## RÉSUMÉ

L'objectif de cette thèse est de montrer que, contrairement aux idées reçues, les mots de basses fréquences peuvent être mis à profit de façon efficace en traitement automatique des langues. Nous les mettons à contribution en alignement sous-phrastique, tâche qui constitue la première étape de la plupart des systèmes de traduction automatique fondée sur les données (traduction probabiliste ou par l'exemple). Nous montrons que les mots rares peuvent servir de fondement même dans la conception d'une méthode d'alignement sous-phrastique multilingue, à l'aide de techniques différentielles proches de celles utilisées en traduction automatique par l'exemple. Cette méthode est *réellement multilingue*, en ce sens qu'elle permet le traitement simultané d'un nombre quelconque de langues. Elle est de surcroît très simple, *anytime*, et permet un passage à l'échelle naturel. Nous comparons notre implémentation, *Anymalign*, à deux ténors statistiques du domaine sur des tâches bilingues. Bien qu'à l'heure actuelle ses résultats sont en moyenne légèrement en retrait par rapport à l'état de l'art en traduction automatique probabiliste par segments, nous montrons que la qualité propre des lexiques produits par notre méthode est en fait supérieure à celle de l'état de l'art.



MOTS-CLÉS : traitement automatique des langues, hapax, multilinguisme, traduction automatique, alignement, événements rares.



## ABSTRACT

*The contribution of low frequencies to multilingual sub-sentential alignment: a differential approach*

The goal of this thesis dissertation is to show that, contrary to preconceived ideas, one can efficiently take advantage of low frequency words in natural language processing. We put them to use in sub-sentential alignment, which constitutes the first step of most data-driven machine translation systems (statistical or example-based machine translation). We show that rare words can be used as a foundation in the design of a multilingual sub-sentential alignment method, using differential techniques similar to those found in example-based machine translation. This method is *truly multilingual*, in that it allows the simultaneous processing of any number of languages. Moreover, it is very simple, *anytime*, and scales up naturally. We compare our implementation, *Anymalign*, to two statistical tools proven in the domain. Although its current results are in average slightly behind those of state of the art methods in phrase-based statistical machine translation, we show that the intrinsic quality of our lexicons is actually superior to that of lexicons produced by state of the art methods.



KEY WORDS: natural language processing, hapax legomenon, multilingualism, machine translation, alignment, rare events.



Laboratoire GREYC – Campus Côte de Nacre, Boulevard du Maréchal Juin – BP 5186 – 14032 CAEN CEDEX



UNIVERSITÉ DE CAEN BASSE-NORMANDIE

U.F.R. DE SCIENCES

ÉCOLE DOCTORALE
STRUCTURE, INFORMATION, MATIÈRE ET MATÉRIAUX

THÈSE

présentée par


M. ADRIEN LARDILLEUX


et soutenue

le 14 septembre 2010

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE CAEN

Spécialité : informatique et applications

*Arrêté du 7 août 2006*

# CONTRIBUTION DES BASSES FRÉQUENCES À L'ALIGNEMENT SOUS-PHRASTIQUE MULTILINGUE : UNE APPROCHE DIFFÉRENTIELLE

MEMBRES DU JURY

M. Christian BOITET, professeur, université de Grenoble *(rapporteur, excusé)*

M. Philippe LANGLAIS, professeur agrégé, université de Montréal *(rapporteur)*

M. François YVON, professeur, université Paris-Sud XI *(rapporteur)*

M$^{me}$ Béatrice DAILLE, professeure, université de Nantes

M. Jacques VERGNE, professeur, université de Caen

M. Andy WAY, professeur associé, Dublin City University

M. Yves LEPAGE, professeur, université de Caen / université Waseda *(directeur)*

Adrien Lardilleux : *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle,* thèse de doctorat, © 2010. Version finale.

Cette thèse a été composée avec LaTeX 2ε en utilisant le style `classicthesis`, disponible via CTAN. La police principale est *Minion*® d'Adobe™. L'arabe a été composé avec ArabTeX et le japonais avec CJK LaTeX. Graphiques par `Gnuplot`.

## PUBLICATIONS

Un certain nombre d'idées et de résultats présentés dans cette thèse ont déjà été publiés dans les articles suivants :

Adrien Lardilleux, Julien Gosme et Yves Lepage : Bilingual Lexicon Induction : Effortless Evaluation of Word Alignment Tools and Production of Resources for Improbable Language Pairs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 252–256, La Valette, mai 2010. URL `http://hal.archives-ouvertes.fr/hal-00488768/fr/`

Adrien Lardilleux : L'alignement sous-phrastique multilingue pour les nuls. In *Actes de la 7e Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication (MajecSTIC 2009)*, Avignon, novembre 2009. URL `http://hal.archives-ouvertes.fr/hal-00439810/fr/`

Adrien Lardilleux, Jonathan Chevelu, Yves Lepage, Ghislain Putois et Julien Gosme : Lexicons or phrase tables ? An investigation in sampling-based multilingual alignment. In *Proceedings of the 3rd Workshop on Example-Based Machine Translation (EBMT3)*, pages 45–52, Dublin, novembre 2009. URL `http://hal.archives-ouvertes.fr/hal-00439806/fr/`

Adrien Lardilleux et Yves Lepage : Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, septembre 2009. URL `http://hal.archives-ouvertes.fr/hal-00439789/fr/`

Adrien Lardilleux et Yves Lepage : Hapax Legomena : Their Contribution in Number and Efficiency to Word Alignment. In Zygmunt

Vetulani et Hans Uszkoreit, éditeurs : *Human Language Technology. Challenges of the Information Society*, volume 5603 de *Lecture Notes in Computer Science*, pages 440–450. Springer Heidelberg, août 2009. URL http://www.springerlink.com/content/d1t04777340u5548/

Adrien Lardilleux et Yves Lepage : Anymalign : un outil d'alignement sous-phrastique libre pour les êtres humains. In *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009) : démonstrations*, Senlis, juin 2009. URL http://hal.archives-ouvertes.fr/hal-00488772/fr/

Adrien Lardilleux et Yves Lepage : A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, pages 125–132, Waikiki, octobre 2008. URL http://hal.archives-ouvertes.fr/hal-00368737/fr/

Adrien Lardilleux et Yves Lepage : Multilingual Alignments by Monolingual String Differences. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling'08)*, pages 55–58, Manchester, août 2008. URL http://www.aclweb.org/anthology/C08-2014

Adrien Lardilleux et Yves Lepage : The contribution of the notion of hapax legomena to word alignment. In *Proceedings of the 3rd Language & Technology Conference (LTC'07)*, pages 458–462, Poznań, octobre 2007. URL http://hal.archives-ouvertes.fr/hal-00252026/fr/





## LISTE DES TABLEAUX

## MERCIS

À Yves Lepage : merci, pour tout. Si je ne commence pas à développer les qualités, d'encadrement ou autres, dont mon directeur de thèse déborde, ce n'est pas parce que l'envie m'en manque. C'est une question de place.

Merci. À très bientôt.

Je remercie Christian Boitet, Philippe Langlais et François Yvon d'avoir bien voulu rapporter sur cette thèse. Merci également à Béatrice Daille, Jacques Vergne et Andy Way d'avoir accepté de faire partie du jury.

Merci aussi à toute l'équipe ISLanD, en particulier Nadine Lucas, Emmanuel Giguet et Jacques Vergne, pour leurs retours réguliers tout au long de ces trois années de thèse. Mention spéciale à Nadine et Emmanuel pour avoir relu le présent document dans des délais particulièrement courts.

Merci enfin à Julien et Julien, Lebranchu et Gosme, mes deux acolytes.

# TABLE DES FIGURES

## SOMMAIRE

# INTRODUCTION

« *More data is better data.*
*— Definitely. What's the limit ?* »

Personne n'a jamais répondu cela. Nous non plus. Bien mal nous en aurait pris : à l'heure où des quantités de données astronomiques sont accessibles et affluent continuellement en masse sur la Toile, il y flotte depuis un certain nombre d'années comme un doux parfum d'infini. En informatique, cela fait naturellement l'affaire des adeptes des procédés fondés sur les données, qui disposent de toujours plus de matière plus ou moins première. Bien entendu, on délègue : c'est la machine qui se charge du traitement de ces masses de données fabuleuses, tâche ingrate s'il en est. Ordinateurs, réseaux et espaces de stockage ont eux aussi progressé, permettant la maîtrise de cet afflux de données virtuelles ; et jusqu'à il y a un certain temps, ces progrès étaient à l'image du déluge de données : exponentiels. Des limites physiques ont cependant été atteintes depuis, et la tendance est maintenant au parallélisme : pour traiter plus de données, il faut plus de ressources physiques. Ce n'est plus à la portée de tout le monde.

Ces considérations nous concernent directement. Nos travaux se rapportent à une tâche fondée sur les données : l'alignement de textes. Elle est liée à la traduction automatique, une des branches les plus — tristement ? — connues du traitement automatique des langues. Dans ce domaine, non seulement les ressources physiques nécessaires au traitement de quantités de données colossales ne sont plus à la portée de tout le monde, mais surtout on ressent un effet de plafonnement : on obtient aujourd'hui moins de résultats qu'on n'ajoute de données. Un virage a donc été entamé : il est désormais plus sage de procéder autrement, en se contentant des données disponibles. Faire plus avec moins : c'est ce que nous proposons dans cette thèse placée sous le signe du recyclage des mots rares. Nous nous intéressons pour cela à

une ressource au potentiel sous-estimé que sont les termes de basses fréquences, et que trop de praticiens excluent de leurs traitements car ils les considèrent statistiquement non significatifs et donc non pertinents. À notre connaissance, nous sommes les premiers à proposer d'utiliser les termes de basses fréquences comme *fondement* de l'alignement. Nos travaux nous ont conduit incidemment à étudier le cas du multilinguisme, c'est-à-dire aux moyens d'aligner un nombre quelconque de langues simultanément grâce aux mots rares. C'est donc à notre connaissance la première fois qu'une méthode *réellement multilingue* entièrement automatisée est proposée pour l'alignement. Une bonne part de nos travaux est donc du défrichage.

La première partie de ce document pose les bases nécessaires à l'élaboration de notre méthode d'alignement, en mettant l'accent sur l'analyse des basses fréquences. La deuxième partie met ces observations à profit et fait la part belle aux expériences et à la pratique. Ces deux parties couvrent en tout six chapitres :

- le chapitre 1 pose les faits : présentation de l'alignement, pistes de travail envisagées, état des basses fréquences en corpus ;
- les basses fréquences sont à l'honneur dans le chapitre 2, qui montre, à travers une expérience d'alignement préliminaire, que la pratique consistant à les rejeter constitue une faute en alignement ;
- le chapitre 3 va plus loin en montrant comment les basses fréquences peuvent être mises à profit pour aligner ;
- nous présentons alors dans le chapitre 4 notre méthode d'alignement et son implémentation : *Anymalign* ;
- cette méthode est passée au crible dans le chapitre 5 ;
- et elle est généralisée au cas du multilinguisme véritable dans le chapitre 6.

Nous concluons par une mise en perspective de nos travaux.

## TABLE DES MATIÈRES

Première partie

# OBSERVATIONS, ANALYSES

*thods in Natural Language Processing (EMNLP 2008)*, pages 572–581, Waikiki, octobre 2008. URL `http://www.mt-archive.info/EMNLP-2008-Zhao.pdf`. (Cité à la page 8.)

George Zipf : *Human Behavior and the Principle of Least Effort : An Introduction to Human Ecology*. Hafner, New York, 1949. 573 pages. (Cité à la page 21.)

Total : 130 références bibliographiques.
La validité de toutes les URL indiquées a été vérifiée au 1er juin 2010.

Stephan Vogel, Hermann Ney et Christoph Tillman : HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, pages 836–841, Copenhague, août 1996. URL http://aclweb.org/anthology-new/C/C96/C96-2141.pdf. (Cité aux pages 12 et 92.)

Dekai Wu et Xuanyin Xia : Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA 1994)*, pages 206–213, Columbia (Maryland, États-Unis), octobre 1994. URL http://www.mt-archive.info/AMTA-1994-Wu.pdf. (Cité à la page 8.)

Hua Wu et Ming Zhou : Synonymous Collocation Extraction Using Translation Information. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pages 120–127, Sapporo, juillet 2003. URL http://www.aclweb.org/anthology/P03-1016. (Cité à la page 9.)

Hao Zhang, Daniel Gildea et David Chiang : Extracting Synchronous Grammar Rules From Word-Level Alignments in Linear Time. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1081–1088, Manchester, août 2008. URL http://www.aclweb.org/anthology/C08-1136. (Cité à la page 8.)

Ying Zhang, Stephan Vogel et Alex Waibel : Interpreting BLEU/-NIST Scores : How Much Improvement Do We Need to Have a Better System ? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 2051–2054, Lisbonne, mai 2004. URL http://www.mt-archive.info/LREC-2004-Zhang.pdf. (Cité à la page 61.)

Bing Zhao et Yaser Al-Onaizan : Generalizing local and non-local word-reordering patterns for syntax-based machine translation. In *Proceedings of the 2008 Conference on Empirical Me-*

# 1

# ÉTAT DES LIEUX

Ce chapitre présente une vue d'ensemble des faits touchant à l'alignement sous-phrastique. Notre approche se plaçant en-dehors du courant dominant, nous nous efforçons de donner une vue synthétique de la tâche afin de bien situer nos travaux par rapport aux approches courantes du domaine. Une présentation détaillée et récente de la notion d'alignement est disponible dans la thèse de Cromières (2010, chap. 1). Nous concluons ce chapitre en introduisant quelques faits relatifs aux mots rares, qui constituent la base de nos travaux.

SOMMAIRE

## 1.1 PRÉSENTATION DE L'ALIGNEMENT SOUS-PHRASTIQUE

### 1.1.1 *But de l'alignement sous-phrastique*

> *Laying aside for the time being the desirability of (idiomatic) word cluster - to - word cluster translation, what we are after at first is to find for each word $f$ in the (French) source language the list of words $\{e_1, e_2, \ldots, e_i\}$ of the (English) target language into which $f$ can translate, and the probability $P(e_i|f)$ that such a translation takes place.* (Brown et coll., 1988, section 3 : « Creating the Glossary, First Attempt »)

Cet énoncé est issu de la première esquisse de traduction automatique « tout statistique » de Brown et coll.. Les auteurs proposaient une nouvelle approche, qualifiée depuis d'*empirique*, à la traduction automatique, où les connaissances nécessaires à la traduction étaient acquises automatiquement à partir de grandes quantités de textes traductions les uns des autres, en lieu et place des approches traditionnelles linguistiques, qualifiées d'*expertes*. L'approche statistique, ou plutôt probabiliste, est de nos jours celle qui connaît la plus grande activité en recherche, la tendance actuelle étant néanmoins à l'intégration de connaissances linguistiques dans ses modèles.

Un des composants des systèmes de traduction automatique consistait — et consiste généralement toujours, sous d'autres noms — en un « glossaire », contenant des correspondances de traductions entre mots ou séquences de mots, telles que *word* ↔ *mot*, *word* ↔ *propos*, *not* ↔ *ne … pas*, *no* ↔ *ne … pas*, etc. (Brown et coll., 1988). C'est à la création d'un tel glossaire que fait référence l'énoncé cité précédemment. L'ambition de Brown et coll. était non seulement d'en constituer automatiquement, mais aussi, à plus long terme, de profiter du fait que les données utilisées étaient constituées de traductions réelles pour apprendre des traductions d'expressions idiomatiques, qui auraient été autrement plus difficiles à rassembler par des moyens humains. À chacune des entrées de ces glossaires serait désormais attribuée une



*(Coling'94)*, pages 297–303, Kyōto, août 1994. URL `http://aclweb.org/anthology-new/C/C94/C94-1048.pdf`. (Cité à la page 98.)

Jörg TIEDEMANN : Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 339–346, Budapest, avril 2003. URL `http://aclweb.org/anthology-new/E/E03/E03-1026.pdf`. (Cité à la page 14.)

Jörg TIEDEMANN : News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing*, 5:237–248, 2009. URL `http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf`. (Cité aux pages 9 et 130.)

Dan TUFIŞ et Ana-Maria BARBU : Lexical token alignment : experiments, results and applications. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 458–465, Las Palmas de Gran Canaria, 2002. URL `http://www.racai.ro/~tufis/papers/Tufis-Barbu-LREC2002.pdf`. (Cité aux pages 13 et 14.)

Peter TURNEY et Michael LITTMAN : Corpus-based Learning of Analogies and Semantic Relations. *Machine Learning*, 60:251–278, septembre 2005. URL `http://arxiv.org/abs/cs/0508103`. (Cité à la page 31.)

Jacques VERGNE : Defining the chunk as the period of the functions length and frequency of words on the syntagmatic axis. In *Proceedings of the 4th Language & Technology Conference (LTC'09)*, pages 85–89, Poznań, novembre 2009. (Cité aux pages 56 et 116.)

David VILAR, Maja POPOVIC et Hermann NEY : AER : Do we need to ''improve'' our alignments ? In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT 2006)*, pages 205–212, Kyōto, décembre 2006. URL `http://www.mt-archive.info/IWSLT-2006-Vilar.pdf`. (Cité aux pages 16 et 97.)

URL http://personalpages.manchester.ac.uk/staff/harold.somers/Carl-Way-ch4.doc. (Cité à la page 157.)

Ankit Srivastava, Sergio Penkale, Declan Groves et John Tinsley : Evaluating syntax-driven approaches to phrase extraction for MT. In *Proceedings of the 3rd Workshop on Example-Based Machine Translation (EBMT3)*, pages 19–28, Dublin, novembre 2009. URL http://www.mt-archive.info/EBMT-2009-Srivastava.pdf. (Cité à la page 111.)

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş et Daniel Varga : The JRC-Acquis : A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147, Gênes, mai 2006. URL http://www.mt-archive.info/LREC-2006-Steinberger.pdf. (Cité aux pages 130, 132 et 141.)

Nicolas Stroppa et Andy Way : MaTrEx : DCU machine translation system for IWSLT 2006. In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT 2006)*, pages 31–36, Kyōtō, novembre 2006. URL http://www.mt-archive.info/IWSLT-2006-Stroppa.pdf. (Cité à la page 23.)

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto et Seiichi Yamamoto : Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 147–152, Las Palmas de Gran Canaria, 2002. URL http://gandalf.aksis.uib.no/lrec2002/pdf/305.pdf. (Cité aux pages 20, 100, 130, 139, 147, 151 et 194.)

Kumiko Tanaka et Kyoji Umemura : Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of the 15th International Conference on Computational Linguistics*



| $f_i$ | $\mathrm{P}(f_i\|e)$ |
|---|---|
| les | 0,267 |
| gens | 0,244 |
| personnes | 0,100 |
| population | 0,055 |
| peuple | 0,035 |
| ⋮ | ⋮ |

(a) Traductions candidates de $e$ = people.

| $e_i$ | $\mathrm{P}(e_i\|f)$ |
|---|---|
| people | 0,781 |
| they | 0,013 |
| those | 0,009 |
| individuals | 0,008 |
| persons | 0,005 |
| ⋮ | ⋮ |

(b) Traductions candidates de $f$ = gens.

Tableau 1 – Exemples d'entrées de glossaires constitués automatiquement à des fins de traduction automatique. Étant donné un mot source, les meilleures traductions, au sens des probabilités, sont présentées. La somme de toutes les probabilités associées aux traductions d'un mot source donné vaut un. Ces exemples sont ceux de Brown et coll. (1988).

probabilité de traduction, indiquant les chances qu'un mot dans une langue source se traduise en un mot donné dans une langue cible. Le tableau 1 ci-dessus en donne deux exemples.

On appellera *alignement* la tâche qui consiste à extraire des traductions de segments textuels à partir de textes traductions les uns des autres. Lorsque ces textes ont préalablement été mis en correspondance à un grain inférieur à celui du texte, typiquement la phrase, et que l'on cherche à extraire des traductions de grain encore inférieur, à savoir tout grain compris entre le caractère et la phrase, on parle d'*alignement sous-phrastique*. C'est sur cette tâche que se concentre cette thèse.

### 1.1.2 *Quelques applications*

L'alignement sous-phrastique a trouvé place dans de nombreuses applications depuis son introduction.

L'application majeure de l'alignement sous-phrastique demeure la traduction automatique. L'alignement constitue le cœur des systèmes de traduction probabiliste depuis les célèbres modèles IBM (Brown et coll., 1993). Le « glossaire » précédemment évoqué, également appelé « dictionnaire probabiliste », et dont les entrées ne contenaient à l'origine que des mots isolés, a depuis été remplacé par des *tables de traductions* dans la terminologie en cours. Celles-ci contiennent des suites de mots, ou segments (*phrases* en anglais) (Koehn et coll., 2003), et prennent récemment la forme de grammaires constituées de règles de dérivations (Chiang, 2007; Zhang et coll., 2008), qui, malgré leur nom, ne sont pas fondées sur des critères linguistiques (Zhao et Al-Onaizan, 2008). Toujours en traduction, l'alignement a trouvé également des applications avec les mémoires de traductions de seconde génération, qui opèrent au niveau sous-phrastique (Planas, 2000), ou les concordanciers (p. ex. Huet et coll., 2009) mettant en évidence les segments textuels traductions les uns des autres au moyen d'une interface adaptée (p. ex. Al-Adhaileh, 2003; Chenon, 2005; Germann, 2008).

Parallèlement à cela, une application des plus naturelles, qui a été engagée dès ses débuts (Brown et coll., 1988), concerne la constitution de ressources dictionnairiques. En effet, le glossaire automatiquement induit pour la traduction automatique probabiliste peut aisément être transformé en ressource utile pour de nombreux systèmes, pour peu que ne soient sélectionnées que les traductions les plus probables de chaque terme. Cela permet la constitution automatique de dictionnaires terminologiques dont le domaine est directement lié aux corpus parallèles dont ils sont extraits. Wu et Xia (1994), et plus récemment Lin et coll. (2008), ont par exemple utilisé l'alignement à cette fin. Cela a en outre permis des applications en lexicographie (p. ex. Klavans et Tzoukermann, 1990).



Yusuke Shinyama et Satoshi Sekine : Named Entity Discovery Using Comparable News Articles. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling'04)*, pages 848–853, Genève, août 2004. URL `http://aclweb.org/anthology-new/C/C04/C04-1122.pdf`. (Cité à la page 31.)

Michel Simard : Text-translation Alignment : Three Languages Are Better Than Two. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 2–11, College Park, 1999. URL `http://acl.ldc.upenn.edu/W/W99/W99-0602.pdf`. (Cité aux pages 117, 118, 129 et 131.)

John Sinclair et J. Ball : Preliminary Recommendations on Text Typology. Rapport technique, Expert Advisory Group on Language Engineering Standards (EAGLE), juin 1996. URL `http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/texttyp.ps.gz`. 36 pages. (Cité à la page 60.)

Frank Smadja, Vasileios Hatzivassiloglou et Kathleen McKeown : Translating Collocations for Bilingual Lexicons : A Statistical Approach. *Computational Linguistics*, 22(1):1–38, mars 1996. URL `http://aclweb.org/anthology-new/J/J96/J96-1001.pdf`. (Cité à la page 13.)

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla et John Makhoul : A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas (AMTA 2006)*, pages 223–231, Cambridge, août 2006. URL `http://www.mt-archive.info/AMTA-2006-Snover.pdf`. (Cité à la page 96.)

Harold Somers : An overview of EBMT. In Michael Carl et Andy Way, éditeurs : *Recent advances in Example-Based Machine Translation*, pages 3–57. Kluwer Academic Publishers, Dordrecht, 2003.

http://www.numdam.org/item?id=MSH_1973__44__41_0. (Cité à la page 21.)

Aaron Phillips et Ralf Brown : Cunei Machine Translation Platform : System Description. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation (EBMT3)*, pages 29–36, Dublin, 2009. URL http://www.mt-archive.info/EBMT-2009-Phillips.pdf. (Cité à la page 94.)

Emmanuel Planas : Extending Translation Memories. In *Proceedings of the fifth workshop of the European Association for Machine Translation (EAMT 2000)*, Laybach, mai 2000. URL http://nl.ijs.si/eamt00/proc/Planas.pdf. 13 pages. (Cité à la page 8.)

Philip Resnik, Mari Broman Olsen et Mona Diab : The Bible as a Parallel Corpus : Annotating the ''Book of 2000 Tongues''. *Computers and the Humanities*, 23(1-2):129–153, 1999. URL http://www.springerlink.com/content/u240g32544t26777/. (Cité aux pages 20, 100, 130, 140 et 148.)

Manabu Sasayama, Fuji Ren et Shingo Kuroiwa : Automatic Extraction of Super-Function From Bilingual Corpus. *Electronic Notes in Theoretical Computer Science*, 225:329–340, janvier 2009. URL http://linkinghub.elsevier.com/retrieve/pii/S1571066108005471. (Cité à la page 15.)

Satoshi Sato : *Example-based Machine Translation*. Thèse de doctorat, université de Kyōto, septembre 1991. 120 pages. (Cité aux pages 119 et 157.)

Bettina Schrader : How does morphological complexity translate ? A cross-linguistic case study for word alignment. In *International Conference on Linguistic Evidence 2006*, Tubingue, février 2006. URL http://www.sfb441.uni-tuebingen.de/LingEvid2006/abstracts/schrader.pdf. 3 pages. (Cité à la page 24.)



L'alignement sous-phrastique est également mis à profit de façon indirecte dans d'autres applications ayant trait au traitement automatique des langues. Il a par exemple été utilisé en désambiguïsation de termes, où l'hypothèse principale réside dans le fait qu'un terme ambigu dans une langue source se traduira probablement de différentes façons dans des langues cibles différentes (p. ex. Brown et coll., 1991a; Ng et coll., 2003; Crego et coll., 2009). De façon similaire, l'alignement permet la recherche de synonymes : si deux termes issus d'une même langue se traduisent par un même terme dans une autre langue, alors ces deux termes ont de grandes chances d'avoir des sens très proches (p. ex. Wu et Zhou, 2003; Manguin et coll., 2007).

Cette liste d'applications n'est pas exhaustive, mais donne une idée du large éventail d'applications possibles de l'alignement. En ce qui nous concerne, nous privilégierons les tâches de production de tables de traductions pour la traduction automatique empirique — ou *fondée sur les données* — et l'induction de dictionnaires. C'est au travers de ces tâches que nous évaluerons les systèmes d'alignement (chapitre 5).

### 1.1.3 *Notre unique matériau : des textes parallèles*

Un des avantages avancés par les praticiens de la traduction automatique empirique est leur moindre coût *a priori* — voir à ce sujet les contre-arguments de Boitet (2008) — : l'unique source de connaissance nécessaire à la création d'un tel système de traduction automatique consiste en un texte source et sa traduction dans une langue cible. On appelle de tels textes des textes parallèles. On peut trouver des textes traductions les uns des autres assez facilement pour certains couples de langues de nos jours ; mentionnons, pour n'en citer qu'un, le corpus OPUS (Tiedemann, 2009), consistant en une large collection de traductions de textes issus du Web. Brown et coll. avaient pour leur part fondé leurs premières expériences sur les débats du parlement canadien, disponibles en français-anglais. La figure 1 page suivante donne un exemple de texte parallèle.

| | |
|---|---|
| ⋮ | ⋮ |
| Du point de vue du Règlement, ces amendements sont parfaitement réglementaires. Par conséquent, nous passons au vote de la proposition de règlement. (Le Parlement approuve la proposition de la Commission) | In accordance with the Rules of Procedure, they are perfectly permissible. Therefore, we shall now proceed to the vote on the proposed regulation. (Parliament approved the Commission's proposal) |
| ⋮ | ⋮ |

FIGURE 1 – Extrait de la partie français-anglais du corpus parallèle Europarl (Koehn, 2005). Les deux textes sont traductions l'un de l'autre.

La tâche d'alignement sous-phrastique présuppose que les textes parallèles aient préalablement été mis en correspondance au grain de la phrase. Lorsque cette mise en correspondance n'a pas été réalisée dès la production du corpus, une phase d'alignement des phrases est nécessaire. Cette étape est résolue de longue date, des heuristiques simples basées sur la longueur des phrases (Gale et Church, 1991a), utilisant éventuellement des points d'ancrage (Brown et coll., 1991b) ou un lexique construit à la volée (Chen, 1993) ayant permis d'atteindre des taux de réussite avoisinant les 100 %. La figure 2 page ci-contre présente le résultat d'une telle étape. Le grain aligné en sortie n'est cependant pas systématiquement la phrase, car une phrase dans une langue source peut très bien se traduire par plusieurs phrases dans une langue cible. Le grain de base auquel nous ferons référence par la suite sera par conséquent l'*énoncé*, constitué éventuellement de plus d'une phrase.

Ces textes parallèles alignés au grain énoncé, que nous appellerons par la suite abusivement *corpus* parallèles, constituent donc notre unique matériau. Plus précisément, nous ferons le choix de nous maintenir dans un cadre applicatif où l'alignement sous-phrastique se voudra :



*18th International Conference on Computational Linguistics (Coling'00)*, pages 1086–1090, Sarrebruck, août 2000. URL `http://aclweb.org/anthology-new/C/C00/C00-2163.pdf`. (Cité aux pages 15 et 97.)

Franz OCH et Hermann NEY : Statistical multi-source translation. In *Proceedings of the 8th Machine Translation Summit (MT Summit VIII)*, pages 253–258, Saint-Jacques-de-Compostelle, septembre 2001. URL `http://www.mt-archive.info/MTS-2001-Och-1.pdf`. (Cité à la page 132.)

Franz OCH et Hermann NEY : A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, mars 2003. URL `http://www.aclweb.org/anthology/J/J03/J03-1002.pdf`. (Cité aux pages 12, 14, 30, 91 et 158.)

Kishore PAPINENI, Salim ROUKOS, Todd WARD et Wei-Jing ZHU : BLEU : a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphie, 2002. URL `http://www.aclweb.org/anthology/P02-1040`. (Cité à la page 95.)

Michael PAUL : Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT 2008)*, pages 1–17, Hawaï, octobre 2008. URL `http://www.mt-archive.info/IWSLT-2008-Paul.pdf`. (Cité aux pages 139 et 147.)

Michael PAUL : Overview of the IWSLT 2009 Evaluation Campaign. In *Proceedings of the 6th International Workshop on Spoken Language Translation (IWSLT 2009)*, pages 1–18, Tōkyō, décembre 2009. URL `http://www.mt-archive.info/IWSLT-2009-Paul.pdf`. (Cité à la page 139.)

Micheline PETRUSZEWYCZ : L'histoire de la loi d'Estoup-Zipf : documents. *Mathématiques et Sciences Humaines*, 44:41–56, 1973. URL

Yayoi NAKAMURA-DELLOYE : *Alignement automatique de textes parallèles français-japonais*. Thèse de doctorat, université Denis Diderot, Paris VII, décembre 2007. URL http://tel.archives-ouvertes.fr/tel-00259276/fr/. 580 pages. (Cité à la page 59.)

Luka NERIMA et Eric WEHRLI : Generating Bilingual Dictionaries by Transitivity. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, pages 2584–2587, Marrakech, mai 2008. URL http://www.mt-archive.info/LREC-2008-Nerima.pdf. (Cité à la page 98.)

Hwee Tou NG, Bin WANG et Yee Seng CHAN : Exploiting Parallel Texts for Word Sense Disambiguation : An Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pages 455–462, Sapporo, juillet 2003. URL http://www.aclweb.org/anthology/P03-1058. (Cité aux pages 9 et 132.)

Vassilina NIKOULINA : *Modèle de traduction statistique à fragments enrichi par la syntaxe*. Thèse de doctorat, université Joseph Fourier, Grenoble, mars 2010. 145 pages. (Cité à la page 13.)

Eiji NISHIMOTO : Defining New Words in Corpus Data : Productivity of English Suffixes in the British National Corpus. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci 2004)*, Chicago, août 2004. URL http://www.cogsci.northwestern.edu/cogsci2004/papers/paper505.pdf. 6 pages. (Cité aux pages 24 et 25.)

Franz OCH : Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 160–167, Sapporo, juillet 2003. URL http://www.aclweb.org/anthology/P03-1021. (Cité à la page 95.)

Franz OCH et Hermann NEY : A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the*



| | |
|---|---|
| Du point de vue du Règlement, ces amendements sont parfaitement réglementaires. | In accordance with the Rules of Procedure, they are perfectly permissible. |
| Par conséquent, nous passons au vote de la proposition de règlement. | Therefore, we shall now proceed to the vote on the proposed regulation. |
| (Le Parlement approuve la proposition de la Commission) | (Parliament approved the Commission's proposal) |

FIGURE 2 – Le même texte parallèle qu'à la figure 1, après mise en correspondance au niveau des phrases. Ces traductions constituent la matière première de l'alignement sous-phrastique.

NON SUPERVISÉ : tout est automatique, l'utilisateur humain n'intervient pas dans le processus d'alignement ;

ENDOGÈNE : aucune connaissance autre que celle intrinsèquement contenue dans les textes parallèles n'est utilisée. Cela implique que nous ne travaillerons que sur des formes surfaciques.

## 1.2 PRINCIPALES APPROCHES

La littérature concernant l'alignement sous-phrastique est particulièrement importante. Nous en donnons ici les grandes lignes. La majorité des méthodes intègrent une part de statistiques et peuvent être réparties en deux catégories : l'approche « estimative », qui est celle qui a été introduite par Brown et coll. (1988), et l'approche « associative », introduite par Gale et Church (1991b).

### 1.2.1 *Approche estimative*

La première approche, dite « estimative », consiste à construire à partir des données un *modèle* du corpus parallèle bilingue de départ dont

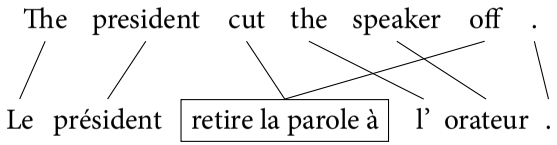



FIGURE 3 – Alignement entre un énoncé anglais et sa traduction française issus d'Europarl sous forme de liens entre mots.

les paramètres sont estimés en partant d'un ensemble d'hypothèses. Ce modèle doit permettre une maximisation globale de la relation de traduction dans son ensemble, en considérant non pas chacune des relations de traduction individuellement mais des ensembles de telles relations. Concrètement, pour chaque couple d'énoncés source-cible d'un corpus parallèle, on cherche à déterminer les meilleurs *liens* entre les mots de l'énoncé source et ceux de l'énoncé cible, comme l'illustre la figure 3 ci-dessus. Une table de traductions — ou dictionnaire probabiliste ou glossaire — peut ensuite être dérivée à partir de l'ensemble des couples d'énoncés du corpus d'entraînement et de ces liens.

Les plus connus de ces modèles sont les modèles IBM, proposés par Brown et coll. (1993). Ceux-ci sont au nombre de cinq, de complexité croissante, chacun introduisant des paramètres permettant d'affiner les résultats du précédent, tels la position absolue ou relative des mots, ou encore leur déplacement au cours du processus de traduction. Ces modèles constituent encore aujourd'hui, et ce depuis leur introduction il y a une vingtaine d'années, une référence en la matière. Bien que le domaine ait connu une grande activité aux cours de ces années, peu d'améliorations ont été définitivement adoptées par la communauté. Seules celles ayant été intégrées dans des solutions « clefs en main » libres d'utilisation sont en effet couramment employées ; citons par exemple le modèle caché de Markov de Vogel et coll. (1996), désormais couramment utilisé en lieu et place du modèle IBM2 dans le célèbre outil GIZA++ (Och et Ney, 2003).

L'approche estimative est intimement liée à la traduction automatique probabiliste dont elle constitue un pilier. Il est donc naturel que



*dings of the 20th International Conference on Computational Linguistics (Coling'04)*, pages 219–225, Genève, août 2004. URL `http://aclweb.org/anthology-new/C/C04/C04-1032.pdf`. (Cité à la page 13.)

Dan MELAMED : Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. In *Proceedings of the Third Workshop on Very Large Corpora (VLC'95)*, pages 184–198, Boston (Massachusetts, États-Unis), juin 1995. URL `http://aclweb.org/anthology-new/W/W95/W95-0115.pdf`. (Cité aux pages 13 et 14.)

Robert MOORE : Association-Based Bilingual Word Alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 1–8, Ann Arbor, juin 2005. URL `http://www.aclweb.org/anthology/W/W05/W05-0801.pdf`. (Cité aux pages 14 et 17.)

Robert MOORE : What Do Computational Linguists Need to Know about Linguistics ? In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, pages 41–42, Athènes, mars 2009. URL `http://www.aclweb.org/anthology/W09-0109`. (Cité à la page 17.)

Makoto NAGAO : A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, Lyon (France), 1984. URL `http://www.mt-archive.info/Nagao-1984.pdf`. (Cité aux pages 119 et 157.)

Makoto NAGAO et Shinsuke MORI : A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, pages 611–615, Kyōto, août 1994. URL `http://aclweb.org/anthology-new/C/C94/C94-1101.pdf`. (Cité à la page 33.)

Adam Lopez : Tera-Scale Translation Models via Pattern Matching. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 505–512, Manchester, août 2008. URL http://www.aclweb.org/anthology/C08-1064. (Cité à la page 18.)

Yanjun Ma, Nicolas Stroppa et Andy Way : Alignment-guided chunking. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2007)*, pages 114–121, Skövde, septembre 2007a. URL http://www.mt-archive.info/TMI-2007-Ma.pdf. (Cité à la page 56.)

Yanjun Ma, Nicolas Stroppa et Andy Way : Bootstrapping Word Alignment via Word Packing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 304–311, Prague, juin 2007b. URL http://www.aclweb.org/anthology/P07-1039. (Cité à la page 55.)

Udi Manber et Gene Myers : Suffix Array : A New Method for On-Line String Searches. *SIAM Journal on Computing*, 22:935–948, 1993. URL http://www.cs.arizona.edu/people/udi/suffx.ps. (Cité à la page 33.)

Jean-Luc Manguin, Jörg Tiedemann et Lonneke van der Plas : Extraction de synonymes à partir d'un corpus multilingue aligné. *Texte et corpus*, (3):151–161, septembre 2007. URL http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_vanderplas_tiedemann_manguin.pdf. (Cité à la page 9.)

Daniel Marcu et Daniel Wong : A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139, Philadelphie, juillet 2002. URL http://www.aclweb.org/anthology/W02-1018. (Cité à la page 13.)

Evgeny Matusov, Richard Zens et Hermann Ney : Symmetric Word Alignments for Statistical Machine Translation. In *Proceedings*



les deux domaines aient tendance à prendre les mêmes virages : on peut ainsi noter le parallèle entre l'apparition de la traduction automatique probabiliste par segments (Koehn et coll., 2003) et celle des modèles d'alignement justement fondés sur des segments (Marcu et Wong, 2002; Matusov et coll., 2004), tentant d'en finir avec la sempiternelle multiplicité 1-n caractérisant les modèles IBM et d'en réduire la complexité par la même occasion. De la même façon, l'hybridation des systèmes de traduction probabiliste avec les systèmes par transfert a vu naître des modèles d'alignement hybrides adaptés (Gildea, 2003). Plus récemment, des efforts ont été entrepris pour enrichir la traduction automatique probabiliste par des informations syntaxiques (Koehn et Hoang, 2007; Nikoulina, 2010), parallèlement à l'apparition de modèles d'alignement intégrant la syntaxe (DeNero et Klein, 2007).

### 1.2.2 *Approche associative*

La seconde approche, dite « associative », a été initialement proposée comme une alternative en réponse à l'apparente complexité de l'approche estimative. Alors que la première est plus orientée vers la traduction du fait de son intégration forte avec la traduction automatique probabiliste, la seconde se veut plus générale en ce sens qu'elle se focalise généralement sur la seule extraction de traductions. Elle permet éventuellement, en second lieu, de déduire des liens à la façon de la figure 3. Ces deux étapes sont inversées dans l'approche estimative. Le principe est de produire une liste de traductions candidates soumises à un test d'indépendance statistique. Celles dont la mesure d'associativité révèle leur dépendance seront considérées comme des traductions.

Parmi les mesures utilisées, on trouve par exemple l'information mutuelle (Fung et Church, 1994), le pourcentage de plus longue sous-séquence commune (Melamed, 1995), le coefficient de Dice (Smadja et coll., 1996), des mesures de log-vraisemblance (Tufiş et Barbu, 2002) ou encore le cosinus (Giguet et Luquet, 2006), sur lequel nous reviendrons dans le chapitre suivant. La méthode de Fung et Church

(1994) a ceci de particulier qu'elle ne nécessite même pas que les textes soient préalablement mis en correspondance au niveau des phrases. L'approche associative se veut d'autre part particulièrement bien adaptée à l'intégration de connaissances linguistiques, qu'il s'agisse d'heuristiques tels que les cognats (Melamed, 1995), ou au contraire d'un recours à de véritables étiqueteurs grammaticaux, analyseurs syntaxiques, ou détecteurs d'entités nommées (Tiedemann, 2003).

Contrairement à l'approche estimative qui consiste à résoudre un problème de maximisation globale, l'approche associative extrait les traductions candidates indépendamment les unes des autres, ce qui en ferait un processus de maximisation locale, ou glouton (Tufiş et Barbu, 2002). Tufiş et Barbu soulignent ainsi que la complexité des méthodes associatives est généralement quadratique en la taille du vocabulaire des textes d'entrée, tandis que les approches estimatives peuvent atteindre une complexité exponentielle. Elles se veulent par nature plus simples, voire *trop*, ce qui se répercuterait surtout sur la couverture du corpus de départ. Och et Ney (2003) avancent que les méthodes associatives, qu'ils qualifient d'« heuristiques », ne peuvent pas rivaliser avec les méthodes estimatives qui reposent sur « l'estimation de paramètres dans le cadre d'une théorie mathématique bien fondée » [1]. Des expériences réalisées par Moore (2005) ont pourtant montré qu'une position aussi catégorique n'était pas légitime, ce que confirment les résultats que nous présentons dans cette thèse.

Comme dit plus haut, la littérature en alignement sous-phrastique est considérable. Nous ne la détaillons pas davantage pour une raison simple : notre approche ne se rattache pas vraiment à l'existant. En fait, de la même façon que l'approche estimative se veut très proche de la traduction automatique probabiliste, notre approche se rapproche de *l'autre* traduction automatique empirique : la traduction par l'exemple. Elle n'en emprunte cependant que certaines techniques, et ne s'y intègre pas de façon aussi manifeste que l'alignement estimatif avec

1 « *the well-founded mathematical theory that underlies their parameter estimation* » (Och et Ney, 2003)



archives-ouvertes.fr/tel-00004372. 388 pages. (Cité à la page 117.)

Yves Lepage et Étienne Denoual : Purest ever example-based machine translation : Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282, 2005. ISSN 0922-6567. URL http://hal.archives-ouvertes.fr/hal-00260994/fr/. (Cité aux pages 59, 94, 117 et 157.)

Yves Lepage, Adrien Lardilleux et Julien Gosme : Commonality across vocabulary structures as an estimate of the proximity between languages. In *Proceedings of the 4th Language & Technology Conference (LTC'09)* , pages 457–461, Poznań, octobre 2009. URL http://hal.archives-ouvertes.fr/hal-00447067/fr/. (Cité à la page 132.)

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese et Omar Zaidan : Joshua : An Open Source Toolkit for Parsing-Based Machine Translation. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 135–139, Athènes, mars 2009. URL http://www.aclweb.org/anthology/W09-0424. (Cité à la page 94.)

Percy Liang, Ben Taskar et Dan Klein : Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 104–111, New York City, juin 2006. URL http://www.aclweb.org/anthology/N/N06/N06-1014. (Cité aux pages 17 et 92.)

Dekang Lin, Shaojun Zhao, Benjamin Van Durme et Marius Paşca : Mining Parenthetical Translations from the Web by Word Alignment. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL'08 : HLT)*, pages 994–1002, Columbus (Ohio, États-Unis), juin 2008. URL http://www.aclweb.org/anthology/P/P08/P08-1113.pdf. (Cité à la page 8.)

URL http://aclweb.org/anthology-new/P/P07/P07-2045.pdf. (Cité à la page 92.)

Philipp Koehn, Franz Och et Daniel Marcu : Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 48–54, Edmonton, 2003. URL http://aclweb.org/anthology-new/N/N03/N03-1017.pdf. (Cité aux pages 8, 13, 81, 82, 84 et 113.)

Mathieu Lafourcade et Christian Boitet : UNL lexical selection with conceptual vectors. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1958–1964, Las Palmas de Gran Canaria, 2002. URL http://www.mt-archive.info/LREC-2002-Lafourcade.pdf. (Cité à la page 31.)

Philippe Langlais, Fabrizio Gotti et Guihong Cao : NUKTI : English-Inuktitut Word Alignment System Description. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 75–78, Ann Arbor, juin 2005. URL http://www.aclweb.org/anthology/W/W05/W05-0810.pdf. (Cité à la page 26.)

Charlotte Lecluze : Méthode d'alignement sémantique multilingue appliquée à une collection de multidocuments : un apport aux systèmes d'aide à la traduction. Mémoire de 2de année de Master Recherche, université de Caen, septembre 2007. (Cité aux pages 117 et 131.)

Anne Lemoine : Segmentation de corpus multilingue par analyse sémantique : une nouvelle approche. Mémoire de 2de année de Master Recherche, université de Caen, 2006. (Cité à la page 117.)

Yves Lepage : De l'analogie rendant compte de la commutation en linguistique. Mémoire d'habilitation à diriger des recherches, université Joseph Fourier, mai 2003. URL http://tel.



la traduction probabiliste. À l'origine, la traduction par l'exemple n'intégrait pas d'alignement sous-phrastique à proprement parler. Elle se concentrait plutôt sur l'extraction de patrons dans le but de traduire à la volée une phrase particulière : il n'y avait pas de « compilation » géante comme en traduction probabiliste mais plutôt un effort de généralisation des exemples à disposition (p. ex. Brown, 1999; Sasayama et coll., 2009). En fait, des techniques d'alignement sous-phrastique, typiquement associatives, ont été utilisées par le passé pour faire de la traduction par l'exemple ; citons par exemple Brown (1997). Nous proposons de faire le contraire : utiliser des techniques de traduction par l'exemple pour faire de l'alignement sous-phrastique.

### 1.2.3 *Constats, propositions*

CONSTAT N° 1 *Les méthodes actuelles sont efficaces et éprouvées.*
Il suffit de faire tourner un aligneur tel que GIZA++ pour s'en rendre compte : l'alignement sous-phrastique marche. Nous pouvons d'ailleurs sans trop prendre de risques avancer que le problème est considéré comme réglé : depuis que les taux de précision d'alignement ont commencé à friser la perfection il y a quelques années — d'après le critère le plus répandu qu'est AER (Och et Ney, 2000) —, on constate un certain ralentissement d'activité dans ce domaine de recherche. Les méthodes répandues actuellement sont éprouvées, et proposer une énième méthode d'alignement, noyée dans la masse, ne présentera guère d'intérêt si celle-ci n'offre pas :

- soit un gain significatif en termes de qualité des résultats ;
- soit de nouvelles perspectives prometteuses.

Sur le premier point, nous ne nous donnons pas pour pour objectif de faire *mieux*, mais *au moins aussi bien*. Sur le second point en revanche, nous proposons d'apporter des contributions visibles.

CONSTAT N° 2 *C'est la course aux résultats.*
Il s'agit d'une question de méthodologie. De la même façon que Callison-Burch et coll. (2006) ont dénoncé une course au score Bleu dans le

domaine de la traduction automatique, Vilar et coll. (2006) ont montré qu'AER ne constituait pas non plus une finalité. La fiabilité d'une mesure d'évaluation n'a rien d'évident, et concentrer ses efforts sur des progrès selon un critère dont on risque d'oublier qu'il n'offre qu'*une* vue *partielle* de la « qualité » d'un système, peut mener à la création de systèmes brillants de par leurs scores, mais finalement assez ternes au regard de leurs résultats effectifs. Un score parfait en AER ne signifie hélas pas qu'un alignement est parfait, car en dehors des cas où les langues sont très proches, tout alignement « de référence » est de toute façon discutable. L'alignement de la figure 3 page 12 n'y déroge pas. Giguet (1996, section 2.1) plaide pour une autre façon de faire : « Le but n'est pas l'efficacité, celle-ci doit résulter d'une bonne analyse linguistique. » [2] C'est la méthodologie à laquelle nous adhérons. Sans prétendre à l'analyse linguistique profonde, nous proposons de construire notre alignement à partir d'*observations* de faits des textes.

CONSTAT N° 3 *Le « meilleur » alignement est toujours contestable.* Les alignements « de référence » établis par les humains étant eux-même discutables, il est difficile d'imaginer a fortiori que la machine puisse proposer « la » bonne solution. C'est pourtant ce qui est implicitement admis lorsqu'une méthode cherche à établir l'unique combinaison des meilleurs « liens » entre les mots d'un énoncé source et d'un énoncé cible. Beaucoup de méthodes sont capables de proposer plusieurs solutions, sous forme des *n* meilleures candidates, mais la limite entre les « bonnes » et les « mauvaises » est soumise à un seuil souvent arbitraire, et, en pratique, une seule est finalement conservée. C'est à notre sens une limitation induite par la définition même de la tâche d'alignement sous-phrastique sous forme de liens : nous ne devrions pas rechercher « la » meilleure solution puisqu'il en existe plusieurs. À la figure 3, l'alignement *cut … off* ↔ *retire la parole à* est correct dans ce contexte, mais nous aurions également pu admettre *cut … off* ↔ *retire la parole*, voire, de façon plus discutable, *cut* ↔ *retire la parole*. Idéalement, une méthode devrait être capable d'énumérer

2 Notre traduction.



Helsinki, août 1990. URL http://aclweb.org/anthology-new/C/C90/C90-3031.pdf. (Cité à la page 8.)

Philipp KOEHN : Statistical Significance Tests for Machine Translation Evaluation . In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 388–395, Barcelone, juillet 2004. URL http://www.aclweb.org/anthology-new/acl2004/emnlp/pdf/Koehn.pdf. (Cité à la page 61.)

Philipp KOEHN : Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, septembre 2005. URL http://www.mt-archive.info/MTS-2005-Koehn.pdf. (Cité aux pages 10, 19, 20, 100, 130, 148 et 194.)

Philipp KOEHN, Alexandra BIRCH et Ralf STEINBERGER : 462 Machine Translation Systems for Europe. In *Proceedings of the twelfth Machine Translation Summit (MT Summit XII)*, pages 65–72, Ottawa, août 2009. URL http://www.mt-archive.info/MTS-2009-Koehn-1.pdf. (Cité aux pages 118 et 132.)

Philipp KOEHN et Hieu HOANG : Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, juin 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1091.pdf. (Cité à la page 13.)

Philipp KOEHN, Hieu HOANG, Alexandra BIRCH, Chris CALLISON-BURCH, Marcello FEDERICO, Nicola BERTOLDI, Brooke COWAN, Wade SHEN, Christine MORAN, Richard ZENS, Chris DYER, Ondrej BOJAR, Alexandra CONSTANTIN et Evan HERBST : Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, juin 2007.

tawa, Canada, août 2009. URL http://www.mt-archive.info/MTS-2009-He.pdf. (Cité à la page 96.)

Dan Hirschberg : A Linear Space Algorithm for Computing Maximal Common Subsequences. *Communications of the ACM*, 18(6):341–343, juin 1975. URL http://www.cs.zju.edu.cn/people/yedeshi/Alinearspace-p341-hirschberg.pdf. (Cité à la page 152.)

Stéphane Huet, Julien Bourdaillet et Philippe Langlais : Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. In *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN/RECITAL 2009)*, Senlis, juin 2009. URL http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_44.pdf. (Cité à la page 8.)

Paul Jaccard : Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. (Cité à la page 30.)

Mark Johnson : How the Statistical Revolution Changes (Computational) Linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, pages 3–11, Athènes, mars 2009. URL http://www.aclweb.org/anthology/W09-0103. (Cité à la page 17.)

Hideki Kashioka, Takehiko Maruyama et Hideki Tanaka : Building a Parallel Corpus for Monologues with Clause Alignment. In *Proceedings of the ninth Machine Translation Summit (MT Summit IX)*, pages 216–223, Nouvelle Orléans, 2003. URL http://www.mt-archive.info/MTS-2003-Kashioka.pdf. (Cité à la page 59.)

Judith Klavans et Evelyne Tzoukermann : The BICORD System : Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics (Coling'90)*, pages 174–179,



plusieurs alignements corrects à partir d'un même couple d'énoncés. C'est bien pourquoi nous ne chercherons jamais à établir de *liens* entre mots, mais directement à extraire des traductions candidates, telles que celles listées précédemment, à partir de couples d'énoncés. Ce que nous nous « contentons » de faire est donc en quelque sorte la première phase de l'alignement associatif.

CONSTAT N° 4 *Les méthodes actuelles sont toutes bilingues.*
Depuis son introduction il y a une vingtaine d'années, l'alignement a toujours été un processus bilingue, le plus souvent orienté — ou *asymétrique* —, d'un énoncé source vers un énoncé cible. Nous proposons d'aligner davantage de langues simultanément. Nous aborderons ce point en détail dans le chapitre 6. Nous nous contentons pour l'instant de mettre cet objectif en avant comme un argument supplémentaire en faveur d'un alignement sans *liens* entre mots : l'établissement de liens multilingues entre mots est difficilement concevable lorsque plus de deux langues sont en jeu, alors qu'une traduction candidate entre plusieurs langues, telle *cut the speaker off* ↔ *retire la parole à l'orateur* ↔ *interrumpe al orador*, ici avec l'espagnol comme troisième langue, n'est en rien un problème.

CONSTAT N° 5 *La complexité appelle les complications.*
La traduction et l'alignement sont des processus complexes. Les *modèles* faisant le vœu pieu de les représenter — voir à ce sujet les discussions de Moore (2009) et de Johnson (2009) sur les limitations des modèles statistiques — sont donc naturellement complexes. Les plus éprouvés sont malheureusement vite devenus *compliqués* à force d'améliorations, ce qui tend à les rendre difficilement exploitables par nombre de praticiens, à moins qu'un outil libre de droits soit à disposition. Liang et coll. (2006) soulignent ainsi que la plupart des utilisateurs ont recours au célèbre outil GIZA++ comme une boîte noire, sans chercher à en comprendre le fonctionnement. Moore (2005) précise pour sa part qu'il n'aurait pas pu optimiser tous les paramètres des modèles de GIZA++ en un temps raisonnable dans ses expériences. En ce qui

nous concerne, nous prônons des méthodes accessibles, au moins aux personnes censées les utiliser. Les systèmes les plus épurés sont de surcroît les plus stables. Définir des méthodes simples qui marchent est cependant tout sauf simple, car cela oblige à tout remettre à plat, quitte à faire une croix — en toute connaissance de cause — sur des acquis de longue date qui ont fait leurs preuves. Pour beaucoup, simplicité rime avec naïveté — (re)voir à ce sujet la note du bas de la page 14. Certains relecteurs d'article ont ainsi trouvé nos travaux *naïfs* ; mais d'autres *simples et astucieux*[3].

CONSTAT N° 6 *Le passage à l'échelle n'est pas naturel.*
Du fait de la complexité — pas de la complication cette fois — des modèles d'alignement, en particulier estimatifs, leur passage à grande échelle nécessiterait des modifications importantes. Étrangement, à notre connaissance, le problème a peu été abordé en tant que tel : la réponse consiste généralement en un recours à des ressources physiques plus importantes — mémoire, espace de stockage, processeurs. Le problème de l'*accès* à de grandes quantités de données compilées a par contre bien été abordé en traduction automatique empirique, en particulier en ayant recours à des techniques issues de la traduction par l'exemple (p. ex. Brown, 2004; Lopez, 2008). Nous proposons de prendre en compte le passage à l'échelle dès la phase d'alignement. Nous verrons que grâce à notre approche, ce passage sera tout naturel : nous n'aurons pour ainsi dire rien à faire.

CONSTAT N° 7 *Les mots rares passent toujours à la trappe.*
Ce septième constat est à nos yeux capital. Souvent victimes de *seuils*, surtout avec les méthodes d'alignement associatives, les mots rares ne montrent pas leur potentiel. Alors que certains travaux mettent en œuvre des techniques spécifiques pour leur traitement (p. ex. Ahrenberg et coll., 1998), nous allons montrer qu'au contraire les mots rares peuvent servir de pilier à l'alignement.

3 Nos traductions.



Ulrich GERMANN : Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-08 : HLT Demo Session*, pages 20–23, Columbus (Ohio, États-Unis), juin 2008. URL `http://www.aclweb.org/anthology/P/P08/P08-4006.pdf`. (Cité à la page 8.)

Emmanuel GIGUET : The Stakes of multilinguality : Multilingual text tokenization in Natural Language Diagnosis. In *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence Workshop : Future issues for Multilingual Text Processing (PRICAI'96)*, Cairns, août 1996. URL `http://users.info.unicaen.fr/~giguet/pricai96/GiguetPricai96wshp.pdf`. 5 pages. (Cité aux pages 16, 116 et 121.)

Emmanuel GIGUET et Pierre-Sylvain LUQUET : Multilingual Lexical Database Generation from Parallel Texts in 20 European Languages with Endogenous Resources. In *Proceedings of the Coling/ACL 2006 Main Conference Poster Sessions*, pages 271–278, Sydney, juillet 2006. URL `http://aclweb.org/anthology-new/P/P06/P06-2035`. (Cité aux pages 13, 24, 31 et 118.)

Daniel GILDEA : Loosely Tree-Based Alignment for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 80–87, Sapporo, juillet 2003. URL `http://www.aclweb.org/anthology/P03-1011`. (Cité à la page 13.)

Thomas GREEN : The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18(4):481–496, 1979. URL `http://dx.doi.org/10.1016/S0022-5371(79)90264-0`. (Cité à la page 23.)

Zellig HARRIS : From phonemes to morphemes. *Language*, 31(2):190–222, 1955. (Cité à la page 56.)

Yifan HE et Andy WAY : Improving the Objective Function in Minimum Error Rate Training. In *Proceedings of the twelfth Machine Translation Summit (MT Summit XII)*, pages 238–245, Ot-

Stefan Evert et Anke Lüdeling : Measuring morphological productivity : Is automatic preprocessing sufficient ? In *Proceedings of the Conference on Corpus Linguistics 2001 (CL2001)*, pages 167–175, Lancaster (Royaume-Uni), 2001. URL http://www.ims.uni-stuttgart.de/projekte/corplex/paper/evert/EvertLuedeling2001.pdf. (Cité à la page 24.)

Cameron Shaw Fordyce : Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of the 4th International Workshop on Spoken Language Translation (IWSLT 2007)*, pages 1–12, Trente, octobre 2007. URL http://www.mt-archive.info/IWSLT-2007-Fordyce.pdf. (Cité aux pages 139, 147 et 151.)

Pascale Fung et Kenneth Church : K-vec : A New Approach for Aligning Parallel Texts. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, volume 2, pages 1096–1102, Kyōto, août 1994. URL http://aclweb.org/anthology-new/C/C94/C94-2178.pdf. (Cité à la page 13.)

William Gale et Kenneth Church : A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pages 177–184, Berkeley (Californie, États-Unis), juin 1991a. URL http://www.aclweb.org/anthology/P91-1023. (Cité à la page 10.)

William Gale et Kenneth Church : Identifying Word Correspondences in Parallel Texts. In *Proceedings of the fourth DARPA workshop on Speech and Natural Language*, pages 152–157, Pacific Grove, février 1991b. URL http://www.aclweb.org/anthology/H/H91/H91-1026.pdf. (Cité à la page 11.)

Qin Gao et Stephan Vogel : Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus (Ohio, États-Unis), juin 2008. URL http://www.aclweb.org/anthology/W/W08/W08-0509.pdf. (Cité à la page 91.)



| LANGUE | FORMES | HAPAX | MOTS/ÉN. | CAR./ÉN. |
|---|---|---|---|---|
| anglais | 58 342 | 39 % | 29 ± 17 | 161 ± 94 |
| espagnol | 93 043 | 40 % | 31 ± 18 | 173 ± 101 |
| finnois | 292 894 | 54 % | 21 ± 12 | 163 ± 95 |
| français | 73 695 | 37 % | 32 ± 19 | 180 ± 104 |

Tableau 2 – Caractéristiques du corpus parallèle Europarl utilisé : taille du vocabulaire, proportion d'hapax, nombre moyen et écart-type de mots et de caractères par énoncé. Les quatre parties sont en correspondance au niveau des énoncés et sont constituées des 354 645 énoncés du corpus original ayant tous la même traduction anglaise.

Tous ces constats sont liés. En nous attaquant à l'un des problèmes mentionnés, nous nous attaquons à tous les autres. Notre cheminement sera le suivant : à partir d'observations (n° 2) sur les mots rares (n° 7) en corpus, nous poserons les bases d'une méthode d'alignement dont la simplicité (n° 5) permettra naturellement l'extraction de multiples traductions candidates (n° 3), le multilinguisme (n° 4) et le passage à l'échelle (n° 6). Ce seront là nos contributions (n° 1).

## 1.3 LES MOTS RARES SONT BIEN FRÉQUENTS

Les mots rares constituent le point de départ de nos travaux. Nous proposons dans cette section un récapitulatif, avec leurs conclusions, d'un certain nombre d'observations.

### 1.3.1 *Choix des corpus d'étude*

Les expériences que nous présenterons tout au long de cette thèse auront principalement recours au corpus Europarl (Koehn, 2005), version 3, constitué des débats du parlement européen en 11 langues. Nous

avons choisi ce corpus principalement parce que les énoncés qui le composent correspondent à des discours prononcés en situation réelle et sont relativement longs (voir tableau 2 page précédente) comparés à ceux d'autres corpus tels que le BTEC (Takezawa et coll., 2002), où la longueur moyenne des énoncés est de 6,5 mots (proche de 10 mots dans nos propres échantillons). Il est aussi plus difficile à traiter que des corpus dont les énoncés sont aussi longs mais dont la forme obéit toujours aux mêmes conventions, tels la Bible (Resnik et coll., 1999), ou dont le domaine est très fermé, tels MÉTÉO (Chandioux et Guéraud, 1981). Il est de surcroît d'une taille conséquente, ce qui permet des études sur de grandes masses de données. Nous avons choisi deux couples de langues « extrêmes » pour réaliser nos expériences, dans l'optique de définir des limites entre lesquelles la plupart des autres couples de langues se situeront. Pour ce faire, nous nous sommes basé sur les premières expériences de traduction automatique de Koehn (2005) utilisant ce corpus. La distance ou proximité des langues que nous évoquons ici est donc attestée au moins sur le corpus que nous utilisons.

Notre premier couple est constitué des deux langues les plus éloignées, d'après les expériences de Koehn, parmi celles offertes par Europarl : l'anglais, langue d'origine germanique, isolante et morphologiquement pauvre, et le finnois, langue ouralienne, agglutinante et morphologiquement riche. Il s'agit de notre couple « difficile ». Il existe typiquement un déséquilibre en termes de longueur en nombre de mots entre un énoncé et sa traduction, comme le montre le couple d'énoncés réel suivant :

| | |
|---|---|
| The next item is the joint debate on the following reports: *(11 mots typographiques, 60 caractères)* | Esityslistalla on seuraavana yhteiskeskustelu seuraavista mietinnöistä: *(6 mots typographiques, 74 caractères)* |

À l'opposé, le second couple que nous avons choisi est constitué de langues les plus proches possible, toujours d'après les expériences de Koehn. Il s'agit du français et de l'espagnol, toutes deux langues romanes synthétiques. Du fait de leur proximité, nous sommes en



URL http://www.aclweb.org/anthology/N/N07/N07-1020.pdf. (Cité à la page 56.)

Fathi Debili et Hadhémi Achour : Voyellation automatique de l'arabe. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (Coling-ACL'98)*, pages 42–49, Montréal, août 1998. URL http://www.aclweb.org/anthology/W/W98/W98-1006.pdf. (Cité à la page 159.)

John DeNero et Dan Klein : Tailoring Word Alignments to Syntactic Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 17–24, Prague, juin 2007. URL http://www.aclweb.org/anthology/P07-1003. (Cité à la page 13.)

Étienne Denoual : *Méthodes en caractères pour le traitement automatique des langues*. Thèse de doctorat, université Joseph Fourier, Grenoble, septembre 2006. URL http://tel.archives-ouvertes.fr/tel-00107056/fr/. 186 pages. (Cité aux pages 117 et 151.)

Lee R. Dice : Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945. URL http://www.esajournals.org/doi/abs/10.2307/1932409. (Cité à la page 30.)

Bonnie Dorr, Lisa Pearl, Rebecca Hwa et Nizar Habash : Improved Word-Level Alignment : Injecting Knowledge about MT Divergences. Rapport technique LAMP-TR-082, CS-TR-4333, UMIACS-TR-2002-15, université du Maryland, College Park, février 2002. URL http://lampsrv02.umiacs.umd.edu/pubs/TechReports/LAMP_082/LAMP_082.pdf. 10 pages. (Cité aux pages 57 et 59.)

Hervé Déjean : *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse de doctorat, université de Caen, décembre 1998. URL http://tel.archives-ouvertes.fr/tel-00169572/fr/. 246 pages. (Cité à la page 56.)

David Chiang : Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, 2007. URL http://aclweb.org/anthology-new/J/J07/J07-2003.pdf. (Cité aux pages 8 et 94.)

Ilyas Cicekli : Similarities and Differences. In *Proceedings of SCI2000*, pages 331–337, Orlando, juillet 2000. URL http://www.cs.bilkent.edu.tr/~ilyas/PDF/sci2000.pdf. (Cité à la page 118.)

Ilyas Cicekli et Halil Altay Güvenir : Learning Translation Rules from a Bilingual Corpus. In *Proceedings of the 2nd International Conference on New Methods in Language Processing (NeMLaP-2)*, pages 90–97, université de Bilkent, Ankara, 1996. URL http://www.mt-archive.info/NEMLAP-1996-Cicekli.pdf. (Cité à la page 70.)

Josep Maria Crego, Aurélien Max et François Yvon : Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multi-source par renforcement lexical. In *Actes de la 16e conférence sur le traitement automatique des langues naturelles (TALN 2009)*, Senlis, juin 2009. URL http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_100.pdf. 10 pages. (Cité aux pages 9 et 132.)

Fabien Cromières : Sub-Sentential Alignment Using Substring Co-Occurrence Counts. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, pages 13–18, Sydney, juillet 2006. URL http://www.aclweb.org/anthology/P/P06/P06-3003.pdf. (Cité aux pages 24 et 33.)

Fabien Cromières : *Vers un plus grand lien entre alignement, segmentation et structure des phrases*. Thèse de doctorat, université Joseph Fourier, Grenoble, janvier 2010. 337 pages. (Cité aux pages 5 et 30.)

Sajib Dasgupta et Vincent Ng : High-Performance, Language-Independent Morphological Segmentation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT'07)*, pages 155–163, Rochester (New York, États-Unis), avril 2007.

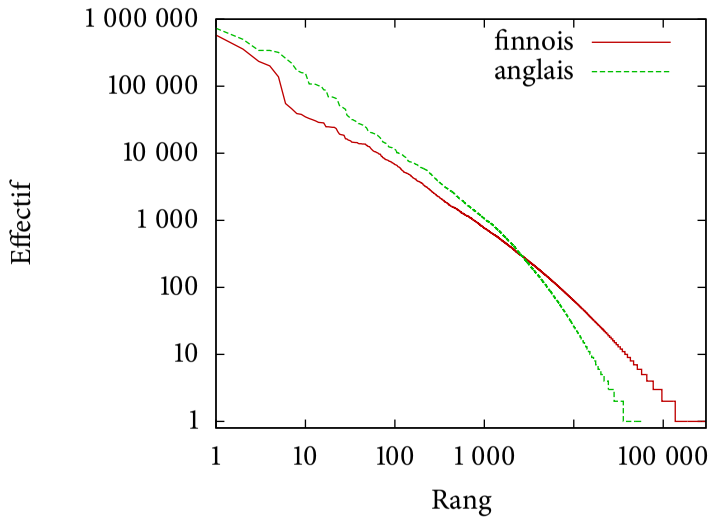


droit d'espérer de bien meilleurs résultats en français-espagnol qu'en anglais-finnois sur toute tâche traitant des couples de langues. Notons que le choix de ces langues pour leur proximité est discutable ; de façon extrême, nous aurions très bien pu utiliser la *même* langue en source et en cible, mais cela est à notre avis excessif dans la mesure où même des méthodes primitives donneraient d'excellents résultats : une comparaison ne présenterait alors guère d'intérêt puisque toutes les approches mèneraient aux mêmes excellents résultats. Un couple de langues tel que français-espagnol nous met à l'abri d'une telle dérive.

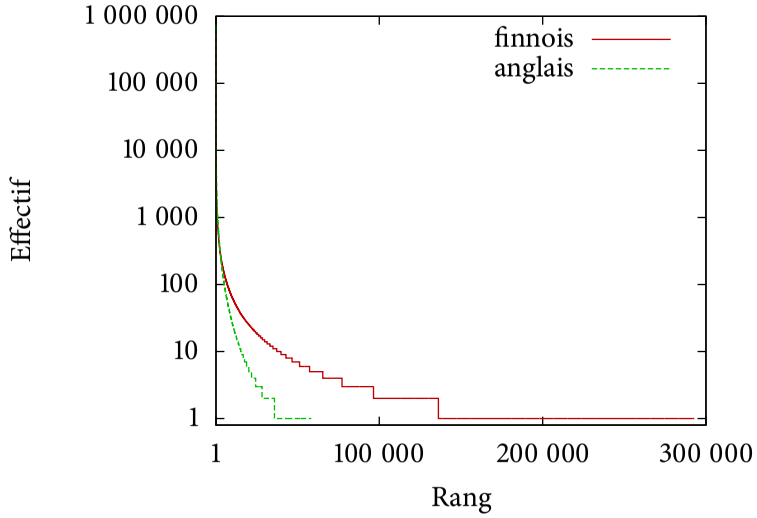
### 1.3.2 *Premières observations*

> Sans doute, le nombre total des mots d'une langue est très grand. Au dire de certains philologues, on n'en compterait pas moins de 90 000 dans la langue française. Mais il s'en faut que tous soient d'un usage courant. Il en est un petit nombre qui reviennent à tout instant et qui forment comme une petite troupe active, toujours en avant, toujours prête à servir, tandis que les autres constituent des réserves et même d'immenses [troupes] territoriales rarement dérangées. (Petruszewycz, 1973)

Petruszewycz reprend ici le paragraphe « Le nombre et la fréquence des mots usuels » de la 7e édition du fascicule de J.-B. Estoup : *Exposé théorique de la méthode pour l'acquisition de la vitesse*. Inventeur de la sténographie moderne, J.-B. Estoup établit une relation entre le rang des mots d'un texte ordonnés par ordre décroissant de fréquences d'apparition et cette fréquence : le produit *rang* × *effectif* est grosso modo constant. Le tracé de la courbe rang-effectif a donc la forme d'une hyperbole. Reprise par Zipf (1949), cette loi est désormais archi-connue en traitement automatique des langues sous le nom de loi d'(Estoup-)Zipf. On la représente généralement avec une échelle logarithmique, comme dans le graphique (a) de la figure 4 page suivante, la courbe prenant alors la forme d'une droite.




(a) Échelle logarithmique en ordonnée et en abscisse.




(b) Échelle logarithmique en ordonnée uniquement.

FIGURE 4 – Illustrations de la loi d'Estoup-Zipf. Les mots de notre corpus Europarl, en finnois et en anglais, ont été triés du plus fréquent au moins fréquent. Les mots rares apparaissent clairement comme majoritaires sur le second graphique.



Chris CALLISON-BURCH, Miles OSBORNE et Philipp KOEHN : Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 249–256, Trente, avril 2006. URL `http://aclweb.org/anthology-new/E/E06/E06-1032.pdf`. (Cité aux pages 15, 95 et 96.)

Bruno CARTONI : Constance et variabilité de l'incomplétude lexicale. In *Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles (TALN/RECITAL 2006)*, pages 661–669, Louvain, avril 2006. URL `http://www.issco.unige.ch/en/staff/bruno/recital_BC_2006.pdf`. (Cité à la page 26.)

John CHANDIOUX et Marie-France GUÉRAUD : MÉTÉO : un système à l'épreuve du temps. *Meta : journal des traducteurs*, 26(1):18–22, 1981. URL `http://id.erudit.org/iderudit/002213ar`. (Cité à la page 20.)

Simon CHAREST, Éric BRUNELLE, Jean FONTAINE et Bertrand PELLETIER : Élaboration automatique d'un dictionnaire de cooccurrences grand public. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN/RECITAL 2007)*, pages 283–292, Toulouse, juin 2007. URL `http://www.irit.fr/~Dominique.Longin/TALN2007/Actes_TALN2007_volume1.pdf`. (Cité à la page 124.)

Stanley CHEN : Aligning Sentences in Bilingual Corpora using Lexical Information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 9–16, Columbus (Ohio, États-Unis), juin 1993. URL `http://www.aclweb.org/anthology/P93-1002`. (Cité à la page 10.)

Christophe CHENON : *Vers une meilleure utilisabilité des mémoires de traduction, fondée sur un alignement sous-phrastique*. Thèse de doctorat, université Joseph Fourier, Grenoble, octobre 2005. URL `http://tel.archives-ouvertes.fr/tel-00012126/fr/`. 228 pages. (Cité à la page 8.)

Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL http://aclweb.org/anthology-new/J/J93/J93-2003.pdf. (Cité aux pages 8, 12 et 92.)

Peter Brown, Jennifer Lai et Robert Mercer : Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pages 169–176, Berkeley (Californie, États-Unis), juin 1991b. URL http://www.aclweb.org/anthology/P91-1022. (Cité à la page 10.)

Ralf Brown : Automated Dictionary Extraction for ''Knowledge-Free'' Example-Based Translation. In *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97)*, pages 111–118, Santa Fe (Nouveau-Mexique, États-Unis), juillet 1997. URL http://www.mt-archive.info/TMI-1997-Brown.pdf. (Cité à la page 15.)

Ralf Brown : Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32, Chester (Royaume-Uni), août 1999. URL http://www.mt-archive.info/TMI-1999-Brown.pdf. (Cité à la page 15.)

Ralf Brown : A Modified Burrows-Wheeler Transform for Highly-Scalable Example-Based Translation. In *Machine Translation : From Real Users to Research*, volume 3265 de *Lecture Notes in Computer Science*, pages 27–36. Springer Heidelberg, 2004. URL http://www.cs.cmu.edu/~ralf/papers/amta2004.pdf. (Cité à la page 18.)

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz et Josh Schroeder : (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, juin 2007. URL http://www.aclweb.org/anthology/W/W07/W07-0718.pdf. (Cité à la page 96.)



Nous nous intéressons ici à cette loi dans un unique but : mettre en évidence la « petite troupe active » et les « immenses territoriales rarement dérangées » dont fait mention Estoup. L'échelle logarithmique du graphique (a) est adaptée pour mettre en évidence la relation rang/fréquence, mais biaise l'apréciation que nous avons des quantités réelles de mots mises en jeu. Par conséquent, nous la présentons à nouveau avec une échelle semi-logarithmique au graphique (b) de la figure 4. À l'extrême gauche du graphique, la « petite troupe active », constituée typiquement de mots grammaticaux — ou mots-outils ou mots fonctionnels ou mots « vides » ou *mots fréquents* : pronoms, déterminants, etc. —, est *vraiment* petite. C'est justement leur faible nombre qui rend ces mots utiles dans certaines tâches fondées sur des marqueurs syntaxiques (Green, 1979), comme l'ont utilisé par exemple Stroppa et Way (2006) en traduction automatique. À l'inverse, les « immenses territoriales », constituées typiquement de mots lexicaux — ou mots pleins ou *mots rares* : noms, adjectifs, verbes, etc. —, sont *vraiment* immenses et occupent la quasi-totalité du graphique.

La limite entre mots grammaticaux et mots lexicaux est bien évidemment arbitraire, puisque sujette à un *seuil* que nous ne cherchons pas ici à définir ; seule l'allure générale des courbes nous intéresse. Tout est en fait une question d'échelle : étant donnée la grande taille de nos textes, une échelle arithmétique ne nous permettrait même pas de voir les courbes car elles seraient confondues avec les axes du graphique, les mots grammaticaux sur l'axe des ordonnées et les mots lexicaux sur l'axe des abscisses. Cela est plus flagrant avec certaines langues que d'autres ; en particulier, une langue agglutinante telle que le finnois compte davantage de mots rares et moins de mots fréquents qu'une langue isolante telle que l'anglais. L'atteste la différence de pente entre les deux « droites » du graphique (a). Dans tous les cas, l'écrasante majorité de mots rares apparaît comme une évidence.

### 1.3.3 *Les hapax : la bête noire du TAL*

Un *hapax*[4] est un mot qui n'apparaît qu'une seule fois dans un texte. Les hapax constituent le dernier « palier », le plus étendu, sur les graphiques de la figure 4 page 22. Une croyance répandue est que :

> Les *hapax legomenon* et autres événements dits rares présentent un problème intéressant pour les applications fondées sur des corpus : du fait de leur faible fréquence, ils n'apportent pas suffisamment d'information statistique pour des applications telles que l'alignement de mots ou la traduction automatique probabiliste. (Schrader, 2006)[5]

Par définition, les hapax sont écartés des données dans les approches qui filtrent les mots de faible fréquence. Les méthodes associatives d'alignement de mots sont particulièrement touchées, car elles reposent justement sur un test de significativité statistique. Par exemple, Cromières (2006) définit une borne inférieure sur les fréquences pour considérer un mot pour l'alignement. Giguet et Luquet (2006) définissent un seuil proportionnel à l'inverse de la longueur du terme.

En plus de leur faible nombre d'occurrences, un aspect supposé négatif des hapax est qu'ils comprennent néologismes et mots mal orthographiés (Schrader, 2006). Les néologismes devraient être considérés comme des mots à part entière. La quantité de mots mal orthographiés dépend quant à elle de la qualité du corpus utilisé. D'après Nishimoto (2004), qui interprète les résultats d'Evert et Lüdeling (2001), chaque erreur n'a lieu qu'une seule fois en moyenne dans un corpus. Les mots mal orthographiés sont ainsi typiquement hapax, mais leur proportion parmi l'ensemble des hapax demeure très faible. Ils ne posent de toute

4 Du grec « ἅπαξ λεγόμενον » */hápax legómenon/* '[dit] une seule fois'. Comme il est couramment pratiqué en informatique, nous utilisons ce mot par extension, abusive, pour désigner un mot unique en corpus. Pour la suite de notre propos donc, un hapax est un mot qui n'apparaît qu'une seule fois dans un corpus. Pour être plus précis, nous déterminerons les hapax dans des corpus monolingues, c'est-à-dire des sous-ensembles d'Europarl. Nous verrons plus loin que ce nouveau cadre peut aussi être remis en question.

5 Notre traduction.



Hervé Blanchon et Christian Boitet : Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. *TAL*, 48(1):33–65, 2007. URL `http://www.atala.org/IMG/pdf/TAL-2007-48-1-02-Blanchon.pdf`. (Cité à la page 96.)

Christian Boitet : Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes. In *Actes de la 15e conférence sur le Traitement Automatique des Langues Naturelles (TALN/RECITAL 2008)*, Avignon, juin 2008. URL `http://www-clips.imag.fr/geta/christian.boitet/pages_personnelles/zArticles_sur_la_TAO_pdf/TALN-08-ArchiTA.080218.v7-final.pdf`. (Cité à la page 9.)

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki et Kentaro Ogura : Design and Construction of a machine-tractable Japanese-Malay Dictionary. In *Proceedings of the eighth Machine Translation Summit (MT Summit VIII)*, pages 53–58, Saint-Jacques-de-Compostelle, septembre 2001. URL `http://www.mt-archive.info/MTS-2001-Bond.pdf`. (Cité à la page 98.)

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Fredrick Jelinek, Robert Mercer et Paul Roossin : A Statistical Approach to Language Translation. In *Proceedings of the 12th International Conference on Computational Linguistics (Coling'88)*, pages 71–76, Budapest, 1988. URL `http://aclweb.org/anthology-new/C/C88/C88-1016.pdf`. (Cité aux pages 6, 7, 8, 9 et 11.)

Peter Brown, Stephen Della Pietra, Vincent Della Pietra et Robert Mercer : Word-Sense Disambiguation using Statistical Methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pages 264–270, Berkeley (Californie, États-Unis), juin 1991a. URL `http://www.aclweb.org/anthology/P91-1034`. (Cité aux pages 9 et 132.)

Peter Brown, Stephen Della Pietra, Vincent Della Pietra et Robert Mercer : The Mathematics of Statistical Machine Translation :

Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, 1999. URL http://www.clsp.jhu.edu/ws99/final/Stat_Machine_Translation.pdf. 42 pages. (Cité à la page 92.)

Harald Baayen et Richard Sproat : Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics*, 22(2):155–166, juin 1996. URL http://aclweb.org/anthology-new/J/J96/J96-2001.pdf. (Cité à la page 26.)

Colin Bannard et Chris Callison-Burch : Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, juin 2005. URL http://aclweb.org/anthology-new/P/P05/P05-1074.pdf. (Cité à la page 116.)

Emily Bender : Linguistically Naïve != Language Independent : Why NLP Needs Linguistic Typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, pages 26–32, Athènes, mars 2009. URL http://www.aclweb.org/anthology/W09-0106. (Cité à la page 116.)

Lasse Bergroth, Harri Hakonen et Timo Raita : A Survey of Longest Common Subsequence Algorithms. In *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00)*, pages 39–48, A Coruña, 2000. URL http://biotec.icb.ufmg.br/cabi/artigos/seminarios2/subsequence_algorithm.pdf. (Cité à la page 152.)

Hervé Blanchon : Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral — Une bataille contre le bruit, l'ambiguïté et le manque de contexte. Mémoire d'habilitation à diriger des recherches, université Joseph Fourier, Grenoble, décembre 2004. URL http://www-clips.imag.fr/geta/herve.blanchon/Pdfs/HDR.pdf. 356 pages. (Cité à la page 95.)



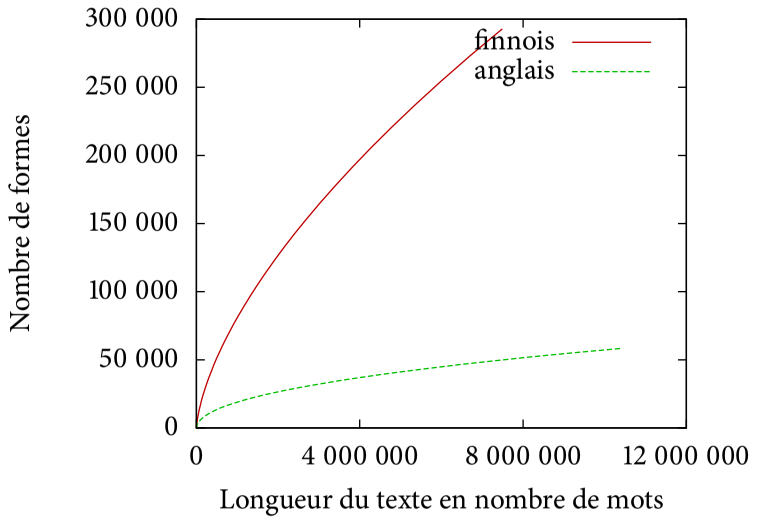



Figure 5 – Nombre de formes dans un texte en fonction de sa longueur en nombre de mots. Tant que la courbe croît, de nouveaux mots, rares, apparaissent.

façon pas de problème pour l'alignement de mots à partir d'un corpus parallèle : si un mot à l'origine déjà hapax se trouve mal orthographié, le seul impact que cela aura est que ce mot se retrouvera mal orthographié dans l'alignement résultant, sans que l'alignement à proprement parler ne soit affecté. S'il s'agit au contraire d'un mot fréquent, on peut s'attendre à ce que le score de l'alignement résultant soit très faible, puisque l'erreur ne se produisant qu'une seule fois (Nishimoto, 2004), les autres instances du mot seront correctement orthographiées, donc bien alignées et avec des scores autrement meilleurs. Même chose pour les néologismes : si des néologismes sont utilisés dans les deux langues, les alignements d'hapax en résultant doivent être valides ; et si au contraire des néologismes dans une langue se trouvent être traduits par des mots attestés plus d'une fois dans l'autre langue, les scores réultants seront très faibles, donc aisément détectables.

Mais les faits sont têtus. Les hapax ont beau être habituellement rejetés, ils sont pourtant omniprésents. Cartoni (2006) rappelle que les hapax représentent généralement 40 % des mots d'un corpus. Ce nombre reflète principalement deux axes. Le premier concerne la *richesse du vocabulaire*, c'est-à-dire la quantité de formes différentes utilisées dans un texte. Des décomptes sur les pièces les plus lues de Shakespeare ont par exemple montré qu'elles contenaient en moyenne 58 % d'hapax [6]. Le second axe concerne le *degré de synthèse* de la langue : isolante, synthétique ou polysynthétique. Plus une langue est synthétique, plus ses mots sont fléchis, et par conséquent plus le nombre de formes est grand. La proportion d'hapax augmente en conséquence. Au graphique (b) de la figure 4, le dernier palier de la courbe correspondant au finnois contient plus de 50 % d'hapax. De façon encore plus remarquable, Langlais et coll. (2005) rapportent plus de 80 % d'hapax sur un corpus d'inuktitut, une langue très synthétique du Canada. Dans ce dernier cas, rejeter les hapax reviendrait à ne considérer que 20 % des données, ce qui dégraderait évidemment la qualité de toute tâche subséquente. À tout cela s'ajoute le fait que la proportion d'hapax est relativement constante quelle que soit la taille d'un texte, car en en augmentant la taille, de nouveaux mots, hapax, font leur apparition, comme le rappelle la courbe de croissance de la figure 5 page précédente. Cette relation qu'entretiennent les hapax avec les mots inconnus (Baayen et Sproat, 1996; Cartoni, 2006) les rend utiles pour estimer par exemple le comportement des systèmes de traduction automatique sur les mots inconnus. En définitive, les mots rares (d'un corpus monolingue) sont abondants : ils forment de grandes populations de faible fréquence et ce phénomène est omniprésent, indépendamment des langues et de la taille du corpus.

6 Décomptes disponibles à l'adresse : `http://www.mta75.org/curriculum/English/Shakes/index.html` (page consultée le 1er juin 2010).

## BIBLIOGRAPHIE

Eneko Agirre, Mikel Lersundi et David Martinez : A Multilingual Approach to Disambiguate Prepositions and Case Suffixes. In *Proceedings of the ACL'02 Workshop on Word Sense Disambiguation : Recent Successes and Future Directions*, pages 1–8, Philadelphie, juillet 2002. URL `http://www.aclweb.org/anthology/W02-0801`. (Cité à la page 116.)

Lars Ahrenberg, Mikael Andersson et Magnus Merkel : A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-Coling 98)*, volume 1, pages 29–35, Montréal (Québec, Canada), août 1998. URL `http://www.aclweb.org/anthology/P98-1004`. (Cité à la page 18.)

Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein et Jörg Tiedemann : Evaluation of Word Alignment Systems. In *Proceedings of the Second International Conference on Language Resources ans Evaluation (LREC 2000)*, pages 1255–1261, Athènes, 2000. URL `http://www.cse.unt.edu/~rada/wa/papers/EvalWASystems.pdf`. (Cité à la page 97.)

Mosleh Al-Adhaileh : *Synchronous Structured String-Tree Correspondence (S-SSTC) and its applications for machine translation*. Thèse de doctorat, Universiti Sains Malaysia, Pulau Pinang, 2003. 163 pages. (Cité à la page 8.)

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz Och, David Purdy, Noah Smith et David Yarowsky : Statistical Machine Translation : Final Report. Rapport technique, Johns Hopkins University 1999 Summer

peut ainsi être considéré comme faux à cause d'un unique caractère. Notre système semble ne pas être capable d'aligner autant de *mots* que GIZA++ dans ces conditions, mais il reste néanmoins capable d'aligner des chaînes beaucoup plus diverses, dont nous ne pouvons évaluer la qualité que qualitativement faute de référence adaptée. Le tableau 18 page précédente présente quelques alignements produits par notre système dans les mêmes conditions que l'expérience précédente. Les premiers ont été choisis manuellement et les derniers par échantillonnage. Comme nous l'avons mentionné précédemment, et malgré la couverture inférieure à la référence du domaine suggérée par la figure 30, notre méthode peut théoriquement aligner n'importe quelle chaîne de caractères pour peu que les données d'entrée le permettent, qu'il s'agisse d'un unique caractère ou de phrases entières. Nous constatons que le cas où l'alignement est proche de la traduction espérée en ne différant que de quelques caractères uniquement n'est pas rare, y compris en arabe où les résultats chiffrés étaient médiocres.



RÉSUMÉ

Ce chapitre a présenté une vue d'ensemble des faits qui entourent l'alignement sous-phrastique. Les principaux points sont les suivants :

- nous nous plaçons dans le cadre d'un alignement sous-phrastique endogène et non supervisé, en partant de corpus parallèles dont les énoncés ont préalablement été mis en correspondance ;
- nos travaux se situeront relativement en marge des deux grands courants, à savoir l'approche estimative et l'approche associative, et apporteront des réponses à des points que ces approches n'ont pas abordés — ou pas pu aborder — depuis leur introduction il y a une vingtaine d'années ;
- notre point de départ sera les mots rares. Ils sont couramment reniés, et pourtant omniprésents. Nous les mettrons au cœur de notre contribution.

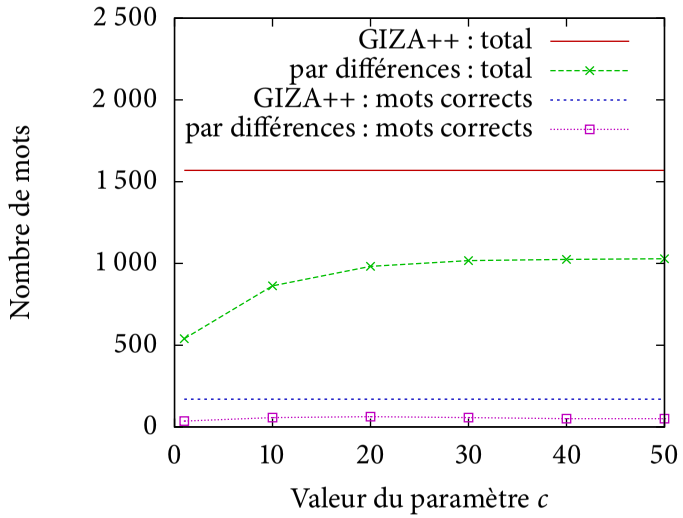
| ANGLAIS | ARABE | JAPONAIS |
|---|---|---|
| . '.' | . /./ '.' | 。 /./ '.' |
| ? '?' | ؟ /?/ '?' | か 。 /*ka* ./ '?' |
| Wh 'Qu…' | اين /*āyn*/ 'où' | 何 /*nani*/, /nan/ 'quoi', 'qu…' |
| here 'ici' | هنا /*hnā*/ 'ici' | ここ /*koko*/ 'ici' |
| I 'd like 'Je voudrais' | اريد /*āryd*/ 'Je voudrais' | 下さい /*kudasai*/ 's'il vous plaît' |
| What 'Quoi, Que' | ماذاحضر لكم /*māḏāḥḍr lkm*/ 'qu'avez-vous' | 何 /*nani*/, /nan/ 'quoi', 'qu…' |
| How 'Comment' | كمخصمه /*kmẖṣmh*/ 'comment, combien' | どのくらい /*donokurai*/ 'combien' |
| How much 'Combien' | كمخصمه /*kmẖṣmh*/ 'comment, combien' | いくら /*ikura*/ 'combien' |
| Thank you 'Merci' | شكرا /*škrā*/ 'merci' | ありがとう /*arigatou*/ 'merci' |

(a) Exemples choisis manuellement.

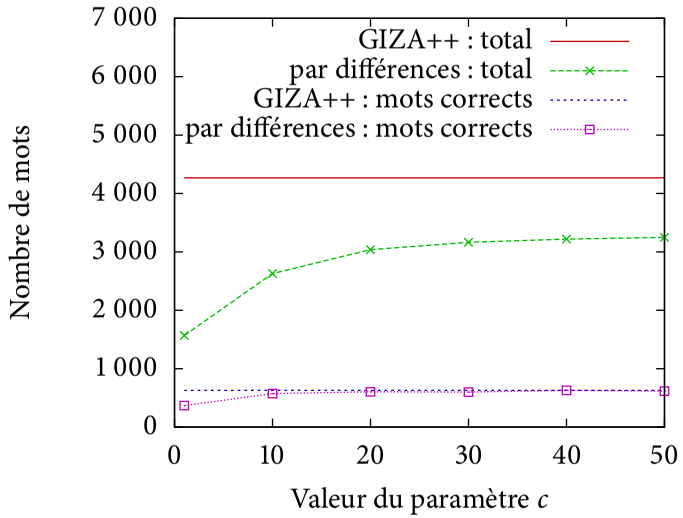
| ANGLAIS | ARABE | JAPONAIS |
|---|---|---|
| Ice 'Glace' | ايس كريم /*āys krym*/ 'glace' | 氷を /*koori o*/ 'de la glace' |
| station 'poste (de police)' | *(chaîne vide)* | 署 /*syo*/ 'poste (de police)' |
| picture book 'livre d'images' | اريد كتاباصورا /*āryd ktābāṣwrā*/ 'Je voudrais un livre d'images' | 絵本 /*ehon*/ 'livre d'images ' |
| a really beautiful dress 'une très belle robe' | هيهذا فستان جميل حقا /*hyhḥḍā fstān ǧmyl ḥqā*/ 'c'est une très belle robe' | とてもきれいなドレス /*totemo kirei na doreṣu*/ 'une très belle robe' |

(b) Exemples obtenus par échantillonnage.

TABLEAU 18 – Exemples d'alignements obtenus par application itérative de différences de chaînes. La langue source est l'anglais. Le paramètre $c$ a été positionné à 20 dans cette expérience. Les traductions arabes et japonaises ont été produites simultanément.




(a) Anglais → arabe




(b) Anglais → japonais

FIGURE 30 – Comparaison des résultats produits par GIZA++ et l'application itérative de différences de chaînes. Le nombre total de mots produits par GIZA++, ainsi que le nombre de ses mots trouvés dans les dictionnaires de référence, sont supérieurs sur les deux tâches, mais cet avantage est moindre en japonais.

# 2

## LA FACE CACHÉE DES MOTS RARES

DANS le chapitre précédent, nous avons vu que les mots rares étaient souvent sous-utilisés — si ce n'est totalement rejetés — car jugés peu fiables, et ce bien qu'ils représentent la moitié du vocabulaire de tout texte. Le but de ce chapitre est de montrer que les meilleurs alignements obtenus à partir de corpus parallèles sont en réalité massivement constitués de mots rares. À cette fin, nous analysons les résultats d'une expérience d'alignement bilingue mot-à-mot à l'aide d'une technique d'alignement associative par corrélation.

SOMMAIRE

## 2.1 MÉTHODE D'ALIGNEMENT

### 2.1.1 *Choix de la méthode*

L'expérience que nous proposons ci-après n'a pas pour but de produire les meilleurs alignements qui soient, mais de mettre en évidence certaines caractéristiques des alignements obtenus par la plupart des méthodes. Notre choix pour l'alignement se porte donc ici sur une mesure de corrélation « simple » qui associe typiquement à chaque couple (*source*, *cible*) un ou plusieurs scores reflétant la probabilité que *source* et *cible* soient de bonnes traductions l'un de l'autre. Dans leur plus simple expression, de telles approches sont bien évidemment très limitées en termes de qualité des résultats, car elles ne permettent l'alignement que d'un grain fixe, par exemple le mot : tous les alignements sont de multiplicité 1-1.

Comme le font remarquer Och et Ney (2003), le choix d'une mesure de corrélation pour l'alignement bilingue est généralement assez arbitraire, car ces mesures produisent typiquement des résultats de qualité comparable. Parmi les mesures les plus courantes, on peut citer les coefficients de Jaccard (Jaccard, 1901) et de Dice (Dice, 1945), le coefficient de corrélation linéaire, ou encore le test du $\chi^2$ et celui du $G^2$. Cromières (2010, chap. II) donne une description théorique de bon nombre d'entre elles en s'appuyant sur les tableaux de contingence, et montre que, contrairement à ce que l'usage laisserait supposer, le choix de la méthode peut avoir en pratique une influence non négligeable sur la qualité des résultats. La qualité des résultats n'étant pas notre priorité dans ce chapitre, notre choix demeure purement pragmatique et se concentre donc sur la simplicité de mise en œuvre de la méthode pour l'exploitation de ses résultats.

Ces méthodes sont fondées sur le décompte du nombre d'énoncés où un segment source et un segment cible apparaissent simultanément, et parfois sur le nombre de fois où ils n'apparaissent *pas* ensemble. Certaines ne se contentent pas d'opérer sur la simple présence ou absence des segments dans un énoncé, mais prennent en compte le nombre



plets d'énoncés alignés en anglais, arabe et japonais. Nous effectuons nos évaluations de l'anglais vers l'arabe d'une part et de l'anglais vers le japonais d'autre part. Notre système traite naturellement les deux langues cible simultanément. Les énoncés sont segmentés en mots, y compris en japonais, et les ponctuations sont séparées des mots par des espaces. Cela n'est pas requis par notre approche puisqu'elle agit au niveau du caractère : dans nos expériences, les résultats obtenus avec ou sans segmentation en mots sont similaires en japonais. Les résultats présentés ci-après sont obtenus avec une pré-segmentation, pour l'unique raison que GIZA++ ne peut traiter que des mots. Celui-ci est utilisé avec ses paramètres par défaut, qui donnent typiquement de bons résultats, en enchaînant cinq itérations des modèles IBM 1, HMM, IBM 3 et IBM 4. Nous testons différentes valeurs du paramètre $c$ décrit à la section précédente avec notre système (1, 10, 20, 30, 40 et 50). Le filtrage des chaînes de caractères cibles mal formées s'effectue sur les trigrammes en arabe et sur les bigrammes en japonais. Les deux lexiques bilingues de référence, anglais-arabe et anglais-japonais, sont issus du site *XDXF* évoqué page 99. Les formes arabes y étant lemmatisées, la partie arabe du corpus d'entraînement a été lemmatisée également par la méthode de Debili et Achour (1998).

La figure 30 page suivante présente les résultats en fonction du paramètre $c$. La courbe correspondant à GIZA++ est constante car ce paramètre n'y a pas d'équivalent. Sur la tâche anglais-japonais, notre approche obtient ses meilleurs résultats pour $c = 40$, où le nombre d'alignements trouvés dans le dictionnaire de référence est quasiment identique à celui de GIZA++ (628 contre 629 pour le second). Sur la tâche anglais-arabe, les meilleurs résultats sont obtenus pour $c = 20$, mais notre approche n'obtient que 63 alignements corrects contre 170 pour GIZA++. Dans les deux cas, le nombre total de mots alignés par notre système est inférieur à celui de GIZA++. Notons que les alignements produits par les deux systèmes n'appartenant pas aux dictionnaires de référence, majoritaires pour les deux systèmes sur la figure, ne sont pas nécessairement erronés, car nous avons recours à des comparaisons exactes. Un alignement produit par notre système

termes eux-mêmes, ce qui permet théoriquement l'alignement de toute chaîne, quelles que soient sa longueur et sa fréquence, à condition que ses contextes apparaissent au moins par morceaux dans d'autres énoncés du corpus. Elle compense également l'absence volontaire de connaissance sur les mots par un processus itératif faisant intervenir des différences de chaînes de caractères contigus. Le nombre d'énoncés nécessaires à l'alignement d'une chaîne dépend ainsi de sa longueur et de la longueur des énoncés constituant le corpus : ce nombre est strictement inférieur au nombre de caractères des énoncés où la chaîne à aligner apparaît. En fin de compte, en sélectionnant les énoncés nécessaires et suffisants à un alignement donné, comme on le fait parfois en traduction automatique par l'exemple, nous constituons en réalité un sous-corpus de petite taille où les termes à aligner et leurs contextes n'ont que de faibles effectifs, comme nous l'avons évoqué au début du chapitre 3. Vue sous cet angle, la traduction automatique par l'exemple reposerait par essence sur les termes de basses fréquences, et ce depuis ses débuts il y a une vingtaine d'années, sans que cela ait jamais réellement été mis en avant.

## D.3 ÉVALUATION

Nous comparons notre approche à GIZA++ (Och et Ney, 2003), outil de référence en matière d'alignement de mots par approche statistique. Notre protocole est ici rudimentaire : nous nous contentons d'aligner chacun des mots source du corpus d'entraînement et de ne conserver que la traduction dont la probabilité associée est la plus élevée, et comptons le nombre de couples de traductions de mots en commun avec un lexique bilingue de référence. Les sorties des deux systèmes sont ainsi comparables, hormis le fait que celles du nôtre ne sont pas nécessairement constituées de mots typographiques isolés tels qu'attestés dans la partie cible du corpus d'entraînement, du fait du traitement au niveau du caractère.

Nous utilisons le corpus d'entraînement de la campagne d'évaluation de systèmes de traduction automatique IWSLT 2007, soit 20 000 tri-



d'occurences au sein d'un même énoncé. C'est le cas de la méthode dite du cosinus, que nous utiliserons pour notre expérience. Cette méthode est classique et est couramment utilisée pour des tâches variées, telles que la découverte d'entités nommées (Shinyama et Sekine, 2004), le traitement de vecteurs conceptuels pour des tâches sémantiques (Lafourcade et Boitet, 2002; Turney et Littman, 2005) et bien sûr l'alignement bilingue (Giguet et Luquet, 2006). Ayant recours au nombre réel d'occurences des segments, on peut espérer a priori de meilleurs résultats qu'avec une autre méthode reposant sur la simple présence ou absence des segments. Nous verrons dans les sections suivantes que le bénéfice n'est pas nécessairement significatif.

### 2.1.2 *Description détaillée de la méthode*

Nous décrivons la méthode du cosinus en utilisant le mot comme unité textuelle de traitement. La méthode consiste à se placer dans un espace vectoriel dont le nombre de dimensions est égal au nombre de couples d'énoncés (*source*, *cible*) dans le corpus parallèle bilingue considéré. Pour chacune des deux langues qui constituent ce corpus parallèle, on associe à chaque mot $m$ un vecteur $\vec{m}$ dont la $i$-ème composante est le nombre d'occurrences de $m$ dans le $i$-ème énoncé. Puis, pour chaque couple de mots $(m_s, m_c)$ issus respectivement de la langue source et de la langue cible, on calcule l'angle entre leurs vecteurs associés $\vec{m}_s$ et $\vec{m}_c$ :

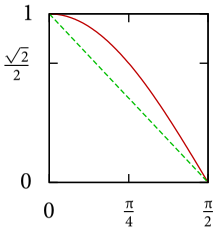
$$\text{angle}(\vec{m}_s, \vec{m}_c) = \text{acos}\left(\frac{\vec{m}_s \cdot \vec{m}_c}{\| \vec{m}_s \| \times \| \vec{m}_c \|}\right) \tag{2.1}$$

où $\vec{u} \cdot \vec{v}$ est le produit scalaire des vecteurs $\vec{u}$ et $\vec{v}$ et $\| \vec{v} \|$ la norme du vecteur $\vec{v}$. Le résultat est le score de l'alignement $(m_s, m_c)$. Ce score est toujours un nombre positif entre zéro et $\pi/2$ (inclus) car toutes les composantes de $\vec{m}_s$ et $\vec{m}_c$ sont positives.

Dans la plupart des travaux où cette méthode est utilisée, le passage à l'angle n'est pas effectué : le calcul se limite au calcul du cosinus de l'angle, entre parenthèses dans l'égalité 2.1. Nous choisissons de

considérer l'angle formé par les deux vecteurs plutôt que son cosinus car, comme nous le verrons par la suite, cela permet une interprétation plus claire de certaines configurations d'alignement. Cela implique de raisonner en termes de distances à la place des similarités qui sont généralement plus usitées : les meilleurs angles sont donc les plus faibles, proches de zéro, et les plus mauvais les plus élevés, proches de $\pi/2$.

Notons qu'une déformation est introduite lors de la conversion entre l'angle et son cosinus, car la fonction cosinus (resp. arc cosinus) n'est pas linéaire entre zéro et $\pi/2$ (resp. entre zéro et un). Cette déformation est visible sur le graphe ci-contre, où l'angle est en abscisses. Le cosinus de l'angle, représenté par la courbe en trait continu, est toujours supérieur à une fonction qui transformerait linéairement la distance représentée par l'angle en une mesure de similarité (en pointillés). Par conséquent, à un cosinus élevé ne correspond pas nécessairement un angle faible. Par exemple, un angle « moyen » de $\pi/4$ a un cosinus de $\sqrt{2}/2 \approx 0{,}71$, ce qui est bien supérieur à la valeur d'un cosinus « moyen » de 0,5. Le cosinus est donc plus optimiste que l'angle lui-même quant à la qualité d'un alignement. Cela revêt une importance toute particulière lorsque les scores des alignements sont considérés de façon absolue, par exemple un seuil qui ne conserverait que les alignements dont le score équivaut à au moins la moitié du score maximal, car un même alignement peut être conservé en se fiant à son cosinus alors qu'il ne le serait pas en se fiant à son angle. En ce qui nous concerne, cela n'aura qu'un impact mineur sur nos résultats car nous nous contentons d'ordonner les alignements en fonction de leurs scores, l'ordre étant préservé d'une mesure à l'autre. Nous garderons simplement à l'esprit que cette déformation se traduit par un décalage de populations d'alignements vers des valeurs d'angle plus élevées par rapport à l'utilisation plus classique du cosinus.






tous les énoncés menant à l'alignement d'une chaîne particulière. Cela résulterait en un grand nombre de calculs superflus, où la plupart des LCSubstr seraient très courtes, donc peu fiables, comme c'est le cas dans l'exemple précédent aux étapes $n = 2$ et $n = 3$. La segmentation d'un énoncé en deux parties inégales telle que nous la désirons suppose que le grain le plus fin soit pertinent vis-à-vis de l'alignement, ce qui peut ne plus être le cas en deçà d'une certaine taille en caractères. Pour limiter ces pertes de temps et de qualité, nous introduisons deux paramètres :

1. seules les $c$ LCSubstr plus longues qu'un seuil prédéfini sont examinées. Ce seuil est fixé arbitrairement à la moitié de la plus longue LCSubstr. Différentes valeurs de $c$ sont testées dans les expériences de la section suivante. Ce paramètre n'est utilisé qu'en langue source.
2. la bonne correction des chaînes est vérifiée en testant si tous les n-grammes de caractères qui les composent sont attestés dans les données de départ. Un tel filtrage est utilisé par exemple en traduction automatique par analogie (Lepage et Denoual, 2005), et a un rôle proche de celui des modèles de langue utilisés en traduction automatique probabiliste. Ce paramètre est utilisé dans les langues cible.

Le score d'un alignement est obtenu en divisant le nombre de fois qu'il a été obtenu par le nombre d'énoncés qui ont été nécessaires à son obtention, car en pratique, plus ce nombre est faible, plus les LCSubstr sont longues, donc fiables. Étant donnée une chaîne à aligner, nous ne conservons à la fin que sa traduction ayant obtenu le score le plus élevé. Cela constitue une perte, mais notre but n'est pas ici l'exhaustivité, et nous nous contenterons d'évaluer le meilleur candidat pour chaque chaîne.

Procéder par différences successives entre énoncés est une opération relativement courante dans le domaine de la traduction automatique par l'exemple (Nagao, 1984; Sato, 1991; Somers, 2003). L'approche que nous avons présentée ici s'en différencie toutefois par le fait qu'elle se focalise sur les contextes des termes à aligner plutôt que sur les

### 2.1.3 *Interprétation de la méthode*

Intuitivement, il est tentant de croire que plus un angle est faible, meilleur est l'alignement correspondant. Autrement dit, cette mesure est à interpréter comme une distance de traduction. Les alignements auxquels nous nous intéressons sont donc ceux dont l'angle est proche de zéro.

Théoriquement, un angle nul implique que $\vec{m}_s$ et $\vec{m}_c$ sont colinéaires, c'est-à-dire que les deux mots $m_s$ et $m_c$ apparaissent systématiquement dans les mêmes énoncés et que leur nombre d'occurrences est proportionnel : $\vec{m}_s = \lambda \vec{m}_t$, avec $\lambda \in \mathbb{N}^*$. En pratique, $\lambda$ vaut 1 la plupart du temps dans nos données. L'interprétation a priori d'un angle nul serait que les deux mots sont de parfaites traductions l'un de l'autre, autrement dit qu'ils sont lexicalement équivalents, mais cela n'est pas forcément vrai (voir section suivante).

À l'opposé, un angle de $\pi/2$ implique que le cosinus est nul. Cela se produit lorsque toutes les composantes du produit scalaire sont nulles, autrement dit quand l'intersection des ensembles d'énoncés dans lesquels les mots source et cible apparaissent simultanément est vide. Cela se produit presque systématiquement lorsqu'on considère tous les couples de mots possibles. Par conséquent, une implémentation efficace d'une telle méthode ne calcule bien évidemment pas les angles entre tous les couples de mots, mais uniquement entre ceux qui apparaissent au moins une fois dans le même énoncé. Cela restreint les vecteurs aux seules dimensions nécessaires : les énoncés où le couple de mots n'apparaît pas ne sont pas conservés en mémoire. Concernant l'implémentation, les tableaux de suffixes introduits par Manber et Myers (1993) puis Nagao et Mori (1994) constituent une structure de données efficace pour le calcul des angles à la volée. Ils ont été utilisés par exemple par Cromières (2006) pour calculer des co-occurences de chaînes.

| $n$ | $A_n$ | $E_n$ | LCSubstr($A_n$,$E_n$) | $\widehat{A}_n$ | LCSubstr($\widehat{A}_n$,$\widehat{E_n}$) |
|---|---|---|---|---|---|
| 0 | Is␣this␣a␣train␣for␣*Chicago*? | C | ␣train␣for␣ | この列車はシカゴ行きですか。 | の列車は |
| 1 | Is␣this␣a*Chicago*? | B | Is␣this␣ | こシカゴ行きですか。 | ですか。 |
| 2 | a*Chicago*? | B | ? | こシカゴ行き | こ |
| 3 | a*Chicago* | C | a | シカゴ行き | 行き |
| 4 | *Chicago* | | | シカゴ | |

TABLEAU 17 – Alignement d'une chaîne de caractères par application itérative de différences de chaînes. Dans cet exemple, nous recherchons la traduction japonaise de l'anglais *Chicago*. La chaîne *Chicago* ne peut pas être modifiée au cours du traitement et n'est donc pas utilisée pour calculer les LCSubstr. La traduction obtenue est $\widehat{A}_4$ = シカゴ */sikago/*, ce qui est correct.

## 2.2 EXPÉRIENCE D'ALIGNEMENT

### 2.2.1 *Exemples d'alignements*

Nous appliquons la méthode du cosinus à notre corpus parallèle Europarl : anglais-finnois d'une part, et espagnol-français d'autre part. Comme cela est couramment pratiqué dans le domaine, nous avons converti tous les caractères du corpus en minuscules et inséré des espaces entre mots et ponctuations, de sorte que les ponctuations soient considérées comme des mots à part entière. La proportion d'hapax dans ce corpus est comprise entre 37 % pour la partie française et 54 % pour la partie finnoise, ce qui est conforme à ce qu'on trouve généralement dans la littérature (voir chapitre précédent). Le tableau 3 page suivante donne des exemples d'alignements obtenus par la méthode du cosinus sur le corpus espagnol-français en fonction de leurs angles. La qualité des alignements finnois-anglais a tendance à être inférieure, mais ils sont répartis de façon comparable.

D'une façon générale, et comme prévu, plus l'angle associé à un alignement est faible, plus cet alignement a de chances d'être correct : on trouve davantage d'alignements corrects dans l'échantillon d'alignements d'angles nuls (3/5) que dans celui d'alignements d'angles élevés (0/5). Parmi les alignements d'angles nuls, certains sont néanmoins clairement erronés. Par exemple, l'alignement entre *acrimonias* et *essouflerait* illustre le fait que même un alignement d'angle nul n'implique pas nécessairement que les mots sont de bonnes traductions. Il représente simplement la présence réciproque des deux mots dans les mêmes énoncés. Dans ce corpus, les mots espagnols *acrimonias* et *desinflaría* et leurs traductions respectives *acrimonies* et *essouflerait* sont tous les quatre hapax et apparaissent dans le même énoncé :

| | |
|---|---|
| […] el tan importante esfuerzo de cohesión económica y social se *desinflaría* poco a poco y […] la unión caería en […] las *acrimonias* nefastas para el desarrollo de la unión . | […] l' effort si important de cohésion économique et sociale s' *essoufflerait* peu à peu et […] l' union retomberait dans […] les *acrimonies* néfastes pour le développement de l' union . |



la contrainte que la chaîne à aligner ne doit pas être altérée pendant le traitement itératif, c'est-à-dire qu'elle ne doit pas être incluse dans une LCSubstr. Ainsi, en partant de $A_0$ contenant *Chicago*, nous effectuons à chaque étape :

$$\begin{aligned} A_{n+1} &= A_n \ominus E_n \\ \widehat{A}_{n+1} &= \widehat{A}_n \ominus \widehat{E_n} \end{aligned}$$

où $E_n$ est le premier énoncé dans la liste de tous les énoncés triés selon la longueur de leur LCSubstr avec $A_n$. En d'autres termes, parmi tous les énoncés anglais $E_i$, nous sélectionnons celui qui a la plus longue sous-chaîne commune avec $A_n$, et supprimons cette sous-chaîne de $A_n$. Les différences correspondantes sont effectuées au même moment dans la (les) langue(s) cible(s). L'application itérative peut être vue comme un compromis entre la suppression d'une unique LCSubstr et celle de la LCS : les LCSubstr étant la plupart du temps petites par rapport à la taille des énoncés, leur suppression correspond en fait directement à une segmentation en deux parties inégales telle que détaillée dans la section 3.2.2 page 56. Cette approche consiste ainsi par essence à supprimer itérativement de petites unités textuelles d'une plus grande, de façon linéaire, à l'opposé d'une méthode qui segmenterait un énoncé selon un arbre binaire équilibré par exemple. En pratique, nous utilisons des tableaux de suffixes pour rechercher efficacement les énoncés partageant les plus longues LCSubstr.

Le tableau 17 page suivante détaille les étapes d'une exécution du traitement itératif. Dans cet exemple, nous aboutissons à un résultat correct alors que des traductions intermédiaires sont erronées : l'égalité $\widehat{\text{LCSubstr}(A_n, E_n)} = \text{LCSubstr}(\widehat{A}_n, \widehat{E}_n)$ est fausse pour $n = 2$ et $n = 3$. Nous comptons en fait sur le nombre de chemins qui ont mené à un alignement pour en confimer la validité : plusieurs $A_0$ sont possibles pour une même chaîne de départ à aligner, et plusieurs $E_n$ sont possibles pour chaque $n$. Par conséquent, sur de grandes quantités de données, plusieurs solutions peuvent être obtenues pour une même chaîne, chacune pouvant être obtenue un certain nombre de fois. En pratique, il n'est pas possible de traiter les différences de chaînes entre

un texte est $-\log(p)$ où $p$ est la probabilité d'occurrence de $c$. Si la langue était un code, dans la théorie de l'information de Shannon la longueur d'un mot $m$ serait liée à sa fréquence dans la langue : longueur($m$) $\approx -\log(\text{freq}(m))$. Davantage d'expériences seraient à mener pour déterminer le critère le plus efficace, ce qui dépasse le cadre des présentes expériences. Nous nous en tenons donc aux plus longues sous-chaînes communes.

### D.2 APPLICATION ITÉRATIVE

Supposons à présent que nous désirions extraire la traduction de l'anglais *Chicago* en japonais à partir des couples d'énoncés suivants :

$A_0$ = Is␣this␣a␣train␣for␣Chicago?
'Est-ce un train pour chicago ?'

$\widehat{A}_0$ = この列車はシカゴ行きですか。
*/kono ressya wa sikago yuki desu ka./*

B = Is␣this␣price␣correct? 'Ce prix est-il correct ?'

$\widehat{B}$ = この値段で正しいですか。
*/kono nedan de tadasii desu ka./*

C = What␣track␣does␣the␣train␣for␣Boston␣start␣from?
'De quelle voie le train pour Boston part-il ?'

$\widehat{C}$ = ボストン行きの列車は何番から出ますか。
*/bosuton yuki no ressya wa nanban kara de masu ka./*

Avec une application directe de la méthode décrite précédemment, nous n'avons aucune garantie que *Chicago* corresponde à une différence de chaînes. Une façon de résoudre ce problème est d'appliquer la méthode de façon itérative. Nous appliquons les différences de chaînes aux énoncés où *Chicago* apparaît, notés $A_i$, dans le but de les réduire progressivement à la seule chaîne *Chicago*. En appliquant le même traitement en parallèle sur les énoncés cible, ces chaînes devraient également être réduites aux seules traductions de *Chicago*. Nous imposons



| | ESPAGNOL | | FRANÇAIS | ANGLE |
|---|---|---|---|---|
| 122 319 | se '(pron. pers.)' | 2 | apparenter | 1,57 |
| 2 782 | según '(prép. ou adv.)' | 263 | transfrontalière | 1,57 |
| 1 974 | dice '[il] dit' | 42 | jugées | 1,57 |
| 357 | finalidad 'but, finalité' | 218 | manifester | 1,57 |
| 148 | maíz 'maïs' | 4 | ensilé | 1,46 |

(a) Alignements dont l'angle est strictement supérieur à $\pi/4 \approx 0{,}79$, a priori de mauvaise qualité.

| | ESPAGNOL | | FRANÇAIS | ANGLE | |
|---|---|---|---|---|---|
| 1 | tragicómico 'tragi-comique' | 2 | tragi-comique | 0,79 | ✓ |
| 2 | gallegos 'gallicien' | 1 | réadaptations | 0,79 | |
| 1 | sito 'situé' | 2 | berlin-est | 0,79 | |
| 1 | eu.observer.com '(URL)' | 2 | gourmande | 0,79 | |
| 472 | deuda 'dette' | 442 | dette | 0,55 | ✓ |

(b) Alignements dont l'angle est compris entre zéro (exclus) et $\pi/4 \approx 0{,}79$ (inclus), a priori de qualité moyenne.

| | ESPAGNOL | | FRANÇAIS | ANGLE | |
|---|---|---|---|---|---|
| 1 | supercentralización 'surcentralisation' | 1 | surcentralisation | 0,00 | ✓ |
| 1 | consonancias 'consonances' | 1 | consonances | 0,00 | ✓ |
| 1 | acrimonias 'acrimonies' | 1 | essouflerait | 0,00 | |
| 1 | josefina '(nom propre)' | 1 | joséphine | 0,00 | ✓ |
| 1 | -b4-0430 '(identifiant)' | 1 | -b4-0433 | 0,00 | ✓ |

(c) Alignements d'angle nul, a priori de bonne qualité.

TABLEAU 3 – Échantillons d'alignements espagnol-français obtenus par la méthode du cosinus. Les mots sont précédés de leur effectif en corpus. Les angles sont exprimés en radians. On trouve davantage de bons alignements (signalés par un ✓) parmi ceux dont l'angle est faible.

Dans une telle configuration, chaque hapax source se trouve aligné avec chaque hapax cible avec un angle nul.

Ces alignements mot-à-mot ne doivent donc pas être mal interprétés. Les angles obtenus ne reflètent pas la qualité des alignements, mais constituent une mesure de cohésion entre les mots de deux langues différentes. Les alignements dont l'angle est faible ne sont en fait qu'un surensemble des véritables « bons » alignements, parmi tous les couples de mots possibles. Le but de la section suivante est de montrer que les mots rares contribuent beaucoup à cet ensemble.

### 2.2.2 *Distribution des alignements*

La distribution des couples de mots issus des données en fonction de leurs angles est présentée à la figure 6 page ci-contre sous forme d'histogramme. Les alignements sont répartis en fonction de leurs angles dans des classes de même taille. Trois grandes populations émergent.

La première et plus importante, non visible sur la figure, est l'ensemble des couples dont l'angle vaut $\pi/2$. Elle est constituée de couples de mots qui n'apparaissent jamais dans les mêmes énoncés. Nous dénombrons près de 7 milliards de tels couples en espagnol-français et plus de 17 milliards en anglais-finnois. Ces alignements ne peuvent raisonnablement pas constituer de bonnes traductions.

La deuxième population s'étend de $\pi/4$ à $\pi/2$. Elle est constituée de couples de mots qui ne sont a priori pas de bonnes traductions du fait de leurs mauvais scores, comme le montrent les exemples de l'échantillon (a) du tableau 3 page précédente. L'échantillon (b) montre néanmoins que certains alignements d'angle $\pi/4 \approx 0{,}79$ sont corrects.

La troisième population contient les couples dont l'angle est proche de zéro. En fait, bien que la première classe de l'histogramme s'étende au-delà de zéro, elle ne contient que des alignements d'angle nul, et ce pour les deux couples de langues. Nous dénombrons 34 838 alignements espagnol-français dans cette classe (resp. 49 633 en finnois-anglais), dont 84 % contiennent des hapax dans les deux langues (resp.



Nous supposons à nouveau pour simplifier que les LCSubstr sont uniques.

Nous allons à présent appliquer cette différence simultanément sur les énoncés anglais précédents et sur leurs traductions en japonais. Ci-après, $\widehat{\text{A}}$ signifie « traduction de A » :

$$\begin{aligned}
\widehat{\text{A}} &= \text{ドーナツを下さい。} \\
&\quad \textit{/dōnatu o kudasai./} \\
\widehat{\text{B}} &= \text{普通サイズを下さい。} \\
&\quad \textit{/hutuu saizu o kudasai./} \\
\text{LCSubstr}(\widehat{\text{A}},\widehat{\text{B}}) &= \text{を下さい。} \\
\widehat{\text{A}} \ominus \widehat{\text{B}} &= \text{ドーナツ} \\
\widehat{\text{B}} \ominus \widehat{\text{A}} &= \text{普通サイズ}
\end{aligned}$$

En appliquant simultanément les différences entre A et B d'une part et entre $\widehat{\text{A}}$ et $\widehat{\text{B}}$ d'autre part, nous obtenons :

$$\begin{aligned}
\text{,␣please.} &\leftrightarrow \text{を下さい。} \\
\text{I␣would␣like␣a␣donut} &\leftrightarrow \text{ドーナツ} \\
\text{Regular␣size} &\leftrightarrow \text{普通サイズ}
\end{aligned}$$

Ce faisant, nous faisons l'hypothèse que les trois chaînes calculées dans la langue source (LCSubstr(A,B), A ⊖ B et B ⊖ A) sont traductions des chaînes correspondantes dans la langue cible (LCSubstr($\widehat{\text{A}},\widehat{\text{B}}$), $\widehat{\text{A}} \ominus \widehat{\text{B}}$ et $\widehat{\text{B}} \ominus \widehat{\text{A}}$). Soit :

$$\begin{aligned}
\widehat{\text{LCSubstr}(\text{A},\text{B})} &= \text{LCSubstr}(\widehat{\text{A}},\widehat{\text{B}}) \\
\widehat{\text{A} \ominus \text{B}} &= \widehat{\text{A}} \ominus \widehat{\text{B}} \\
\widehat{\text{B} \ominus \text{A}} &= \widehat{\text{B}} \ominus \widehat{\text{A}}
\end{aligned}$$

D'autres opérations seraient envisageables en remplacement de la plus longue sous-chaîne commune. Nous pourrions définir par exemple un critère fondé sur la quantité d'information, car on peut considérer que la quantité d'information d'une chaîne est liée à sa longueur. En effet, la quantité d'information d'une chaîne *c* dans

mune (LCS : *Longest Common Subsequence*) (Hirschberg, 1975; Bergroth et coll., 2000). Étant données deux chaînes A et B, il est toujours possible de déterminer leur(s) plus longue(s) sous-séquence(s) commune(s). Pour simplifier, et bien que cela soit faux dans le cas général, nous supposons que l'opération de LCS n'a qu'une seule solution. Les caractères de cette sous-séquence ne sont pas nécessairement contigus. Considérons par exemple les énoncés anglais suivants, où les espaces sont mises en évidence par le caractère « ␣ » et ont le même statut que tous les autres caractères :

A = I␣would␣like␣a␣donut,␣please. 'Je voudrais un beignet, s'il vous plaît.'

B = Regular␣size,␣please. 'Taille standard, s'il vous plaît.'

La plus longue sous-séquence commune à A et B est ici :

LCS(A,B) = ul␣ie,␣please.

de longueur 14 caractères. Nous pouvons alors définir leurs différences :

A ⊖ B = I␣wodlk␣a␣donut

B ⊖ A = Regarsz

où A ⊖ B = A − LCS(A,B). Plusieurs caractères se trouvent isolés, ce qui résulte en une chaîne mal formée dont l'intérêt est limité. Pour éviter d'aboutir à de telles chaînes, nous remplacons la LCS par la plus longue sous-*chaîne* commune (LCSubstr : *Longest Common Substring*), qui, elle, est formée de caractères contigus. Dans l'exemple précédent, la LCSubstr est :

LCSubstr(A,B) = ,␣please.

de longueur 9 caractères, et beaucoup plus sensée. En supprimant cette sous-chaîne de A et B, nous obtenons :

A ⊖ B = I␣would␣like␣a␣donut

B ⊖ A = Regular␣size



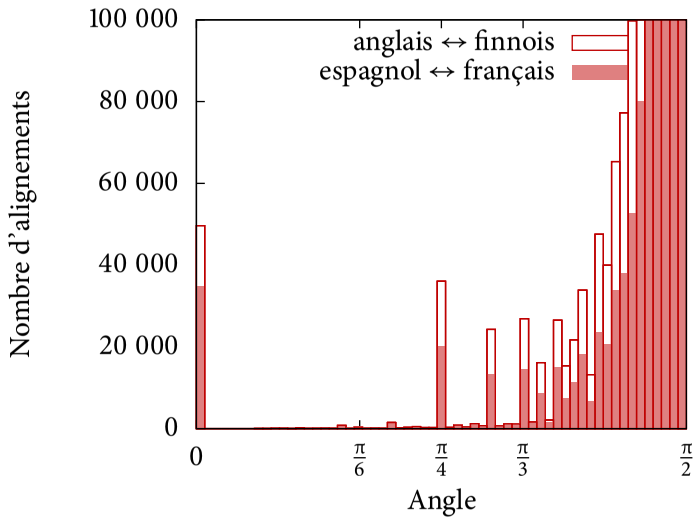



FIGURE 6 – Distribution des alignements obtenus par la méthode du cosinus en fonction de leurs angles.

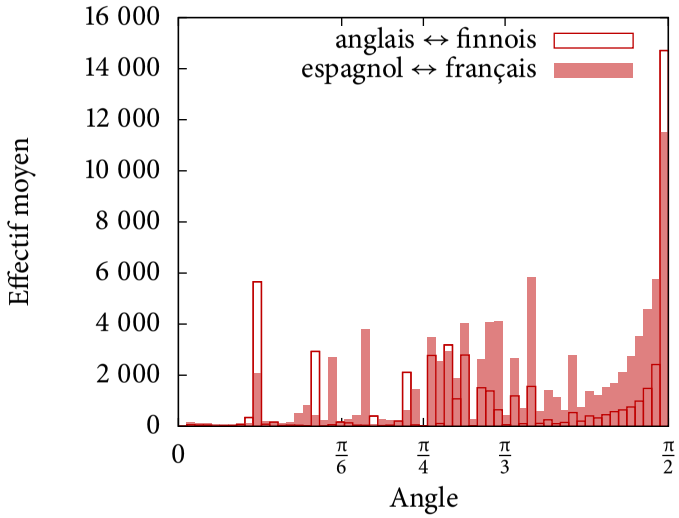



FIGURE 7 – Effectifs moyens des mots composants les alignements obtenus par la méthode du cosinus.

92 %). En écartant les hapax, comme il est couramment pratiqué de façon volontaire ou non, cette population serait quasiment vide : l'efficacité de la méthode du cosinus en termes de bons candidats de traduction ne serait que d'environ 5 000 couples pour 350 000 énoncés au départ, lesquels contiennent plusieurs milliers de formes dans chaque langue — voire plusieurs centaines de milliers dans le cas du finnois.

Si les hapax sont responsables d'une grande proportion des meilleurs alignements, c'est parce que deux hapax apparaissant dans les parties source et cible d'un même énoncé sont, par définition, alignés avec un angle nul.

### 2.2.3 *Effectifs en alignement*

Outre l'existence des trois populations évoquées précédemment, la distribution des alignements de la figure 6 page précédente met en évidence certains intervalles où ne se trouve quasiment aucun alignement. Le plus remarquable est l'intervalle $]0\,;\,\pi/4[$, auquel viennent s'ajouter d'autres de plus en plus petits après $\pi/4$. En fait, les angles des alignements ne peuvent théoriquement prendre que certaines valeurs particulières, car les angles sont calculés à partir des effectifs des mots, qui ne peuvent naturellement prendre que des valeurs discrètes. Cela implique une distribution inégale des alignements en fonction de leurs angles : tous les angles ont tendance à prendre les mêmes valeurs, ce qui se traduit par des pics de population sur l'histogramme, en particulier en zéro et en $\pi/4$.

Nous avons vu précédemment que les mots rares, hapax en tête, étaient à l'origine de l'écrasante majorité des alignements d'angles nuls. D'une façon plus générale, ce sont eux qui sont à l'origine de ces pics. Pour s'en convaincre, nous déterminons l'effectif moyen des mots composant les alignements de chacune des classes précédentes. Ces effectifs moyens sont visibles à la figure 7 page précédente. On constate que les effectifs des mots constituant des alignements dont l'angle correspond à un pic de population à la figure 6 sont très faibles,

# D

## UNE MÉTHODE D'ALIGNEMENT MULTILINGUE PAR DIFFÉRENCES MONOLINGUES

Cette annexe présente une de nos premières tentatives d'alignement multilingue, basée sur l'utilisation parallèle d'opérations monolingues. Notre but est ici de montrer qu'en sélectionnant les énoncés appropriés d'un corpus, toute chaîne, quelle que soit sa longueur, peut être alignée par différences successives. À cette fin, nous adoptons un point de vue négatif en nous concentrant sur les contextes d'une chaîne de caractères plutôt que sur ses propres occurrences. Nous n'utiliserons ainsi comme données d'entrée que les énoncés où ces contextes apparaissent pour calculer les alignements. Afin de démontrer l'universalité de cette approche, nous choisissons de travailler en caractères, dont les bénéfices en traitement automatique des langues ont été explorés par Denoual (2006, chap. II). Nous utilisons à cette occasion un extrait du BTEC (Takezawa et coll., 2002), distribué lors de la campagne d'évaluation de traduction automatique IWSLT 2007 (Fordyce, 2007), et dont une partie est en japonais. Nous traitons ainsi le texte en caractères sans nous soucier de l'existence de caractères de séparation entre mots.

### D.1 DIFFÉRENCES DE CHAÎNES

Pour introduire l'opération que nous utilisons, nous partons d'une technique similaire bien connue, la plus longue sous-séquence com-

| | dan | eng | fin | fra | spa | swe | zho |
|---|---|---|---|---|---|---|---|
| dan | | + 3 | −10 | −15 | + 9 | +8 | − 4 |
| eng | −15 | | −10 | +13 | + 2 | −6 | − 5 |
| fin | + 2 | +36 | | +70 | +53 | +7 | +11 |
| fra | −15 | 0 | − 2 | | + 1 | −3 | + 5 |
| spa | − 9 | +15 | + 3 | +13 | | −2 | +15 |
| swe | − 4 | + 7 | −18 | +19 | + 7 | | − 1 |
| zho | −13 | +16 | 0 | +58 | +31 | +3 | |

TABLEAU 16 – Gain relatif en f-mesure (pourcentages) en utilisant Anymalign à la place de MGIZA++. Nous observons un gain moyen de 7 % par rapport à MGIZA++.



en particulier en zéro, π/4 et π/3. À l'inverse, les effectifs des mots des alignements issus des intervalles où se trouvent peu d'alignements peuvent être très élevés. En fait, pour certaines valeurs, telles qu'aux abords des pics, ils *doivent* être très élevés. Un alignement d'angle (π/4 + ε), où ε est très petit, fait nécessairement intervenir des mots fréquents. C'est le cas par exemple du dernier alignement de l'échantillon (b) du tableau 3 page 35 — les quatre autres ayant un angle de π/4 exactement et ne faisant intervenir que des mots rares. La même chose se produit aux abords de zéro, bien que cela ne soit pas visible sur la figure du fait de la quasi-absence d'alignements dans cette zone : les deuxième, troisième et quatrième classes ne contiennent qu'un seul alignement chacune, typiquement des alignements de ponctuations fréquentes, pour les deux couples de langues. Cela est normal dans la mesure où, comme nous l'avons vu au chapitre précédent, les mots fréquents ne représentent qu'une petite partie du vocabulaire d'un texte. Ainsi, les alignements d'angles nuls, présumés les meilleurs, ne sont obtenus qu'à partir de mots rares, correspondant aux formes les plus nombreuses, suivis de près par les alignements obtenus à partir des mots les plus fréquents, correspondant malheureusement aux formes les moins nombreuses.

Pour confirmer ces faits, nous visualisons les angles des alignements et leur distribution, cette fois en fonction des effectifs des mots source et cible qui les composent. La figure 8 page suivante montre qu'en effet seuls les alignements de couples de mots (rare, rare) et (fréquent, fréquent) ont des angles supérieurs à π/6. Les alignements faisant intervenir un mot de fréquence « intermédiaire », ici entre 10 occurrences et 100 000 occurences environ, ont tous pour leur part des angles supérieurs à π/3, ce qui se traduit par la surface blanche occupant la majeure partie de la figure. Notons au passage la diagonale qui se profile entre les zones en bas à gauche et en haut à droite de la figure, indiquant que les mots ont tendance à s'aligner avec des mots de même effectif, ce qui est naturel en soi. La figure 9 page suivante montre quant à elle que les alignements les plus nombreux sont de type (rare, rare), (rare, fréquent) ou (fréquent, rare), alors que ceux du type

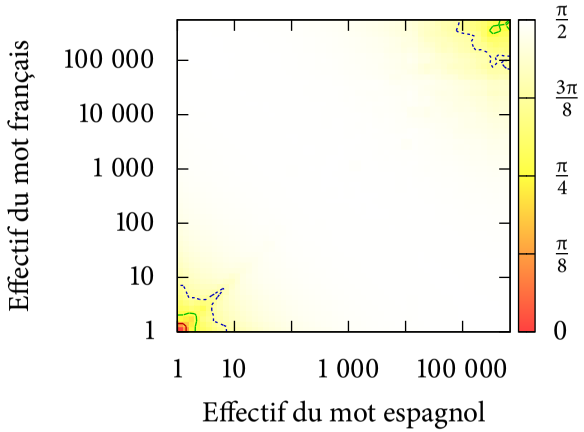



FIGURE 8 – Angle moyen des alignements obtenus par la méthode du cosinus en fonction des effectifs des mots qui les composent. Les meilleurs scores ne sont obtenus qu'à partir de mots très rares ou très fréquents. Rappellons que les meilleurs scores sont les plus faibles. Les courbes de niveau correspondent aux multiples de π/8.

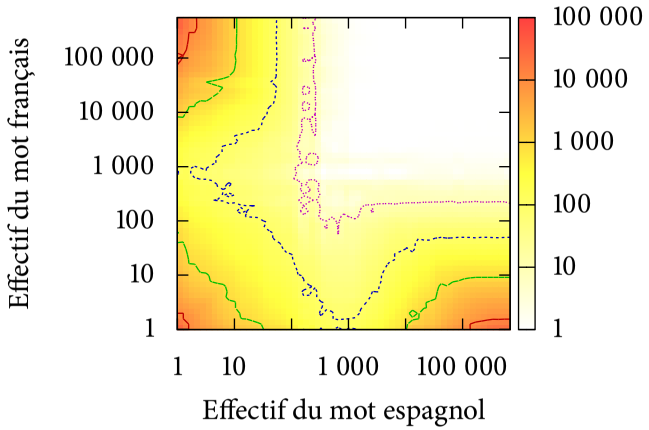



FIGURE 9 – Distribution des alignements obtenus par la méthode du cosinus en fonction des effectifs des mots qui les composent. Les alignements les plus nombreux font intervenir au moins un mot rare. Les courbes de niveau correspondent aux puissances de dix.



| | dan | eng | fin | fra | spa | swe | zho |
|---|---|---|---|---|---|---|---|
| dan | | 46 | 32 | 35 | 37 | 51 | 29 |
| eng | 39 | | 27 | 36 | 42 | 36 | 26 |
| fin | 40 | 34 | | 25 | 28 | 36 | 26 |
| fra | 33 | 43 | 25 | | 45 | 29 | 24 |
| spa | 39 | 46 | 28 | 46 | | 34 | 27 |
| swe | 48 | 43 | 32 | 31 | 33 | | 25 |
| zho | 19 | 18 | 17 | 15 | 17 | 17 | |

TABLEAU 14 – F-mesures (pourcentages) obtenues par MGIZA++ sur 42 couples de langues. La langue source est indiquée dans la première colonne et la langue cible dans la première ligne.

| | dan | eng | fin | fra | spa | swe | zho |
|---|---|---|---|---|---|---|---|
| dan | | −24 | −10 | −19 | −16 | − 7 | −21 |
| eng | −27 | | − 8 | − 7 | − 8 | −21 | − 7 |
| fin | − 5 | + 7 | | +26 | +25 | + 1 | + 2 |
| fra | −19 | − 1 | + 5 | | − 8 | −13 | − 9 |
| spa | −20 | − 4 | +10 | − 9 | | −19 | +15 |
| swe | − 7 | −13 | − 3 | − 6 | −10 | | −11 |
| zho | −25 | −13 | + 2 | + 1 | − 9 | −24 | |

TABLEAU 15 – Gain relatif en f-mesure (pourcentages) en utilisant BerkeleyAligner à la place de MGIZA++. Nous observons une perte moyenne de 7 % par rapport à MGIZA++.

| TÂCHE | ANYMALIGN | MGIZA++ |
|---|---|---|
| fr → en | 0,25 | 0,29 |
| fr → es | 0,32 | 0,36 |
| de → el | 0,15 | 0,16 |
| el → de | 0,14 | 0,16 |
| en → fi | 0,11 | 0,12 |
| fi → en | 0,16 | 0,21 |

TABLEAU 13 – Scores BLEU obtenus par le système Moses à partir des tables de traductions produites par Anymalign et MGIZA++ sur des extraits d'Europarl (Koehn, 2005). Les corpus d'entraînement sont constitués de 200 000 couples d'énoncés longs (environ 30 mots en moyenne). Ici, Anymalign est en retrait de 2,8 points BLEU en moyenne.

## C.2 EN INDUCTION DE LEXIQUES BILINGUES

Nous présentons les résultats des comparaisons des tables de traductions produites par Anymalign, MGIZA++ et BerkeleyAligner avec des lexiques bilingues de référence en utilisant le corpus de la Bible (Resnik et coll., 1999). Près de 30 000 énoncés longs (environ 29 mots en moyenne) en sept langues sont utilisés en entrée des aligneurs. Nous comparons les f-mesures obtenues par Anymalign et BerkeleyAligner relativement à MGIZA++ sur chacun des 42 couples de langues de notre corpus.



(fréquent, fréquent) sont quasiment inexistants. Cela est — toujours — dû à la singularité des mots fréquents, et à l'abondance des mots rares. Avec une méthode fondée sur les mots rares, nous pourrions donc produire des alignements de bonne qualité, *et* en grande quantité.

## 2.3 LE CAS DES HAPAX

Les mots rares étant à l'origine de la majorité des alignements les plus prometteurs, et étant de surcroît les plus abondants, nous proposons d'étudier de plus près la forme sous laquelle ils se manifestent le plus volontiers : l'hapax.

### 2.3.1 *Distribution des hapax*

Comme illustré précédemment (échantillon (c) du tableau 3 page 35), certains alignements d'angle nul, alignements qui sont principalement dûs aux hapax, ne sont pas valides. Cela se produit car, lorsqu'un énoncé contient plus d'un hapax, chaque hapax source se trouve aligné avec chaque hapax cible avec un angle nul. D'un autre côté, s'il n'y a qu'un seul hapax source et un seul hapax cible dans un énoncé, l'alignement résultant devrait — a priori — être correct. Pour déterminer si cette configuration est courante, nous étudions la distribution des hapax dans nos données.

Les fréquences d'apparition des hapax dans notre corpus sont présentées au tableau 4 page suivante. Bien que le nombre d'hapax représente généralement près de la moitié des formes, ils n'apparaissent que dans 5 à 9 % des énoncés, à l'exception du finnois où les hapax sont monnaie courante : ils apparaissent en effet dans 31 % des énoncés. Plus important, la plupart (entre 71 et 85 %) des énoncés qui contiennent un hapax n'en contiennent qu'un seul, la moyenne étant proche de 1,2 hapax par énoncé. Le cas où un énoncé source et sa traduction ne contiennent tous deux qu'un hapax ne devrait donc rien avoir d'extraordinaire.

Parmi les alignements d'angle nul constitués exclusivement d'hapax, une quantité non négligeable (6 230 en espagnol-français soit 21 %,

| LANGUE | AU MOINS 1 HAPAX | 1 HAPAX | HAPAX/ÉN. |
|---|---|---|---|
| espagnol | 9 % | 85 % | 1,2 ± 0,7 |
| français | 5 % | 85 % | 1,2 ± 0,8 |
| anglais | 5 % | 85 % | 1,2 ± 0,8 |
| finnois | 31 % | 71 % | 1,4 ± 0,9 |

TABLEAU 4 – Fréquences d'apparition des hapax dans notre corpus Europarl : pourcentage d'énoncés contenant au moins un hapax, pourcentage d'énoncés contenant un hapax exactement parmi ceux contenant au moins un hapax, et nombre moyen d'hapax par énoncé.

7 268 en finnois-anglais soit 16 %) correspond en fait à des alignements d'hapax provenant d'énoncés contenant exactement un hapax dans les deux langues. Ces alignements suffisent donc à couvrir 12 % du vocabulaire anglais, 7 % du vocabulaire espagnol, 8 % du vocabulaire français, mais seulement 2 % du vocabulaire finnois. Le tableau 5 page suivante présente des échantillons de ces alignements. Les alignements espagnol-français sont sans doute parmi les meilleurs qui soient : rejeter ces alignements sous prétexte qu'ils sont constitués d'hapax serait une faute. La qualité n'est malheureusement pas aussi bonne du côté des alignements anglais-finnois : la plupart sont erronés [1], ce qui s'explique par le fait qu'on tente de trouver des correspondances lexicales au grain mot entre une langue isolante, l'anglais, et une langue agglutinante, le finnois, le tout à l'aide d'une méthode volontairement « naïve ». Cela étant dit, si une méthode aussi simple que celle du cosinus permet d'obtenir d'aussi bons résultats avec les hapax sur des langues morphologiquement proches, on est en droit d'espérer que des méthodes

1 Nous ne parlons pas finnois. Les traductions données au tableau 5 ont été obtenues à partir d'un dictionnaire, des alignements phrastiques du corpus Europarl et de *Google Traduction*…

# C

# RÉSULTATS D'EXPÉRIENCES COMPLÉMENTAIRES

Cette annexe complète le chapitre 5. Nous donnons un aperçu des résultats obtenus en traduction automatique et en induction de lexiques bilingues sur quelques tâches supplémentaires.

## C.1 EN TRADUCTION AUTOMATIQUE

| TÂCHE | ANYMALIGN | MGIZA++ |
|---|---|---|
| IWSLT 2007 : ja → en | 0,46 | 0,45 |
| IWSLT 2008 : ar → en | 0,37 | 0,41 |
| IWSLT 2008 : zh → en | 0,32 | 0,32 |
| IWSLT 2008 : zh → es | 0,25 | 0,24 |

TABLEAU 12 – Scores BLEU obtenus par le système Moses à partir des tables de traductions produites par Anymalign et MGIZA++ sur des extraits du BTEC (Takezawa et coll., 2002). Les corpus d'entraînement sont ceux distribués lors des campagnes d'évaluation IWSLT (Fordyce, 2007; Paul, 2008) et sont constitués de 20 000 à 40 000 couples d'énoncés courts (environ 10 mots en moyenne). Ici, Anymalign est en retrait d'un point BLEU en moyenne.

```
# Main loop (cont.)
    # For each group of words (cont.)
        # For each line of the subcorpus (cont.)

            # We get alignments only if both the source
            # and the target parts actually contain words.
            # If so, increase alignment count.

            if sourceAl and targetAl:
                alignment = "%s\t%s" % (" ".join(sourceAl),
                                        " ".join(targetAl))
                if alignment not in allAlignments:
                    allAlignments[alignment] = 0
                allAlignments[alignment] += 1

            if sourceCont and targetCont:
                alignment = "%s\t%s" % (" ".join(sourceCont),
                                        " ".join(targetCont))
                if alignment not in allAlignments:
                    allAlignments[alignment] = 0
                allAlignments[alignment] += 1

# End of main loop

# Sort all alignments according to their count
# and output everything
allAlignments = allAlignments.items()
allAlignments.sort(key=lambda x:x[1], reverse=True)
for alignment, count in allAlignments:
    print "%s\t%i" % (alignment, count)
```



| ESPAGNOL | FRANÇAIS | |
|---|---|---|
| descolonizar ‘décoloniser’ | décolonisé | ✓ |
| predeterminarán ‘prédéterminer’ | prédéterminerer *(sic)* | ✓ |
| wallner ‘(nom propre)’ | wallner | ✓ |
| h-0818 ‘(identifiant)’ | h-0818 | ✓ |
| fantoche ‘fantoche’ | oubliette | |
| burns ‘(nom propre)’ | burns | ✓ |
| pseudojurídicos ‘pseudo-juridiques’ | pseudo-juridiques | ✓ |
| h-0484 ‘(identifiant)’ | h-0484 | ✓ |
| antimaastrichtiana ‘anti-maastrichtienne’ | anti-maastrichtienne | ✓ |
| archiconocidos ‘archiconnus’ | archiconnus | ✓ |

(a) Espagnol ↔ français

| ANGLAIS | FINNOIS | |
|---|---|---|
| non-racist ‘non raciste’ | popular ‘[parti] populaire’ | |
| unsurpassed ‘inégalé’ | charmi ‘charme (qualité)’ | |
| south-north ‘nord-sud’ | pohjoinen-etelä-politiikan ‘politique nord-sud’ | |
| h-0246 ‘(identifiant)’ | h-0246 ‘(identifiant)’ | ✓ |
| h-0621 ‘(identifiant)’ | h-0621 ‘(identifiant)’ | ✓ |
| 58.8 ‘58,8 [%]’ | valkoturskalajien ‘espèces de poisson blanc’ | |
| anti-women ‘phallocrate’ | saudi-arabialla ‘l’Arabie Séoudite’ | |
| featherweight ‘poids-plume’ | höyhensarjalainen ‘un poids-plume’ | ✓ |
| h-0466 ‘(identifiant)’ | h-0466 ‘(identifiant)’ | ✓ |
| 25.7 ‘25,7 [millions]’ | paluuavusta ‘pour l’aide au retour’ | |

(b) Anglais ↔ finnois

TABLEAU 5 – Échantillons d’alignements d’hapax issus d’énoncés ne contenant qu’un seul hapax. Un seul alignement est erroné dans l’échantillon espagnol-français : *fantoche* ↔ *oubliette*. La majorité (6/10) des alignements anglais-finnois sont quant à eux erronés.

| LANGUE | FORMES | OCC./ÉN. = 1 | MOYENNE |
|---|---|---|---|
| espagnol | 93 043 | 82 368 (89 %) | 1,011 ± 0,079 |
| français | 73 695 | 63 544 (86 %) | 1,012 ± 0,079 |
| anglais | 58 342 | 49 265 (84 %) | 1,014 ± 0,081 |
| finnois | 292 894 | 277 104 (95 %) | 1,007 ± 0,065 |

TABLEAU 6 – Quantités de mots dont le nombre d'occurrences par énoncé vaut un. La plupart des mots n'apparaissent qu'une seule fois par énoncé.

d'alignement plus évoluées feront au moins aussi bien. Nous nous y intéresserons davantage dans les chapitres suivants.

### 2.3.2 *Hapax en corpus, hapax en énoncé*

Nous nous sommes intéressés jusqu'ici au cas des *hapax en corpus*, mots qui n'apparaissent qu'une seule fois au sein de chaque ensemble monolingue de notre corpus parallèle. Nous montrons à présent que la quasi-totalité des mots d'un tel ensemble monolingue sont en fait des *hapax en énoncé*, mots qui n'apparaissent qu'une seule fois au sein d'un énoncé. Cette propriété nous permettra de justifier une simplification de la méthode du cosinus. Cette simplification est couramment employée, mais nous n'en avons trouvé aucune justification. C'est ce que nous proposons ici. Nous mettrons en outre cette simplification à profit dans les chapitres suivants.

Nous déterminons le nombre d'occurrences moyen des mots dans les énoncés où ils apparaissent en divisant le nombre total d'occurrences d'un mot dans le corpus par le nombre d'énoncés dans lesquels il apparaît. Un hapax en corpus étant nécessairement un hapax en énoncé dans l'énoncé où il apparaît, le nombre d'hapax en énoncé dans un énoncé donné est supérieur ou égal au nombre d'hapax en corpus. Les résultats sont présentés dans le tableau 6 ci-dessus. En moyenne,



```
# Main loop (cont.)

    # For each group of words, make a new pass on the subcorpus
    # to extract alignments and their contexts
    for vec in vec_words:
        sourceWords, targetWords = vec_words[vec]
        if not targetWords:
            # target part is empty -> no alignment
            continue

        sourceSet = set(sourceWords) # Speed up searches
        targetSet = set(targetWords)

        # For each line of the subcorpus
        for lineId in vec:
            sourceSentence, targetSentence = corpus[lineId]
            # Same words as in <sourceSet>, but ordered
            sourceAl = []
            targetAl = []
            # Complementary of <sourceAl> on the line
            sourceCont = []
            targetCont = []

            for word in sourceSentence:
                if word in sourceSet:
                    sourceAl.append(word)
                else:
                    sourceCont.append(word)

            for word in targetSentence:
                if word in targetSet:
                    targetAl.append(word)
                else:
                    targetCont.append(word)
```

```
# Main loop
for i in xrange(NB_SAMPLES):
    # Select a random subcorpus
    subcorpusSize = random.randrange(0, len(corpus))
    selection = random.sample(xrange(len(corpus)),
                              subcorpusSize)

    # Assign to each word of the subcorpus
    # the line ids it appears on
    sourceWord_vec = {} # {string:  [lineNo, ...], ...}
    targetWord_vec = {}
    for lineId in selection:
        sourceSentence, targetSentence = corpus[lineId]

        for word in sourceSentence:
            if word not in sourceWord_vec:
                sourceWord_vec[word] = []
            sourceWord_vec[word].append(lineId)

        for word in targetSentence:
            if word not in targetWord_vec:
                targetWord_vec[word] = []
            targetWord_vec[word].append(lineId)

    # Group words according to the lines they appear on
    # {tupleOfLineNos:  ([srcWord, ...], [tgtWord, ...]), ...}
    vec_words = {}
    for word in sourceWord_vec:
        vec = tuple(sourceWord_vec[word])
        if vec not in vec_words:
            vec_words[vec] = ([], [])
        vec_words[vec][0].append(word)
    for word in targetWord_vec:
        vec = tuple(targetWord_vec[word])
        if vec in vec_words:
            vec_words[vec][1].append(word)
        # else:  there will not be any alignment
        # since the source part is empty
```



près de 9 mots sur 10 ont un nombre d'occurrences valant un dans les énoncés où ils apparaissent, hapax en corpus inclus. Les autres sont typiquement des mots-outils, très fréquents mais peu nombreux. Par exemple, la forme la plus fréquente dans la partie française de notre corpus Europarl, « de », qui apparaît dans près de 70 % des énoncés, a un nombre d'occurrences moyen de 2,17 dans ces énoncés. Dans l'ensemble cependant, le nombre d'occurrences moyen d'un mot par énoncé est de 1,011, ce qui est très proche de un : la quasi-totalité des mots sont des hapax en énoncé dans les énoncés où ils apparaissent.

Ces résultats dépendent bien entendu directement de la longueur des énoncés du corpus parallèle utilisé : le nombre d'hapax en énoncé augmente avec la longueur de l'énoncé. La définition d'énoncé, rappelons-le, est proche de celle de la phrase. Notons néanmoins que les énoncés du corpus que nous utilisons ici sont relativement longs : près de 30 mots en moyenne. Les chiffres que nous avons obtenus dans cette expérience demeurent donc valables avec un corpus dont les énoncés sont de longueur conséquente.

### 2.3.3 *Simplification de l'alignement à l'aide des hapax*

Les mots étant presque tous des hapax en énoncé, renseigner la présence ou l'absence d'un mot dans un énoncé plutôt que son nombre d'occurrences suffit amplement lors du calcul de l'angle entre deux mots. Cela revient à assimiler tous les mots du corpus à des hapax en énoncé. Cela est bien entendu faux pour les mots les plus fréquents, mais ceux-ci étant très peu nombreux, nous allons voir que leur influence est négligeable. Avec cela, les composantes des vecteurs $\vec{m}_s$ et $\vec{m}_c$ prennent leurs valeurs dans l'ensemble {0, 1}, et l'égalité 2.1 page 31 se simplifie en :

$$\text{angle}(\vec{m}_s, \vec{m}_c) = \text{acos}\left(\frac{|\,E_s \cap E_c\,|}{\sqrt{|\,E_s\,| \times |\,E_c\,|}}\right) \tag{2.2}$$

où $E_s$ est l'ensemble des énoncés du corpus source où $\vec{m}_s$ apparaît (même chose pour $E_c$). Le cosinus est souvent exprimé sous cette forme

lorsque les éléments des vecteurs ne peuvent prendre que des valeurs binaires. L'expression du cosinus simplifié, c'est-à-dire l'argument de acos, s'apparente à celle du coefficient de Dice :

$$\text{dice}(E_s, E_c) = \frac{2\,|\,E_s \cap E_c\,|}{|\,E_s\,| + |\,E_c\,|} = \frac{|\,E_s \cap E_c\,|}{\frac{1}{2}(|\,E_s\,| + |\,E_c\,|)}$$

Le dénominateur de l'expression du cosinus est en fait la moyenne géométrique des cardinaux de $E_s$ et $E_c$, tandis que celui de l'expression du coefficient de Dice en est la moyenne arithmétique, qui est toujours supérieure. Le cosinus simplifié est donc toujours supérieur ou égal au coefficient de Dice, lequel est lui-même supérieur ou égal au coefficient de Jaccard [2]. On a donc :

$$0 \leqslant \text{jaccard} \leqslant \text{dice} \leqslant \text{cosinus simplifié} \leqslant 1$$

Les trois coefficients sont donc monotoniquement équivalents : ils fourniront les mêmes résultats s'ils ne sont utilisés que pour ordonner les alignements en fonction de leurs scores. Le cosinus apporte ainsi peu à l'alignement bilingue par rapport aux coefficients de Jaccard et de Dice, plus simples, car son principal avantage, qui est de prendre en compte le nombre d'occurrences d'un mot dans un énoncé, est atténué par le fait que tous les mots tendent à être des hapax en énoncé.

Nous montrons à présent empiriquement qu'une telle simplification n'altère pas la qualité des alignements produits. Nous effectuons pour cela une comparaison systématique entre les alignements obtenus par la méthode du cosinus originale et la version simplifiée, la première servant de référence. Remarquons d'abord que la simplification ne produit théoriquement pas de nouvel alignement au sein des deux populations dont l'angle est différent de $\pi/2$ par rapport à la méthode originale, ce qui est vérifié en pratique. Ses effets se limitent à une modification de l'angle pour les couples de mots où l'un des mots au moins n'est pas un hapax en énoncé. L'angle d'un couple (hapax en corpus, hapax en corpus), quant à lui, ne change pas.

2 $\text{jaccard}(E_s, E_c) = \frac{|E_s \cap E_c|}{|E_s \cup E_c|} = \frac{\text{dice}(E_s, E_c)}{2 - \text{dice}(E_s, E_c)}$

# B

## MINIMALIGN.PY

```
#!/usr/bin/python
"""minimalign.py:  minimal version of anymalign.py

Adrien Lardilleux <Adrien.Lardilleux@info.unicaen.fr>
http://users.info.unicaen.fr/~alardill/anymalign/
"""

import sys
import random

NB_SAMPLES = 10 # The larger, the more alignments

# Read input file and load bicorpus into memory, as a list
# of pairs of sentences (1 sentence = 1 list of words)
sourceFile = open(sys.argv[1], 'r')
targetFile = open(sys.argv[2], 'r')
corpus = zip((line.split() for line in sourceFile),
             (line.split() for line in targetFile))
sourceFile.close()
targetFile.close()

# Simple counter {alignmentString:  integerCount, ...}
allAlignments = {}
```

Sur les dizaines de millions d'angles d'alignements inférieurs à $\pi/2$, les écarts relatifs entre l'angle original et celui obtenu par la méthode simplifiée sont de 0,07 % ± 0,67 % en espagnol-français et de 0,08 % ± 0,83 % en finnois-anglais. Cette différence en angle entre la méthode du cosinus originale et la méthode simplifiée est négligeable et ne devrait pas affecter la qualité de tâches subséquentes. La version simplifiée est également beaucoup plus rapide que l'originale : dans notre expérience, le temps nécessaire à l'alignement de tous les couples de mots passe de l'ordre de l'heure à celui de la minute. La vitesse ne constitue pas une priorité en alignement sous-phrastique, car cette tâche n'est généralement effectuée qu'une seule fois en amont d'autres traitements ; mais de deux systèmes offrant des performances égales, autant privilégier le plus rapide. Nous ne nous priverons donc pas d'assimiler tous les mots d'un texte à des hapax en énoncé dans les chapitres suivants.

## RÉSUMÉ

Nous avons mis en évidence dans ce chapitre les trois points suivants :

- les alignement mot-à-mot de meilleurs scores sont constitués de mots très rares ou très fréquents, mais pas de mots de fréquence intermédiaire ;
- les alignements constitués de deux mots rares constituent la majorité des alignements mot-à-mot de meilleurs scores ;
- dans un corpus parallèle, même dont la longueur des énoncés est grande, il est inutile de prendre en compte le nombre d'occurrences exact des mots car ils ont presque tous tendance à être des hapax en énoncé.

Sortie XML (TMX : *Translation Memory eXchange*) : 22 langues à partir de 80 000 énoncés du JRC-Acquis (Steinberger et coll., 2006).

```
<?xml version="1.0"?>
<tmx version="1.4">
<header creationtool="anymalign" creationtoolversion="2.3 (July 20th 2009)"
 datatype="plaintext" segtype="phrase" adminlang="en-us" srclang="*all*"
 o-tmf="none" />
<body>
<tu>
 <prop type="freq">12</prop>
 <prop type="probas">1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
 </prop>
 <prop type="lexWeights">1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
  1.000000</prop>
 <tuv xml:lang="bg"><seg>&#1095;&#1083;&#1077;&#1085;</seg></tuv>
 <tuv xml:lang="cs"><seg>&#269;l&#225;nek</seg></tuv>
 <tuv xml:lang="da"><seg>artikel</seg></tuv>
 <tuv xml:lang="de"><seg>artikel</seg></tuv>
 <tuv xml:lang="el"><seg>&#940;&#961;&#952;&#961;&#959;</seg></tuv>
 <tuv xml:lang="en"><seg>article</seg></tuv>
 <tuv xml:lang="es"><seg>art&#237;culo</seg></tuv>
 <tuv xml:lang="et"><seg>artikkel</seg></tuv>
 <tuv xml:lang="fi"><seg>artikla</seg></tuv>
 <tuv xml:lang="fr"><seg>article</seg></tuv>
 <tuv xml:lang="hu"><seg>cikk</seg></tuv>
 <tuv xml:lang="it"><seg>articolo</seg></tuv>
 <tuv xml:lang="lt"><seg>straipsnis</seg></tuv>
 <tuv xml:lang="lv"><seg>pants</seg></tuv>
 <tuv xml:lang="mt"><seg>artikolu</seg></tuv>
 <tuv xml:lang="nl"><seg>artikel</seg></tuv>
 <tuv xml:lang="pl"><seg>artyku&#322;</seg></tuv>
 <tuv xml:lang="pt"><seg>artigo</seg></tuv>
 <tuv xml:lang="ro"><seg>articolul</seg></tuv>
 <tuv xml:lang="sk"><seg>&#269;l&#225;nok</seg></tuv>
 <tuv xml:lang="sl"><seg>&#269;len</seg></tuv>
 <tuv xml:lang="sv"><seg>artikel</seg></tuv>
</tu>
```

⋮

(× 200 alignements [`<tu>` : *translation unit*] obtenus en une minute)

Sortie texte visionnée dans un tableur : 12 langues à partir de 5 000 énoncés de la Bible (Resnik et coll., 1999). Les scores ont été supprimés.

| ceb | dan | grc | eng | fin | fra | ind | lat | spa | swe | vie | zho |
|---|---|---|---|---|---|---|---|---|---|---|---|
| pedro | peter | πετρος | peter | pietari | pierre | petrus | petrus | pedro | petrus | phêrô | 彼得 |
| pablo | paulus | παυλος | paul | paavali | paul | paulus | paulus | pablo | paulus | phaolô | 保羅 |
| pilato | pilatus | πιλατος | pilate | pilatus | pilate | pilatus | pilatus | pilato | pilatus | philatô | 彼拉多 |
| zabulon | sebulons | ζαβουλων | zebulon | sebulonin | zabulon | zebulon | zabulon | zabulón | sabulons | zabulôn | 西布倫 |
| alfeo | alfæus | αλφαιου | alphaeus | alfeuksen | alphée | alfeus | alphei | alfeo | alfeus | alphê | 亞勒腓 |
| moises | moses | μωσης | moses | mooses | moïse | musa | moses | moisés | moses | môsê | 摩西 |
| simon | simon | σιμων | simon | simon | simon | simon | simon | simón | simon | simôn | 西門 |
| pilato | pilatus | πιλατος | pilate | pilatus | pilate | pilatus | pilatus | pilato | pilatus | philatô | 彼拉多 |
| arquipo | arkippus | αρχιππου | archippus | arkippukselle | archippe | arkhipus | archippo | arquipo | arkippus | archippô | 亞基布 |
| marta | martha | μαρθα | martha | martta | marthe | marta | martha | marta | martha | martha | 馬大 |
| juan | johannes | ιωαννης | john | johannes | jean | yohanes | johannes | juan | johannes | yoan | 約翰 |
| elisabet | elisabeth | ελισαβετ | elisabeth | elisabet | élisabeth | elisabet | elisabeth | elisabet | elisabet | elisabet | 沙伯 |
| herodes | herodes | ηρωδης | herod | herodes | hérode | herodes | herodes | herodes | herodes | hêrôđe | 希律 |
| igsoon | brødre | αδελφοι | brethren | veljet | frères | saudara-sauda | fratres | hermanos | bröder | hõi | 弟兄們 |
| escriba | skriftkloge | γραμματεις | scribes | kirjanoppineet | scribes | ahli-ahli | scribae | escribas | skriftlärde | ký luc | 文士 |
| fariseo | farisæerne | φαρισαιοι | pharisees | farisukset | pharisiens | farisi | pharisaei | fariseos | fariséerna | biệt phái | 法利賽 |
| cordero | lammets | αρνιου | lamb | karitsan | agneau | domba | agni | cordero | lammets | chiên | 羔羊 |
| capernaum | kapernaum | καπερναουμ | capernaum | kapernaumin | capernaüm | kapernaum | capharnaum | capernaúm | kapernaum | capharnaum | 迦百農 |
| dios | guds | θεου | god | jumalan | dieu | allah | dei | dios | guds | chúa | 天數 |
| damgo | dræm | οναρ | dream | unessa | songe | mimpi | somnis | sueños | drömmen | mộng | 夢中 |
| jesus | jesus | ιησους | jesus | jeesus | jésus | yesus | iesus | jesús | jesus | đức yêsu | 耶穌 |
| aleluya | halleluja | αλληλουια | alleluia | halleluja | alléluia | haleluya | alleluia | aleluya | halleluja | hôi lê lu gia | 哈利路亞 |
| escriba | skriftkloge | γραμματεις | scribes | kirjanoppineet | scribes | ahli-ahli taurat | scribae | escribas | skriftlärde | ký luc | 文士 |
| fariseo | farisæerne | φαρισαιοι | pharisees | farisukset | pharisiens | farisi | pharisaei | fariseos | fariséerna | biệt phái | 法利賽 |
| maria | maria | μαριαμ | mary | maria | marie | maria | maria | maría | maria | maria | 馬利亞 |
| tarso | tarsus | ταρσον | tarsus | tarsolainen | tarse | tarsus | tarsum | tarso | tarsus | tarsô | 大數 |
| galilea | galilæa | γαλιλαιας | galilee | galilean | galilée | galilea | galilaeae | galilea | galileen | galilê | 加利利 |
| dios | guds | θεου | god | jumalan | dieu | allah | dei | dios | guds | thiên | 神 |
| juan | johannes | ιωαννην | john | johanneksen | jean | yohanes | iohannem | juan | johannes | yoan | 約翰 |
| cornelio | kornelius | κορνηλιος | cornelius | kornelius | corneille | kornelius | cornelius | cornelio | kornelius | cornêliô | 哥尼流 |
| hari | konge | βασιλευς | king | kuningas | roi | raja | rex | rey | konung | vua | 王 |
| saulo | saulus | σαυλος | saul | saul | saül | saulus | saulus | saulo | saulus | saulô | 掃羅 |
| derbe | derbe | δερβην | derbe | derbeen | derbe | derbe | derbe | derbe | derbe | đerbê | 特庇 |
| dalmanuthas | dalmanutha | δαλμανουθα | dalmanutha | dalmanutan se | dalmanutha | dalmanutha | lystram | dalmanuta | dalmanuta | đalmanutha | 多馬 |
| herodes | herodes | ηρωδης | herod | herodes | hérode | herodes | herodes | herodes | herodes | hêrôđe | 希律 |
| benjamin | benjamin | βενιαμιν | benjamin | benjaminin | benjamin | benyamin | benjamin | herodes | benjamins | benyamin | 便雅憫 |
| judas | judas | ιουδας | judas | judas | judas | yudas | judas | judas | judas | yuđa | 猶大 |
| siria | syrien | συριαν | syria | syriaan | syrie | siria | syriam | siria | syrien | syri | 敘利亞 |
| agripa | agrippa | αγριππα | agrippa | agrippa | agrippa | agripa | agrippa | agripa | agrippa | agrippa | 亞基帕 |
| dragon | dragen | δρακων | dragon | lohikäärme | dragon | naga | draco | dragón | draken | rồng | 龍 |
| sambingay | lignelse | παραβολην | parable | vertauksen | parabole | perumpamaan | parabolam | parabola | liknelse | vi dụ | 比喻 |
| egipto | ægypten | αιγυπτω | egypt | egyptissä | égypte | mesir | aegypto | egipto | egypten | cập | 埃及 |
| elias | elias | ηλιας | elias | elias | élie | elia | helias | elías | elias | êlya | 以利 |
| santiago | jakob | ιακωβον | james | jaakobin | jacques | yakobus | iacobum | jacobo | jakob | yacôbê | 雅各 |
| salome | salome | σαλωμη | salome | salome | salomé | salome | salome | salomé | salome | salômê | 羅米 |
| magdalena | magdalene | μαγδαληνη | magdalene | magdalaena | magdala | magdalena | magdalene | magdalena | magdala | magđala | 大拉 |
| babilonia | babylon | βαβυλωνι | babylon | babylonissa | babylone | babel | babylon | babilonia | babylon | babylôn | 巴比倫大 |
| silas | silas | σιλας | silas | silas | silas | silas | silas | silas | silas | sila | 西拉 |
| cornelio | kornelius | κορνηλιος | cornelius | kornelius | cornelie | kornelius | cornelius | cornelio | kornelius | cornêliô | 哥尼 |

… (× 200 alignements obtenus en une minute)

# 3

## VERS DE L'ALIGNEMENT BASSES FRÉQUENCES

Fort de nos résultats sur l'impact des mots de faible effectif en alignement, nous allons à présent montrer qu'ils peuvent servir de fondation à une méthode d'alignement complète. Nous prenons ainsi délibérément le contre-pied des approches affirmant que les termes de faible effectif ne peuvent mener qu'à de mauvais résultats. Ce chapitre s'articule autour de deux questions : *comment* aligner, et *quoi*.

SOMMAIRE

## 3.1 COMMENT ALIGNER AVEC LES BASSES FRÉQUENCES ?

### 3.1.1 *Briser un vieux cercle vicieux…*

Nous avons vu que les mots rares sont souvent rejetés des tâches d'alignement parce que de moindre significativité statistique. Ainsi, seuls les mots ayant un nombre d'occurrences suffisant sont utilisés pour aligner. Pour aligner les mots rares, une solution évidente est la suivante : il suffit d'augmenter la quantité de données (réelles) en entrée, de sorte que l'effectif de chacun de ces mots rares augmente également. Devenus plus fréquents, ils peuvent être alignés. Cependant, en ajoutant de nouvelles données, de nouveaux mots rares sont apparus, mots qu'on est bien évidemment incapable d'aligner… à moins d'ajouter de nouvelles données. Cela est sans fin, car comme nous l'avons vu au chapitre 1 avec la figure 5 page 25, la proportion d'hapax dans un texte est constante, quelle que soit sa taille, dès que celle-ci est supérieure à quelques énoncés. À l'inverse, une méthode reposant sur l'exploitation des mots rares ne nécessiterait pas l'ajout perpétuel de données, bien au contraire : *supprimer* des données en entrée, de sorte que les mots fréquents deviennent rares, suffirait pour s'acquitter de la tâche d'alignement. L'ajout de données est potentiellement infini, pas leur suppression.

Par conséquent, nous ne considérerons pas un corpus comme un tout constitué d'un nombre fini d'événements, à chacun de ces événements étant associée une probabilité fixée une fois pour toutes. En supprimant des données du corpus d'entrée, comme nous en avons l'intention, c'est un *nouveau corpus* qui est constitué. Nous avons donc en fait à disposition de multiples corpus d'entrée possibles. Le nombre de sous-corpus d'un corpus de $n$ énoncés est exactement $2^n - 1$, chaque sous-corpus possédant son propre ensemble d'événements et de probabilités associées. En un certain sens, supprimer des données revient à en ajouter encore davantage ! En outre, les événements rares gagnent en pertinence, car ils se produisent dorénavant dans plusieurs sous-corpus, donc plusieurs fois. Autrement dit, plus aucun événement n'est rare.

# A

## EXEMPLES DE SORTIES D'ANYMALIGN

Sortie HTML visionnée dans un navigateur Web : 5 langues à partir de 20 000 énoncés du BTEC (Takezawa et coll., 2002), distribués lors des campagnes d'évaluation de traduction automatique IWSLT (Fordyce, 2007; Paul, 2008, 2009). Les parties chinoise et japonaise ont préalablement été segmentées en mots.

| No | Freq. | Translation probabilities | Lexical weights | es | en | ar | zh | ja |
|---|---|---|---|---|---|---|---|---|
| 1 | 3279 | 0.71 0.72 0.71 0.72 0.94 | 1.00 1.00 1.00 1.00 0.64 | . | . | . | 。 | 。 |
| 2 | 457 | 0.10 0.10 0.10 0.10 0.93 | 1.00 1.00 1.00 1.00 0.41 | . | . | . | 。 | です。 |
| 3 | 161 | 0.72 0.65 0.61 0.85 0.65 | 0.99 0.95 0.94 0.75 0.82 | Dónde | Where | أين | 哪 | どこ |
| 4 | 143 | 0.03 0.03 0.03 0.03 0.86 | 1.00 1.00 1.00 1.00 0.35 | . | . | . | 。 | ます。 |
| 5 | 125 | 0.93 0.94 0.92 0.93 0.91 | 0.99 0.95 0.99 0.90 0.88 | Japón | Japan | اليابان | 日本 | 日本 |
| 6 | 81 | 0.95 0.95 0.98 0.98 0.95 | 0.99 0.98 1.00 1.00 1.00 | Tokio | Tokyo | طوكيو | 东京 | 東京 |
| 7 | 70 | 0.02 0.02 0.02 0.02 0.93 | 0.99 1.00 0.98 0.99 0.63 | . | . | . | 。 | です |
| 8 | 57 | 0.01 0.01 0.01 0.01 0.55 | 0.99 1.00 0.98 0.99 0.71 | . | . | . | 。 | を |
| 9 | 54 | 0.71 0.71 1.00 0.68 0.69 | 1.00 1.00 1.00 0.97 1.00 | pasaporte | passport | جواز | 护照 | パスポート |
| 10 | 48 | 0.01 0.01 0.01 0.01 0.79 | 0.99 1.00 0.98 0.99 0.66 | . | . | . | 。 | の |
| 11 | 44 | 0.01 0.01 0.01 0.01 0.90 | 0.99 1.00 0.98 0.99 0.84 | . | . | . | 。 | が |
| 12 | 43 | 0.98 0.98 0.98 0.98 1.00 | 0.88 1.00 0.98 0.89 0.97 | aeropuerto | airport | المطار | 机场 | 空港 |
| 13 | 34 | 1.00 1.00 1.00 1.00 1.00 | 1.00 0.98 1.00 1.00 1.00 | Chicago | Chicago | شيكاغو | 芝加哥 | シカゴ |
| 14 | 33 | 0.75 0.97 0.59 0.59 0.60 | 1.00 0.84 0.99 0.94 1.00 | Nueva York | New York | نيويورك | 纽约 | ニューヨーク |
| 15 | 33 | 0.01 0.01 0.01 0.01 0.59 | 0.99 1.00 0.98 0.99 0.69 | . | . | . | 。 | に |
| 16 | 32 | 0.01 0.01 0.01 0.46 0.01 | 1.00 1.00 1.00 1.00 0.64 | . | . | . | 了。 | 。 |
| 17 | 29 | 0.94 0.94 0.94 1.00 0.94 | 0.96 1.00 1.00 0.98 1.00 | Boston | Boston | بوسطن | 波士顿 | ボストン |
| 18 | 27 | 0.96 0.96 1.00 0.96 0.96 | 1.00 0.97 1.00 1.00 1.00 | Londres | London | لندن | 伦敦 | ロンドン |
| 19 | 26 | 0.90 0.96 0.90 0.93 0.90 | 1.00 1.00 1.00 1.00 0.97 | Tanaka | Tanaka | تاناكا | T a n a k a | タナカ |
| 20 | 22 | 0.10 0.09 0.08 0.54 0.09 | 0.99 0.95 0.94 0.26 0.82 | Dónde | Where | أين | 在 | どこ |
| 21 | 21 | 1.00 1.00 1.00 1.00 1.00 | 1.00 1.00 1.00 1.00 1.00 | Yamada | Yamada | يامادا | Y a m a d a | ヤマダ |
| 22 | 21 | 0.00 0.00 0.00 0.51 0.01 | 1.00 1.00 1.00 1.00 0.64 | . | . | . | 的。 | 。 |
| 23 | 20 | 1.00 0.91 0.87 0.95 0.87 | 0.96 0.99 0.73 0.72 0.86 | hoy | today | اليوم | 今天 | 今日 |
| 24 | 19 | 0.90 1.00 0.86 0.86 0.86 | 1.00 1.00 1.00 1.00 1.00 | Miami | Miami | ميامي | 迈阿密 | マイアミ |

⋮

(× 30 000 alignements obtenus en 10 secondes)

Le nombre de sous-corpus possibles est bien entendu trop important pour que tous soient traités, car ce nombre est exponentiel en la taille du corpus de départ. Mais tous les sous-corpus n'auront bien évidemment pas le même intérêt : par exemple, un sous-corpus constitué du corpus de départ moins un énoncé n'apportera sans doute pas grand chose si le corpus de départ est grand. Un sous-corpus constitué d'un seul énoncé ne nous apprendra rien de nouveau non plus. Il s'agira donc de trouver un juste milieu, en ne sélectionnant par exemple que les énoncés nécessaires et suffisants à un alignement donné. Nous verrons au chapitre suivant que les sous-corpus de relativement petite taille se révéleront être les plus productifs.

### 3.1.2 *… et les avantages qui en découlent*

Outre leur grande disponibilité, les mots de faible effectif présentent a priori de nombreux avantages en alignement. Pourtant, personne, à notre connaissance, ne semble avoir jamais tenté de remettre en cause la croyance selon laquelle ces mots ne seraient que sources de problèmes. Voici quelques-uns de ces avantages.

FAIRE PLUS SIMPLE Intuitivement, les traitements nécessaires à l'alignement de mots rares devraient se révéler plus simples que pour des mots fréquents. Déterminer si deux mots fréquents apparaissant *à peu près* dans les mêmes énoncés doivent être alignés ou non est typiquement un problème de décision, dont les meilleurs résultats sont obtenus en ayant recours à un certain bagage mathématique. Le problème ne se pose tout simplement pas avec des hapax, avec lesquels le principe du tout ou rien s'applique : les deux mots sont présents simultanément ou ils ne le sont pas. C'est là tout le bagage mathématique nécessaire et suffisant à leur alignement.

SOULAGER LA MÉMOIRE La quantité de données à traiter simultanément diminue lorsqu'on en supprime en entrée. Nous pouvons donc traiter des corpus de taille importante sans nous soucier outre

mesure des ressources en mémoire requises par la machine pour leur traitement, car seules de petites parties en seront traitées à la fois. Toute machine, dans les limites du raisonnable, fera donc l'affaire. Un supercalculateur sera superflu.

PARALLÉLISER FACILEMENT L'alignement reposant sur le traitement de sous-corpus, nous pouvons sans difficulté traiter séparément plusieurs parties d'un corpus d'entrée sur différents processeurs ou différentes machines. Il suffit pour cela de s'assurer de l'indépendance des calculs ainsi répartis et de la possibilité d'intégration des résultats des différents processus.

FAIRE PLUS RAPIDE Il s'agit de la conséquence logique des trois points précédents. Moins de données en entrée, puisqu'on en supprime, implique moins de traitements à effectuer. Des traitements plus simples sont — en général, et a priori ! — plus rapides à exécuter. Et la parallélisation des traitements peut bien entendu grandement diminuer les temps de calcul.

DÉSAMBIGUÏSER GRATUITEMENT Autre avantage propre aux hapax : ils ne peuvent avoir qu'un seul et unique sens dans le texte où ils apparaissent. Une certaine désambiguïsation est ainsi opérée implicitement dès lors que nous supprimons des données d'entrée, car des mots deviennent hapax à travers ce processus.

PERMETTRE LE MULTILINGUISME, LE VRAI Comme nous le verrons par la suite, l'exploitation des termes de faible effectif permettra l'alignement d'un nombre quelconque de langues simultanément. Ce point sera abordé en détail au chapitre 6.

### 3.1.3 *Tout est alignable*

Le dernier avantage relatif à l'utilisation des mots rares que nous mentionnons, et qui n'est pas des moindres, est leur bonne prédispo-

# Troisième partie

# ANNEXES

Nous avons donc indiqué certaines pistes susceptibles de mener à des améliorations de notre méthode :

- recombiner les alignements produits pour obtenir des alignements constitués de mots de fréquences différentes, ou effectuer un traitement équivalent *avant* la phase d'alignement. Des expériences préliminaires d'alignement de n-grammes de mots ont abouti directement à une amélioration décisive de nos résultats en traduction automatique fondée sur les segments. Il s'agit de la piste à laquelle nous donnons le plus d'importance.
- combiner nos alignements avec ceux obtenus par une autre méthode. Nous avons mentionné l'existence d'une différence de contenu entre nos alignements et ceux obtenus par exemple par MGIZA++. Cette différence est aisément concevable dans la mesure où ces approches sont radicalement différentes. Établir leur complémentarité permettrait une amélioration de la couverture des alignements.
- dépasser le traitement par mots typographiques. Nous avons vu que notre méthode était sensible au degré de synthétisme des langues, sensibilité qui est naturellement atténuée dans les sous-corpus de petite taille sur lesquels opère notre méthode. Un traitement en caractères serait à notre sens idéal, car cela permettrait de neutraliser implicitement tout problème de segmentation ou de divergence entre langues.

Cette thèse a dressé une recette complète pour effectuer de l'alignement sous-phrastique multilingue. Beaucoup de principes énoncés ne se limitent pas à l'alignement, et peuvent aisément être transposés à d'autres domaines, les deux grands axes étant le lien entre monolinguisme et multilinguisme, d'une part, et la réduction de la quantité de données à traiter d'autre part. Nous sommes partisan de l'approche consistant à sortir du cadre classique pour résoudre un problème, en observant les faits, comment ils sont généralement exploités, et surtout comment ils ne sont *pas* exploités. Nous nous efforcerons de continuer de cultiver la différence, dans tous les sens du terme.

sition pour s'aligner. En supprimant des données du corpus d'entrée, tous les mots peuvent a priori devenir des hapax. En admettant alors que tout hapax s'aligne, nous pouvons aligner n'importe quel mot : il suffit pour cela de se placer dans un sous-corpus dans lequel un mot source donné est hapax, et sa traduction en langue cible aura de grandes chances d'être également hapax dans ce sous-corpus. Nous pouvons même faire en sorte que seuls ces mots soient hapax dans l'énoncé, de façon à ce qu'aucune alternative ne soit possible lors de l'alignement. Ce faisant, nous nous placerions artificiellement dans la situation où nous alignions des hapax issus d'énoncés ne contenant qu'un seul hapax, comme c'était le cas dans le tableau 5 page 43.

Nous montrons qu'en pratique il est possible d'aligner non seulement tous les couples de mots (*source*, *cible*) nécessaires, mais aussi davantage. Il est en effet aisé de déterminer s'il existe un sous-corpus tel qu'un mot source et un mot cible sont les seuls hapax dans cet énoncé — autrement dit, s'il existe un sous-corpus ne contenant pas les deux mots à rendre hapax mais contenant chacun des autres mots de l'énoncé. Pour cela, il suffit de déterminer, pour chaque mot d'un énoncé, l'ensemble des énoncés où il apparaît, et d'en retirer l'ensemble des énoncés où le mot à rendre hapax apparaît. Si aucun des ensembles obtenus n'est vide, alors ce sous-corpus existe dans le corpus utilisé : il s'agit de l'union de ces ensembles. S'il existe un tel sous-corpus dans les deux langues et si leur intersection n'est pas vide, alors, en se plaçant dans cette intersection, les deux mots seront les seuls hapax dans leur énoncé, donc alignables.

Nous effectuons cette expérience sur notre corpus Europarl. Idéalement, le nombre de couples de mots alignables parmi l'ensemble des couples de mots (*source*, *cible*) d'un énoncé devrait être au moins égal à la longueur de cet énoncé en nombre de mots, que nous notons $n$, ce qui signifie dans le meilleur des cas que chacun des mots a été aligné. La quantité moyenne de couples de mots pour lesquels il existe un sous-corpus tel que ces deux mots sont les seuls hapax dans leur énoncé est tracée en fonction de la taille du corpus sur la figure 10 page suivante. Elle est très rapidement supérieure à $n$, dès

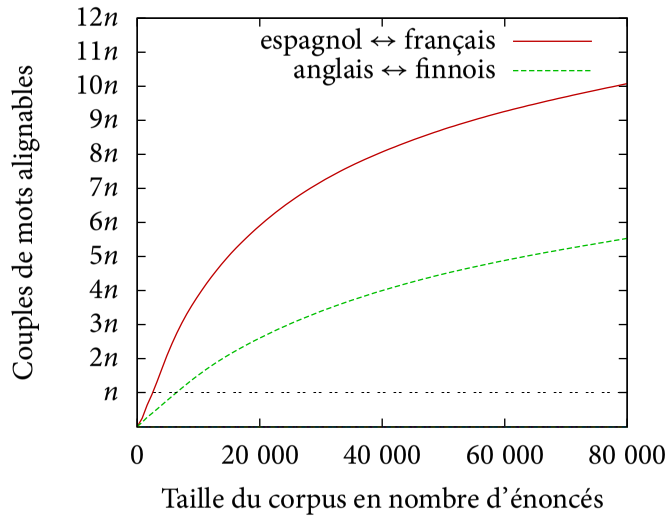



FIGURE 10 – Nombre moyen de couples de mots issus d'un énoncé et de sa traduction pour lesquels il existe un sous-corpus tel que les deux mots sont seuls hapax dans cet énoncé. Plus le corpus est grand, plus ce nombre est élevé. Il dépasse très rapidement le nombre de mots $n$ dans un énoncé.

un nombre d'énoncés relativement modeste : environ 5 000, selon le couple de langues. Jusqu'à six fois plus d'alignements pourraient être obtenus aux limites de ce graphique, et le nombre de couples semble croître logarithmiquement ; théoriquement jusqu'à $n^2$ pour des tailles de corpus encore supérieures. Au total, tous les mots peuvent donc être alignés, a priori, en devenant hapax. Par conséquent, un alignement fondé sur les basses fréquences est tout à fait concevable.

# CONCLUSION

REJETER les mots de basse fréquence n'est pas une bonne idée en alignement. Nous l'avons clairement montré tout au long de cette thèse : ils sont omniprésents, facilement alignables, et peuvent mener à de très bons résultats. Nous les avons donc fortement mis à contribution dans la conception d'une méthode d'alignement multilingue, par laquelle les mots fréquents sont rendus rares dans des sous-corpus constitués par échantillonnage. Cette approche permet naturellement une mise en œuvre à grande échelle. En fait, si nous n'avons que très peu abordé cet aspect, c'est pour une raison très simple : la méthode *est* le passage à l'échelle. À *moindre* échelle : plutôt que d'enrichir une méthode pour traiter de plus grandes quantités de données, nous avons pris le problème à l'envers en accordant les données avec la méthode. Supprimer des données nous a permis de faire plus simple et vraiment plus rapide. Le multilinguisme a été rendu possible par le monolinguisme. Traiter plusieurs langues ensemble est aisé dès lors que tous les traitements en jeu sont effectifs sur une langue unique, car cela permet leur mise en œuvre sur toutes les langues désirées en parallèle — ou sur toutes les langues confondues, ce à quoi nous sommes parvenu dans la conception d'Anymalign. Nous avons ainsi été en mesure d'aligner un nombre quelconque de langues simultanément, et cela s'avère plus efficace qu'un classique traitement par couples.

Nous avons obtenu, avec une méthode très simple qui se positionne clairement hors du courant dominant, des résultats d'une qualité rivalisant avec ceux de l'état de l'art. En fait, la qualité propre des lexiques produits par notre méthode est supérieure en moyenne. Utilisés comme tables de traductions, ils mènent à des résultats un peu moins bons en traduction automatique probabiliste par segments, mais nous savons *pourquoi* et *comment* résoudre ce problème : ils ne constituent pas en l'état, comme nous l'imaginions, de vraies tables de traductions, car ils ne contiennent pas d'alignements de mots de fréquences différentes.

## 3.2 QUOI ALIGNER ?

### 3.2.1 *D'un point de vue pratique : pré-segmentation*

Dans le chapitre précédent, nous avons vu que les « meilleurs » alignements qu'étaient les alignements d'hapax issus d'énoncés ne comportant qu'un seul hapax étaient de très bonne qualité entre espagnol et français, mais médiocres entre anglais et finnois (voir tableau 5 page 43). Ces deux dernières langues ayant des mots (orthographiques) de structure et de longueur très différentes, un mot de l'une peut ne pas avoir de *mot* traduction dans l'autre. Le mot finnois *valkoturskalajien* 'espèces de poisson blanc' aurait par exemple dû être aligné avec quatre mots anglais : *species of white fish*, ce qui était impossible car la méthode employée se bornait à produire des alignements de multiplicité 1-1. Inversement, aucun des quatre mots anglais n'aurait trouvé de traduction finnoise à moins d'une segmentation préalable de *valkoturskalajien*. Une « bonne » pré-segmentation pour l'alignement bilingue en serait une par laquelle toute unité textuelle trouverait sa traduction dans l'autre langue. Une façon de faire serait d'optimiser le grain de traitement des deux langues de façon à ce qu'ils aient grosso modo même taille. Dans le cas du couple anglais-finnois, nous pourrions segmenter les mots finnois, langue agglutinante, de sorte que les unités finnoises à aligner correspondent davantage à la notion de mot en anglais. Le contraire est également possible : grossir le grain de l'anglais, langue la plus isolante, en considérant non plus le mot comme unité textuelle mais plusieurs mots (voir p. ex. le *word packing* de Ma et coll., 2007b). Ceux-ci peuvent former des groupes linguistiquement motivés (cohérents en termes de dépendances : syntagmes, propositions, voire phrases) ou non (simples n-grammes de mots). De façon plus générale, l'unité de traitement pourrait très bien être la *sous-séquence* de mots d'un énoncé, potentiellement discontinue.

Cependant, grouper les mots de façon pertinente linguistiquement nécessite bien souvent des ressources propres aux langues traitées. Cela sort du cadre applicatif que nous nous sommes fixé, à savoir traiter les

corpus de façon endogène, et par des approches les plus génériques possibles. Des méthodes de segmentation sans ressources existent, citons entre autres les travaux de Déjean (1998), qui propose un algorithme de découverte de morphèmes inspiré des travaux de Harris (1955), ainsi qu'un algorithme de segmentation en syntagmes ; et plus récemment de Dasgupta et Ng (2007) sur la segmentation en morphèmes et de Vergne (2009) sur le chunking. Ces algorithmes se limitent cependant typiquement aux écritures alphabétiques. En outre, comme le rappellent Ma et coll. (2007a), chaque *couple* de langues traité implique une nouvelle segmentation : on ne segmentera pas un énoncé allemand de la même façon selon qu'on a l'intention de l'aligner avec du finnois ou de l'anglais. Cela se révélerait particulièrement problématique si nous voulions aligner plus de deux langues simultanément, comme nous réussirons à le faire par la suite.

Pour ces raisons purement méthodologiques, nous ferons donc le choix de n'avoir recours à aucune segmentation autre que celle fournie par la typographie d'une langue : caractères et mots, visuellement identifiables. La segmentation en mots présente de surcroît l'avantage d'être pertinente linguistiquement. Seuls de très légers pré-traitements seront éventuellement appliqués aux langues le permettant, tels l'insertion de caractères d'espacement entre les caractères de ponctuation et les mots à proprement parler *(tokenization)*, ce à quoi nous avons déjà eu recours dans les expériences précédentes. Cette pratique est courante en traduction automatique, car elle permet d'améliorer très simplement les résultats. Les langues dont l'écriture ne sépare pas les mots par des caractères d'espacement, telles le chinois où le japonais, seront pré-segmentées en mots à l'aide un analyseur externe, sans quoi nous ne pourrions pas les traiter du tout.

### 3.2.2 *D'un point de vue théorique : divergence*

Un traitement en mots, bien que naturel, soulève néanmoins un problème bien connu en traduction automatique, qui est celui de la divergence entre langues : les mêmes idées ne s'expriment pas avec



De façon directe, nous pourrions effectuer des mesures directement sur un fichier de sortie d'Anymalign. Cela peut être aussi élémentaire qu'un simple décompte des alignements en fonction des langues en entrée. Les propriétés des alignements vues à la figure 29 page 128 peuvent pour cela être mises à profit : avec filtrage sur le nombre de langues couvertes, les langues proches produisent naturellement plus d'alignements différents que des langues éloignées ; et sans filtrage, la tendance s'inverse du fait du grand nombre de traductions possibles d'un même mot. Par exemple, puisque les trois langues du triplet espagnol-français-italien appartiennent à la même famille, romane, alors le nombre d'alignements obtenus couvrant ces trois langues sera supérieur à celui du triplet espagnol-français-polonais, cette dernière appartenant aux langues slaves. De la même façon, le nombre de fois qu'un même alignement a été obtenu doit être plus élevé avec un groupe constitué de langues proches qu'avec un autre constitué de langues choisies aléatoirement. Nous pourrions ainsi introduire la possibilité de travailler réellement sur les familles plutôt que sur les — toujours — traditionnels couples de langues.

#### RÉSUMÉ

L'alignement sous-phrastique véritablement multilingue est une réalité :

- toutes les opérations mises en jeu sont monolingues ;
- notre approche permet le traitement d'un nombre quelconque de langues simultanément, y compris d'une seule, et cela s'avère en définitive plus efficace qu'un traditionnel traitement par couples de langues ;
- du fait de notre position de « pionnier » en alignement multilingue, les applications tirant parti d'alignements multilingues sont peu nombreuses à l'heure actuelle. Il ne reste plus qu'à faire évoluer les choses.

un processus d'alignement-affinement itératif où davantage de langues mènent à de meilleurs alignements. Cela se rapproche des tâches de désambiguïsation s'appuyant sur des traductions dans diverses langues pour décider du sens d'un terme (p. ex. Brown et coll., 1991a; Ng et coll., 2003). Ce problème se pose naturellement dans l'autre sens en traduction automatique, un terme pouvant se traduire de différentes façons selon le sens qu'il revêt — pour reprendre un exemple archi-connu : *avocat* en français peut donner en anglais *avocado* ou *lawyer* selon le contexte.

Cette approche a été intégrée dans les systèmes de traduction probabiliste, sous le nom de traduction « multi-source » (Och et Ney, 2001; Crego et coll., 2009; Koehn et coll., 2009). Bien que ne correspondant pas à une situation de traduction réelle, elle a montré pouvoir améliorer la qualité des traductions au moins pour les langues couvertes par des corpus parallèles multilingues. Dans cette approche, des choix lexicaux sont faits en se basant sur les *sorties* de systèmes de traduction auxiliaires. Avec des alignements multilingues, cette désambiguïsation pourrait être effectuée en amont.

### 6.3.3 *Classification de langues*

Enfin, une autre application susceptible de profiter d'alignements multilingues concerne la typologie des langues. L'étude des alignements devrait permettre la mise en évidence de certains phénomènes linguistiques.

De façon indirecte, on peut simplement appliquer une méthode externe sur un de nos lexiques multilingues. Par exemple, dans (Lepage et coll., 2009), un lexique multilingue produit par Anymalign a été utilisé comme support pour évaluer les distances (ou plutôt pseudo-distances) entre langues : à partir de correspondances multilingues, le nombre d'analogies entre mots communes à deux langues a servi de fondement pour attester de la proximité de ces langues. Une étude sur neuf langues du JRC-Acquis (Steinberger et coll., 2006) a ainsi conforté la connaissance de la proximité relative entre ces langues.



les même structures d'une langue à l'autre. Le professeur Vauquois l'a illustré par l'exemple anglais-français suivant, devenu un classique :

Elle lui plaît. ↔ He likes her.

Établir des correspondances entre mots dans un tel exemple n'est pas pertinent, car l'information n'est pas distribuée de la même façon entre les mots des énoncés anglais et français. D'après Dorr et coll. (2002), ces divergences sont monnaie courante : les auteurs montrent dans une expérience que 35 % des phrases d'un corpus espagnol-anglais en comportent. Naturellement, cette proportion est d'autant plus importante que les langues sont éloignées. Les liens créés par l'alignement entre les mots d'énoncés comportant de telles divergences ne peuvent être qu'approximatifs, et il peut alors être préférable de n'en rien faire, ou plutôt d'établir ces liens à un grain plus fin ou plus gros, qui, lui, sera pertinent.

Nous montrons que théoriquement, la faisabilité d'un alignement dépend du grain de traitement utilisé. Admettons qu'un énoncé se compose d'une suite d'unités alignables minimales, que nous appellerons pour l'occasion *alignèmes*. Pour simplifier, considérons le cas utopique où les énoncés source et cible se composent du même nombre d'alignèmes et où chacun a directement un unique équivalent dans l'autre langue. De telles unités sont typiquement organisables en une hiérarchie de grains. Selon le niveau auquel nous nous plaçons dans cette hiérarchie, le grain sera plus ou moins gros, le plus fin étant l'alignème et le plus gros l'énoncé dans sa totalité, en passant éventuellement par le mot, le syntagme, la proposition, la phrase. Lorsque les deux langues sont très éloignées, un alignème et son équivalent ne seront pas regroupés dans les mêmes unités de grain supérieur dans leurs langues respectives ; autrement dit, l'existence d'équivalences à un grain $g$ n'implique pas nécessairement l'existence d'équivalences à un grain $g + 1$. Le contraire est également vrai : au grain le plus gros, à savoir l'énoncé dans sa totalité, l'existence d'une équivalence est garantie, puisqu'il ne s'agit que de l'énoncé complet dans l'autre langue ; mais cette garantie n'a pas lieu d'être à un grain inférieur.

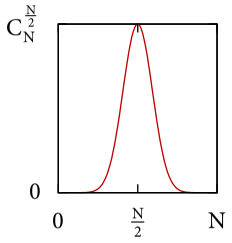
Pour s'en convaincre, considérons l'exemple suivant, où les deux lignes correspondent à un même énoncé dans deux langues différentes. Les alignèmes sont assimilés aux caractères et sont directement équivalents : « *a* » en italique est traduction de « a » en romain, et ainsi de suite. Les alignèmes sont regroupés en mots, séparés par des espaces. Notre hiérarchie de grains est donc caractère < mot < énoncé. Le mot fait office de grain de traitement.

ab cd ef g
*a bc de fg*

Dans cette configuration, aucune combinaison de mots autre que l'énoncé dans sa totalité ne peut être alignée. Théoriquement donc, un mauvais grain peut mener à l'impossibilité d'aligner quoi que ce soit, ou pire, à la certitude que tous les alignements produits seront mauvais.

### 3.2.3 *De la découpe d'un énoncé*

La probabilité que l'équivalent d'une unité existe dans l'autre langue dépend en fait du nombre d'alignèmes dont ces unités sont constituées. Pour un énoncé composé de N alignèmes, cette probabilité sera inversement proportionnelle au nombre de sous-séquences de $n$ alignèmes ($1 \leqslant n \leqslant N$), qui n'est rien d'autre que le nombre de combinaisons de $n$ objets parmi N, et ce dans les deux langues. La cloche ci-contre en rappelle l'allure. Lorsque $n$ s'approche de la moitié de la longueur de l'énoncé, le nombre de façons de combiner les alignèmes augmente de façon importante, et avec elle diminue la probabilité d'existence d'une unité textuelle équivalente dans l'autre langue. Concrètement, cela revient à dire qu'en partitionnant un énoncé ainsi que sa traduction en deux parties égales, il est hautement improbable que les parties ainsi créées soient alignables entre elles, car les équivalents des alignèmes d'une partie ont davantage de chances d'être dispersés entre les deux parties de l'autre langue. Les segmentations assurant






Notons qu'Anymalign dispose d'un avantage supplémentaire relatif à la constitution de ressources : à partir d'un unique fichier de sortie multilingue, il permet de créer des « vues » sur des sous-ensembles de langues. Pratiquement, il suffit de supprimer une ou plusieurs langues d'un fichier de sortie de l'aligneur au format texte. Chacune des langues se présentant sous la forme de « colonnes », cette opération est typiquement l'affaire d'un simple appel à la commande Unix `cut`. Puis, tant que les décomptes associés aux alignements sont conservés, il est possible de recalculer l'intégralité des probabilités de traduction. Il est ainsi possible d'offrir des vues (statiques) adaptées aux désirs d'un utilisateur à partir d'un unique fichier faisant office de base de données.

### 6.3.2 *Traduction automatique*

Traditionnellement, la traduction automatique a toujours été un processus bilingue, le plus souvent orienté. Outre le fait que traiter plusieurs langues simultanément s'avère plus efficace que traiter tous les couples séparément (section 6.2.3), des alignements multilingues pourraient permettre la traduction d'une langue vers plusieurs autres simultanément. C'est d'ailleurs en ce sens que nous avons redéfini le calcul de nos scores à la section 6.2.1. Nous avons déjà évoqué le fait que les concepts monolingues que nous utilisons pour l'alignement étaient directement issus de la traduction automatique par l'exemple, ce côté monolingue n'ayant simplement jamais été mis en avant dans la littérature. De la même façon que nous avons pu généraliser l'alignement au cas du multilinguisme, nous pouvons imaginer un système de traduction — par l'exemple ? — multilingue.

Une autre application potentielle est le *renforcement* de connaissances bilingues. Il s'agissait en fait de l'intention première de Simard (1999) lorsqu'il proposa de réaliser des alignements multilingues : ces alignements ne constituaient pas une finalité en soi, mais servaient de source d'information pour conforter des choix d'alignement bilingues. Lecluze (2007) a étendu plus avant cette philosophie à travers

péenne (Koehn, 2005; Steinberger et coll., 2006), les textes religieux tels que la Bible (Resnik et coll., 1999) et certains corpus constitués spécialement par des humains tels le BTEC (Takezawa et coll., 2002) ou les messages système d'environnements informatiques tels KDE (Tiedemann, 2009). Nous pensons néanmoins que le « peu » disponible mérite d'être exploité dès maintenant, ne serait-ce que parce que beaucoup plus de matériau pourrait très bien voir le jour à l'avenir.

### 6.3.1 *Constitution de ressources multilingues*

Une application pour laquelle Anymalign est d'ores et déjà opérationnel est la constitution de lexiques multilingues. On peut trouver de tels lexiques sur la Toile ; citons à titre d'exemple le projet MAGUS [3], répertoriant une centaine de noms d'animaux dans plus de cinquante langues classés par familles, le dictionnaire Babel [4], contenant des entrées réparties en thèmes dans une cinquantaine de langues et permettant de visualiser des traductions dans quatre langues simultanément, ou encore le « dictionnaire universel » du site *Dicts.info* [5] permettant de visualiser des résultats en trois langues.

Constitués manuellement, la plupart de ces lexiques sont typiquement confinés à un domaine bien particulier et contiennent relativement peu de termes. Anymalign permet non seulement de produire de véritables dictionnaires multilingues sans effort à partir de corpus parallèles, mais leur couverture des langues est de surcroît relativement bonne, car directement liée aux corpus utilisés, lesquels couvrent de nombreux domaines du fait de leur nature (Resnik et coll., 1999; Koehn, 2005). Nous avons déjà mis plusieurs lexiques multilingues en ligne [6]. Des échantillons sont visibles à l'annexe A page 139.

3 http://www.informatika.bf.uni-lj.si/magus.html
4 http://projetbabel.org/forum/babel/
5 http://www.dicts.info/ud.php
6 http://users.info.unicaen.fr/~alardill/anymalign/lexicons/
(Pages consultées le 1er juin 2010)



l'alignement seraient plutôt celles des extrémités, là où la courbe est au plus bas : près du grain le plus fin ou du grain le plus gros. L'alignement du plus gros grain, à savoir les énoncés complets, n'apporte rien car il ne s'agit que de notre connaissance de départ. Resterait donc l'alignement du grain le plus fin.

Cette vue est théorique à l'extrême, et suppose que les langues n'aient presque rien en commun, voire qu'elles ne soient que des codes aléatoires. C'est heureusement loin d'être le cas en pratique. Elle caricature cependant ce à quoi peut mener une pré-segmentation pour l'alignement. On peut admettre sans problème que pratiquement toute proposition française est alignable avec une proposition anglaise. Cet alignement est beaucoup moins évident entre des propositions françaises et japonaises, comme le montrent les travaux de Nakamura-Delloye (2007). Kashioka et coll. (2003) montrent aussi dans une expérience que près de 10 % des propositions japonaises d'un corpus japonais-anglais n'ont pas d'équivalent en anglais. Pour traiter ces problèmes, Dorr et coll. (2002) proposent de « réparer » les divergences en rapprochant structurellement un énoncé de sa traduction avant l'alignement à proprement parler, mais cette approche nécessite des ressources propres à chaque langue. Lepage et Denoual (2005) démontrent d'autre part que le recours à l'analogie proportionnelle appliquée au grain caractère permet de neutraliser *implicitement* les divergences entre langues dans le cadre de la traduction automatique, en s'affranchissant de l'étape d'alignement ; mais notre tâche consiste précisément à faire de l'alignement, et par conséquent à *expliciter* les relations de traduction.

Nous retiendrons donc principalement que la segmentation des énoncés la moins risquée théoriquement est celle qui les coupe en deux parties les plus inégales possibles, soit une unité minimale d'une part et toutes les autres d'autre part. Ayant conclu à la section 3.2.1 que notre unité minimale serait le mot (typographique), la découpe d'un énoncé que nous viserons donc autant que faire se peut sera un mot d'une part et ses contextes d'autre part.

## 3.3 LEVÉE DES DERNIERS VERROUS

Le début de ce chapitre a montré qu'un alignement fondé sur les termes de basse fréquence n'avait rien d'une chimère. La façon dont l'alignement a été abordé jusqu'ici induisait cependant un certain nombre de limitations, que nous levons dans cette section.

### 3.3.1 *Comment : à la recherche de nouveaux alignements*

La première limitation concerne la constitution des sous-corpus. Telle que nous l'avons opérée précédemment, elle était guidée par le besoin d'aligner un couple de mots (*source*, *cible*) donné, ce qui risque au total de se révéler assez lent. Plutôt que de partir d'aspirants alignements pour aller vers des sous-corpus, partir de sous-corpus déjà constitués pour en extraire des alignements pourrait se révéler autrement plus efficace. En outre, la nécessité de savoir à l'avance quelles unités sources doivent être alignées constitue en elle-même une restriction. Avec la méthode du cosinus et celle consistant à déterminer des sous-corpus où les mots à aligner étaient les seuls hapax dans un énoncé, nous nous étions contenté d'aligner tous les couples de mots (*source*, *cible*) possibles. Mais, en procédant ainsi, aucune *découverte* de chaînes ou séquences pertinentes vis-à-vis de l'alignement n'est malheureusement possible.

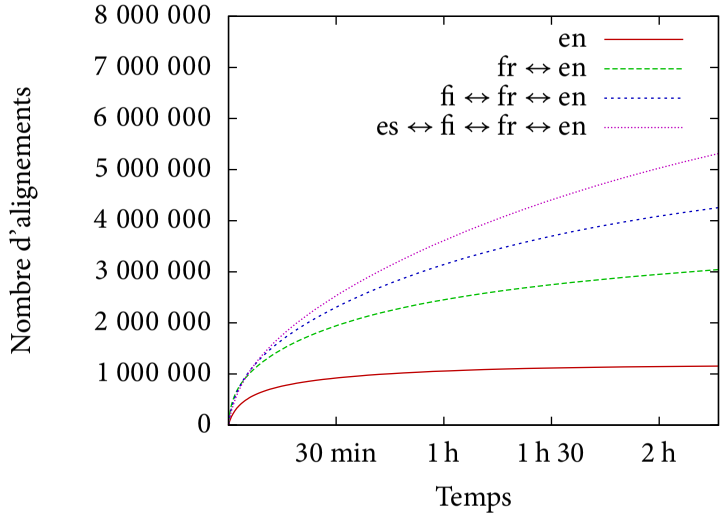
L'approche qui consiste à traiter tous les sous-corpus d'un corpus de départ n'est pas raisonnable car le nombre de sous-corpus est exponentiel en la taille de ce corpus. L'approche qui nous paraît non seulement la plus simple mais surtout la plus précise, parce qu'elle n'altère en aucune manière la distribution naturelle des mots du corpus de départ, est de constituer des sous-corpus en procédant par échantillonnage : puisque le corpus de départ est censé constituer un échantillon d'une langue (Sinclair et Ball, 1996), un échantillon de celui-ci devrait également constituer un échantillon de cette langue. Et en multipliant ces sous-corpus, nous nous rapprocherons de la représentativité de l'échantillon initial.



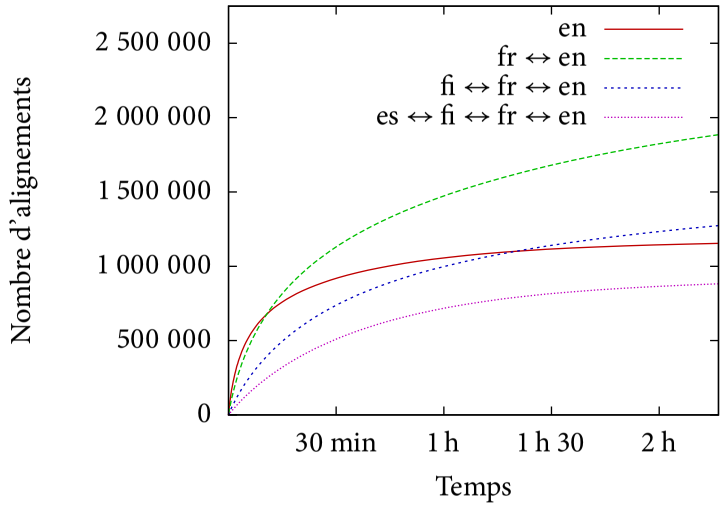
que par une seule langue, sont bien considérés comme deux alignements distincts. Dans le pire des cas, la polysémie des mots peut entraîner un nombre d'alignements exponentiel en le nombre de langues. Sans en arriver là, le graphique (a) de la figure 29 page précédente confirme qu'en l'absence de toute stratégie de filtrage, le nombre d'alignements en sortie augmente bien avec le nombre de langues. Les courbes semblent montrer que cette augmentation est de plus en plus faible quand le nombre de langues croît, mais cela reflète en fait l'augmentation du temps nécessaire à leur traitement : davantage de langues implique bien davantage d'alignements, mais conformément à la figure 28, l'extraction d'un alignement est d'autant plus lente que le nombre de langues — ou la longueur des énoncés — est élevé. À l'inverse, nous assistons à une diminution lorsque ne sont conservés que les alignements couvrant toutes les langues, comme le montre le graphique (b) de la figure 29, car nous ne conservons dans ce cas que l'« intersection » de toutes les langues. Les alignements restants étant typiquement parmi les meilleurs, et nécessitant naturellement moins de stockage, nous avons fait de ce comportement le comportement par défaut d'Anymalign. Au total, il est toujours plus rapide de traiter toutes les langues simultanément plutôt que tous les couples séparément.

## 6.3 APPLICATIONS EN PERSPECTIVE

À ce jour, les applications véritablement multilingues ne sont pas légion. En fait, à notre connaissance, c'est un premier pas que nous proposons avec Anymalign. Nous suggérons néanmoins quelques pistes qui pourraient d'ores et déjà profiter d'alignements réellement multilingues, qui sont naturellement différents — meilleurs ? Cela reste à démontrer — d'alignements multilingues qui auraient été obtenus à partir d'alignements bilingues (voir p. ex. la méthode de Simard, 1999). Nous sommes conscient de la relative rareté des corpus parallèles réellement multilingues à ce jour ; les seuls disponibles en plus de deux langues sont typiquement les textes de la commission euro-




(a) Sans filtrage sur le nombre de langues : tous les alignements sont comptabilisés. Plus le nombre de langues est grand, plus la quantité d'alignements produits est élevée.




(b) Avec filtrage sur le nombre de langues : seuls les alignements couvrant toutes les langues sont comptabilisés. Le cas monolingue mis à part, plus le nombre de langues est grand et plus la quantité d'alignements produits est faible.

FIGURE 29 – Influence du nombre de langues sur la quantité d'alignements produits en fonction du temps. La courbe en trait continu *(en)* est identique sur les deux graphiques.



Nos sous-corpus seront donc créés à partir d'énoncés tirés aléatoirement du corpus de départ, avec remise. Cela n'est pas sans rappeler les méthodes de *bootstrap* par rééchantillonnage, utilisées principalement, en traduction automatique, pour juger de la pertinence des résultats (Koehn, 2004; Zhang et coll., 2004). Nous pourrions mettre ces méthodes à profit pour améliorer la représentativité de nos sous-corpus, mais nous verrons dans les chapitres suivants qu'un appareil mathématique relativement élémentaire donnera d'ores et déjà de bons résultats ; nous conservons donc ces techniques statistiques pour des améliorations futures. Du fait de l'aspect aléatoire de l'échantillonnage, nous pouvons nous attendre à ce que deux expériences identiques effectuées à partir du même corpus d'entrée produisent des résultats différents en pratique. Ces différences sont minimes. La couverture du corpus d'entrée ne pourra pas non plus être garantie. Mais comme nous l'avons déjà évoqué, ce problème peut aisément être contourné en extrayant des alignements à partir de nombreux sous-corpus. Malgré le très grand nombre potentiel de sous-corpus nécessaires, nous verrons que notre approche est finalement une des plus rapides qui soient, car ayant le potentiel pour produire des alignements quasiment instantanément.

### 3.3.2 *Quoi : multiplicité des alignements*

La deuxième limitation concerne la multiplicité des alignements, qui étaient jusqu'à présent toujours de type 1-1. Naturellement, cela ne permet pas d'atteindre des résultats de très bonne qualité, car une telle multiplicité ne permet pas de rendre compte de toutes les relations de traduction, bien souvent plus complexes, entre les mots de deux langues. Elle n'est valide que si l'alignement est précédé d'une segmentation appropriée, ce que nous ne nous permettons pas, comme dit plus haut. Cela nous a d'ailleurs déjà posé problème dans nos expériences d'alignement passées :

– avec la méthode du cosinus, à partir de ce couple d'énoncés issu d'Europarl :

| | |
|---|---|
| […] el tan importante esfuerzo de cohesión económica y social se *desinflaría* poco a poco y […] la unión caería en […] las *acrimonias* nefastas para el desarrollo de la unión . | […] l' effort si important de cohésion économique et sociale s' *essoufflerait* peu à peu et […] l' union retomberait dans […] les *acrimonies* néfastes pour le développement de l' union . |

nous étions incapables de séparer l'alignement de l'hapax espagnol *desinflaría* avec l'hapax français *essoufflerait* et l'alignement de l'hapax espagnol *acrimonias* avec l'hapax français *acrimonies*. À la place, nous alignions chacun des deux hapax source avec chacun des deux hapax cible, soit quatre alignements dont deux erronés, et ce nombre aurait été supérieur si le nombre d'hapax l'avait été également.

– aligner correctement des hapax en corpus à partir d'un énoncé en comportant plusieurs, comme dans l'exemple précédent, aurait été impossible en tentant de déterminer un sous-corpus où ils auraient été les seuls hapax dans cet énoncé, car les hapax en corpus de cet énoncé n'auraient pas pu se désolidariser les uns des autres puisqu'ils auraient toujours eu le même effectif.

Pour éviter cette situation, nous adopterons dorénavant l'approche la plus naturelle qui soit, qui est celle qui consiste à ne rien faire. Notre grain de base sera toujours le mot, mais nos alignements seront systématiquement de multiplicité *m*-*n*. Là où nous alignions des *hapax*, ou des *séquences hapax* comme nous l'avons évoqué, nous alignerons désormais des *séquences* d'*hapax*, ordonnées. Par exemple, à partir du couple d'énoncés précédent, nous produirons l'alignement *desinflaría … acrimonias* ↔ *essouflerait … acrimonies*, sans chercher à savoir par quel mot cible se traduit quel mot source. Les mots d'une telle séquence pourront éventuellement être alignés séparément à partir d'un autre sous-corpus, pour peu qu'ils n'apparaissent pas systématiquement dans les mêmes énoncés, comme c'est le cas des séquences d'hapax. Dans le cas contraire, il est plus prudent de faire le choix de conserver ces séquences intactes.

Des heuristiques, telles que le recours à la position des mots dans un énoncé, pourraient permettre d'affiner l'alignement, mais cela dépen-



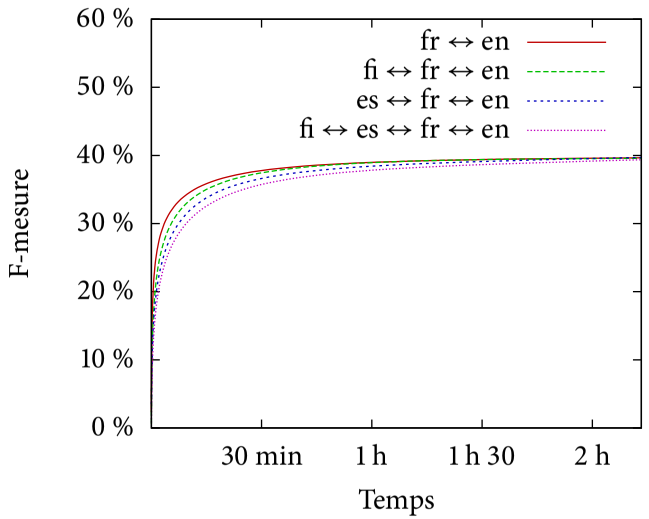



FIGURE 28 – Impact du nombre de langues sur la qualité des alignements. Les scores indiqués sont ceux du couple anglais → français. Le temps nécessaire pour atteindre un score donné semble évoluer linéairement avec le nombre de langues. Les courbes ont été lissées par souci de lisibilité.

constant pour atteindre la même qualité. Ce temps varie bien évidemment en fonction de la langue, et en particulier du nombre de mots dont sont constitués ses énoncés : dans notre expérience, l'ajout du finnois a un impact plus faible que l'ajout de l'espagnol car la longueur moyenne de ses énoncés est moindre. Au total, il est plus rapide de traiter toutes les langues d'un corpus simultanément plutôt que par couples, le temps nécessaire à la seconde approche étant naturellement quadratique en le nombre de langues, contre un temps linéaire pour la première.

Ensuite, comme nous pouvions nous y attendre, plus le nombre de langues est élevé et plus le nombre d'alignements en sortie l'est également. En effet, deux alignements multilingues, même s'ils ne diffèrent

rapidement, comme en attestent les courbes rendant compte du comportement d'Anymalign en fonction du temps du chapitre précédent, rapidité à laquelle contribue en grande partie la fonction d'optimisation de l'échantillonnage que nous avons définie. Finalement, les alignements que nous extrayons rendent compte d'une réalité de corpus, qui est celle de la cohésion des mots dans un contexte plus ou moins multilingue.

### 6.2.3 *Multilingue > bilingue*

Nous montrons dans cette section que d'un point de vue pratique, il est préférable de traiter plusieurs langues simultanément plutôt que par couples. Remarquons d'abord que raisonner en terme de nombre de langues n'est pas nécessairement pertinent. Plus exactement, il ne s'agit que d'une façon de modifier un autre facteur, qui est celui de la longueur des énoncés en nombre de mots. En effet, notre processus étant en fait monolingue, ajouter une langue au corpus de départ revient simplement à augmenter la longueur de ses énoncés. Cette équivalence s'avère vérifiée en pratique : le traitement de deux langues nécessite autant de temps que le traitement d'une seule langue dont les énoncés sont constitués des énoncés de ces deux langues mis bout à bout.

Nous évaluons d'abord l'impact du nombre de langues sur le temps nécessaire à leur traitement. Pour cela, nous recourons au second protocole d'évaluation que nous avons défini au chapitre précédent, qui consiste à comparer les alignements produits avec un lexique bilingue de référence. Nous déterminons donc la F-mesure associée à l'ensemble des alignements obtenus à partir d'un couple de langues donné en fonction du temps. La même évaluation est effectuée en ajoutant deux langues au corpus de départ, d'abord séparément puis simultanément. Ce à quoi nous nous intéressons alors est le temps nécessaire pour atteindre un même score. Les résultats sont présentés dans la figure 28 page ci-contre. Bien que cela soit difficile à déceler à vue d'œil sur la figure, les chiffres montrent que l'ajout d'une nouvelle langue implique à chaque fois un surplus de temps plus ou moins



drait à notre avis trop du couple de langues étudié, et nous préférons demeurer dans un cadre le plus générique possible. Nous verrons ainsi au chapitre 5 que notre approche a l'heureuse tendance à surpasser l'état de l'art sur des couples de langues éloignées, tandis que l'état de l'art, faisant usage de modèles de distortion et autres optimisations, donne de meilleurs résultats sur des langues plutôt proches, surtout lorsque l'anglais est en jeu.

### 3.3.3 *Oublions les basses fréquences*

Enfin, un des enseignements du chapitre 2 était que les meilleurs alignements étaient constitués d'une grande quantité de mots rares d'une part, et de mots fréquents en moindre quantité d'autre part, l'utilisation des mots de fréquence intermédiaire étant moins fructueuse (figures 8 et 9 page 40). Nous avons donc mis en évidence la faisabilité d'un alignement recourant exclusivement aux termes de basse fréquence, et en particulier aux hapax.

Il existe néanmoins un cas où la « transformation » des mots en hapax n'est pas possible, celui des termes de très haute fréquence, tels que le point de fin de phrase considéré comme un mot à part entière. Son alignement est difficile, car les seuls sous-corpus au sein desquels il peut devenir hapax ne peuvent être constitués que d'à peine un énoncé, puisque ce mot apparaît dans la quasi-totalité des énoncés. Mais dans de tels sous-corpus, tous les mots ont un effectif de un : le point n'est pas le seul hapax et par conséquent ne peut pas être aligné. Le recours aux termes de haute fréquence pour l'alignement serait alors inévitable. Pourrait-on concilier termes de haute et basse fréquence ?

Les mots très fréquents et très rares sont en fait bien liés, mais le critère qui les lie n'a rien à voir avec leurs effectifs : il s'agit moins d'un critère monolingue que d'un critère multilingue. Leur point commun est justement qu'ils s'alignent bien par des méthodes n'ayant recours qu'à leurs distributions dans un corpus. Il s'agit de termes qui, d'une langue source à une langue cible, s'expriment sous une forme véritablement équivalente — si l'on peut dire ; c'est ce qu'une interpré-

tation superficielle de l'égalité de leurs distributions pourrait laisser entendre —, au moins dans le corpus utilisé. Ces termes n'y sont pas ambigus. Pratiquement, ils apparaissent tout simplement strictement dans les mêmes énoncés, que cela soit en source ou en cible.

Les termes que nous chercherons à aligner par la suite ne seront donc ni des termes de haute, ni des termes de basse fréquence, mais des termes partageant exactement la même distribution dans un corpus ou sous-corpus, indépendemment de leurs effectifs. Par distribution, nous entendons la répartition d'un mot dans les énoncés du corpus, et non ses différents contextes observés. En pratique, ces mots seront constitués majoritairement de mots peu fréquents, typiquement des hapax, et dans une moindre mesure de mots très fréquents, typiquement des ponctuations, mais cela n'aura dorénavant aucune importance pour nous.

La figure 11 page ci-contre illustre comment nous procéderons pour extraire toutes les séquences de mots apparaissant strictement dans les mêmes énoncés à partir d'un (sous-)corpus jouet arabe-français. Il s'agit de nos premiers « vrais » alignements. Le nombre de fois qu'un mot apparaît dans un énoncé n'est pas pris en compte, car nous avons vu dans le chapitre 2 que ce n'était pas nécessaire : n'est prise en compte que la présence ou l'absence des mots dans les énoncés.

RÉSUMÉ

Nous avons à présent toutes les cartes en main. En définitive :

- l'alignement fondé sur des mots de basse fréquence est non seulement *faisable*, mais surtout *profitable*.
- pour des raisons aussi bien pratiques que théoriques, notre grain de traitement sera le mot typographique. Nos alignements seront constitués de séquences de mots extraits des énoncés de façon déséquilibrée.
- l'alignement des mots de basse fréquence ne constitue plus un *moyen*, mais une *conséquence* : ce que nous rechercherons, ce sont



| COLLOCATIONS | DÉC. |
|---|---|
| aujourd' hui | 584 |
| états membres | 478 |
| union européenne | 304 |
| chers collègues | 101 |
| nations unies | 96 |
| j' ai | 87 |
| parlement européen | 80 |
| de la | 55 |
| c' est | 40 |
| qu' il | 37 |
| s' agit | 37 |
| ad hoc | 32 |
| vifs applaudissements | 29 |
| ne pas | 28 |
| je voudrais | 26 |

(a) En filtrant les séquences discontinues.

| COLLOCATIONS | DÉC. |
|---|---|
| de _ . | 994 |
| ( _ ) | 957 |
| , _ . | 942 |
| monsieur _ président | 612 |
| de _ de _ . | 510 |
| , _ , _ . | 501 |
| aujourd' hui | 494 |
| la _ . | 448 |
| états membres | 440 |
| le _ . | 389 |
| mesdames _ messieurs | 337 |
| ne _ pas | 331 |
| « _ » | 291 |
| union européenne | 252 |
| et _ . | 246 |

(b) Sans filtrage sur les sorties.

TABLEAU 11 – Exemples de collocations obtenues en exécutant Anymalign sur un corpus monolingue. L'aligneur a été exécuté pendant dix secondes sur la partie française de notre corpus Europarl. Comme pour des alignements multilingues, les décomptes indiquent le nombre de fois que les collocations ont été obtenues, et le tiret bas (_) indique une discontinuité.

Cette approche du multilinguisme a été implantée dans Anymalign dès ses débuts, et c'est ainsi que toutes les expériences bilingues présentées dans les chapitres précédents ont en fait été réalisées. Le tableau 10 page précédente donne des exemples d'alignements multilingues réels avec leurs scores associés.

### 6.2.2 *Des alignements monolingues ou des collocations multilingues ?*

Rien n'empêche notre méthode d'être appliquée sur un corpus monolingue. Cela est possible car — encore — tous nos traitements sont monolingues. Lorsqu'une seule langue est en jeu, ce que nous recherchons n'est en fait rien d'autre qu'une certaine forme de collocations (voir p. ex. Charest et coll., 2007), car l'essence même de notre méthode est de rechercher des mots partageant la même distribution. Autrement dit, notre méthode d'alignement consiste à extraire des « collocations multilingues ».

Puisque notre méthode consiste en fin de compte à extraire des collocations, nous pourrions songer à remplacer notre critère de détection, c'est-à-dire la recherche de séquences de mots partageant strictement la même distribution dans des sous-corpus aléatoires, par des méthodes plus classiques de détection de collocations, à l'aide de scores tels que l'information mutuelle, et de les appliquer sur notre corpus alingue. Cela n'est en fait pas raisonnable, car nous perdrions alors l'avantage le plus décisif de notre approche par rapport à celles fondées sur l'attribution de scores, qui est qu'elle permet une détection très rapide et peu coûteuse des alignements — ou des collocations. Inutile de passer en revue tous les couples de mots possibles pour leur attribuer des scores : les groupes de mots pertinents apparaissent d'eux-mêmes. En fait, il serait plus judicieux de faire le contraire : utiliser notre méthode d'alignement pour extraire des collocations. Le tableau 11 page suivante donne des exemples de telles collocations.

Dans l'ensemble, du fait de l'intervention d'un processus aléatoire dans nos traitements, les résultats produits par notre méthode constituent clairement une approximation. Ils convergent néanmoins très



CORPUS D'ENTRÉE :

| | ARABE | | FRANÇAIS |
|---|---|---|---|
| 1 | قهوة ، من فضلك . | ↔ | Un café , s'il vous plaît . |
| 2 | هذه قهوة ممتازة . | ↔ | Ce café est excellent . |
| 3 | شاي ثقيل . | ↔ | Un thé fort . |

⇓

| CHACUN DES MOTS : | | | APPARAÎT DANS LES ÉNONCÉS : |
|---|---|---|---|
| | ↔ | Un | 1 3 |
| قهوة | ↔ | café | 1 2 |
| من فضلك ، | ↔ | , s'il vous plaît | 1 |
| . | ↔ | . | 1 2 3 |
| هذه _ ممتازة | ↔ | Ce _ est excellent | 2 |
| شاي ثقيل | ↔ | thé fort | 3 |

FIGURE 11 – Extraction de séquences de mots partageant la même distribution dans un (sous-)corpus. Chaque ligne du mini-corpus arabe-français est un couple d'énoncés traductions les uns des autres. Nous faisons l'hypothèse que les séquences de mots qui apparaissent strictement aux mêmes énoncés sont traductions les unes des autres. Les discontinuités sont indiquées par un tiret bas (_). Ici, l'article français *Un* ne sera pas aligné parce qu'aucun mot arabe n'apparaît strictement aux mêmes énoncés — il sera au mieux aligné avec la chaîne vide. Les alignements sont triés ici par ordre d'apparition des mots dans le corpus.

les séquences de mots partageant la même distribution, qui se trouvent être majoritairement constituées de mots rares.

| ANGLAIS ($e$) | FRANÇAIS ($f$) | ALLEMAND ($g$) | DÉC. | PROB. DE TRAD. | | | POIDS LEXICAUX | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $P(f,g\|e)$ | $P(e,g\|f)$ | $P(e,f\|g)$ | $L(f,g\|e)$ | $L(e,g\|f)$ | $L(e,f\|g)$ | |
| loud applause ↔ | vifs applaudissements ↔ | lebhafter beifall | 122 | 0,73 | 0,76 | 0,83 | 0,94 | 0,99 | 0,99 | ✓ |
| loud applause ↔ | vifs applaudissements ↔ | starker beifall | 24 | 0,14 | 0,14 | 0,82 | 0,94 | 0,99 | 0,90 | ✓ |
| loud applause ↔ | vifs applaudissements ↔ | ( lebhafter beifall ) | 12 | 0,07 | 0,09 | 0,67 | 0,94 | 0,99 | 0,06 | |
| loud applause ↔ | applaudissements prolongés ↔ | lebhafter beifall | 8 | 0,05 | 0,17 | 0,05 | 0,92 | 0,99 | 0,99 | ✓ |
| loud applause ↔ | ↔ | beifall | 1 | 0,01 | 0,00 | 0,01 | 0,84 | 1,00 | 0,99 | |

TABLEAU 10 – Exemples d'alignements multilingues avec leurs probabilités de traduction et poids lexicaux. Nous avons utilisé à cette occasion la partie allemande du corpus Europarl (*e*, *f*, *g* : *English*, *French*, *German*). Les alignements anglais-français-allemand présentés sont tous ceux dont la partie anglaise est *loud applause*, obtenus en exécutant notre système pendant cinq minutes sur le même échantillon de 20 000 énoncés d'Europarl que celui du tableau 7 page 83. Sur la première ligne, la première probabilité de traduction est P(*vifs applaudissements*, *lebhafter beifall* | *loud applause* = 122/(122+24+12+8+1) = 0,73. Le premier poids lexical est L(*vifs applaudissements*, *lebhafter beifall* | *loud applause*) = meilleure traduction pour *loud* × meilleure traduction pour *applause* = D(*vifs* | *loud*) × D(*beifall* | *applause*) = 0,94.

cadre de l'Union européenne où un texte de départ en anglais, français ou allemand est traduit vers une vingtaine d'autres langues. Nous ne calculerons donc qu'un seul score par langue, qui reflète les chances que la partie de l'alignement dans cette langue se traduise par toutes les autres simultanément. Il s'agit toujours de scores source-cible, mais la cible est constituée de l'ensemble de toutes les langues autres que la langue source.

Formellement, pour des alignements en L langues, une probabilité de traduction est calculée pour chaque langue $i$ ($1 \leqslant i \leqslant \mathrm{L}$) d'un alignement multilingue. Comme avant, ce score est la probabilité que la séquence de mots $s_i$ se traduise en le reste de l'alignement, soit le décompte de l'alignement, $\mathrm{C}(s_1, \ldots, s_\mathrm{L})$, divisé par la somme des décomptes de tous les alignements où $s_i$ apparaît, $\mathrm{C}(s_i)$ :

$$\mathrm{P}(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_\mathrm{L} | s_i) = \frac{\mathrm{C}(s_1, \ldots, s_\mathrm{L})}{\mathrm{C}(s_i)}$$

Dans le cas d'un alignement bilingue, ces scores correspondent bien au traditionnel couple P(*source*|*cible*) et P(*cible*|*source*). Dans le cas où les données ne sont constituées que d'une seule langue (voir sections suivantes), la probabilité est toujours $\mathrm{C}(s_1)/\mathrm{C}(s_1) = 1$.

De la même façon, les poids lexicaux s'obtiennent d'abord en définissant la distribution de probabilité D fondée sur les fréquences des mots dans le corpus :

$$\mathrm{D}(m_j | m_i) = \frac{\mathrm{C}(m_i, m_j)}{\mathrm{C}(m_i)} \qquad \text{avec } 1 \leqslant i \leqslant \mathrm{L}$$

puis en recherchant la meilleure traduction possible d'un mot $m_i$ issu d'une séquence $s_i$ parmi les mots de *toutes* les autres langues, selon la distribution D. Puis, comme dans le cas bilingue, le poids lexical d'un alignement pour la langue $i$ est le produit de toutes les probabilités conservées, après détermination de la meilleure traduction pour chaque mot de $s_i$ :

$$\mathrm{L}(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_\mathrm{L} | s_i) = \prod_{m_i \in s_i} \max_{m_j \in \bigcup_{i \neq j} s_j} \mathrm{D}(m_j | m_i)$$

Deuxième partie

# MISE EN ŒUVRE

## 6.2 ANYMALIGN : *ANY-NUMBER-OF-LANGUAGES*

### 6.2.1 *Généralisation au multilinguisme*

Nous généralisons [2] les définitions établies au chapitre 4 au cas du multilinguisme. Les exemples et définitions vus alors, bilingues, ne constituaient en fait que des cas particuliers. Le premier changement est le remplacement du corpus parallèle bilingue d'entrée par un corpus alingue. Le regroupement des mots en fonction de leurs distributions, ainsi que l'extraction des alignements par différence, s'en trouve simplifiés, car nous n'avons plus qu'une langue à traiter. Le principe est exactement le même que celui illustré par la figure 12 page 71, la seule différence résidant dans le fait que les séparations entre langues, notées par le caractère « ↔ » dans la figure 12, sont totalement absentes du processus, et ne sont rétablies qu'en sortie. En fait, les seuls véritables changements concernent l'attribution des scores. Le décompte des alignements reste inchangé, mais le calcul des probabilités de traduction ainsi que celui des poids lexicaux nécessitent une adaptation.

Traditionnellement, on attribue deux scores à chaque alignement, l'un reflétant les chances que la source se traduise par la cible et l'autre les chances que la cible se traduise par la source. Conformément à cela, nous pourrions attribuer à chacun de nos alignements autant de scores qu'il y a de couples de langues, c'est-à-dire calculer pour chaque langue d'un alignement les chances que la partie de l'alignement dans cette langue se traduise par chacune des autres parties dans les autres langues. Cela n'est cependant pas pertinent à notre sens, car le côté multilingue de notre approche perdrait alors tout son intérêt, puisque nous n'apprendrions rien de plus qu'avec une approche bilingue : autant aligner tous les couples de langues séparément. Nous sommes davantage intéressé par la traduction d'une langue source vers plusieurs langues cible simultanément, car cela correspond plus à une situation réelle de traduction. Le cas se présente en traduction non littéraire, avec par exemple la traduction de manuels d'utilisation, ou dans le

2 « *We can say that multilinguality implies generalization.* » (Giguet, 1996, section 1)

CORPUS PARALLÈLE MULTILINGUE :

| | 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 . |

⇓

CORPUS ALINGUE :

| | | |
|---|---|---|
| 1 | ١. ١فضلك ١من ١، ١قهوة | Un$_2$ café$_2$ ,$_2$ s'il$_2$ vous$_2$ plaît$_2$ .$_2$ One$_3$ coffee$_3$ ,$_3$ please$_3$ .$_3$ |
| 2 | ١. ١ممتازة ١قهوة ١هذه | Ce$_2$ café$_2$ est$_2$ excellent$_2$ .$_2$ This$_3$ coffee$_3$ is$_3$ excellent$_3$ .$_3$ |
| 3 | ١. ١ثقيل ١شاي | Un$_2$ thé$_2$ fort$_2$ .$_2$ One$_3$ strong$_3$ tea$_3$ .$_3$ |

FIGURE 27 – Assimilation d'un corpus multilingue à un corpus monolingue. Le corpus de départ est celui de la figure 11 page 65 auquel nous avons ajouté la traduction anglaise. Dans le nouveau corpus, les mots ont été distingués par indiçage sur les langues (1 pour l'arabe, 2 pour le français et 3 pour l'anglais), ce qui permet ici de distinguer les points et virgules français et anglais. Les séparations entre les langues ont été supprimées.

par la suite en tant que corpus *alingue*. Un exemple de tel corpus est présenté à la figure 27 ci-dessus. Ce corpus est le point de départ de tous les traitements. Au final, il n'est même plus nécessaire d'effectuer nos opérations monolingues sur toutes les langues séparément : indexation des mots et différences de chaînes ne sont réalisées que sur l'unique langue fictive constituée de la concaténation de toutes les langues de départ. En définitive, l'approche multilingue devient *plus simple* qu'une approche bilingue !

# 4

# ANYMALIGN : L'ALIGNEMENT PAR ÉCHANTILLONNAGE

ANYMALIGN est le logiciel d'alignement sous-phrastique que nous avons réalisé dans le cadre de cette thèse. Il repose essentiellement sur les mots rares, qu'il fait émerger dans des sous-corpus façonnés par échantillonnage. Ce chapitre présente son fonctionnement en détail. Ce fonctionnement est *simple*.

SOMMAIRE

## 4.1 VUE D'ENSEMBLE

### 4.1.1 *Les bases*

Le fonctionnement de base d'Anymalign se résume en deux propositions : tirer des sous-corpus du corpus de départ par échantillonnage et rechercher dans chacun d'eux les séquences de mots de même distribution.

### 4.1.2 *Extraction des alignements*

Concrètement, l'extraction des alignements consiste simplement à indexer chacun des mots du corpus et à réunir ceux qui apparaissent strictement dans les mêmes énoncés dans un même groupe, comme c'était le cas dans la figure 11, où chaque ligne du second tableau constituait un groupe. En outre, si les séquences de mots apparaissant dans les mêmes énoncés sont de bonnes traductions, comme nous en faisons l'hypothèse, alors les parties restantes de ces énoncés ont de grandes chances d'être de bonnes traductions également. Il s'agit là d'un principe couramment utilisé en traduction automatique par l'exemple : Cicekli et Güvenir (1996) font ainsi l'hypothèse que les parties similaires entre des phrases source sont traductions des parties similaires entre les phrases cible correspondantes, et qu'il en va de même pour les parties qui diffèrent. La principale différence avec notre travail est que nous ne traitons pas les énoncés deux par deux, mais par sous-corpus, pouvant être composés aussi bien de nombreux énoncés que d'un seul.

En définitive, chaque groupe de mots partageant la même distribution est susceptible de produire deux alignements par énoncé où il apparaît :

1. la séquence de mots constituée du groupe de mots lui-même, en préservant l'ordre des mots de l'énoncé ;
2. le complémentaire de cette séquence dans l'énoncé, c'est-à-dire ses contextes, ordonné également.



nous pouvons sans problème effectuer l'opération en parallèle sur davantage de langues :

su *içeceğim* ↔ *I will drink* water ↔ *Je vais boire* de l'eau
çay *içeceğim* ↔ *I will drink* tea ↔ *Je vais boire* du thé

Utiliser des opérations monolingues à des fins multilingues n'est donc pas une nouveauté, et remonte au moins aux premières expériences de Nagao (1984), reprises par la suite par Sato (1991). Nous proposons simplement d'appliquer ces opérations sur davantage de langues simultanément, plutôt que de nous limiter au traditionnel bilinguisme. Chacune des langues étant traitée isolément, le résultat d'une opération dans l'une n'influe aucunement sur les autres. Nous donnons à titre illustratif à l'annexe D page 151 le détail d'une des premières expériences par laquelle nous avons expérimenté l'utilisation parallèle d'opérations monolingues à des fins d'alignement multilingue.

### 6.1.3 *Plus fort : multilingue = monolingue*

Les concepts mis en jeu dans Anymalign sont également monolingues. D'une part, nous utilisons une opération de différence similaire à celle citée précédemment pour extraire les alignements. D'autre part, nous nous contentons de déterminer les distributions des mots dans le corpus, ce qui est réalisé séparément dans chaque langue.

En fait, nous pouvons simplifier encore davantage ce processus en assimilant notre corpus d'entrée multilingue, composé de plusieurs parties parallèles, à un unique corpus monolingue. Cette transformation peut être effectuée en distinguant toutes les formes de surface des mots en fonction de leur langue d'origine. Nous distinguons ainsi les mots de même graphie mais issus de langues différentes. Les séparations entre langues n'ont plus de raison d'être : elles sont purement et simplement supprimées et seront rétablies après le processus d'alignement, selon l'origine des mots.

Ce corpus étant une abstraction de plusieurs langues ne faisant intervenir aucune connaissance sur celles-ci, nous y faisons référence

dratique en le nombre de langues, soit par le recours à une langue naturelle comme « pivot », dont l'impact négatif sur la qualité des résultats est malheureusement assuré [1]. Dans le cas de l'alignement, Simard (1999) montre comment obtenir des alignements multilingues à partir d'alignements bilingues, mais son approche nécessite de déterminer quelles sont les deux langues les plus « similaires », ce qui implique de tester tous les couples. Giguet et Luquet (2006) produisent quant à eux des alignements en vingt langues, mais seuls les couples impliquant l'anglais sont alignés.

Nous préférons contourner ces obstacles plutôt que de chercher à les surmonter. Au lieu de chercher à créer des alignements multilingues en recombinant des alignements bilingues, nous produirons directement des alignements dans toutes les langues désirées. Aussi contradictoire que ceci puisse paraître, nous pensons que la seule façon d'atteindre ce multilinguisme véritable est de n'avoir recours qu'à des opérations *monolingues*. Le principe consiste à appliquer la même opération sur toutes les traductions d'un corpus en parallèle. Une telle opération peut tout simplement consister en une différence de chaînes, bien connue dans le domaine de la traduction par l'exemple, et que nous avons déjà évoquée à la section 4.1.2 page 70. Reprenons pour l'illustrer un exemple turc-anglais de Cicekli (2000) :

su *içeceğim* ↔ *I will drink* water
çay *içeceğim* ↔ *I will drink* tea

Nous nous contentons de rechercher les parties identiques d'un exemple à l'autre, ici en italique, au sein d'une des deux langues. La même chose est effectuée *séparément* dans l'autre langue, et nous faisons l'hypothèse que les parties identiques sont traductions les unes des autres. L'opération de différence elle-même est bien monolingue, et

1 Dans certains cas, le contraire peut également se produire : dans une expérience, Koehn et coll. (2009) montrent que les scores d'un système de traduction automatique probabiliste peuvent être meilleurs en utilisant l'anglais comme « pivot » que par traduction directe. La principale hypothèse formulée par les auteurs est qu'il s'agit d'un artefact dû au fait que les différentes traductions du corpus parallèle utilisé, Europarl, sont en fait toutes issues de l'anglais.



*Corpus d'entrée :* voir figure 11 page 65.

⇓

*Extraction des séquences de mots de même distribution et de leurs contextes :*

| LES MOTS : | APPARAISSENT DANS LES ÉNONCÉS : | D'OÙ NOUS EXTRAYONS : |
|---|---|---|
| قهوة ↔ café | 1 | قهوة ↔ café |
| | | من فضلك ، ↔ Un _ , s'il vous plaît . |
| | 2 | قهوة ↔ café |
| | | هذه _ ممتازة . ↔ Ce _ est excellent . |
| ⋮ | ⋮ | ⋮ |

⇓

*Collecte des alignements et décompte :*

| ARABE | | FRANÇAIS | DÉCOMPTE |
|---|---|---|---|
| قهوة | ↔ | café | 2 |
| من فضلك ، . | ↔ | Un _ , s'il vous plaît . | 1 |
| هذه _ ممتازة . | ↔ | Ce _ est excellent . | 1 |
| | | ⋮ | ⋮ |

FIGURE 12 – Extraction et décompte d'alignements à partir d'un (sous-)corpus. À partir du corpus d'entrée précédent, 2 alignements peuvent être extraits pour chacun des 6 groupes de la figure 11, soit 12 alignements distincts.

Ce principe est illustré par la figure 12 page précédente. Un alignement peut être obtenu plusieurs fois, à partir de différents sous-corpus ou de différents énoncés. Le résultat est une liste d'alignements accompagnés du nombre de fois qu'ils ont été obtenus. Dans le cas général, la méthode produit en sortie des séquences de mots discontinues. Elles peuvent ensuite être filtrées selon des critères particuliers tels que la contiguïté des mots, la présence ou l'absence de mots dans une des langues, ou de façon plus générale leur utilité sur une tâche bien précise. Par exemple, des séquences de mots discontinues peuvent constituer des patrons utiles en traduction automatique, mais il peut être préférable de ne conserver que les plus courts, car plus un patron est long, plus les chances de le rencontrer dans une phrase à traduire seront faibles.

### 4.1.3 *Un processus infini et plat*

Le processus décrit ci-dessus — tirer un sous-corpus aléatoirement, en extraire des alignements, recommencer — est potentiellement infini. Il peut s'arrêter à la demande ou si certaines conditions sont remplies, telles que le temps écoulé, la couverture du corpus de départ par les alignements, ou encore le nombre de nouveaux alignements obtenus par seconde.

Il serait tentant d'ajouter aux énoncés de départ les alignements nouvellement obtenus, à travers un processus incrémental, par exemple afin d'améliorer la couverture des alignements. Nous y renonçons pour les raisons suivantes :

- le nombre d'énoncés en entrée augmenterait de façon exponentielle, jusqu'à atteindre le nombre total de sous-séquences de mots des énoncés de départ. Cela n'est techniquement pas souhaité.
- en modifiant les données d'entrée, nous risquons d'altérer la distribution naturelle des mots du corpus de départ. Contrairement à l'ajout de données *réelles* dont il était question à la section 3.1.1 page 50, cela risquerait de créer des configurations particulièrement alambiquées, et en particulier des situations de



Ces deux sens sont distincts et une méthode peut relever plus ou moins de l'un ou de l'autre. Cela dit, les méthodes *réellement multilingues*, que cela relève du quantitatif ou du qualitatif, sont rares, voire inexistantes. En ce qui concerne la seconde acception, les seules méthodes capables de traiter indifféremment n'importe quelle langue sont à notre connaissance celles qui opèrent au niveau du caractère (Lepage, 2003; Lepage et Denoual, 2005; Denoual, 2006). En effet, nombre d'approches, bien qu'endogènes, utilisent certains a priori sur les langues, tels la présence d'une écriture alphabétique ou tout simplement de mots typographiques, alors que ces notions ne sont pas pertinentes dans toutes les langues. Anymalign n'est pas plus multilingue selon ce sens que la plupart des approches endogènes dans la mesure où il repose actuellement sur l'existence de mots typographiques. C'est selon le premier sens qu'il se distingue et que nous en revendiquons le multilinguisme véritable : nous allons voir comment notre approche permet l'alignement d'un nombre quelconque de langues simultanément, sans étape intermédiaire, et surtout sans que n'intervienne le moindre traitement par *couples* de langues. Notons que la notion même d'alignement multilingue n'est pas une nouveauté en soi, car elle a déjà été évoquée par Simard (1999), et plus récemment par Lemoine (2006) et Lecluze (2007) d'après les pistes proposées par le professeur Vergne. Nous contribuons en proposant une méthode d'alignement multilingue d'ores et déjà opérationnelle. Le multilinguisme auquel nous référerons par la suite sera toujours quantitatif.

### 6.1.2 *Des opérations monolingues pour un traitement multilingue*

Les approches les plus couramment utilisées sont bilingues par nature. Beaucoup ont sans doute subi l'influence des travaux portant sur la traduction automatique, qui étaient à l'origine asymétriques, dans l'optique de traduire d'une langue source vers une langue cible. Lorsque des résultats multilingues sont nécessaires, un traitement par couples de langues s'avère alors nécessaire. Cela se traduit soit par le traitement de tous les couples de langues, dont le nombre est qua-

## 6.1 ALIGNEMENTS SANS FRONTIÈRES

### 6.1.1 *Qu'est-ce qu'une méthode multilingue ?*

Les approches qualifiées de « multilingues » sont monnaie courante en traitement automatique des langues. Nous distinguons deux sens à ce terme. Dans la première acception, une méthode est considérée comme multilingue dès qu'elle met en jeu plusieurs langues dans un même traitement. Nous parlerons de multilinguisme « quantitatif ». Typiquement, on trouve de telles méthodes en recherche d'information translingue, en traduction automatique, ou encore dans certaines approches de désambiguïsation (p. ex. Agirre et coll., 2002) ou de reformulation (p. ex. Bannard et Callison-Burch, 2005). Les méthodes purement bilingues, comme c'est souvent le cas de ce qui touche de près ou de loin à la traduction automatique, constituent typiquement les plus élémentaires des méthodes multilingues quantitatives. Les qualifier de multilingues est à notre avis abusif, mais cela est malheureusement d'usage courant.

Dans la seconde acception, une méthode peut être qualifiée de multilingue si elle est suffisamment générique pour être validée sur plusieurs langues. Nous parlerons de multilinguisme « qualitatif ». Il peut tout aussi bien s'agir de méthodes ayant été appliquées en pratique sur une seule langue, par exemple le chunking sans marqueur (Vergne, 2009) ou une « simple » *tokenization* (Giguet, 1996), que plusieurs, comme en traduction automatique empirique. Ces méthodes sont typiquement endogènes : elles ne nécessitent aucune connaissance sur les langues autre que celles contenues dans le texte étudié, ce qui leur assure une certaine « portabilité » d'une langue à l'autre. Bender (2009) rappelle que, contrairement à ce que l'on pourrait penser, de telles méthodes ne font pas fi de toute connaissance linguistique, mais recourent au contraire à des invariants linguistiques induits par l'étude de la typologie des langues. Les propriétés qu'elles mettent en jeu relèvent donc plus du *langage* que des langues.



sur-représentativité de certains phénomènes. Cela biaiserait notre processus d'alignement.

- il est peu probable que les alignements nouvellement extraits soient de même qualité que les énoncés de départ. Plus nous ajouterons de données non fiables aux données de départ, moins les alignements extraits seront fiables. Une solution serait d'attribuer un score à chacun, le score maximal étant attribué aux couples d'énoncés du corpus de départ, et éventuellement de n'ajouter que les meilleurs, mais cela compliquerait le processus pour obtenir au total des bénéfices pas nécessairement significatifs.

Comme nous le verrons par la suite, un tel processus itératif n'est de toute façon pas nécessaire car la couverture de nos alignements est très rapidement supérieure à celle des alignements obtenus par les systèmes classiques. Nous nous contenterons donc à chaque fois de réutiliser notre connaissance de départ pour constituer nos sous-corpus et en extraire des alignements.

## 4.2 DÉTERMINATION DES TAILLES DE SOUS-CORPUS OPTIMALES

Le seul véritable paramètre que nécessite notre méthode est la taille, primordiale, des sous-corpus à partir desquels extraire les alignements. Le nombre de sous-corpus possibles dépend de leur taille et suit une gaussienne similaire à la cloche de la page 58 : pour un corpus de départ de taille N, il n'existe qu'un seul sous-corpus de taille N, un seul sous-corpus de taille zéro, mais un maximum de sous-corpus de taille N/2. Les sous-corpus que nous constituerons ne seront pas majoritairement de taille N/2 pour autant, car ce qui compte n'est pas tant le nombre de sous-corpus possibles que le nombre d'alignements corrects qu'ils sont susceptibles de produire. Nous montrons expérimentalement que les sous-corpus de petite taille doivent être privilégiés si l'on veut obtenir les meilleurs résultats possibles. Les expériences décrites ci-après sont réalisées sur le couple espagnol-français de notre corpus Europarl. Les

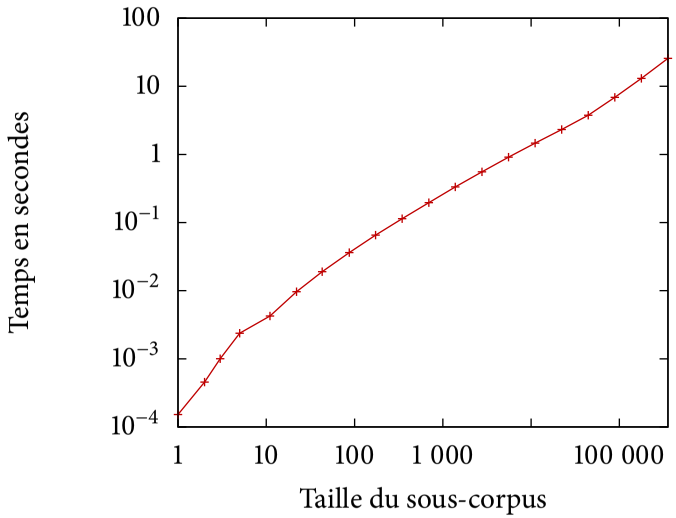



FIGURE 13 – Temps nécessaire au traitement d'un sous-corpus en fonction de sa taille en nombre d'énoncés. Le temps de traitement semble augmenter quasi-linéairement avec la taille du sous-corpus.

alignements obtenus sont filtrés de sorte que seuls les alignements comportant des mots en source et en cible soient conservés.

### 4.2.1 *Davantage d'alignements avec de petits sous-corpus*

D'abord, d'un point de vue pratique, plus un sous-corpus est petit, plus il est rapide à traiter. La figure 13 ci-dessus montre que le temps de traitement d'un sous-corpus est approximativement linéaire en le nombre d'énoncés dont il est constitué. Traiter 1 000 sous-corpus de 100 énoncés prendra donc typiquement autant de temps que traiter un seul sous-corpus de 100 000 énoncés. Pour une juste comparaison entre les différentes tailles de sous-corpus, les mesures que nous effectuerons à l'avenir seront fonction du temps de traitement total plutôt que du nombre de sous-corpus traités.

# 6

# DU MULTILINGUISME EN ALIGNEMENT

Nous étendons le concept d'alignement bilingue à celui d'alignement multilingue. Nous montrons comment cette généralisation peut être conduite en abordant le problème sous un angle monolingue. Cette généralisation — cette *simplification* — est appliquée à notre méthode. Nous proposons enfin des applications susceptibles de tirer parti d'alignements réellement multilingues.

SOMMAIRE

– Anymalign fournit des résultats de qualité légèrement inférieure à l'état de l'art dans la tâche de constitution de tables de traductions à des fins de traduction automatique ayant recours aux segments, mais surpasse aisément ses concurrents dans les tâches de constitution de lexiques dès que la taille du corpus d'entrée est suffisamment grande ;
– la raison principale est que le fonctionnement propre d'Anymalign en fait un spécialiste de l'extraction des mots de même fréquence. Pour faire sauter cet ultime verrou, nous devrons dans le futur recombiner les alignements déjà produits afin d'aligner davantage de n-grammes, ou faire évoluer notre méthode vers l'indexation de séquences de mots plutôt que de se contenter de l'indexation de mots isolés.



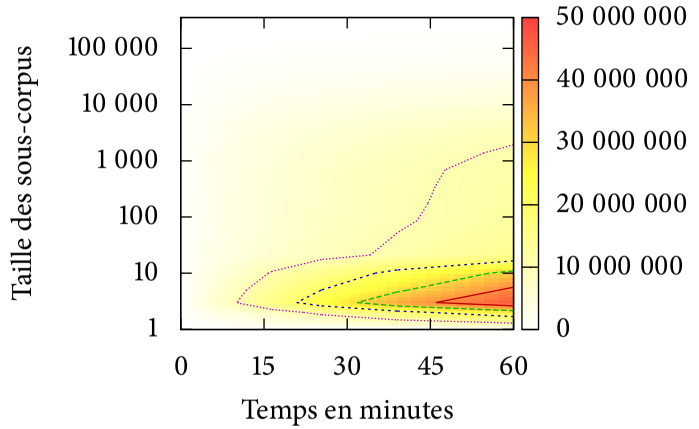



FIGURE 14 – Nombre d'alignements distincts obtenus par notre méthode en fonction du temps et de la taille des sous-corpus. Les sous-corpus de petite taille donnent davantage d'alignements, plus rapidement. Les quatre courbes de niveau correspondent aux multiples de 10 millions d'alignements.

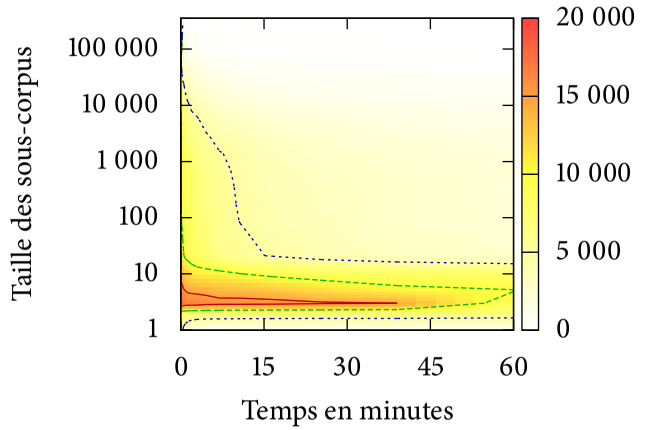



FIGURE 15 – Nombre de nouveaux alignements distincts obtenus par notre méthode en fonction du temps et de la taille des sous-corpus. Le nombre de nouveaux alignements diminue avec le temps. Les trois courbes de niveau correspondent aux multiples de 5 000.

Ensuite, les sous-corpus de petite taille sont beaucoup plus productifs en alignements distincts que ceux de grande taille, comme le montre la figure 14 page précédente. Typiquement, plus le temps écoulé est important — donc plus le nombre de sous-corpus traités est important —, plus le nombre d'alignements obtenus est important. Comme nous pouvions nous y attendre, à l'extrémité basse de la figure, le nombre d'alignements obtenus par des sous-corpus constitués d'un seul énoncé tend vers le nombre d'énoncés du corpus de départ, soit N = 354 645 : ces alignements sont ni plus ni moins les alignements d'énoncés de départ. Les sous-corpus pour lesquels le nombre d'alignements augmente le plus rapidement sont constitués de deux à dix énoncés environ. Le nombre d'alignements distincts décroît ensuite progressivement lorsque la taille des sous-corpus augmente. Ainsi, à l'extrémité haute de la figure, et bien que cela n'y soit pas discernable à l'œil, l'unique sous-corpus de taille N produit l'intégralité de ses 45 548 alignements possibles en une seule fois après le temps nécessaire à son traitement, soit environ 25 secondes, pour ne plus produire aucun nouvel alignement.

Notons que du point de vue de l'implémentation, il ne sera pas nécessaire de constituer aléatoirement les sous-corpus de tailles extrêmes, car leur traitement exhaustif est non seulement possible, mais surtout bien plus rapide : il est inutile de tirer plusieurs fois le sous-corpus de taille N ou de compter sur le hasard pour finir par tirer tous ceux de taille 1. La figure 15 page précédente montre la dérivée en fonction du temps du graphique de la figure 14, soit le nombre de nouveaux alignements distincts en fonction du temps et de la taille des sous-corpus. Quelle que soit la taille des sous-corpus, le nombre de nouveaux alignements diminue en fonction du temps, mais ce phénomène est moindre avec les petits sous-corpus. Au total, les petits sous-corpus peuvent bien produire davantage d'alignements.



En définitive, cela peut mener à la *non*-extraction d'alignements aussi simples que

la maison ↔ the house

à cause de la différence de fréquence entre le déterminant et le nom. Pourtant, la méthode produira bien séparément les deux alignements

la ↔ the  maison ↔ house

car typiquement, la méthode construira des sous-corpus où les deux déterminants partagent la même distribution (idem pour les deux noms).

Pour faire sauter cet ultime verrou, il ne nous resterait en fait qu'à recombiner les alignements produits par Anymalign afin d'en produire de plus longs. Cela n'est en fait rien de plus que ce qui a été introduit par la traduction probabiliste par segments (Koehn et coll., 2003), où les alignements de chaînes de mots sont obtenus en combinant des alignements mot-à-mot. Dans l'exemple ci-dessus, refaire une passe sur le corpus d'origine pour détecter la succession entre le déterminant et le nom, suivie d'un recalcul des probabilités de traduction à partir du nouvel ensemble d'alignements, suffirait pour s'acquitter de cette tâche. Une autre approche consisterait à ne plus fonder notre alignement sur l'indexation de mots, mais sur l'indexation de *séquences* de mots. Dans des expériences préliminaires, le simple recours à des n-grammes de mots sur un corpus parallèle français-anglais a suffi pour améliorer nos scores en traduction automatique de façon très significative. Cette piste de recherche constitue actuellement notre priorité.

## RÉSUMÉ

Le comportement de notre méthode se démarque manifestement de celui des approches statistiques :

– le côté *anytime* d'Anymalign le rend toujours nettement plus rapide que les outils auxquels nous l'avons comparé ;

en est fortement peuplée. Pour confirmer ou réfuter cette observation, nous nous concentrons sur les bigrammes de mots dans une seconde expérience. Nous évaluons à présent le nombre de bigrammes source d'une table de traductions en fonction du nombre d'occurrences des deux mots dont ils sont composés. Nous traçons alors les distributions correspondantes, et les comparons à la distribution naturelle des bigrammes dans la partie source de notre corpus Europarl. Les résultats sont présentés dans les figures 25 et 26 pages 110 et 111. La distribution des bigrammes de la table de traductions produite par MGIZA++ est très proche de celle d'Europarl. La distribution d'Anymalign est par contre très différente : aucun bigramme n'est visible dans les régions en haut à gauche et en bas à droite de la figure, alors qu'il s'agit de zones très denses dans les distributions de Moses et d'Europarl. Cela montre que l'approche par échantillonnage serait spécialisée dans l'alignement de termes de même fréquence.

### 5.3.3 *Le dernier verrou*

Pour bien saisir l'origine de cette spécialisation de notre méthode, il suffit de se remémorer son fonctionnement : elle produit en sortie des séquences de mots partageant exactement la même distribution dans un sous-corpus. La raison pour laquelle des mots de fréquences différentes ne sont pas alignés ensemble est que les mots de haute fréquences, par exemple un point de fin de phrase, et les mots de basse fréquence, typiquement un hapax, ne partagent fatalement jamais la même distribution. La seule configuration dans laquelle un point et un hapax pourraient partager la même distribution serait celle d'un sous-corpus d'un seul énoncé. Mais dans ce cas, *tous* les mots partageraient la même distribution : la méthode ne pourrait extraire que l'énoncé dans son intégralité, ce qui ne fournirait aucune information nouvelle. Cela est d'autant plus vrai sur de longs énoncés, tels que ceux que l'on trouve dans le corpus Europarl.



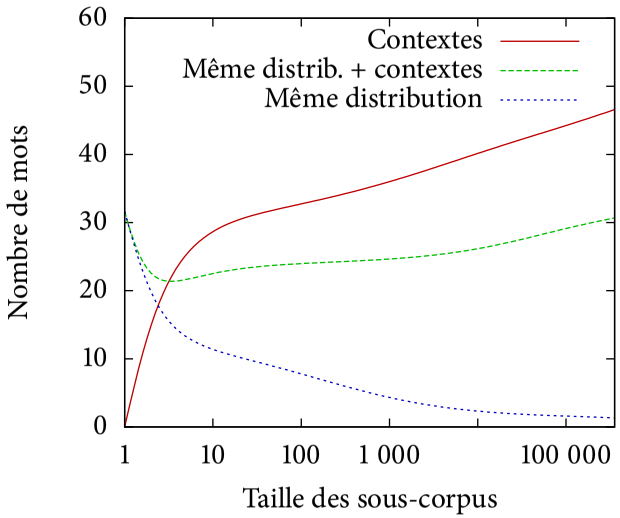



FIGURE 16 – Longueur moyenne des alignements obtenus par notre méthode en fonction de la taille des sous-corpus dont ils ont été extraits. Plus les sous-corpus sont petits, plus les séquences de mots de même distribution sont longues et plus leurs contextes sont courts. La tendance s'inverse avec de grands sous-corpus. La courbe rassemblant tous les alignements a tendance à suivre celle des contextes car ceux-ci sont plus nombreux. Les parties source et cible des alignements sont ici confondues, mais nous ne nous intéressons qu'à l'allure générale des courbes.

### 4.2.2 *De meilleurs alignements avec de petits sous-corpus*

La taille des sous-corpus influe également sur la qualité des alignements extraits, ce qui est primordial. Plus les sous-corpus sont grands, plus les séquences de mots partageant la même distribution sont petites, et plus leurs contextes sont grands. Inversement, plus ils sont petits, plus les séquences sont grandes et plus les contextes sont petits. La figure 16 ci-dessus confirme cette tendance. Or, nous avons vu au chapitre précédent qu'il pouvait être théoriquement préférable de segmenter un énoncé en deux parties les plus inégales possibles, ce qui

est directement faisable à partir de sous-corpus de tailles extrêmes, a fortiori à partir des sous-corpus constitués d'un seul énoncé : les séquences de mots de même distribution sont alors toujours constituées des énoncés complets et leurs contextes sont vides. Il s'agit d'ailleurs ici du seul type d'alignement dont la validité est absolument certaine, puisque c'est notre connaissance de départ. Plus généralement, moins un sous-corpus comprend d'énoncés, plus les alignements extraits ont de chances d'être valides.

Pour l'illustrer, considérons un cas extrême : l'alignement à partir d'un corpus parallèle japonais-anglais, sans que la partie japonaise ne soit segmentée en mots. Chaque énoncé japonais n'est donc constitué que d'un seul mot typographique, qui est presque toujours un hapax en corpus. En appliquant notre méthode au sous-corpus constitué du corpus entier, chaque hapax source japonais ne sera aligné qu'avec les éventuels hapax cible anglais, qui ne couvrent assurément pas tous les mots de l'énoncé. Le seul alignement correct serait pourtant l'alignement d'un mot-énoncé source avec l'énoncé cible correspondant dans sa totalité, alignement que les méthodes statistiques dont nous nous distinguons sont tout à fait capables d'obtenir — nous en avons fait l'expérience. Avec notre méthode, cela n'est en fait possible qu'à partir de sous-corpus de très petite taille : lorsque la taille du sous-corpus tend vers un, tous les mots tendent à devenir hapax en énoncé, et par conséquent l'alignement tend à devenir correct. Ce cas est extrême, mais montre bien qu'en termes de qualité d'alignement, c'est à partir des sous-corpus les plus petits possibles que notre méthode donne le meilleur d'elle-même. Ce n'était pas un hasard si le nombre d'alignements d'hapax erronés était si grand dans le tableau 5 page 43 en anglais-finnois comparé à l'espagnol-français : la nature différente du mot typographique des langues source et cible, *combinée avec la grande taille du corpus* à partir duquel les alignements étaient extraits, n'y étaient pas étrangères.

Pour conclure cette étude, nous effectuons une dernière expérience faisant autorité : nous comptons le nombre d'alignements présents dans un dictionnaire espagnol-français de référence en fonction de la



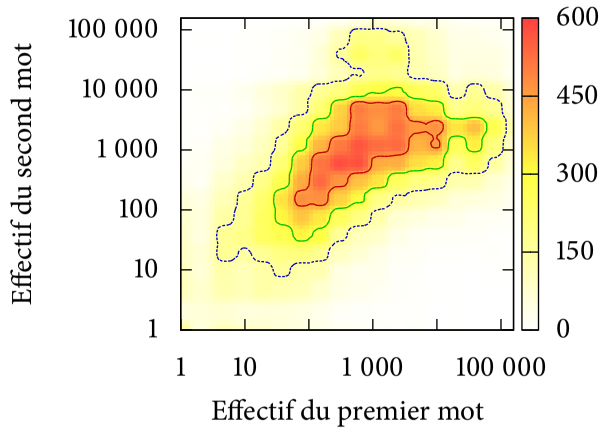
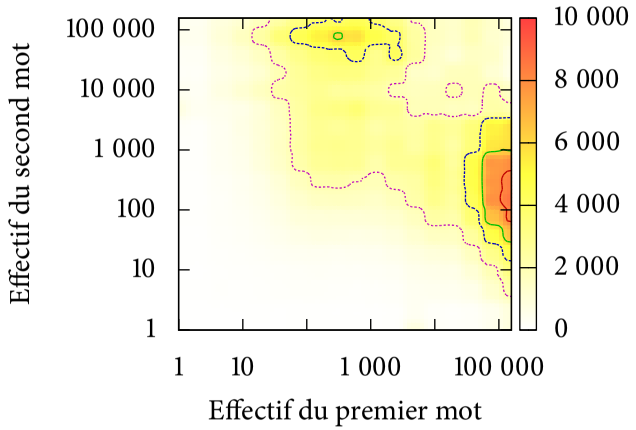


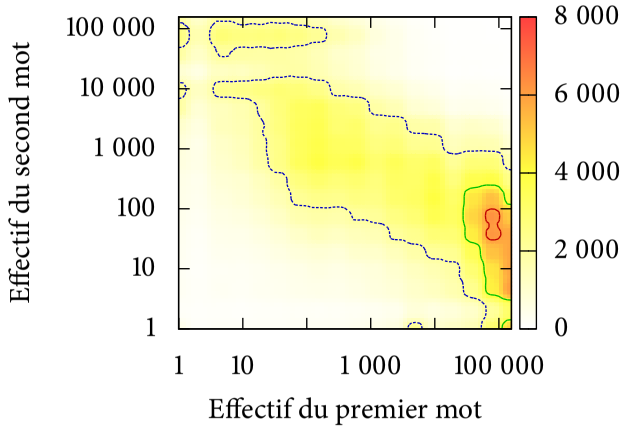

FIGURE 26 – Distribution des bigrammes de la partie française de la table de traductions obtenue à partir d'Anymalign en fonction des effectifs des mots qui les composent. Les bigrammes sont majoritairement constitués de mots de même fréquence, d'où l'esquisse d'une diagonale partant du bas à gauche et allant vers le haut à droite. Les courbes de niveau correspondent au multiples de 150.

laquelle un système de traduction automatique basé sur des segments serait handicapé s'il est construit à partir de notre méthode ne serait pas une question de *qualité* des n-grammes, mais plutôt de *quantité* : la méthode n'aligne simplement pas de n-grammes avec $n \geq 2$ en nombre suffisant. Il semblerait également que les alignements produits soient différents : par exemple, plus de 30 % des bigrammes produits par Anymalign ne sont pas obtenus à partir de MGIZA++. Par conséquent, nous pourrions gagner à combiner les tables de traductions obtenues par les deux systèmes, comme l'ont par exemple tenté Srivastava et coll. (2009) ; nous gardons cette expérience pour des recherches futures.

Une inspection manuelle du contenu des tables de traductions suggère que ce que n'aligne pas Anymalign est en fait constitué par des séquences de mots de différentes natures, telles qu'un mot suivi d'une ponctuation, alors que la table de traductions produite par MGIZA++




(a) Europarl




(b) MGIZA++/Moses

FIGURE 25 – Distributions des bigrammes des parties françaises d'Europarl et de la table de traductions obtenue à partir de MGIZA++ en fonction des effectifs des mots qui les composent. Les deux distributions sont similaires : les bigrammes sont majoritairement constitués de mots de fréquences très différentes, d'où l'esquisse d'une diagonale partant du haut à gauche et allant vers le bas à droite sur les deux graphiques. Les courbes de niveau correspondent au multiples de 2 000.



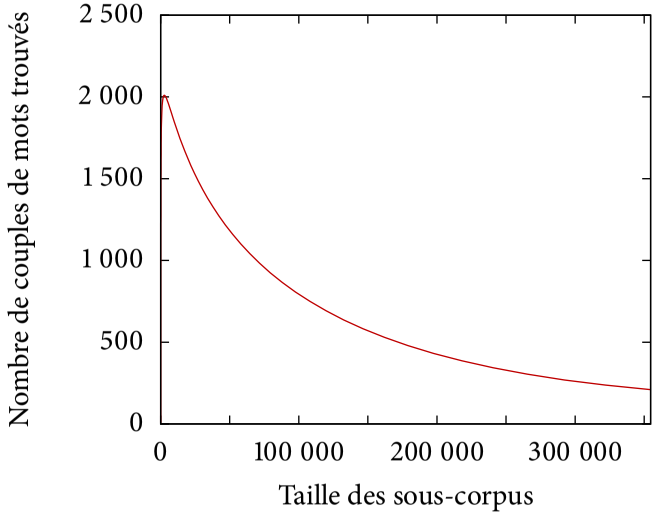



FIGURE 17 – Nombre d'alignements de mots trouvés dans un dictionnaire de référence en fonction de la taille des sous-corpus. Le système a été exécuté pendant 25 secondes dans cette expérience, ce qui correspond au temps nécessaire au traitement du sous-corpus de taille N = 354 645. La courbe augmente très rapidement jusqu'à atteindre son maximum pour des sous-corpus d'environ 1 000 énoncés, avant de chuter progressivement.

taille des sous-corpus. Ce dictionnaire est principalement constitué d'unigrammes. Les résultats sont présentés à la figure 17 ci-dessus. Le maximum est atteint pour des tailles de sous-corpus d'environ 1 000 énoncés dans cette expérience. Cela montre que les sous-corpus de — relativement, cette fois — petite taille peuvent bien produire des alignements à la fois courts et de bonne qualité.

### 4.2.3 *Optimisation de l'échantillonnage*

La conclusion des expériences précédentes est que les sous-corpus de petite taille doivent être privilégiés, car ils sont plus rapides à traiter, pro-

duisent davantage d'alignements, et les alignements qu'ils produisent sont plus fiables. Pour des raisons de couverture, les sous-corpus de taille plus conséquente ne sont pas à négliger pour autant, mais nous traiterons simplement davantage de petits sous-corpus. Idéalement, le choix des tailles de sous-corpus doit être tel que chacun des énoncés doit apparaître dans le plus grand nombre possible de sous-corpus les plus variés possibles. Nous proposons donc de déterminer une distribution par laquelle l'échantillonnage doit assurer une certaine couverture du corpus de départ. Cette distribution est uniquement fonction de la taille des sous-corpus. Une fois déterminée la taille du prochain sous-corpus à traiter, les énoncés le constituant seront piochés aléatoirement dans le corpus de départ selon une distribution uniforme.

Nous notons $x_k$ le nombre de sous-corpus de taille $k$ à traiter. $x_k$ est défini comme suit : il doit garantir que la probabilité qu'aucun des énoncés d'un sous-corpus de $k$ énoncés ne soit jamais choisi soit inférieure à un certain seuil $t$. Ainsi, $t$ est un indicateur de la couverture du corpus d'entrée : plus il est proche de zéro, meilleure est la couverture. Soit $n$ la taille du corpus d'entrée bilingue ($1 \leqslant k \leqslant n$) :

- la probabilité qu'un énoncé donné soit choisi dans un échantillon de taille $k$ est $k/n$ ;
- la probabilité que cet énoncé ne soit pas choisi est $1 - k/n$ ;
- la probabilité qu'aucun des $k$ énoncés ne soit choisi est $(1 - k/n)^k$ ;
- la probabilité qu'aucun de ces $k$ énoncés ne soit jamais choisi est $(1 - k/n)^{kx_k}$.

Le nombre de sous-corpus de taille $k$ à constituer par échantillonnage est ainsi contraint par $(1 - k/n)^{kx_k} \leqslant t$, ce qui donne après résolution :

$$x_k \geqslant \frac{\log t}{k \log (1 - k/n)}$$

Le traitement de $x_k$ sous-corpus aléatoires de taille $k$ garantit ainsi la couverture voulue du corpus d'entrée.

Plutôt que de définir à l'avance un degré de couverture particulier, ce qui implique un nombre fixe de sous-corpus à traiter, nous déduisons



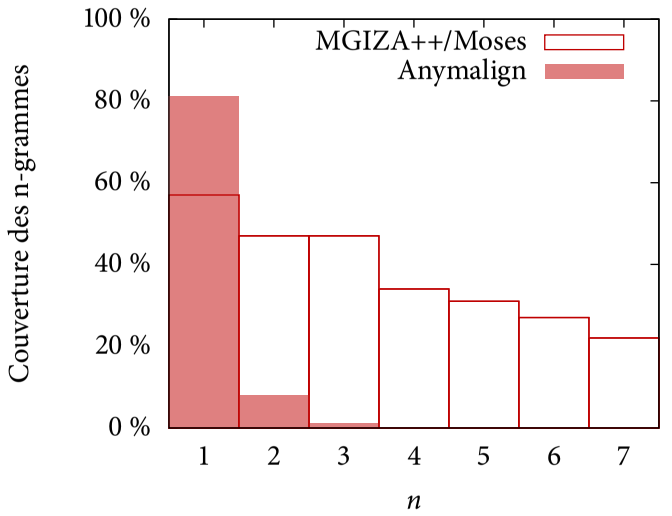



FIGURE 24 – Couverture de la partie française de notre corpus d'entraînement Europarl par les tables de traductions produites par MGIZA++ et Anymalign. La couverture des unigrammes est clairement meilleure avec Anymalign qu'avec MGIZA++. Anymalign est par contre manifestement très en retrait pour toutes les tailles supérieures de n-grammes.

Cependant, Anymalign est loin derrière sur tous les n-grammes restants. Dans l'ensemble, la table de traductions produite à partir de MGIZA++ est quatre fois plus grande que celle d'Anymalign : plus de quatre millions d'entrées dans la première contre à peine plus d'un million dans la seconde. En pratique, nous avons montré dans une expérience que ces différences étaient d'autant plus prononcées que le corpus d'entraînement était grand. La tendance s'inverse même sur des corpus de « petite » taille, en particulier sur des échantillons de quelques dizaines de milliers d'énoncés du BTEC — où, soit dit en passsant, nous avons obtenu de meilleurs résultats que l'état de l'art en traduction automatique. Ces résultats suggèrent que la raison pour

notre but n'est pas ici de produire des traductions de qualité. Comme dans la section 5.2.3, MGIZA++ est exécuté avec trois itérations, BerkeleyAligner avec deux, et Anymalign pendant deux heures.

Les nouveaux scores sont de 0,68 pour Anymalign, 0,68 pour MGIZA++ et 0,67 pour BerkeleyAligner. Comme prévu, les scores obtenus en traduction automatique sont moins bons que précédemment : TER augmente d'environ 0,04 pour MGIZA++ et BerkeleyAligner. Le point le plus remarquable est que le score d'Anymalign demeure inchangé, et est désormais au même niveau que ceux des deux autres outils, alors qu'il était inférieur de façon significative dans notre première expérience. Il s'agit donc bien d'une question de *segments* : le décodeur de Moses n'aurait en fait sélectionné que des unigrammes dans la table de traductions au cours de notre première expérience (section 5.2.1). Les alignements d'unigrammes d'Anymalign seraient-ils bons au point que le décodeur ne voie qu'eux, ou au contraire, ses alignements de segments seraient-ils de moindre qualité ?

### 5.3.2 *Spécialiste des mots de même fréquence*

Nous examinons la raison pour laquelle l'approche par échantillonnage ne permettrait pas l'alignement correct de n-grammes. Dans ce but, nous examinons minutieusement le contenu des tables de traductions produites par Anymalign, exécuté pendant deux heures, et le comparons avec celui obtenu à partir de MGIZA++, exécuté avec trois itérations. Nous nous intéressons tout particulièrement à la différence entre les unigrammes et les n-grammes plus longs. Par conséquent, dans une première expérience, nous comptons simplement le nombre d'entrées source dans les tables de traductions produites par MGIZA++/Moses et Anymalign, pour chaque n-gramme de longueur 1 à 7. Ces valeurs correspondent aux longueurs minimale et maximale par défaut des segments dans la table de traductions de Moses.

Les résultats sont présentés sur la figure 24 page suivante. La couverture de la table de traductions d'Anymalign est nettement supérieure sur les unigrammes : plus de 80 % du vocabulaire source est couvert.

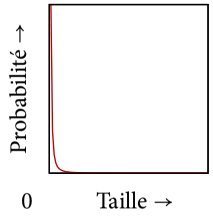


sons du résultat précédent une distribution de probabilité pour tirer aléatoirement la taille du prochain sous-corpus à traiter :

$$\mathrm{p}(k) = \frac{-1}{k \log\left(1 - k/n\right)}$$

Le numérateur, log(t), a été remplacé par –1 parce que $t$ est une constante : $t \leqslant 1 \Rightarrow \log(t) \leqslant 0$. Cette égalité n'étant pas normalisée, nous la normaliserons lors de l'implémentation de façon à ce que la somme suivante soit vérifiée :

$$\sum_{i=1}^{n} \mathrm{p}(k) = 1$$




L'allure de cette distribution est donnée par le graphe ci-contre. Elle privilégie grandement les sous-corpus de petite taille, dont nous avons précédemment montré qu'ils nous seraient bénéfiques. La comparaison entre les résultats obtenus avec cette distribution et ceux obtenus avec une distribution uniforme sera examinée dans le chapitre suivant.

## 4.3 FINITIONS

La dernière étape pour rendre nos alignements directement utilisables est de leur attribuer des scores. Pratiquement, nous en faisons une *table de traductions*, qui, dans la terminologie actuelle, consiste en une liste d'alignements auxquels sont affectés un certain nombre de traits. Nos traits sont ceux qui ont été initialement proposés par Koehn et coll. (2003) : les probabilités de traduction, fondées sur la fréquence des alignements obtenus, et les poids lexicaux, obtenus à partir des décomptes des mots dans le corpus.

### 4.3.1 *Probabilités de traduction*

Les probabilités de traduction sont calculées à partir du nombre de fois que chacun des alignements a été obtenu. Il s'agit de probabilités

observées. Elles reflètent la probabilité qu'une séquence de mots dans une langue se traduise en une séquence de mots dans l'autre langue.

Formellement, nous calculons un score pour chaque langue $i \in \{1 ; 2\}$ d'un alignement bilingue. Ce score est la probabilité que la séquence de mots $s_i$ se traduise par le reste de l'alignement. Il est obtenu en divisant le décompte de l'alignement, $\mathrm{C}(s_i, s_i)$, par la somme des décomptes de tous les alignements où $s_i$ apparaît, $\mathrm{C}(s_i)$ :

$$\mathrm{P}(s_j|s_i) = \frac{\mathrm{C}(s_i, s_j)}{\mathrm{C}(s_i)} \qquad \text{avec } i \neq j$$

Le tableau 7 page ci-contre donne un exemple sur des données réelles.

En plus d'attester du nombre de traductions possibles d'une séquence de mots, les probabilités de traduction constituent notre principal indicateur de qualité des alignements : une probabilité très petite indique typiquement un mauvais alignement. Donc, dans le cas où un mauvais alignement serait produit, il obtiendrait systématiquement de mauvais scores.

### 4.3.2 *Poids lexicaux*

Les poids lexicaux ont été proposés par Koehn et coll. pour valider la qualité des alignements. Ils correspondent en quelque sorte à un score de confiance. On sait qu'ils améliorent légèrement la qualité des traductions obtenues par méthodes probabilistes. Étant donné un alignement bilingue, il s'agit d'observer par quels mots cible se traduit chacun des mots source. Lorsqu'un mot source se traduit par plusieurs mots cible, c'est la moyenne de leurs probabilités de traduction qui est utilisée. Un poids lexical source-cible est alors le produit de ces scores. Le même principe est appliqué de la cible vers la source, et le résultat est un couple de poids lexicaux compris entre zéro et un. Nous adaptons cette technique avec deux changements majeurs.



minue de façon significative, d'où une baisse de la F-mesure. Le rappel d'Anymalign augmente de la même façon, mais sa précision est stable : cela signifierait que notre méthode est moins sensible au bruit. Notons que BerkeleyAligner est capable d'obtenir un score non nul à partir d'un corpus d'entrée constitué d'un unique couple d'énoncés, mais il « triche » : sa table de traductions contient des alignements entre à peu près tous les segments source et cible. À l'opposé, MGIZA++ impose une taille minimale de corpus pour pouvoir fonctionner, d'où l'absence de point d'abscisse inférieure à 100 énoncés environ.

## 5.3 EXAMEN DU CONTENU DES ALIGNEMENTS

Un des faits les plus marquants de la section précédente est la grande différence de qualité entre les résultats obtenus lors de nos deux évaluations : Anymalign produit de bien meilleurs alignements d'unigrammes dès que la taille du corpus d'entrée est suffisamment grande, mais mène à des résultats légèrement moins bons en traduction automatique. Nous mettons les origines de ces disparités en évidence dans cette section.

### 5.3.1 *Spécialiste des unigrammes*

Nous avons vu dans les expériences précédentes qu'Anymalign produit de bien meilleurs résultats sur la tâche d'induction de lexiques bilingues que sur celle de production de tables de traductions pour la traduction automatique probabiliste par segments. Les lexiques bilingues de référence que nous utilisons étant principalement constitués d'unigrammes, nous concluons naturellement qu'Anymalign produit de meilleurs alignements d'unigrammes.

Pour le vérifier, nous répétons l'évaluation de traduction automatique réalisée à la section 5.2.1, mais en ne considérant cette fois que les unigrammes. Cela implique de créer des tables de traductions constituées uniquement de couples de mots (*source*, *cible*). Le décodeur ne fait plus alors que du mot-à-mot, renouant avec les débuts de la traduction probabiliste, ce qui se répercute sur la qualité des résultats, mais

FIGURE 23 – Comportement des aligneurs en fonction de la taille du corpus d'entrée. Les points de l'extrémité droite de la figure correspondent aux points dont l'abscisse vaut environ 2 h sur la figure 21 page 102. Anymalign est le plus régulier des trois aligneurs : il semble toujours donner de meilleurs résultats lorsque la quantité de données augmente. La qualité des résultats de MGIZA++ et BerkeleyAligner diminue à partir d'une certaine taille.

commencer à fournir de bons résultats qu'avec des corpus d'entrée relativement grands.

Le comportement des deux autres aligneurs semble également surprenant : leurs scores augmentent jusqu'à des tailles que l'on pourrait qualifier de petites, c'est-à-dire environ 200 énoncés pour BerkeleyAligner et 1 000 pour MGIZA++, avant de redescendre, alors qu'Anymalign semble plafonner. Donc, contrairement à certaines idées reçues, des outils statistiques peuvent fournir de meilleurs résultats sur de petits corpus. En fait, le rappel des deux outils continue d'augmenter légèrement sur les « grandes » tailles de corpus, mais leur précision di-



| ESPAGNOL ($e$) | | FRANÇAIS ($f$) | DÉCOMPTE | PROB. DE TRAD. | | POIDS LEXICAUX | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $P(f\|e)$ | $P(e\|f)$ | $L(f\|e)$ | $L(e\|f)$ | |
| interesantes 'intéressants' | ↔ | intéressants | 10 | 0,59 | 0,50 | 0,33 | 0,80 | ✓ |
| fascinante 'fascinant' | ↔ | intéressants | 8 | 0,33 | 0,40 | 0,50 | 0,20 | ✓ |
| quinquenal estadístico 'programme quinquennal statistique' | ↔ | intéressants | 1 | 0,50 | 0,05 | 0,08 | 0,20 | |
| interesante 'intéressant' | ↔ | intéressants | 1 | 0,02 | 0,05 | 0,03 | 0,20 | |

TABLEAU 7 – Exemples d'alignements avec leurs probabilités de traduction et leurs poids lexicaux. Les alignements espagnol-français présentés sont tous ceux dont la partie française est le mot *intéressants*, obtenus en exécutant notre système pendant deux minutes sur un échantillon de 20 000 énoncés d'Europarl. Sur la première ligne, la seconde probabilité de traduction est $P(\textit{interesantes} \mid \textit{intéressants}) = 10/(10+8+1+1) = 0{,}50$. Sur la troisième ligne, le premier poids lexical est $L(\textit{intéressants} \mid \textit{quinquenal estadístico})$ = meilleure traduction pour *quinquenal* × meilleure traduction pour *estadístico* = $D(\textit{quinquenal} \mid \textit{intéressants}) \times D(\textit{estadístico} \mid \textit{intéressants}) = 0{,}08$. Le détail des effectifs des mots en corpus n'est pas donné.

D'abord, comme nous ne disposons au départ d'aucun alignement mot-à-mot, nous définissons une distribution de probabilité D fondée sur les fréquences des mots dans le corpus :

$$\mathrm{D}(m_j|m_i) = \frac{\mathrm{C}(m_i, m_j)}{\mathrm{C}(m_i)} \qquad \text{avec } i \neq j$$

Ensuite, notre approche basée sur l'échantillonnage ne crée pas de *liens* entre les mots à la façon des méthodes probabilistes. Dans l'exemple de l'aligement d'hapax espagnol-français *desinflaría … acrimonias* ↔ *essouflerait … acrimonies*, nous pourrions nous attendre à ce que le mot espagnol *desinflaría* soit lié au mot français *essouflerait*, et *acrimonias* à *acrimonies*. Notre méthode ne permet pas cela ; à la place, la séquence *desinflaría … acrimonias* est considérée comme traduction de la séquence *essouflerait … acrimonies* en entier. Par conséquent, là où Koehn et coll. (2003) calculaient la *moyenne* des probabilités des mots reliés, nous calculons le *maximum* des probabilités de *tous* les liens possibles, soit de tous les mots source vers tous les mots cible.

Formellement, au sein d'un alignement, nous recherchons la meilleure traduction possible d'un mot $m_i$ issu d'une séquence $s_i$ (dans la langue $i$) parmi tous les mots de l'autre langue, selon la distribution D, et conservons sa probabilité. Le poids lexical d'un alignement pour la langue $i$ est le produit de toutes les probabilités conservées, après détermination de la meilleure traduction pour chaque mot de $s_i$ :

$$\mathrm{L}(s_j|s_i) = \prod_{m_i \in s_i} \max_{m_j \in s_j} \mathrm{D}(m_j|m_i) \qquad \text{avec } i \neq j$$

Cela correspond en quelque sorte à la probabilité que tous les mots aient une bonne traduction selon la distribution D. Le tableau 7 en donne un exemple.

### 4.3.3 *Implémentation*

Nous terminons ce chapitre en décrivant brièvement l'implémentation de notre méthode. La simplicité a été au cœur de la conception d'Anymalign. Simplicité d'utilisation d'abord. Anymalign a moins les



### 5.2.3 *En fonction de la quantité de données en entrée*

Dans une dernière expérience, nous étudions le comportement des aligneurs en fonction de la taille du corpus d'entraînement qui leur est fourni en entrée. Il est bien connu dans le domaine de la traduction automatique probabiliste que l'ajout de données en entrée constitue le meilleur moyen d'améliorer les scores [Bleu] *(« More data is better data »)*. Nous avons pu confirmer cet adage au cours de nos expériences : pour un jeu de test fixé, les scores [Bleu] croissent typiquement de façon logarithmique lorsque la quantité de données en entrée augmente. Pour analyser simplement le comportement des aligneurs, nous proposons plutôt d'évaluer ici la qualité des lexiques produits sur des petites tailles de corpus.

La figure 23 page suivante présente la F-mesure obtenue par les trois tables de traductions en fonction de la taille du corpus d'entrée espagnol-français. MGIZA++ a été exécuté avec trois itérations, BerkeleyAligner avec deux, et Anymalign pendant deux heures, ce qui correspond plus ou moins aux meilleurs résultats obtenus par les deux premiers outils dans nos expériences précédentes. Étant donné qu'Anymalign tire parti des termes de basse fréquence, en les extrayant dans des *petits sous-corpus* de surcroît, nous pourrions nous attendre à ce qu'il produise de meilleurs résultats sur les petites tailles de corpus d'entrée. Ce n'est pas le cas : ses résultats sont en fait inférieurs pour toutes les tailles de corpus de moins de 5 000 énoncés environ. Il reprend l'avantage sur les « grands » corpus d'entrée. Cela traduit en fait la sensibilité d'Anymalign vis-à-vis de la longueur des énoncés du corpus d'entrée : plus les énoncés sont longs, plus la taille des sous-corpus doit l'être également pour que les alignements puissent être extraits, car c'est de la variété des sous-corpus tirés que notre méthode tire son potentiel. À cela s'ajoute le fait que notre méthode ignore volontairement la position des mots, ce qui lui assure une portabilité sans faille d'un couple de langues à un autre ; c'est dans cette expérience qu'elle en paye le prix. Les énoncés du corpus Europarl utilisés ici étant relativement longs, plus de 30 mots en moyenne, elle ne peut en fait

dans le tableau 9. Les résultats de la comparaison sont présentés dans la figure 21 page 102. Les résultats d'Anymalign sont désormais nettement supérieurs à ceux des deux autres aligneurs : sa F-mesure converge très rapidement vers une valeur de 58 %, contre un maximum de 51 % pour MGIZA++ et 46 % pour BerkeleyAligner. Même les scores obtenus avec la version d'Anymalign sans optimisation de l'échantillonnage leur sont toujours supérieurs ; ceux-ci convergent néanmoins beaucoup plus lentement. Si le premier point de la courbe de BerkeleyAligner a l'air étrangement placé, c'est parce qu'il produit de nombreux liens erronés lorsqu'exécuté pour une unique itération, ce qui résulte en une taille bien supérieure de la table de traductions construite par Moses, d'où un surplus de temps de traitement. Ce résultat justifie le choix des concepteurs de l'outil de proposer deux itérations par défaut.

Étrangement, les scores de MGIZA++ et BerkeleyAligner ne semblent pas s'améliorer avec le nombre d'itérations ; ils décroissent même légèrement. La figure 22 page précédente, donnant le détail des rappels et précisions utilisés dans le calcul des F-mesures de la figure 21, permet d'en saisir la raison : il s'agit principalement d'une perte de précision, le rappel étant beaucoup plus stable. De la même façon, la précision d'Anymalign décroît constamment, mais son rappel augmente plus que sa précision ne diminue. Ses résultats convergent toujours très rapidement.

Au total, et en conformité avec l'ensemble des expériences que nous avons pu mener en sus des deux présentées ici, Anymalign est un peu en retrait sur les tâches de traduction automatique ayant recours à Moses, MGIZA++ et BerkeleyAligner produisant des résultats comparables. Par contre, les lexiques qu'il produit semblent de bien meilleure qualité. Nous examinerons l'origine de ces différences dans la section 5.3. Dans tous les cas, Anymalign est beaucoup plus rapide : il est le seul à pouvoir fournir des résultats utilisables quasiment instantanément. L'optimisation de l'échantillonnage définie au chapitre précédent permet en outre une convergence très nette de ses résultats.



caractéristiques d'un programme que d'un script : il est constitué d'un unique fichier texte écrit en Python, à la fois code source (n'ayant recours qu'à la bibliothèque standard) et exécutable. En ce qui concerne l'installation, il est donc directement utilisable sur n'importe quel système disposant d'un interpréteur Python, soit la (quasi-)totalité des systèmes d'exploitation. Nombre de systèmes sont d'ailleurs directement livrés avec un tel interpréteur de nos jours, comme c'est le cas avec la majorité des distributions Linux et Mac OS X. Le coût de déploiement est donc quasi-nul. L'interface est dans la plus pure tradition des outils Unix en ligne de commande : pas d'interface graphique conviviale, mais un simple appel au programme avec ses éventuels paramètres au moyen d'un terminal. Ses entrées et sorties sont de simples fichiers texte. Anymalign dispose d'un certain nombre d'arguments de contrôle, mais ils n'ont aucune incidence sur la qualité des alignements obtenus. Ce sur quoi ils agissent ne relève que du pratique : quantité de mémoire utilisée, emplacement des fichiers temporaires (au nombre de deux), filtrage des sorties, format des fichier en sortie, minuteur, etc. À l'exécution, Anymalign est mono-processus et n'a pas recours aux *threads*. La façon dont le parallélisme est géré est probablement une des plus pures qui soient : on lance la *même* ligne de commande sur divers processeurs ou machines. Du fait de l'échantillonnage, les sorties de chacun sont différentes. Ces sorties peuvent ensuite être fusionnées en appelant une ultime fois le programme avec l'option appropriée. Faire tourner Anymalign une heure sur dix machines aboutira aux même résultats qu'en le faisant tourner dix heures sur une seule machine.

Simplicité de développement ensuite. Python est un langage de haut niveau, faisant sans doute partie des langages de programmation les plus confortables. En contrepartie, Anymalign est certainement plus lent à l'exécution que s'il avait été écrit dans un langage de bas niveau tel que C. Mais à l'échelle humaine, cela ne sera pas flagrant, pour la simple raison que la méthode ne présente aucune section calculatoirement critique. Le gros du travail consiste à indexer les mots en fonction des énoncés où ils apparaissent, ce dont les structures de données de base de Python, et plus particulièrement ses tables de hachage, qui

constituent une des bases du langage, s'accommodent très bien. La méthode est intrinsèquement rapide : l'influence du langage servant à l'implémenter, qu'il soit de haut ou bas niveau, est insignifiante. Avec Python, nous privilégions portabilité, confort de programmation, simplicité de déploiement et d'utilisation. Et comme nous le verrons dans le chapitre suivant, Anymalign est de toute façon plus rapide que les outils d'alignement de référence, écrits dans des langages de plus bas niveau.

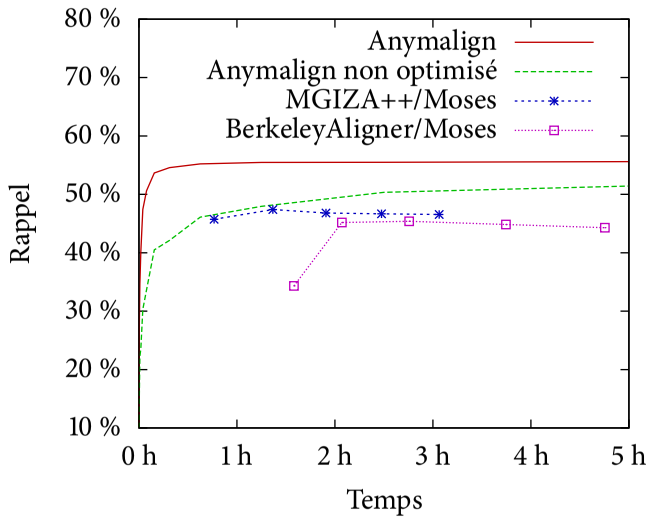
La première version du programme rendue publique s'appelait Malign, pour *Multilingual Aligner*. Nous avons profité d'une récriture complète pour lui donner un nom qui le rendrait plus visible par les moteurs de recherche, Malign étant trop commun en anglais. Anymalign — *any* pour *any-time*, *any-language*, *any-number-of-languages* (à suivre), *any-number-of-processes*, *any-user*, etc. — est un néologisme sur la Toile. Notre implémentation est actuellement disponible en ligne à l'adresse `http://users.info.unicaen.fr/~alardill/anymalign/`, sous les termes de la licence GPL. Entre juin 2009 et juin 2010, elle a été téléchargée une bonne centaine de fois. Nous avons également proposé une version minimale, nommée *Minimalign*, qui constitue un aligneur complet en 100 lignes de code. Elle est librement disponible à la même adresse. Nous la listons à l'annexe B page 143.

### RÉSUMÉ

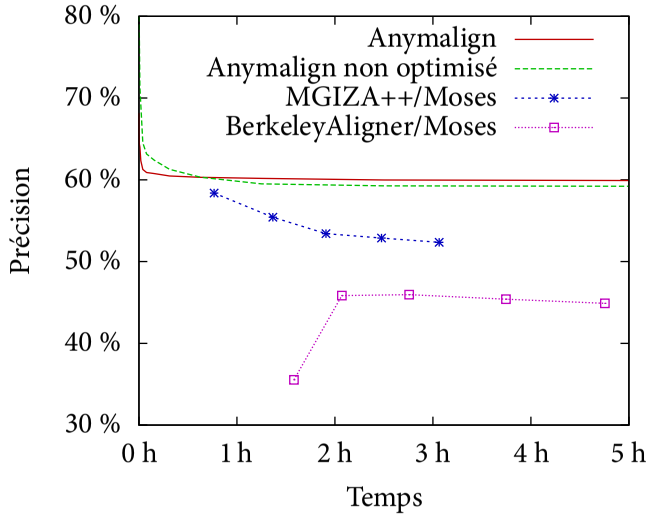
Anymalign, c'est :

1. construire un sous-corpus par échantillonnage. La taille de ce sous-corpus est déterminée par une distribution de probabilité qui privilégie les petites tailles ;
2. extraire de ce sous-corpus toutes les séquences de mots de même distribution et leurs contextes ;
3. répéter les étapes 1 et 2 jusqu'à ce que certaines conditions soient remplies, par exemple un certain temps écoulé, un certain degré de couverture du corpus d'entrée atteint, ou tout simplement une interruption de l'utilisateur ;






(a) Rappel




(b) Précision

FIGURE 22 – Détail des rappels et précisions des F-mesures de la figure 21 page précédente.

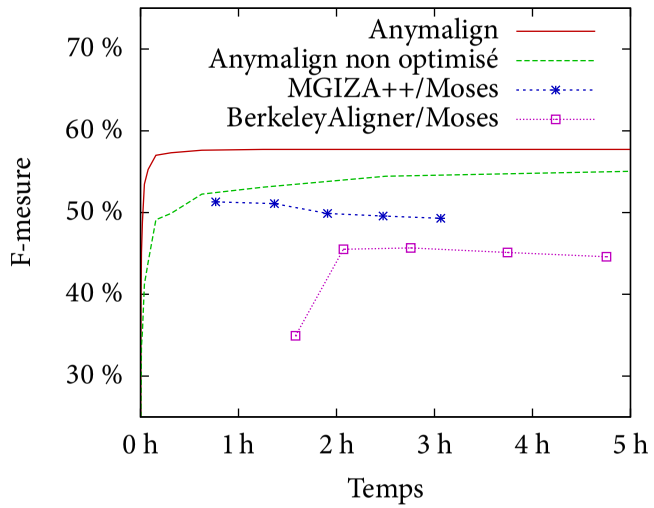



FIGURE 21 – Comportement des aligneurs sur une tâche d'induction de lexique bilingue. Anymalign fournit de bien meilleurs résultats que MGIZA++ et BerkeleyAligner, beaucoup plus rapidement. Rappelons qu'il n'est pas question de poids lexicaux dans cette évaluation : seules les probabilités de traduction sont prises en compte.

de calcul, ce qui s'exprime par le fait que la courbe correspondant aux résultats sans poids lexicaux se situe sous la courbe principale pendant la première demi-heure. Quant à la courbe correspondant à Anymalign sans optimisation d'échantillonnage, elle est loin derrière toutes les autres, et ne converge pas aussi nettement : l'optimisation que nous avons définie au chapitre précédent est clairement bénéfique.

### 5.2.2 *En induction de lexiques bilingues*

Dans la deuxième expérience, nous comparons les tables de traductions avec le lexique bilingue de référence espagnol-français mentionné



4. récolter tous les alignements obtenus à l'étape 2, compter le nombre global de fois qu'ils ont été obtenus, et leur attribuer des scores afin d'en faire une table de traductions.

Nous voilà prêts à nous mesurer à l'état de l'art.

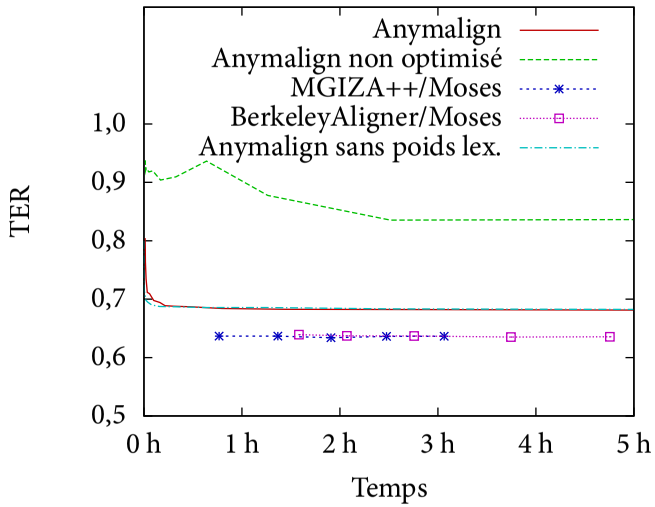



FIGURE 20 – Comportement des aligneurs sur une tâche de traduction automatique. Rappelons que les scores TER les plus faibles sont les meilleurs. MGIZA++ et BerkeleyAligner produisent des résultats comparables. Anymalign est très légèrement moins bon, et l'ajout des poids lexicaux ne l'aide que très peu. Sans optimisation de l'échantillonnage, ses scores sont mauvais. Bien que cela ne soit pas la priorité pour un aligneur, Anymalign est le plus rapide des trois, et de loin.

le calcul des poids lexicaux. Les meilleurs résultats, avec un score de 0,64, sont obtenus à partir des tables de traductions de MGIZA++ et BerkeleyAligner, le premier étant un peu plus rapide. Leurs scores sont relativement constants quel que soit le nombre d'itérations.

Les scores d'Anymalign, version « complète », convergent très rapidement vers une valeur de 0,68, ce qui est inférieur aux scores des deux autres aligneurs. Les poids lexicaux n'améliorent ici que très légèrement les scores d'Anymalign (de 0,005 environ) ; leur impact varie en fait selon le couple de langues, mais ils ont rarement amélioré un score TER de plus de 0,01 dans nos expériences. Ils nécessitent plus de temps

utilise 2 giga-octets pour traiter notre corpus d'entraînement raccourci. Pour le même traitement, MGIZA++ requiert environ 350 méga-octets et Anymalign 200 méga-octets. Les expériences réalisées sont asymétriques : nous évaluons à chaque fois le couple espagnol → français. Nous ne présentons des résultats que pour ce seul couple, mais nous avons aussi effectué de nombreuses expériences sur des corpus de types variés et sur de nombreux couples de langues, incluant 11 langues européennes avec Europarl (Koehn, 2005), 12 langues avec la Bible (Resnik et coll., 1999) et 6 langues avec le BTEC (Takezawa et coll., 2002). Nous insistons sur le fait que les résultats présentés ici sont représentatifs de l'ensemble des résultats obtenus au cours des expériences que nous avons menées. Des résultats complémentaires sont présentés à l'annexe C page 147. D'une façon générale, Anymalign semble moins sensible à la distance entre les deux langues utilisées que les deux autres outils. Enfin, les différences de scores observées dans ces expériences étant relativement grandes, nous n'avons pas jugé opportun d'effectuer de tests de significativité.

### 5.2.1 *En traduction automatique*

Dans un premier temps, nous étudions le comportement des aligneurs en fonction du temps. MGIZA++ et BerkeleyAligner sont exécutés en faisant varier le nombre d'itérations de leurs modèles par défaut, soit de 1 à 5 itérations, et leurs temps d'exécution sont mesurés. Anymalign pouvant être arrêté à tout moment, nous lui faisons répéter la même expérience pour des durées variées, allant de la seconde au temps d'exécution des deux autres aligneurs. 500 couples d'énoncés du corpus Europarl français-anglais sont utilisés pour le développement et 500 autres pour le décodage.

La figure 20 page suivante présente les scores TER obtenus par le système de traduction automatique Moses avec les tables de traductions produites par les trois aligneurs. Anymalign est décliné en trois versions afin de mesurer la contribution des deux optimisations que nous avons proposées, à savoir l'optimisation de l'échantillonnage et

# 5

# ÉVALUATION

Ce chapitre présente des expériences détaillées d'évaluation de notre méthode d'alignement. Nous confrontons les alignements obtenus par notre outil, Anymalign, avec ceux obtenus par l'état de l'art en alignement sous-phrastique bilingue. Nous définissons à cette fin deux protocoles d'évaluation. Nous analysons enfin les résultats de cette évaluation afin de déterminer les avantages et les points faibles de notre méthode.

SOMMAIRE

## 5.1 DESCRIPTION DE L'ÉVALUATION

Les évaluations présentées dans ce chapitre sont toutes *bilingues*, donc nécessairement réductrices vis-à-vis d'Anymalign, qui, nous le montrerons enfin au chapitre suivant, est capable de multilinguisme véritable. En fait, une des raisons pour lesquelles nous avons choisi de ne pas aborder le multilinguisme plus tôt est qu'il n'existe à notre connaissance ni aligneur ni protocole d'évaluation multilingue. Le présent chapitre est donc le dernier placé dans le cadre du bilinguisme.

### 5.1.1 *Outils sur le banc d'essai*

Nous n'évaluons pas Anymalign de façon absolue mais en le comparant à deux autres outils biens connus dans le domaine. Il s'agit de logiciels libres relativement récents. Le tableau 8 page ci-contre récapitule certaines de leurs caractéristiques. Bien que les trois aligneurs soient tous capables d'effectuer des traitements en parallèle, pour que la comparaison soit juste, nous ne les exécuterons que sur un seul processeur. Sauf mention contraire, nous nous contenterons d'utiliser tous leurs paramètres par défaut. Il est assurément possible d'obtenir de meilleurs résultats en optimisant certaines options en vue de la réalisation d'une tâche particulière. Mais, comme nous l'avons dit au chapitre 1, rares sont en pratique les utilisateurs qui maîtrisent tous les paramètres de ces outils, et la plupart se contentent généralement des valeurs par défaut, qui donnent typiquement de bons résultats.

ANYMALIGN `http://users.info.unicaen.fr/~alardill/anymalign/`
Nous en avons décrit le fonctionnement dans le chapitre précédent. Il se distingue de ceux des deux autres outils sur plusieurs points :

- la méthode sous-jacente se rapproche du courant de la traduction automatique par l'exemple, les autres se positionnant clairement dans le courant de la traduction probabiliste ;
- il ne crée pas de liens entre les mots source et cible d'un couple d'énoncés, mais produit directement des traductions. Cela constitue



| LANGUES | ORIGINE | ENTRÉES | APR. FILTR. |
|---|---|---|---|
| anglais ↔ finnois | *XDXF* | 30 188 | 7 531 |
| anglais ↔ français | *XDXF* | 104 775 | 20 104 |
| anglais ↔ espagnol | *XDXF* | 27 639 | Non utilisé |
| espagnol ↔ français | Jointure | 70 770 | 13 617 |

TABLEAU 9 – Caractéristiques des lexiques bilingues de référence utilisés pour nos évaluations : origine, nombre total d'entrées (couples source-cible) et nombre d'entrées effectivement utilisées lors de l'évaluation après filtrage par les corpus Europarl bilingues correspondants. Le dictionnaire anglais-espagnol n'a pas été utilisé directement, mais a servi à produire le dictionnaire espagnol-français par jointure sur sa partie anglaise et sur celle du dictionnaire anglais-français.

erronées produites par l'opération de jointure sur la partie anglaise disparaissent lors du filtrage par le corpus parallèle. Au final, ce protocole nous permet d'évaluer les aligneurs même sur des couples de langues peu dotés. Le tableau 9 ci-dessus présente les tailles des dictionnaires utilisés pour ce second protocole d'évaluation. Nous avons principalement eu recours à des dictionnaires issus du site *XDXF* [2] *(XML Dictionary Exchange Format)*.

## 5.2 RÉSULTATS DES EXPÉRIENCES

Les expériences présentées ci-après, et jusqu'à la fin du présent chapitre, sont réalisées sur un échantillon de 100 000 énoncés de notre corpus Europarl espagnol-français. Nous nous limitons à ce sous-ensemble principalement pour accélérer les traitements, mais également pour des raisons de mémoire : la mémoire vive de la machine sur laquelle ces expériences sont réalisées est rapidement saturée par BerkeleyAligner, qui

2 `http://xdxf.sourceforge.net/`

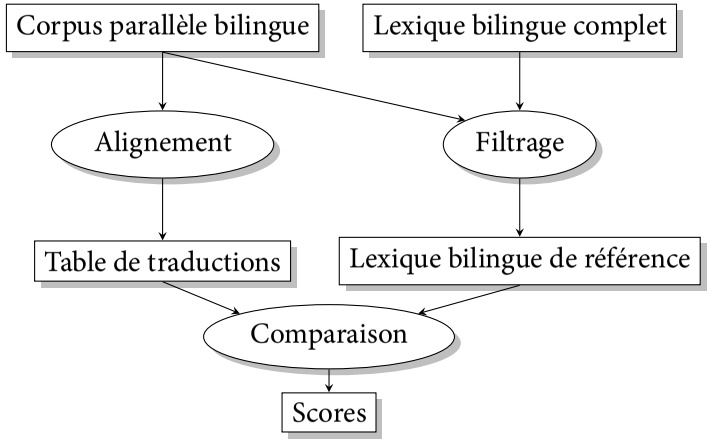



FIGURE 19 – Vue d'ensemble du protocole d'évaluation par comparaison de lexiques bilingues.

bilingue aisément trouvable pour bon nombre de couples de langues, l'anglais étant assurément la langue la mieux dotée.

Il est important de noter que la qualité de ce lexique bilingue ne revêt pas d'importance particulière. En effet, il est systématiquement filtré par le corpus bilingue d'entraînement, ce qui revient en quelque sorte à prendre l'intersection de ces deux ressources. La qualité du lexique bilingue de référence résultant est ainsi garantie, quelle que soit l'origine du lexique bilingue initial. Par conséquent, nous pouvons compiler des lexiques à partir de diverses sources, car les entrées erronées ou hors domaine sont naturellement filtrées. Ce dernier point présente un avantage non négligeable : il permet de constituer des lexiques bilingues entre de nouveaux couples de langues par transitivité, comme l'ont expérimenté par exemple Tanaka et Umemura (1994), Bond et coll. (2001) ou Nerima et Wehrli (2008), mais sans que le moindre post-traitement ne soit nécessaire. Par exemple, le lexique bilingue de référence utilisé ci-après (section 5.2.2) a été constitué en utilisant l'anglais comme langue « pivot ». Les nombreuses entrées



| ALIGNEUR | SOURCES | PARALLÉLISME | OPTIONS |
|---|---|---|---|
| Anymalign v. 2.3 | Python ≈ 1 500 lignes | Multi-processus sur 1 ou plusieurs machines | 17 |
| MGIZA++ v. 0.6.3 | C++ ≈ 30 000 lignes | Multi-*threads* sur 1 machine | 59 |
| BerkeleyAligner v. 2.1 | Java ≈ 50 000 lignes | Multi-*threads* sur 1 machine | 81 |

TABLEAU 8 – Les aligneurs sur le banc d'essai. Le nombre de lignes de code source indiqué est grossier et inclut les commentaires. Il existe une version de MGIZA++, nommée PGIZA++, dédiée aux *clusters* de calcul, mais elle ne bénéficie pas du même degré d'activité que MGIZA++.

tue un manque pour certaines applications pour lesquelles la position des mots joue un rôle important, en particulier en traduction automatique probabiliste. Pour toutes les autres, cela constitue un gain de temps ;

– il peut être arrêté à tout moment au cours de son exécution. Le temps d'exécution n'influe pas sur la *qualité* des alignements produits, mais sur leur *couverture* : plus il est exécuté longtemps, plus le nombre d'alignements en sortie est élevé. Il est donc *anytime* à sa façon.

Du fait de ce dernier point, nous commencerons en pratique par exécuter les deux autres aligneurs, mesurerons leur temps de traitement, et exécuterons Anymalign pendant la même durée. Afin de mesurer l'impact de la stratégie de sélection des tailles de sous-corpus, Anymalign sera exécuté deux fois dans les expériences à venir : une fois avec la distribution privilégiant les petits sous-corpus que nous avons définie dans le chapitre précédent, et une fois avec une distribution uniforme.

MGIZA++ `http://www.cs.cmu.edu/~qing/`

Proposé par Gao et Vogel (2008), MGIZA++ est une amélioration du programme GIZA++ (Och et Ney, 2003), lui-même une amélioration

du programme GIZA (Al-Onaizan et coll., 1999), lequel implémente les indétrônables modèles IBM (Brown et coll., 1993). Bien qu'apparus il y a bientôt vingt ans, les modèles IBM servent toujours de pilier à la traduction automatique probabiliste. Comme nous l'avons déjà vu au chapitre 1, de nombreuses améliorations ont été proposées au cours des années passées, mais rares sont celles qui ont été adoptées définitivement par la communauté, la principale étant le HMM de Vogel et coll. (1996), intégré à GIZA++. MGIZA++ introduit la possibilité de paralléliser les traitements et supprime quelques bogues par rapport à GIZA++. Par défaut, il enchaîne 5 itérations du modèle IBM1, 5 du HMM, 5 du modèle IBM3 et 5 du modèle IBM4. Le nombre d'itérations de ces modèles est le seul paramètre que nous ferons varier, de 1 à 5 : plus ce nombre est élevé, meilleurs sont censés être les résultats. Nous fixerons toujours le même nombre d'itérations pour chaque modèle. Ces modèles sont asymétriques, et doivent donc être exécutés deux fois, de la source vers la cible et de la cible vers la source, puis leurs résultats sont symétrisés, afin de fournir les meilleurs résultats possibles. Nous déléguons cette étape aux scripts du jeu d'outils de Moses (Koehn et coll., 2007). Une telle symétrisation n'est pas nécessaire avec Anymalign, dont les alignements sont naturellement symétriques.

BERKELEYALIGNER `http://nlp.cs.berkeley.edu/Main.html#WordAligner`
Introduit par Liang et coll. (2006), BerkeleyAligner n'entraîne que des modèles statistiques simples, typiquement les modèles IBM1 et IBM2 et le HMM, mais *conjointement* de la source vers la cible et de la cible vers la source, afin de produire de meilleurs résultats qu'en les entraînant de façon asymétrique comme le fait MGIZA++. Les alignements produits sont ainsi directement symétriques. L'outil est capable de prendre en compte la syntaxe des phrases si elle lui est fournie, et il en existe une version permettant la supervision de l'alignement. Pour que la comparaison avec les deux autres outils soit juste, nous ne ferons pas usage de ces fonctionnalités avancées. Par défaut, l'outil enchaîne 2 itérations du modèle IBM1 et 2 du HMM. Comme pour MIGZA++, nous ferons varier ce nombre de 1 à 5.



pondent à une entrée du lexique bilingue de référence, divisée par le nombre A d'entrées source distinctes parmi ces alignements :

$$\mathrm{P} = \frac{\mathrm{S}}{\mathrm{A}}$$

RAPPEL : même chose que pour la précision, mais nous divisons par le nombre D d'entrées source distinctes dans le lexique bilingue de référence :

$$\mathrm{R} = \frac{\mathrm{S}}{\mathrm{D}}$$

F-MESURE : moyenne harmonique de la précision et du rappel :

$$\frac{1}{\mathrm{H}} = \frac{1}{2}\left(\frac{1}{\mathrm{P}} + \frac{1}{\mathrm{R}}\right)$$

$$\Rightarrow \mathrm{H} = \frac{1}{\frac{1}{2}\left(\frac{1}{\mathrm{P}} + \frac{1}{\mathrm{R}}\right)} = \frac{2}{\frac{\mathrm{A}}{\mathrm{S}} + \frac{\mathrm{D}}{\mathrm{S}}} = \frac{2\mathrm{S}}{\mathrm{A} + \mathrm{D}}$$

La figure 19 page suivante donne une vue d'ensemble des données et traitements impliqués dans cette évaluation. Nous avons proposé ce protocole afin d'effectuer une évaluation rapide et peu coûteuse des aligneurs. Jusqu'à il y a peu, cette évaluation était couramment menée avec AER *(Alignment Error Rate)* (Och et Ney, 2000). D'autres propositions ont été également faites (p. ex. Ahrenberg et coll., 2000), mais on leur préfère maintenant l'évaluation par une tâche subséquente, typiquement la traduction automatique, telle que celle que nous avons décrite dans la section précédente, car il a été montré qu'une amélioration en AER ne menait pas nécessairement à une amélioration des résultats sur cette tâche (Vilar et coll., 2006). D'autre part, AER nécessite des alignements de référence, qui sont relativement rares, et toujours au format « lien » tels qu'en sortie de BerkeleyAligner ou MGIZA++, ce à quoi Anymalign coupe court. Le protocole que nous proposons ne nécessite qu'une unique ressource, qui est un lexique

Blanchon et Boitet, 2007). Parmi les autres critères existants, dont les plus éprouvés ont été passés en revue par Callison-Burch et coll. (2007), notre choix s'est porté sur TER *(Translation Edit Rate)* (Snover et coll., 2006), principalement pour sa clarté et son indépendance vis-à-vis des langues. TER reflète le nombre d'opérations nécessaires à un être humain pour transformer une phrase produite par un système de traduction automatique en une des phrases de référence correspondantes issues de la partie cible du corpus de test. Les scores sont typiquement compris entre zéro, indiquant l'égalité des traductions, et un. Ils peuvent néanmoins être supérieurs si les références sont plus courtes que la phrase candidate. Au total, notre processus de traduction utilise donc deux critères : BLEU pour le développement et TER pour l'évaluation finale. Nous conservons le premier principalement pour des raisons pratiques, et parce qu'il est généralement admis qu'il reste adapté pour mesurer les améliorations d'un même système (Callison-Burch et coll., 2006), ce à quoi est dédiée la phase de développement où nous le faisons intervenir. He et Way (2009) ont en outre montré que les meilleurs scores finals selon un critère donné n'étaient pas nécessairement obtenus en effectuant la phase de développement à l'aide ce même critère.

### 5.1.3 *Protocole 2 : induction de lexiques bilingues*

Notre second protocole consiste à comparer les tables de traductions avec un lexique bilingue de référence, en pondérant les alignements par leurs probabilités de traduction. Dans un premier temps, nous filtrons le lexique de façon à ce que les références ne contiennent que des entrées qui peuvent être réellement produites par les aligneurs à partir du corpus d'entraînement. En pratique, une entrée n'est conservée que s'il s'agit d'une sous-séquence d'un couple d'énoncés du corpus d'entraînement correspondant. Nous calculons ensuite les trois scores suivants :

PRÉCISION : somme S des probabilités de traduction source-cible associées aux alignements d'une table de traductions qui corres-



| *Anymalign* | *MGIZA++* | *BerkeleyAligner* |
|---|---|---|
| ENTRÉE : CORPUS PARALLÈLE BILINGUE | | |
| Alignement (`anymalign.py`) | Entraînement source-cible (`mkcls` + `mgizapp`) | Entraînement conjoint (`berkeleyaligner.jar`) |
| | Entraînement cible-source (`mkcls` + `mgizapp`) | |
| | Symétrisation (Moses : `symal`) | |
| | Extraction des alignements (Moses : `extract`) | |
| | Attribution de scores (Moses : `score`) | |
| SORTIE : TABLE DE TRADUCTIONS | | |

FIGURE 18 – Les trois programmes d'alignement dans leurs chaînes de traitement respectives. Anymalign est auto-suffisant. Nous n'utilisons Moses avec BerkeleyAligner que pour extraire les alignements et leur attribuer des scores. MGIZA++ doit de plus être exécuté une seconde fois et ses résultats doivent être symétrisés.

La figure 18 ci-dessus récapitule les programmes nécessaires à la création d'une table de traductions complète à partir d'un corpus parallèle pour chacun des trois aligneurs. Par la suite, les temps d'exécution seront mesurés en incluant toutes les étapes depuis l'introduction du corpus parallèle dans le système jusqu'à l'obtention de la table de traductions en sortie.

### 5.1.2 *Protocole 1 : traduction automatique*

Notre premier protocole d'évaluation consiste à utiliser les tables de traductions obtenues à partir des aligneurs comme principale source de connaissance dans le cadre d'une tâche de traduction automatique empirique, et d'évaluer la « qualité » des traductions obtenues. Parmi

les systèmes de traduction librement disponibles sur la Toile, notre choix s'est porté sur Moses, que nous utilisons déjà comme module de post-traitement pour MGIZA++ et BerkeleyAligner. Son moteur de traduction est fondé sur des modèles probabilistes par segments, plus précisément des n-grammes de mots, dont sont justement constituées nos tables de traductions. Parmi les autres choix qui eussent été possibles, citons principalement Joshua (Li et coll., 2009), moteur de traduction fondé sur des modèles probabilistes hiérarchiques (Chiang, 2007) ; Cunei (Phillips et Brown, 2009), moteur de traduction hybride prenant ses racines à la fois dans la traduction probabiliste et dans la traduction par l'exemple ; et Homomorphism (Lepage et Denoual, 2005), moteur de traduction purement par l'exemple. Ce dernier n'est pas adapté à notre tâche car il ne nécessite aucun alignement sous-phrastique. Cunei nécessite des alignements sous forme de liens, tels que produits par BerkeleyAligner ou par la symétrisation des alignements de MGIZA++, et n'est donc pas utilisable avec Anymalign en l'état. Quant à Joshua, il pourrait être compatible avec Anymalign puisque basé sur des modèles discontinus, mais il nécessite des correspondances entre les discontinuités source et cible des alignements, sous forme de variables, ce dont nos alignements font actuellement abstraction.

L'avantage de Moses est qu'il nous permet d'intégrer nos alignements *en l'état* dans une tâche de traduction automatique. Étant donné qu'il a recours à des n-grammes de mots, nous filtrons nos alignements de façon à ce que ceux contenant des discontinuités soient écartés. Le moteur est utilisé avec ses paramètres par défaut ; nous remplaçons simplement sa table de traductions par l'une des trois tables produites respectivement par nos trois aligneurs. De façon classique, notre tâche de traduction automatique probabiliste comprend trois étapes :

ENTRAÎNEMENT : extraire d'un corpus parallèle, appelé corpus d'entraînement, les données nécessaires à la traduction de phrases dont on s'attend à ce qu'elles soient du même type que celles de ce corpus d'entraînement. Concrètement, il s'agit principalement de la création de nos tables de traductions par nos trois aligneurs.

DÉVELOPPEMENT : optimisation des poids associés à chacun des paramètres impliqués dans le processus de traduction dans le but de produire les meilleures traductions possibles. Les paramètres incluent les traits associés à nos alignements : probabilités de traduction et poids lexicaux. Cette étape est réalisée automatiquement par MERT *(Minimum Error Rate Training)* (Och, 2003). Pratiquement, elle nécessite un petit corpus parallèle, dit de développement, dont la partie source est traduite par le moteur et dont les traductions obtenues sont comparées à la partie cible, qui fait office de référence. Plusieurs références peuvent éventuellement être présentes. Ce processus est itératif et continue jusqu'à ce que le gain selon le critère d'évaluation des phrases traduites, qui par défaut dans Moses est BLEU (Papineni et coll., 2002), ne soit plus significatif. Les paramètres optimisés sont conservés pour la dernière étape.

DÉCODAGE : la traduction à proprement parler. Comme pour la phase de développement, un petit corpus parallèle du même domaine que le corpus d'entraînement, dit de test, est nécessaire pour réaliser une évaluation objective. La partie source est fournie en entrée du système de traduction, constitué du « moteur », de la table de traductions et des paramètres optimisés. Les traductions obtenues sont comparées à la partie cible et évaluées selon un critère particulier.

Bien que BLEU demeure à ce jour le critère d'évaluation le plus usité en traduction automatique, nous décidons de ne pas l'utiliser pour évaluer nos traductions finales car ses limites ont été montrées à plusieurs reprises[1] (p. ex. Blanchon, 2004; Callison-Burch et coll., 2006;

1 Pour en combler certaines lacunes, l'implémentation « officielle » de BLEU, `mteval-v13.pl`, intègre depuis la version 13 un lissage sur les décomptes des n-grammes. Les phrases candidates de moins de quatre mots ne se voient plus systématiquement attribuer un score nul pour la seule raison qu'elles comportent moins de quatre mots.