par fragments. Nous avons montré que ces différences ont principalement deux causes : les différences de fréquences des mots qui composent les séquences à aligner (cause propre à la méthode), et les fréquences de mots utiles à ces tâches (cause propre à la tâche). Pour pallier le premier problème, nous avons proposé une généralisation de la phase d'indexation de notre méthode, en ne considérant non plus le mot comme unité, mais le n-gramme. Le résultat de cette généralisation est un fort accroissement du nombre de n-grammes en sortie, qui mène à des gains très significatifs en traduction automatique par fragments (jusqu'à +7 points BLEU sur le couple arabe-anglais). Notre méthode fait désormais jeu égal avec l'état de l'art sur des tâches « simples » de traduction automatique (BTEC), et nous avons pratiquement divisé son retard par deux sur des tâches plus difficiles (Europarl). Pour aller plus loin, nous envisageons d'étudier le cas de l'alignement des mots fréquents, dont nous avons montré qu'ils étaient moins bien alignés que les mots rares par notre méthode, ainsi que la question de sa condition d'arrêt.

[6]Cela soulève une autre question, qui est celle de la condition d'arrêt d'Anymalign. Les présentes expériences montrent que nos critères actuels sont insuffisants, ne serait-ce que pour effectuer une juste comparaison avec d'autres outils.

Remerciements

Les travaux présentés dans cet article ont été partiellement financés par le projet Cap Digital SAMAR.

Références

BROWN P., COCKE J., DELLA PIETRA S., DELLA PIETRA V., JELINEK F., MERCER R. & ROOSSIN P. (1988). A Statistical Approach to Language Translation. In *Proceedings of Coling'88*, p. 71–76, Budapest.

BROWN P., DELLA PIETRA S., DELLA PIETRA V. & MERCER R. (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, **19**(2), 263–311.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

FORDYCE C. S. (2007). Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of IWSLT 2007*, p. 1–12, Trente.

FUNG P. & CHURCH K. (1994). K-vec : A New Approach for Aligning Parallel Texts. In *Proceedings of Coling'94*, volume 2, p. 1096–1102, Kyōto.

GALE W. & CHURCH K. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the fourth DARPA workshop on Speech and Natural Language*, p. 152–157, Pacific Grove.

GAO Q. & VOGEL S. (2008). Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, p. 49–57, Columbus (Ohio, USA).

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, p. 79–86, Phuket.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, p. 177–180, Prague.

KOEHN P., OCH F. & MARCU D. (2003). Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003*, p. 48–54, Edmonton.

LARDILLEUX A. (2010). *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. PhD thesis, université de Caen Basse-Normandie. 204 pages.

LARDILLEUX A., CHEVELU J., LEPAGE Y., PUTOIS G. & GOSME J. (2009). Lexicons or phrase tables ? An investigation in sampling-based multilingual alignment. In *Proceedings of EBMT3*, p. 45–52, Dublin.

LARDILLEUX A. & LEPAGE Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. In *Proceedings of AMTA 2008*, p. 125–132, Waikiki.

LARDILLEUX A. & LEPAGE Y. (2009). Sampling-based multilingual alignment. In *Proceedings of RANLP 2009*, p. 214–218, Borovets.

MANDELBROT B. (1954). Structure formelle des textes et communication. *Word*, **10**, 1–27.

MELAMED D. (2000). Models of Translational Equivalence among Words. *Computational Linguistics*, **26**(2), 221–249.

MONTEMURRO M. (2004). A generalization of the Zipf-Mandelbrot Law in Linguistics. *Nonextensive Entropy : interdisciplinary applications*. 12 pages.

MOORE R. (2005). Association-Based Bilingual Word Alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, p. 1–8, Ann Arbor.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, p. 311–318, Philadelphie.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*, p. 223–231, Cambridge.

TAKEZAWA T., SUMITA E., SUGAYA F., YAMAMOTO H. & YAMAMOTO S. (2002). Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. In *Proceedings of LREC 2002*, p. 147–152, Las Palmas de Gran Canaria.

VERGNE J. (2009). Defining the chunk as the period of the functions length and frequency of words on the syntagmatic axis. In *Proceedings of LTC'09*, p. 85–89, Poznań.

ZIPF G. (1965). *The Psycho-Biology of Language : An Introduction to Dynamic Philology*. Classic Series. Cambridge, USA : The MIT Press. Fist edition 1935.