Université Sorbonne Nouvelle – Paris 3

Master Ingénierie Linguistique

Typologie pour l'alignement multilingue

Encadrants
François Yvon
Isabelle Tellier

Membres du jury
Alexandre Allauzen
Thierry Poibeau

Pierre Godard

Mémoire de recherche de Master 2

Stage effectué au

LIMSI - CNRS
Équipe Traitement du Langage Parlé

avril-septembre 2014

Table des matières

Venons maintenant à la cinquiesme reigle, que doibt observer ung bon traducteur. Laquelle est de si grand' vertu, que sans elle toute composition est lourde, & mal plaisante. Mais qu'est-ce, qu'elle contient ? Rien autre chose, que l'observation des nombres oratoires : c'est asscavoir une liaison, & assemblement des dictions avec telle doulceur, que non seulement l'ame s'en contente, mais aussi les oreilles en sont toutes ravies, & ne se faschent iamais d'une telle harmonie de langage.

Estienne Dolet
La Manière de bien traduire d'une langue en aultre, 1540

Qui peut échapper à ce que dit le mot désir ? Ni le vêtement, ni le silence, ni la nuit, ni les fards, ni même les pensées volontaires ne dissimulent tout à fait la honte des fantasmes qui nous affolent. La femme ou l'homme qui implorerait pitié pour son désir implorerait en vain.

Agustina Izquierdo
L'amour pur, 1993

Introduction

Le Traitement Automatique des Langues (TAL), dont les problématiques s'étendent de la linguistique à l'informatique théorique et qui constitue une branche de la recherche en intelligence artificielle, à évolué résolument depuis un peu plus d'une dizaine d'années vers l'utilisation de méthodes d'apprentissage statistique. Ces techniques s'appuient en effet sur de très grandes quantités de données, que la preste expansion des technologies de l'information et de la communication ont rendues aujourd'hui disponibles.

La traduction automatique, discipline qui naît à la fin des années 1940, en particulier avec l'idée de Warren Weaver d'appliquer au problème de la traduction les concepts de la théorie de l'information de Claude Shannon, a elle-même suivi ce mouvement la ramenant à ses origines statistiques.

Nous nous proposons, dans cette étude, d'étudier différentes approches visant à identifier automatiquement des liens de correspondance entre des textes en relation de traduction réciproque dans différentes langues. Des corpus parallèles de grande dimension, construits par alignement de ces textes au niveau phrastique, sont disponibles ou peuvent être construits automatiquement de manière efficace.

De nombreux traitements – la traduction automatique bien sûr, mais aussi l'extraction automatique de lexiques bilingues, la désambiguïsation sémantique, le transfert d'annotations – requièrent cependant un alignement de nature plus fine, soit au niveau des mots eux-mêmes, soit au niveau de groupes formés par ces mots.

Nous tenterons de dresser, à travers l'étude de divers travaux, une typologie pour l'alignement dans un contexte multilingue, que nous envisagerons sous trois angles principaux : les objectifs visés, la représentation mathématique des objets considérés, les méthodes (modèles et algorithmes). À partir des enseignements tirés, nous chercherons ensuite à proposer des stratégies pour améliorer un alignement bilingue lorsque nous disposons d'alignements supplémentaires avec le même texte.

Nous commençons par une discussion sur la notion même d'alignement (premier chapitre), puis revenons sur le cas bilingue (deuxième chapitre) pour étudier ensuite la problématique multilingue [1] (troisième chapitre). En dernier lieu, nous rendons compte des expériences que nous avons conduites, en particulier autour de l'idée du filtrage des liens les moins sûrs (quatrième chapitre), ce qui conclura ce travail.

1. Nous appelons *multilingues*, ou *multi-parallèles*, uniquement les corpus parallèles faisant intervenir au moins trois langues.

Chapitre 1

L'alignement

Nous abordons dans ce chapitre certaines questions générales relatives à l'alignement de textes en relation de traduction mutuelle, et les problèmes qui se posent dans un cadre computationnel.

1.1 Quels objets ?

La première question à se poser est de savoir pour quel type d'unités nous souhaitons réaliser des alignements, c'est-à-dire quels sont les ensembles de définition des relations matérialisées par des liens d'alignement. En matière de texte, les niveaux naturels sont ceux du paragraphe (ou de toute autre structuration sémantique du texte comme les chapitres ou les sections), de la phrase, du syntagme, ou du mot. Nous nous cantonnerons ici aux niveaux phrastiques et sous-phrastiques, en notant que si la phrase est une unité linguistique relativement bien définie, le sens linguistique des notions de syntagme et de mot peuvent varier. En pratique, on travaillera donc sur des *tokens* plutôt que sur des mots, et sur des *segments* plutôt que sur des syntagmes.

Notons également que si la tâche d'alignement phrase à phrase profite d'une opportune monotonie [1], ce n'est nullement le cas de l'alignement mot à mot ni de l'alignement de segments. Lorsque l'on parle de segments, il faudra par ailleurs préciser s'ils peuvent ou non être discontinus [2] sur la séquence textuelle.

Considérant un corpus comprenant N textes ou ensembles de textes constituant les systèmes linguistiques [3] $S_1, \ldots, S_N$, nous noterons V_n avec $n \in [1 \ldots N]$ l'ensemble des unités considérées dans le texte correspondant au n-ième système S_n, c'est-à-dire son vocabulaire indexé [4].

1.2 Quels liens ?

Une fois définis les ensembles de travail, la seconde question consiste à se demander quels types de relations entre des éléments de ces ensembles nous voulons considérer. La littérature parle d'*alignements* et de *liens* [5]. Mais cela recouvre parfois des réalités

1. La fonction associant l'indice d'une phrase source à l'indice de la phrase cible correspondante est essentiellement croissante, malgré des phénomènes d'inversion.

2. Dans le cas discontinu, le terme *locution* est peut-être alors préférable au mot *segment*.

3. Le terme de *langue* n'est pas suffisamment précis dans la mesure où nous serons amenés à travailler sur des paraphrases ou des variantes d'un texte dans une même langue, c'est-à-dire des traductions *intralangues*.

4. Deux occurrences du même mot dans le texte constituent des unités différentes.

5. Notons la parenté un peu trompeuse de ces mots : le mot *alignement*, qui vient de ligne, évoque l'idée de séquence, de relation d'ordre, quand le mot *lien* fait plutôt penser à une relation binaire – ce

un peu différentes et il faut donc préciser ce que représentent ces liens d'alignement et comment on peut envisager de les représenter.

Pour répondre à la première question, nous nous contenterons, faute de mieux pour l'instant, de dire que nous cherchons à capturer des rapports de traduction mutuelle entre différentes unités textuelles de surface. Or, il semble raisonnable d'affirmer que ce que nous appelons *traduire* implique la mise en correspondance d'unités non surfaciques, aux niveaux syntaxique et sémantique, et il faut garder à l'esprit le caractère très restrictif du seul alignement des formes graphiques.

1.2.1 Relations

Un premier formalisme pour représenter un alignement [6] multilingue consiste à définir une relation n-aire A sur les ensembles $V_1, \ldots, V_N$ (correspondant à des ensembles de mots ou de groupes de mots sur les différents systèmes linguistiques considérés), c'est-à-dire un sous-ensemble du produit cartésien $V_1 \times \cdots \times V_N$.

Dans le cas bilingue, cela revient bien sûr à considérer une relation binaire sur les éléments de V_1 et V_2. Mais on peut, même dans un contexte multilingue, préférer s'intéresser à des relations binaires et définir la relation $\mathcal{R}$ sur $V \times V$ avec $V = \bigcup_{n=1}^{N} V_n$, ou encore une série de relations binaires $\mathcal{R}_{ij}$ sur $V_i \times V_j$, avec $i, j \in [1 \ldots N]$. Il y a un avantage important à cela : les relations binaires peuvent être caractérisées par des propriétés bien définies, et en particulier la *réflexivité*, la *symétrie*, et la *transitivité* [7].

Le fait de savoir si une relation définie sur des unités textuelles est réflexive a peu d'importance. Il semble par ailleurs raisonnable d'admettre la symétrie dans le cas général (i.e. si l'unité a est en relation de traduction avec b, alors b est en relation de traduction avec a). En revanche, la transitivité doit être discutée, et particulièrement si l'on considère une relation sur l'union de tous les V_n, plutôt que des relations binaires distinctes par paires de langues.

Ces relations pourront par ailleurs être représentées par un hypercube de dimension n, c'est-à-dire dans le cas particulier de la dimension 2 et à supposer que la relation soit symétrique, une matrice, et en dimension 3, un cube.

1.2.2 Graphes

Un autre formalisme équivalent [8] aux relations binaires, celui des graphes, permet de représenter des liens d'alignement. Un graphe $G = (V, E, \gamma)$ est défini par un ensemble V de sommets, un ensemble E d'arêtes et une fonction d'incidence $\gamma : E \to V \times V$ associant à chaque arête un couple de sommets. Un graphe simple est ainsi entièrement défini par le couple (V, E) et peut être représenté par une relation binaire $\mathcal{R}$ sur $V \times V$. Le point de vue consistant à voir la nature des liens comme spécifique à une paire de langues particulière (série de relations $\mathcal{R}_{ij}$) conduirait ici à considérer des graphes indépendants pour chaque paire de langues. Par ailleurs, si l'on considère des liens non symétriques, il faudra les représenter par des arcs [9], c'est-à-dire des arêtes orientées.

Pour un alignement bilingue, la représentation naturelle prendra la forme d'un graphe biparti, c'est-à-dire un graphe $G = (V, E)$ dont l'ensemble des sommets V admet une

qui n'est pas incompatible, mais différent.

6. On n'emploiera ce terme que pour désigner un ensemble de liens réalisés ou réalisables au niveau phrastique, et non pour désigner un lien particulier entre deux unités.

7. Pour pouvoir parler de réflexivité, de symétrie ou de transitivité, il faut obligatoirement considérer l'union des vocabulaires indexés comme ensembles de définition des relations binaires.

8. Si l'on ne considère que des graphes simples, c'est-à-dire qui n'ont ni liens multiples ni boucles.

9. La terminologie de la théorie des graphes n'est pas complètement fixée, mais le terme d'*arc* semble néanmoins réservé au cas des graphes orientés.

partition en deux sous-ensembles disjoints V_1 et V_2 tels que chaque arête de E ait une extrémité dans V_1 et l'autre dans V_2.

Notons enfin que les hypergraphes (Berge, 1970), qui généralisent les graphes en autorisant les arêtes à relier plus de deux sommets, permettent d'effectuer un parallèle analogue pour les relations n-aires. Il est par ailleurs possible d'encoder simultanément plusieurs alignements binaires dans un hypergraphe biparti (Liu *et al.*, 2013).

La Figure 1.1 synthétise les trois représentations les plus fréquentes des alignements, sous forme de graphe, de matrice, ou de séquence d'étiquettes.

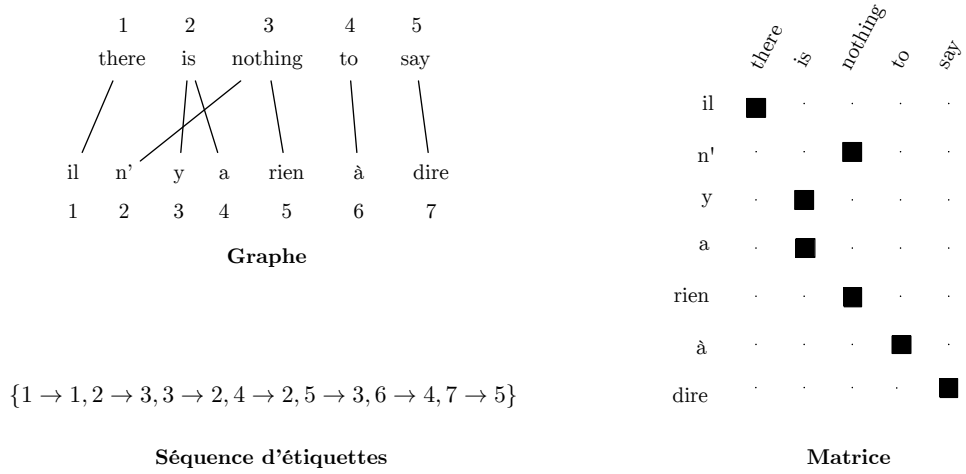


FIGURE 1.1 – Diverses représentations des alignements.

1.3 Quels objectifs ?

Les trois niveaux d'alignement que nous considérons dans ce document (phrase, segment, mot) sont bien sûr interdépendants mais répondent à des objectifs différents. Si un alignement au niveau de la phrase est un préalable à tout traitement sous-phrastique, l'alignement au niveau des mots est généralement réalisé avant l'alignement au niveau des segments, car il sert précisément de base pour l'extraction de ces segments. Nous nous intéressons ici exclusivement à des alignements de niveau sous-phrastique.

1.3.1 Évaluation extrinsèque

L'alignement se présente le plus fréquemment comme une tâche support, et son évaluation relève alors naturellement d'une mesure d'impact sur le score de la tâche principale. Pour la traduction automatique, par exemple, qui repose sur ces alignements de mots et de segments pour l'extraction de tables de traduction, on pourra mesurer un gain de qualité des alignements par un gain du score BLEU (Papineni *et al.*, 2002) ou METEOR (Banerjee et Lavie, 2005).

Ce type d'évaluation pose néanmoins le problème de sa sensibilité à la tâche considérée, et la « qualité » de l'alignement se trouve seulement quantifiée de manière très relative. Lorsque l'on cherche à apprendre des liens de traduction entre des mots pour construire un système de traduction automatique, il peut se trouver qu'il soit plus efficace du point de vue de la performance du système complet d'augmenter le rappel sur ces liens – quitte à aussi augmenter le bruit – en ajoutant une grande quantité de liens peu sûrs et en alignant ainsi des mots d'une manière peu intuitive. En revanche, dans

le cas de l'extraction automatique d'un lexique bilingue, ou si l'on cherche à fournir en temps réel à un lecteur utilisant un dispositif de lecture augmenté des traductions de mots ou de fragments du texte source, la situation est toute différente et la précision de l'alignement devient déterminante.

1.3.2 Évaluation intrinsèque

Afin de mesurer la performance intrinsèque d'un programme qui réalise automatiquement des alignements, il faut disposer d'alignements de référence produits par des annotateurs humains selon des règles bien définies.

Un usage courant introduit par (Och et Ney, 2003) pour l'alignement bilingue est de distinguer les liens *sûrs* des liens *possibles*[10], et de calculer ensuite la précision p et le rappel r selon :

$$r = \frac{|A \cap S|}{|S|} \quad \text{et} \quad p = \frac{|A \cap P|}{|A|}$$

avec A l'ensemble des liens d'alignement extraits, S l'ensemble des liens sûrs, et P l'ensemble des liens possibles, puis de considérer la mesure AER[11] définie par :

$$\mathrm{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

À noter que cette métrique, sévèrement critiquée dans (Fraser et Marcu, 2007), ne satisfait pas à la propriété importante de la F-mesure qui est de pénaliser les déséquilibres entre le rappel et la précision. Il est donc possible de doper un score AER en favorisant la précision, c'est-à-dire par exemple en prédisant très peu de liens. Par ailleurs, bien qu'elle soit utilisée comme mesure intermédiaire de performance en traduction automatique, cette métrique ne présente pas de corrélation substantielle avec la métrique BLEU et son usage dans ce contexte est donc peu informatif.

1.3.3 Évaluation multilingue

Lorsque l'on dispose de corpus parallèles dans au moins trois langues, de nouvelles stratégies peuvent être envisagées, et les moyens de mesurer leurs résultats seront donc à définir.

Dans le cadre d'une approche que nous appellerons **symétrique**[12], toutes les langues disponibles dans le corpus jouent un rôle comparable. L'apprentissage statistique de liens d'alignement, que ces liens soient à proprement parler multilingues ou qu'il s'agisse d'une série d'alignements bilingues, s'attache au moins en principe à traiter en parallèle toutes les langues. En ce cas, l'évaluation posera le problème délicat de la définition d'une métrique puisque nous ne disposons pas, à ce jour, de métrique standard pour l'alignement multilingue symétrique – ceci rendant difficile toute comparaison avec d'autres travaux.

Mais l'ajout d'une ou de plusieurs langues à un corpus bilingue peut également, dans une approche dite par contraste **asymétrique**, être motivé par l'espoir d'améliorer la qualité des liens d'alignement d'une paire de langues particulière. Il sera alors possible d'utiliser les métriques usuelles de l'alignement bilingue.

10. Noter que les liens « sûrs » sont également « possibles ».
11. *Alignment Error Rate*.
12. À ne pas confondre avec la propriété de symétrie des relations binaires évoquées plus haut.

Chapitre 2

Le cas bilingue

La notion d'alignement bilingue à laquelle se réfère la quasi-totalité de la littérature sur le sujet a été introduite par les auteurs de (Brown *et al.*, 1990) puis précisée dans (Brown *et al.*, 1993). Nous rappelons ici les aspects qui touchent à la nature des alignements que ces modèles autorisent.

2.1 Modélisation générative canonique

Historiquement, la notion d'alignement apparaît dans le cadre des premières méthodes probabilistes de traduction automatique à base de **mots** développées dans les laboratoires d'IBM au début des années 90. Nous en redonnons les principaux traits en nous appuyant sur (Brown *et al.*, 1993) ainsi que sur la présentation qui en est faite dans (Allauzen et Yvon, 2011) et dans (Brunning, 2010).

2.1.1 Premières hypothèses

La probabilité $\mathrm{P}(\mathbf{f} \mid \mathbf{e})$ d'observer la séquence de mots *source* $\mathbf{f} = f_1^J = f_1, \ldots, f_J$ conditionnellement à l'observation d'une séquence de mots *cible*[1] $\mathbf{e} = e_1^I = e_1, \ldots, e_I$ est modélisée grâce à la marginalisation d'une variable cachée d'alignement A qui se réalise dans l'espace $\mathcal{A}$:

$$\mathrm{P}(\mathbf{f} \mid \mathbf{e}) = \sum_{A \in \mathcal{A}} \mathrm{P}(\mathbf{f}, A \mid \mathbf{e}) \tag{2.1}$$

Une forme assez générale (voir section 1.2.1) pour l'ensemble $\mathcal{A}$ est fournie par l'ensemble des matrices binaires $A = (a_{ij})$, avec $a_{ij} = 1$ si le mot f_j est aligné avec le mot e_i et $a_{ij} = 0$ sinon.

Pour des raisons computationnelles, cependant, les auteurs de (Brown *et al.*, 1993) restreignent l'espace $\mathcal{A}$ à l'ensemble des vecteurs aléatoires $\mathbf{A} = A_1, \ldots, A_J$ où chaque variable aléatoire A_j pour $j \in [1, \ldots, J]$ se réalise dans l'espace $\{0, 1, \ldots, I\}$. En notant $\mathbf{a} = a_1^J = a_1, \ldots, a_J$ la réalisation de $\mathbf{A}$, a_j représente l'indice du mot de la phrase cible $\mathbf{e}$ avec lequel f_j est aligné[2], et autrement dit que le mot f_j est aligné avec le mot e_{a_j}. Ceci revient à envisager l'alignement comme un étiquetage de séquence.

1. Remarquons la relative confusion qui règne dans les usages terminologiques, expliquée par l'utilisation de modèles à canal bruité suivant $\mathrm{P}(\mathbf{e} \mid \mathbf{f}) \propto \mathrm{P}(\mathbf{f} \mid \mathbf{e})\,\mathrm{P}(\mathbf{e})$. Nous choisissons d'appeler *source* la langue à traduire dans le problème initial.

2. L'indice 0 correspond à l'introduction du concept de mot `null` dans la phrase cible afin de traiter le cas où les mots de la phrase source ne sont alignés avec aucun mot dans la phrase cible.

2.1.2 Représentation des alignements

Dans ce cadre théorique, il est uniquement possible de réaliser des alignements de type « 1-1 » ou « 1-n » mais non de type « m-n » (voir la Figure 2.1).

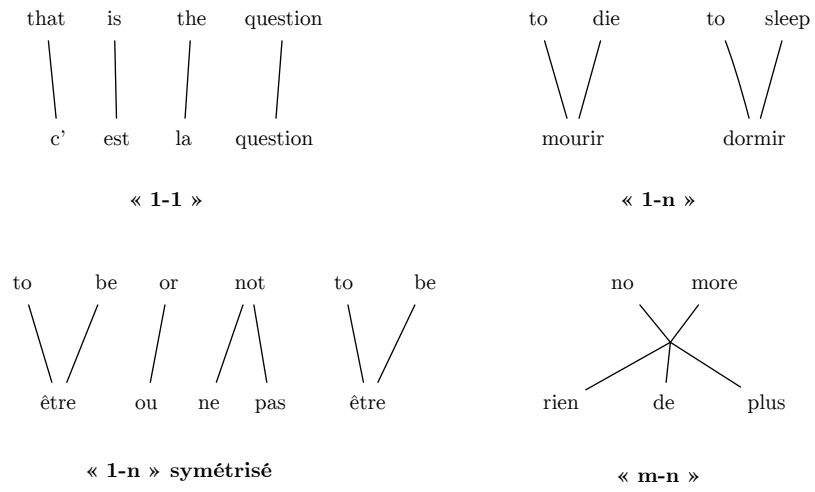


FIGURE 2.1 – Types d'alignements pour les modèles IBM. Les alignements de type « m-n » ne sont pas pris en charge.

Il s'agit donc d'un type d'alignement souvent qualifié d'*asymétrique*. Ici, l'asymétrie ne signifie toutefois pas que l'on considère une relation binaire sous-jacente qui soit asymétrique, mais une relation binaire qui soit *fonctionnelle* (c'est-à-dire pour une relation binaire $\mathcal{R}$ sur $V_1 \times V_2$ le fait de vérifier la propriété : $\forall x \in V_1$, et $y, z \in V_2,\ x\mathcal{R}y \wedge x\mathcal{R}z \implies y = z$).

Les systèmes de traduction automatique qui utilisent l'apprentissage non supervisé issu des modèles IBM ont par conséquent le plus souvent recours à l'entraînement d'un modèle dans chaque « sens », de la langue source vers la langue cible et de la langue cible vers la langue source. Des méthodes heuristiques dites de symétrisation (Och *et al.*, 1999 ; Koehn *et al.*, 2003) permettent de construire de nouveaux alignements en ajoutant, par expansion au voisinage des liens contenus dans l'intersection des ensembles de liens obtenus dans chaque sens, les liens présents seulement dans l'union des alignements. Sont ensuite extraites, à partir de ces alignements symétrisés, des tables de traduction constituées de segments.

2.1.3 Hypothèses de modélisation pour les variables latentes

Pour modéliser le terme $\mathrm{P}(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$, de nouvelles hypothèses doivent être faites afin de rendre possible, en terme de complexité, l'estimation des paramètres du modèle ainsi que l'inférence sur de nouvelles données. Il existe une littérature abondante sur ce sujet et nous ne donnons ici que de brefs repères concernant les hypothèses faites dans les modèles IBM 1, IBM 2 et HMM, qui servent de base à de nombreux modèles plus complexes.

Il peut être utile, pour avoir une intuition de ces modèles génératifs, de considérer la décomposition suivante en deux sous-modèles relatifs, respectivement, à la *distorsion* (de la place des mots) et à la *traduction*[3] :

3. Rappelons que ces modèles, aujourd'hui exclusivement utilisés pour réaliser de manière non supervisée des alignements mot à mot avaient initialement vocation à être des modèles complets de traduction.

$$\mathrm{P}(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \underbrace{\mathrm{P}(\mathbf{a} \mid \mathbf{e})}_{\text{distorsion}} \underbrace{\mathrm{P}(\mathbf{f} \mid \mathbf{a}, \mathbf{e})}_{\text{traduction}}$$

Observons par ailleurs qu'il est possible d'écrire, par dérivation en chaîne de la règle de Bayes, et sans perte de généralité, la factorisation plus fine suivante :

$$\mathrm{P}(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \mathrm{P}(J \mid \mathbf{e}) \prod_{j=1}^{J} \mathrm{P}(a_j \mid a_1^{j-1}, f_1^{j-1}, J, \mathbf{e})\, \mathrm{P}(f_j \mid a_1^{j}, f_1^{j-1}, J, \mathbf{e}) \tag{2.2}$$

Les modèles présentés ci-après correspondent à des hypothèses qui permettent de simplifier cette factorisation.

IBM 1 Dans ce modèle, on fait l'hypothèse que chaque alignement a_j est choisi indépendamment des autres (a) et que les alignements sont distribués de manière uniforme (b) sur toutes les positions possibles de la phrase cible, de 0 à I, ce qui se traduit par

$$\mathrm{P}(\mathbf{a} \mid \mathbf{e}) \overset{\text{a}}{\equiv} \prod_{j=1}^{J} \mathrm{P}(a_j \mid \mathbf{e}) \overset{\text{b}}{\equiv} \frac{1}{(I+1)^J} \ .$$

Par ailleurs, une fois les alignements déterminés, les mots sources sont supposés ne dépendre que du mot cible sur lequel ils sont alignés, et partant

$$\mathrm{P}(\mathbf{f} \mid \mathbf{a}, \mathbf{e}) \equiv \prod_{j=1}^{J} \mathrm{P}(f_j \mid e_{a_j}) \ ,$$

d'où l'on déduit, en ignorant [4] la probabilité de la longueur J de la phrase source, la forme complète décrivant le modèle IBM 1 :

$$\mathrm{P}(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \mathrm{P}(f_j \mid e_{a_j}) \ .$$

La probabilité de générer $\mathbf{f}$ à partir de $\mathbf{e}$ dans ce modèle est ensuite obtenue par marginalisation de la variable d'alignement, comme définie par l'équation (2.1).

Un problème bien connu de ce modèle est l'effet, appelé *garbage collector*, qui tend à aligner les mots rares de la phrase cible sur de trop nombreux mots de la phrase source, et pour lequel divers remèdes sont présentés dans (Moore, 2004). Ce modèle souffre aussi de ne pas capturer de dépendance entre les positions des mots qui sont alignés, ce qui induit des effets de distorsion.

IBM 2 Le modèle IBM 2 introduit cette dépendance, faisant défaut au modèle IBM 1, entre la valeur de a_j et la position j du mot correspondant dans la phrase cible en réécrivant le premier terme du produit dans (2.2) selon :

$$\mathrm{P}(a_j \mid a_1^{j-1}, f_1^{j-1}, J, \mathbf{e}) \equiv \mathrm{P}(a_j | j, J, I),$$

Ceci permet de prendre en compte, par exemple, le fait que les liens d'alignement se trouvent le plus souvent autour de la diagonale (en considérant une représentation matricielle de l'alignement). L'hypothèse faite sur les dépendances lexicales est identique

4. En utilisant ce modèle pour aligner les mots de phrases dont la longueur est connue (observée), et non comme un modèle de traduction, ceci ne pose pas de problème.

au modèle IBM 1 et – en ignorant derechef la probabilité de J – la forme complète décrivant le modèle IBM 2 s'écrit :

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

Les modèles subséquents (IBM 3, IBM 4 et IBM 5) également introduits par (Brown *et al.*, 1993) corrigent – au prix d'une complexité nettement accrue – certaines limitations du modèle IBM 2. Une nouvelle variable latente de *fertilité*, en particulier, est introduite et modélise la propension d'un mot cible à être aligné avec plusieurs mots sources (IBM 3). La dépendance en position pour ces groupes de mots issus d'un même mot cible est capturée dans une phase de réordonnancement (IBM 4) et les déficiences des modèles IBM 3 et IBM 4 sont traitées par le modèle IBM 5 dont la complexité rend l'utilisation moins attractive en pratique. Le modèle IBM 4 constitue la *baseline* « état-de-l'art » pour la plupart des travaux sur l'alignement mot à mot bilingue.

HMM Un autre modèle, élaborant sur ceux présentés plus haut, a été introduit par (Vogel *et al.*, 1996) avec la volonté de prendre en compte la monotonie locale des alignements habituellement observée. Ainsi, le terme de distorsion est approximé par une hypothèse de dépendance markovienne d'ordre 1

$$\mathrm{P}(\mathbf{a} \mid \mathbf{e}) \equiv \prod_{j=1}^{J} \mathrm{P}(a_j \mid a_{j-1}, \mathbf{e}) \ ,$$

d'où l'on peut déduire, en faisant pour le terme de traduction une hypothèse identique à celle des modèles IBM 1 et IBM 2, et en ignorant toujours le terme $\mathrm{P}(J \mid \mathbf{e})$, la forme suivante pour le modèle HMM :

$$\mathrm{P}(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \prod_{j=1}^{J} \mathrm{P}(a_j \mid a_{j-1}, \mathbf{e})\, \mathrm{P}(f_j \mid e_{a_j}) \ .$$

En pratique, le terme $\mathrm{P}(a_j \mid a_{j-1}, \mathbf{e})$ prend la forme d'une fonction de la largeur des sauts $|a_j - a_{j-1}|$, et il faut ajouter un état supplémentaire pour chaque état du modèle de Markov caché afin de prendre en compte les alignements vides (voir note 2) sans rompre l'hypothèse markovienne.

2.1.4 Estimation des paramètres

Les corpus parallèles ne contiennent que très rarement, et le cas échéant en très faible quantité, d'annotations pour l'alignement au niveau des mots. L'apprentissage de ces alignements est donc une tâche d'apprentissage essentiellement non supervisé [5]. Afin d'estimer les paramètres d'un modèle d'alignement, c'est-à-dire entraîner le modèle, l'algorithme itératif EM [6] (Dempster *et al.*, 1977) est convoqué pour trouver la valeur des paramètres qui maximisent la vraisemblance des données malgré l'absence des variables latentes dans les données observées.

L'intuition sous-jacente est qu'il est possible, lorsque les données d'alignement sont disponibles (données *complètes*), d'estimer les paramètres du modèle en formant le lagrangien associé au programme de maximisation sous contrainte de la vraisemblance.

5. Les modèles dit discriminants néanmoins, que nous n'évoquons que très brièvement, apprennent des alignements de manière supervisée. Ils compensent la faible quantité des données disponibles par l'exploitation d'une grande richesse de traits associés à ces données.

6. *Expectation-Maximization*

Inversement, les estimateurs de maximum de vraisemblance pour les paramètres permettent de calculer les probabilités a posteriori des liens d'alignement. En l'absence de données complètes et en initialisant les paramètres arbitrairement [7], on pourra utiliser ces probabilités a posteriori comme « pseudo-comptes » pour ré-estimer les paramètres du modèle maximisant la vraisemblance.

Plus formellement, en notant $\mathcal{C} = \{(\mathbf{e}^{(n)}, \mathbf{f}^{(n)}), n = 1 \dots n_D\}$ une collection de n_D paires de phrases en relation de traduction, $\mathcal{A}^{(n)}$ l'ensemble des alignements possibles pour la n-ième paire, et $\boldsymbol{\theta}$ l'ensemble des paramètres du modèle, la log-vraisemblance que nous cherchons à maximiser est donnée par :

$$\ell(\boldsymbol{\theta}) = \log\left(\prod_{n=1}^{n_D} \mathrm{P}(\mathbf{f}^{(n)} \mid \mathbf{e}^{(n)}; \boldsymbol{\theta})\right) = \sum_{n=1}^{n_D} \log(\mathrm{P}(\mathbf{f}^{(n)} \mid \mathbf{e}^{(n)}; \boldsymbol{\theta})) \tag{2.3}$$

$$= \sum_{n=1}^{n_D} \log(\sum_{\mathbf{a}^{(n)} \in \mathcal{A}^{(n)}} \mathrm{P}(\mathbf{f}^{(n)}, \mathbf{a}^{(n)} \mid \mathbf{e}^{(n)}; \boldsymbol{\theta})) \ . \tag{2.4}$$

On peut prouver que les valeurs de $\boldsymbol{\theta}$ qui améliorent la fonction auxiliaire $Q_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$ définie par

$$Q_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \sum_{n=1}^{n_D} \sum_{\mathbf{a}^{(n)} \in \mathcal{A}^{(n)}} \mathrm{P}(\mathbf{a}^{(n)} \mid \mathbf{e}^{(n)}, \mathbf{f}^{(n)}; \boldsymbol{\theta}') \log(\mathrm{P}(\mathbf{f}^{(n)}, \mathbf{a}^{(n)} \mid \mathbf{e}^{(n)}; \boldsymbol{\theta})) \tag{2.5}$$

améliorent également la log-vraisemblance (et donc la vraisemblance) des données observées. L'algorithme EM, justifié par cette propriété de la fonction auxiliaire – plus facile à optimiser que la log-vraisemblance – opère ainsi itérativement selon les deux étapes suivantes :

— Étape E : Les probabilités a posteriori $\mathrm{P}(\mathbf{a}^{(n)} \mid \mathbf{e}^{(n)}, \mathbf{f}^{(n)}; \boldsymbol{\theta}')$ des alignements selon les paramètres courants $\boldsymbol{\theta}'$ du modèle sont calculées.

— Étape M : La maximisation de la fonction auxiliaire $Q_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$ fournit la nouvelle valeur des paramètres $\hat{\boldsymbol{\theta}}$, avec $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$, qui remplacent $\boldsymbol{\theta}'$ dans l'itération suivante.

2.1.5 Extraction des alignements

Une fois que l'algorithme EM a convergé, on dispose d'un modèle flanqué de paramètres qui maximisent [8] la vraisemblance des données partielles observées. Il est alors possible [9] d'extraire les alignements les plus probables selon

$$\hat{\mathbf{a}}^{(n)} = \arg\max_{\mathbf{a}^{(n)} \in \mathcal{A}^{(n)}} \mathrm{P}(\mathbf{a}^{(n)} \mid \mathbf{e}^{(n)}, \mathbf{f}^{(n)}) \ .$$

On peut également prédire tous les liens d'alignement correspondant à une probabilité a posteriori supérieure à un seuil ainsi que le proposent (Liang *et al.*, 2006) dans

7. Les valeurs initiales des paramètres auront néanmoins un impact fort sur la convergence de l'algorithme.

8. Il s'agit le plus souvent d'un maximum local, la log-vraisemblance n'étant le plus souvent pas concave. Le modèle IBM 1, néanmoins, admet un maximum de vraisemblance global et les paramètres correspondant à ce maximum sont couramment utilisés pour initialiser les paramètres de modèles plus complexes, l'opération pouvant être reproduite à nouveau en construisant ainsi successivement plusieurs étages de modélisation.

9. Le problème est toutefois extrêmement difficile en toute généralité, et un algorithme de recherche devra tirer parti des hypothèses simplificatrices du modèle considéré, ou procéder par heuristique (recherche autour de la diagonale, par exemple).

le cadre de l'apprentissage joint de deux modèles d'alignement asymétriques pour lesquels les auteurs cherchent à faire concorder les prédictions, de manière à réaliser une symétrisation pendant l'apprentissage, et non via une heuristique de symétrisation a posteriori.

2.2 Approches complémentaires

Nous présentons dans cette section des approches développées pour remédier à l'asymétrie des alignements IBM, ou intégrer des connaissances plus riches dans les modèles d'alignement.

2.2.1 Régularisation des probabilités a posteriori des alignements

Il est difficile de garantir certaines propriétés désirables pour les alignements, comme la symétrie ou la bijectivité, sans rendre insoluble numériquement l'apprentissage de modèles intégrant dans leur « histoire générative » ces contraintes. Une méthode introduite par (Graça *et al.*, 2007), et détaillée spécifiquement pour la tâche d'alignement de mots dans (Graça *et al.*, 2010), consiste à modifier l'étape E de l'algorithme EM afin de *régulariser* itérativement la distribution des paramètres sans complexifier le modèle sous-jacent.

Au lieu de calculer pendant l'étape E de l'algorithme EM les probabilités a posteriori $\mathrm{P}(\mathbf{a} \mid \mathbf{x})$ des liens d'alignement selon les paramètres courants du modèle (en notant $\mathbf{x} = (\mathbf{e}, \mathbf{f})$ pour alléger la notation et suivre l'usage dans la littérature), on calcule ces probabilités selon une distribution $q(\mathbf{a} \mid \mathbf{x})$ qui vérifie :

$$\arg\min_{q} \mathrm{KL}(q(\mathbf{a} \mid \mathbf{x}) || \, \mathrm{P}(\mathbf{a} \mid \mathbf{x}; \boldsymbol{\theta}')) \quad \text{t.q.} \quad \mathbb{E}_q[\mathbf{f}(\mathbf{x}, \mathbf{a})] \leq \mathbf{b} \tag{2.6}$$

avec $\mathbf{f}$ un vecteur de fonctions et $\mathbf{b}$ un vecteur de constantes[10], et $\mathrm{KL}(q||p) = \mathbb{E}_q[\log \frac{q(\cdot)}{p(\cdot)}]$ la divergence de Kullback-Leibler. Cette divergence, quoique ne constituant pas une distance à proprement parler, fournit une mesure de comparaison entre deux distributions (d'autant plus proches que la divergence est petite). Elle est toujours positive, et s'annule lorsque les distributions sont identiques.

Le problème de minimisation défini par (2.6) consiste à projeter la distribution a posteriori définie à chaque itération de l'algorithme EM sur un espace de distributions respectant certaines contraintes.

À titre d'exemple, pour contraindre pendant l'apprentissage les alignements à satisfaire, *en espérance*, une forme de bijectivité (alignements « 1-1 ») on pourra définir pour chaque mot cible e_i une fonction $f_i(\mathbf{x}, \mathbf{a}) = \sum_j \mathbb{1}(a_j = i)$ et introduire la contrainte $\mathbb{E}[\mathbf{f}(\mathbf{x}, \mathbf{a})] \leq 1$.

Pour garantir, d'autre part, la symétrie des alignements, c'est-à-dire l'identité de la probabilité a posteriori des liens pour chaque sens d'alignement, on imposera que des fonctions associées à chaque lien et comptant $+1$ pour le sens source-cible et -1 pour le sens cible-source, aient une espérance nulle sous un modèle de mélange des distributions dans les deux sens (voir (Graça *et al.*, 2010) pour plus de précisions).

2.2.2 Modèles discriminants

Une classe très importante de modèles d'alignement, et qui est seulement mentionnée sans développement ici, est la classe des modèles discriminants, comme les classifieurs à

10. Notons qu'une contrainte d'égalité peut se spécifier à l'aide des deux contraintes $\mathbb{E}_q[f(\mathbf{x}, \mathbf{a})] \leq b$ et $\mathbb{E}_q[-f(\mathbf{x}, \mathbf{a})] \leq -b$.

maximum d'entropie (ou classifieurs d'entropie maximale) ou encore les champs conditionnels aléatoires, qui permettent de prendre en compte des traits beaucoup plus riches que ceux que sont capables de capturer les modèles génératifs.

On trouvera dans (Allauzen et Wisniewski, 2009) une étude approfondie de ces modèles. Nous citons ici une description concise et éclairante donnée par les auteurs de l'esprit dans lequel sont construits ces modèles : *« Les modèles discriminants proposent de modéliser directement la probabilité sur laquelle la prédiction est fondée, c'est-à-dire la probabilité conditionnelle des étiquettes connaissant les observations. Ils traitent donc une tâche plus simple que l'estimation de la probabilité jointe, puisqu'il n'est plus nécessaire d'estimer la probabilité de l'observation qui n'intervient pas directement dans la prise de décision. De plus, ces modèles permettent de prendre en compte des caractéristiques arbitraires, alors que les modèles génératifs, pour des questions d'efficacité, imposent des hypothèses d'indépendance fortes entre les caractéristiques et n'ont donc qu'une expressivité réduite. ».* On pourra également se référer aux travaux de (Berger *et al.*, 1996 ; Ayan et Dorr, 2006 ; Blunsom et Cohn, 2006).

Chapitre 3

Le cas multilingue

Nous l'avons mentionné brièvement en section 1.3.3, de nouveaux types de stratégies sont envisageables lorsque l'on dispose, dans un corpus parallèle, d'au moins trois langues (ou « systèmes linguistiques », voir section 1.1). L'alignement de séquences multiples est un problème plus large, qui intéresse en particulier la bio-informatique pour l'alignement de séquences de gènes ou de protéines (Gusfield, 1997). Dans ce chapitre, nous dressons un état de l'art des différents travaux qui s'attaquent à la problématique de l'alignement multilingue de séquences textuelles selon trois axes : les objectifs, les représentations formelles, et les méthodes convoquées. Puis nous dressons une typologie de l'alignement multilingue et proposons quelques pistes de travail.

3.1 Notations

Nous notons, comme précédemment, $\mathbf{e}$ une séquence d'unités du système S_1 dans le vocabulaire (indexé) V_1 et $\mathbf{f}$ une séquence d'unités du système S_2 dans le vocabulaire V_2, et ajoutons la notation $\mathbf{g} = g_1^K = g_1, \ldots, g_K$ pour une séquence d'unités du système S_3 dans le vocabulaire V_3.

Nous notons $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ un alignement de $\mathbf{e}$ avec $\mathbf{f}$ (ou $A^{\mathbf{e}\leftarrow\mathbf{f}}$ pour l'alignement des unités de $\mathbf{f}$ sur les unités de $\mathbf{e}$, et $A^{\mathbf{f}\leftarrow\mathbf{e}}$ respectivement, si l'alignement ne correspond pas à une relation binaire symétrique). $A^{\mathbf{e},\mathbf{f},\mathbf{g}}$ désigne un alignement de $\mathbf{e}$, $\mathbf{f}$ et $\mathbf{g}$. Si la réalisation de l'alignement a pour forme un vecteur, ce vecteur sera noté $\mathbf{a}$.

En cas de besoin, nous généraliserons la notation avec $\mathbf{u^{(n)}} = u_1^{(n)}, \ldots, u_{L_n}^{(n)}$ représentant une séquence d'unités du système S_n dans le vocabulaire V_n. Par ailleurs, $\mathcal{V}(\mathbf{u})$ représentera l'ensemble défini par l'union de toutes les unités contenues dans $\mathbf{u}$. Rappelons que les unités considérées peuvent être des phrases, des mots, ou des groupes de mots.

3.2 État de l'art

Nous proposons, dans cette section, une synthèse de plusieurs travaux de référence s'attaquant au problème de l'alignement d'au moins trois langues.

3.2.1 Première tentative d'alignement multilingue

- article de référence : (Simard, 1999)
- unités alignées : phrases
- nombre de langues : 3

- type de méthode : associative / asymétrique (et tentative pour une méthode symétrique)
- objectif : améliorer l'alignement bilingue

Objectifs L'un des premiers travaux cherchant à exploiter, dans le cadre de l'alignement de phrases, l'information fournie par un troisième texte en relation de traduction mutuelle avec deux autres textes est réalisé par (Simard, 1999). L'objectif principal, bien que d'autres aspects soient envisagés et explorés, est d'améliorer la qualité d'un l'alignement phrase à phrase bilingue.

Modélisation L'auteur définit un alignement *bilingue* $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ entre les séquences $\mathbf{e}$ et $\mathbf{f}$ – qui sont ici des *textes*, les unités alignées étant des phrases – comme une relation binaire sur $(\mathcal{V}(\mathbf{e}) \cup \mathcal{V}(\mathbf{f})) \times (\mathcal{V}(\mathbf{e}) \cup \mathcal{V}(\mathbf{f}))$ [1], c'est-à-dire un ensemble de couples. Cette relation binaire est posée comme étant réflexive, symétrique, et transitive.

Un alignement *multilingue* est d'abord défini comme une relation n-aire (voir 1.2.1), puis sera ensuite vu indifféremment [2] comme une union de relations binaires, c'est-à-dire :

$$A^{\mathbf{e},\mathbf{f},\mathbf{g}} = A^{\mathbf{e}\leftrightarrow\mathbf{f}} \cup A^{\mathbf{f}\leftrightarrow\mathbf{g}} \cup A^{\mathbf{g}\leftrightarrow\mathbf{e}}$$

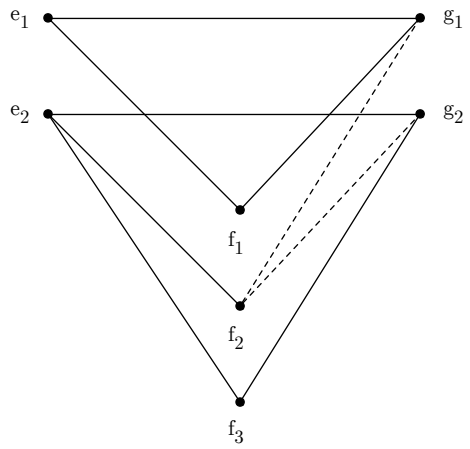


FIGURE 3.1 – Intuition quant à l'utilité d'un alignement multilingue

Nous reproduisons en Figure 3.1 un exemple donné par l'auteur pour donner l'intuition de la démarche. En cas d'incertitude sur les liens (f_2, g_1) ou (f_2, g_2) dans l'alignement $A^{\mathbf{f}\leftrightarrow\mathbf{g}}$, les alignements $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ et $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$ suggèrent de valider le lien (f_2, g_2) et de rejeter le lien (f_2, g_1).

Méthodes Les alignements multilingues sont réalisés par combinaison d'alignements de plus faibles degrés, en l'espèce des alignements bilingues. Nous donnons les principales étapes de la procédure :

1. Calcul des alignements bilingues $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$, $A^{\mathbf{f}\leftrightarrow\mathbf{g}}$, et $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$ avec une méthode du type de celle introduite par (Gale et Church, 1991).

1. V_1 ou V_2, selon les notations introduites en section 1.1 représenteraient ici le vocabulaire d'une *collection* de textes dans le système linguistique S_1 ou S_2.
2. Quoique cela ne soit pas a priori équivalent.

2. Identification de la paire de séquences la plus « similaire » à partir des scores fournis par l'algorithme d'alignement bilingue.
3. Alignement de la troisième séquence sur l'alignement déterminé à l'étape précédente. Il faut préciser ici que l'on aligne alors des unités avec un alignement, ce qui n'est pas nécessairement facile à formaliser. En pratique, l'alignement déjà réalisé est vu comme une séquence d'unités, et le nouvel alignement est réalisé comme dans le cas bilingue grâce à la programmation dynamique, à ceci près que le score des pseudo-triplets est calculé comme la somme des scores des unités deux à deux.

Un redécoupage a posteriori est aussi réalisé afin de corriger un « bruit de transitivité » (*transitivity noise*) : dans le cadre de ce travail, en effet, chaque unité e_i alignée avec une unité f_j sera *de facto* alignée avec toutes les unités elles-mêmes alignées avec f_j. Cet effet est illustré en Figure 3.2 (en considérant des mots et non des phrases).

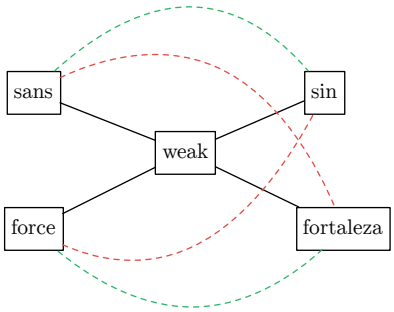


FIGURE 3.2 – Illustration du bruit transitif (en rouge, les liens non souhaitables résultant de la transitivité de la relation binaire)

Résultats et perspectives La méthode implémentée permet d'améliorer la F-mesure pour des alignements bilingues par rapport à des méthodes purement bilingues en réduisant l'erreur de 20%.

Par ailleurs, une optimisation est implémentée par le biais d'une segmentation de la séquence trilingue alignée en recherchant des « points d'accord » entre des groupes d'unités formés par fermeture transitive sur la relation binaire d'alignement. Quoique n'améliorant pas la performance des alignements bilingues, ceci permet de réduire le temps de calcul en limitant la profondeur de l'espace de recherche.

Une implémentation, grâce à cette segmentation, d'une méthode véritablement trilingue, utilisant la programmation dynamique dans le cube de l'espace de recherche, donne des résultats très décevants et pose la question (voir la note 2) de la modélisation choisie initialement, à savoir le remplacement d'un modèle de traduction trilingue par trois modèles de traduction bilingues.

Enfin, la généralisation de cette méthode à plus de trois langues, bien que non triviale, semble envisageable. Pour ce qui est d'une adaptation à l'alignement de mots et non de phrases, la principale difficulté serait de trouver une méthode pour la correction du bruit de transitivité. La méthode employée dans (Simard, 1999) n'est pas détaillée mais semble reposer sur la monotonie des alignements phrastiques.

3.2.2 Logique du pont

- article de référence : (Kumar *et al.*, 2007)
- unités alignées : mots (et subséquemment, segments)

- nombre de langues : N
- type de méthode : paramétrique / asymétrique
- objectif principal : améliorer l'alignement bilingue et un système de traduction bilingue

Objectifs Il s'agit d'utiliser l'information fournie par un corpus multilingue pour apprendre plusieurs alignements mot à mot bilingues pour une paire de langues particulière, via des langages pivots (*bridge languages*), et de les injecter dans un système de traduction bilingue. Plutôt que de définir un modèle à proprement parler multilingue, il s'agit donc de *combiner* différents modèles trilingues, eux-mêmes construits sur des modèles bilingues.

Modélisation Le cadre formel de modélisation des alignements bilingues est celui, suivant (Brown *et al.*, 1993), que nous avons introduit en section 2.1.1. Nous considérons des alignements $\mathbf{a}^{\mathbf{e}\leftarrow\mathbf{f}}$ de type « 1-n », non symétriques, entre une séquence source $\mathbf{f}$ et une séquence cible $\mathbf{e}$. Une séquence $\mathbf{g}$ choisie dans l'une des langues supplémentaires disponibles, et correspondant à la traduction de $\mathbf{e}$ et de $\mathbf{f}$, sert de *pont* entre les deux séquences.

Méthodes Les probabilités postérieures $\mathrm{P}(a_j^{\mathbf{e}\leftarrow\mathbf{f}} = i \mid \mathbf{e}, \mathbf{f})$, qui forment une matrice $(I+1) \times J$, sont calculées selon l'expression

$$\mathrm{P}(a_j^{\mathbf{e}\leftarrow\mathbf{f}} = i \mid \mathbf{e}, \mathbf{f}) = \sum_{k=0}^{K} \mathrm{P}(a_j^{\mathbf{g}\leftarrow\mathbf{f}} = k \mid \mathbf{g}, \mathbf{f}) \, \mathrm{P}(a_k^{\mathbf{e}\leftarrow\mathbf{g}} = i \mid \mathbf{g}, \mathbf{e})$$

cette simplification s'obtenant en faisant les hypothèses suivantes :

1. Il n'existe qu'une seule traduction $\mathbf{g}$ correspondant à la paire de phrases $\mathbf{e}, \mathbf{f}$,
2. $\mathbf{a}^{\mathbf{e}\leftarrow\mathbf{g}}_{\mathbf{a}_j^{\mathbf{g}\leftarrow\mathbf{f}}} = i = \mathbf{a}_j^{\mathbf{e}\leftarrow\mathbf{f}}$, ce qui revient à considérer que la relation associée aux liens est transitive,
3. Un alignement entre $\mathbf{f}$ et $\mathbf{g}$ ne dépend pas de $\mathbf{e}$.

Ceci permet d'obtenir, en plus de la probabilité a posteriori « directe », la probabilité a posteriori des liens entre $\mathbf{e}$ et $\mathbf{f}$ à partir d'un produit matriciel sur les probabilités postérieures $\mathrm{P}(a_j^{\mathbf{g}\leftarrow\mathbf{f}} = k \mid \mathbf{g}, \mathbf{f})$ et $\mathrm{P}(a_k^{\mathbf{e}\leftarrow\mathbf{g}} = i \mid \mathbf{g}, \mathbf{e})$, pour peu que l'on complète la seconde matrice par une colonne correspondant à $k = 0$ (non déterminée par le modèle d'alignement bilingue).

Dans un second temps, il s'agit de combiner différentes « versions » des probabilités a posteriori des liens entre $\mathbf{e}$ et $\mathbf{f}$ selon plusieurs langues jouant ce rôle de pont, comme représenté en Figure 3.3. En pratique, un poids identique est donné à chaque probabilité a posteriori lors de l'interpolation.

Enfin, la probabilité a posteriori interpolée P_{int} est utilisée pour prédire les alignements selon le critère du maximum a posteriori (MAP) [3] :

$$a_{MAP}^{\mathbf{e}\leftarrow\mathbf{f}}(j) = \arg\max_i \mathrm{P}_{int}(a_j = i \mid \mathbf{e}, \mathbf{f})$$

3. L'estimateur MAP de a_j contraste avec l'alignement dit de Viterbi défini pour la séquence entière d'alignement par

$$\mathbf{a}_V = \arg\max_{\mathbf{a}} \mathrm{P}(\mathbf{a} \mid \mathbf{e}, \mathbf{f}) = \arg\max_{\mathbf{a}} \mathrm{P}(\mathbf{a}, \mathbf{f} \mid \mathbf{e}).$$

Dans le cas des modèles IBM 1 ou IBM 2, les deux critères sont équivalents, mais ce n'est plus le cas pour un modèle HMM par exemple.

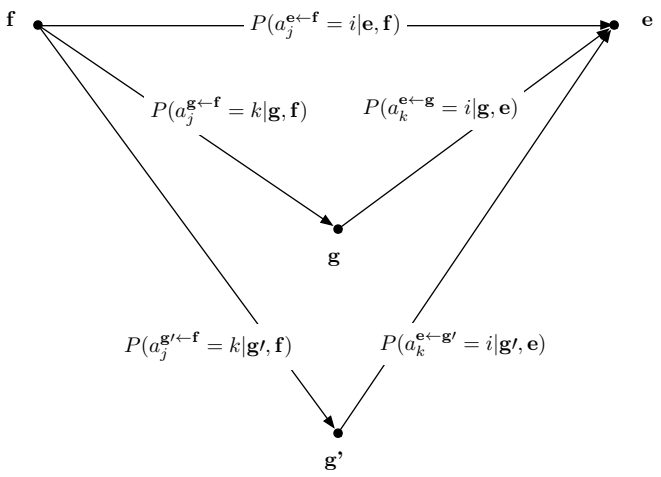


FIGURE 3.3 – Exemple d'utilisation de deux langues-pivots pour calculer selon trois chemins des probabilités a posteriori qui seront ensuite interpolées.

Résultats et perspectives Cette méthode, utilisée avec quatre langues-pivots et une interpolation sur l'alignement direct et sur les alignements réalisés via les pivots exhibe des taux d'erreur (au sens de la mesure AER, voir section 1.3.2) plus élevés que l'alignement direct [4], et ceci dans les deux sens d'alignement réalisés séparément. Chaque alignement « pivot » est individuellement moins bon, néanmoins, que l'alignement « combiné », avec des disparités importantes selon la langue choisie.

En revanche, lorsque ces alignements supplémentaires sont utilisés pour diversifier les hypothèses d'un système de traduction, des gains sont obtenus sur le score BLEU, en particulier pour un système utilisant un décodeur par consensus. Nous n'en donnons pas le détail ici.

Concernant la qualité des alignements, qui nous intéresse spécifiquement dans ce document, les auteurs de (Kumar *et al.*, 2007) font l'hypothèse qu'un choix plus fin de la pondération au moment de l'interpolation serait à même d'améliorer les résultats, mais cette piste n'est pas explorée.

Il nous semble également possible que ces résultats décevants soient liés de nouveau au bruit de transitivité évoqué dans (Simard, 1999).

Un aspect séduisant de la méthode est que la combinaison de matrices de probabilités a posteriori pour les alignements est agnostique du modèle d'apprentissage. Dans (Kumar *et al.*, 2007), ces probabilités sont estimées après six itérations d'un modèle IBM-1 (Brown *et al.*, 1993) puis six itérations d'un modèle HMM (Vogel *et al.*, 1996 ; Deng et Byrne, 2005), mais tout autre modèle d'alignement est envisageable.

Enfin, bien qu'asymétrique, la méthode n'est pas limitée en nombre de langues-pivots et pourrait tirer parti de corpus « massivement » parallèles.

3.2.3 Alignement par échantillonnage

- article de référence : (Lardilleux *et al.*, 2011)
- unités alignées : mots et groupes de mots (segments éventuellement discontinus)
- nombre de langues : N

4. On lit pourtant de manière étonnante en section 6 de (Kumar *et al.*, 2007) : *« Despite its simplicity, the system combination gives improvements in alignment and translation performance. »*

- type de méthode : associative / symétrique
- objectif principal : réaliser des pseudo-alignements sous-phrastiques

Objectifs L'idée centrale introduite par (Lardilleux et Lepage, 2008, 2009) et développée dans (Lardilleux *et al.*, 2011) est que la faible fréquence des mots représente une information aussi valable que leur haute fréquence pour apprendre des liens [5] de correspondance entre des groupes sous-phrastiques. Les hapax, en particulier, sont systématiquement filtrés par les méthodes statistiques traditionnelles, alors qu'ils constituent une part très importante du contenu textuel. Le but est, partant, de construire une méthode d'alignement basée sur l'identification de profils distributionnels *identiques* (et non pas seulement *similaires* au sens d'une mesure cosinus ou de Dice par exemple) en échantillonnant des sous-corpus de différentes tailles à l'intérieur d'un corpus multilingue.

Contrairement à des approches de type paramétriques qui procèdent par optimisation globale des paramètres d'un modèle donné sur un vaste corpus en s'appuyant essentiellement sur sa redondance, c'est-à-dire la grande fréquence de certains de ses motifs, la méthode employée ici s'inscrit dans le prolongement des approches *associatives* qui cherchent à maximiser localement des scores d'association entre des entités.

Modélisation Étant donné un corpus multilingue aligné au niveau phrastique et composé de N systèmes linguistiques S_n, un alignement correspond à une relation n-aire sur $V_1 \times \cdots \times V_N$, avec les ensembles V_n définis comme des ensembles de mots et de groupes de mots, éventuellement discontinus, sur chacun des systèmes S_n.

Aucun lien n'est réellement construit au niveau des seuls mots, et l'on pourrait par conséquent arguer qu'il ne s'agit pas véritablement d'« alignements » au sens où la littérature emploie le plus communément ce mot.

Méthodes Un aspect clef de la méthode – qui cherche donc à identifier des profils distributionnels identiques entre les différents textes – est d'échantillonner différentes tailles de sous-corpus afin de capturer plusieurs effets complémentaires. En particulier, *soustraire* des données, dans ce cadre, peut s'avérer être une stratégie tout à fait légitime et efficace.

Voici les principales étapes de la boucle correspondant à ce traitement :

1. Échantillonnage d'un sous-corpus, dont la taille t_e en nombre de phrases suit une loi proportionnelle à
$$\frac{-1}{t_e \log\left(1 - \frac{t_e}{t_{tot}}\right)} \quad ,$$
avec t_{tot} le nombre global de phrases du corpus multilingue. Ceci favorise les sous-corpus de taille réduite, les plus à même de tirer parti des faibles fréquences, tout en autorisant des échantillons plus larges, permettant une grande couverture du corpus, et la capture des hapax plus « globaux » ;
2. Pour chaque mot de chaque langue contenu dans le corpus échantillonné, le vecteur de sa distribution dans l'espace vectoriel défini par les phrases du sous-corpus est calculé (nombre d'occurrences du mot pour chaque phrase) ;
3. Regroupement de tous les mots, langues confondues, ayant des vecteurs identiques sur le sous-corpus ;

5. Au sens le plus général : nous verrons que l'approche présentée ici ne construit pas véritablement des liens mais plutôt des scores d'association.

4. L'ordre initial des mots est restauré, à l'intérieur de chaque groupe, par langue et par phrase, et un compteur est incrémenté, non seulement pour le groupe réordonné, mais aussi pour le groupe « complémentaire » dans la phrase initiale.

Le critère d'arrêt de l'échantillonnage sera soit une durée, soit l'absence de nouveaux « liens ».

Une fois l'extraction des alignements réalisée, et afin de pouvoir les comparer à des méthodes existantes, une table de traduction est construite en attribuant des scores aux alignements. Les scores habituellement utilisés [6] sont adaptés étant donné l'absence de liens mot à mot. Il faut noter que cela oblige à ignorer les alignements de segments discontinus qui sont construits, et qui constituent pourtant un atout important de cette nouvelle approche, car ceux-ci ne sont pas pris en charge par les autres méthodes d'alignement.

Résultats et perspectives Deux campagnes de test sont réalisées pour comparer l'implémentation de cette méthode, `Anymalign`, avec les outils d'alignement état-de-l'art `MGIZA++` [7] et `BerkeleyAligner` [8].

Sur une tâche de traduction automatique, `Anymalign` obtient des scores plus bas que les méthodes classiques, ce qui semble provenir du nombre assez faible de n-grammes de taille supérieure à 2 extraits par `Anymalign` relativement aux deux autres méthodes. Cet effet sera corrigé par un raffinement de la méthode décrit dans (Lardilleux *et al.*, 2013) qui permet d'obtenir, sur la tâche de traduction, des performances équivalentes aux systèmes état-de-l'art.

Sur une tâche d'extraction de lexique bilingue, `Anymalign` donne en revanche des résultats nettement supérieurs aux deux autres systèmes.

Pour résumer, la nouvelle méthode d'alignement introduite ici est performante, *anytime* [9], non-directionnelle, véritablement multilingue, et simple à mettre en œuvre en comparaison des méthodes concurrentes.

En revanche, elle n'est pas très solidement assise théoriquement (l'échantillonnage mis en œuvre ressort plus de l'heuristique que du modèle statistique), et ne fournit pas de liens, qui peuvent être utiles pour différentes tâches, au niveau des mots.

3.2.4 Clustering de similarités

- article de référence : (Mayer et Cysouw, 2012)
- unités alignées : mots
- nombre de langues : N
- type de méthode : associative / symétrique
- objectif principal : comparaison des langues et reconstruction phylogénétique

Objectifs De même qu'en biologie de l'évolution des séquences de gènes sont comparées afin de révéler d'éventuelles parentés entre les espèces, les auteurs de (Mayer et Cysouw, 2012) souhaitent comparer des séquences de mots disponibles dans des corpus massivement parallèles [10] afin de reconstruire automatiquement des arbres phylogénétiques.

6. Ces scores font intervenir la probabilité des liens d'alignement et les poids lexicaux, selon une méthode introduite par (Koehn *et al.*, 2003).

7. `http://www.kyloo.net/software/doku.php/mgiza:overview`

8. `https://code.google.com/p/berkeleyaligner/`

9. C'est-à-dire que l'algorithme est capable de fournir un résultat valide même s'il est interrompu.

10. Corpus multilingues portant sur de très nombreuses langues, éventuellement plusieurs centaines ; cela implique souvent, a contrario, des corpus relativement courts.

Modélisation Dans ce travail, un alignement multilingue est défini pour chaque phrase multi-parallèle par un ensemble de clusters regroupant les mots des différentes langues représentés dans cette phrase. Sauf à définir un sous-graphe complet pour chaque cluster, on ne disposera pas ici, ainsi qu'en section 3.2.3, de liens au niveau des mots.

Méthodes En premier lieu, une matrice de similarité entre tous les mots de toutes les langues, et basée sur la cooccurrence de ces mots dans des paires de phrases en traduction mutuelle, est construite. Dans un deuxième temps, et pour chaque phrase (c'est-à-dire toutes les versions de cette phrase), le sous-ensemble des scores de similarité des mots apparaissant dans cette phrase particulière sert de base à un partitionnement par un algorithme de clustering [11] de tous les mots de toutes les langues apparaissant dans cette phrase. Chaque cluster multilingue constitue, dans ce contexte, un *alignement*.

Nous ne rendons pas compte ici de l'usage qui est fait de ces alignements afin de comparer les langues entre elles en construisant une nouvelle mesure de similarité.

Résultats et perspectives Cette méthode semble donner des résultats raisonnables quant à la mesure d'une proximité entre différentes langues. Pour ce qui concerne la qualité des alignements, aucune évaluation n'est proposée et il est par conséquent difficile de se prononcer. L'attrait de la méthode repose essentiellement sur sa capacité à passer à l'échelle sur un très grand nombre de langues, ainsi que sur la nature symétrique des pseudo-alignements qu'elle construit.

3.2.5 Apprentissage conjoint pour quatre langues

- articles de référence : (Filali et Bilmes, 2005)
- unités alignées : mots
- nombre de langues : 4
- type de méthode : paramétrique / asymétrique
- objectif principal : améliorer l'alignement bilingue

Objectifs La question posée par ce travail est de savoir si l'information fournie par des alignements auxiliaires lorsque l'on dispose d'un corpus multilingue est à même d'améliorer les résultats d'un modèle d'alignement bilingue.

Modélisation Le formalisme pour l'alignement utilisé ici est le formalisme standard des modèles IBM, présenté en section 2.1.1. La sémantique de la variable latente d'alignement est néanmoins différente car elle régit l'alignement des deux langues principales mais aussi celui des deux langues auxiliaires comme nous le détaillons ci-après.

Méthodes Une phrase $\mathbf{e} = e_1^I$ est d'abord alignée, à l'aide d'un modèle IBM ou HMM, avec sa traduction $s_1^{I'}$ dans l'une des langues disponibles, et la séquence $\mathbf{s} = s_1^I$ est alors *construite* à partir des mots de $s_1^{I'}$ qui s'alignent le mieux avec les mots de e_1^I. On peut voir la séquence s_1^I comme une séquence de tags [12] pour la séquence e_1^I. De même, une séquence $\mathbf{g} = g_1^J$ est construite pour taguer la séquence $\mathbf{f} = f_1^J$.

Une dérivation similaire à celle donnée en équation (2.2) permet d'écrire en toute généralité [13] :

11. *Affinity Propagation*, qui n'oblige pas à faire d'hypothèse sur le nombre de classes.
12. Les auteurs appellent le modèle présenté ici *Alignment-tag Model*.
13. Nous corrigeons l'équation (4) de l'article car la probabilité, conditionnellement à l'observation de la phrase anglaise (et dans ce cas-ci espagnole), de la longueur de la phrase française ne doit pas être sous le produit.

$$\begin{aligned}\mathrm{P}(\mathbf{f},\mathbf{g},\mathbf{a} \mid \mathbf{e},\mathbf{s}) = \mathrm{P}(J \mid \mathbf{e},\mathbf{s}) \prod_{j=1}^{J} (& \mathrm{P}(a_j \mid a_1^{j-1}, f_1^{j-1}, g_1^{j-1}, J, \mathbf{e}, \mathbf{s}) \\ & \times \mathrm{P}(g_j \mid a_1^{j}, f_1^{j-1}, g_1^{j-1}, J, \mathbf{e}, \mathbf{s}) \\ & \times \mathrm{P}(f_j \mid a_1^{j}, f_1^{j-1}, g_1^{j}, J, \mathbf{e}, \mathbf{s}))\end{aligned}$$

Sous certaines hypothèses d'indépendance, et en adoptant une modélisation HMM (une dérivation équivalente peut être obtenue simplement pour les modèles IBM 1-4), cette dérivation peut s'écrire :

$$\mathrm{P}(\mathbf{f},\mathbf{g},\mathbf{a} \mid \mathbf{e},\mathbf{s}) = \mathrm{P}(J \mid \mathbf{e},\mathbf{s}) \prod_{j=1}^{J} \mathrm{P}(a_j \mid a_{j-1}, I, J)\, \mathrm{P}(f_j \mid e_{a_j})(\mathrm{P}(g_j \mid s_{a_j}))^{\alpha}$$

avec $\alpha \in [0, 1]$ un exposant qui permet de réguler l'importance du facteur $\mathrm{P}(g_j \mid s_{a_j})$. Notons que la variable d'alignement $\mathbf{a}$ sert désormais non seulement à aligner $\mathbf{e}$ avec $\mathbf{f}$ mais aussi la séquence de tags $\mathbf{g}$ avec $\mathbf{s}$. Ce dernier modèle augmenté de la probabilité $\mathrm{P}(f_j \mid e_{a_j})(\mathrm{P}(g_j \mid s_{a_j}))^{\alpha}$ est alors utilisé pour apprendre de nouveaux alignements entre $\mathbf{e}$ et $\mathbf{f}$.

Résultats et perspectives La performance de cette méthode d'alignement est évaluée selon la métrique AER définie en section 1.3.2, dans un scénario de traduction de l'anglais vers le français. Les meilleurs résultats sont obtenus en choisissant l'espagnol à la fois pour $\mathbf{s}$ et pour $\mathbf{g}$.

À quantité de données égale, la nouvelle méthode induit une baisse de l'ordre de 2% en valeur absolue sur l'erreur AER. Les données multilingues étant plus rares que les données bilingues, les auteurs observent également qu'un modèle IBM 4 amélioré [14] obtient un taux d'erreur comparable à la nouvelle méthode seulement lorsqu'il dispose d'au moins quatre fois plus de données d'entraînement.

Outre les bonnes performances de ce modèle, son attrait tient aussi dans une description probabiliste au niveau génératif même, ce qui autorise une bonne interprétation des hypothèses faites et des résultats obtenus. La méthode ouvre également la voie à une méthode itérative qui réinjecterait, pour déterminer de nouvelles séquences de tags, les alignements obtenus lors d'une précédente itération.

3.2.6 Modèle bayésien pour l'alignement d'un corpus massivement parallèle

- articles de référence : (Östling, 2014)
- unités alignées : mots
- nombre de langues : N
- type de méthode : paramétrique [15] / symétrique
- objectif principal : réaliser des alignements multilingues

Objectifs Le but de ce travail est d'apprendre simultanément des alignements mot à mot véritablement multilingues à partir d'un corpus massivement parallèle.

14. L'alignement est réalisé dans les deux sens, puis est symétrisé par intersection des liens, et enfin complété par une heuristique d'expansion (Och et Ney, 2003).

15. Quoique « non paramétrique » au sens bayésien.

Modélisation Le problème de l'alignement multilingue est ici formalisé comme un problème d'alignement de chaque langue avec une représentation commune *interlingua*.

Méthodes Une représentation, commune à toutes les langues, de concepts est générée par un *Chinese Restaurant Process* (CRP) [16]. Des alignements entre chaque langue et les concepts de cette représentation commune sont ensuite produits par un modèle d'alignement inspiré des modèles bayésiens utilisés pour l'alignement bilingue. La dérivation de ce modèle, assez technique et encore hors de notre portée, est omise ici.

Résultats et perspectives Le problème difficile de l'évaluation multilingue, déjà évoqué, s'envisage ici comme une comparaison des clusters constitués, d'une part, par les concepts communs et les alignements appris par le modèle, et d'autre part par l'utilisation des numéros de Strong [17].

La performance du modèle joint est comparée à une *baseline* basée sur un modèle comparable mais dans lequel une langue particulière (en pratique, ici, soit l'anglais, soit le mandarin) sert de représentation commune. Le modèle joint permet d'obtenir des gains significatifs.

3.2.7 Autres approches en rapport avec la problématique multilingue

Nous donnons ici quelques pointeurs vers des méthodes qui, quoique ne relevant pas de l'alignement multilingue, utilisent des approches pouvant inspirer une telle démarche.

Combinaisons Le travail présenté dans (Ayan et Dorr, 2006) utilise, ainsi que la méthode décrite dans (Kumar *et al.*, 2007) et présentée en section 3.2.2, des combinaisons d'alignements produits indépendamment, mais dans un contexte bilingue.

À partir d'alignements prédits par des méthodes paramétriques usuelles (combinaison de modèles IBM et HMM), un classifieur log-linéaire est appris sur les liens eux-mêmes, en faisant intervenir des *features* sur leur voisinage, leur fertilité, leur monotonie, et les parties du discours associées. Quoique conçue dans un cadre d'alignement bilingue, cette méthode pourrait être adaptée au problème de l'alignement multilingue. Pour autant, la difficulté réside dans la nécessité de disposer d'un corpus annoté (éventuellement de petite taille, cela étant) pour l'apprentissage du modèle discriminant.

Toujours dans un contexte bilingue, on retrouve cette approche par combinaison dans (Tu *et al.*, 2012) qui procèdent par sélection, raffinage , et combinaison d'alignements obtenus par différentes méthodes, et dans (Ayan *et al.*, 2005) qui ont recours à des réseaux de neurones pour apprendre l'alignement combiné.

Symétrie des alignements Un nombre important de travaux cherchent à remédier à la nature asymétrique des alignements type IBM. Citons en particulier (Och *et al.*, 1999) puis (Koehn *et al.*, 2003) qui introduisent les heuristiques de symétrisation évoquées en sections 2.1.2, ou (Matusov *et al.*, 2004) qui maximisent la probabilité des alignements

16. Processus stochastique discret à valeur dans les partitions de l'ensemble {1, 2, ..., n} et dont la distribution de probabilité permet de capturer une connaissance a priori sur la structure de ces partitions. L'analogie plaisante qui donne son nom à ce processus est la suivante : des clients s'attablent successivement dans un restaurant (chinois) qui dispose d'un nombre infini de tables. La probabilité qu'un client s'assoie à une table déjà occupée est proportionnelle à son nombre d'occupants, tandis que la probabilité qu'il s'assoie à une nouvelle table est, elle, proportionnelle à un facteur de concentration.

17. De James Strong, théologien américain ayant dirigé au XIX^{e} siècle la réalisation d'une concordance exhaustive de la Bible – le corpus multilingue utilisé dans ce travail reposant, en effet, sur de multiples traductions du Nouveau Testament. Les annotations « Strong » sont disponibles pour 9 traductions dans 7 langues à l'intérieur du corpus considéré.

symétriques par des algorithmes de recherche dans des graphes, ou encore via un véritable modèle bidirectionnel (DeNero et Macherey, 2011) très récemment rendu utilisable en pratique par (Chang *et al.*, 2014).

Par ailleurs, la méthode de régularisation a posteriori de (Graça *et al.*, 2007, 2010) présentée en section 2.2.1 permet de s'assurer, en espérance, de la bijectivité mais aussi de la symétrie des liens d'alignement [18].

Ces efforts dans le sens de la symétrisation des liens d'alignement nous semblent à même de réduire le bruit de transitivité qui apparaît par combinaison de liens (voir section 3.2.1).

Graphes et segments L'étude des alignements s'appuyant sur le formalisme de la théorie des graphes permet de tenter de poser des fondements solides en lieu et place de méthodes heuristiques qui n'ont pour preuve que l'évidence empirique. Dans (Martzoukos *et al.*, 2013), les auteurs montrent que l'heuristique de cohérence (*consistency*) suivant (Och *et al.*, 1999) et assortie d'un calcul de scores pour les segments extraits est mal fondée théoriquement, et que ces scores ne reposent pas sur une mesure de probabilité définie sur la tribu engendrée par les composantes connexes [19] des graphes correspondant aux alignements.

Un autre travail récent (Liu *et al.*, 2013) propose une représentation des alignements par des hypergraphes bipartis assortis de pondérations et qui permettent d'exploiter des relations entre les différents liens d'alignement pour extraire des segments.

Comme le font remarquer (Martzoukos *et al.*, 2013), voir les alignements comme des graphes revient à penser à ces alignements comme à des partitions plutôt qu'à des fonctions.

3.2.8 Synthèse de l'état de l'art

Un tableau récapitulatif de certains traits des méthodes étudiées est donné dans la Table 3.1. Nous contrastons les méthodes faisant intervenir l'estimation des paramètres d'un modèle probabiliste, avec les méthodes reposant sur des scores d'association statistiques (« Param. / Assoc. »). L'objectif principal de la méthode, donnant ou non un rôle symétrique aux différentes langues (« Sym. / Asym. ») – ce qui concorde également avec le type d'alignement produit en sortie du programme (bilingue ou multilingue) – est également fourni.

3.3 Typologie pour l'alignement multilingue

Nous proposons ici un inventaire typologique des méthodes identifiées pour l'alignement multilingue.

Comme nous venons de le voir, il est possible d'établir une ligne de partage entre des méthodes que nous appelons *symétriques* [20] et celles que nous appelons *asymétriques*.

Les méthodes symétriques relèvent d'une logique de la combinaison de plusieurs modèles d'alignement bilingues pour construire un nouveau modèle bilingue enrichi par l'information fournie par différentes paires de langues. Nous en donnons un diagramme de principe en Figure 3.4.

18. Remarquons que, dans le cas de la symétrie faisant intervenir le mélange de deux distributions (voir section 2.2.1), cela relève aussi d'une logique de la combinaison.

19. Les paires de segments d'intérêt pour la traduction automatique statistique se trouvent faire partie de cette tribu, ou σ-algèbre.

20. Encore une fois, il faut se méfier de ce terme que nous utilisons, faute de mieux, car il peut induire une confusion avec la « symétrie » des relations binaires d'une part, mais aussi la « symétrie »

Réf.	Unités alignées	Nb. langues	Param./ Assoc.	Sym./ Asym. (objectif)
(Simard, 1999)	phrases	3	assoc.	asym. (bilingue)
(Filali et Bilmes, 2005)	mots	4	param.	asym. (bilingue)
(Kumar *et al.*, 2007)	mots	N	param.	asym. (bilingue)
(Lardilleux *et al.*, 2011)	groupes de mots	N	assoc.	sym. (multilingue)
(Mayer et Cysouw, 2012)	mots	N	assoc.	sym. (multilingue)
(Östling, 2014)	mots	N	param.	sym. (multilingue)

TABLE 3.1 – Synthèse de l'état de l'art pour l'alignement de plus de deux langues

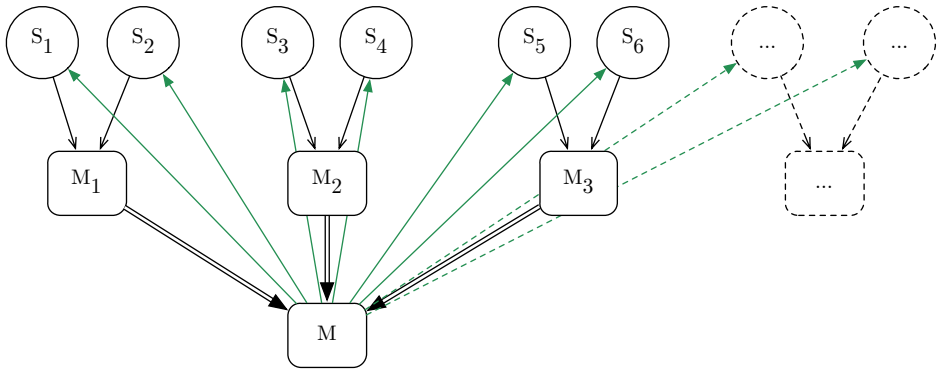


FIGURE 3.4 – Schéma de principe pour la combinaison de modèles bilingues dans une approche asymétrique. Les modèles M_1, M_2 et M_3 sont appris sur les systèmes (langues) S_1, S_2, S_3, S_4, S_5 et S_6, et combinés dans un modèle bilingue M qui opère (flèches vertes) éventuellement ensuite sur plusieurs des paires initiales.

Par contraste, les méthodes symétriques tentent de capturer *simultanément* l'information disponible dans les corpus multilingues, selon le schéma donné en Figure 3.5.

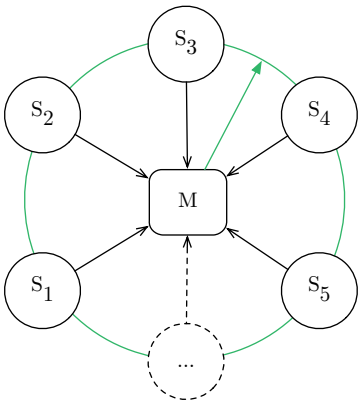
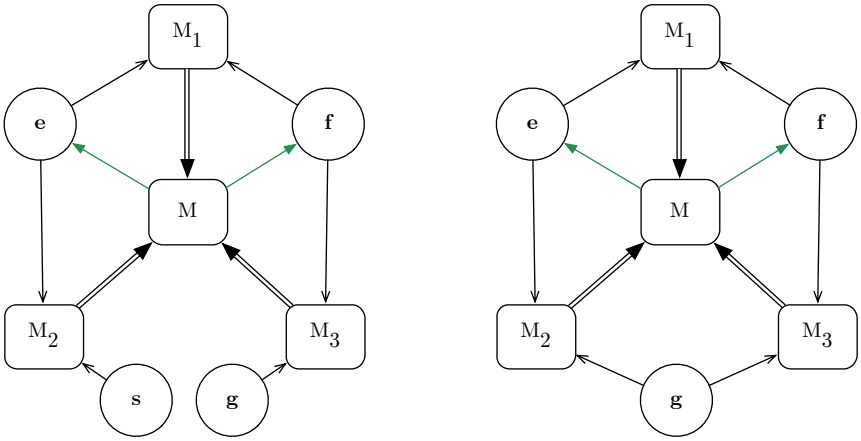


FIGURE 3.5 – Schéma de principe pour l'apprentissage symétrique d'un modèle d'alignement multilingue.

Ce partage cache pourtant une complexité sous-jacente assez grossièrement synthétisée par le critère paramétrique/associatif. Si l'on met en regard (Figure 3.6) la méthode présentée par (Filali et Bilmes, 2005) avec celle présentée par (Kumar *et al.*, 2007), les schémas de principes se donnent comme quasiment identiques [21].



(Filali et Bilmes, 2005) **(Kumar et al., 2007)**

FIGURE 3.6 – Comparaison des diagrammes de principe pour deux méthodes non comparables

Or, la première méthode (Filali et Bilmes, 2005) fait intervenir une réécriture du modèle génératif et correspond à l'apprentissage d'une loi jointe (sous des hypothèses simplificatrices fortes) faisant intervenir plus de deux langues, quand la seconde (Kumar *et al.*, 2007) découple la contribution de chaque langue supplémentaire, et ceci sans jamais

des alignements qui, suivant les cas, peut revêtir plusieurs sens.

21. Et si l'on se souvient que les meilleurs résultats de la première méthode s'obtiennent pour $\mathbf{s} = \mathbf{g}$, les schémas sont alors parfaitement identiques.

apprendre les paramètres d'un modèle génératif faisant intervenir plus de deux langues. Nous proposons donc en Figure 3.7 une hiérarchie plus fine des méthodes identifiées.

Concernant ce que nous appelons « sortie bilingue » et « sortie multilingue » pour rendre compte des méthodes de nature symétrique ou asymétrique, il s'agit de la sortie *naturelle*, étant donné qu'il est bien sûr possible de produire une sortie multilingue en additionnant des sorties bilingues provenant de méthodes asymétriques.

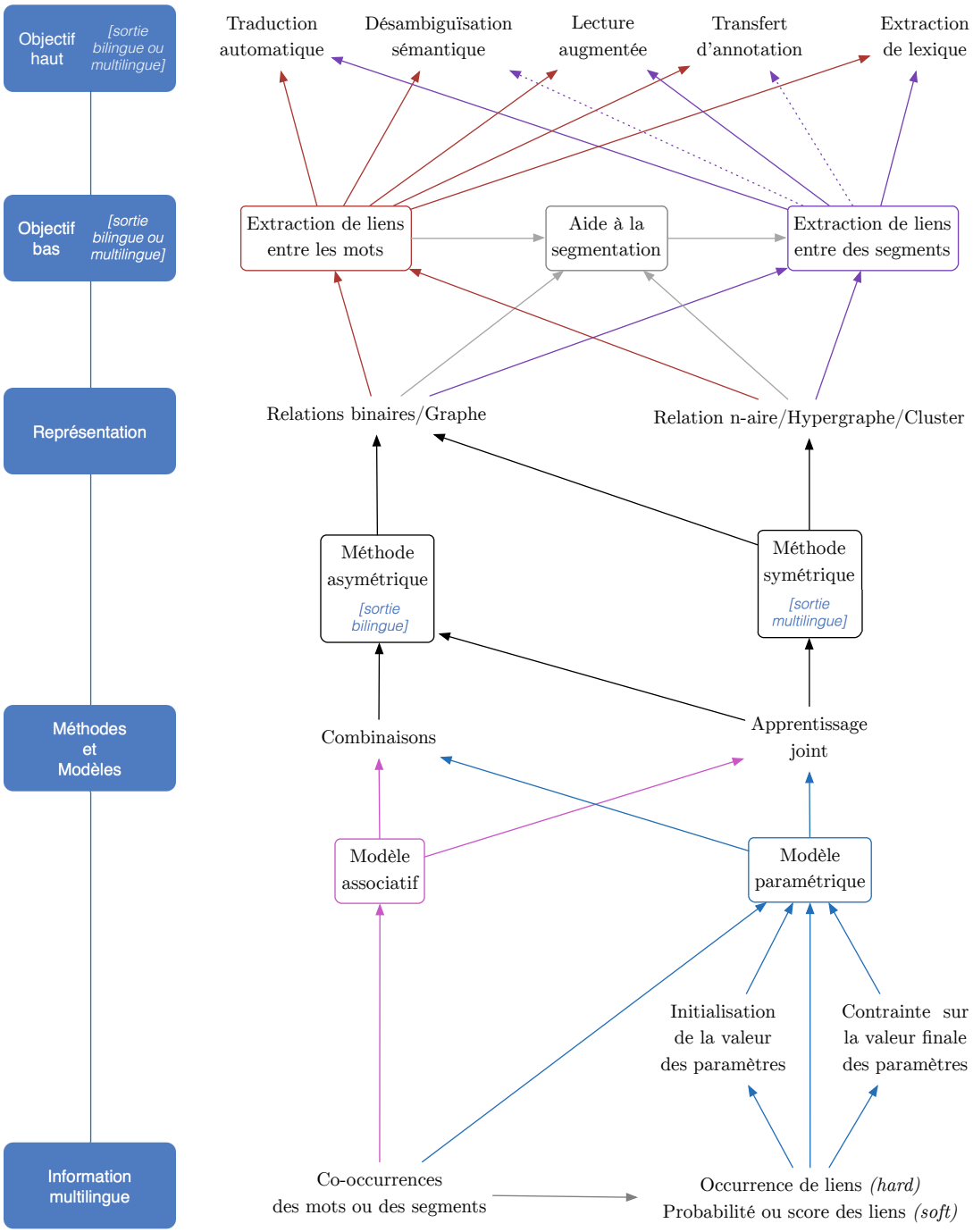


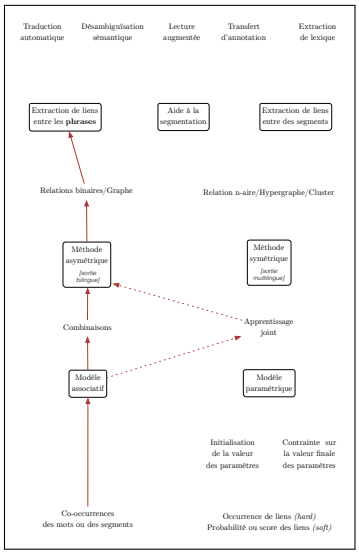
FIGURE 3.7 – Typologie pour l'alignement multilingue

Ce que nous appelons, dans ce diagramme, « information multilingue » peut correspondre soit aux données brutes contenues dans un corpus multilingue, soit à des

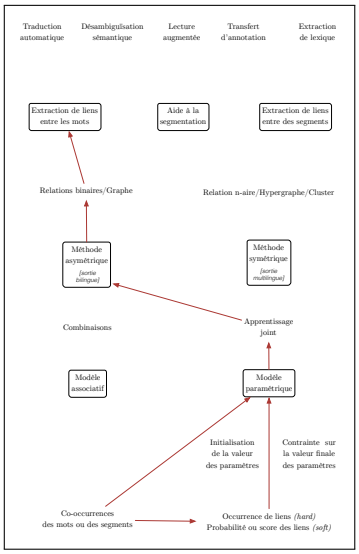
informations extraites à partir de modèles monolingues ou bilingues. Il peut s'agir dans ce cas d'occurrences de lien, ou de probabilités pour la présence de ces liens. Dans le cas d'un modèle paramétrique, cette information pourra être intégrée sous forme de contrainte pendant l'apprentissage.

Par « combinaison », nous entendons toutes les opérations de sélection des liens (ou des probabilités de présence de lien dans le cas paramétrique). Il s'agira par exemple d'heuristiques du type intersection-expansion, ou bien d'interpolation de probabilités provenant de traitements différents.

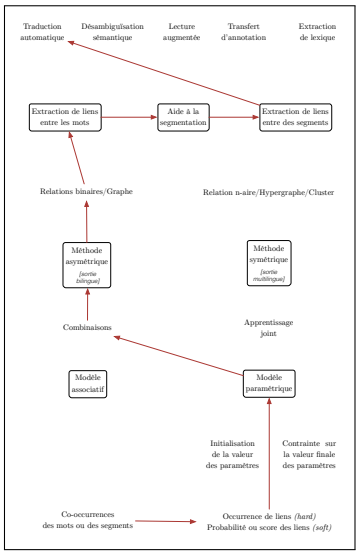
Nous donnons en Figure 3.8 les chemins, dans cette hiérarchie typologique, des différents articles de la Table 3.1.



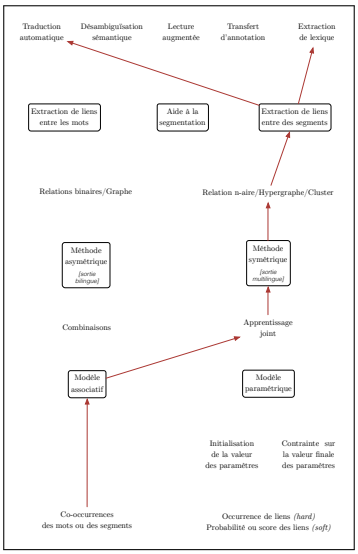
(a) Simard



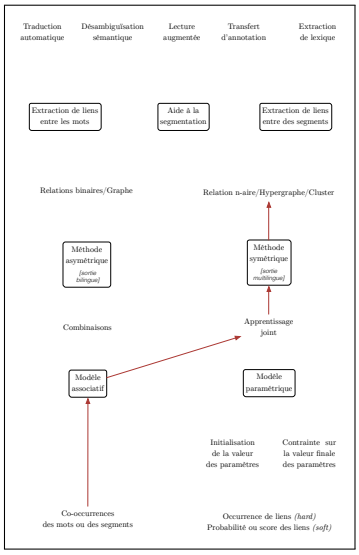
(b) Filali



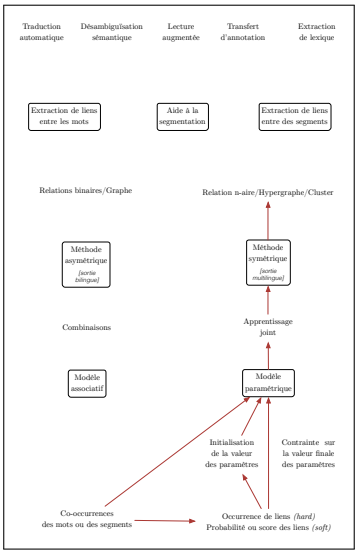
(c) Kumar



(d) Lardilleux



(e) Mayer



(f) Östling

FIGURE 3.8 – Comparaison typologique

Chapitre 4

Expériences

Les questions soulevées par l'étude théorique des problèmes d'alignements dans un cadre multilingue sont nombreuses et l'espace pour le travail, très ouvert. Dans ce chapitre, nous rendons compte de la direction de travail que nous avons choisi d'emprunter dans un premier temps, et du travail qui a été réalisé dans ce sens.

4.1 Direction de travail

La question de la transitivité des relations d'alignement, et le bruit qu'elle peut induire (voir en particulier la section 3.2.1), nous a semblé très tôt une question particulièrement intéressante. D'une certaine manière, la question de la transitivité des liens formalise partiellement une question plus vaste : étant donné un alignement entre deux textes mutuellement traduits, quelles informations utiles, et de nature à améliorer l'alignement initial, peut nous apporter un nouvel alignement de l'un de ces textes avec une autre traduction ?

Dans cette optique, qui se rattache à ce que nous avons appelé plus haut une approche par *combinaison* d'alignements bilingues, le cas particulier de l'alignement de deux traductions d'un texte dans la même langue nous a semblé, en réduisant un facteur de complexité, prometteur. Si les deux textes cibles appartiennent à la même langue, il semble en effet plus simple de déterminer des critères de similarité entre certains mots. En admettant en outre, ce qui semble raisonnable, la transitivité des liens – c'est-à-dire, encore une fois, l'idée que de l'information « passe » dans ces liens – ceci permettrait, par contraposition, d'invalider certains liens créant des chemins entre des mots ne présentant aucune similarité.

Un exemple construit à partir d'une paire de traductions provenant du corpus Europarl (Koehn, 2005) et d'une hypothèse de traduction plus littérale, illustre (Figure 4.1) cette intuition.

Ainsi, en notant $\mathcal{E}$ un texte source, et $\mathcal{F}$ et $\mathcal{G}$ deux traductions de ce texte dans la même langue, nous nous proposons d'étudier les moyens d'améliorer l'alignement $A^{\mathcal{E}\leftrightarrow\mathcal{F}}$ en tirant parti de la connaissance de l'alignement $A^{\mathcal{E}\leftrightarrow\mathcal{G}}$. Dans la suite de ce travail, nous faisons le choix de l'anglais pour la langue source (texte $\mathcal{E}$), et du français pour la langue cible (textes $\mathcal{F}$ et $\mathcal{G}$).

4.2 Constitution d'un corpus d'étude

Les travaux pionniers de Brown *et al.* (1990) en matière d'alignement statistique bilingue au niveau des mots s'appuyaient sur le Hansard canadien, constitué par la transcription des débats du parlement fédéral canadien et tenu en anglais et en français.

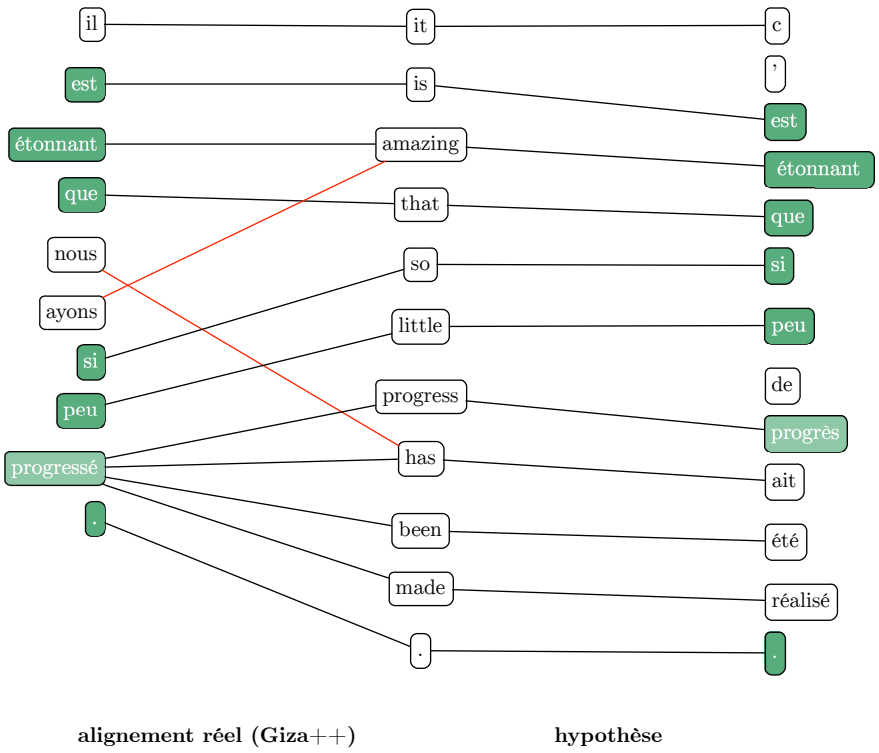


FIGURE 4.1 – Alignements de deux traductions dans la même langue et liens erronés sur des mots apparaissant seulement dans l'une des deux paraphrases

Aujourd'hui, les travaux qui requièrent des corpus multilingues font majoritairement usage de transcriptions de débats ou de résolutions au Parlement européen [1] ou aux Nations Unies [2]. De nombreux auteurs ont également recours à des corpus constitués à partir de textes religieux qui ont la particularité d'être traduits dans de très nombreuses langues. C'est le cas des travaux déjà cités de Simard (1999), Mayer et Cysouw (2012) et Östling (2014).

Pour autant, la particularité de l'axe de travail choisi – ainsi que notre goût modéré pour la prose parlementaire ou confessionnelle – nous ont conduits à nous intéresser plutôt à des textes de la littérature. Après avoir un temps souhaité constituer un corpus de traductions multiples en français de pièces de William Shakespeare, nous avons finalement constitué un corpus autour de deux traductions en français d'*Alice's Adventures in Wonderland* de Lewis Carroll. Nous détaillons les différentes étapes de ce travail dans la présente section.

4.2.1 Travailler sur des textes littéraires

Les textes littéraires ont la réputation (justifiée) d'être difficiles à aligner automatiquement. Ils possèdent également la propriété rare d'être souvent, du moins pour les textes qui font référence, disponibles dans plusieurs traductions pour une langue donnée. Ces deux aspects combinés en font des candidats très naturels pour un travail cherchant à améliorer la qualité de l'alignement mot à mot de deux textes en tirant parti de l'infor-

1. Voir en particulier `http://www.statmt.org/europarl/`, et (Koehn, 2005) cité plus haut.
2. Voir par exemple `http://www.euromatrixplus.net/multi-un/`, ou encore `http://www.uncorpora.org/`.

mation supplémentaire fournie par une traduction supplémentaire dans la langue cible.

4.2.2 Alignement phrase à phrase

En préalable à tout travail d'alignement à un niveau plus fin, il faut disposer d'un corpus parallèle aligné au niveau des phrases. La construction automatique particulière d'un corpus de textes littéraires et les difficultés qui lui sont afférentes font l'objet d'une étude exhaustive dans (Yu *et al.*, 2012). Les auteurs de cet article utilisent pour cette tâche un classifieur à maximum d'entropie appris sur des alignements « 1-1 » très fiables produits par la méthode d'alignement décrite dans (Moore, 2002). De cette manière, il est possible d'atteindre une bonne couverture en rappel, tout en conservant des niveaux très élevés de précision. Pour notre étude, il est également nécessaire de disposer, au moins pour une partie du corpus, de plusieurs variantes de traduction dans une même langue.

De William Il a semblé tentant dans un premier temps de constituer un corpus à partir de pièces de Shakespeare. De multiples traductions existent en français, ce qui nous importe principalement à ce stade, mais aussi dans de nombreuses autres langues, ouvrant des perspectives diverses d'extension de l'exploration méthodologique. *The sky's the limit.*

Malheureusement, il est vite apparu qu'un alignement automatique phrase à phrase présentait, dans le cas de Shakespeare, des difficultés supplémentaires à celles habituellement rencontrées pour cette tâche. Tout d'abord le texte source de la traduction peut varier suivant les traductions [3]. D'autre part, le vers shakespearien non rimé [4] est souvent rendu par de la prose en français, ce qui rend l'alignement en phrase particulièrement délicat. À noter aussi que le poète passe très librement et sans crier gare du vers rimé à la prose et de la prose au vers non rimé. D'ailleurs, l'alignement des vers – qui pourraient passer pour des unités faciles à identifier et donc aligner – peut se révéler éminemment retors car la restitution des vers, par la recherche d'une analogie de métrique et de prosodie, impose souvent au traducteur des contorsions diverses qui conduisent à des traductions plus libres, et partant, plus difficile à faire correspondre à la source.

Ajoutons enfin que l'imprimerie du XVII^{e} siècle montre une fâcheuse tendance à substituer des caractères en fonction de leur usure (par exemple le « u » et le « v »), ajoutant à l'instabilité des graphies de l'*Early Modern English*, ou anglais élisabéthain, au sein duquel s'amorce une standardisation de l'orthographe, de la grammaire, et de la prononciation, qui donneront naissance à l'anglais moderne.

À Lewis Pour toutes ces raisons, et puisque l'alignement phrase à phrase ne constituait pas à proprement parler l'objet de ce travail, il a semblé plus raisonnable de chercher dans d'autres directions. Force est de constater une grande rareté des sources poly-traduites pour lesquelles un alignement phrase à phrase a déjà été réalisé. Un site intéressant [5] propose de telles ressources, mais sans que l'on y trouve néanmoins plusieurs traductions pour une même langue cible.

3. Il existe, pour l'édition des pièces de Shakespeare, plusieurs versions, en particulier différents « Quarto » qui sont les éditions les plus anciennes mais peu fiables, le « First Folio » (1623), le « Second Folio » (1632), ainsi que d'autres versions postérieures.

4. En anglais *blank verse* ; à distinguer du vers libre car le *blank verse* chez Shakespeare est soumis à la métrique du pentamètre ïambique, un vers de dix syllabes composé de cinq ïambes (couples de deux syllabes respectivement non accentuée et accentuée).

5. `http://www.farkastranslations.com/bilingual_books.php`.

C'est le corpus « Alice au pays des mesures » [6] (réalisé sous la direction de Serge Fleury) qui nous a fourni un point de départ avec deux traductions françaises d'*Alice's Adventures in Wonderland* de Lewis Carroll (1865) ; la première est collective et non signée (2004), la seconde est celle d'Henri Bué (1869).

4.2.3 Plongement

Afin de pouvoir utiliser certains algorithmes d'apprentissage non supervisé (voir section 2.1.4), il est nécessaire de plonger le texte source anglais (`alice.en`) ainsi que la traduction que l'on veut aligner (`alice.fr1` ou `alice.fr2`) dans un corpus plus vaste afin d'atteindre une taille qui soit significative statistiquement.

Nous avons utilisé pour cela un corpus parallèle de textes français et anglais, essentiellement du dix-neuvième siècle, construit par François Yvon lors d'une collaboration avec le metteur en scène Jean-François Perret [7]. Un système de traduction automatique statistique produisant des traductions du *Walden* d'Henry David Thoreau avait été développé à cette occasion, nécessitant l'alignement de nombreux textes littéraires (nous renvoyons ici de nouveau à (Yu *et al.*, 2012), cité plus haut) dont nous avons pu ici bénéficier. Nous donnons en Table 4.1 la liste des œuvres de ce corpus.

4.2.4 Préparation pour l'alignement mot à mot

Quelques traitements supplémentaires doivent être réalisés avant de pouvoir procéder à l'apprentissage des liens d'alignement entre les mots.

Il faut d'abord concaténer les différents textes du corpus dans un unique fichier par langue. Nous détaillons plus loin (section 4.3.2) les trois concaténations réalisées.

Une segmentation en *tokens* (tokénisation) est également requise, ainsi qu'un filtrage après tokénisation des phrases trop longues ou trop courtes [8] – nous choisissons ici d'exclure les phrases de moins de 2 mots et de plus de 100 mots – couplé avec divers nettoyages des textes (suppression d'éventuelles balises, de répétitions du caractère d'espacement, etc.). À la suite de ces traitements, nous disposons d'un corpus dont nous donnons quelques chiffres en Table 4.2.

4.3 Alignement mot à mot bilingue

4.3.1 Programmes utilisés

`MGiza++` Une implémentation des différents modèles IBM et HMM (voir section 2.1.3) qui constituent toujours, en matière d'alignement mot à mot, une référence état-de-l'art, est fournie par le programme `MGiza++` [9] (Gao et Vogel, 2008) que nous utilisons ici. Ce programme est une ré-implémentation du programme `Giza++` [10] (Och et Ney, 2003) qui lui adjoint, en particulier, la possibilité d'effectuer les calculs de manière parallèle.

Ainsi que nous l'indiquions en section 2.1.2, il est d'usage d'entraîner ces modèles dans les deux sens, de la langue source vers la langue cible et réciproquement, ce qui permet

6. `http://www.tal.univ-paris3.fr/mkAlign/`.

7. *Re : Walden* a été présenté au théâtre de la Colline début 2014, ainsi que, dans une forme antérieure, au festival Paris-Villette en juin 2012 et au Festival d'Avignon en juillet 2013. Ce travail a également donné naissance à une installation au Fresnoy – Studio national des arts contemporains, début 2013.

8. Pour *Alice*, cela implique de filtrer simultanément les trois textes.

9. `http://www.kyloo.net/software/doku.php/mgiza:overview` (lien déjà donné en note 7 de la page 27).

10. `https://code.google.com/p/giza-pp/`

Auteur	Titre	Année de publication
Daniel Defoe	*Robinson Crusoe*	1719
Jonathan Swift	*Gulliver's Travels*	1726
Richard Walter	*Anson's Voyage Round the World*	1748
Jean-Jacques Rousseau	*Les Confessions*	1782-1789
Jane Austen	*Sense and Sensibility*	1811
Jane Austen	*Emma*	1815
Johann Wolfgang von Goethe	*Voyage en Italie*	1816-1817
James Fenimore Cooper	*The Last of the Mohicans*	1826
Washington Irving	*A Tour on the Prairies*	1832
Ralph Waldo Emerson	*Nature*	1836
Stendhal	*Mémoires d'un touriste*	1838
Charles Darwin	*The Voyage of the Beagle*	1839
Nathaniel Hawthorne	*The Birth-Mark*	1843
Théophile Gautier	*Voyage en Espagne*	1843
Charles Dickens	*Pictures from Italy*	1846
Charlotte Brontë	*Jane Eyre*	1847
William Makepeace Thackeray	*Vanity Fair*	1847-1848
Francis Parkman, Jr.	*The Oregon Trail*	1849
Henry David Thoreau	*On the Duty of Civil Disobedience*	1849
Nathaniel Hawthorne	*The Scarlet Letter*	1850
Herman Melville	*Bartleby, the Scrivener*	1853
Henry David Thoreau	*Walden*	1854
Richard Francis Burton	*Personal Narrative of a Pilgrimage to El-Medinah and Meccah*	1855
Henry David Thoreau	*A Plea for Captain John Brown*	1859
Jules Verne	*Cinq semaines en ballon*	1863
Jules Verne	*Voyage au centre de la Terre*	1864
Lewis Carroll	***Alice's Adventures in Wonderland***	**1865**
Jules Verne	*De la Terre à la Lune*	1865
Mark Twain	*The Innocents Abroad*	1869
Henry Morton Stanley	*How I Found Livingstone*	1872
Jules Verne	*Le Tour du monde en quatre-vingts jours*	1872
Robert Louis Stevenson	*Travels with a Donkey in the Cévennes*	1879
Robert Louis Stevenson	*Henry David Thoreau : His Character and Opinions*	1880
Mark Twain	*Life on the Mississippi*	1883
Henry James	*A Little Tour in France*	1884
Pierre Loti	*Au Maroc*	1890
Mary Kingsley	*Travels in West Africa*	1897
George Gissing	*By the Ionian Sea*	1901
Pierre Loti	*L'Inde sans les Anglais*	1903
Pierre Loti	*La mort de Philæ*	1909
John Muir	*Travels in Alaska*	1915
Alexandra David-Néel	*Voyage d'une Parisienne à Lhassa*	1927
Thor Heyerdahl	*Kon-Tiki*	1948

TABLE 4.1 – Listes des œuvres du corpus `walden`.

	# phrases	# mots		
		en		fr
corpus `walden` (hors `alice`)	129 112	2 802 035		2 906 727
		en	fr_1	fr_2
`alice`	686	21 499	22 954	23 787

TABLE 4.2 – Caractéristiques du corpus.

d'obtenir, après symétrisation, des alignements de meilleure qualité. Les alignements que nous utilisons correspondent aux sorties symétrisées de trois modèles IBM 4 appris sur les trois corpus.

`postCAT` Afin de pouvoir, dans la suite du travail, contraster les alignements produits par `MGiza++`, nous avons également appris différents modèles HMM sous la contrainte d'une régularisation des probabilités a posteriori des liens d'alignement (voir section 2.2.1), à l'aide du programme `postCAT`[11] (Graça *et al.*, 2007). Ces contraintes garantissent, en espérance, la symétrie (agrément dans les deux sens) ou la bijectivité (type « 1-1 ») des alignements produits.

4.3.2 Méthodologie

En pratique, nous travaillons sur trois variantes du corpus, correspondant respectivement à la première traduction française, à la seconde, et à la juxtaposition des deux traductions. Pour fixer les idées, nous donnons en Table 4.3 le schéma de concaténation de ces trois variantes.

Variante du corpus	Fichiers concaténés
`alice.1`	`alice.en` `alice1.fr` corpus parallèle `walden`
`alice.2`	`alice.en` `alice2.fr` corpus parallèle `walden`

`alice.1and2`	`alice.en` `alice1.fr`
	`alice.en` `alice2.fr`
	corpus parallèle `walden`

TABLE 4.3 – Les trois corpus de travail.

Les deux premières variantes sont naturelles et permettent d'obtenir des alignements pour chaque traduction d'*Alice* avec le texte source. La dernière variante permet, elle, de réaliser un alignement qui servira de *baseline* à nos efforts pour exploiter l'information contenue dans le second alignement pour améliorer le premier. En versant, en effet, la

11. `http://www.seas.upenn.edu/~strctlrn/CAT/CAT.html`

seconde traduction (tout en dupliquant le texte de Carroll), nous exploitons déjà durant l'apprentissage, de manière certes grossière, une information supplémentaire de nature à améliorer le premier alignement.

4.4 Étude des alignements combinés

4.4.1 Notations

Nous disposons donc d'un corpus $\mathcal{C}(\mathcal{E}, \mathcal{F}, \mathcal{G}) = \{(\mathbf{e}^{(n)}, \mathbf{f}^{(n)}, \mathbf{g}^{(n)}), n = 1 \dots n_D\}$ de n_D triplets de phrases. Lorsqu'il n'y a pas d'ambiguïté, nous omettons l'indexation (n) des triplets pour faciliter la lecture.

Pour chaque triplet, nous disposons désormais également d'alignements, au niveau des mots et sous forme de matrices, $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ et $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$. Nous notons l'alignement combiné [12] de la manière suivante :

$$A^{\mathbf{f}\leftrightarrow\mathbf{e}\leftrightarrow\mathbf{g}} = A^{\mathbf{e}\leftrightarrow\mathbf{f}} \cup A^{\mathbf{e}\leftrightarrow\mathbf{g}},$$

et avec $\mathbf{e} = e_1^I$, $\mathbf{f} = f_1^J$ et $\mathbf{g} = g_1^K$, nous écrivons les termes des matrices binaires d'alignement selon :

$$A^{\mathbf{e}\leftrightarrow\mathbf{f}} = (a_{ij})_{1\leq i\leq I, 1\leq j\leq J} \quad \text{et} \quad A^{\mathbf{e}\leftrightarrow\mathbf{g}} = (b_{ik})_{1\leq i\leq I, 1\leq k\leq K}.$$

Ces alignements pourront éventuellement être obtenus de manière hétérogène [13].

4.4.2 Visualisation

Pour commencer l'étude des données, nous éprouvons rapidement le besoin de visualiser les alignements. Il existe des outils [14] qui affichent sous forme de matrices les alignements produits dans des formats standards. Nous avons également pu utiliser avec profit un script [15] produisant le code LaTeX (TikZ) nécessaire à la création du diagramme d'alignement. Il est enfin possible d'utiliser un outil d'alignement manuel comme `Yawat` (Germann, 2008) afin de visualiser les alignements selon un paradigme de coloration des mots alignés plutôt que de représentation des arêtes.

Lorsqu'il s'agit de visualiser de manière dynamique des alignements combinés, un grand chagrin emplit néanmoins l'âme. C'est pourquoi nous avons entrepris de développer un outil de visualisation et d'analyse de ces alignements. Le programme, écrit en Python et élaboré à partir des librairies `NetworkX` [16] et `matplotlib` [17], traite les alignements comme des graphes et peut en donner une représentation à la volée comme celle présentée en Figure 4.2. Cette représentation peut en outre inclure visuellement certaines informations sur la nature des liens.

4.4.3 Mesures d'agrément

La première question qui se pose lorsque l'on cherche à exploiter l'alignement $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$ pour améliorer $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ relève de la mesure de leur accord ou de leur désaccord. Intuitivement, si les deux alignements « disent » toujours la même chose – ou toujours des choses

12. La notation diffère ici d'une précédente notation, $A^{\mathbf{e},\mathbf{f},\mathbf{g}} = A^{\mathbf{e}\leftrightarrow\mathbf{f}} \cup A^{\mathbf{f}\leftrightarrow\mathbf{g}} \cup A^{\mathbf{g}\leftrightarrow\mathbf{e}}$, en l'absence d'alignement entre $\mathbf{f}$ et $\mathbf{g}$.

13. $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ pourra par exemple être obtenu avec `MGiza++` et $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$ avec `postCAT`, ou une configuration différente de `MGiza++`.

14. Par exemple, `picaro` : `https://code.google.com/p/picaro/`

15. Thank you for this, and much more, Nicolas Pécheux !

16. `https://networkx.github.io/`

17. `http://matplotlib.org/`

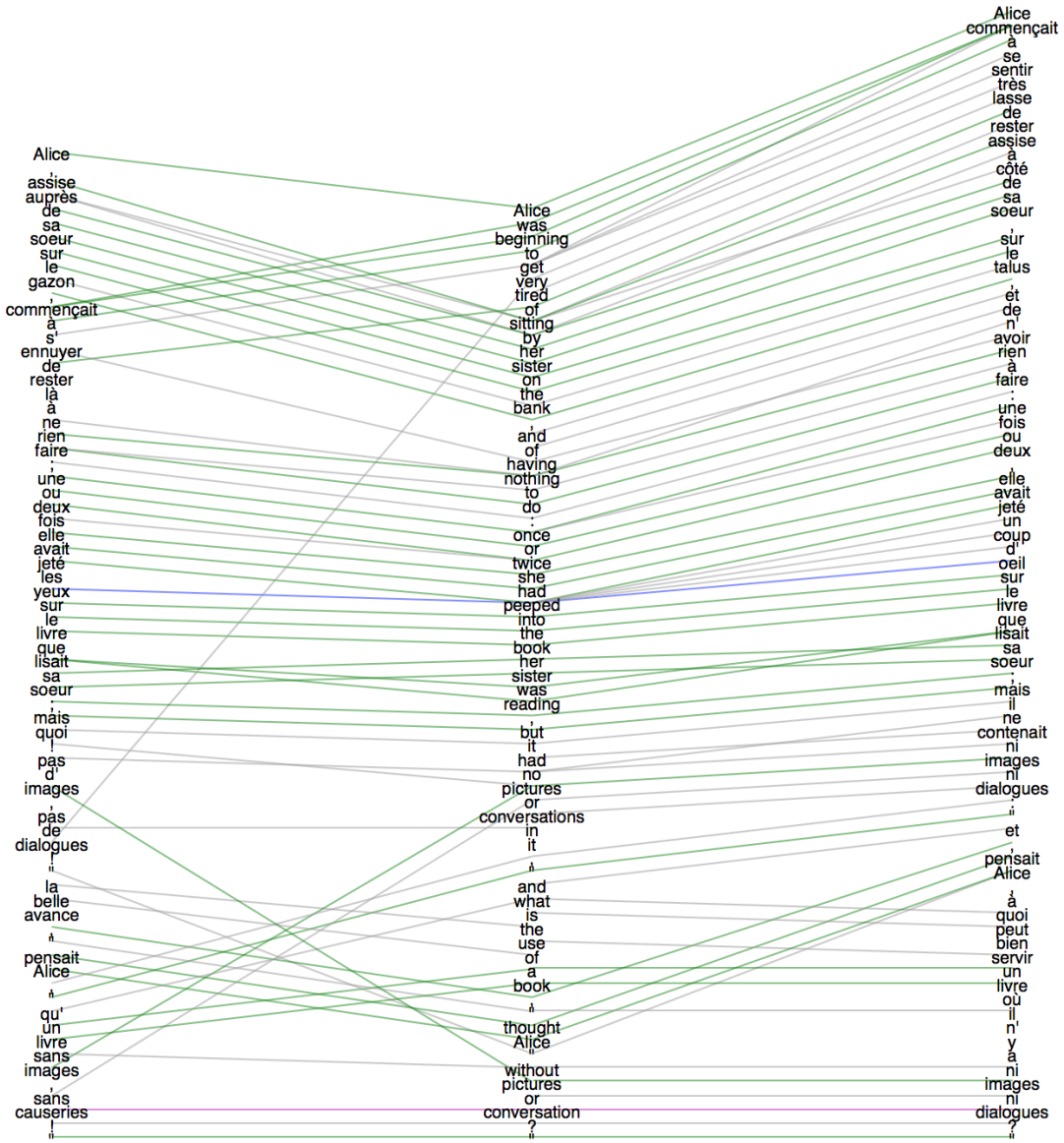


FIGURE 4.2 – Visualisation de deux alignements produits par `MGiza++` et combinés.

différentes – alors il y a peu à attendre de la direction de travail choisie. C'est donc la première analyse que nous avons réalisée.

4.4.3.1 Le point de vue de la source

Nous introduisons deux fonctions de score pour les liens entre mots source et cible. La première est définie sur les mots de la première cible :

$$\sigma_a(e_i, f_j) = a_{ij} \ \max_k(b_{ik} \ \mathrm{sim}(f_j, g_k))$$

et la seconde sur les mots de la seconde cible :

$$\sigma_b(e_i, g_k) = b_{ik} \ \max_j(a_{ij} \ \mathrm{sim}(f_j, g_k))$$

avec $\mathrm{sim}(\cdot, \cdot)$ une fonction de similarité lexicale entre deux mots. En pratique, nous utiliserons les similarités assez rudimentaires suivantes :

- $\mathrm{sim}_1(u, v) = \begin{cases} 1 & \text{si } u \text{ et } v \text{ ont des formes identiques} \\ 0 & \text{sinon.} \end{cases}$
- $\mathrm{sim}_2(u, v) = \begin{cases} 1 & \text{si } u \text{ et } v \text{ ont des formes identiques} \\ 0.7 & \text{si } u \text{ et } v \text{ ont seulement le même lemme} \\ 0 & \text{sinon.} \end{cases}$
- $\mathrm{sim}_3(u, v) = \begin{cases} 1 & \text{si } u \text{ et } v \text{ ont des formes identiques} \\ 0.7 & \text{si } u \text{ et } v \text{ ont seulement le même lemme} \\ 0.5 & \text{si } u \text{ et } v \text{ ont seulement des lemmes synonymes} \\ 0 & \text{sinon.} \end{cases}$

Lorsque le lemmatiseur [18] n'a pas été en mesure de désambiguïser la forme et retourne une liste de lemmes possibles, nous adaptons les critères de similarité (existence d'un lemme commun ou d'une paire de lemmes synonymes).

La fonction σ_a (resp. σ_b) quantifie, lorsqu'un lien existe entre le mot source et le mot cible, l'agrément de ce lien avec un autre lien présent dans le second (resp. premier) alignement.

Nous définissons ensuite un score d'agrément α par mot source selon :

$$\alpha(e_i) = \begin{cases} \frac{\sum_j \sigma_a(e_i,f_j)+\sum_k \sigma_b(e_i,g_k)}{\sum_j a_{ij}+\sum_k b_{ik}} & \text{si } \sum_j a_{ij} + \sum_k b_{ik} \geq 0 \\ 1 & \text{sinon.} \end{cases}$$

avec $\sum_j a_{ij} + \sum_k b_{ik}$ le nombre de liens – vers des mots de la première ou de la seconde cible – incidents au mot source e_i.

Nous donnons en Table 4.4 la moyenne et l'écart-type de ces scores d'agrément par mot source pour les 686 paires de phrases d'`alice`, suivant différentes combinaisons d'alignements. Le premier alignement est toujours celui réalisé avec `MGiza++` pour la première traduction. Le second correspond à l'alignement appris pour la seconde traduction, soit avec `MGiza++` également, ou bien avec `postCAT` en appliquant la contrainte de bijectivité ou de symétrie. Nous donnons en bas de la table, pour contraster, les mêmes mesures

18. Nous utilisons le TreeTagger (Schmid, 1994) (`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`) et le wrapper PYTHON écrit par Laurent Pointal (`http://perso.limsi.fr/pointal/dev:treetaggerwrapper`).

d'agrément pour un alignement combiné construit avec deux exemplaires de la première traduction. La synonymie est testée à l'aide d'un dictionnaire que nous avons construit à partir du Wiktionnaire français [19] dont un *dump* a été traité et converti au format XML [20].

alignements	sim_1		sim_2		sim_3	
en-fr_1/en-fr_2	μ	σ	μ	σ	μ	σ
`MGiza`/`MGiza`	**0,508**	0,485	**0,524**	0,480	**0,529**	0,477
`MGiza`/`pCAT` (bij)	0,467	0,486	0,482	0,482	0,487	0,480
`MGiza`/`pCAT` (sym)	0,478	0,483	0,494	0,478	0,498	0,475
en-fr_1/en-fr_1	μ	σ	μ	σ	μ	σ
`MGiza`/`pCAT` (bij)	0,768	0,395	0,769	0,395	0,769	0,394
`MGiza`/`pCAT` (sym)	0,792	0,375	0,792	0,375	0,792	0,375

TABLE 4.4 – Moyenne et écart-type des scores d'agrément $\alpha(e_i)$ par mot source.

Ceci nous permet de répondre à la question soulevée plus haut. Il n'y a pas d'accord ou de désaccord systématique entre les deux alignements, et l'ordre de grandeur est celui d'un accord dans la moitié des cas. Les valeurs de l'écart-type révèlent par ailleurs que l'immense majorité de ces scores vaut soit 0, soit 1.

D'autre part, le changement de mesure de similarité a un impact non négligeable mais modeste. Dans le cas (bas de table) d'un alignement combiné avec deux fois la même traduction française, le changement de mesure de similarité est quasi indécelable, ce qui est conforme à l'intuition [21]. Ces résultats permettent également d'exhiber le fait que la contrainte de symétrie induit, en sortie de `postCAT`, des alignements plus proches de `MGiza++` que la contrainte de bijectivité.

4.4.3.2 Le point de vue de la cible

Une intuition forte, et qui adopte le point de vue des mots cibles plutôt que des mots sources, peut s'exprimer de la manière suivante : des mots identiques (ou similaires) dans les deux traductions devraient être alignés avec les mêmes mots sources. Par conséquent, il faudrait pouvoir compter les liens du premier alignement qui sont en quelque sorte ajoutés par rapport au second alignement, et réciproquement, afin de pouvoir accéder à une information complémentaire à celle obtenue pour les mots sources.

La difficulté tient au fait que nous n'avons pas à notre disposition de moyen simple de réaliser automatiquement un alignement monolingue. De plus, l'éventuelle répétition de certains mots et l'absence de garantie, le cas échéant, que l'ordre d'apparition de mots répétés équivalents soit le même dans les deux traductions, **f** et **g**, complique encore la tâche.

19. `http://fr.wiktionary.org`

20. `http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html`, laboratoire CLLE-ERSS (CNRS-Université de Toulouse-Le Mirail).

21. Pour qu'une similarité plus fine fasse une différence lorsque les deux textes cibles sont identiques, il faut un désaccord d'alignement qui mette en relation des mots ayant par exemple un lemme commun mais correspondant à des parties différentes de la phrase.

Nous construisons, pour contourner ces problèmes, des clusters de mots pour **f** et **g** au sens du critère de similarité $\mathrm{sim}(\cdot,\cdot)$ défini dans la section précédente. Pour chaque cluster de **f**, nous sommons tous les liens incidents aux mots de ce cluster, et identifions les mots sources atteints par ces liens. Nous identifions le cluster correspondant dans **g**, s'il existe, et procédons de même. Nous disposons alors des mots sources « atteints » à la fois par les mots du premier cluster dans **f** et de son pendant dans la seconde traduction **g**.

Ceci nous permet de calculer une valeur approchée du nombre de liens de $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ qui agréent avec $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$, du nombre de liens que le premier alignement ajoute par rapport au second, et le nombre de liens qui manquent.

Formellement, soit $\mathcal{C}_1^L = \{\mathcal{C}_1, \ldots, \mathcal{C}_l, \ldots, \mathcal{C}_L\}$ l'ensemble des clusters de mots construits à partir de la séquence de mots f_1^J pour satisfaire :

$$\begin{cases} \forall u, v \in \mathcal{C}_l & \mathrm{sim}(u,v) > 0 \\ \forall u \in \mathcal{C}_{l_1}\ \forall v \in \mathcal{C}_{l_2} & \mathrm{sim}(u,v) = 0 \end{cases}$$

Nous construisons de la même manière un ensemble de clusters $\mathcal{D}_1^M$ sur **g**. Pour chaque cluster $\mathcal{C}_l$, le cluster équivalent $\mathcal{D}_m$ doit satisfaire à :

$$\exists u \in \mathcal{C}_l\ \exists v \in \mathcal{D}_m \quad \text{t.q.} \quad \mathrm{sim}(u,v) > 0$$

La relation $\mathcal{R}$ définie sur $\mathcal{V}(\mathbf{f}) \times \mathcal{V}(\mathbf{g})$ par :

$$\forall j,k \quad \mathcal{R}(f_j, g_k) \iff \mathrm{sim}(f_j, g_k) > 0$$

est transitive si la fonction sim est l'une de celles définies plus haut, et par conséquent dans ce cas, le cluster équivalent $\mathcal{D}_m$, s'il existe, est unique. Notons toutefois que le problème soulevé plus haut d'une éventuelle ambiguïté sur le lemme nous obligeant à traiter une liste de lemme, compromet cette garantie d'unicité. En pratique, nous n'avons détecté ce problème lié à la lemmatisation qu'une poignée de fois sur le corpus, et avons choisi dans ce cas de nous contenter du premier cluster équivalent identifié.

Nous pouvons à présent définir, pour chaque cluster $\mathcal{C}_l$ qui possède un cluster équivalent $\mathcal{D}_m$ non vide, le nombre de liens $\beta(\mathcal{C}_l)$ qui agréent, le nombre de liens $\gamma(\mathcal{C}_l)$ ajoutés dans $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$ par rapport à $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$, et le nombre de liens $\delta(\mathcal{C}_l)$ de $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$ manquants dans $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$, selon :

$$\begin{aligned} \beta(\mathcal{C}_l) &= \sum_{i,j,k} a_{ij} b_{ik}\ \mathbb{1}(f_j \in \mathcal{C}_l)\ \mathbb{1}(g_k \in \mathcal{D}_m) \\ \gamma(\mathcal{C}_l) &= \sum_{i,j} a_{ij}\ \mathbb{1}(f_j \in \mathcal{C}_l) \quad - \quad \beta(\mathcal{C}_l) \\ \delta(\mathcal{C}_l) &= \sum_{i,k} b_{ik}\ \mathbb{1}(g_k \in \mathcal{D}_m) \quad - \quad \beta(\mathcal{C}_l) \end{aligned}$$

Nous donnons en Table 4.5 la somme de ces mesures (respectivement $\mathrm{B}, \Gamma, \Delta$ pour $\beta(\cdot), \gamma(\cdot), \delta(\cdot)$) pour toutes les phrases de $\mathcal{F}$ (`alice.1.fr`) aligné avec `MGiza++` et combiné avec $\mathcal{G}$ (`alice.2.fr`), ce dernier texte étant aligné avec les méthodes déjà présentées pour la Table 4.4. Ces sommes sont divisées [22] par le nombre de liens contenus dans tous les alignements $A^{\mathbf{e}\leftrightarrow\mathbf{f}}$. De nouveau, nous donnons également en bas de table les mesures correspondantes pour l'alignement combiné produit à partir de deux alignements de $\mathcal{F}$ avec $\mathcal{E}$.

22. Il ne s'agit pas d'une véritable normalisation puisque les liens comptés par $\delta(\cdot)$ proviennent de $A^{\mathbf{e}\leftrightarrow\mathbf{g}}$. Ce comptage n'est pas symétrique et se fait du point de vue du premier alignement en parcourant les clusters de mots similaires pour chaque **f**. Autrement dit, $\sum_{\mathcal{F}} \delta(\cdot) \neq \sum_{\mathcal{G}} \gamma(\cdot)$.

alignements	sim_1			sim_2			sim_3		
en-fr_1/en-fr_2	B	Γ	Δ	B	Γ	Δ	B	Γ	Δ
`MGiza`/`MGiza`	**0,519**	0,141	0,130	**0,546**	0,163	0,151	**0,555**	0,173	0,167
`MGiza`/`pCAT` (bij)	0,451	**0,209**	0,117	0,475	**0,233**	0,134	0,484	**0,245**	0,146
`MGiza`/`pCAT` (sym)	0,479	0,181	**0,141**	0,504	0,204	**0,161**	0,513	0,215	**0,177**
en-fr_1/en-fr_1	B	Γ	Δ	B	Γ	Δ	B	Γ	Δ
`MGiza`/`pCAT` (bij)	0,739	0,261	0,142	0,740	0,260	0,143	0,740	0,260	0,144
`MGiza`/`pCAT` (sym)	0,790	0,210	0,170	0,790	0,210	0,170	0,789	0,211	0,171

TABLE 4.5 – Mesures B, Γ, et Δ d'accord et de désaccord des liens de deux alignements du point de vue de la première cible ($\mathcal{F}$).

Les scores d'agrément du point de vue des mots sources donnés en Table 4.4 sont très cohérents avec la mesure d'accord B, ce qui est conforme à l'intuition, les deux approches mesurant fondamentalement la même information. Il y a pourtant des différences dans les démarches. L'approche du point de vue de la source est symétrique quant aux cibles, ce qui n'est plus le cas ici. Par ailleurs, un mot source resté non aligné des deux côtés aura un score d'agrément maximal dans l'approche de la section précédente, tandis qu'il sera tout simplement ignoré ici dans le calcul de B. La construction des clusters de similarité côté cible est aussi à même de créer des effets d'agrégation ayant tendance à augmenter l'accord.

Notons que la mesure d'agrément de la Table 4.4 correspond à une moyenne sur tous les mots cibles. Pour avoir une mesure côté source plus immédiatement comparable à B, il faudrait procéder un peu différemment, et calculer une autre moyenne, sur un nombre de liens global et non de mots.

On observe également que les colonnes B et Γ somment à valeur constante, pour un critère de similarité donné, quelle que soit la méthode d'alignement entre $\mathcal{E}$ et $\mathcal{G}$ (deuxième alignement). Lorsque $\mathcal{F} = \mathcal{G}$ (bas de table), cette somme,

$$\mathrm{B} + \Gamma = \sum_{\substack{\mathbf{f}\in\mathcal{F} \\ \mathcal{C}_l\in\{\mathcal{C}_l \mid \exists\mathcal{D}_m \text{ équiv.}\}}} \beta(\mathcal{C}_l) + \gamma(\mathcal{C}_l) \tag{4.1}$$

$$= \sum_{\substack{\mathbf{f}\in\mathcal{F} \\ \mathcal{C}_l\in\{\mathcal{C}_l \mid \exists\mathcal{D}_m \text{ équiv.}\}}} \sum_{i,j} a_{ij}\ \mathbb{1}(f_j \in \mathcal{C}_l)\ , \tag{4.2}$$

vaut 1. En revanche, dans le cas général ($\mathcal{F} \neq \mathcal{G}$), cette quantité mesure l'accord maximal possible étant donnés le premier alignement et le texte de la seconde traduction. Cette quantité augmente à mesure que l'on utilise une mesure de similarité de moins en moins restrictive.

Considérée séparément, la colonne Γ (resp. Δ) donne la proportion – par rapport au nombre de liens du premier alignement – de liens prédits seulement par le premier (resp. second) alignement.

L'ensemble de ces mesures nous permet d'avancer dans l'étude avec la preuve que l'accord entre deux alignements correspondant à des traductions différentes, quelle que soit la manière de le mesurer, ne concerne pas seulement une partie marginale des données, et devrait pouvoir se prêter à une exploitation visant à améliorer le premier alignement.

4.4.4 Probabilités a posteriori des liens

Les mesures d'agrément que nous venons de présenter invitent à imaginer une méthode de filtrage des liens non renforcés par l'agrément. Dans cette optique, disposer d'un score de confiance des liens eux-mêmes serait précieux. La probabilité a posteriori $\mathrm{P}(a_{ij} \mid \mathbf{e}, \mathbf{f})$, au sens des paramètres du modèle estimés à l'issue de l'apprentissage, constitue certainement le score de confiance le plus naturel.

Si l'on dispose, à l'issue de l'entraînement d'un modèle IBM 4, des probabilités a posteriori $\mathrm{P}(\mathbf{a} \mid \mathbf{e}, \mathbf{f})$ des *n-best* alignements les plus probables, il est possible, en notant $\mathcal{N}$ l'ensemble de ces alignements, de calculer une approximation de la probabilité a posteriori de chaque lien en re-normalisant selon :

$$\mathrm{P}(a_{ij} \mid \mathbf{e}, \mathbf{f}) \approx \frac{\sum_{\mathbf{a} \in \mathcal{N}} a_{ij} \, \mathrm{P}(\mathbf{a} \mid \mathbf{e}, \mathbf{f})}{\sum_{\mathbf{a} \in \mathcal{N}} \mathrm{P}(\mathbf{a} \mid \mathbf{e}, \mathbf{f})}$$

Les auteurs de (Liu *et al.*, 2009) – dans le cadre d'un travail portant sur l'extraction de bi-segments pour la traduction automatique – mettent à disposition un programme [23] permettant de construire les matrices de probabilité des liens a posteriori à partir des fichiers de sortie produits par `MGiza++` (ou `Giza++`) [24]. Un autre outil [25], que nous n'avons pas encore pu utiliser, permet d'extraire le même type de matrices en sortie de `postCAT`.

Nous donnons en Figure 4.3 un alignement combiné à partir de deux alignements fournis par `MGiza++` et dont les liens sont décorés par l'approximation de leur probabilité a posteriori. On peut y observer que les liens erronés (<*je*—*what*>, <*me*—*feeling*>, <*sens*—*shutting*>, <*on*—*be*>, par exemple dans l'alignement de gauche) semblent correspondre à des niveaux de probabilité a posteriori plus faible. On y observe également que ce critère n'est pas infaillible, comme le montre la probabilité très élevée du lien <"—*Alice*> dans l'alignement de droite.

4.4.5 Corrélation entre l'agrément et la probabilité

À partir d'un alignement combiné produit par `MGiza++` sur les deux traductions, et pour chaque lien du premier alignement dans chaque phrase, nous recueillons $\sigma_a(e_i, f_j)$ et $\mathrm{P}(a_{ij} \mid \mathbf{e}, \mathbf{f})$, puis calculons le coefficient de Pearson correspondant. Nous obtenons un résultat de 0.167 avec une valeur-p quasi nulle, ce qui indique, étant donné la taille non négligeable de l'échantillon (21 728 liens dans le premier alignement), une corrélation positive mais faible.

Ceci semble une information encourageant à utiliser de manière conjointe les deux informations. C'est ce que nous faisons pour élaborer une première méthode très simple de filtrage des liens.

4.5 Filtrage

Comme nous l'avons déjà évoqué, certaines tâches requérant une information d'alignement sous-phrastique sont parfois peu sensibles, dans une certaine limite, à la qualité des alignements et ont tendance à privilégier la quantité disponible de liens prédits, c'est-à-dire le rappel plutôt que la précision. Mais, nous l'évoquions également, l'exigence de précision est beaucoup plus forte si l'on veut construire par exemple un système de lecture

23. `http://nlp.csai.tsinghua.edu.cn/~ly/wam/wam.html`

24. Les alignements en sortie de `MGiza++` sont asymétriques, et sont symétrisés par des heuristiques déjà mentionnées. Il n'est pas complètement clair pour nous, néanmoins, de savoir comment la re-normalisation peut elle-même s'effectuer sur l'alignement symétrisé dont a priori nous ne disposons par des probabilités a posteriori.

25. `https://code.google.com/p/geppetto/`

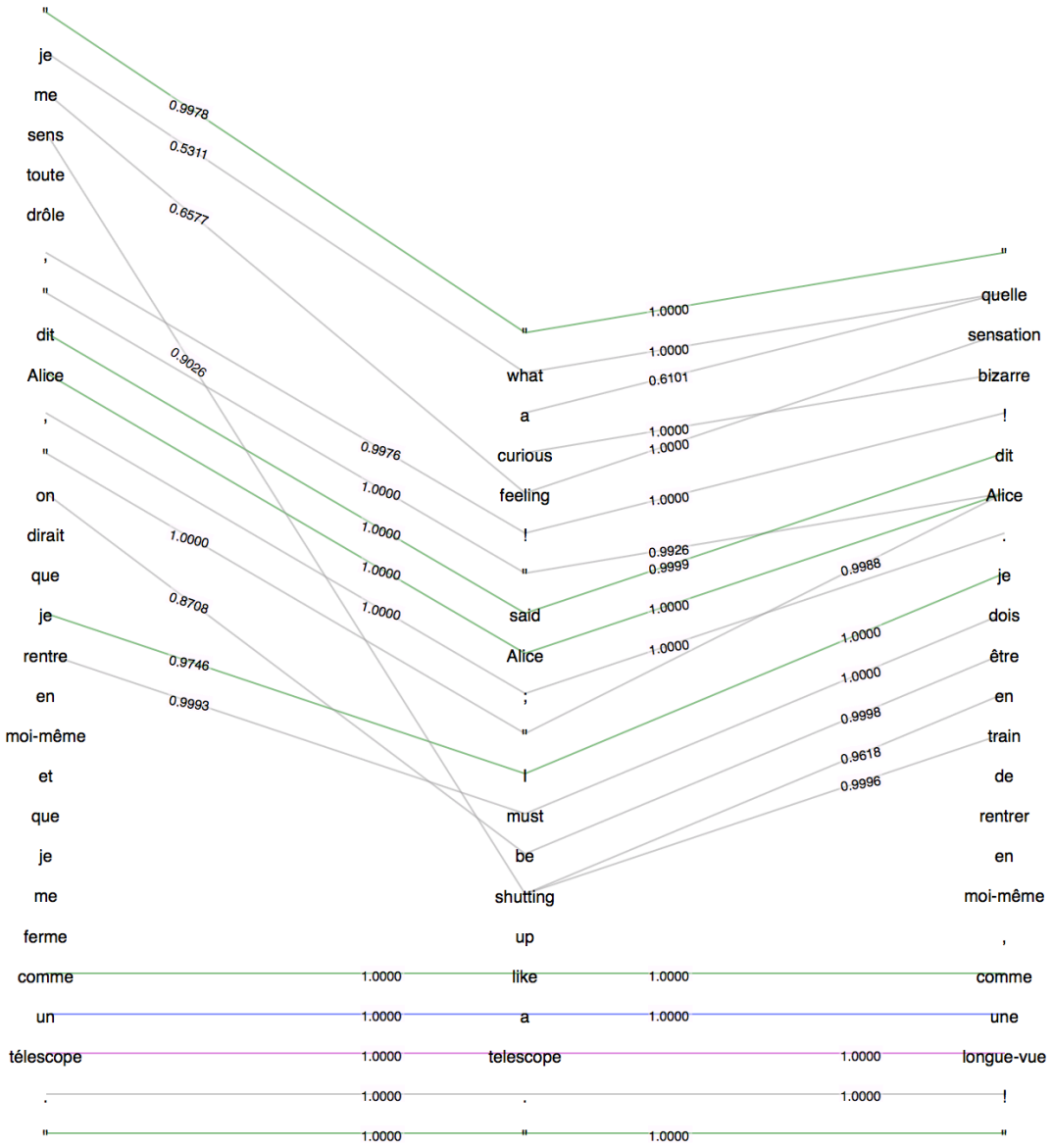


FIGURE 4.3 – Décoration des liens par leur probabilité a posteriori (approximation).

augmentée pour le lecteur imparfaitement bilingue. Pour aller dans ce sens, et exploiter certaines connaissances acquises lors de l'exploration analytique des données dont nous rendions compte dans la section précédente, nous avons implémenté une méthode de filtrage à l'intérieur d'un alignement combiné. Nous présentons dans cette section les moyens mis en œuvre pour pouvoir évaluer ce filtrage, puis nous en donnons les premiers résultats.

4.5.1 Alignements de référence

Pour mesurer l'éventuelle amélioration de la qualité de l'alignement, une étude d'erreur sur des exemples choisis plus ou moins au hasard n'était plus satisfaisante à ce stade, et nous avions donc besoin d'un alignement de référence.

Méthode d'annotation L'alignement manuel au niveau des mots est une tâche réputée difficile et coûteuse en temps. Pour donner un cadre de référence à cette annotation, nous avons tenté de suivre au plus près les recommandations du projet Blinker (Melamed, 1998a,b).

La philosophie générale de ce guide d'annotation est de privilégier l'alignement le plus fin possible, c'est-à-dire de décomposer le plus possible les unités linguistiques, en maintenant toutefois la relation d'équivalence entre les mots ou les groupes de mots alignés [26]. Par ailleurs, ne doivent être écartés de l'alignement que les mots dont la suppression améliorerait, du point de vue de l'annotateur, la similarité de traduction entre la phrase source et la phrase cible. Nous ne faisons pas, dans ce cadre-ci, de distinction de catégorie entre les liens (sûr, possible, paraphrase, dépendant du contexte,...).

Certains cas problématiques fréquents sont traités en détail par le guide d'annotation : la négation à deux termes du français, certains problèmes de rattachement de la préposition, l'absence fréquente des déterminants français dans le texte anglais, les difficultés liées au passage à la voix passive, le traitement des possessifs, en sont quelques exemples.

Pour autant, et particulièrement lorsqu'il s'agit d'un texte littéraire dont la traduction invite plus volontiers à l'invention, la tâche est souvent ardue et subjective. N'ayant pas les moyens de réaliser des annotations indépendantes afin de mesurer des scores d'agrément entre les annotateurs, nous avons procédé par annotation successive de trois annotateurs en tentant de rendre l'annotation de plus en plus homogène et en discutant les cas problématiques.

À l'aide d'un outil d'alignement manuel, `Yawat` (Germann, 2008) dont nous avons parlé en section 4.4.2, nous avons ainsi annoté les 100 premières paires de phrases formées par le texte de Lewis Carroll et la première traduction dont nous disposons. Cet alignement manuel représente un peu plus de 6 000 liens, c'est-à-dire environ 60% de liens supplémentaires par rapport à l'alignement réalisé par `MGiza++`, ce qui explique des niveaux de précision relativement élevés sur les alignements automatiques, et des niveaux plus faibles de rappel.

Évaluation La Table 4.6 donne les valeurs de précision, de rappel et de F-mesure [27] pour l'alignement automatique de la première traduction d'*Alice* sur les 100 premières paires de phrases, ainsi que les valeurs correspondantes pour notre *baseline* consistant à entraîner un modèle IBM 4 en versant la seconde traduction dans le corpus (voir la

26. Les expressions figées, par exemple, doivent être alignées intégralement avec les mots correspondants de l'autre texte.

27. Voir la section 1.3.2, avec l'ensemble P des liens possibles égal ici à l'ensemble S des liens sûrs puisque nous ne faisons pas de distinction entre ces deux classes de liens.

section 4.3.2). Nous contrastons également ces résultats avec la performance de deux modèles HMM entraînés sous contrainte de bijectivité ou de symétrie (voir la section 2.2.1).

alignements	précision (%)	rappel (%)	F-mesure (%)
`MGiza` (1)	79,36	49,98	61,33
`MGiza` (1and2)	79,31	**51,33**	62,32
`pCAT` (bij)	**84,69**	49,46	**62,45**
`pCAT` (sym)	76,83	50,42	60,89

TABLE 4.6 – Précision, rappel, et F-mesure pour différentes méthodes d'alignement automatique sur les 100 paires de phrases annotées manuellement.

La *baseline* fait apparaître un gain non négligeable en rappel, et un léger recul en précision difficile à interpréter, mais l'augmentation d'un point de F-mesure permet néanmoins d'affirmer que cette méthode améliore l'alignement standard réalisé avec `MGiza++`. Quoique l'objectif de ce travail ne soit pas à proprement parler de comparer les mérites de diverses méthodes d'alignement bilingue, mais d'examiner comment améliorer le résultat de ces méthodes en tirant parti d'un alignement supplémentaire, nous observons la performance clairement supérieure en précision de l'alignement réalisé avec `postCAT` sous contrainte de bijectivité, et ce sans recul important du rappel.

4.5.2 Seuils de filtrage

Nous avons à ce stade les moyens de filtrer les liens a_{ij}, selon un seuil t sur les $\mathrm{P}(a_{ij} \mid \mathbf{e}, \mathbf{f})$ et/ou un seuil u sur les $\sigma_a(e_i, f_j)$, et de mesurer l'évolution conséquente de la précision, du rappel, et de la F-mesure. Nous effectuons ces expériences sur le premier alignement réalisé avec `MGiza++`, puisque nous ne disposons pas encore des probabilités a posteriori des liens en sortie de `postCAT`.

La Figure 4.4 matérialise la performance d'un filtrage sur ce premier alignement consistant à écarter les liens ayant un niveau de probabilité a posteriori inférieur à un seuil t, sans tenir compte pour l'instant du score d'agrément σ_a.

Il est clair qu'en procédant de la sorte on ne peut rien espérer d'autre, du point de vue du rappel, qu'une dégradation de la performance. La question est de savoir quelle est l'ampleur réciproque de la chute en rappel et de la progression en précision. Dans le cas de figure présenté en Figure 4.4, la progression de la précision au seuil de filtrage le plus sévère ($t = 1.0$) est de l'ordre de 10 % et la perte en rappel de l'ordre de 15 %.

La Figure 4.5, quant à elle, illustre le phénomène équivalent, mais en activant un seuil u sur $\sigma_a(e_i, f_j)$, avec la mesure de similarité sim_3, ceci faisant intervenir un second alignement réalisé lui aussi avec `MGiza++`. En faisant varier ce seuil entre les différentes valeurs possibles au regard de la définition de sim_3, c'est-à-dire $u = 0$, $u = 0,5$, $u = 0,7$, ou $u = 1$, on observe qu'il n'y a pratiquement aucun bénéfice en précision à augmenter le seuil au-delà de 0,5 et que le faire n'a pour effet que de dégrader plus encore le rappel. L'expérience précédente correspondait à un seuil nul, et la Figure 4.5 que nous présentons à présent correspond, elle, à un seuil $u = 0,5$.

Même sans disposer des probabilités a posteriori des liens en sortie de `postCAT`, nous pouvons cependant étudier l'impact de la méthode utilisée pour le second alignement dans notre opération de filtrage, puisque celle-ci influe sur les scores d'agrément $\sigma_a(e_i, f_j)$,

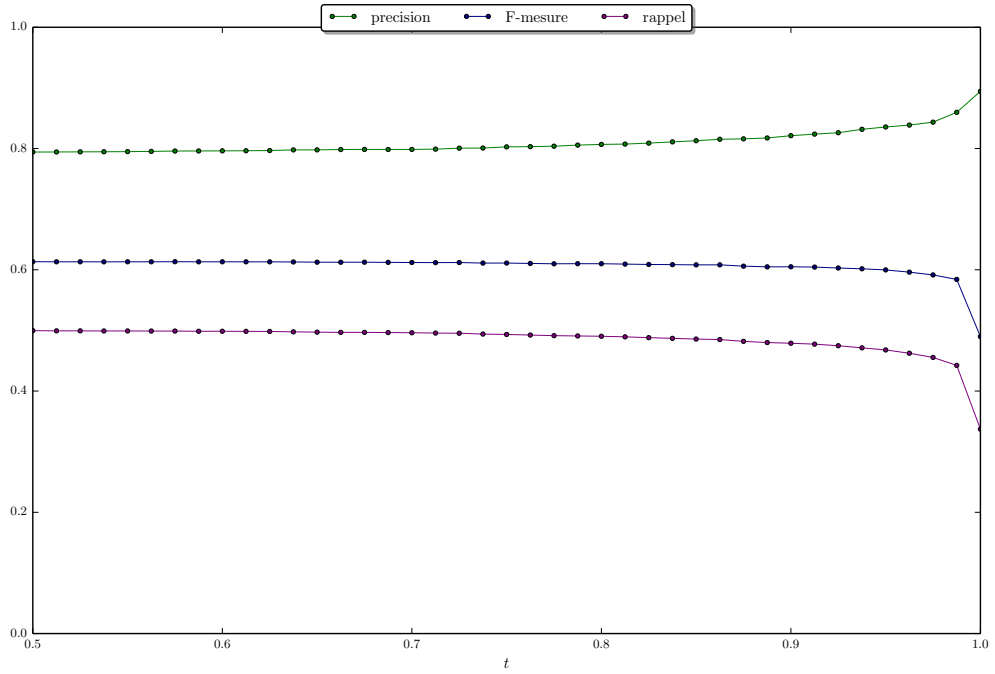


FIGURE 4.4 – Précision, rappel, et F-mesure en fonction du seuil d'acceptation t pour $\mathrm{P}(a_{ij} \mid \mathbf{e}, \mathbf{f})$.

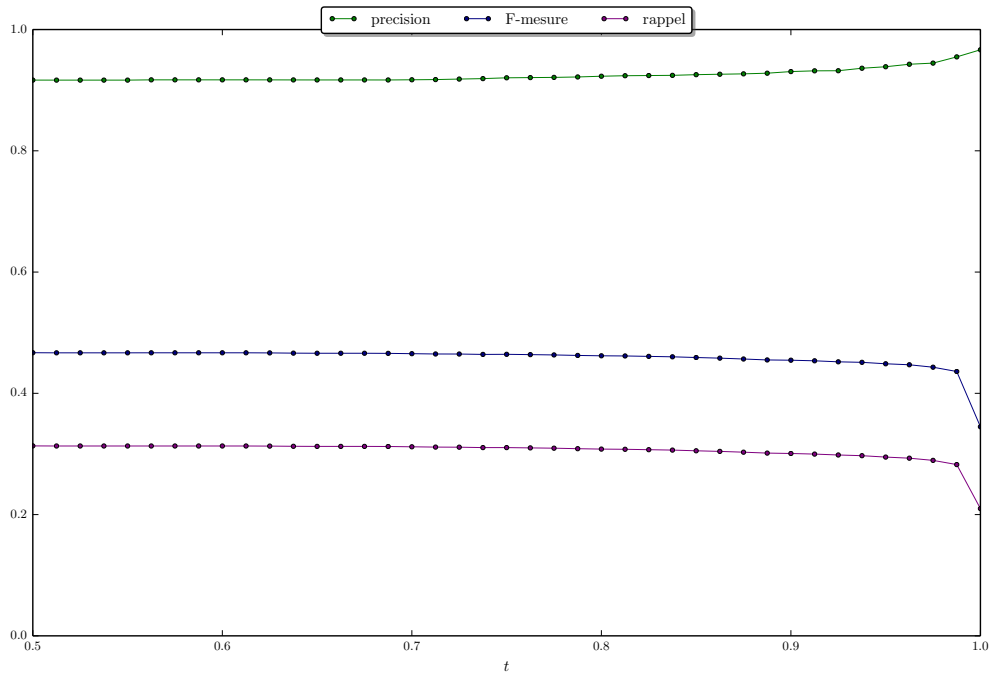


FIGURE 4.5 – Précision, rappel, et F-mesure en fonction du seuil d'acceptation t, avec un seuil supplémentaire $u = 0.5$ pour $\sigma_a(e_i, f_j)$. (Deuxième alignement : `MGiza++`).

et qu'à ce stade nous n'exploitons pas le niveau de probabilité a posteriori des liens de ce second alignement.

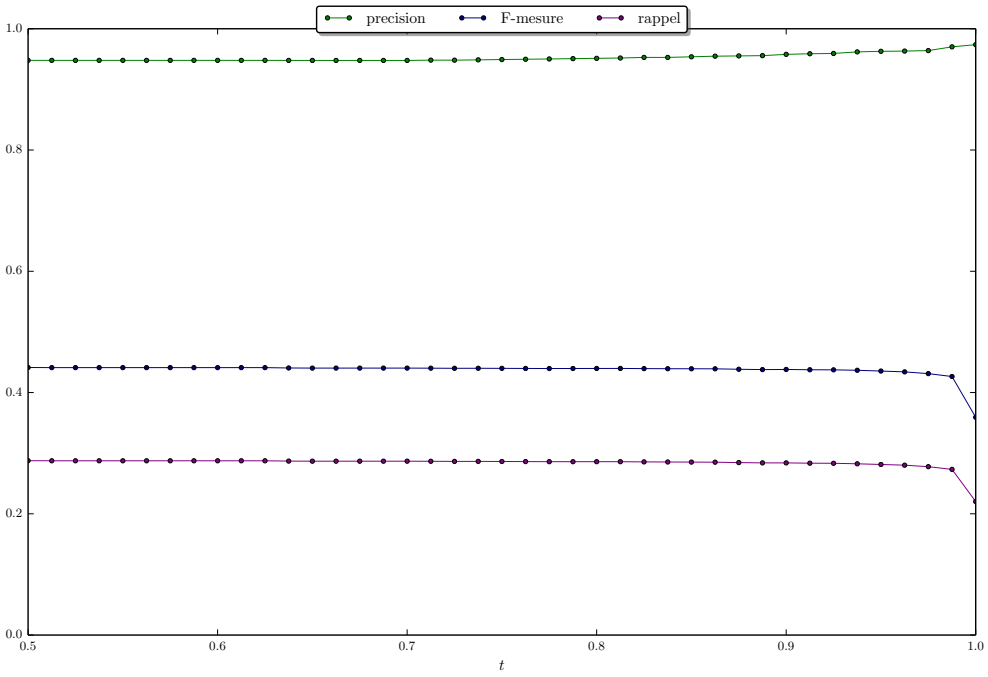


FIGURE 4.6 – Précision, rappel, et F-mesure en fonction du seuil d'acceptation t, avec un seuil supplémentaire $u = 0.5$ pour $\sigma_a(e_i, f_j)$. (Deuxième alignement : `postCAT` avec une contrainte de bijectivité).

La Figure 4.6 rend compte de cette troisième expérience avec une précision très élevée, mais qui se paie par une nouvelle dégradation en rappel.

4.5.3 Synthèse

Pour donner une vision globale de certains des effets discutés plus haut, nous reportons, dans la Table 4.7, les niveaux de précision, rappel et F-mesure des différentes méthodes pour les 100 premières phrases du corpus.

Pour le filtrage, nous avons choisi un seuil $t = 0.95$ et un seuil $u = 0.5$. Les mentions « 1and2 » renvoient à la *baseline* décrite plus haut, et ce pour des alignements produits par les deux programmes que nous avons utilisés, `MGiza++` et `postCAT`. Dans le cas de `postCAT` (ici, tous les résultats correspondent à une contrainte de bijectivité), le seuil t n'est pas activé, faute de disposer des probabilités nécessaires [28]. Enfin, les mentions « + `MGiza` » et « + `postCAT` » indiquent la méthode d'alignement utilisée pour le second alignement (deuxième traduction) de façon à pouvoir calculer les scores $\sigma_a(e_i, f_j)$.

Les résultats de cette méthode de filtrage simple montrent la capacité de cette approche à extraire des liens très fiables. Du point de vue du rappel, le paysage se montre plus aride, et les gains en précision induisent un dépeuplement conséquent de notre alignement. Encore une fois, l'usage que l'on fait de ces alignements doit guider l'évaluation

28. Nous observons aussi que la *baseline* « 1and2 » pour `postCAT` n'augmente pas la performance par rapport à l'alignement standard, et la dégrade même un peu, ce que nous avons un peu de peine à expliquer.

alignements	précision (%)	rappel (%)	F-mesure (%)
non filtrés (ref.)	79,36	49,98	61,33
`MGiza` (1and2)	79,31	**51,33**	62,32
`MGiza` filtré			
$t = 0.95, u = 0$	83,55	46,78	59,98
$t = 0.95, u = 0.5$ + `MGiza`	93,86	29,48	44,87
$t = 0.95, u = 0.5$ + `pCAT`	**96,28**	28,13	43,54
non-filtré (ref.)	84,69	49,46	**62,45**
`pCAT` (1and2)	84,62	49,46	62,43
`pCAT` filtrés			
$u = 0.5$ + `MGiza`	95,95	29,63	45,28
$u = 0.5$ + `pCAT`	94,69	29,96	45,52

TABLE 4.7 – Synthèse de la précision, du rappel, et de la F-mesure pour une méthode de filtrage des alignements.

de la performance. Une information de très grande confiance même sur à peine plus d'un quart des liens pourra dans certains cas se révéler utile.

4.5.4 Perspectives

Pour conclure ce chapitre, nous donnons ici un début de liste de questions autour desquelles nous aimerions continuer à travailler.

Sur le filtrage

- Comment utiliser un moyen plus fin de mesurer la similarité lexicale monolingue ? Comment réaliser un alignement automatique monolingue de qualité ?
- Est-ce que certains types de partie du discours agréent plus fréquemment ?
- Comment peut-on utiliser les liens très sûrs que nous avons extraits pour initialiser (*bootstrapping*) un modèle d'apprentissage existant ou un nouveau modèle ?
- Comment adapter la méthode de filtrage présentée ici à une deuxième traduction qui soit dans une autre langue ?
- Étant donné que le filtrage n'est exécuté que localement (au niveau de chaque phrase), serait-il intéressant d'essayer de capturer des phénomènes récurrents à plus grande échelle ?
- À l'inverse du filtrage, comment « compléter » le premier alignement avec la connaissance des liens du second alignement « manquants » dans le premier ?

Sur l'extraction Une direction de travail complémentaire, et à laquelle nous n'avons pas encore pu nous consacrer relève de la question de la segmentation : est-ce que la donnée d'un ou plusieurs alignements supplémentaires peut être utile à la segmentation des phrases ?

En recherchant des composantes connexes sur un graphe formé à partir des plusieurs graphes bipartis représentant des alignements bilingues, on pourrait par exemple extraire

des segments non directement identifiables dans le seul cas bilingue. Dans la Figure 4.7, qui illustre cette intuition, on peut voir que la donnée de l'alignement français ↔ anglais pourrait permettre d'extraire la paire de segments (*boules de neige*, *bolas de nieve*) non atteignable par le seul alignement français ↔ espagnol.

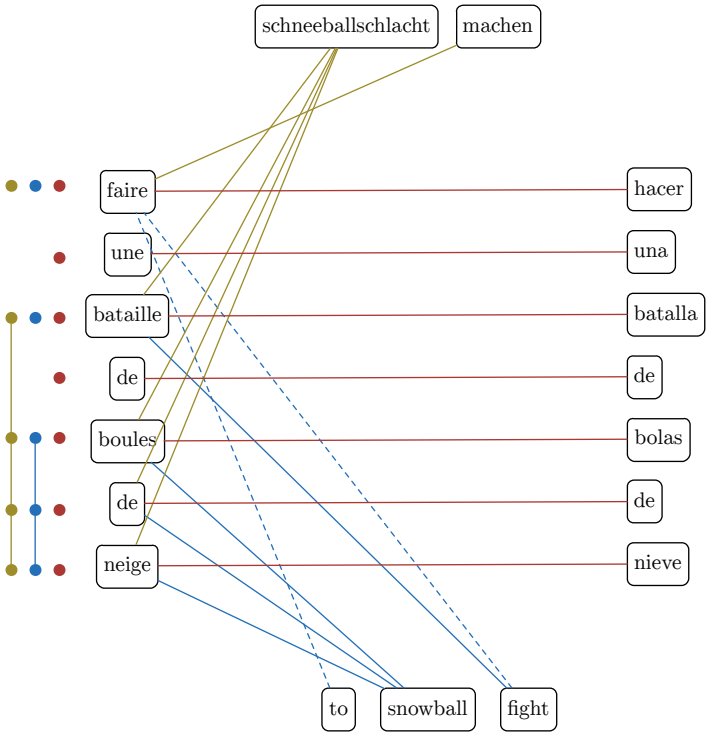


FIGURE 4.7 – Composantes connexes multilingues et extraction de nouveaux segments pour une paire de langues

Conclusion

Le travail présenté dans ce mémoire nous a permis de nous confronter à de nombreux problèmes théoriques, et quoique toujours intimidés par certains développements à ce jour encore hors de notre portée, nous avons désormais acquis les connaissances nécessaires à la lecture d'une grande partie de la littérature du domaine.

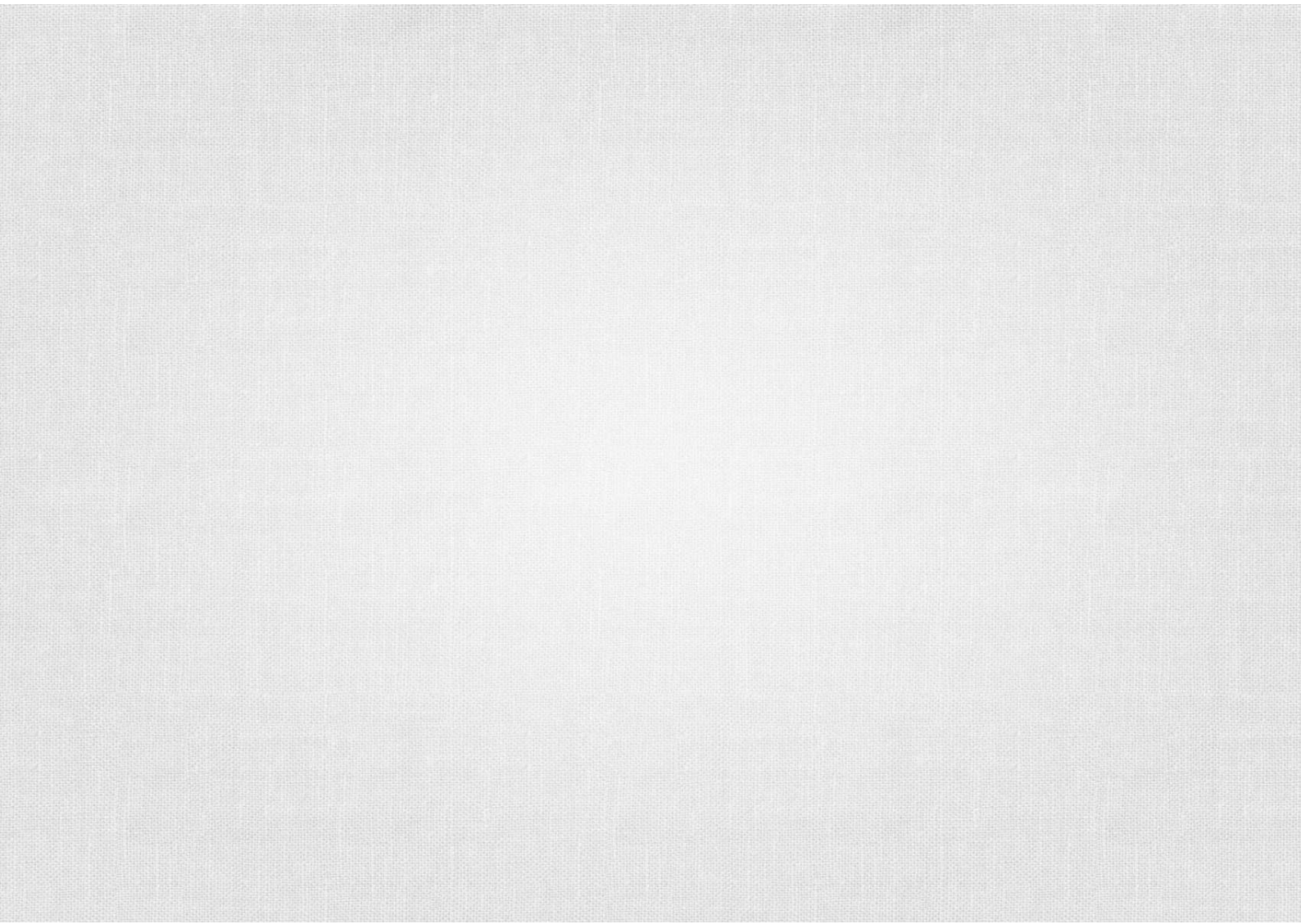
L'acquisition des compétences plus techniques nécessaires à la conduite d'expériences sur des données réelles a également mobilisé une grande quantité de notre énergie, et nous achevons ce stage avec la maîtrise de nombreux nouveaux outils.

Les expériences réalisées sont des ébauches qui nous semblent prometteuses pour la tâche d'alignement en général. L'alignement des textes littéraires, d'autre part, offre aussi beaucoup d'espace tant pour le jeu scientifique que poétique.

Le choix du texte de Lewis Carroll, pourtant, fut un choix de raison à un moment du travail où il était important de commencer à travailler sur des données, difficiles à trouver ou à constituer. C'est sans doute le regret le plus important que nous formons sur le travail de ces derniers mois.

D'autres textes viendront, et le travail, quelle qu'en soit la manière, se poursuivra.

Remerciements



Robert Rauschenberg, *White Painting* [three panel], 1951 (détail).

Ma chaleureuse gratitude pour François Yvon et Isabelle Tellier qui ont encadré ce travail, Alexandre Allauzen, Nicolas Pécheux, Hélène Bonneau-Maynard, Guillaume Wisniewski, Marianna Apidianaki, Éric Bilinski, Natalia Segal, Thomas Lavergne, Aurélien Max, Li Gong et Yong Xu pour leur aide patiente, et le plaisir des conversations. Ma gratitude également pour Serge Fleury et sa généreuse disponibilité tout au long de l'année. Et, naturellement, pour les Annotateurs Anonymes.

Table des figures

Liste des tableaux

Bibliographie

Allauzen, A. et Wisniewski, G. (2009). Modèles discriminants pour l'alignement mot à mot. *Traitement Automatique des Langues*, 50(3) :173–203.

Allauzen, A. et Yvon, F. (2011). Méthodes statistiques pour la traduction automatique. *Gaussier, E. et Yvon, F., éditeurs : Modèles statistiques pour l'accès à l'information textuelle*, 7 :271–356.

Ayan, N. F. et Dorr, B. J. (2006). A maximum entropy approach to combining word alignments. *In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 96–103. Association for Computational Linguistics.

Ayan, N. F., Dorr, B. J. et Monz, C. (2005). Neuralign : Combining word alignments using neural networks. *In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 65–72. Association for Computational Linguistics.

Banerjee, S. et Lavie, A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. *In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Berge, C. (1970). *Graphes et hypergraphes.* Monographies universitaires de mathematiques. Dunod.

Berger, A. L., Pietra, V. J. D. et Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1) :39–71.

Blunsom, P. et Cohn, T. (2006). Discriminative word alignment with conditional random fields. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 65–72. Association for Computational Linguistics.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. et Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2) :79–85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. et Mercer, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational linguistics*, 19(2) :263–311.

Brunning, J. (2010). *Alignment Models and Algorithms for Statistical Machine Translation.* Thèse de doctorat.

Chang, Y.-W., Rush, A. M., DeNero, J. et Collins, M. (2014). A constrained viterbi relaxation for bidirectional word alignment. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1481–1490, Baltimore, Maryland. Association for Computational Linguistics.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1) : 1–38.

DeNero, J. et Macherey, K. (2011). Model-based aligner combination using dual decomposition. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 420–429. Association for Computational Linguistics.

Deng, Y. et Byrne, W. (2005). HMM word and phrase alignment for statistical machine translation. *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 169–176, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Filali, K. et Bilmes, J. (2005). Leveraging multiple languages to improve statistical MT word alignments. *Machine translation*, 1 :2.

Fraser, A. et Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational linguistics*, 33(3) :293–303.

Gale, W. A. et Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Berkeley, California, USA. Association for Computational Linguistics.

Gao, Q. et Vogel, S. (2008). Parallel implementations of word alignment tool. *In Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics.

Germann, U. (2008). Yawat : yet another word alignment tool. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies : Demo Session*, pages 20–23. Association for Computational Linguistics.

Graça, J., Ganchev, K. et Taskar, B. (2007). Expectation maximization and posterior constraints. *In NIPS*, volume 20, pages 569–576.

Graça, J., Ganchev, K. et Taskar, B. (2010). Learning tractable word alignment models with complex constraints. *Computational Linguistics*, 36(3) :481–504.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences : computer science and computational biology*. Cambridge University Press.

Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. *In MT summit*, volume 5, pages 79–86.

Koehn, P., Och, F. J. et Marcu, D. (2003). Statistical phrase-based translation. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Kumar, S., Och, F. J. et Macherey, W. (2007). Improving word alignment with bridge languages. *In EMNLP-CoNLL*, pages 42–50.

Lardilleux, A. et Lepage, Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. *Proceedings of AMTA 2008*, pages 125–132.

Lardilleux, A. et Lepage, Y. (2009). Sampling-based multilingual alignment. *In Proceedings of Recent Advances in Natural Language Processing*, pages 214–218.

Lardilleux, A., Lepage, Y. et Yvon, F. (2011). The contribution of low frequencies to multilingual sub-sentential alignment : a differential associative approach. *International Journal of Advanced Intelligence*, 3(2) :189–217.

Lardilleux, A., Yvon, F. et Lepage, Y. (2013). Generalizing sampling-based multilingual alignment. *Machine translation*, 27(1) :1–23.

Liang, P., Taskar, B. et Klein, D. (2006). Alignment by agreement. *In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Liu, Q., Tu, Z. et Lin, S. (2013). A novel graph-based compact representation of word alignment. *In ACL (2)*, pages 358–363.

Liu, Y., Xia, T., Xiao, X. et Liu, Q. (2009). Weighted alignment matrices for statistical machine translation. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2-Volume 2*, pages 1017–1026. Association for Computational Linguistics.

Martzoukos, S., Florêncio, C. C. et Monz, C. (2013). Investigating connectivity and consistency criteria for phrase pair extraction in statistical machine translation. *In The 13th Meeting on the Mathematics of Language*, page 93.

Matusov, E., Zens, R. et Ney, H. (2004). Symmetric word alignments for statistical machine translation. *In Proceedings of the 20th international conference on Computational Linguistics*, page 219. Association for Computational Linguistics.

Mayer, T. et Cysouw, M. (2012). Language comparison through sparse multilingual word alignment. *In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62. Association for Computational Linguistics.

Melamed, I. D. (1998a). Annotation style guide for the blinker project. *arXiv preprint cmp-lg/9805004*.

Melamed, I. D. (1998b). Manual annotation of translational equivalence : The blinker project. *arXiv preprint cmp-lg/9805005*.

Moore, R. C. (2002). *Fast and accurate sentence alignment of bilingual corpora*. Springer.

Moore, R. C. (2004). Improving IBM word-alignment model 1. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 518. Association for Computational Linguistics.

Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1) :19–51.

Och, F. J., Tillmann, C., Ney, H. et others (1999). Improved alignment models for statistical machine translation. *In Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.

Östling, R. (2014). Bayesian word alignment for massively parallel texts.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Simard, M. (1999). Text-translation alignment : Three languages are better than two. *In Proc. of EMNLP/VLC*, pages 2–11.

Tu, Z., Liu, Y., He, Y., van Genabith, J., Liu, Q. et Lin, S. (2012). Combining multiple alignments to improve machine translation. *In COLING (Posters)*, pages 1249–1260. Citeseer.

Vogel, S., Ney, H. et Tillmann, C. (1996). HMM-based word alignment in statistical translation. *In Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Yu, Q., Max, A. et Yvon, F. (2012). Aligning bilingual literary works : a pilot study. *NAACL-HLT 2012*, page 36.