Mémoires partagées d'alignements sous-phrastiques bilingues

Johan Segura

UNIVERSITÉ DE MONTPELLIER II

Mémoires partagées d'alignements sous-phrastiques bilingues

par
Johan Segura

Sciences et Techniques du Languedoc
Spécialité d'informatique et de recherche opérationnelle

Thèse présentée à l' Université des Sciences et Techniques du Languedoc
en vue de l'obtention du grade de Docteur
en informatique

Novembre, 2012

UNIVERSITÉ DE MONTPELLIER II
Université des Sciences et Techniques du Languedoc

Cette thèse intitulée:

Mémoires partagées d'alignements sous-phrastiques bilingues

présentée par:

Johan Segura

a été évaluée par un jury composé des personnes suivantes:

Violaine Prince,	directeur de recherche
Christian Boitet,	rapporteur
Christian Retoré,	rapporteur
Yves Lepage,	examinateur
Christophe Paul,	examinateur
Jean-Philippe Prost,	examinateur

TABLE DES MATIÈRES

TABLE DES MATIÈRES . . . . . **v**

LISTE DES FIGURES . . . . . **ix**

CHAPITRE 1 : PROBLÉMATIQUE ET ÉTAT DE L'ART **7**

1.1 L'alignement 7

1.1.1 Notions 7

1.1.2 L'alignement sous-phrastique 10

1.1.3 Des modèles de différentes expressivités 11

1.2 De la divergence en traduction 22

1.2.1 Premier obstacle : les expressions 23

1.2.2 La préservation du sens derrière la forme de surface 25

1.2.3 La divergence observée face à la divergence souhaitée 29

1.3 Positionnement 31

CHAPITRE 2 : L'ENSEMBLE DES ALIGNEMENTS **35**

2.1 Discussion en faveur d'un modèle d'alignement adapté aux divergences 35

2.2 Un modèle adapté 40

2.2.1 Généralités 40

2.2.2 L'espace formel des alignements sous-phrastiques 43

2.2.3 Les sous-ensembles de modèles existants 52

2.3 Comparaison entre alignements 60

2.3.1 Avant-propos 60

2.3.2 Les distances des transferts 66

2.3.3 La distance des divisions 75

CHAPITRE 3 : ALIGN^{IT} : UNE APPROCHE COLLABORATIVE ET DES OUTILS AUTOMATIQUES **79**

3.1 Mise en place d'un outil d'alignement collaboratif 79

3.1.1 Motivations et orientations envisagées 79
3.1.2 Align^{It} : Présentation de l'interface homme-machine 91
3.2 L'alignement sous-phrastique à base d'exemples 100
3.2.1 Avant-propos 100
3.2.2 Architecture générale 101
3.2.3 L'analyse syntaxique en renfort 108

CHAPITRE 4 : MODÈLE DE REPRÉSENTATION THÉORIQUE POUR DES ALIGNEMENTS STRUCTURÉS 117
4.1 Une structure expressive pour décrire les exemples 117
4.1.1 Une correspondance entre l'arbre et la phrase : la SSTC 117
4.1.2 Une correspondance bilingue entre SSTC : la S-SSTC 122
4.2 Les structures bilingues dans Align^{It} 131

CHAPITRE 5 : UNE MÉMOIRE DE FRAGMENTS POUR L'ALIGNEMENT À BASE D'EXEMPLES 135
5.1 Constitution d'une mémoire d'exemples 135
5.1.1 Fragments formels 135
5.1.2 Les mémoires d'alignements 148
5.1.3 La taille (potentielle) de l'ensemble des fragments compatibles . 155
5.2 Reconstruction à base de fragments 160
5.2.1 Modélisation du problème 161
5.2.2 Complexités 165
5.2.3 Bilan 169

CHAPITRE 6 : CADRE EXPÉRIMENTAL 173
6.1 Ressources utilisées 173
6.1.1 Des corpus parallèles 173
6.1.2 Les analyseurs syntaxiques 181
6.2 Quelques expériences 184
6.2.1 La reconstruction par des fragments courts 184

6.2.2 Accords et désaccords d'une approche non experte 187

CHAPITRE 7 : CONCLUSION 191

7.1 Synthèse 191

7.2 Perspectives 195

7.2.1 Apports techniques souhaitables 195

7.2.2 R-utilisateurs trilingues 196

7.2.3 Fragmentations 196

7.2.4 Des solutions du côté de la bioinformatique 198

7.2.5 Le mot de la fin 199

BIBLIOGRAPHIE 201

I.1 Corpus *news commentary* xv

I.2 Corpus *DW2* xix

II.1 Corpus *news commentary* xxiii

II.2 Corpus *DW2* xxvii

RÉSUMÉ xxxi

LISTE DES FIGURES

1.1 Un corpus parallèle bilingue 8
1.2 Représentations d'un alignement sous-phrastique 9
1.3 Un alignement de mots non-injectif posant problème 13
1.4 Deux types de groupes non contigus : cas monolingue et cas bilingue 17
1.5 transformation d'un alignement discontinu 20
1.6 Deux alignements ne sont pas atteints pour deux phrases de longueur 4 : "inside-out" 22
1.7 Un idiotisme aligné révélant une discontinuité et des groupements 24
1.8 Deux exemples de collocations alignées 25
1.9 Cas de divergences syntaxiques participant au transfert "naturel" . 26
1.10 Les sept cas de divergence lexico-sémantique de B. Dorr 27
1.11 Un cas de traduction approchée dû à une traduction via une langue pivot. 29
1.12 Récapitulatif des phénomènes de divergence observés 30
1.13 Un exemple d'alignement manuel proposé par le guide du Blinker 32

2.1 Un alignement dégénéré et des alternatives "bien formées" 36
2.2 On peut vouloir aligner des unités plus fines que le mot 37
2.3 Un alignement plus grossier résoud le problème. 37
2.4 Exemple de groupement non contigu de grande longueur 38
2.5 Exemple de configuration *"inside-out"* rencontrée 39
2.6 Élargir les éléments de base réduit la divergence apparente de l'alignement. 39
2.7 Diagramme de Hasse des partitions sur un ensemble à quatre éléments 43
2.8 Un exemple d'alignement : chaque couleur indique une partition partielle 44
2.9 Les alignements triviaux $\ell_{\mathbb{O}}$ et $\ell_{\mathbb{1}}$ 45

2.10 L'espace des alignements s'injecte dans celui des graphe bipartis . 45
2.11 L'opérations d'affinement sur un exemple 48
2.12 L'opérations d'élargissement sur un exemple 49
2.13 Des alignements de moins en moins fins 49
2.14 Diagramme de Hasse de $\mathscr{A}(3,2)$ 51
2.15 Les cardinalités par rapport à la longueur (n,n) (échelle *log*) . . . 58
2.16 Alignements éventuellement partiels pour ces trois ensembles . . 58
2.17 L'espace total inclut tous les autres 59
2.18 L'espace des graphes bipartis couvrants (échelle *log*) 59
2.19 Démultiplication des liens pour mesures pondérées 61
2.20 Les cinq transformations élémentaires pour un ensemble à 16 éléments 64
2.21 Les cinq transformations élémentaires pour une biphrase de longueur $(4,4)$ 65
2.22 *Transfert de singleton* et *division unitaire* sur une partition 67
2.23 Analogie alignement-partition pour le calcul de la distance unitaire $\not\vdash_1$ 71
2.24 Transférer 6 prématurément produit une opération illicite 72
2.25 Le transfert étendu unitaire et la suppression unitaire sont licites . 72
2.26 Analogie alignement-partition pour le calcul de la distance $\not\vdash_2$ . . 73
2.27 Une chaîne minimale illicite permise par l'analogie 74
2.28 5 qui doit être transféré ailleurs accompagne provisoirement 2 . . 75

3.1 Alignements bons ou mauvais selon le guide d'annotation de Blinker 81
3.2 Dans le cas des omissions, le guide préconise un alignement couvrant 83
3.3 L'alignement proposé peut être incompatible avec les préférences du guide de Blinker 84
3.4 Un compromis est d'élargir les deux alignements 85
3.5 Le travail d'alignement est nominatif 92
3.6 Un corpus et son descriptif 92

3.7 Une biphrase interactive dans Align^{It} 93
3.8 L'ascenseur horizontal en cas de phrase trop longue 94
3.9 Il est possible de décaler une des deux phrases toute seule 95
3.10 Créer un lien simple 95
3.11 Sélectionner un groupe entier par un clic appuyé et le survol des mots 96
3.12 Un double clic sera plus rapide si les mots sont liés ensembles . . 96
3.13 Le clic droit sert aussi à effacer des liens 97
3.14 On peut effacer plus de liens grâce aux touches *Suppr* et *Echap* . . 97
3.15 Les options de navigation dans l'espace de travail 99
3.16 schéma d'une approche d'alignement à base d'exemples 103
3.17 Un alignement manuel collaboratif ouvert 104
3.18 Des fragments d'alignements 106
3.19 Un étiquetage permettra aussi de former des patrons syntaxiques . 107
3.20 Mémoire partielle, alignement partiel 110
3.21 Un patron syntaxique bilingue pour lier les "*fourberies*" 112
3.22 Une mémoire de traduction/d'alignement simplifiée 113
3.23 L'alignement est direct par juxtaposition des fragments disponibles 114
3.24 Un cas d'ambiguité 115
3.25 L'insertion de "patrons syntaxiques" permet de résoudre l'ambiguité 115

4.1 Exemple de SSTC 119
4.2 Exemple de S-SSTC 124
4.3 Un exemple de S-SSTC impropre 126
4.4 Un exemple de S-SSTC non dominante 127
4.5 Alternatives propre et dominante 127
4.6 Cas d'inversion de dominance 129
4.7 L'inversion de dominance dans *"Le docteur lui soigne les dents"/"The doctor treats his teeth"* 130
4.8 Construction des liens ℓ_{node} 132

5.1 Le plongement de ℓ' dans B forme ℓ 137
5.2 La projection de ℓ dans B' forme ℓ' 138
5.3 ℓ' est un sous-alignement de ℓ 139
5.4 ℓ' est un fragment de l'alignement ℓ ($\ell' \triangleleft \ell$) 140
5.5 ℓ' est un fragment compatible avec ℓ 141
5.6 ℓ_1 et ℓ_2 sont composables, ℓ_3 et ℓ_4 ne le sont pas 142
5.7 Les fragments réutilisés sont positionnés selon l'alignement en cours 147
5.8 Exemple ℓ à fragmenter 148
5.9 Fragmentation B 149
5.10 Fragmentation W 149
5.11 Fragmentation X 150
5.12 Fragmentation unilingue par les syntagmes, de type $X1$ et $W1$ . . 151
5.13 Fragmentation bilingue par les syntagmes 153
5.14 Hiérarchie des différentes fragmentations 154
5.15 Pré-alignement découpant la biphrase en $L+1$ 158
5.16 Le problème 2 sur un ensemble de fragments contigus (les liens internes ne sont pas pris en compte) 163
5.17 Deux fragments différents compatibles avec la biphrase mais identiquement positionnés 164
5.18 Le même problème sous sa forme "rectangle" 167
5.19 Le problème 4 monotone sur un ensemble de fragments contigus . 169
5.20 Un cas de chassé-croisé respectant fortement l'hypothèse de cohésion 170

6.1 Pour deux fichiers *.srt* ayant le même timing, l'alignement est naturel 175
6.2 La table graphique des caractères d'un jeu *NES* dans un éditeur de tuiles 176
6.3 Le code hexadécimal à gauche est converti à droite via la table . . 176
6.4 Des marqueurs permettant de séparer les segments, aident à les aligner 178

6.5 Les quatres corpus utilisés. Les langues originales supposées sont en gras 181
6.6 Quelques structures issues des analyseurs évoqués 183
6.7 Les résultats des premières expériences menées 185
6.8 Quatres expériences comparant différents annotateurs 188
6.9 Les distances cumulées exprimées par rapport au nombre de mots parcourus en situation d'alignements indépendants 189
6.10 Les distances cumulées exprimées par rapport au nombre de mots parcourus en situation de post-édition 189

INTRODUCTION

Cette thèse s'inscrit dans le cadre du Traitement Automatique des Langues (TAL), plus précisément dans le domaine de la Traduction Automatique (TA), et son sujet concerne plus spécifiquement l'*alignement sous-phrastique multilingue en situation de corpus parallèle*. L'objet au centre de la TA, le traducteur automatique, est aujourd'hui un outil connu de tous. La qualité des traducteurs grand public bien que critiquée, forme d'ores et déjà une brèche dans cette fameuse "barrière de la langue". La tâche d'une machine de traduction automatique est de porter un texte d'une langue vers une autre en essayant de préserver au mieux le sens (on parle de *texte source* au départ, et de *texte cible* en résultat). On peut retenir trois grandes "*familles*" en TA. La *Traduction Automatique à Base de Règles* (TABR) qui correspond à l'approche historique, utilise généralement des dictionnaires et des outils d'analyse (morphologique, syntaxique et/ou sémantique) qui seront utilisés de concert avec un ensemble de règles de transformation afin de produire la phrase cible (éventuellement via un pivot). La *Traduction Automatique Statistique* (TAS) et la *Traduction Automatique à Base d'Exemples* (TABE) s'appuient sur des ressources textuelles bilingues qui feront office de référence pour les processus de traduction. Les ressources en question sont ce que l'on appelle des *bitextes* ou *corpus bilingues parallèles*. Il s'agit d'un texte disponible en deux langues dont les deux versions sont appariées au niveau des phrases. Parmi les plus connus (et les plus utilisés) incluant le français, on peut citer les débats du parlement canadien formant le corpus *HANSARD* ou les *Acquis Communautaires* regroupant des textes de loi de l'UE.

Un outil de TAS parcourra ces grandes ressources parallèles pour entraîner des méthodes statistiques de résolution, tandis qu'un outil de TABE les utilisera comme base de connaissances pour traduire par analogie (le principe étant, pour une situation nouvelle, de retrouver des cas similaires déjà observés par le passé).

L'*alignement de données bilingues* consiste en l'appariement d'éléments équivalents par le moyen de liens qui commentent le transfert au sein d'un bitexte. Les unités linguistiques alignées peuvent être de tailles variables, en commençant par deux textes, puis par leurs paragraphes, en passant par leurs phrases, jusqu'à des *granularités* plus fines : on

parle dans ce cas d'un **alignement sous-phrastique**. Des méthodes automatiques très efficaces existent pour répondre au problème de *l'alignement phrastique*. Pourtant, le niveau de la phrase constitue une barrière en deça de laquelle la notion d'*unité* (caractère, mot, expression polylexicale, syntagme,...) est moins évidente et le problème de l'alignement y est plus ardu.

Historiquement, l'alignement sous-phrastique a été introduit par le biais de la TAS dans [22] comme une étape interne au processus de traduction. Il est, par ailleurs, envisagé comme un problème indépendant dans [58]. Dans ce deuxième article, le terme d'*"alignement"* est réservé à l'appariement de phrases, tandis que *"correspondance"* sera préféré pour l'appariement de mots. En deux décennies, de nombreux termes ont été employés, parfois ambigus d'une approche à l'autre. Nous utiliserons ici le terme d'*alignement* pour désigner la mise en relation de tous types d'éléments textuels, en précisant toutefois la granularité lorsque nécessaire. En l'absence de précision, il sera question d'*alignement sous-phrastique*.

L'*alignement sous-phrastique bilingue en situation de corpus parallèle* est une problématique de recherche très active en TAL et de nombreux travaux concernent les difficultés techniques et théoriques qu'il soulève. À travers le temps, les modèles ont évolué, d'un alignement de mots [58], vers un alignement d'unités plus longues [82], tentant de tenir compte des phénomènes de transfert d'une plus grande complexité. Des efforts ont été faits pour permettre une plus grande expressivité en agrandissant la taille des unités liées [164],[69], en incluant des discontinuités [13] ou en tenant compte de chassés-croisés importants [6]. Parallèlement, les avancées techniques ont été accompagnées d'apports théoriques fixant les limites claires de faisabilité des modèles d'alignement statistiques ([81],[48]). Par ailleurs de nombreuses approches ont mis à profit l'utilisation d'informations linguistiques pour accroître l'expressivité [115] comme pour réduire le coût combinatoire. Il s'agit là d'autres approches, plus globales, parmi lesquelles se dégagent des systèmes tels que l'analyse syntaxique bilingue simultanée [154] pour laquelle l'alignement de mots est une conséquence, ou l'analyse à base d'exemples par transfert syntaxique ([4], [153]) pour laquelle l'alignement est une nécessité. L'introduction de la syntaxe au cœur de la problématique de l'alignement fait apparaître

des alignements hiérarchiques entre structure et chaîne [159], mais aussi entre structure et structure [98], pouvant ainsi proposer des appariements plus adaptés aux paires de langues considérées. Si dans l'ensemble l'alignement sous-phrastique statistique prédomine, des modélisations plus intuitives s'inspirant de la théorie des graphes existent ([60], [97]), ainsi que des résolutions à base de règles [109].

Le travail présenté ici concerne l'alignement sous-phrastique, questionne sa capacité à faire face à la variabilité des langues naturelles et tente d'apporter des éléments de réponse. Les études théoriques des modèles statistiques prévoient des coût combinatoires exorbitants pour un traitement sur l'ensemble exhaustif des configurations possibles. Désormais, proposer un compromis entre expressivité et complexité algorithmique semble faire consensus parmi les différentes approches existantes. Ainsi, on peut dire que les aligneurs, en tant qu'outils d'observation des biphrases, s'imposent des œillères pour des raisons pratiques. Par ailleurs, différents travaux sur la *divergence* entre langues offrent un recul permettant de s'interroger sur les phénomènes que les aligneurs ne peuvent pas capturer. En se limitant au contexte sous-phrastique, nous avons tâché d'imaginer un modèle très permissif, autorisant des appariements de groupes de tailles variables, discontinus, et éventuellement très intriqués.

Au commencement de cette thèse, il était question d'étudier la contribution de l'analyse syntaxique dans une problématique d'alignement sous-phrastique qui repose traditionnellement sur des informations de surface. Pour cela, nous souhaitions récolter et réutiliser des patrons syntaxiques bilingues de tailles et de formes libres, très similaires à des règles, pour les aligner. L'écriture d'une grammaire de règles de priorité était cependant hors de question. Une application itérative gloutonne selon différents critères (fréquence, longueur, groupes syntaxiques prioritaires) a été testée, mais ne fonctionnait pas vraiment. Un patron, relativement long, et mal positionné empêchait l'application d'au moins deux autres, ce qui créait, en cas d'erreur, des alignements à la fois faux et vides. Un traitement récursif sur des unités plus fines via l'utilisation d'analyses syntaxiques profondes a été envisagé, mais rendait l'approche dépendante d'outils encore rares pour beaucoup de langues, et peu robuste aux erreurs d'analyse.

En se ramenant à un problème de maximisation, les approches statistiques permettent

d'appréhender les liens dans leur globalité plutôt qu'individuellement. Par ailleurs, nous avons observé des similitudes entre des alignements manuels provenant de corpus de référence et certains de nos patrons syntaxiques, ce qui nous a mené à envisager, comme critère global, celui de la *couverture maximale*. En effet, une application juste des patrons aura tendance à peupler harmonieusement la biphrase considérée, alors qu'au contraire les erreurs seront génératrices de vides. Un alignement se basant sur un critère de couverture présentera également l'avantage de traiter simultanément le problème de la segmentation en unités bilingues sous-phrastique et celui de l'appariement, qui sont bien souvent séparés dans les approches d'alignement de segments sous-phrastiques. Cette démarche, qui adapte un axe principal de la TABE, orientera l'ensemble du travail décrit ici.

Le chapitre 1 abordera les travaux existant dans le thème de l'alignement sous-phrastique en insistant sur les efforts des différents modèles pour tenir compte de phénomènes de complexité croissante (discontinuités, longs syntagmes,...). Nous rapprocherons ensuite la problématique de l'alignement de différentes études sur la divergence entre langues. En confrontant des alignements produits par des outils automatiques à des alignements motivés linguistiquement, nous dégagerons le thème général, à savoir celui de l'expressivité des modèles d'alignement sous-phrastique. Plus précisémment, notre travail s'articulera autour de la mise en place d'une plate-forme collaborative permettant l'alignement manuel via une interface similaire à celle du projet *Blinker* [95] connu pour avoir contribué à créer une ressource de grande qualité. Nous souhaitons aider la tâche de création d'alignements de qualité en l'assistant par des outils automatiques évolutifs reposant sur une mécanique à base d'exemples. Cela nous conduit alors à un double objectif, la création d'une ressource librement accessible et la création d'outils d'alignement automatique hautement expressifs. Une soin particulier a été accordé à la formalisation de la notion d'alignement, et une partie importante des travaux présentés consiste à proposer des opérations canoniques, des métriques et des représentations structurelles. Cet apport théorique accompagnera l'ensemble de notre propos, et nous permettra d'argumenter en faveur des axes défendus et sous-tendant l'architecture générale de l'approche. Le chapitre 3 sera consacré à la description du cadre collaboratif accompagnant notre ap-

proche, et abordera le thème de la qualité en argumentant en faveur d'une approche non experte. L'outil d'acquisition et son fonctionnement seront présentés. L'outil collaboratif sera le premier élément d'une approche plus large d'*alignement à base d'exemples* pour laquelle nous proposerons une architecture capable de tirer avantage d'analyseurs morphosyntaxiques de différentes profondeurs. Nous définirons au chapitre 4 les modèles de représentation utilisés pour unifier de manière cohérente les biphrases alignées avec les deux analyses syntaxiques associées, les *S-SSTC*[1]. Le chapitre 5 présentera en détail deux composantes nécessaires à la partie automatique de cette architecture d'*alignement à base d'exemples* : une mémoire de fragments et des algorithmes d'alignement reposant sur cette mémoire. Les fragments d'alignement qui avaient été déjà appréhendés par des exemples seront formalisés, ainsi que les traitements nécessaires à la formation d'une mémoire de fragments. Nous proposerons plusieurs types de fragmentation d'expressivités égales. Les conséquences des choix de fragmentation sur les problèmes de résolution seront observées en termes de complexité. Nous y présenterons les algorithmes de résolution, des heuristiques, ainsi que différentes stratégies possibles. Finalement, nous présenterons au chapitre 6 les ressources utilisées ainsi que les résultats d'expériences menées dans le cadre de cette thèse.

[1] Synchronous Structured String-Tree Correspondance

CHAPITRE 1

PROBLÉMATIQUE ET ÉTAT DE L'ART

1.1 L'alignement

1.1.1 Notions

Aligner consiste à mettre en relation des éléments semblables selon certains critères dans un contexte donné. Il existe dans le cadre du TAL plusieurs domaines présentant des problèmes d'alignement. La *reconnaissance de l'écriture manuscrite*, qui permet de substituer aux claviers les nombreux outils tactiles, associe des représentations picturales à des chaînes de caractères. La *reconnaissance automatique de la parole*, permettant d'étendre l'utilisation de certaines interfaces homme-machine à la voix, s'intéresse à aligner des séquences "observées" de vecteurs acoustiques avec du texte. Nous nous intéresserons ici à l'alignement de données textuelles bilingues, c'est-à-dire à la mise en relation des éléments qui sont en relation de traduction. On appelle *bitexte* ou *corpus bilingue* (respectivement *multitexte* ou *corpus multilingue*) *une paire* (respectivement *un n-uplet*) de textes étant en relation de traduction. Le corpus multilingue le plus célèbre est sans doute la pierre de Rosette, présentant un décret du pharaon gravé en deux langues, l'égyption ancien (en écritures hyéroglyphique et démotique) et en grec ancien.

On appellera *biphrase* une paire de phrases en relation de traduction. En pratique, il ne s'agit pas d'une association complètement symétrique car l'une traduit l'autre. Le terme de *biphrase* sera étendu à l'appariement de deux groupes de plusieurs phrases, car comme on le voit sur la figure 1.1, dans un bitexte, une phrase peut souvent être la traduction de plusieurs autres. *Aligner un bitexte* au niveau des phrases revient à le segmenter en biphrases ([59],[83]) et le résultat de cette opération est ce que l'on appelle un *corpus parallèle bilingue*.

La même problématique portée à un niveau inférieur, l'*alignement sous-phrastique*, consiste à apparier des "*unités plus fines que la phrase*". L'unité visée ici devient moins claire : certaines approches alignent des *chaînes de caractères* [15], d'autres traitent

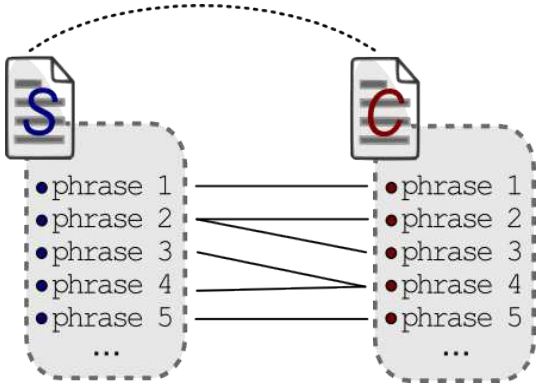


Figure 1.1 – Un corpus parallèle bilingue

d'alignement de *mots* [58], de *groupes de mots* [164] ou encore de *syntagmes* [69] [126]. Les notions de mot ou de syntagme semblent parfois inadaptées selon la langue et le système d'écriture considérés et même si l'on peut s'accorder à définir des unités minimales de l'ordre du caractère, elles ne sont en général pas très porteuses de sens [143]. L'alignement sous-phrastique pose donc d'emblée un problème de *segmentation* [65] ou plutôt de *fragmentation*. Le terme de *segmentation* est généralement utilisé pour désigner la division d'un texte en phrases ou groupements de phrases. Il n'est pas rare que son utilisation soit étendue au cas de la division d'une phrase en sous-unités contigües (intervalles de caractères ou de mots). Nous lui préfèrerons le terme plus général de **fragmentation**, qui selon le contexte, pourra désigner la division d'une phrase ou d'une biphrase en sous-éléments non nécessairement contigus. Nous proposerons en section 2.2.2 une définition formelle des alignements, basée sur la *fragmentation*. On remarque qu'une fragmentation bilingue induira deux fragmentations monolingues, mais le contraire n'est en général pas vrai.

Il existe deux problématiques proches auxquelles fait référence le terme ambigu d'*alignement sous-phrastique*. La première, reposant sur des *corpus bilingues parallèles*, est plus proche de la TA, l'autre, reposant sur des *corpus bilingues comparables*, concerne la constitution de ressources bilingues. Il convient généralement de préciser le cadre considéré. Les *corpus comparables bilingues* sont des textes en plusieurs langues partageant sujet, domaine ou style, et pour lesquels une grande partie du vocabulaire est

commune (dans chacune des langues présentes). Par exemple, les articles de Wikipedia en plusieurs langues forment un *corpus multilingue comparable*. Ils peuvent sembler d'un intérêt moindre par rapport aux corpus parallèles, mais sont plus nombreux et plus faciles à constituer ([50], [120]). Les techniques mises en jeu et les applications sont évidemment différentes, et nous parlerons ici toujours, sans le préciser, d'*alignement sous-phrastique en situation de corpus parallèle bilingue*.

Il y a deux manières équivalentes de représenter une *biphrase alignée*. La première utilise des *matrices binaires* où les lignes et les colonnes sont indicées par les *unités sous-phrastiques*, et dans lesquelles un **1** désigne une correspondance, un **0** l'absence de correspondance. La deuxième représentation emploiera les formes naturelles d'un graphe biparti dont les nœuds sont les "mots" et les liens les correspondances. Nous utiliserons la représentation de type graphe tout au long du manuscrit, en simplifiant visuellement les ensembles complètement liés (appelés *bicliques*[1] en théorie des graphes) comme on peut le voir en figure 1.2.

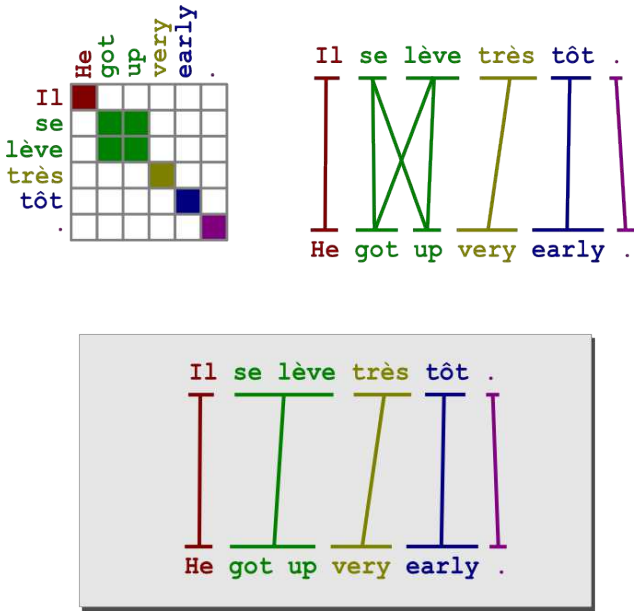


Figure 1.2 – Représentations d'un alignement sous-phrastique

[1] aussi *graphe biparti complet* qui a le mérite d'être plus explicite

1.1.2 L'alignement sous-phrastique

1.1.2.1 Lien avec la traduction statistique

La première idée d'utiliser une machine afin d'aborder le problème de la traduction est communément attribuée à Warren Weaver dans son mémorandum [149] de 1949. L'effort de traduction y est envisagé comme un problème de décodage : une phrase cible *C* est vue comme une phrase source *S* encodée.

> *It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code".*

Après seulement deux années, les jalons sont plantés dans le premier état de l'art de *traduction mécanisée* [14]. Les directions suivies dès lors verront naître les systèmes de TA dits de première génération, dans un paradigme à base de règles. Un panorama du TAL assez précis aborde cette période dans [21] et un historique de la TA peut être trouvé dans [68]. Il faudra alors attendre 1990 pour voir resurgir une approche statistique de la TA portée par l'amélioration grandissante des outils informatiques, ainsi que la multiplication des ressources textuelles numériques. Le cadre devient favorable pour adapter à la traduction le modèle du canal bruité déjà utilisé en reconnaissance de la parole (décrit dans [12]), dans l'article fondateur [22]. C'est également l'apparition du problème de l'*alignement sous-phrastique* qui fait alors partie intégrante du modèle de traduction. Dans cette nouvelle approche, des modèles sont entraînés sur des bitextes, et des alignements entre mots sont formés. Ces travaux ouvrent la voie à la *Traduction Automatique Statistique* (**TAS**)[2] qui occupe toujours une place centrale dans le domaine de la **TA**. Parallèlement, on s'intéresse à la construction automatique de corpus bilingues parallèles qui sont un prérequis impératif pour les modèles nouveaux de la TAS [77] [57]. Une automatisation totale fait parti des promesses de la TAS, revendiquant un fonctionnement indépendant (contrairement aux systèmes nécessitant une maintenance des règles de traduction). La problématique de l'*alignement sous-phrastique* est rapidemment étudiée comme une problématique indépendante, notamment dans les travaux de Gale et Church

[2] *Statistical Machine Translation* en anglais (**SMT**)

[58] où sont envisagées des applications autres que la traduction (essentiellement les concordanciers bilingues et la désambiguïsation translingue).

1.1.2.2 Applications

Les applications premières de l'alignement sous-phastique se trouvent en TA. Un tel alignement sert à entraîner des modèles de traduction et à construire des tables bilingues de groupes de mots pour la TAS. Certaines approches de TABE y ont également recours pour construire une mémoire par la fragmentation d'exemples [5] [165] [26]. La technique d'alignement sous-phrastique pourra être mêlée plus marginalement à des approches à base de règles, en permettant d'assurer la maintenance et le renouvellement de règles par des méthodes d'apprentissage ([28], [107]). Bien que fortement associé à la TA, l'alignement a également démontré une utilité plus générale car on l'utilise dans de nombreux problèmes du TAL, tels que :

-L'amélioration de concordanciers [67]
-Les outils d'aide à la traduction [62]
-La désambiguïsation lexicale [36]
-La construction de ressources lexicales ou terminologiques [58] [45]
-La recherche d'informations translingue ([102])
-L'analyse syntaxique d'une langue par rapport à une autre via le transfert [153] [5]
-La génération de paraphrases [114]
-L'expansion de requêtes [118]
-Les systèmes de question-réponse [53]

La pertinence d'un alignement dépend aussi de l'application souhaitée. Nous verrons dans la partie suivante que les outils existants proposent des représentations et des niveaux d'expressivité variés.

1.1.3 Des modèles de différentes expressivités

Les travaux concernant l'alignement continuent d'évoluer pour prendre en compte des phénomènes de plus en plus complexes. Certains modèles tâchent de reconnaître des

croisements importants qui peuvent apparaître lorsque l'on considère deux langues très différentes. La traduction ne se faisant pas mot à mot, certaines approches cherchent à traiter le problème de l'alignement de groupes de mots. Une difficulté supplémentaire vient du fait que des unités élémentaires de la phrase peuvent aussi faire sens de manière non contigüe.

1.1.3.1 Alignements monotones

Un alignement sera dit **monotone** si aucune paire de liens ne se croise. Typiquement, les *alignements de phrases* au sein d'un *bitexte* sont monotones. Même si cette contrainte n'est pas toujours vraie en pratique, les algorithmes visant l'alignement des phrases ont tendance à se baser sur cette hypothèse simplificatrice. Différents travaux sur les bitextes constatent la "simplicité" de l'*alignement de phrases* relativement à un *alignement de mots*, et obtiennent des résultats de bonne qualité en n'utilisant que peu ou pas d'informations lexicales. Les travaux décrits dans [77] fonctionnent itérativement en se basant sur un alignement rudimentaire entre mots. Un autre type d'approche consiste à ne tenir compte que des longueurs des phrases [23] [57]. Nous pouvons aussi donner l'exemple de travaux utilisant des marqueurs externes comme des ancres tels que les indices temporels dans les sous-titres de films [138] [145].

Si l'on observe un alignement monotone à un niveau sous-phrastique, c'est que la traduction est littérale. Le phénomène se produit plus souvent entre langues proches (gaélique irlandais/gaélique écossais, castillan/catalan, l'ensemble des dialectes vénitiens, le français et certaines langues régionales, etc). Le parallélisme de telles langues proches, donne lieu à des traitement translingues utiles par exemple pour des langues peu dotées [122] [30]. Un exemple à la fois original et très significatif de la simplification apportée par l'hypothèse de monotonie est celui de travaux d'alignement réalisés entre le français et le "*français SMS*" (le corpus de Louvain [54]). Les biphrases de ce corpus ne sont pas en relation de traduction, mais plutôt de correction, ce qui a permis un alignement au niveau des caractères optimal de coût polynomial dans [15] (voir aussi [42]) car les opérations élémentaires pour transformer le français correct en français SMS sont des substitutions, des insertions et des suppressions mais jamais de permutation.

1.1.3.2 Alignement de mots

Des situations de croisement interviennent dès lors que l'on aligne des unités plus fines que la phrase entre deux langues quelconques. Nous abordons ici l'*alignement de mots* (ou *alignement simple*) dans lequel on fait l'hypothèse que les correspondances sont injectives [58] [137] [97] (un mot est aligné au plus à un autre, éventuellement à aucun). Réduire l'alignement à un paradigme *injectif*, ne semble pas très pertinent, d'un point de vue linguistique, mais les différentes approches liant simplement des mots ont souvent l'avantage d'être plus précises et moins difficiles. Produire un alignement simple constitue parfois la première étape de traitements visant à apparier des unités plus longues par propagation [107] ou par discrimination [58]. Les unités liées sont alors appelées des *ancres sûres*. Par ailleurs, le problème de l'alignement de mots peut se ramener à un problème connu de *couplage dans un graphe biparti*, pour lequel il existe des solutions polynomiales, notamment par la *méthode hongroise* [97] [136].

Lorsque l'hypothèse d'injectivité n'est pas faite, des alignements entre groupes de mots peuvent apparaître, ce qui présente à la fois un avantage et une faiblesse. Certes, il est alors possible de rendre compte de correspondances plus complexes et plus légitimes (les verbes anglais à particule sont un bon exemple), mais il sera parfois difficile d'interpréter les raisons d'un alignement groupant. En effet, des liens groupants pourront refléter la présence d'un syntagme cohérent, mais tout aussi bien indiquer une hésitation de la part du système, auquel cas il faut envisager chaque lien comme une branche possible d'une alternative (voir figure 1.3). On parle parfois de "*liens sûrs*" pour désigner les liens simples et de "*liens flous*" ou "*probables*" pour les liens multiples [82] [38].

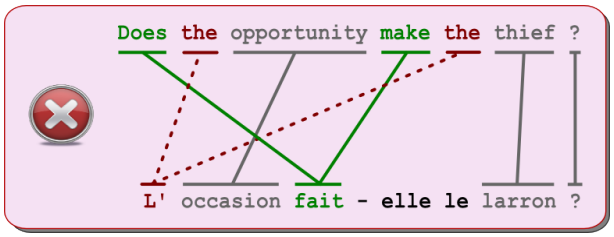


Figure 1.3 – Un alignement de mots non-injectif posant problème

1.1.3.3 Alignements asymétriques

L'article fondateur [22] introduit un type d'alignement asymétrique obtenu lorsque l'hypothèse d'injectivité n'est pas faite. La traduction y est introduite comme un problème de maximisation. L'élément de départ est une phrase source $s_1^J = (s_1, \ldots, s_J)$ et l'on cherche à produire une traduction $c_1^I = (c_1, \cdots, c_I)$. Pour chaque phrase, il est possible d'évaluer sa pertinence selon une mesure de probabilité P s'appliquant sur les suites de mots. La phrase c_1^I recherchée est une de celles qui maximisera la probabilité conditionnelle $P(c_1^I|s_1^J)$:

$$\hat{C}_1^J = \arg\max_{c_1^I} \left(P(c_1^I|s_1^J)\right) = \arg\max_{c_1^I} \left(P(c_1^J) \cdot P(s_1^J|c_1^I)\right)$$

La formule de Bayes est utilisée pour séparer les deux aspects du problème que sont la *bonne formation* de la phrase cible produite et le *réordonnancement* des mots au travers de la traduction. Le modèle de langue évalue de manière statistique la bonne formation d'une phrase dans la langue cible par le calcul de $P(c_1^I)$. Une hypothèse de localité est faite sur la dépendance d'un mot par rapport à ses prédécesseurs dans la phrase selon un modèle de n-grammes [25] :

> *"In an n-gram langage model, we treat two histories as equivalent if they end in the same n - 1 words"*

Le *modèle de traduction* donne une estimation statistique du transfert entre la langue source et la langue cible par le calcul de $P(s_1^J|c_1^I)$. La notion d'**alignement** est introduite pour permettre de décomposer le modèle et de l'évaluer :

$$P(s_1^J|c_1^I) = \sum_{a_1^J} P(s_1^J, a_1^J|c_1^I)$$

Pour les besoins du modèle théorique, les alignements considérés par le modèle étaient alors représentés par des fonctions $a : [\![1,J]\!] \mapsto [\![1,I]\!]$, ce qui impose donc qu'un mot source puisse être lié à plusieurs mots cible, mais pas le contraire. On parle d'alignements "*un à plusieurs*" ou *asymétriques*.

Dans leur article [24] les auteurs décrivent différents modèles de traduction connus comme étant les *quatre modèles successifs d'IBM* permettant de tenir compte de phénomènes de plus en plus complexes. Pour chaque modèle, il existe un alignement a entre S et C qui a la meilleure probabilité relativement aux paramètres du dit modèle, que l'on appelle l'alignement de Viterbi. Pour *IBM-1* et *IBM-2*, l'alignement de Viterbi est calculé de manière exacte, pour les autres, de manière approchée. Les travaux de [144] apportent une variante au modèle *IBM-2* pour renforcer la dépendance locale entre les unités alignées grâce à un modèle *HMM* (*Hidden Markov Model*). Ces modèles sont implémentés dans la boîte à outils *Giza++*[3] [106] qui reste un outil de référence dans de nombreux travaux.

La forme asymétrique de ces alignements a été critiquée, notamment parce qu'elle dégraderait la qualité des traductions [155]. Les alignements asymétriques de *Giza++* sont souvent utilisés après des traitements heuristiques de symétrisation produisant un alignement de mots ou un alignement de groupes de mots. En renversant l'ordre source-cible des langues, on obtient deux types d'alignements "*un à plusieurs*" $a_{S\mapsto C}$ et $a_{C\mapsto S}$. L'opération d'intersection qui consiste à ne conserver que les liens présents dans les deux sens produit un alignement de mots $a_{\cap}$ bien plus précis (utilisé dans [107] par exemple). Une opération d'*union étendue*, décrite dans [82] a été notamment utilisée pour produire des tables de groupes de mots pour des approches d'*alignement de groupes de mots.*

La faisabilité d'un système de TAS a été étudiée d'un point de vue purement théorique ; il semblerait que la résolution du problème ainsi posé soit d'un coût élevé. Trouver une phrase maximisante, en considérant un réordonnancement quelconque des mots, est un problème NP-difficile [81]. Cette difficulté donne lieu à nombre d'adaptations parmi lesquelles des approches heuristiques, des approximations, et l'expérimentation de modèles alternatifs.

1.1.3.4 Alignement de groupes de mots

Les modèles statistiques classiques ont été adaptés afin de tenir compte de phénomènes plus larges encapsulant des dépendances locales. Différents travaux se sont

[3]Disponible à l'adresse `http://code.google.com/p/giza-pp/downloads/list`

donc intéressés à lier des groupes de mots sous différentes appellations comme "*templates*", "*spans*", "*classes*", "*shallow phrases*, "*concepts*" ou "*n-grammes de mots*". Formellement, il s'agit de *sous-ensembles contigus de mots* (nous employerons le terme d'*intervalle de mots*). L'ensemble des alignements d'*intervalles de mots* sur une biphrase correspond donc à des liens simples entre intervalles de mots source et cible. Il y est parfois fait référence comme l'ensemble des *alignements bijectifs d'intervalles de mots* (*Bijective Phrase Alignments* en anglais[4]).

Des modèles s'appuient sur une *fragmentation bilingue* statistique des biphrases [106] [7]. Dans ce type d'approche la *"bifragmentation"* permet de renforcer la localité des alignements et de diviser le problème global sur des "îlots". Par ailleurs, [82] envisage plutôt de s'attaquer au problème en recourant à des *tables d'intervalles de mots* (*phrase table*) constituées automatiquement (notamment en utilisant les heuristiques de symétrisation des modèles IBM évoquées ci-dessus). Les auteurs préconisent l'utilisation d'intervalles de mots assez courts (trois mots maximum), le gain au delà étant minime. L'approche est étendue dans [40] à des *segments hiérarchiques*, c'est-à-dire des intervalles de mots pouvant en contenir d'autres. Ils partagent des caractéristiques avec des groupes de mots non contigus et s'apparentent à des règles de réordonnancement de sous-intervalles.

Le problème de l'alignement de groupes de mots souffre a fortiori de difficultés similaires à celles des modèles asymétriques. En effet, il est démontré que le problème d'optimisation est NP-difficile dans [48]. Les solutions envisagées dans différents travaux tentent donc de contourner le problème en se plaçant par exemple sur des espaces restreints (voir par exemple [39]) ou en proposant des solutions approchées [91]. Les travaux les plus récents tentent de s'extraire de la traditionnelle expressivité "asymétrique" parmi lesquels on peut citer Anymalign [87], qui met à profit les mots de basse fréquence, ou encore TransTree [37] qui propose une structuration hiérarchique, les arbres d'amphigrammes, raffinant les bi-segments en correspondances traductionnelles gigognes.

[4]Nous prendrons soin d'appeler *intervalle de mots* des ensembles de mots contigus, même si le terme anglais fait plutôt référence à des syntagmes qui désigne, selon nous, un ensemble plus général

1.1.3.5 Alignement de groupes non contigus

Une phrase n'est pas toujours une succession de syntagmes contigus. En fait, les syntagmes non contigus sont plutôt communs dans le langage naturel : nous pouvons donner les exemples connus de la négation en français ou des verbes à particule en anglais. Pour ces cas simples, la raison du défaut de contiguïté est *monolingue*. Dès lors que nous nous plaçons dans le cas plus complexe de la bilingualité, les unités non contigües apparaissent d'autant plus, notamment à cause des omissions et des anaphores. Nous pouvons observer ces deux cas sur la figure 1.4.

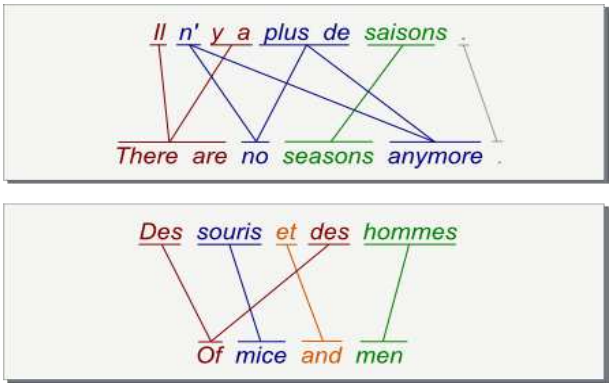


Figure 1.4 – Deux types de groupes non contigus : cas monolingue et cas bilingue

Dans le premier exemple, les groupes non contigus intriqués | *Il ◇ y a* | *There are* | et | *ne ◇ ◇ plus de* | *no ◇ anymore* | sont l'appariement de deux groupes monolingues non contigus. Dans le deuxième, on voit que pour aucune des deux langues il n'y a de groupe monolingue non contigu, pourtant le *"bi-intervalle"* | *Des ◇ ◇ des* | *Of* | présente un trou de longueur 2 côté source, due à la disparition d'une préposition en anglais (relativement au français). Le défaut de contiguïté est dit *bilingue* car il dépend ici de l'opposition des deux langues.

Les systèmes utilisant des tables d'intervalles de mots, évoqués précédemment, ne peuvent pas tenir compte des phénomènes non contigus autrement qu'en les englobant dans des intervalles plus longs, sur des cas précis ([91]).

Certains travaux se sont intéressés aux phénomènes non contigus en proposant des

modèles tenant compte d'intervalles utilisant des *joker* (notés "◇" dans les exemples ci-avant). On peut citer les approches de [134] et [13] plus récemment.

1.1.3.6 Alignements croisants

Les croisements sont une difficulté majeure dans l'alignement sous-phrastique puisque nous l'avons vu, chercher une solution sans contrainte de réordonnancement est un problème NP-difficile.

Dans les modèles IBM [24], cette restriction se fait via un paramètre dit de *distorsion* grâce auquel tous les alignements ne sont pas équiprobables et dépendent notamment des positions des mots dans la biphrase. Les modèles 4 et 5, en complexifiant les paramètres, tiennent compte de distorsions relatives, permettant à des mots liés localement de conserver leur proximité en passant à l'autre langue. Des approches utilisant des tables d'intervalles de mots [82] ou des segmentations de la biphrase [139] entraînent également des modèles de distorsion relative pour réordonner les intervalles. Des travaux proches entraînent leur paramètre de distorsion grâce à un pré-alignement de mots [6] utilisé pour décrire les permutations locales entre les langues.

Une approche différente consiste, pour une des deux langues, à réordonner dans chaque phrase des ensembles de mots afin de rendre les deux langues plus "semblables". L'intérêt est alors de pouvoir y appliquer des techniques d'*alignement monotone* ou presque (voir par exemple [75], [41], [66]). On parle de techniques de *monotonisation*. Des analyses syntaxiques permettent un réordonnancement motivé linguistiquement (voir section suivante 1.1.3.7).

1.1.3.7 Alignement et syntaxe

La traduction se fait par modifications locales et permutations de syntagmes, ce qui implique l'importance de considérer, d'une part, des unités longues et d'autre part des réordonnancements entre syntagmes. C'est du moins l'hypothèse que semblent faire beaucoup de travaux concernant l'alignement sous-phrastique. Cette *hypothèse de cohésion* est vérifiée expérimentalement dans [56] sur un corpus aligné à la main. Les résultats

confirment le bien-fondé de l'hypothèse sur la paire de langues testée, c'est-à-dire l'anglais et le français. Y sont évoqués les travaux de [161] qui observent la forte cohérence des syntagmes nominaux qui, au travers de la traduction, ont tendance à rester groupés, même si d'importants déplacements internes peuvent se produire (les langues considérées incluaient l'anglais et le tchèque où l'ordre des mots est relativement libre). Les modifications internes aux syntagmes via la traduction sont appelés des *divergences de constituants* et les permutations de syntagmes des *divergences de dépendances* dans les travaux de [19] et [113]. Ces deux derniers mesurent expérimentalement la portée de l'hypothèse de cohésion entre, d'une part, le français et l'anglais et d'autre part, le français et l'allemand. Les auteurs se basent sur une grammaire de règles de transformation d'arbres syntaxiques écrites manuellement pour passer d'une langue à l'autre sur les corpus de test. Il y est constaté une plus grande proportion de divergences de dépendances entre l'allemand et le français, tandis que pour le français et l'anglais, plus proches, les divergences de constituants sont majoritaires.

Aborder l'alignement sous-phrastique (et la traduction) en considérant l'hypothèse de cohésion justifie l'utilisation d'analyses syntaxiques. Bien sûr, l'approche devient alors tributaire des outils d'analyse disponibles qui peuvent être de types différents, dépendent des ressources de la langue, et commettent des erreurs. Il existe des analyseurs de différentes profondeurs : racineur[5], lemmatiseur, étiqueteur morphosyntaxique, fragmenteur[6], analyseurs syntaxiques produisant une structure profonde en constituants ou en dépendances et/ou des informations sémantiques,... Ils permettent d'enrichir par des informations linguistiques plus ou moins fines les formes de surface. Leur application est répandue au travers du TAL. En ce qui concerne l'alignement sous-phrastique, la syntaxe offre notamment des perspectives intéressantes pour tenir compte des difficultés d'expressivité déjà évoquées : l'alignement de groupes éventuellement non contigus et les croisements importants.

[5] *stemmer* en anglais
[6] *chunker* en anglais

Alignement de constituants : les structures syntaxiques ont servi de base à de nombreux travaux souhaitant étendre l'alignement de mots à des unités plus longues comme les *chunks*[7] ou les syntagmes [73]. Des paires de chunks bilingues déduits d'alignements hiérarchiques sur des arbres d'analyse servent notamment à étendre le vocabulaire pour des approches statistiques et à dépasser le simple alignement de mots [69] [147] [148]. Cependant, l'utilisation de syntagmes dans des tables pour la TAS a fait débat, notamment car dans [82] on l'évalue comme détériorant le processus de traduction par rapport à une approche purement statistique [8]. D'autres approches nuancent cette affirmation en proposant des modélisations similaires au problème, utilisant favorablement des analyses syntaxiques [121] ou des segments hiérarchiques "*syntaxiquement motivés*" [40].

Alignements non contigus : il y a assez peu de travaux recourant à la syntaxe pour traiter exclusivement des problèmes de non-contiguïté en traduction ou en alignement. En général, sont abordés conjointement des problèmes de réordonnancement et de non-contiguïté comme c'est par exemple le cas pour l'alignement de segments hiérarchiques.

On peut tout de même citer [44] où les phénomènes non contigus sont prétraités par des règles syntaxiques de séparation. De même que des traitements de monotonisation visent à réduire les croisements, les règles de dédoublement réduisent les phénomènes non contigus : par exemple, la biphrase | *je ne veux pas* | *I don't want* | sera transfomée par une des règles en | *je ne veux pas* | *I don't*$_1$ *want don't*$_2$ | se résolvant par un alignement monotone, comme on peut le voir à la figure 1.5.

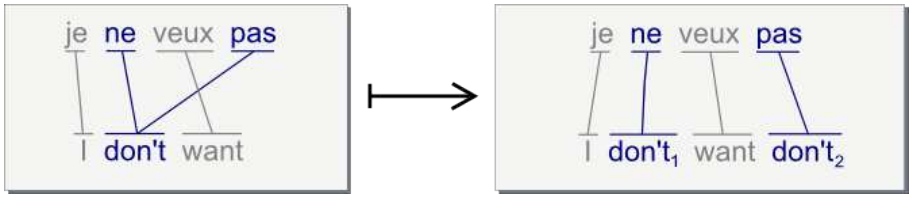


Figure 1.5 – transformation d'un alignement discontinu

[7] Par le terme *chunk*, on désigne une séquence formée de mots contigus et qui n'est pas récursive

[8] On pourra remarquer que les courbes de *KoehnOM03* comparant les processus avec ou sans analyse syntaxique expriment leurs performances par rapport à la taille des corpus d'entraînement, ce qui désavantage clairement la table syntaxique, naturellement plus petite puisque obtenue par filtrage. Une visualisation par rapport à la taille des tables donne un résultat moins tranché.

La syntaxe sert à généraliser des transformations purement lexicales pour plus de généricité. On remarque que, dans cet exemple, déplacer la négation du côté français règle autant le problème que le dédoublement du côté anglais. De plus, pour des cas comme celui de la négation, il n'est pas rare que des analyseurs syntaxiques non-projectifs réunissent les deux parties.

Nous n'aborderons pas le thème complexe de la détection des *anaphores grammaticales* qui participent aussi grandement à la présence de groupes non contigus. Un exemple : *"Hachiko a attendu **son maître** jusqu'à la fin. Bien sûr, **il** n'est jamais reparu."*

Alignements croisants : nous avons vu que l'ordre des mots pose un problème de complexité pour l'alignement sous-phrastique et qu'il convient de limiter les permutations envisagées dans la phase de résolution. Mais l'ordre des mots est aussi et avant tout de nature linguistique, notamment régi par des lois syntaxiques. Une restriction sur les permutations prenant en compte les syntaxes des deux langues est donc une idée qui semble légitime et adaptée.

Il est alors assez naturel d'observer que de nombreux travaux abordant l'alignement de l'anglais avec le japonais (où le verbe est situé à la fin de la phrase) recourent à des représentations syntaxiques ([93] [159] [158]) quand par ailleurs, des travaux visant à aligner le français et l'anglais (une paire de langues syntaxiquement plus proches) s'en passent plus souvent. Certaines approches statistiques sont étendues pour tenir compte de structures syntaxiques [159] ou hiérarchiques [40] tandis que d'autres les utilisent séparément en prétraitement pour réordonner les syntagmes et y appliquer des méthodes d'alignement plus monotones [157] [158] (monotonisation par la syntaxe). Notons que certaines approches se dégagent entièrement des considérations sur l'ordre relatif des deux langues en propageant itérativement les alignements entre les nœuds des deux structures suivant leurs dépendances et/ou dominances [107] [63] [98].

Si l'analyse syntaxique peut aider à l'alignement, il existe des approches doublement motivées, rapprochant les deux problématiques. Les ITG (*Inversion Transduction Grammar*) [154] sont des grammaires hors-contexte bilingues (applicables à deux phrases en relation de traduction). Leurs règles permettent de décrire simultanément les ordres des

constituants dans les deux langues, pour construire une structure syntaxique double induisant un alignement. Les solutions d'alignement apportées ont l'avantage d'une part d'être calculables en temps polynomial, et d'autre part de rendre compte de chassés-croisés importants. Ces solutions sont appliquées au couple anglais/chinois dans l'article d'origine. En contrepartie, certains alignements ne sont pas exprimables car on ne peut les déduire d'aucun arbre d'analyse. L'espace des alignements que les ITG permettent d'atteindre est l'ensemble des alignements injectifs de mots, privé des configurations dites "*inside-out*", c'est-à-dire des permutations $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}$ et $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$ [151]. Nous pouvons les observer à la figure 1.6 parmi des alignements déduits d'arbres ITG valides.

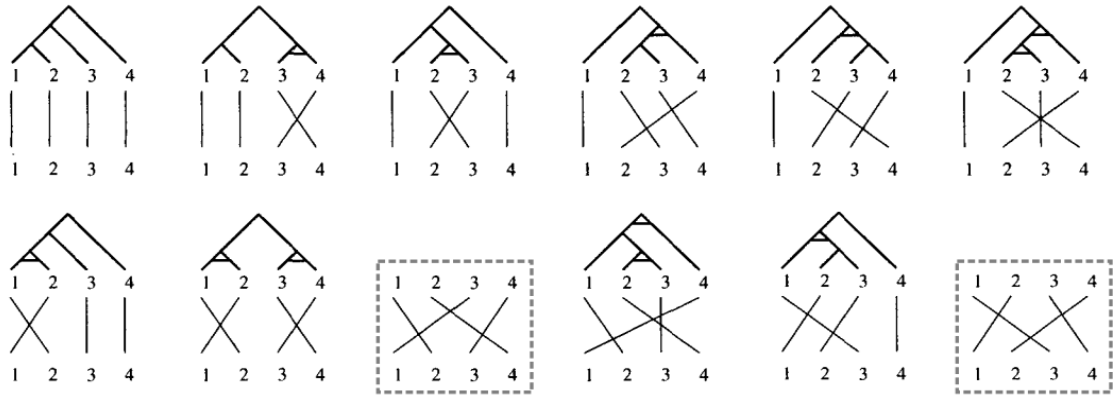


Figure 1.6 – Deux alignements ne sont pas atteints pour deux phrases de longueur 4 : "inside-out"

1.2 De la divergence en traduction

Traduire une phrase dans une autre langue se fait rarement sans détour. En effet, deux langues naturelles ne possèdent jamais des lexiques en correspondance parfaite, ni même des catégories grammaticales parfaitement similaires et bien souvent, la traduction naturelle d'une phrase produit un résultat d'une structure assez différente de la structure initiale. On peut dire que la traduction ne se fait pas mot à mot en général et que c'est la raison pour laquelle un alignement sous-phrastique bilingue n'est pas toujours une succession de liens simples et parallèles. On parlera de **phénomène**

de divergence lorsque la traduction naturelle d'une langue dans une autre produit des formes dissemblables. Nous nous focaliserons sur des phénomènes localisés au sein de la phrase. Certains travaux insistent sur l'importantce des phénomènes co-textuels faisant intervenir plus d'une phrase [83] mais ils sortent du cadre de notre travail. L'étude de la divergence prend ses racines dans la linguistique constrastive dont les motivations applicatives sont de faciliter l'apprentissage d'une langue en la comparant à la langue maternelle de l'apprenant. Nous abordons ici ces phénomènes linguistiques, intervenant à un niveau sous-phrastique, qui ont tendance à complexifier l'effort de traduction. Les phénomènes de divergence qui sont une difficulté pour l'apprenant le sont tout autant pour des systèmes de traduction ou des systèmes d'alignement : ils représentent des verrous que les modèles actuels prennent difficilement en compte, au mieux partiellement. En situation d'alignement sous-phrastique bilingue, des phénomènes de divergence se cachent derrière des liens groupe à groupe, des liens non contigus, des groupes de liens croisés, etc.

1.2.1 Premier obstacle : les expressions

Un travail de traduction entre deux langues, même très proches, se fera toujours au travers du prisme de la divergence. Elle sera à première vue lexicale, mais se révélera en général également grammaticale et stylistique. L'exemple le plus évident de situation dans laquelle une traduction ne peut être littérale est celui des expressions idiomatiques telles que "*Manger les pissenlits par la racine*" se traduisant par "*to kick the bucket*" en anglais. Il s'agit de formes figées, non-ambigües et métaphoriques. Elles sont donc généralement non-compositionnelles [9] bien que parfois suffisement imagées pour qu'un apprenant de la langue puisse les comprendre en contexte comme dans "*You can't put the clock back*" signifiant "*On ne peut pas revenir en arrière*". Cependant, sans qu'il ne la connaisse au préalable, il est peu probable d'imaginer l'apprenant employer une telle expression. Une expression idiomatique ne se traduit pas nécessairement par une autre expression idiomatique, ce qu'illustre l'exemple "*poser un lapin à quelqu'un*" se

[9] Une expression est compositionelle si le sens global se déduit directement du sens des mots la composant : le sens global est transparent.

traduisant en anglais par un verbe à particule : "*to stand somebody up*". Dans une problématique d'alignement, un idiotisme devant être appréhendé comme un ensemble, il est raisonnable de penser qu'un résultat correct consiste à lier l'expression entière avec sa traduction (voir figure 1.7), un alignement plus fin ne pouvant se concevoir.

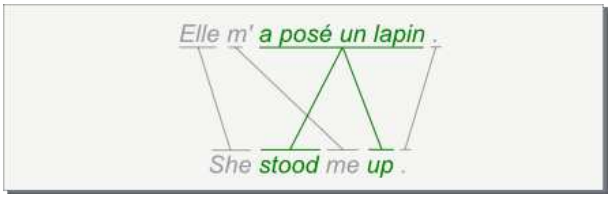


Figure 1.7 – Un idiotisme aligné révélant une discontinuité et des groupements

Parmi les difficultés que rencontre l'apprenant d'une langue, les **collocations** sont des expressions faisant parti d'un savoir collectif nécessaire à un parler courant. Sans être autant figées[10] que les idiotismes, ce sont des rapprochements de termes privilégiés qui forment un ensemble compositionnel comme *"un film grand public","une lutte acharnée"* (nominales), *"courir les filles"* ou *"manger à sa faim"* (verbales). Elles peuvent paraître d'une grande banalité car parfois difficiles à distinguer d'un simple syntagme et leur emploi est souvent inconscient. Pourtant elles n'ont rien de fortuit, ce qui nous est rappelé lorsque par exemple un anglophone apprenant le français dit souhaiter nous *secouer la main*. Les rapprochements entre termes dépendent de la langue et souvent la traduction littérale d'une collocation donne lieu à des expressions disgracieuses bien que rafraîchissantes. Elles sont une difficulté à prendre en compte car leur présence au sein de la langue n'est pas marginale : une étude de Nesi [103] tend à prouver qu'une collocation est susceptible d'apparaître à chaque phrase. Toutefois, les collocations présentent l'avantage de n'être qu'assez peu ambigües [160]. Leur impact sur la langue est abordé dans différents domaines du TAL notamment l'analyse syntaxique [128],[150], de manière moins conséquente en traduction [129]. Une collocation peut se traduire en un ensemble syntaxiquement très proche comme dans "*battre un record*" se traduisant plus naturellement en anglais par "*to break a record*" (littéralement "*casser un record*")

[10] Une expression est dite figée si elle est peu suceptible de variations syntaxiques, on pourra alors parler de locution.

ou encore "*un vin piqué*" se traduisant en espagnol plutôt par "*un vino rancio*" (littéralement "*un vin rance*"). Pour ces exemples, on peut alors parler de **divergence lexicale**, les structures restant donc intactes, ce qui se traduit en terme d'alignement par des liens simples parallèles comme observé sur le premier exemple de la figure 1.8. Des travaux se sont intéressés à la détection de paires de collocations en relation de traduction aux structures syntaxiques proches [129]. Mais une collocation en langue source ne se traduisant pas forcément par une collocation en langue cible, on peut s'attendre à un équivalant de traduction syntaxiquement différent pour des exemples comme "*tenir ses promesses*" se traduisant simplement par le verbe "*to deliver*". La divergence des structures dessine un alignement groupant (deuxième exemple de la figure 1.8).

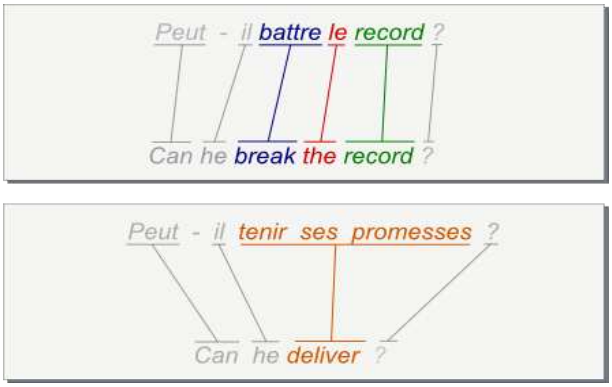


Figure 1.8 – Deux exemples de collocations alignées

1.2.2 La préservation du sens derrière la forme de surface

La traduction d'une collocation au sein d'une phrase produira une déformation locale, mais la divergence entre deux langues comprend des phénomèmes plus larges qui peuvent être inhérents notamment aux grammaires des deux langues. On peut donner, des exemples tels que l'inversion adjectif-nom pour l'anglais, le pronom personnel au datif placé après le verbe en anglais, le placement du verbe en fin de phrase pour le japonais, l'omission des pronoms sujets en espagnol (relativement à la langue française). De tels phénomènes indépendants de l'information lexicale des deux langues et quasiment promus au statut de règles, justifient des constructions de grammaires bilingues telles

que les ITG [154]. Ils sont désignés comme étant des cas de **divergence syntaxique** dans [51]. On peut considérer que de telles "divergences" participent en fait au transfert syntaxique naturel entre deux langues. Les quelques exemples figure 1.9 rendent compte de ce type de divergence dont les alignements présentent des groupement, des défauts de contiguïté ou des croisements. Des travaux tels que ceux de Yamada et Knight [159] sont parmi les premiers à tenir compte de la divergence syntaxique dans un modèle statistique en proposant un système de traduction automatique dont le résultat est produit par transformations élémentaires sur un arbre syntaxique source. L'approche transformationnelle est caractéristique de l'approche historique.

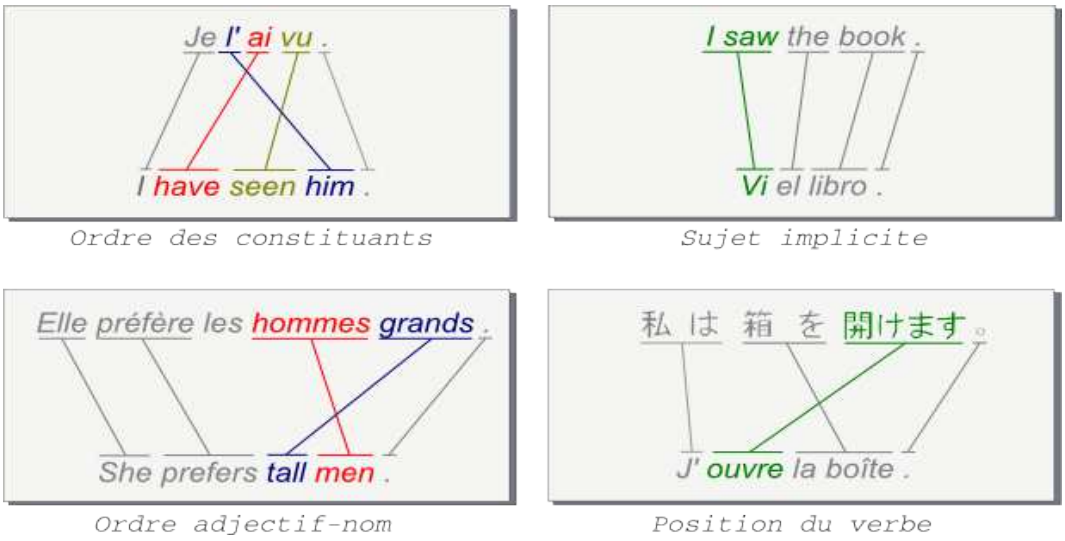


Figure 1.9 – Cas de divergences syntaxiques participant au transfert "naturel"

Pour autant, les phénomènes de divergence entre deux langues ne sauraient se résumer à un ensemble de transformations quasiment figées. Il existe des phénomènes plus diffus dont la cause est lexicale (donc très localisée comme pour les collocations) mais qui peuvent impliquer des chamboulements importants dans la structure traduite. La phrase "*John a traversé la rivière à la nage*" se traduira plus naturellement par "*John swam across the river*" et pourtant on ne peut dégager une règle générale d'ordre syntaxique qui décrirait ce chassé-croisé entre "nager" et "traverser". La transformation ici dépend essentiellement de la notion de *déplacement en traversant*, qui pourra tout au

plus dégager une règle s'appliquant à l'action de traverser par différents moyens tels que "à pied", "à vélo", etc. Dorr [52] propose un cadre formel pour étudier des types de divergences qu'elle regroupe sous le terme de divergence lexico-sémantique[11].

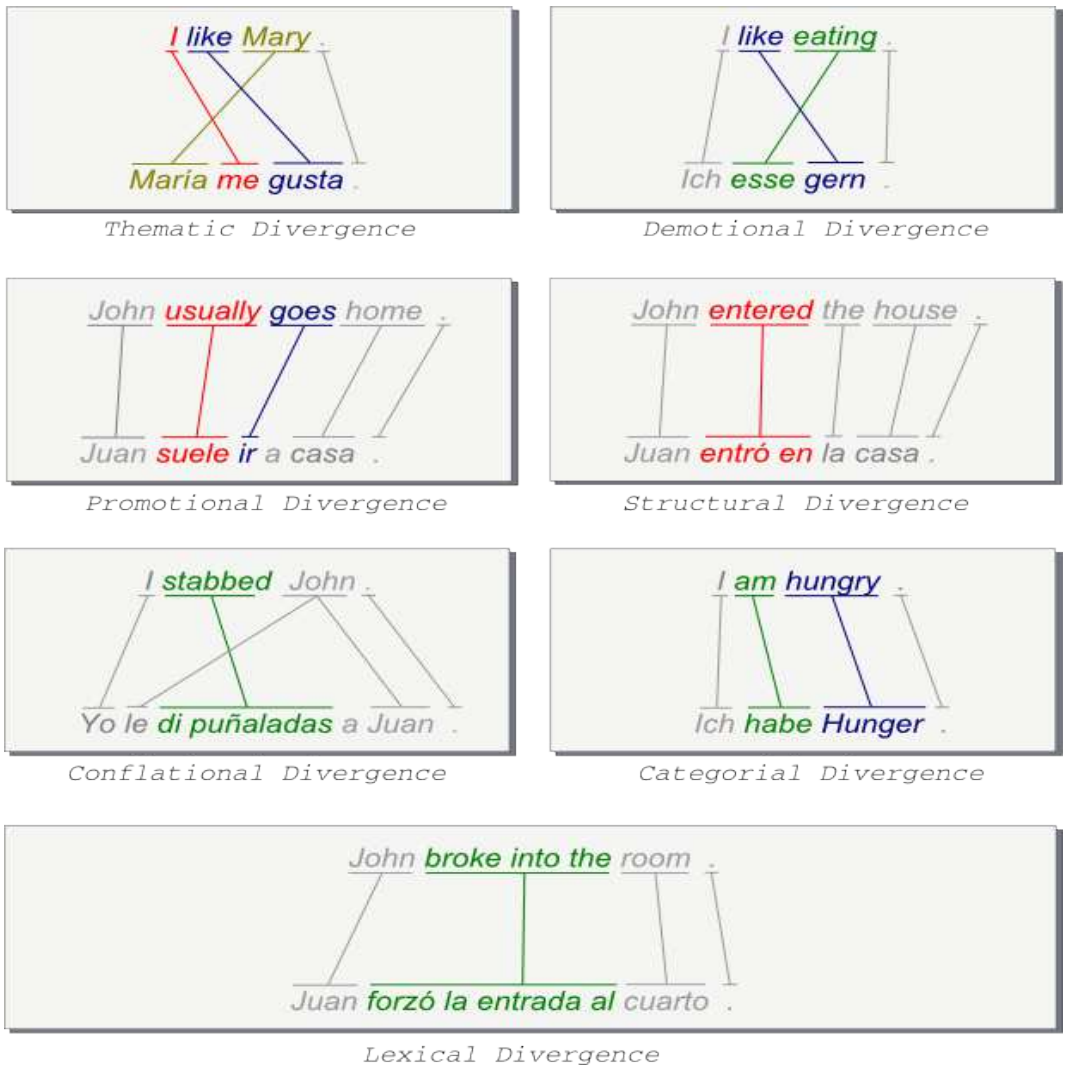


Figure 1.10 – Les sept cas de divergence lexico-sémantique de B. Dorr

L'étude arrive à dégager sept phénomènes atomiques qui peuvent par composition couvrir un ensemble plus vaste (on peut voir des exemples alignés en figure 1.10). Dans

[11] La divergence lexico-sémantique se différencie des cas de divergences dont une résolution impliquerait des connaissances syntaxiques générales, spécifiques ou pragmatiques en tenant compte du contexte.

cette étude, les divergences lexico-sémantiques sont décrites comme des exceptions au transfert naturel entre une représentation logique de la phrase (via une LCG[12]) et une structure syntaxique. L'utilisation des LCG comme représentation indépendante du langage offre un cadre théorique permettant notamment de justifier que les cas de divergences lexico-sémantiques induisant des croisements dérivent des trois cas de base que sont les divergences thématique, promotionnelle et démotionnelle. On remarque que les alignements de la figure 1.10, page 27, ne suffisent pas toujours à capturer un de ces phénomènes de divergence. C'est en fait parce que la divergence lexico-sémantique est plutôt vue comme un ensemble de modifications entre une représentation sémantique et une forme syntaxique, ce qui n'apparaît pas forcément en surface. Cette taxonomie concerne des phrases en relation de traduction qui partagent une même représentation sémantique sous-jacente. Or, la représentation sémantique se déduisant de la surface, n'est pas toujours connue et peut s'avérer incomplète. En général, l'effort de traduction ne peut donc se dédouaner complètement de quelques **décalages sémantiques** lorsque la préservation du sens d'origine n'est pas possible. Ce type de phénomène, étudié par Kameyama [74], et qu'il appelle "*traduction décalée*"[13], peut être utilisé pour défendre l'idée selon laquelle "*la traduction est souvent affaire d'approximation du sens dans la langue source*". Il démontre à plusieurs reprises l'aspect asymétrique du processus de traduction par des exemples tels que la phrase "*I like this drawing*", se traduisant en japonais par "*watashi wa kono e ga suki desu*". De l'anglais vers le japonais, la traduction est simple mais en sens inverse c'est différent car le mot "*e*" peut aussi bien signifier "*drawing*"(dessin) que "*painting*"(peinture). Hors contexte, un traducteur peut proposer un décalage en traduisant par "*I like this picture*". Mais cette astuce par affaiblissement du sens n'est pas toujours possible, et on ne peut pas toujours se passer d'une analyse du contexte. Un exemple simple est celui du vouvoiement français vu depuis l'anglais : le choix entre le "tu" et le "vous" n'est pas toujours déductible du contexte proche.
L'alignement de ces exemples de traductions décalées ne semble pas poser de problème tandis que l'effort de traduction requiert le plus haut niveau d'analyse contextuelle. En

[12]Lexical Conceptual Structure : voir [71] [72]
[13]"translation mismatch" : en anglais dans l'article original

effet, les difficultés de l'alignement sous-phrastique ne sont pas toujours les mêmes que celles de la traduction automatique (notamment lorsqu'il est question d'ambiguïté (nous le verrons en section 3.2.3.3).

1.2.3 La divergence observée face à la divergence souhaitée

Loin d'être une version simplifiée de la traduction automatique, *l'alignement sous-phrastique en situation de corpus multilingues parallèles*[14] présente ses propres défis, notamment celui de pouvoir faire face à la grande variabilité des ressources parallèles. Pour différentes raisons, la qualité des corpus traduits peut être sujette à caution et former un "*bruit*" qu'il sera en général difficile de différencier des phénomènes de divergence plus licites. En fait, nous considérerons toute irrégularité comme un phénomène de divergence et préférerons parler de *traduction approchée* plutôt que de traduction bruitée, considérant qu'une solution d'alignement est toujours possible même pour des cas de grande divergence. Dans l'exemple 1.11 dont on peut contester la qualité de traduction, il n'en demeure pas moins qu'une solution d'alignement viable est proposée pour l'intervalle "*the one who carries on the bloodline*" qui correspond effectivement à *"l'actuelle descendante"*.

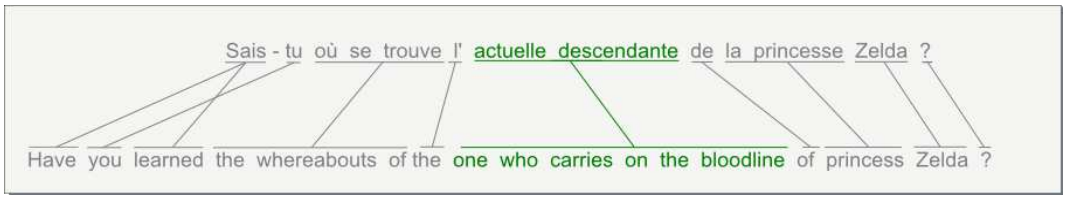


Figure 1.11 – Un cas de traduction approchée dû à une traduction via une langue pivot.

Des cas de traduction approchée comme celui-ci donnent lieu à des équivalents que l'on peut qualifier d'expressions paraphrastiques, dans le sens où le syntagme traduit effectif ressemble à une paraphrase du syntagme traduit attendu. Les corpus que nous utilisons sont de différents types (voir la partie 6.1.1 pour une présentation détaillée) et les

[14] un corpus multilingue parallèle est un ensemble de textes dans plusieurs langues dont les phrases en relation de traduction partagent un même index

cas de traduction approchée peuvent se justifier pour diverses raisons : le type du corpus influencera grandement la fidélité de la traduction et un roman sera vraisemblablement traduit avec plus de libertés qu'un texte de loi. On parlera dans ces cas de *divergence stylistique*. Il y a aussi le cas des traductions "amateurs" venant grossir la masse des traductions accessibles par Internet de qualités très variables. De plus, souvent dans le cas de corpus multilingues, la mise en correspondance de deux langues est artificielle car une ou plusieurs langues pivot peuvent séparer la langue source de la langue cible (c'est le cas pour l'exemple 1.11 page 29 puisqu'il provient d'un jeu vidéo japonais traduit dans plusieurs langues européennes). Un autre aspect est à prendre en compte dans notre approche expérimentale : les analyses syntaxiques associées à chaque phrase des corpus seront issues d'outils automatiques (voir partie 6.1.2) présentant donc d'inévitables erreurs. Une partie de notre approche traitera de la mise en correspondance des structures d'analyse de phrases en relation de traduction (un alignement arbre-à-arbre). Ainsi, la divergence observée dépendra d'irrégularités entre structures et donc des analyseurs utilisés.

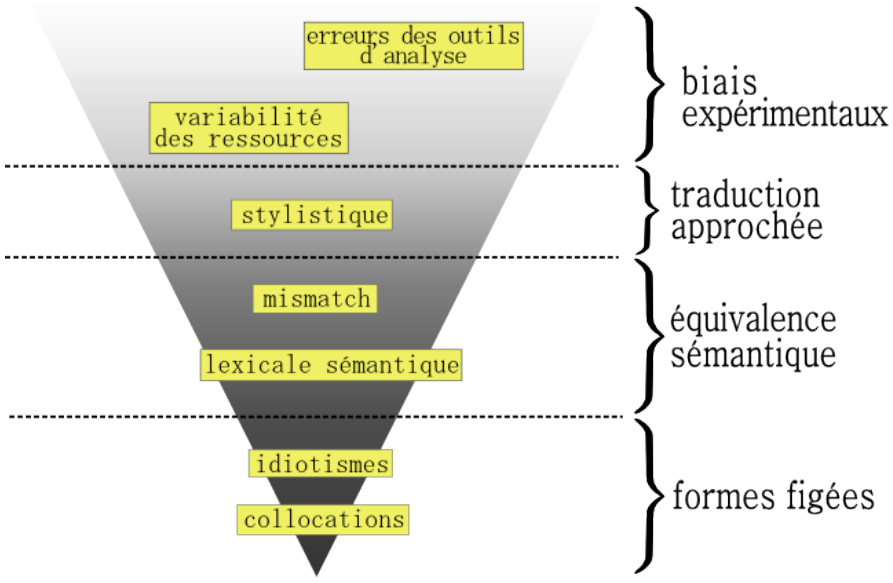


Figure 1.12 – Récapitulatif des phénomènes de divergence observés

1.3 Positionnement

Les travaux menés dans le cadre de cette thèse naissent d'un questionnement sur l'expressivité des modèles de représentation pour l'alignement sous-phrastique. L'observation conjointe d'outils existants et d'études sur la divergence au travers d'une problématique d'alignement met en évidence l'écart entre les phénomènes *exprimables* et ceux que l'on souhaiterait *exprimer*.

L'*expressivité*, telle qu'on l'entendra ici, sera liée à la notion d'*espace d'alignement*, c'est-à-dire l'ensemble des alignements qu'un outil est capable de construire à partir de n'importe quelle biphrase. Par exemple, un aligneur de mots sera dit moins expressif qu'un aligneur de syntagmes, car l'espace d'alignement du premier sera strictement inclus dans celui du deuxième.

La position classique adoptée par les modèles d'alignement est généralement de proposer un compromis entre l'*expressivité* du modèle et la *complexité* des algorithmes de résolution. Pour prendre un exemple, le coût algorithmique d'un modèle d'alignement de *n*-grammes sera d'autant plus faible que *n* est petit, mais également moins expressif. En général, il s'agit de proposer des espaces suffisament petits pour être parcourus en temps acceptable, ou ayant des propriétés de régularité telles qu'il soit possible de les parcourir de manière *"intelligente"*. Les alignements formés par un outil dépendent généralement de l'approche, des ressources linguistiques, des modèles mathématiques sous-jacents, et forment un *espace d'alignement* clairement circonscrit et motivé par la pratique.

Par ailleurs, les différentes études sur la divergence, que nous revisitons au travers d'une problématique d'alignement sous-phrastique, sont empreintes d'une réalité plus linguistique. Il en ressort essentiellement que les difficultés de mise en correspondance bilingues donnent lieu à des alignements complexes et peu réguliers. Au total, des cas de divergence d'importance croissante semblent dessiner un espace d'alignement ouvert (fig. 1.12 page 30). Pour en rendre compte, les modèles nécessitent une plus grande expressivité afin de lier des groupes longs, non contigus et intriqués. Les outils existants dans leur ensemble peuvent proposer des solutions à différents niveaux de complexité, mais ne se montrent pleinement satisfaisants pour aucun d'entre eux.

Le défi consiste donc à proposer une approche permettant la construction d'alignements en situation de corpus parallèle pouvant rendre compte de phénomènes de divergence variés. Dans le cadre du projet Blinker (pour *Bilingual Linker* [95]), Dan Melamed propose et étudie un protocole afin de créer une ressource bilingue (français/anglais) extraite de la Bible. Elle est alignée manuellement à un niveau sous-phrastique par sept experts à l'aide d'une interface adaptée (les mots et groupes de mots sont "liables" à la souris). Un guide d'alignement est produit pour orienter les experts vers ce qui est considéré comme un alignement correct (nous abordons ce probléme en section 3.1.1.1) dont on peut observer un exemple en figure 1.13 illustrant bien l'idée selon laquelle un alignement peut exprimer de manière pertinente le transfert, même dans des situations extrêmes.

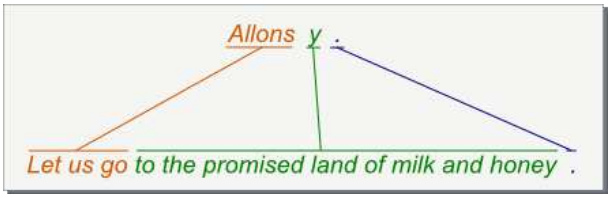


Figure 1.13 – Un exemple d'alignement manuel proposé par le guide du Blinker

Les alignements manuels peuvent être complexes et s'adaptent de manière très expressive à des phénomènes de grande divergence. Atteindre la qualité d'alignements manuels par des outils automatiques actuels semble peu probable [130]. Nous souhaitons étudier et proposer une approche différente succeptible de produire des alignements d'une structure rappelant l'expressivité observée dans le *corpus Blinker*[15]. La participation d'annotateurs sera requise et se dessine alors le problème majeur du coût d'un tel projet. Nous rappelons que le projet de Dan Melamed a nécessité en moyenne plus de 20 heures par personne pour aligner les mêmes 250 (bi-)versets (composés d'environ 7500 mots pour l'anglais et 8000 pour le français). Cette ressource de qualité est donc coûteuse en temps, mais aussi en argent puisque les experts étaient payés à l'heure.

Différents axes qui nous permettront de proposer des solutions à cet obstacle évident

[15]disponible ici `http://nlp.cs.nyu.edu/blinker/`

ont été défendus dès 1965 dans l'article de Martin Kay concernant les *rôles adéquates de l'homme et de la machine dans la traduction* [76]. L'auteur y propose une interface utilisateur pour le problème de la traduction qui *"n'existe pas et n'existera peut être jamais"*... pourtant bien des points sembleront familiers à un utilisateur d'outils de traduction en ligne d'aujourd'hui. L'article insiste clairement sur la nécessité d'une approche partagée et non marchande. L'effort de traduction est envisagé via une possiblement longue collaboration homme-machine par cette probable interface appelée la "*Translator's Amanuensis*"[16] aux fonctionnalités extensibles. Il souhaite défendre une approche où la machine prendrait graduellement et imperceptiblement les rênes dans le processus de traduction. L'issue de cette approche incrémentale serait l'émancipation des robots, aboutissant à un outil de traduction automatique autonome ayant atteint sa maturité, vision chère à l'imaginaire asimovien. Il insiste sur la patience et la modestie que demande cette voie en prononçant le fameux *"Little steps for little feet !"*[17].

Nous souhaitons adapter les préceptes des *petits pieds* à l'alignement sous-phastique dans un outil d'annotation du type Blinker. Une partie du travail décrit en 3.2 consiste en la mise en place d'une architecture pour des outils évolutifs accompagnant l'annoteur, dans un partenariat bénéfique à la fois pour les deux parties : l'humain et la machine. L'interface que nous proposons 3.1.2 permet une participation collaborative en ligne transparente via une participation non-experte. En s'améliorant, la partie automatique de l'outil allègerait la tâche de ses utilisateurs qui en retour seraient capables de lui fournir d'autant plus de données pour progresser. L'objectif visé est alors double : construire une ressource de qualité et entraîner un outil d'alignement automatique reposant sur une mémoire d'alignement. Si le robot doit ultimement être autonome, il est important qu'en attendant, à l'instar du cône de divergence décrit avant (voir figure page 1.12 30), il demeure ouvert.

Si le vecteur directeur est une expressivité non compromise, il est nécessaire en tout premier lieu de circonscrire une notion d'alignement adéquate. Il sera donc primordial de discuter les prérequis pour un modèle général (section 2.1), recouvrant les repré-

[16] *Le copiste du traducteur.*
[17] *De petits pas pour de petits pieds.*

sentations existantes et tenant compte des idiosyncrasies liées à une approches collaborative. Pour cela, une partie importante de notre travail consiste à formaliser un **espace des alignements** général (section 2.2.2) ainsi que des notions relatives permettant de le munir d'opérations canoniques ainsi que de métriques adaptées à l'observation des données. Nous nous sommes attelé durant cette thèse à étudier les difficultés inhérentes à cette expressivité, les circonscrire, et à proposer des solutions pratiques ou formelles. Le cadre théorique omniprésent accompagnera l'ensemble du propos, étayant notre argumentaire et sous-tendant l'architecture d'une approche collaborative basée sur une mémoire d'alignements(section 3.2). Les éléments du modèle sauront tirer avantage d'informations provenant d'analyseurs syntaxiques automatiques, en plaçant toutefois les prérequis technologiques à un niveau raisonnablement peu élevé. Les divergences rencontrées dépendant de la paire de langue concernée, de la qualité des corpus mais aussi des erreurs d'analyse, les solutions envisagées devront être adaptables à différentes langues et se montrer robustes.

CHAPITRE 2

L'ENSEMBLE DES ALIGNEMENTS

Ce chapitre est consacré à la discussion (section 2.1) et à la mise en place d'un environnement formel (section 2.2.2) pour la notion d'alignement afin d'accompagner l'ensemble de notre propos. Il offrira une visibilité précise des espaces d'alignements (section 2.2.3) afin d'insister sur l'importance de l'expressivité des modèles. Cela nous permettra notamment de relativiser la notion de qualité, en proposant une vision plus graduée des désaccords entre annotateurs (section 3.1.1.1). Enfin, nous proposerons une alternative aux mesures d'alignements de type erreur/accord en définissant trois distances transformationnelles sur l'espace des alignements (section 2.3), reflétant intuitivement l'effort de post-annotation par le biais de différentes interfaces graphiques d'alignement manuel.

2.1 Discussion en faveur d'un modèle d'alignement adapté aux divergences

Autour d'une discussion guidée par des exemples, nous tentons de circonscrire et de préciser le type d'alignements dont il est question ici. En d'autres termes, quelle limite peut-on légitimement s'autoriser ? En évoquant les différents phénomènes de divergence au chapitre précédent, nous avons mentionné la volonté de voir, au sein d'une biphrase alignée, la possibilité de lier des groupes de mots. La première contrainte que l'on souhaite imposer est celle-ci : *un mot lié n'appartiendra qu'à un seul groupe de mots tous liés entre eux*[1]. Il s'agit d'une hypothèse de "*bonne formation*" classique dans des approches de traduction formalisant des représentations alignées. On peut citer par exemple le "*Alignment Well-Formedness Criteria*" dans l'approche d'alignement de sous-arbres de [140] ou encore la contrainte d'unicité dans la définition des S-SSTC de Tang [5] qui sont des alignements sur des structures bilingues complexes. Un "alignement" ne satisfaisant pas cette hypothèse serait selon nous une forme dégénérée sans grande valeur

[1] l'interaction avec l'outil Align$^{\text{It}}$ respectera cette contrainte

linguistique.

On observe en figure 2.1 un alignement dégénéré et quelques alternatives bien formées : dans l'alignement dégénéré, les mots cible s_1 et s_3 sont liés à c_1, mais ne partagent pas le même ensemble de voisins (c_3 est voisin de s_3 mais pas de s_1). Le groupe de liens rouges ne forme donc pas une unité de sens. Remarquons qu'aucun des modèles décrits dans l'état de l'art en section 1.1 (mots, ITG, asymétriques, intervalles de mots, ...) ne produit de forme dégénérée. La contrainte, bien qu'implicite, est assez universelle.

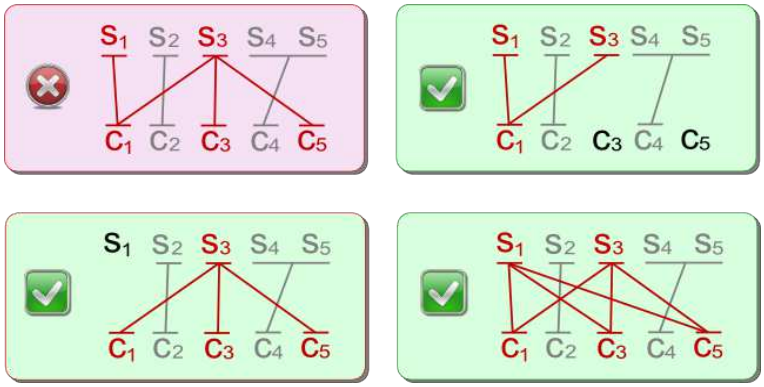


Figure 2.1 – Un alignement dégénéré et des alternatives "bien formées"

Dans un soucis de généricité, nous souhaitons proposer un modèle d'alignement pouvant tenir compte d'une **segmentation arbitraire en unités minimales** que nous appelons abusivement des "mots". Cela nous permettra d'intégrer une segmentation imposée par des analyseurs automatiques et ainsi de pouvoir décomposer des phrases écrites dans des langues à morphologie compositionnellement riches telles que l'allemand, des langues agglutinantes telles que le tchèque ou le basque, où dans des systèmes d'écriture sans blanc typographique tels que le chinois, le japonais ou le thaï.

On observe en effet sur l'exemple 2.2 (emprunté à Y. Lepage dans [88]) que "*Atravesó*" (correspondant au verbe espagnol "*atravesar*", signifiant "*traverser*", au passé parfait, 3$^{\text{ème}}$ personne du singulier, correspond légitimement au groupe non contigu formé de "*It*" (pour la 3$^{\text{ème}}$ personne du singulier), le suffixe "*-ed*" (comme marqueur de temps) et à la particule "*across*" (porteuse du sens). Pourtant, si la segmentation des unités lexicales est fixée par le blanc typographique, le marqueur du temps "*-ed*" fait corps avec

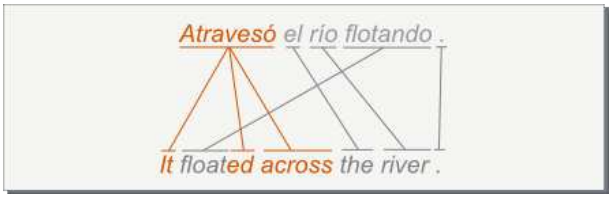


Figure 2.2 – On peut vouloir aligner des unités plus fines que le mot

"*floated*", ce qui rend cet alignement impossible. Deux solutions sont alors envisageable : la première considère que l'unité minimale est le caractère, ce que nous exclurons car nous souhaitons bénéficier d'informations morphosyntaxiques d'analyses automatiques, mais aussi et surtout car, d'un point de vue pratique, il sera plus ergonomique pour un annotateur de cliquer sur des mots que sur des caractères. Nous accepterons donc de nous plier à une segmentation imposée par les outils, et proposons d'employer la deuxième solution qui consiste à effectuer un alignement plus grossier, comme observé en figure 2.3.

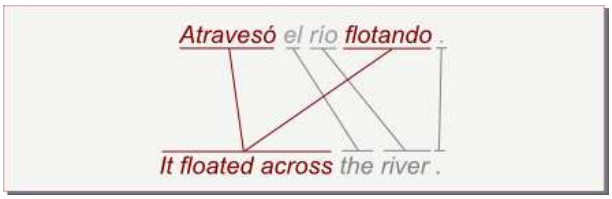


Figure 2.3 – Un alignement plus grossier résoud le problème.

Nous avons observé les résultats d'alignements construits selon ces deux contraintes par des utilisateurs de notre interface d'alignement manuel, présentée plus loin à la partie 3.1.2, afin de répondre à différentes questions concernant l'expressivité. Peut-on fixer une limite arbitraire à la taille des groupes liés ? Peut-on éviter les phénomènes non contigus, ou les limiter par un paramètre ? Il est clair qu'en regardant un ensemble fini de biphrases alignées, on est capable de majorer les nombres de mots liés et de défauts de contiguïté. Il y a pourtant fort à parier que ces valeurs deviennent arbitrairement grandes en augmentant notre corpus de biphrases alignées avec des biphrases de plus en

plus longues. Nous le constatons aisément sur l'exemple de la figure 2.4. La préposition "*d'*" qui doit se répéter à chaque étape de l'énumération en français, n'apparaît qu'une fois en anglais puis est implicite. Cet exemple est intéressant car il rend compte d'une transformation classique entre l'anglais et le français et peut être généralisé en énumérant plus longuement : "*un climat, d'instabilité, d'incertitude, de misère, de dépression, de dégoulinante noirceur, etc.*". Ainsi, le nombre de défauts de contiguïté et la longueur de ce groupement de mots liés n'ont de limite que l'imagination.

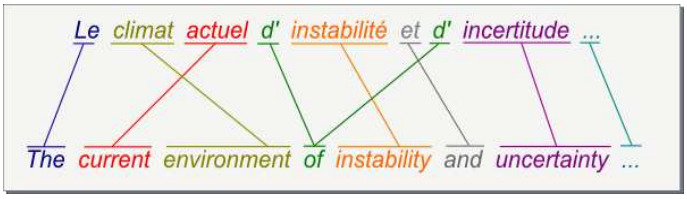


Figure 2.4 – Exemple de groupement non contigu de grande longueur

Une autre piste afin de limiter l'espace d'alignement pourrait être de souhaiter une structuration des liens sous-tendue par une grammaire générative, comme pour le modèle ITG. Nous avons vu dans l'état de l'art (section 1.1.3.7, page 22) que l'espace d'alignement des ITG était l'ensemble des permutations privées de $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}$ et $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$ (ces configurations sont aussi appelées *"inside-out"*). Nous n'entendons pas mener ici une investigation sur la présence de ces configurations entre paires de langues, mais pressentons qu'il y a peu de chances qu'elles soient entièrement absentes d'un couple donné, à plus forte raison si une des deux langues est à ordre totalement libre comme le latin. Nous avons parcouru notre base d'alignements au début de sa constitution pour détecter des configurations *"inside-out"* sur des mots et/ou des groupes de mots et en avons effectivement rencontré à plusieurs reprises, même si leur fréquence est plutôt faible (voir figure 2.5). Priver l'espace total d'un ou plusieurs motifs semble donc une idée assez peu naturelle dans une approche à base d'exemples qui prétend représenter des phénomènes de divergence complexes. L'interaction avec l'outil devra donc permettre de générer n'importe quel alignement non dégénéré. L'ensemble des alignements potentiellement atteints sera désigné comme étant l'**espace total des alignements** (défini

partie 2.2.2). Nous avons vu que les différentes approches existantes produisaient des alignements respectant certaines contraintes strictes, limitant souvent leur expressivité pour des raisons de complexité algorithmique (nous comparons ces différents *espaces d'alignements* dans la partie 2.2.3).

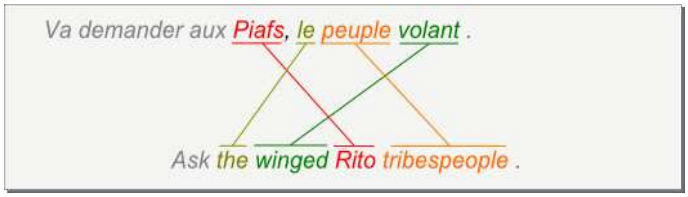


Figure 2.5 – Exemple de configuration *"inside-out"* rencontrée

En s'attaquant à un espace d'alignements libéré de toute contrainte, on s'attend donc à une difficulté élevée. Cependant, l'approche à base d'exemples qui est défendue ici tire parti d'une simplification apportée par l'étape de *fragmentation* (détaillée en partie 5.1.2) grâce à laquelle il est possible de réduire en apparence la divergence d'un alignement. En effet, comme illustré à la figure 2.6, si on l'oberve via différentes segmentations, un alignement devient plus simple à mesure que la taille des fragments augmente (ultimement, le fragment total n'est pas du tout générique pour une longue biphrase).

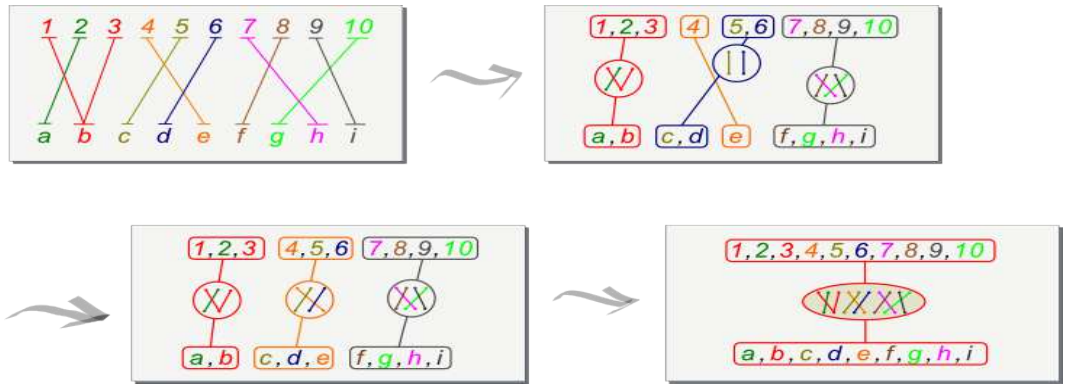


Figure 2.6 – Élargir les éléments de base réduit la divergence apparente de l'alignement.

L'idée est donc de retenir des groupes d'une taille intermédiaire dans une base d'exemples afin de réduire la difficulté des problèmes d'alignement. Ainsi, alors que les approches

classiques proposent systématiquement un compromis entre *expressivité* du modèle et *complexité algorithmique*, une approche à base d'exemples aurait l'avantage de proposer une expressivité potentiellement maximale en plaçant un compromis nouveau entre **généricité** des groupes de mots et *complexité algorithmique*. Une telle approche, en faisant intervenir de manière naturelle l'humain dans la constitution d'une mémoire, serait donc capable non seulement d'exprimer des phénomènes structurels parfois très divergents et hors du champ d'expressivité des modèles classiques, mais également d'en tirer parti.

2.2 Un modèle adapté

2.2.1 Généralités

Nous avons vu que l'alignement sous-phrastique avait tendance à s'orienter vers l'alignement d'unités plus longues que le mot. L'unité visée est difficile à définir et les différents cas de divergence nous encouragent à viser des groupes de taille variable non nécessairement composés de mots contigus. Les *partitions d'un ensemble* nous permettront de définir une fragmentation monolingue très générale. Nous donnons dans cette partie quelques notions générales sur les *ensembles de partitions* qui nous seront utiles dans le reste du chapitre, et dans ce mémoire en général.

Rappelons la définition d'une **partition d'un ensemble** :

Définition 1. *Soit X un ensemble quelconque. Une **partition** $\mathcal{P} = \{P_i\}_{i \in I}$ de X est un ensemble de sous-ensembles de X, i.e. $\mathcal{P} \subset \mathscr{P}(X)$, telle qu'ils sont non vides, recouvrent X et sont deux à deux disjoints :*

1. $\forall i \in I, P_i \neq \emptyset$
2. $X = \bigcup_{i \in I} P_i$
3. $\forall i, j \in I, i \neq j \Rightarrow P_i \cap P_j = \emptyset$

On notera $\Pi(X)$ l'ensemble des partitions de X.

Exemple. *On peut donner deux exemples triviaux de partitions sur X :*

- *la* ***partition grossière*** $\{X\}$ *ne contenant qu'une seule classe.*
- *la* ***partition en singletons*** $\{\{x\}, x \in X\}$ *comportant autant de classes que X contient d'éléments.*

Nous proposons de relâcher la contrainte de recouvrement pour définir la notion de *partition partielle*. Celle-ci sera utilisée en partie 2.2.2 pour établir une définition formelle de l'alignement :

Définition 2. *Soit X un ensemble quelconque. Pour tout* $X' \subset X$ *tel que* $\mathcal{P}$ *est une partition de X,* $\mathcal{P}$ *est aussi appelé une* ***partition partielle*** *de X. On dira que* X' *est le* ***support de*** $\mathcal{P}$*, noté* $supp(\mathcal{P})$.

Les éléments P_i d'une *partition* $\mathcal{P}$ de X sont appelés ses classes, sont disjointes et recouvrent X par union. Une partition $\mathcal{P}$ définit naturellement une relation d'équivalence dont les classes ne sont autres que les classes de $\mathcal{P}$. En divisant les classes d'une partition, on obtient une autre partition "plus fine". On définit la relation de finesse sur les partitions de X :

Définition 3. *Soient* $\mathcal{P}, \mathcal{Q}$ *deux partitions de X. On dira que* $\mathcal{P}$ *est* ***plus fine*** *que* $\mathcal{Q}$ *(noté* $\mathcal{P} \prec \mathcal{Q}$*) si toute classe de* $\mathcal{P}$ *est incluse dans une classe de* $\mathcal{Q}$.

Exemple. $X = \{1,2,3,4,5\}$*, on se donne trois partitions :*

$$\begin{cases} \mathcal{P}_1 &= \{\{1,2\},\{3,5\},\{4\}\} \\ \mathcal{P}_2 &= \{\{1,2\},\{3\},\{5\},\{4\}\} \\ \mathcal{P}_3 &= \{\{1,2\},\{3\},\{4,5\}\} \end{cases}$$

Les relations de finesse sont : $\mathcal{P}_2 \prec \mathcal{P}_1$ *et* $\mathcal{P}_2 \prec \mathcal{P}_3$*, mais il n'y a pas forcément compatibilité car* $\mathcal{P}_1 \not\prec \mathcal{P}_3$ *et* $\mathcal{P}_3 \not\prec \mathcal{P}_1$.

La *relation de finesse* '$\prec$', définie relativement à l'ensemble X, est une *relation d'ordre partiel* sur l'ensemble des partitions de X. Elle admet la *partition grossière* comme élément maximum et la *partition en singletons* comme élément minimum.

Nous définissons deux lois de composition entre partitions sur un même ensemble :

Définition-Propriété 1. *Soient $\mathcal{P}, \mathcal{Q}$ deux partitions de X :*

- *Il existe une partition plus fine que $\mathcal{P}$ et $\mathcal{Q}$. On note $\mathcal{P} \vee \mathcal{Q}$ la partition de X la moins fine de toutes les partitions plus fines que $\mathcal{P}$ et $\mathcal{Q}$. C'est l'infimum de $\mathcal{P}$ et $\mathcal{Q}$. On parle parfois aussi de croisement.*

- *Il existe une partition moins fine que $\mathcal{P}$ et $\mathcal{Q}$. On note $\mathcal{P} \wedge \mathcal{Q}$ la partition de X la plus fine de toutes les partitions moins fines que $\mathcal{P}$ et $\mathcal{Q}$. C'est le supremum de $\mathcal{P}$ et $\mathcal{Q}$.*

Exemple. *En reprenant les partitions $\mathcal{P}_1$, $\mathcal{P}_2$ et $\mathcal{P}_3$ de l'exemple précédent :*

$$\begin{cases} \mathcal{P}_1 \wedge \mathcal{P}_3 &= \mathcal{P}_2 \\ \mathcal{P}_1 \vee \mathcal{P}_3 &= \{\{1,2\},\{3,4,5\}\} \end{cases}$$

L'infimum $\mathcal{P} \wedge \mathcal{Q}$ s'obtient en retenant toutes les intersections non vides entre les classes de $\mathcal{P}$ et $\mathcal{Q}$. Le supremum $\mathcal{P} \vee \mathcal{Q}$ s'obtient par fermeture transitive (itération répétée) de l'opération qui consiste à faire correspondre à chaque élément de X tous ceux qui sont dans la même classe que lui soit dans $\mathcal{P}$, soit dans $\mathcal{Q}$.

L'ensemble des partitions sur un même ensemble X, muni de la relation d'ordre partiel '$\prec$' et des opérations $\wedge$ et $\vee$ est un *treillis* ou *espace réticulé*.

Définition 4. *Un treillis est un ensemble T muni d'une relation d'ordre partiel $\leq$ telle que pour tout élément de $x, y \in T$, l'ensemble $\{x,y\}$ possède une borne supérieure et une borne inférieure. On peut alors munir T de deux lois internes :*

$$\begin{cases} x \cup y &= \sup(\{x,y\}) \\ x \cap y &= \inf(\{x,y\}) \end{cases}$$

On appelle treillis la donnée des quatre éléments $(T, \leq, \cup, \cap)$ vérifiant les propriétés énoncées.

L'ensemble des partitions est un exemple de treillis. Les ensembles totalement ordonnés comme $(\mathbb{R}, \leq, \sup, \inf)$ en sont des exemples plus triviaux. On représente souvent les treillis finis par leur *diagramme de Hasse*, c'est-à-dire en représentant seulement les *liens de finesse directs*. On peut voir en figure 2.7 le diagramme de Hasse de $\Pi(\{1,2,3,4\})$.

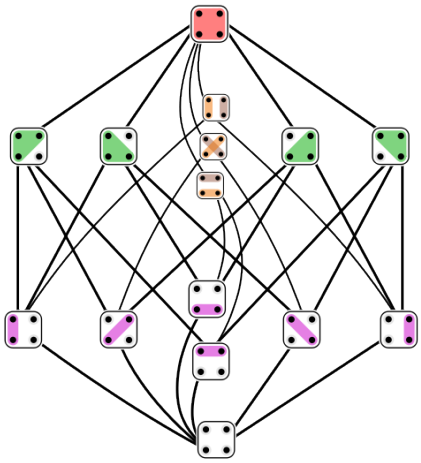


Figure 2.7 – Diagramme de Hasse des partitions sur un ensemble à quatre éléments

2.2.2 L'espace formel des alignements sous-phrastiques

2.2.2.1 Définitions et notations

On définit ici un alignement entre deux listes d'éléments[2] $S = (s_1, \ldots, s_n)$ et $C = (c_1, \ldots, c_m)$. En pratique, ces éléments peuvent être des textes, des paragraphes, des phrases, des mots ou des caractères. Dans notre cas, les éléments s_i et c_j (pour $1 \leq i \leq n$ et $1 \leq j \leq m$) seront des "mots" issus d'une segmentation automatique (les analyseurs utilisés sont présentés partie 6.1.2). On notera $B = (S, C)$ la biphrase formée par S et C. On dira que B est une biphrase de **longueur** (n, m).

Nous avons déjà écarté en partie 2.1 la possibilité de définir simplement un alignement comme un ensemble de liens entre S et C en interdisant les formes *dégénérées*, peu pertinentes linguistiquement (revoir figure 2.1 page 36). La définition souhaitée doit

[2]Le S et le C sont des notations traditionnelles en traduction automatique pour désigner les phrases **s**ource et **c**ible. On emploie également cette notation en alignement malgré l'apparente symétrie.

donc se montrer plus restrictive que de considérer l'ensemble des graphes bipartis entre S et C (noté $\mathscr{B}(S,C)$) qui inclut toutes les formes *dégénérées*. Naturellement, l'ensemble des alignements entre S et C, que nous noterons $\mathscr{A}(S,C)$, doit pouvoir se plonger injectivement dans $\mathscr{B}(S,C)$. La condition de *bonne formation* impose de voir un alignement comme un ensemble de liens simples entre des sous-parties de S et C ou plus formellement une bijection entre une *partition partielle* de $[\![1,n]\!]$ et une *partition partielle* de $[\![1,m]\!]$. L'utilisation des *partitions partielles* permet de modéliser des alignements également partiels.

On peut alors définir un alignement entre S (de longueur n) et C (de longueur m), en considérant deux *partitions partielles* de $[\![1,n]\!]$ et $[\![1,m]\!]$ de même cardinalité k. Elles définissent intuitivement deux découpages en "syntagmes" des phrases S et C que l'on peut mettre en bijection (voir la représenation à la figure 2.8).

Définition 5. *Un* ***alignement*** ℓ *sur une biphrase* $B = (S,C)$ *de longueur* (n,m) *sera la donnée de* 3 *éléments :*

1. *Une partition partielle à k classes de* $[\![1,n]\!]$*, notée* $\mathscr{V} = \{V^1,\ldots,V^k\}$
2. *Une partition partielle à k classes de* $[\![1,m]\!]$ *notée* $\mathscr{W} = \{W^1,\ldots,W^k\}$
3. *une permutation des classes* $\sigma : \mathscr{V} \mapsto \mathscr{W}$ *qui définit le transfert.*

où $0 \leq k \leq min(n,m)$*. On notera* $\ell = (\mathscr{V},\mathscr{W},\sigma)$*.*
On notera $\mathscr{A}(S,C)$ *l'ensemble des alignements entre S et C (On pourra aussi le noter* $\mathscr{A}(n,m)$ *lorsque le contexte ne fait pas explicitement références à S et C).*

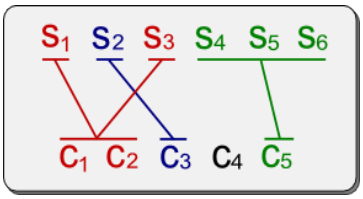


Figure 2.8 – Un exemple d'alignement : chaque couleur indique une partition partielle

On peut donner comme exemples d'alignements triviaux l'*alignement vide* $\ell_{\mathbb{O}} = (\emptyset, \emptyset, \mathbb{O})$ (où $\mathbb{O}$ désigne la fonction vide de $\emptyset$ dans lui même) et l'*alignement grossier* $\ell_{\mathbb{1}} = (S, C, \mathbb{1})$ (où $\mathbb{1}$ désigne la fonction triviale qui à S associe C) qui lie tous les mots ensemble (voir figure 2.9).

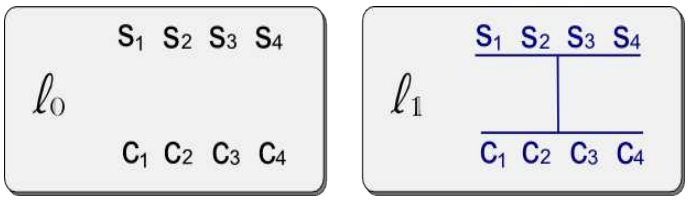


Figure 2.9 – Les alignements triviaux $\ell_{\mathbb{O}}$ et $\ell_{\mathbb{1}}$

Définition 6. *On dira que deux mots $s_i \in S$ **et** $c_j \in C$ **sont liés** selon ℓ si il existe $V \in \mathscr{V}$ et $W \in \mathscr{W}$ vérifiant :*

$$i \in V,\ j \in W \text{ et } \sigma(V) = W$$

Autrement dit, les classes de mots contenant s_i et c_j sont liées par σ.

Remarque 1. *Comme annoncé, l'ensemble des alignements entre S et C, $\mathscr{A}(S,C)$ s'injecte naturellement dans l'ensemble des graphes bipartis de parties S et C. Pour s'en appercevoir, il suffit à partir d'un alignement ℓ donné de construire le graphe biparti associé en plaçant un arc entre tous les mots source et cible liés. Il en résulte un graphe biparti qui est l'union de ses **bicliques** maximales (voir la figure 2.10).*

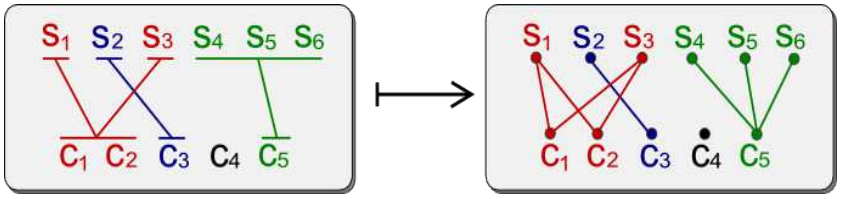


Figure 2.10 – L'espace des alignements s'injecte dans celui des graphe bipartis

Cette remarque nous permettra de donner un majorant, même grossier, de la taille de l'espace que nous considérons.

Définissons les notions intuitives de *couverture* et *d'alignement couvrant* :

Définition 7. *Soit $\ell = (\mathscr{V}, \mathscr{W}, \sigma)$ un alignement sur (S,C).*

- *On dira que le i-ème mot source $s_i \in S$ (resp. le j-ème mot cible $c_j \in C$) est **couvert** par ℓ si il existe $V \in \mathscr{V}$ (resp. $W \in \mathscr{W}$) contenant i (resp. j).*
- *Le **support source** correspond aux positions des mots source couverts par ℓ (noté $supp_S(\ell)$), c'est-à-dire $supp_S(\ell) = supp(\mathscr{V}) = \bigcup_{V \in \mathscr{V}} V$*
- *Le **support cible** correspond aux positions des mots cibls couverts par ℓ (noté $supp_S(\ell)$), c'est-à-dire $supp_C(\ell) = supp(\mathscr{W}) = \bigcup_{W \in \mathscr{W}} W$*
- *Un alignement sera dit **couvrant** si tous les mots source et cible de la biphrase sont couverts par ℓ.*
- *On appelle **couverture** de ℓ le nombre de mots source et cible couverts, c'est-à-dire $couv(\ell) = |supp_S(\ell)| + |supp_C(\ell)|$*

2.2.2.2 Notion de finesse et structure de treillis

Nous avons vu en état de l'art (partie 1.1.3.3) qu'il existe des méthodes heuristiques combinant des alignements asymétriques pour former une alternative symétrique plus fine ou plus grossière [107]. Les deux voies possibles qui dépendent de l'utilisation souhaitée. La première permet de renforcer les informations communes, quitte à réduire les données pour limiter les erreurs. Elle est souvent privilégiée dans des approches d'alignement nécessitant des points d'ancrage sûrs [107]. Elle est analogue à une opération d'intersection. La deuxième possibilité consiste à conserver les informations des deux alignements, quitte à "*lisser*" les parties en désaccord. Moins discriminante, elle retiendra une information floue produite de la superposition de plusieurs informations proches plutôt que de l'ignorer [82]). Elle est analogue à une opération d'union.

Ces deux opérations appliquées aux alignements asymétriques des modèles IBM pour les "symétriser", permettent de se déplacer, soit dans un sous-espace (celui des alignements de mots pour l'intersection) soit dans un niveau supérieur (celui des alignements de groupes de mots pour l'union étendue).

Nous définirons ici deux opérations similaires, l'*affinement* et l'*élargissement*. L'espace $\mathscr{A}(n,m)$ est suffisemment général pour que les deux opérations y soient internes. Elles nous permettront de munir l'ensemble des alignements d'une structure de treillis propre à cette approche à base d'exemples. Cette représentation de l'espace des alignements comme un ensemble structuré et cohérent avec ses opérations naturelles est originale, à notre connaissance. Nous prendrons donc la peine de définir précisément chacune des notions de manière formelle.

Afin d'alléger les notations dans cette partie, nous donnons maintenant quelques notations utilisées pour les différents éléments formels qui ne seront pas systématiquement rappelées dans cette section.

On se donne $B = (S,C)$ une biphrase de longueur (n,m) et considérerons deux alignements ℓ_1 et $\ell_2 \in \mathscr{A}(S,C)$. On notera $\ell_1 = (\mathscr{V}_1, \mathscr{W}_1, \sigma_1)$ et $\ell_2 = (\mathscr{V}_2, \mathscr{W}_2, \sigma_2)$ deux alignements où $\mathscr{V}_1$ et $\mathscr{V}_2$ sont des partitions partielles de $[\![1,n]\!]$. De même, $\mathscr{W}_1$ et $\mathscr{W}_2$ sont des partitions partielles de $[\![1,m]\!]$. Les fonctions σ_1 et σ_2, liant les classes source et cible, seront des permutations respectivement de $\mathscr{V}_1$ vers $\mathscr{W}_1$ et de $\mathscr{V}_2$ vers $\mathscr{W}_2$. Afin d'alléger les notations, nous utiliserons hors contexte les notations V_1 pour désigner un élément de $\mathscr{V}_1$, ainsi que V_2 pour $\mathscr{V}_2$, W_1 pour $\mathscr{W}_1$ et W_2 pour $\mathscr{W}_2$.

On peut définir formellement l'opération d'*affinement* sur $\mathscr{A}(S,C)$. Cette opération permettra d'exprimer par un alignement tiers l'information commune à ℓ_1 et ℓ_2 :

Définition 8. *On définit la loi de composition interne de $\mathscr{A}(S,C)$, appelée* ***opération d'affinement*** *entre ℓ_1 et ℓ_2, de la manière suivante :*

$$\ell_1 \wedge \ell_2 = (\hat{\mathscr{V}}, \hat{\mathscr{W}}, \hat{\sigma}).$$

On définit ses différents éléments comme suit :

$$\begin{cases} \hat{\mathscr{V}} = & \{ \quad V_1 \cap V_2 \quad & \text{tel que} \quad V_1 \cap V_2 \neq \emptyset \quad \text{et} \quad \sigma_1(V_1) \cap \sigma_2(V_2) \neq \emptyset \} \\ \hat{\mathscr{W}} = & \{ \quad \sigma_1(V_1) \cap \sigma_2(V_2) \quad & \text{tel que} \quad V_1 \cap V_2 \neq \emptyset \quad \text{et} \quad \sigma_1(V_1) \cap \sigma_2(V_2) \neq \emptyset \} \\ \hat{\sigma}(V_1 \cap V_2) = \sigma_1(V_1) \cap \sigma_2(V_2), & \text{où} \quad V_1 \cap V_2 \neq \emptyset \quad \text{et} \quad \sigma_1(V_1) \cap \sigma_2(V_2) \neq \emptyset \} \end{cases}$$

$\ell_1 \wedge \ell_2$ définit un alignement de $\mathscr{A}(S,C)$. L'opération d'affinement admet l'alignement grossier $\ell_{\mathbb{1}}$ comme élément neutre.

On remarque que dans l'exemple de la figure 2.11, les éléments s_2 et c_3 ont été "déliés" durant l'opération d'affinement.

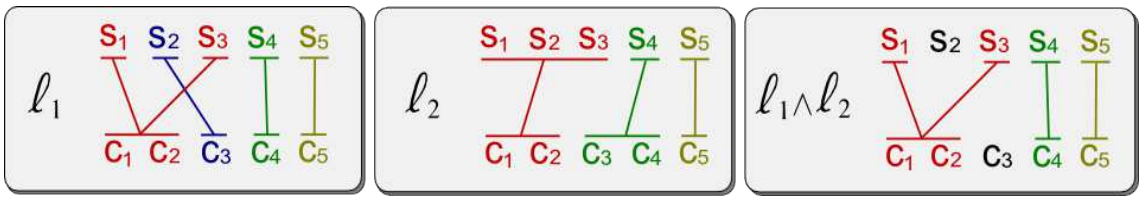


Figure 2.11 – L'opérations d'affinement sur un exemple

Nous définissons l'opération duale qui "englobera" les différences pour obtenir un alignement moins fin que les deux alignements d'origine : l'*élargissement*. Pour cela nous devons utiliser la notion de *chaîne de parties* qui correspond à une suite de parties qui se chevauchent de proche en proche :

Définition 9. *Soient E un ensemble fini et $\mathscr{E} \in \mathcal{P}(E)$ un ensemble de parties de E. On dit que $\mathscr{E}$ est une chaîne sur E si il est possible de numéroter les éléments de $\mathscr{E}$ de sorte que $\mathscr{E} = \{E_1, ..., E_k\}$ et que pour tout $i \in \{1, ..., k-1\}$ on ait :*

$$E_i \cap E_{i+1} \neq \emptyset$$

Définition 10. *On définit la loi de composition interne de $\mathscr{A}(S,C)$ appelée* ***opération d'élargissement*** *entre ℓ_1 et ℓ_2 de la manière suivante :*

$$\ell_1 \vee \ell_2 = (\check{\mathscr{V}}, \check{\mathscr{W}}, \check{\sigma}).$$

On définit ses différents éléments comme il suit :

$$\begin{cases} \check{\mathscr{V}} = \{ \bigcup_{\check{V} \in \check{\mathscr{V}}_1 \cup \check{\mathscr{V}}_2} \check{V} \quad \text{tel que } \check{\mathscr{V}}_1 \cup \check{\mathscr{V}}_2 \text{ ou } \sigma_1(\check{\mathscr{V}}_1) \cup \sigma(\check{\mathscr{V}}_2) \text{ est une chaîne maximale} \} \\ \check{\mathscr{W}} = \{ \bigcup_{\check{W} \in \sigma_1(\check{\mathscr{V}}_1) \cup \sigma_2(\check{\mathscr{V}}_2)} \check{W} \quad \text{tel que } \check{\mathscr{V}}_1 \cup \check{\mathscr{V}}_2 \text{ ou } \sigma_1(\check{\mathscr{V}}_1) \cup \sigma(\check{\mathscr{V}}_2) \text{ est une chaîne maximale} \} \\ \check{\sigma}(\bigcup_{\check{V} \in \check{\mathscr{V}}_1 \cup \check{\mathscr{V}}_2} \check{V}) = \bigcup_{\check{W} \in \sigma_1(\check{\mathscr{V}}_1) \cup \sigma_2(\check{\mathscr{V}}_2)} \check{W} \text{ où } \check{\mathscr{V}}_1 \cup \check{\mathscr{V}}_2 \text{ ou } \sigma_1(\check{\mathscr{V}}_1) \cup \sigma(\check{\mathscr{V}}_2) \text{ est une chaîne maximale} \end{cases}$$

$\ell_1 \vee \ell_2$ est un alignement de $\mathscr{A}(S,C)$. L'opération d'élargissement admet l'alignement vide $\ell_\emptyset$ comme élément neutre.

On observe sur la figure 2.12 que l'opération a regroupé des liens en désaccord, proposant ainsi un consensus formant un alignement moins fin.

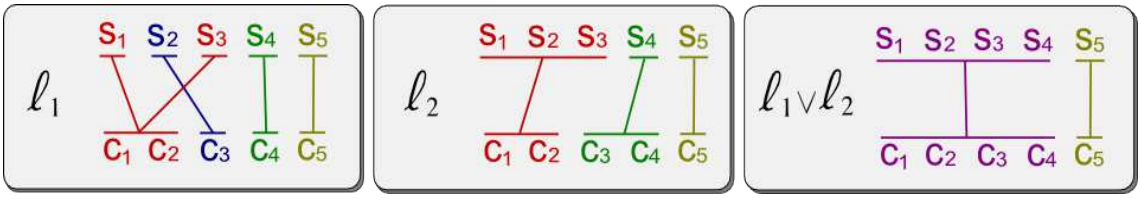


Figure 2.12 – L'opérations d'élargissement sur un exemple

On définit la relation de finesse entre alignements (voir figure 2.13). Intuitivement, un alignement ℓ_1 sera dit plus fin que ℓ_2 si tout couple de mots (s_i, c_j) liés par ℓ_1 est lié par ℓ_2 :

Définition 11. *On dira que ℓ_1 est **plus fin que** ℓ_2 (noté $\ell_1 \prec \ell_2$) si les trois conditions suivantes sont remplies :*

1. *Tout élément de $\mathscr{V}_1$ est inclus dans un élément de $\mathscr{V}_2$*
2. *Tout élément de $\mathscr{W}_1$ est inclus dans un élément de $\mathscr{W}_2$*
3. *Pour tous V_1 de $\mathscr{V}_1$ et V_2 de $\mathscr{V}_2$ tels que $V_1 \subset V_2$, on a $\sigma_1(V_1) \subset \sigma_2(V_2)$*

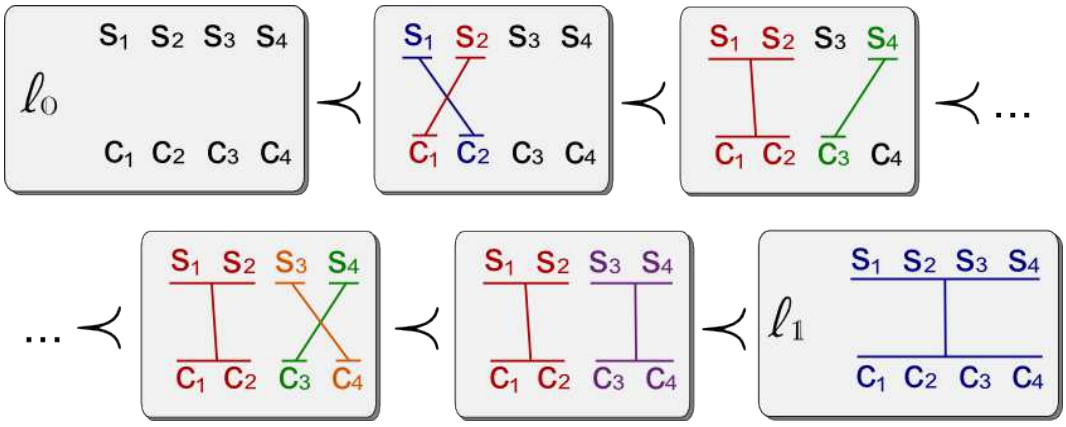


Figure 2.13 – Des alignements de moins en moins fins

Propriété 1. *On a les propriétés suivantes :*

- $\prec$ *est une relation d'ordre partiel sur l'ensemble* $\mathscr{A}(S,C)$.
- $\ell_1 \wedge \ell_2$ *est l'alignement le moins fin parmi les alignements à la fois plus fins que* ℓ_1 *et que* ℓ_2.
- $\ell_1 \vee \ell_2$ *est l'alignement le plus fin parmi les alignements à la fois moins fins que* ℓ_1 *et que* ℓ_2.
- $(\mathscr{A}(V,W), \wedge, \vee, \prec)$ *forme un treillis complet de plus grand élément l'alignement grossier* $\ell_{\mathbb{1}}$ *et de plus petit élément l'alignement vide* ℓ_{O} *(définis au 2.2.2).*

Nous venons de munir l'ensemble des alignements $\mathscr{A}(S,C)$ d'une structure de treillis. La relation de finesse '$\prec$' permet de les comparer lorsqu'ils sont compatibles et les opérations $\wedge$ et $\vee$ permettent de les combiner. On peut observer à la figure 2.14 le diagramme de Hasse représentant l'espace des alignements entre une phrase source de longueur 3 et une phrase cible de longueur 2.

On remarque qu'en général, l'opération $\wedge$ dégrade la couverture, tandis que $\vee$ l'augmente. La dégradation par l'affinement $\wedge$ pourra être assez importante lorsque les alignements présentent un désaccord structurel relativement important. On ne peut en toute généralité énoncer que la propriété suivante :

Propriété 2. *Les supports vérifient les inclusions suivantes :*

- *Le support source (resp. cible) de* $\ell_1 \wedge \ell_2$ *est* ***inclus*** *dans l'intersection des supports source (resp. cible) couverts par* ℓ_1 *et* ℓ_2
- *Le support source (resp. cible) de* $\ell_1 \vee \ell_2$*) est* ***égal*** *à l'union des supports source (resp. cible) couverts par* ℓ_1 *et* ℓ_2

Remarque 2. *La couverture est une fonction croissante de* $\mathscr{A}(n,m)$ *dans* $\mathbb{N}$*, mais n'est pas une valuation car elle n'est pas strictement croissante.*

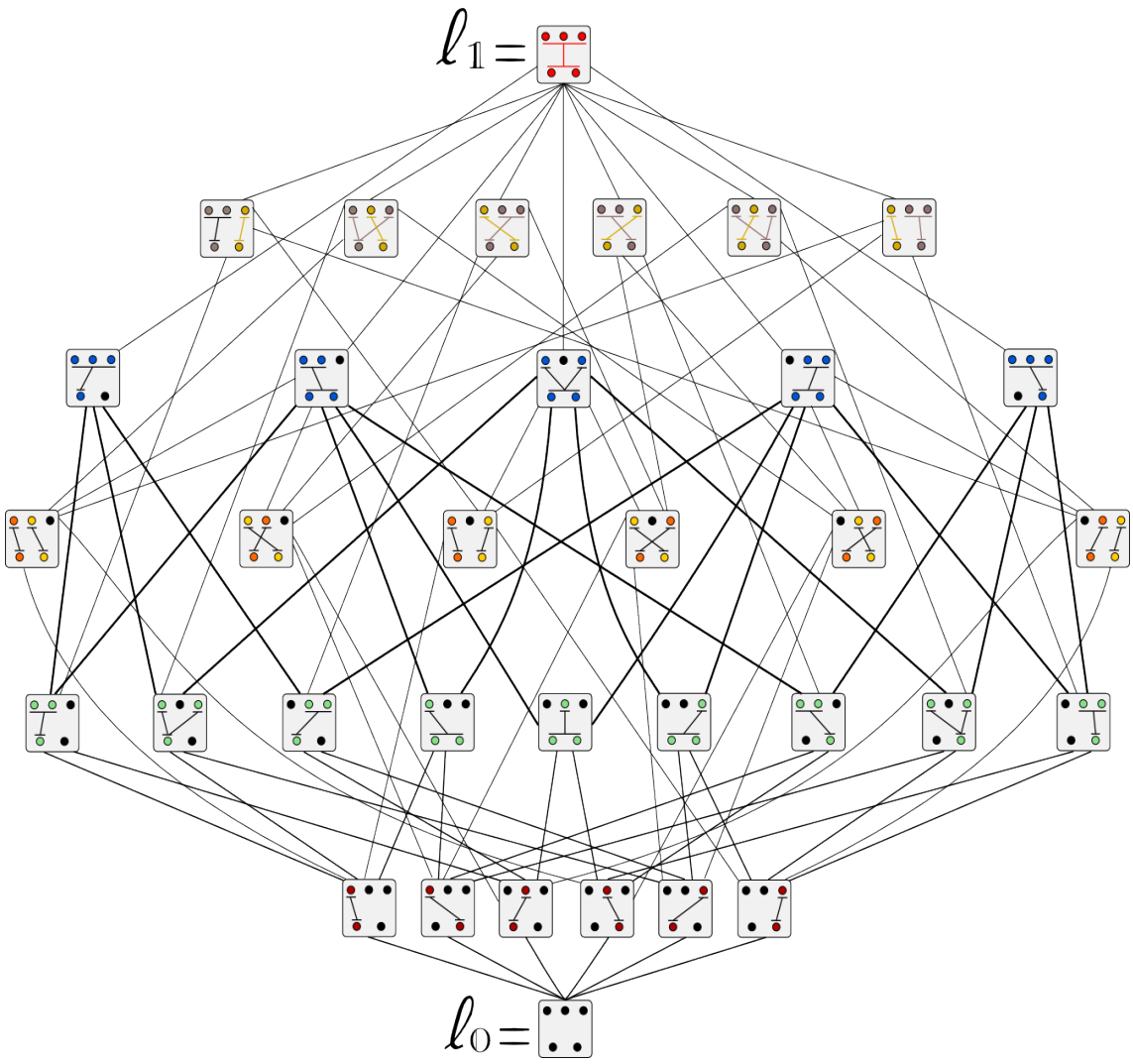


Figure 2.14 – Diagramme de Hasse de $\mathscr{A}(3,2)$

2.2.3 Les sous-ensembles de modèles existants

Les différents modèles d'alignement existants imposent généralement des contraintes afin d'alléger les traitements algorithmiques. Il en résulte des *types d'alignement* variés.

En se basant sur une biphrase $B = (S,C)$ de longueur (n,m) donnée, les différents modèles décrits précédemment au 1.1 (mots, ITG, asymétriques, intervalles de mots,etc.) *opèrent sur des ensembles* que l'on peut injecter naturellement dans $\mathscr{A}(S,C)$. Il peut lui-même être injecté dans $\mathscr{B}(S,C)$, l'ensemble des graphes bipartis sur des nœuds étiquetés par les mots de S et C. En dénombrant les espaces sur lesquels opèrent les différents modèles évoqués, nous espérons ici donner une idée de leur expressivité. Bien sûr, il s'agit d'une description quantitative. En aucun cas il ne faut voir les espaces d'alignements comme l'ensemble des configurations obtenues en confrontant des paires de langues, mais plus vraisemblablement comme l'espace circonscrit par les limites de l'expressivité de chaque modèle. Les différentes études sur la divergence nous encouragent à opter pour des barrières très reculées entourant un très vaste espace. En choisissant $\mathscr{A}(S,C)$ comme l'espace le plus général au lieu de $\mathscr{B}(S,C)$, nous signalons toutefois que cet espace ne doit pas se montrer inutilement trop général. Nous passons en revue quelques exemples de modèles classiques et dénombrons chaque fois les alignements de *couverture maximale* du modèle concerné, pour alléger les formules. Pour dénombrer aussi les *alignements partiels*, il faudra sommer sur toutes les sous-biphrases.

Alignements monotones

On note $\bar{\mathscr{A}}_{monotone}(n,m)$, l'ensemble des alignements monotones couvrant une biphrase de longueur (n,m). Les groupes alignés sont contigus et découpent les côtés source et cible en le même nombre d'intervalles. On remarque qu'il y a $\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$ manières de séparer l'ensemble $[\![1,n]\!]$ en k intervalles, ainsi :

$$|\bar{\mathscr{A}}_{monotone}(n,m)| = \sum_{k=1}^{\min(n,m)} \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix}$$

Pour $m = n$, la formule se simplifie en

$$|\bar{\mathscr{A}}_{monotone}(n,n)| = \begin{pmatrix} 2\cdot(n-1) \\ n-1 \end{pmatrix}$$

Alignements de mots

On note $\bar{\mathscr{A}}_{\text{1-1}}$ l'ensemble des alignements à liens simples de couverture maximale sur une biphrase de longueur (n,m) (non couvrants quand $n \neq m$). On définit l'espace $\mathscr{A}_{\text{1-to-1}}(n,m)$ comme le produit cartésien $[\![1,n]\!] \times [\![1,m]\!]$ qui s'injecte dans $\mathscr{A}(n,m)$. Supposons que $n \leq m$, alors le nombre d'alignements de couverture maximale correspond au nombre d'injections de $[\![1,n]\!]$ dans $[\![1,m]\!]$, c'est-à-dire au nombre d'arrangements A_m^n :

$$|\bar{\mathscr{A}}_{\text{1-to-1}}(n,m)| = \frac{n!}{(m-n)!} \quad (\text{pour } n \leq m)$$

Pour $m = n$, on retrouve $n!$, le nombre de permutations sur un ensemble à n éléments.

Certains modèles opèrent sur un sous-ensemble strict de $\mathscr{A}_{\text{1-to-1}}(n,m)$ comme par exemple celui des ITG dont les alignements excluent les configurations dites "*inside-out*" : on note l'espace d'alignements $\mathscr{A}_{ITG}(n,m)$. Le nombre d'alignements **couvrants** noté $\bar{\mathscr{A}}_{ITG}(n,n)$ est le n-ième nombre de Schröder S_n (voir [163], [131] et [151]) :

$$S_n = \sum_{i=0}^{n} \begin{pmatrix} 2n-i \\ i \end{pmatrix} C_{n-i}$$

où C_k est le k-ième nombre de Catalan :

$$C_k = \frac{1}{k+1} \begin{pmatrix} 2k \\ k \end{pmatrix}$$

Et donc, si on suppose que $n \leq m$:

$$|\bar{\mathscr{A}}_{ITG}(n,m)| = \begin{pmatrix} n \\ m \end{pmatrix} S_n$$

Alignements asymétriques

Il y a deux ensembles d'alignements asymétriques notés $\mathscr{A}_{1-N}(n,m)$ et $\mathscr{A}_{N-1}(n,m)$ opérant sur une biphrase de longueur (n,m). Pour le premier, un élément cible est lié à au plus un élément source et pour le deuxième, ce sont les éléments source qui ne participent qu'à une correspondance. Nous raisonnons sur le premier seulement (par symétrie on peut transposer le raisonnement au second). L'ensemble peut formellement être représenté par l'ensemble des fonctions de $[\![1,m]\!]$ dans $[\![1,n]\!]$, un élément cible n'ayant bien sûr qu'une seule image dans $[\![1,n]\!]$. Nous ne dénombrons pas seulement les alignements de couverture maximale car sur une biphrase de longueur (n,n) il s'agit simplement des fonctions bijectives (il y en a donc $n!$), et l'expressivité est la même que pour un alignement de mots. Le nombre d'alignements vaut donc :

$$|\mathscr{A}_{1-N}(n,m)| = n^m$$

Les alignements asymétriques présentés dans l'état de l'art proviennent des modèles IBM qui reposent sur des modélisations par les n-grammes. Le nombre de mots source atteignables est en fait limité par une constante k, correspondant à une fenêtre glissante. Il est dénombré dans [163] pour $n = m$ qui donne un résultat paramétré par la constante k. On note $\mathscr{A}_{1-k}(n,n)$, l'ensemble de ces alignements, qui représente aussi un ensemble d'alignements non nécessairement couvrants :

$$|\mathscr{A}_{1-k}(n,n)| = \begin{cases} k^{n-k} \cdot k! & si\ n > k \\ n! & si\ n \leq k \end{cases}$$

De même que pour l'ensemble général $\mathscr{A}_{1-N}(n,m)$, la formule pour $\mathscr{A}_{1-k}(n,m)$ dénombre tous les alignements et pas seulement les alignements de couverture maximale. Nous la comparons tout de même quantitativement aux alignements de mots couvrants et aux ITG maximaux dans la figure 2.15, page 58. Le nombre $\mathscr{A}_{1-k}(n,n)$, sans être comparable par inclusion ni à $\bar{\mathscr{A}}_{1-to-1}(n,n)$, ni à $\bar{\mathscr{A}}_{ITG}(n,n)$, reste moins important (quantitativement) que les deux pour $k = 4$. Si k augmente, le nombre d'alignements

asymétriques de type IBM dépasse le nombre d'alignements ITG, mais reste dominé par le nombre d'alignements de mots couvrants, dans la mesure où l'un a une croissance exponentielle et l'autre une croissance polynomiale.

Nous comparons dans la figure 2.16, page 58, les différents espaces d'alignements $\bar{\mathscr{A}}_{1-to-1}(n,n)$, $\bar{\mathscr{A}}_{1-N}(n,n)$ et $\bar{\mathscr{A}}_{span}(n,n)$ (incluant les alignements partiels afin de pouvoir comparer avec les alignements asymétriques). L'espace des alignements asymétriques n'est pas comparable en terme d'inclusion avec l'espace des alignements d'intervalles de mots, puisqu'il permet des défauts de contiguïté. À partir de $n = 12$, l'espace asymétrique comptera plus d'éléments que l'espace des intervalles de mots (sans qu'il s'agisse des mêmes types d'alignements).

Alignements d'intervalles de mots

Nous avons vu que les systèmes d'alignement ont tendance à s'orienter vers un alignement de groupes de mots contigus. Nous notons $\mathscr{A}_{span}(n,m)$, le sous-ensemble des alignements d'intervalles de mots sur une biphrase de longueur (n,m). De même que pour les alignements monotones, on considère le nombre de manières de segmenter en k intervalles, mais ici les liens peuvent se croiser de $k!$ manières :

$$|\bar{\mathscr{A}}_{span}(n,m)| = \sum_{k=1}^{\min(n,m)} \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) \left(\begin{array}{c} m-1 \\ k-1 \end{array} \right) k!$$

Les approches utilisant des tables d'intervalles de mots imposent généralement une longueur inférieure à une constante K. On note $\mathscr{A}_{K-span}(n,m)$ l'ensemble des alignements respectant cette contrainte, la formule les dénombrant devient paramétrée par la constante K :

$$|\bar{\mathscr{A}}_{K-span}(n,m)| = \sum_{k=\lceil \max(n,m)/K \rceil}^{\min(n,m)} \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) \left(\begin{array}{c} m-1 \\ k-1 \end{array} \right) k!$$

On voit que l'espace est réduit, mais reste d'une grandeur asymptotiquement équivalente.

Alignements non contraints

Enfin, on se place sur l'espace total des alignements $\mathscr{A}(n,m)$ incluant tous les précédents. Pour cela, on rappelle la formule explicite du nombre de partitions en k sous-ensembles d'un ensemble à n éléments (appelé un nombre de Stirling de seconde espèce) :

$$\left\{ \begin{array}{c} n \\ k \end{array} \right\} = \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-j} \left(\begin{array}{c} k \\ j \end{array} \right) j^n$$

On note $\bar{\mathscr{A}}(n,m)$ l'ensemble des alignements couvrants sur une biphrase de longueur (n,m). D'après la définition, on peut les dénombrer à l'aide du nombre de partitions de $[\![1,n]\!]$ et $[\![1,m]\!]$ et on obtient la formule suivante :

$$|\bar{\mathscr{A}}(n,m)| = \sum_{k=1}^{\min(n,m)} \left\{ \begin{array}{c} n \\ k \end{array} \right\} \left\{ \begin{array}{c} m \\ k \end{array} \right\} k!$$

Nous voyons sur la figure 2.17, page 59, que l'espace général $\bar{\mathscr{A}}(n,n)$ domine largement les espaces précédents, ce qui est légitime puisqu'il ne présente aucune contrainte sur la contiguïté, les croisements et la taille des groupes.

Graphes bipartis

L'ensemble des alignements $\mathscr{A}(n,m)$ s'injecte naturellement dans l'ensemble des graphes bipartis entre deux ensembles de nœuds étiquetés par les intervalles $[\![1,n]\!]$ et $[\![1,m]\!]$, noté $\mathscr{B}(n,m)$. Nous avons jusqu'ici dénombré des ensembles d'alignements couvrants pour que les cardinalité des ensembles décrits restent exprimables par une formule "simple". L'ensemble $\mathscr{B}(n,m)$ est de cardinalité $2^{n \cdot m}$, mais le cardinal du sous-ensemble des graphes bipartis *couvrants* est plus compliqué à exprimer ; nous n'avons pas pu simplifier la formule obtenue et la donnons en l'état.

Nous avons en fait dénombré l'ensemble des alignements non couvrants en le dé-

composant en les sous-ensembles ne couvrants pas un mot. On pose :

$$\begin{cases} NS, \text{ l'ensemble des graphes ne couvrant pas un mot source} \\ NC, \text{ l'ensemble des graphes ne couvrant pas un mot cible} \end{cases}$$

Cela nous mène à dénombrer le cardinal de l'ensemble couvrant $\bar{\mathscr{B}}(n,m)$ décomposé ainsi :

$$|\bar{\mathscr{B}}(n,m)| = |\mathscr{B}(n,m) \setminus (NS \cup NC)| = 2^{n \cdot m} - |NS| - |NC| + |NS \cap NC|$$

On décompose NS et NC comme l'union des graphes ne couvrant un mot source ou cible. En utilisant le principe d'inclusion-exclusion généralisé, on obtiendra après simplification :

$$\bar{\mathscr{B}}(n,m) = (2^n - 1)^m - \sum_{k=1}^{n} (-1)^k \cdot \begin{pmatrix} n \\ k \end{pmatrix} \cdot (2^{n-k} - 1)^m$$

Cette quantité peut être largement majorée par $2^{n \cdot m}$ qui correspond au cardinal de $\mathscr{B}(n,m)$. La figure 2.18, page 59, donne une comparaison générale de l'expressivité des différents modèles. La courbe représentant $|\bar{\mathscr{B}}(n,n)|$ présente une croissance sous-quadratique en échelle logarithmique (quadratique pour $|\mathscr{B}(n,n)|$) tandis que les autres sont dominées par une croissance simplement exponentielle, relativement "raisonnable".

Les courbes

Les courbes comparant quantitativement les ensembles évoqués expriment les cardinalités par rapport à la longueur d'une biphrase de taille (n,n) (n en abscisse). Nous comparons généralement les espaces maximaux (alignements couvrants), surlignés d'une barre, sauf pour comparer les alignements asymétriques qui possèdent très peu d'alignements couvrants. L'échelle est logarithmique pour mieux visualiser les évolutions. Les même courbes tracées avec une échelle normale seraient illisibles, les courbes dominées apparaissant alors "collées" à l'axe des abscisses.

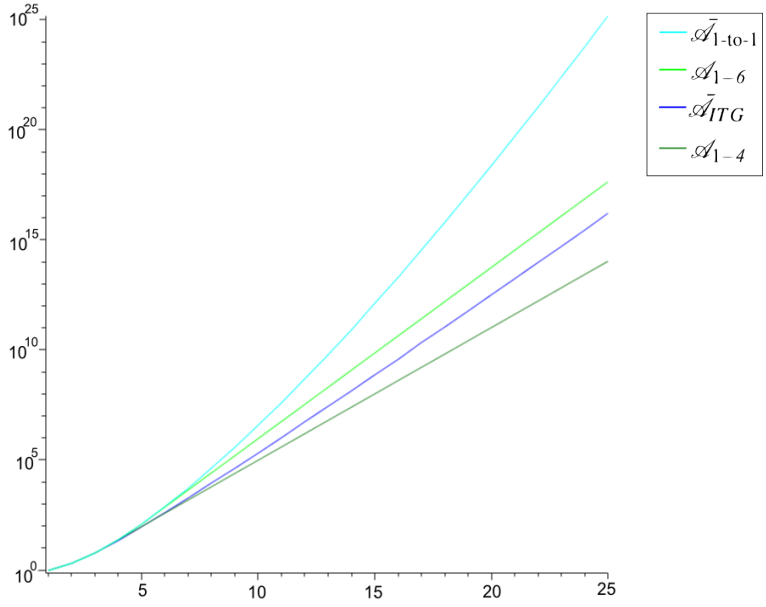


Figure 2.15 – Les cardinalités par rapport à la longueur (n,n) (échelle *log*)

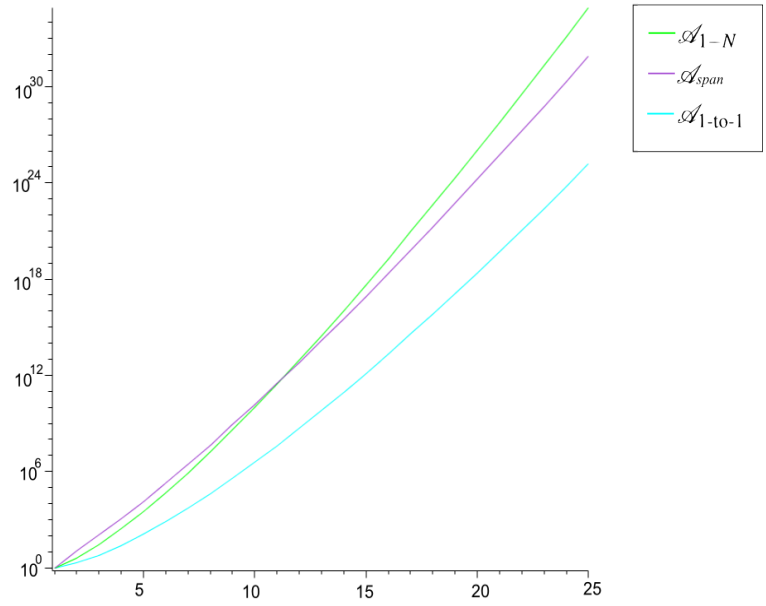


Figure 2.16 – Alignements éventuellement partiels pour ces trois ensembles

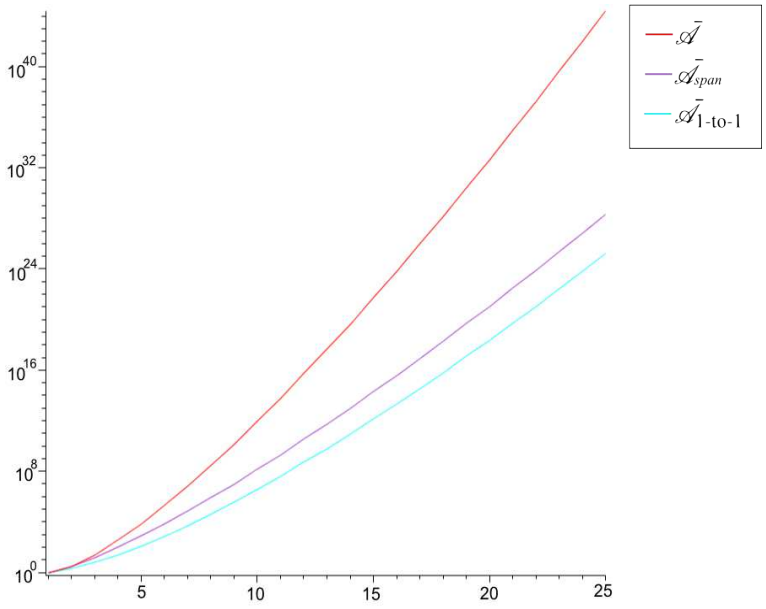


Figure 2.17 – L'espace total inclut tous les autres

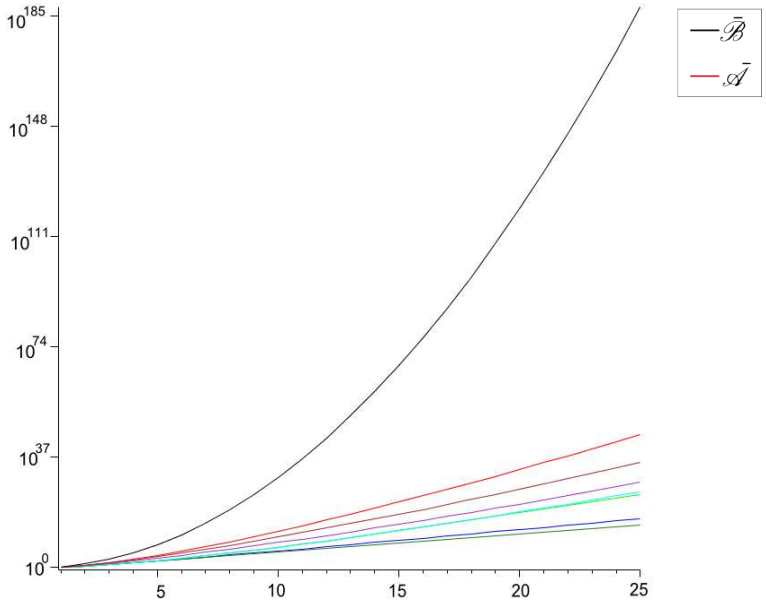


Figure 2.18 – L'espace des graphes bipartis couvrants (échelle *log*)

2.3 Comparaison entre alignements

2.3.1 Avant-propos

La comparaison entre alignements intervient généralement pour mesurer la qualité des résultats d'un outil sur un corpus de référence [82], et aussi pour comparer l'accord entre des annotateurs lors de la constitution manuelle d'un corpus de référence [95].

2.3.1.1 Des mesures d'accord entre alignements

La *précision*, le *rappel* et la *F-mesure* sont des mesures classiques en recherche d'information. Elles sont souvent utilisées en alignement pour comparer des résultats produits automatiquement A à des résultats de référence C (éventuellement produits par des annotateurs). La précision P exprime la proportion d'informations pertinentes parmi les informations trouvées A et le rappel R la proportion d'informations pertinentes retrouvées parmi les informations de référence C.

$$\begin{cases} P &= \frac{|A \cap C|}{|A|} \\ R &= \frac{|A \cap C|}{|C|} \end{cases}$$

Ces mesures sont adaptées pour comparer des alignements de mots $\mathscr{A}_{1-to-1}(n,m)$. Les informations A et C peuvent être vues comme des sous-ensembles du produit cartésien $[\![1,n]\!] \times [\![1,m]\!]$. *La précision exprime la qualité de l'alignement produit et le rappel décrit sa couverture*. Un alignement est considéré comme bon s'il a à la fois une bonne précision et un bon rappel. C'est ce qu'exprime la F-mesure [119] (analogue au coefficient de Dice [49] pour $\alpha = 1$) qui combine les deux par une moyenne harmonique pondérée :

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}} = \left(\frac{R^{-1} + P^{-1}}{2} \right)^{-1} = \frac{2 \cdot P \cdot R}{P + R} \qquad F_\alpha = (1 + \alpha^2) \cdot \frac{P \cdot R}{\alpha^2 P + R}$$

Ces mesures sont généralement plus adaptées à des alignements de mots qu'à des alignements plus généraux pour lesquels différentes adaptations existent.

La mesure *AER*(*Alignment Error Rate*) [82] se base sur la précision, le rappel et le F-score pour évaluer des alignements asymétriques par rapport à un corpus de référence. On rappelle qu'un alignement asymétrique permet de lier un mot source à plusieurs mots cible. Un de ces alignements A peut être représenté comme un ensemble de liens :

$$A = \{(i, a_i) \in \mathbb{N}^2, i \in [\![1, n]\!], a_i \in [\![1, m]\!]\}$$

Dans le corpus de référence, deux types de liens sont différenciés, les liens sûrs S et les liens probables P. Les liens S sont communs aux annotateurs ayant constitué le corpus de référence, contrairement aux liens probables. Ainsi, chaque alignement est séparé en deux sous-ensembles disjoints de liens, $A_{ref} = S \cup P$. Le rappel est exprimé par rapport aux liens sûrs S et la précision par rapport aux liens probables P :

$$rappel = \frac{|A \cap S|}{|S|} \qquad\qquad precision = \frac{|A \cap P|}{|A|}$$

La mesure *AER* s'inspire de la F-mesure :

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

Des mesure pondérées sur des alignements de groupes viennent par ailleurs étendre les mesures d'alignement de mots ou asymétriques en fractionnant des groupes de mots liés en un ensemble de liens simples. Un lien associant un groupe source de longueur p à un groupe cible de longueur q est vu comme $p \cdot q$ liens simples, en nous ramenant à l'ensemble des graphes bipartis (voir figure 2.19).

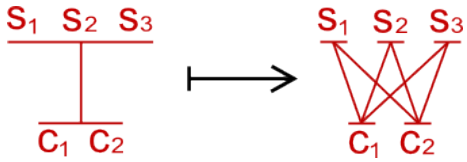


Figure 2.19 – Démultiplication des liens pour mesures pondérées

Les liens provenant de groupes liés prennent alors une importance considérable par

rapport aux liens "réellement" simples. Une solution consiste à pondérer chaque lien pour tenir compte de la taille des groupes dont il est issu.

Dans le projet Blinker [95] sept experts alignaient manuellement un corpus bilingue. Une mesure d'accord inter-annotateurs est proposée pour justifier la qualité du corpus formé. Il s'agit d'une mesure d'accord de Dice adaptée via l'éclatement en liens simples. Le poid d'un lien (u,v) est :

$$w(u,v) = \frac{1}{max(a(u)+a(v))}$$

où $a(u)$ est le nombre de liens sortant du mot u. Pour ses expériences, une mesure similaire est utilisée normalisant la fonction de poids.

Similairement, le *WAA* (Word Alignment Agreement) dans [46] étend la mesure *AER* a des alignements de groupes de mots en se ramenant aussi à l'ensemble des liens simples pondérés. La fonction de poids w est adaptée suivant le constat suivant : pour deux groupes source et cible liés de tailles respectives p et q, on extrait $p \cdot q$ liens (car l'ensemble forme une biclique) et on souhaiterait que le poids cumulé de tous ces liens vaille $\frac{p+q}{2}$ (le poids d'un alignement monotone dont chaque lien a un poids de 1). Pour un lien l issu d'un bi-fragment de taille (p,q), le poids vaut :

$$w(l) = \frac{p+q}{2 \cdot p \cdot q}$$

Ce type d'approche ne considère pas les groupes liés comme des unités, et ce défaut structurel aurait tendance à surévaluer la précision (voir par exemple [108]).

La mesure *CPER* (Consistency Phrase Error Rate) [10] tente d'éviter le fractionnement des alignements en liens simples en considérant les groupes liés comme des éléments uniques (elle reprend la mesure *AER*). Un seul lien est présent par groupe lié, et donc, contrairement aux mesures fractionnant l'alignement, la mesure *CPER* ne tient pas compte du recouvrement de l'alignement produit par rapport à l'alignement de référence. La mesure *CPER* aurait tendance à surévaluer le rappel [108].

Une critique que l'on peut adresser à ces différentes mesures lorsqu'elles comparent

des alignements automatiques à des alignements manuels, est qu'elles ne prennent pas en compte la différence structurelle séparant les deux [130].

2.3.1.2 Le point de vue transformationnel

Nous nous intéressons à des mesures de comparaison évaluant l'effort nécessaire pour "*transformer*" un élément à un autre. Ce type de mesure a l'avantage d'être structurellement proche des modèles décrits et d'être naturellement interprétable. Il n'existe pas, à notre connaissance, d'indices ou de mesures, empiriques ou théoriques, comparant des alignements sous-phrastiques d'un point de vue transformationnel. Nous donnons ici quelques exemples de telles distances sur différents types de données avant de proposer trois distances transformationnelles entre biphrases alignées.

La *distance de Levenshtein* [89] est un exemple connu de distance sur l'ensemble des chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut *supprimer*, *insérer* ou *remplacer* pour passer d'une chaîne à l'autre. Elle est calculable en un nombre quadratique d'opérations (des optimisations existent).

Une mesure plus empirique est utilisée dans [104] pour quantifier la qualité d'une traduction produite par un outil automatique. L'effort de post-édition des résultats pour être corrigés est enregistré et le nombre de touches tapées participe à une mesure d'évaluation :

> *The number of keystrokes is interesting in that it indicates how much work may be involved in post-editing.*

Dans le cas de l'alignement, il s'agirait plutôt du nombre de clics dans une interface d'alignement manuel du type Blinker.

On remarque également que le *coefficient de Dice* (défini page 60) entre deux ensembles A et B peut prendre une interprétation transformationnelle. En effet, l'indice de dissimilarité associé $1 - D(A,B) = 1 - \frac{2 \cdot |A \cap B|}{|A| + |B|}$ est le cardinal normalisé de la différence symétrique $|A \Delta B| = |A| + |B| - 2 \cdot |A \cap B|$. Il s'agit du nombre minimal de *suppressions* et d'*insertions* nécessaires pour passer de A à B.

Dans une approche incluant un effort d'annotation, il est tentant d'évaluer la similarité entre deux alignements en terme d'opérations d'édition pour transformer l'un en l'autre. Nous proposerons dans les parties suivantes deux mesures de similarité entre alignements qui sont en fait des distances exprimant un effort minimal théorique de post-édition. Nous nous inspirons pour cela de distances utilisées (notamment en *psychométrie* et en *clustering*) pour comparer des *partitions d'ensembles de données*. Dans [47], cinq transformations élémentaires sont proposées à partir desquelles sont introduites plusieurs distances d'édition. Les transformations qui portent sur les classes de partitions d'un ensemble fini sont représentées dans la figure 2.20 (la *suppression*, l'*insertion*, la *division*, l'*agrégation* et le *transfert*). Soit $\mathcal{P} = (P_i)_{i=1,\dots,k}$ une partition de X :

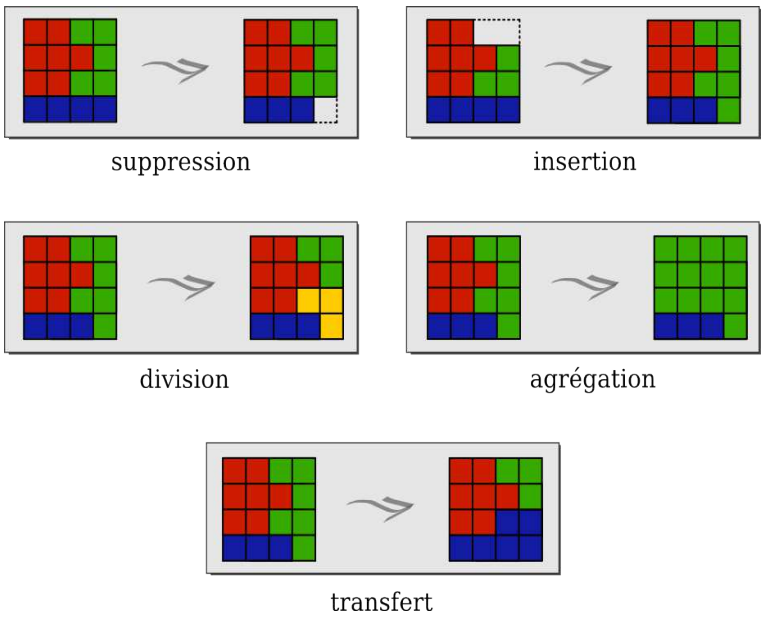


Figure 2.20 – Les cinq transformations élémentaires pour un ensemble à 16 éléments

- La suppression consiste à retirer un ensemble Y d'une classe P_i afin d'obtenir une partition de $X \setminus Y$.

- L'insertion consiste en l'opération inverse d'ajout d'un ensemble Y (extérieur à X) à une classe P_i pour former une partition de $X \cup Y$.

- La division peut être vue comme une *suppression* suivie d'une *insertion* portant toutes les deux sur les mêmes éléments ou comme la division d'une classe en deux.

- L'agrégation est l'opération inverse, qui fusionne deux classes.

- Le transfert peut être vu comme deux opérations successives de *division* et d'*agrégation* sur les mêmes éléments. On remarque qu'un transfert de la classe entière correspond à une opération d'agrégation.

Nous proposons des transformations analogues sur un alignement, représentées dans la figure 2.21. Une opération supplémentaire est introduite pour raffiner l'opération de transfert. On pose $\ell = (\mathscr{V}, \mathscr{W}, \sigma)$ un alignement de $\mathscr{A}(n,m)$.

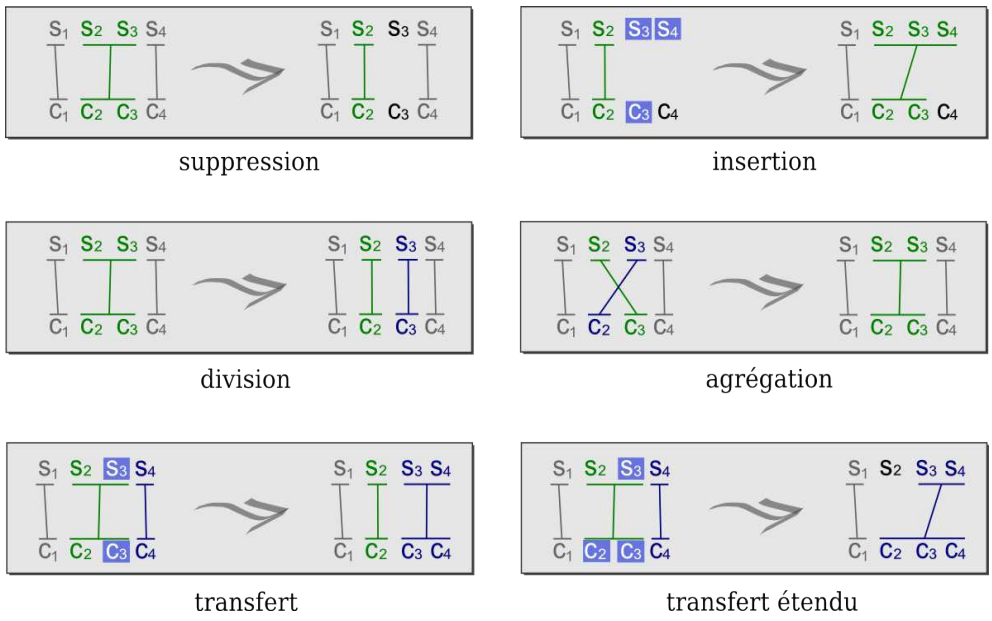


Figure 2.21 – Les cinq transformations élémentaires pour une biphrase de longueur $(4,4)$

- La suppression retire des éléments à un groupe lié (V_i, W_i) (des liens) pour former un sous-alignement de ℓ (l'alignement reste dans $\mathscr{A}(n,m)$, mais est partiel).

- L'insertion, de même que la suppression, reste interne à $\mathscr{A}(n,m)$: en partant d'un alignement partiel, on ajoute des mots non liés (dans aucun des V_j ni des W_j) à un groupe déjà formé (V_i, W_i).

- La division sur un alignement permet de séparer un groupe lié (V_i, W_i) en deux nouveaux groupes. Si $X \subset V_i$ et $Y \subset W_i$, les deux nouveaux groupes seront $(V_i \backslash X, W_i \backslash Y)$ et (X,Y). X et Y doivent être deux sous-ensembles stricts, sans quoi l'opération créerait

des "groupes monolingues" en violation avec la définition 5 page 44. Le résultat produit un alignement plus fin.

- L'agrégation est l'opération inverse qui fusionne deux groupes (V_i, W_i) et (V_j, W_j) en $(V_i \cup V_j, W_i \cup W_j)$. Elle produit un alignement plus grossier.

- Le transfert peut être vu comme deux opérations successives de *division* et d'*agrégation*. Il considère deux groupes (V_i, W_i) et (V_j, W_j) ainsi que $X \subset V_i$ et $Y \subset W_i$ et les remplace par les deux classes $(V_i \backslash X, W_i \backslash Y)$ et $(V_j \cup X, W_i \cup Y)$. On ajoute une condition supplémentaire sur X et Y pour ne pas former de groupes source (ou cible) isolés : $X = P_i \Leftrightarrow Y = W_i$. Ainsi, les deux côtés d'un groupe peuvent être entièrement transférés, mais seulement simultanément, auquel cas il s'agit d'une opération d'agrégation.

- Le transfert étendu généralise l'opération de transfert en permettant de "vider" complètement un groupe (source et/ou cible). Il s'agit d'une transformation plus naturelle pour une interface d'alignement manuel. On reprend la même définition, sauf que l'on retire la condition de simultanéité. Si $X = V_i$, on efface simplement le groupe (V_i, W_i) et on remplace le groupe (V_j, W_j) par $(V_j \cup V_i, W_j \cup Y)$. Les mots numérotés par $W_j \backslash Y$ ont été "déliés".

Nous adaptons aux alignements deux distances d'édition de partitions basées sur les transformations précédentes : la *distance des transferts* et la *distance des divisions*.

2.3.2 Les distances des transferts

Nous adaptons deux distances entre partitions basées sur les opérations de transfert, la première par transfert de singletons et la deuxième par transfert de sous-ensembles. Nous rappelons les définitions et les étapes nécessaires à leur calcul sur les partitions et proposons des distances analogues sur l'ensemble des alignements.

2.3.2.1 Transfert unitaire

La *distance des transferts unitaires* introduite dans [116] est une distance entre deux partitions (mettons $\mathcal{P}$ et $\mathcal{Q}$) d'un ensemble X à n éléments. Ici, les opérations concernées sont *le transfert unitaire* et *la division unitaire*, c'est-à-dire une opération de transfert

qui déplace un seul élément et une division qui crée une classe singleton (figure 2.22).

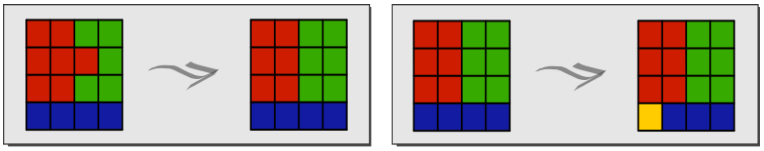


Figure 2.22 – *Transfert de singleton* et *division unitaire* sur une partition

Définition 12. *La distance des transferts unitaire entre* $\mathcal{P}$ *et* $\mathcal{Q}$ *notée* $t_1(\mathcal{P},\mathcal{Q})$ *correspond au nombre minimum d'opérations de transfert unitaire et de division unitaire nécessaires pour passer de* $\mathcal{P}$ *à* $\mathcal{Q}$.

Nous donnons le détail d'une technique de calcul (voir par exemple [34]) de cette distance en se ramenenant à un problème de couplage. Tout d'abord, on relache une contrainte sur la définition de partition en autorisant des classes vides. Cela nous permet de supposer, sans perte de généralité, que les partitions $\mathcal{P}$ et $\mathcal{Q}$ ont le même nombre k de classes, quittes à compléter une des deux par des classes vides. De cette manière, on peut ne considérer que des opérations de transfert, la division correspondra à un transfert vers une classe vide. Nous raisonnerons à partir de l'exemple suivant sur l'ensemble $X = [\![1,8]\!]$:

$$\begin{cases} \mathcal{P} = \begin{array}{|c|c|c|}\hline 1 & 2345 & 678 \\ \hline \end{array} \\ \mathcal{Q} = \begin{array}{|c|c|c|}\hline 12346 & 5 & 78 \\ \hline \end{array} \end{cases} \tag{2.1}$$

Pour cet exemple, $t_1(\mathcal{P},\mathcal{Q}) = 3$, les trois étapes étant (il n'y a pas toujours unicité des opérations de transfert) :

$$\mathcal{P} = \begin{array}{|c|c|c|}\hline 1 & 2345 & 678 \\ \hline \end{array} \begin{array}{l} \overset{1}{\longmapsto} \begin{array}{|c|c|c|}\hline & 12345 & 678 \\ \hline \end{array} \\ \overset{5}{\longmapsto} \begin{array}{|c|c|c|}\hline 5 & 1234 & 678 \\ \hline \end{array} \\ \overset{6}{\longmapsto} \begin{array}{|c|c|c|}\hline 5 & 12346 & 78 \\ \hline \end{array} = \mathcal{Q} \end{array}$$

On définit la *similarité* entre $\mathcal{P}$ et $\mathcal{Q}$, notée $s(\mathcal{P},\mathcal{Q})$ comme le nombre maximum d'éléments qui n'ont pas besoin d'être transférés pour passer d'une partition à l'autre.

La résolution du problème passe par l'estimation de la similarité car elle est liée à la distance de transfert par l'équation :

$$t_1(\mathcal{P},\mathcal{Q}) = n - s(\mathcal{P},\mathcal{Q}) \tag{2.2}$$

Pour calculer $s(\mathcal{P},\mathcal{Q})$, on introduit la matrice de similarité $T \in \mathscr{M}_k(\mathbb{N})$ qui compte le nombre d'éléments communs entre les classes $(P_i)_{i\in[\![1,k]\!]}$ de $\mathcal{P}$ et les classes $(Q_j)_{j\in[\![1,k]\!]}$ de $\mathcal{Q}$:

$$T_{ij} = |P_i \cap Q_j|$$

Sur notre exemple, la matrice de similarité est :

$$T = \begin{pmatrix} 1 & \textcircled{0} & 0 \\ \textcircled{3} & 1 & 1 \\ 1 & 0 & \textcircled{2} \end{pmatrix}$$

Le problème de maximisation de la similarité revient à trouver une correspondance σ bijective entre les classes de $\mathcal{P}$ et $\mathcal{Q}$ (σ est une permutation de $[\![1,k]\!]$) qui devra maximiser le nombre d'éléments non déplacés). La similarité entre $\mathcal{P}$ et $\mathcal{Q}$ est ramenée à un problème de maximisation sur la matrice de similarité :

$$s(\mathcal{P},\mathcal{Q}) = \max_{\sigma}\left(\sum_{i=1}^{k} T_{i,\sigma(i)}\right) \tag{2.3}$$

Sur notre exemple, on voit que les trois opérations sur $\mathcal{P}$ réordonnent implicitement les classes de $\mathcal{Q}$ suivant la permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$ qui maximise bien la similarité. On observe a posteriori sur la matrice de similarité T les valeurs entourées qui correspondent aux indices maximisant $\sum_{i=1}^{k} T_{i,\sigma(i)}$ et vérifions bien d'après la formule 2.2 que $t_1(\mathcal{P},\mathcal{Q}) = 8-(0+3+2) = 3$.

La résolution du problème passe par le calcul de σ (équation 2.3). Il s'agit d'une recherche de *couplage de poids maximum* sur le graphe biparti complet $B_{k,k}$ valué. Il est résoluble par la méthode hongroise en $O(k^3)$ opérations. La construction de la matrice

de similarité est en $O(n+k^2)$, bien qu'il existe un calcul en temps linéaire [112].

2.3.2.2 Transfert d'un ensemble d'éléments

La première situation décrit des opérations de transfert élément par élémént. Une distance analogue plus proche de notre situation devrait pouvoir prendre en compte l'opération de transfert général qui déplace plusieurs éléments d'une classe à une autre (revoir la figure 2.20 page 64).

Définition 13. *La **distance de transfert** entre $\mathcal{P}$ et $\mathcal{Q}$ sera notée $t_2(\mathcal{P},\mathcal{Q})$ et correspond au nombre minimum d'opérations de transfert et de division nécessaires pour passer de $\mathcal{P}$ à $\mathcal{Q}$.*

Une solution au problème précédent n'apporte pas nécessairement une solution à celui-ci. En effet, en reprenant les permutations $\mathcal{P}$ et $\mathcal{Q}$ de l'exemple 2.1 page 67, on s'aperçoit que seulement deux opérations de transfert d'ensembles sont nécessaires pour passer de la partition $\mathcal{P}$ à la partition $\mathcal{Q}$, au lieu de trois, et qu'elles ne portent pas sur les mêmes éléments :

$$\mathcal{P} = \boxed{1}\boxed{2345}\boxed{678} \overset{234}{\longmapsto} \boxed{1234}\boxed{5}\boxed{678} \overset{6}{\longmapsto} \boxed{12346}\boxed{5}\boxed{78} = \mathcal{Q}$$

Le calcul de $t_2(\mathcal{P},\mathcal{Q})$ peut également se ramener à un problème de couplage parfait, mais il existe un moyen plus direct via le calcul de l'infimum de $\mathcal{P}$ et $\mathcal{Q}$. La formule suivante [8] se vérifie :

$$t_2(\mathcal{P},\mathcal{Q}) = |\mathcal{P}\cap\mathcal{Q}| - \min(|\mathcal{P}|,|\mathcal{Q}|) \tag{2.4}$$

En effet, si l'on considère une chaîne de transferts et de divisions, les opérations portent sur les intersections de classes de $\mathcal{P}$ et $\mathcal{Q}$. De même, les éléments non déplacés par la chaîne sont parmi ces intersections. Au final, $t_2(\mathcal{P},\mathcal{Q})$ peut s'interpréter comme la plus petite des deux valeurs : le nombre minimal d'agrégations pour passer de $\mathcal{P}\cap\mathcal{Q}$ à $\mathcal{Q}$ ou le nombre minimal de divisions pour passer de $\mathcal{P}$ à $\mathcal{P}\cap\mathcal{Q}$.

La distance $t_2(\mathcal{P},\mathcal{Q})$ sera évidemment plus petite que $t_1(\mathcal{P},\mathcal{Q})$ et présentera l'avantage d'être calculable en $O(n)$ opérations.

2.3.2.3 Sur les alignements d'une biphrase

Plusieurs adaptations seront nécessaires pour définir et calculer des distances analogues sur l'espace des alignements $\mathscr{A}(n,m)$. Notamment parce que deux alignements n'auront pas forcément la même couverture et qu'il faudra par conséquent considérer les opérations de *suppression* et d'*insertion* en plus des opérations de *transfert* :

Définition 14. *Soient $\ell_1,\ell_2 \in \mathscr{A}(n,m)$. Nous définissons les deux distances suivantes :*

- *La distance d'édition unitaire $t_1(\ell_1,\ell_2)$ correspond au nombre minimum d'opérations **unitaires** parmi la suppression, l'insertion, le **transfert étendu**, à appliquer pour transformer ℓ_1 en ℓ_2.*
- *La distance d'édition $t_2(\ell_1,\ell_2)$ correspond au nombre minimum d'opérations à appliquer parmi la suppression, l'insertion, la **division** et le **transfert**, à appliquer pour transformer ℓ_1 en ℓ_2*

Remarque 3. *La distance d'édition unitaire se base sur des déplacements individuels de mot sauf dans un cas particulier, celui de la suppression lorsque le groupe lié est constitué d'un mot source et d'un mot cible. Dans ce cas, on considèrera que la suppression d'un des mot supprime l'autre et la distance unitaire comptera une unique transformation. On parlera de **suppression unitaire** (voir figure 2.25.b page 72).*

Pour que ces deux distances soient bien définies, il faut préciser qu'il est toujours possible de transformer un alignement en un autre par une succession des opérations élémentaires concernées. t_1 et t_2 sont bien des distances sur $\mathscr{A}(n,m)$, c'est-à-dire qu'elle sont *définies positives*, *symétriques* et vérifient l'*inégalité triangulaire*. Nous pouvons d'ores et déjà affirmer que l'ensemble des alignements est métrique.

La distance d'édition unitaire compte une chaîne d'opérations ne déplaçant qu'un mot à la fois. À partir de deux alignements ℓ_1 et $\ell_2 \in \mathscr{A}(n,m)$, nous construisons deux

partitions analogues $\mathcal{P}_1$ et $\mathcal{P}_2$ sur $X = [\![1, n+m]\!]$. Nous ramènerons ainsi le calcul de $\mathcal{t}_1(\ell_1, \ell_2)$ a celui de $t_1(\mathcal{P}_1, \mathcal{P}_2)$. Dans cette analogie alignement-partition (réutilisée tout le long de cette partie), les mots source sont représentés par les entiers de 1 à n et les mots cible par les entiers de $n+1$ à $n+m$. Les partitions regroupent en classes les mots liés (indifféremment source ou cible). Ça ne suffit pas à construire la partition analogue car il faut aussi prendre en compte les alignements partiels. Pour chaque mot non lié, on ajoute une classe singleton contenant l'entier correspondant. Cette analogie (représentée à la figure 2.23) fonctionne bien car il faut au minimum deux nœuds pour former un lien.

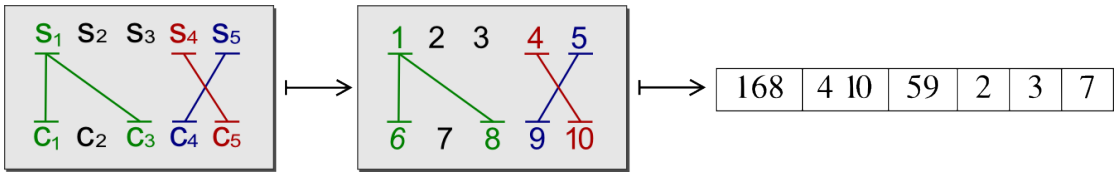


Figure 2.23 – Analogie alignement-partition pour le calcul de la distance unitaire $\mathcal{t}_1$

La valeur $t_1(\mathcal{P}_1, \mathcal{P}_2)$ donne la longueur minimum d'une d'une chaîne d'opérations unitaires transformant $\mathcal{P}_1$ et $\mathcal{P}_2$. Pour les alignements, les opérations élémentaires analogues seront les suivantes :

-Une division créant un singleton correspond à une suppression sur l'alignement

-Un transfert vers/depuis un singleton correspond à une suppression/insertion

-Un transfert depuis une paire correspond à un alignement étendu unitaire

-Dans tout autre cas, un transfert est analogue à un transfert

Mais les opérations sur les partitions sont plus permissives que celles entre alignements car un transfert ou une division peut créer un groupe de mots source liés sans aucun mot cible en face (voir figure 2.24).

Cela pourrait constituer un problème, mais il n'en est rien car dans une chaîne minimale de transferts/divisions unitaires, les éléments déplacés ne le sont qu'une seule fois et l'on peut donc proposer une chaîne produisant le même résultat en déplaçant les éléments dans un ordre arbitraire. On peut alors choisir un ordre qui rend toutes les opérations licites pour l'alignement en conservant un mot source et un mot cible par classe

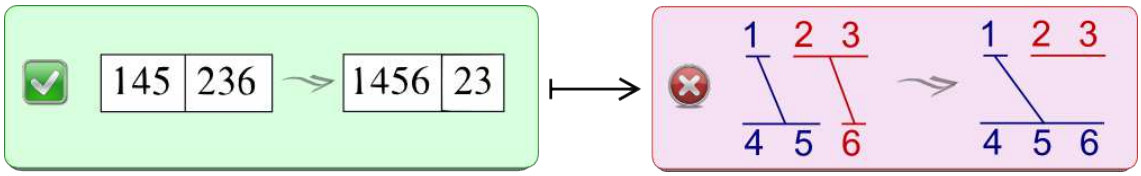


Figure 2.24 – Transférer 6 prématurément produit une opération illicite

jusqu'au dernier moment. Les transfert ou division unitaires sur un ensemble à deux éléments sont analogues à un *transfert étendu unitaire* ou à une *suppression unitaire* introduite par la remarque 3 (voir figure 2.25).

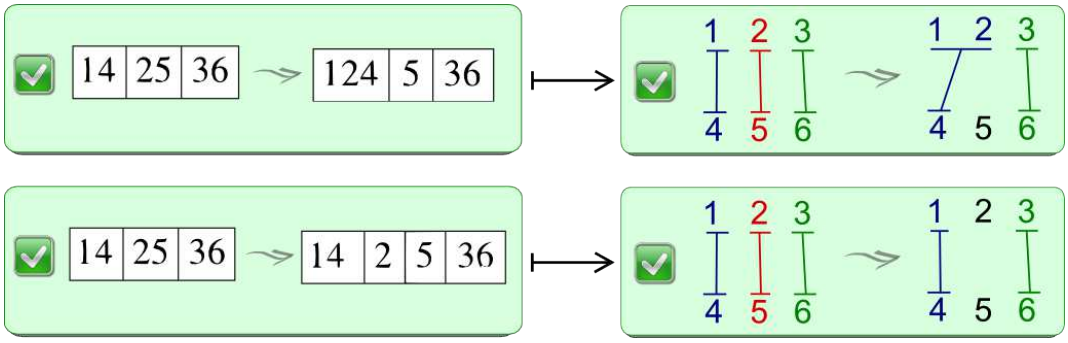


Figure 2.25 – Le transfert étendu unitaire et la suppression unitaire sont licites

Finalement, en construisant les partitions analogues $\mathcal{P}_1$ et $\mathcal{P}_2$, on a :

$$\mathcal{t}_1(\ell_1, \ell_2) = t_1(\mathcal{P}_1, \mathcal{P}_2) \tag{2.5}$$

La *distance d'édition unitaire* $\mathcal{t}_1$ est donc calculable en $O(k^3 + n + m)$ (k étant le nombre de groupes liés) et est naturelle dans le sens où, nous le verrons en partie 3.1.2, elle est liée à l'effort nécessaire à un annotateur pour post-éditer un alignement via une interface d'acquisition de type Blinker.

La distance d'édition $\mathcal{t}_2$ compte une chaîne d'opérations pouvant déplacer plusieurs mots entre deux groupes suivant les cinq transformations élémentaires sauf le *transfert étendu*. Nous utilisons une analogie légèrement différente pour nous ramener, par analo-

gie, à un calcul de *distance de transfert* t_2 entre partitions.

À partir d'un alignement ℓ de $\mathscr{A}(n,m)$, nous construisons une partition analogue $\mathcal{Q}$ de $[\![1, n+m+1]\!]$ de la manière suivante. Les mots source sont représentés par les n premiers entiers et les mots cible par les m derniers. La différence notable intervient sur les mots non liés. Ici, nous avons besoin de prendre en compte des suppressions et des insertions générales. Nous ajoutons donc une seule classe contenant **tous** les mots non liés, indifféremment source ou cible (la "*classe des non liés*" $P_{\emptyset}$). Cette classe, pour être non vide comptera toujours l'entier $n+m+1$ ajouté artificiellement et ne correspondant à aucun mot (il n'affecte en rien la distance t_2 car il n'y aura jamais besoin de le déplacer). La figure 2.26 représente l'analogie sur un exemple.

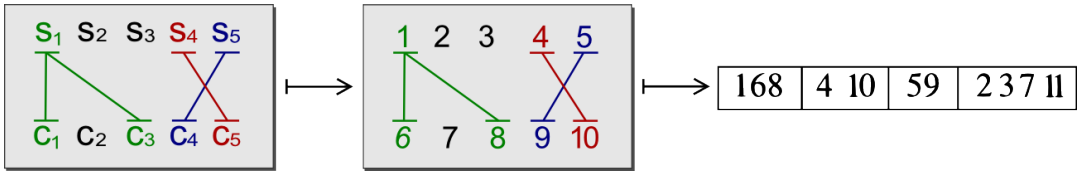


Figure 2.26 – Analogie alignement-partition pour le calcul de la distance t_2

La valeur $t_2(\mathcal{Q}_1, \mathcal{Q}_2)$ donne la longueur minimum d'une chaîne de certaines opérations élémentaires transformant $\mathcal{Q}_1$ en $\mathcal{Q}_2$. Les opérations analogues sur les alignements sont les suivantes :

-Un transfert vers/depuis la classe $P_{\emptyset}$ correspond à une suppression/insertion

-Une division sur $P_{\emptyset}$ correspond à une insertion

-Sinon, un transfert est analogue à un transfert et une division à une division

Remarque 4. *On pourrait penser que la division de la classe des non liés laisse une ambiguïté sur la "nouvelle" classe des non liés. En fait, la classe contenant le mot vide* $n+m+1$ *correspond à la nouvelle classe.*

On voit que la *classe des non liés* joue un rôle central pour représenter les suppressions et insertions. Mais pour fonctionner, l'analogie ne doit admettre qu'une seule classe de mots non liés contrairement à l'analogie précédente qui utilisait les singletons. Ainsi,

les transferts comptés par la distance $\mathcal{L}_2$ n'incluent pas le *transfert étendu* qui, ici, est à distance 2.

De même que pour la *distance unitaire*, il existe des transformations sur les partitions non interprétables dans notre analogie. Une division séparant les éléments source et cible d'un groupe ne se produira pas dans une chaîne minimisant le nombre de transformations car ne peut pas être optimale.

Par contre, un transfert laissant une classe uniquement composée de mots source ou cible pourra se produire et jouer un rôle dans une chaîne minimisante sur les partitions alors qu'elle n'est pas interprétable dans l'analogie (voir figure 2.27) :

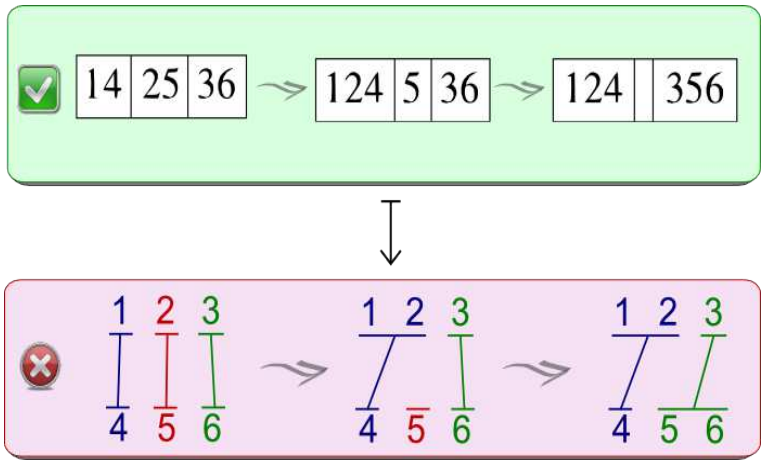


Figure 2.27 – Une chaîne minimale illicite permise par l'analogie

Ce problème peut en fait être contourné car il n'existe pas une unique chaîne de longueur minimum réalisant la distance de transfert t_2. Il est possible de modifier une de nos chaînes de transferts/divisions en conservant un analogue licite pour les alignements. En effet, un transfert laissant derrière lui un groupe de mots source (ou cible) sera toujours suivi d'un ou plusieurs transferts rétablissant la situation car le résultat $\mathcal{P}_2$ est bien formé. On peut donc modifier cette succession d'agrégations de manière à transporter "*en passagers clandestins*" les éléments source (ou cible) vers leur destination finale, sans modifier le nombre d'opérations (voir figure 2.28 qui corrige l'exemple précédent).

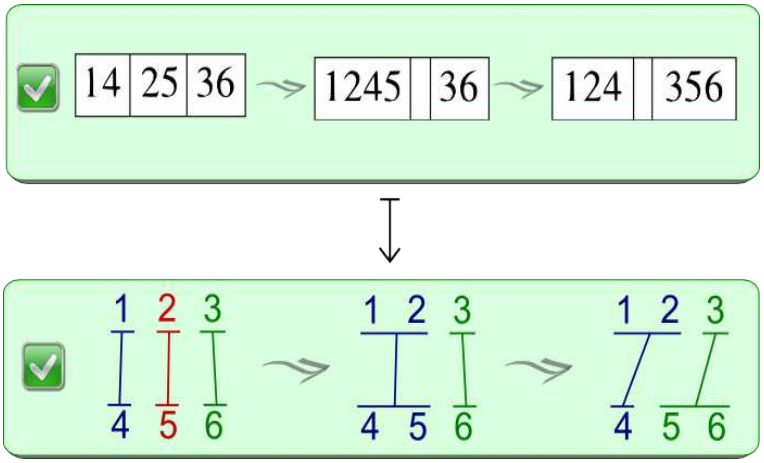


Figure 2.28 – 5 qui doit être transféré ailleurs accompagne provisoirement 2

Finalement, en construisant les partitions analogues $\mathcal{Q}_1$ et $\mathcal{Q}_2$, on a :

$$t_2(\ell_1, \ell_2) = t_2(\mathcal{Q}_1, \mathcal{Q}_2) \tag{2.6}$$

La *distance d'édition* t_2 est calculable en $O(n+m)$. Elle quantifiera mieux que t_1, l'effort nécessaire à un annotateur pour post-éditer un alignement via l'interface de l'outil présenté en section 3.1.2.

Remarque 5. *Les transformations comptées par t_2 ont le même poids, qu'il s'agisse de la suppression d'un simple mot ou d'un long groupe. La distance d'édition unitaire t_1 pourra être préférée si l'on souhaite en tenir compte. La distance d'édition t_2 sera plutôt représentative d'un effort de post-édition et présentera des amplitudes moindres.*

2.3.3 La distance des divisions

Nous adaptons une autre distance entre partitions, basée sur les opérations de division et d'agrégation, la *distance des divisions*.

2.3.3.1 Sur les partitions d'un ensemble

D'après [20], cette distance aurait été utilisée par David Pavy en 1968 pour comparer des comportements schizophènes à des comportements normaux. Sa définition est similaire aux précédentes :

Définition 15. *La distance des divisions entre deux partitions $\mathcal{P}$ et $\mathcal{Q}$, $d(\mathcal{P},\mathcal{Q})$ est le nombre **minimum** d'opérations de division et d'agrégation sur $\mathcal{P}$ nécessaires pour obtenir $\mathcal{Q}$.*

Puisque une opération de transfert peut être vue comme une opération de division suivie d'une opération d'agrégation, on a la majoration $\dot{(}\mathcal{P},\mathcal{Q}) \leq 2 \cdot t_2(\mathcal{P},\mathcal{Q})$. La distance des divisions traduit une transformation de $\mathcal{P}$ en $\mathcal{Q}$ plus décomposée. Elle peut être exprimée simplement par la formule suivante [8] :

$$d(\mathcal{P},\mathcal{Q}) = |\mathcal{P}| + |\mathcal{Q}| - 2 \cdot |\mathcal{P} \cup \mathcal{Q}|$$

Où '$\cup$' représente la borne supérieure de deux partitions pour la relation d'ordre '$\prec$'. Elle se calcule donc en temps linéaire, ce qui la rend appréciable.

2.3.3.2 Sur les alignements d'une biphrase

Nous proposons une distance analogue sur les alignements qui différencie trois niveaux de désaccords. Tout d'abord deux alignements ℓ_1 et ℓ_2 dont les groupes du premier incluent ou sont incluent dans ceux du deuxième seront considérés comme relativement proches car on peut transformer directement l'un en l'autre par des opérations de division ou d'agrégation. Ensuite, deux alignements ayant les mêmes supports mais dont les groupes ont des frontières différentes seront considérés comme un peu plus différents. Pour transformer l'un en l'autre, le plus direct serait d'utiliser des opérations de transfert, mais une pénalité est introduite par le fait qu'il faudra deux opérations ici car nous ne considérons que les agrégations et les divisions. Enfin, des alignements qui n'ont pas les mêmes supports source et cible peuvent être considérés comme éloignés

car il faut insérer et supprimer des mots dans le processus de transformation. Pour pénaliser ce type de différence, nous ne considérerons que des insertions et des suppressions unitaires. Nous appellerons cette distance, la *distance des divisions* :

Définition 16. *Soient ℓ_1 et $\ell_2 \in \mathscr{A}(n,m)$. On définit la distance des divisions $d(\ell_1,\ell_2)$ comme le nombre minimum de divisions, d'agrégations, d'insertions unitaires et de suppressions unitaires à appliquer sur ℓ_1 pour obtenir ℓ_2*

Nous adaptons la distance d sur les partitions pour le calcul de d sur $\mathscr{A}(n,m)$ grâce à l'analogie précédemment utilisée pour la distance t_1. À partir d'un alignement $\ell \mathscr{A}(n,m)$, nous produisons une partition $\mathcal{P}$ de $[\![1,n+m]\!]$ dont les classes singletons correspondent à des mots non liés.

Soient ℓ_1, $\ell_2 \in \mathscr{A}(n,m)$ et $\mathcal{P}_1,\mathcal{P}_2$ les partitions analogues. La valeur de $d(\mathcal{P}_1,\mathcal{P}_2)$ donne la valeur minimum d'une chaîne d'opérations de divisions/agrégations pour former $\mathcal{P}_2$ à partir de $\mathcal{P}_1$. Du point de vue des alignements, les opérations analogues sont les suivantes :

-L'agrégation d'un singleton correspond à une insertion unitaire

-Dans tout autre cas, une agrégation reste une agrégation

-Une division séparant un mot correspond à une suppression unitaire

-Dans tout autre cas, une division (licite) reste une division

Ici encore, une opération de division qui crée une nouvelle classe exclusivement composée de mots source (ou cible) n'est pas interprétable dans l'analogie bien que de telles transformations puissent participer à une chaîne minimisante pour transformer $\mathcal{P}_1$ en $\mathcal{P}_2$. Il sera en fait possible de se ramener à une chaîne d'opérations licites pour l'analogie.

En effet, si l'on se donne une chaîne d'opérations constituée de p divisions et de q agrégations sur une partition, il est toujours possible de construire une chaîne produisant le même résultat en effectuant d'abord q agrégations puis p divisions. Cela est dû au fait que pour une division δ et une agrégation α, il est possible de construire δ' et α' tels que $\delta \circ \alpha = \delta' \circ \alpha'$ (on peur alors faire "commuter" les opérations).

La chaîne ainsi réordonnée est alors interprétable pour les alignements comme une suite d'insertions et d'agrégations produisant $\ell_1 \vee \ell_2$ puis une série de suppressions et de divisions produisant ℓ_2. Une opération de division illicite ne peut intervenir dans une chaîne minimisante ordonnée de la sorte car nécessiterait d'être suivie par une agrégation pour produire le résultat souhaité (qui est bien formé).

Ainsi, la *distance des divisions* entre deux alignements se déduit de son analogue entre deux partitions par la formule :

$$d(\ell_1, \ell_2) = d(\mathcal{P}_1, \mathcal{P}_2) \tag{2.7}$$

Lorsque l'on considère deux alignements ℓ_1 et ℓ_2 ayant les mêmes supports source et cible, la distance pourra se calculer directement :

$$d(\ell_1, \ell_2) = |\ell_1| + |\ell_2| - 2 \cdot |\ell_1 \vee \ell_2| \tag{2.8}$$

où $|\ell|$ est le nombre de groupes liés, c'est-à-dire $|\mathcal{V}| = |\mathcal{W}|$. Cette distance, plus décomposée que la distance d'édition t_2 présentera une plus grande amplitude. Elle aura aussi l'avantage d'être calculable en $O(n+m)$ opérations.

CHAPITRE 3

ALIGN^{IT} : UNE APPROCHE COLLABORATIVE ET DES OUTILS AUTOMATIQUES

Ce chapitre propose un cadre pratique pour créer une ressource multilingue de qualité alignée à une granularité sous-phrastique. Les éléments théoriques proposés précédemment (section 2.2.2) nous permettent d'envisager une solution amorcée par une collaboration en ligne d'annotateurs non-experts (section 3.1). Nous insisterons sur la nécessité de voir une telle approche soutenue par des outils automatiques évolutifs. S'ensuivra une présentation concrète de l'outil nommé Align^{It} [1] en l'état actuel de développement (section 3.1.2). Enfin, la partie 3.2 s'intéressera à la nécessité d'une interaction mutuellement bénéfique entre des annotateurs et des outils automatiques restant à définir pour converger vers une architecture proche de celle des traducteurs à base d'exemples.

3.1 Mise en place d'un outil d'alignement collaboratif

3.1.1 Motivations et orientations envisagées

Des ressources parallèles apparaissent chaque jour, disponibles librement et dans de nombreuses langues. Elles sont d'une grande utilité pour tout travail dans le thème de la traduction automatique. Elles servent notamment à entraîner des modèles de langues pour des machines de traduction statistique et permettent l'extraction terminologique bilingue [57] [141] [156] [45]. En comparaison, des ressources de qualité alignées à un niveau sous-phrastique, telles que le *corpus Blinker*, sont assez rares. Une raison principale à cet état de fait est la relativement bonne qualité des *aligneurs de phrases* (section 1.1.3.1) par rapport aux différents *systèmes d'alignements sous-phrastiques* existants. Construire de telles ressources pourrait être d'une utilité cruciale pour les machines de traduction, mais aussi pour créer des ressources alignées de référence qui pourront par exemple servir à l'évaluation d'autres systèmes d'alignements sous-phrastique [82] ou à

[1] accessible en ligne à l'adresse `www.alignit.fr`

constituer des tables bilingues pour une machine de traduction.

Est-il envisageable alors de s'investir dans la construction d'un nouveau corpus d'alignements manuels ? Comme nous l'avons vu en section 1.3, la construction du corpus Blinker s'est avérée coûteuse en temps et en argent (les experts étaient rémunérés). Les moyens matériels et les efforts mis en place pourraient décourager toute volonté de construire une ressource alignée de qualité. Nous pensons pourtant que cela serait envisageable à moindre coût en tirant profit de trois vecteurs que l'on expose ici : le concours de *non-experts*, une architecture permettant un *travail collaboratif* et le développement d'*outils automatiques évolutifs*.

3.1.1.1 Contribution de non-experts et qualité

Dans l'expérience menée par Dan Melamed, les experts sont sélectionnés et doivent répondre à certains critères fixés par un questionnaire évaluant entre autres leur niveau de langue et la vitesse à laquelle ils pensent pouvoir honorer leur engagement. La sélection est inévitable pour une expérience rémunérée et l'objectif n'était pas tant de produire une ressource que de proposer une méthodologie et de l'évaluer selon des critères d'accord inter-annotateurs. Nous souhaitons envisager une approche similaire à laquelle pourraient participer des utilisateurs non-experts et bénévoles sans que la ressource créée ne subisse une détérioration rédhibitoire. Cela nous semble envisageable pour différentes raisons. Tous d'abord, l'étude des alignements du corpus Blinker a prouvé que les alignements produits par les différents experts étaient très proches malgré la grande divergence des corpus [2] et des profils d'annotateurs différents. Ce qui a été confirmé par la suite dans des projets similaires [99] [90]. Ensuite, nous constatons que l'alignement sous-phrastique demande des connaissances moins importantes que la traduction car il n'est pas demandé de générer mais seulement de reconnaître. Enfin, nous supposons que si l'une des deux langues est parlée couramment par l'annotateur, une connaissance limitée de la seconde est suffisante pour être capable de produire des alignements corrects.

Pour réduire le biais expérimental et augmenter la qualité des alignements, les parti-

[2] les deux versions n'étaient pas en traduction directe mais provenaient d'une langue tierce, probablement le latin

cipants au projet Blinker devaient se plier à une liste de principes directeurs décrits dans le guide [96]. Celui-ci expose différentes stratégies et situations ainsi que des exemples d'alignements que l'on peut considérer comme bons ou mauvais (voir figure 3.1). Dans le cadre d'un travail bénévole, il semble difficile d'imposer la contrainte d'un long cahier des charges à ses contributeurs. Cela ne signifie pas que l'outil créé ne soit pas accompagné d'un guide, mais nous avons remarqué qu'en général il faisait au mieux l'objet d'une lecture en diagonale. Nous avons donc créé un guide complet ainsi qu'une brève introduction par quelques exemples afin de suggérer l'idée contenue dans le guide du Blinker selon laquelle un bon alignement devra être à la fois **précis** et **couvrant**, potentiellement *groupant*, *croisant* et *non contigu*.

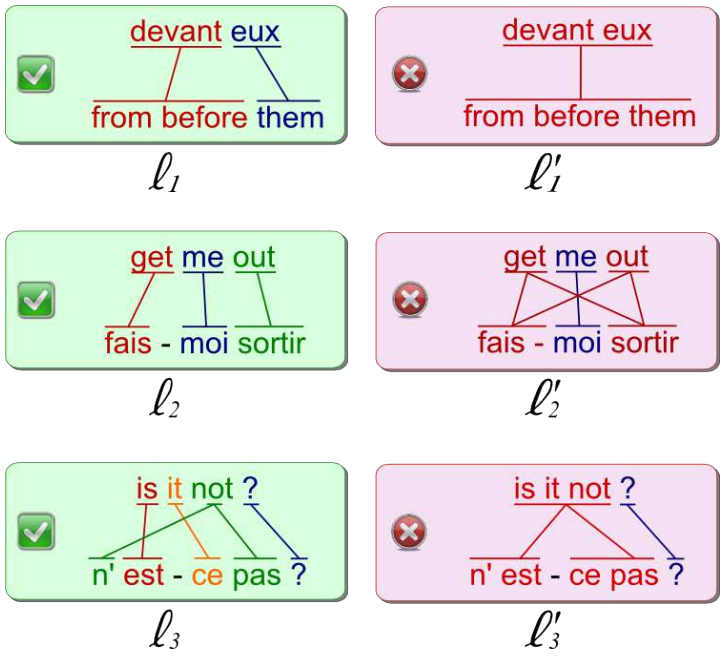


Figure 3.1 – Alignements bons ou mauvais selon le guide d'annotation de Blinker

Nous souhaitons nuancer cette idée selon laquelle il y aurait une bonne manière de mettre en correspondance des éléments en relation de traduction. Certes on peut proposer des directions concernant certaines situations afin de proposer une forme canonique, mais le plus souvent il y a *matière à discussion* plutôt qu'*incompatibilité* (nous précisons plus loin ce que nous entendons par là). Nous pouvons observer, à la figure 3.1,

des exemples de "bons" alignements selon le guide d'annotation du Blinker. Prenons l'exemple de l'alignement ℓ_2 dans lequel *"get"* est lié avec *"fais"*. Ces deux mots partagent la particularité d'être le seul verbe dans leur fragment, mais ici, le verbe anglais est en fait le verbe à particule *"get out"* signifiant *"sortir"*. Une proposition d'alignement différente aurait pu être de lier le fragment non contigu *"get...out"* avec *"sortir"* seulement. Pourtant on sent que le verbe *"fais"* participe à la correspondance avec le verbe anglais car non seulement il contribue à rendre la traduction naturelle en français, mais il porte également la marque de l'impératif portée par *"get out"*. On peut donc proposer de lier *"get...out"* avec *"fais...sortir"*, ainsi l'alignement ℓ'_2 semble être un alignement correct alors que ℓ_2 est discutable. En fait dans ces trois exemples, les liens sont différents, mais pas fondamentalement incompatibles. On remarque que les paires de mots liées par ℓ_1, ℓ_2 et ℓ_3 le sont également dans ℓ'_1, ℓ'_2 et ℓ'_3. Dans la partie théorique 2.2.2.2, nous nous attachions à définir une structure formelle pour l'ensemble des alignements ainsi qu'une relation d'ordre partiel, la *finesse* notée "$\prec$" et la notion de *support*. Les alignements de la figure 3.1 vérifient les deux propriétés suivantes :

$$\begin{cases} \ell_i \prec \ell'_i \\ \ell_i \text{ et } \ell'_i \text{ ont les mêmes supports source et cible} \end{cases} \quad \text{pour } i \in \{1,2,3\}$$

À ce titre, il serait plus correct de considérer que les alignements "*corrects*" et "*incorrects*" de la figure expriment des informations compatibles de différentes granularités, entre le niveau du mot et celui de la phrase. Nous évoquions déjà cette approche par niveaux au chapitre précédent avec les exemples[3] 2.2 et 2.3 page 37. Le premier proposait de lier des unités morphologiques plus fines que le mot, le deuxième s'adaptait pour adopter le mot comme unité minimale.

Dans une situation de disparition d'article ou de préposition ou encore de dédoublement à cause, par exemple, d'une énumération comme dans notre exemple 2.4 page 38, deux possibilités s'offrent à nous : lier les éléments répétés ou les considérer comme non liés. Le guide d'annotation suggère que la bonne manière est de produire l'alignement le

[3]La biphrase en anglais/espagnol était *"Atravesó el rio flotando" / "It floated across the river"*

plus couvrant possible en respectant les dépendances (voir figure 3.2).

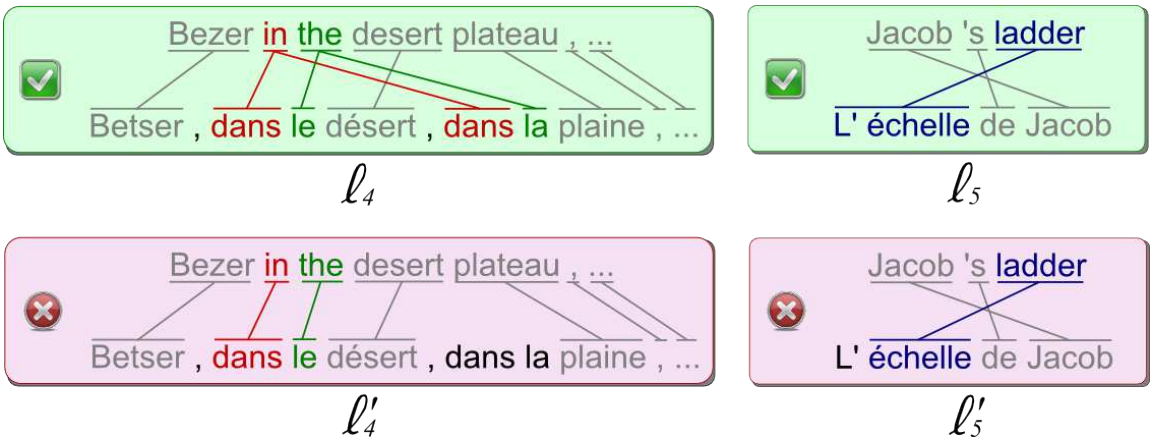


Figure 3.2 – Dans le cas des omissions, le guide préconise un alignement couvrant

Bien que ce point de vue permette une meilleure détection des frontières par les outils automatiques proposés ici (décrits en partie 5.2), la position contraire m'a été exposée et défendue par certains. Il en ressort que les deux opinions découlent d'un point de vue asymétrique inspiré du processus de traduction : hésitant, l'annotateur choisit une des deux langues et se demande de quelle manière il aurait procédé pour obtenir la traduction proposée. Selon son choix, il peut avoir le sentiment que des mots disparaissent lors de la traduction, ou au contraire qu'ils apparaissent[4]. Ces deux points de vue semblent difficiles à départager. Ici encore, en se référant à la relation finesse, il n'y a pas une incompatibilité entre les alignements alternatifs :

$$\begin{cases} \ell_i \prec \ell'_i \\ \text{Les supports souce et cible de } \ell_i \text{ sont inclus dans ceux de } \ell'_i \end{cases} \quad \text{pour } i \in \{4,5\}$$

La relation de finesse s'est avérée un outil d'observation intéressant pour les cinq exemples décrits par les figures 3.1 et 3.2. On a pu voir que si un alignement ℓ est considéré comme bon, un alignement ℓ' plus grossier (i.e. $\ell \prec \ell'$) ne présentera pas d'erreur d'alignement. Au contraire, le risque de réelles erreurs dans un alignement aurait plutôt

[4]L'alignement ℓ_4 présente l'avantage de couvrir l'accord en genre de l'article défini.

tendance à se trouver vers des granularités plus fines lorsque des groupes cohérents sont séparés à tort.

Des alignements sensiblement différents entre annotateurs sont inévitables, experts ou non-experts. Pourtant, comme nous l'avons vu, ce n'est pas forcément un problème fondamental si ces alignements forment une chaîne compatible. La relation de finesse "$\prec$" nous donne ainsi un moyen de regrouper des informations concordantes, mais également de pointer automatiquement des phénomènes complexes ou des erreurs en cas d'incompatibilité : on pourra parler de véritable désaccord dans ces cas.

Un exemple de désaccord plus fondamental est donné à la figure 3.3 où l'alignement considéré comme mauvais par le guide était en général préféré à l'autre (pourtant considéré comme bon par le guide du Blinker) par les annotateurs à qui l'on a proposé d'aligner la biphrase via notre interface graphique.

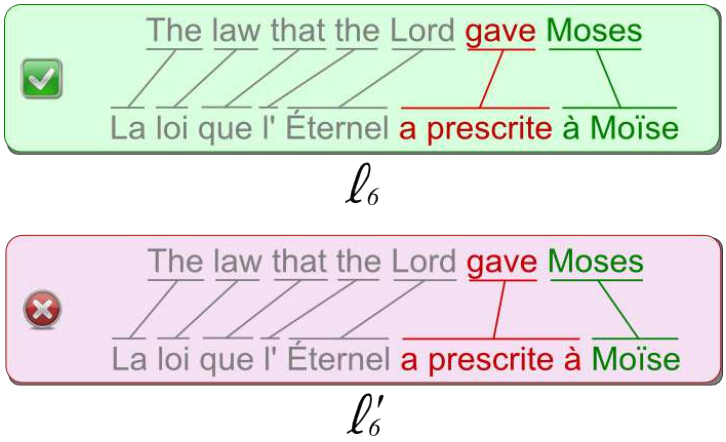


Figure 3.3 – L'alignement proposé peut être incompatible avec les préférences du guide de Blinker

Le guide de Blinker préfère lier la préposition disparaissant "*à*" à l'objet "*Moïse*". Pourtant nos annotateurs auront préféré laisser la préposition non liée dans le doute, ou la lier au groupe verbal (alignement ℓ'_6). Et il est clair que dans une table d'équivalents de traduction, la correspondance "*gave*" / "*a prescrite à*" sera plus générique et composable que la correspondance "*Moses*" / "*à Moïse*".

Cette fois, les deux alignements ℓ_6 et ℓ'_6 sont en désaccord :

$$\begin{cases} \ell_6 \not\prec \ell'_6 \text{ et } \ell_6 \not\succ \ell'_6 \\ \ell_i \text{ et } \ell'_i \text{ ont les mêmes supports source et cible} \end{cases}$$

Ces deux alignements seront pourtant très proches pour n'importe quelle mesure d'accord. Par exemple, pour les distances introduites en partie 2.3, on a :

$$\begin{cases} t_1(\ell_6, \ell'_6) &= 1 \\ t_2(\ell_6, \ell'_6) &= 1 \\ d(\ell_6, \ell'_6) &= 2 \end{cases}$$

Une solution consensuelle consiste à régler ce type d'incompatibilité par l'opération d'élargissement "$\vee$" (voir figure 3.4).

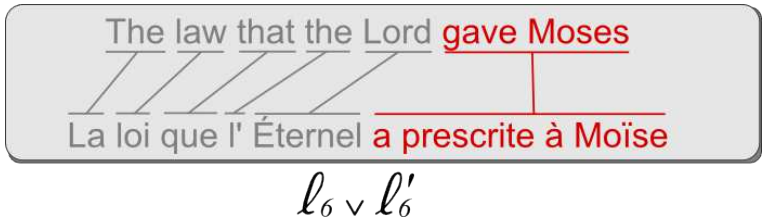


Figure 3.4 – Un compromis est d'élargir les deux alignements

En proposant une structure de treillis pour représenter l'ensemble des alignements, on voit qu'on est alors capable d'aborder les désaccords inter-annotateurs d'une manière naturelle. Cela rend possible de détecter des désaccords structurellement importants de manière plus fine que par le biais de simples mesures d'accords, mais également de les traiter par des opérations adaptées. L'exemple des alignements ℓ_6 et ℓ'_6 illustre assez bien cette idée selon laquelle deux alignements avec une forte forte similarité mais incompatibles pour "$\prec$" seront plus symptomatiques que deux alignements comparables avec une distance éventuellement plus grande.

3.1.1.2 Un "travail" collaboratif

Le magazine *Time* choisit de reconnaître "Vous" comme personnalité de l'année 2006, en soulignant l'importance de *la communauté informatique qui collabore [en ligne] comme jamais auparavant* pour créer des objets nouveaux tels que *Wikipedia*, *Youtube* ou *MySpace*. L'outil développé durant cette thèse est une plate-forme collaborative, aidant à la construction d'une ressource alignée de qualité dans plusieurs langues, contrôlable et modifiable par des annotateurs non-experts. Ceux-ci peuvent, via une interface web, parcourir phrase après phrase des corpus multilingues proposés et les aligner (à un niveau sous-phrastique) à la souris. Pour parler de véritable collaboration, il faut favoriser l'interopérabilité. Nous avons choisi une approche totalement transparente dans le sens où les alignements de chacun sont visibles par tous. Un enregistrement de l'utilisateur est tout de même nécessaire afin d'associer chaque alignement à son annotateur. Ainsi, il est possible de contribuer en créant un alignement à partir d'une biphrase vierge de tout lien ou de se servir de l'alignement d'un autre comme base de départ. Dans les deux cas, sauvegarder l'alignement se fait sur son espace de travail personnel, on ne pourra jamais modifier celui d'un autre et inversement. Les informations sont stockées sur le serveur hébergeant le site et seuls quelques calculs nécessaires à l'affichage des liens sont gérés côté client. L'approche web permet un accés immédiat et ne demande aucune installation. Finalement, la ressource créée sera téléchargeable en l'état à tout instant, directement sur le site sous un format XML[5].

L'outil Align^{It} est donc un prototype permettant l'alignement collaboratif via un site web. Les approches collaboratives nécessitent généralement la valorisation de l'outil dans un modèle économique et/ou social propre à entretenir l'intervention de contributeurs. Dans le cadre de cette thèse nous nous sommes limités à étudier l'outil en soi, ses prérequis théoriques et techniques ainsi que d'explorer des possibilités d'automatisation. Nous parlons donc de prototype dans le sens où l'outil n'est qu'une vitrine (bien que la ressource créée soit bien réelle) et pourrait être au cœur d'une forme finalisée dans le type de différentes plates-formes collaboratives (une transformation nécessitant

[5] pas dans la version actuelle, mais c'est prévu

quelques apports, mais aucune modification fondamentale).

On peut donner l'exemple d'approches dites de *sciences participatives*[6] utiles pour la récolte de données scientifiques. Il peut y avoir un travail de communication opéré en amont au travers des média classiques (journaux, radios, TVs internet) et/ou des compensations financières. Les approches les plus connues sont en biologie/médecine pour des études sur la santé ou encore le test de médicaments ou de produits cosmétiques. On peut aussi trouver des exemples plus originaux comme celui du projet *sud4science* [110] qui pour former un corpus de plus de 90000 SMS a ouvert une vaste campagne de communication autour de la récolte de SMS, récompenses à la clé. Le projet Blinker est également assimilable aux sciences participatives et pour lesquelles les annotateurs sont rémunérés. Mais de nombreux projets aux budgets limités recherchent aussi l'intervention de bénévoles en imaginant d'autres motivations. C'est le cas des approches de type **jeu avec un but**[7]. Nous pouvons citer, dans le domaine du TAL, le *Phrase Detective*[8] de Jon Chamberlain [32] qui propose, au travers de mini-jeux, de créer une ressource annotée sémantiquement. Il s'agit en général de repérer dans un corpus, des co-références, ce qui est appelé "*trouver le coupable*". Le *JeuxDeMots*[9] de Mathieu Lafourcade [85] est un autre exemple de *GWAP* pour lequel les joueurs contribuent à la construction d'un réseau lexical. Principalement, les jeux proposent de lister des mots ou expressions qui partagent une relation avec un mot de départ (relation dont la nature peut être entre autres la synonymie, l'antonymie, l'hyperonymie ou plus simplement une idée associée). La motivation qui est ici le jeu, peut revêtir d'autres formes comme celle d'outil pédagogique : nous citons à ce titre *PtitClic-Kids* [162][10] proposé à des enseignants pour travailler sur le vocabulaire avec leurs élèves. Puis nous pouvons évoquer simplement les outils non ambivalents de *travail collaboratif* pour lesquels l'objet produit est le but premier des contributeurs. L'exemple le plus proche de notre problématique est celui des outils de *Traduction Assistée par Ordinateur (TAO)*[11] englobant une dimension collabo-

[6] On parle aussi de *science citoyenne* ou de *science collaborative*
[7] On parle plus souvent de *Game With A Purpose* ou *GWAP*
[8] accessible sur `http://anawiki.essex.ac.uk/phrasedetectives/`
[9] accessible sur `www.jeuxdemots.org`
[10] PtitClic est à la base un jeu basé sur le réseau JeuxDeMots visant à le renforcer
[11] *CAT tools* en anglais pour *Computer Aided Translation*

rative tel que TRANSBey [17].

Ainsi, la frontière devient parfois mince entre un annotateur, un joueur et un élève ou entre un superviseur, un joueur gradé et un enseignant. Mais qu'il s'agisse d'une note, d'un score ou d'une mesure de performance les approches précédentes posent le problème de l'évaluation des données. Lorsqu'il existe une hiérarchie des annotateurs, la ressource créée par les *subalternes/joueurs/élèves* peut être validée par les *chefs/joueurs gradés/enseignants*. Dans l'état prototypal actuel de Align$^{\text{It}}$, il n'existe pas de statuts différents pour les annotateurs, ce qui est envisageable, mais on note que la charge de validation manuelle pourrait s'avérer assez importante. On peut alors exploiter avantageusement la structure en treillis théorique de l'espace d'alignement (partie 2.2.2.2). Une validation partielle automatique est possible en comparant les alignements entre eux par la relation de finesse "$\prec$". En effet, considérons un alignement ℓ validé par un *annotateur gradé* et un alignement ℓ' moins fin ($\ell \prec \ell'$) : l'alignement ℓ' peut être considéré comme structurellement correct avec un score inférieur à celui de ℓ décroissant par rapport à la distance des divisions $d(\ell, \ell')$.

Dans des approches plus horizontales, l'alternative envisagée est souvent la validation par accord mutuel des utilisateurs. À l'instar des célèbres *"j'aime / j'aime pas"*, l'accord renforce, le désaccord affaibli. Dans le *Phrase Detective*, les *Detective Conference* sont une manière ludique de donner son opinion sur des désaccords observés. De la même manière *JeuxDeMots* donne la possibilité au joueur lésé d'entamer un procès en cas de désaccord. Ces manières analogues de procéder permettent aussi de détecter et de traiter les erreurs. On a déjà observé à ce sujet que l'ensemble des alignements sur une biphrase en particulier était muni d'une structure agréablement adaptée aux désaccords entre annotateurs. Elle permet via l'utilisation de la relation de finesse de détecter et de séparer les types de désaccord en 3 catégories :

1. Les alignements sont compatibles selon "$\prec$" et ont même support

2. Les alignements sont compatibles selon "$\prec$" et ont des supports différents

3. Les alignements sont incompatibles

La troisième catégorie est la plus significative. La relation "$\prec$" permet alors de dégager des tendances, tandis que la structure en treillis permettrait automatiquement de pointer localement les incompatibilités et la distance des divisions de lister par priorité les désaccords des plus divergents aux plus anecdotiques.

Dans Align$^{\text{It}}$, les utilisateurs peuvent occuper plusieurs rôles en créant ou modifiant les alignements d'autres utilisateurs. Cette approche transparente offre une visibilité claire de la ressource en construction. Elle permet dans le cadre d'un travail en collaboration (par groupe ou de manière opportuniste), la vérification et la correction de son état courant. Pourtant, l'effort commun reste substantiel, mais ce serait sans compter sur la participation désintéressée de certains de ses membres : les R-utilisateurs.

3.1.1.3 Un effort accompagné par des outils automatiques

Le projet *Penn TreeBank* [92] a permis la construction à la main d'un corpus anglais de phrases étiquetées et parenthésées (c'est-à-dire munis d'une structure syntaxique). Des outils automatiques ont également été utilisés pour permettre aux annotateurs de travailler en post-édition et il est observé expérimentalement que corriger les étiquettes produites par un outil automatique est au moins deux fois plus rapide que d'étiqueter une phrase vierge. La post-édition constitue une alternative à la construction de ressources de qualité par des annotateurs et une approche purement automatique produisant d'énormes ressources, peu contrôlables.

L'intégration d'outils automatiques dans des approches similaires telles que pour la TAO n'est pas nouvelle. Nous proposons, dans cette approche collaborative, la possibilité de mêler des outils automatiques existants à la communauté des annotateurs. Ils formeraient ce que nous appelons les R-utilisateurs[12]. Les annotateurs peuvent ainsi voir leur charge de travail personnel allégée relativement à la qualité des aligneurs automatiques considérés. En post-éditant les résultats d'un outil automatique existant, il y a bien un gain, mais il ne peut être que d'un facteur constant, du moins pour un outil n'évoluant pas. Pour répondre à cette problématique, nous avons tâché durant cette thèse d'imaginer

[12]En l'état actuel, nous avons travaillé avec *Giza++*, mais ne l'avons pas encore intégré comme utilisateur.

des solutions d'accompagnement par des outils automatiques capables de s'améliorer au fil du temps. Il serait alors possible de parler réellement d'accéleration du temps de traitement.

Cette idée d'un partenariat entre les annotateurs humains et les outils automatiques est au cœur de notre problématique car nous souhaitons voir s'appliquer à la tâche de l'alignement les préceptes du *Translator's Amanuensis* [76], à savoir la progressive émancipation des outils automatiques. Pour cela, nous étudions différentes voies possibles pour des outils dits "*à base d'exemples*" reposant sur une mémoire à laquelle ils se réfèrent pour proposer des alignements inspirés d'une expérience passée. La mémoire que nous souhaitons utiliser est bien sûr celle de l'outil Align$^{\text{It}}$ comprenant les alignements produits par les annotateurs [127]. De cette manière, l'espoir est de voir l'échange entre les annotateurs et l'outil automatique devenir mutuellement bénéfique : le robot s'améliorant permettrait à l'annotateur de produire de plus en plus rapidemment et avec un effort moindre une ressource alignée qui en contrepartie enrichirait la mémoire de nouveaux cas utiles pour le robot. Nous discutons plus en profondeur de cette approche d'*Alignement à Base d'Exemples* [125] dans la partie 3.2 ainsi que des solutions choisies.

La motivation principale de l'intégration d'outils automatiques est bien sûr la diminution de l'effort nécessaire à produire la fameuse ressource mais on peut en tirer d'autres avantages dignes d'être mentionnés ici. Tout d'abord, de la même manière que l'approche collaborative permet la post-édition, nous souhaitons proposer des outils autant capables d'aligner des biphrases neuves que des biphrases pré-alignées, accentuant ainsi le comportement antropomorphe. Cela permet une intégration naturelle d'une approche d'alignement à base d'*ancrages sûrs*. Les liens d'ancrage sont autant de contraintes supplémentaires pour l'outil d'alignement qui peuvent lui éviter de propager des erreurs. On peut citer à ce sujet l'approche caractéristique de S. Ozdowska dont les performances de l'aligneur *AliBi* dépendent de liens de départs à partir desquels se forme itérativement l'alignement selon des règles de propagation. Dans son approche, l'alignement initial est produit selon deux méthodes. La première étant une heuristique basée sur les cognats, ces mots proches entre deux langues et facilement reconnaissables. La deuxième heuristique utilise la symétrisation par intersection d'alignements automa-

tiques issus de *Giza++* entraîné selon le modèle *IBM4*. Il résulte de l'intersection un alignement très précis à liens simples, mais peu couvrant, intéressant comme alignement de départ.

Par ailleurs, intégrer les outils proposés dans Align$^{\text{It}}$ permet également de détourner le travail de post-édition en système d'évaluation simple. Le chronométrage du temps à aligner permet de mesurer l'évolution des algorithmes à base d'exemples avec l'avancement de la mémoire et les distances introduites rendent aussi compte de l'évolution en termes d'opérations nécessaires à finaliser un alignement.

Enfin, nous espérons tirer un effet fédérateur de l'utilisation d'outils automatiques. On se rappelle le guide d'annotation du projet Blinker dont le but était de réduire le biais expérimental en dirigeant les experts vers une bonne manière d'aligner. Dans notre approche recourant à d'éventuels non-experts, il est difficile de fixer une direction commune stricte et il est connu que le désaccord augmente avec le nombre de participants. Une direction proposée par le guide et qui est très intéressante dans la construction d'une telle ressource est de construire des alignements le plus couvrant possible. Les outils automatiques proposés que nous décrivons plus tard, reposent sur la combinaison de liens mémorisés pour produire un alignement qui, parmi d'autres critères, devra présenter une couverture maximale. Ainsi, les outils fonctionneront mieux si les alignements mémorisés sont très couvrants, mais ils auront tendance aussi à proposer comme solution des alignements le plus couvrant possible. Nous nous retrouvons encore dans une situation bouclante entre l'annotateur humain et l'annotateur robot. Un des effets de bord intéressant est donc qu'en utilisant les outils automatiques, l'annotateur humain se laisse suggérer un comportement, sans qu'il soit besoin de recourir à un guide explicatif.

3.1.2 Align$^{\text{It}}$: Présentation de l'interface homme-machine

Align$^{\text{It}}$ est un outil d'alignement collaboratif en ligne. Il peut être vu comme un outil d'annotation pour l'alignement utilisable en ligne et dont les utilisateurs partagent une même mémoire. Le site est accessible en français et en anglais et nous présentons ici l'interface d'acquisition et ses fonctionnalités. La majorité des informations délivrées dans cette partie 3.1.2 se trouvent sur le site dans la partie "*tutoriel*".

3.1.2.1 Premiers pas

En premier lieu l'utilisateur doit choisir un identifiant. Cela permet aux différents alignement récoltés d'être nominatifs et d'écarter par exemple un utilisateur malveillant des autres. La procédure d'inscription est volontairement légère et ne demande qu'un *pseudonyme*, *un mot de passe*[13] et une *adresse e-mail*. Une fois le formulaire d'inscription rempli et validé, l'utilisateur peut se connecter et cocher une option pour garder la session active via un cookie, ce qui lui évitera d'avoir à se reconnecter à chaque visite.

Figure 3.5 – Le travail d'alignement est nominatif

Une fois connecté, un nouvel onglet, marqué "*Aligner*", permet d'accéder à l'outil. L'utilisateur peut alors choisir un corpus et enregistrer ses propres alignements sur son espace de travail qui est consultable par tous comme nous allons le voir.

3.1.2.2 Choix du corpus

En cliquant sur "*Aligner*", l'utilisateur se retrouve sur la page de choix du corpus. Les corpus disponibles sont divisés en phrases qui sont sont stockées dans autant de tables qu'il y a de langues disponibles. On observe la présentation d'un corpus en figure 3.6. L'icône dans la barre de titre indique le type du corpus (ici un corpus journalistique).

Figure 3.6 – Un corpus et son descriptif

[13] le mot de passe est encodé et n'apparaît pas clairement dans la base de donnée

Des paires de langues sont proposées, mais on peut spécifiquement choisir une paire non proposée : dans l'exemple ci-dessus, il y a quatre langues disponibles, donc 12 combinaisons possibles. Nous ne proposons par défaut que les 3 paires de langues incluant le français, puis une troisième option ouvrant un menu à choix multiples pour les autres cas. Un bref descriptif du corpus est affiché.

3.1.2.3 Espace de travail

Lorsqu'un nouvel utilisateur a choisi son corpus, la première biphrase apparaît ainsi que quelques autres informations (voir la figure 3.7).

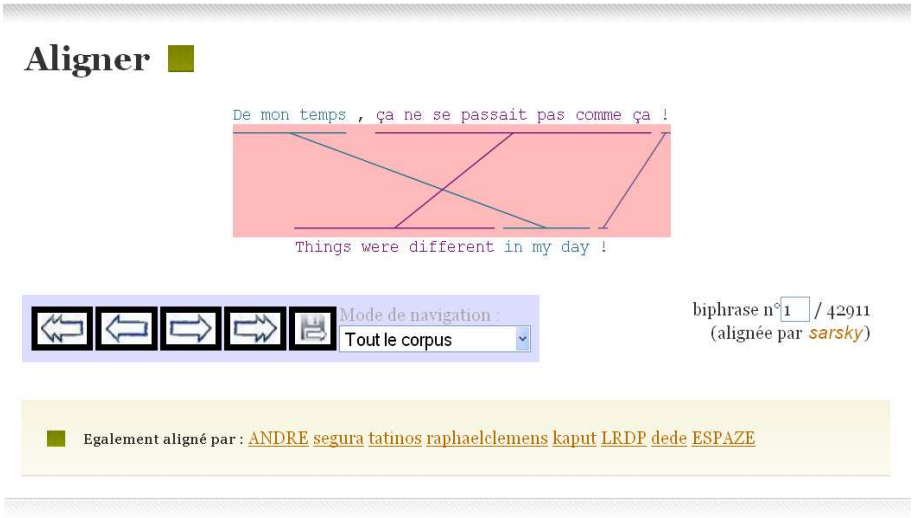


Figure 3.7 – Une biphrase interactive dans Align^{It}

Comme la biphrase a déjà été alignée par un autre utilisateur, des liens sont affichés et le fond du cadre est rouge. Le numéro de la biphrase apparaît à droite ainsi que la longueur du corpus. En dessous, on peut lire le pseudonyme de l'utilisateur ayant proposé cet alignement. On voit, en bas de l'écran, les autres utilisateurs ayant déjà aligné cette biphrase : on peut visualiser l'alignement de chacun en cliquant sur les pseudonymes. Il reste deux objets graphiques à détailler : tout d'abord le groupe des liens qui est le centre de l'outil, puis le cadre bleu de navigation qui permet de se déplacer dans le corpus.

3.1.2.4 Interactions à la souris et au clavier

On commence par remarquer l'affichage horizontal des biphrases et des liens dans l'espace de travail qui tranche avec l'interface du Blinker. En effet, cette dernière proposait une représentation verticale ce qui présente le défaut d'une lecture peu naturelle. L'ergonomie s'en était retrouvée critiquée par certains utilisateurs dont l'un admettait par mail (rapporté dans le papier) avoir parfois commis de mauvais alignements à cause de configurations compliquées nécessitant une manipulation fastidieuse :

> *"I do at times throw up my hands in frustration at how hard it is ... to link a word at the veeeeeeeery bottom. I reckon you just may get extra "not-linked"s because of this".*

Si la biphrase est trop longue il faut utiliser un scrolling horizontal. C'est réalisable de deux manières : tout d'abord à la souris en effectuant un "*attrapé-glissé*" (ou *grab and scroll* en anglais) de la partie centrale contenant les liens, ou en utilisant les flèches ← et → du clavier (voir figure 3.8).

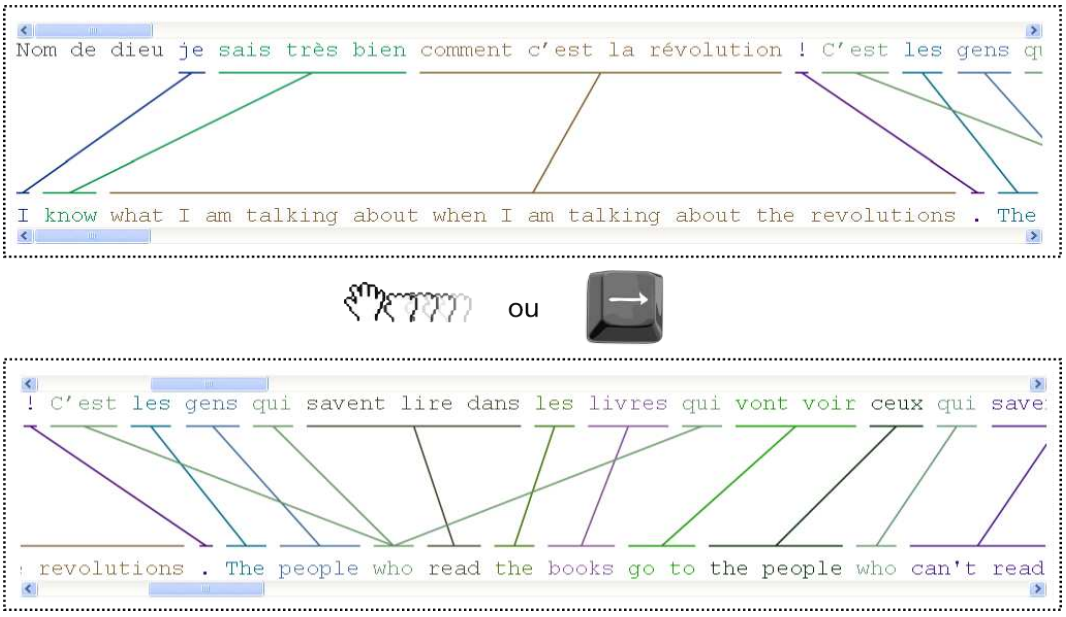


Figure 3.8 – L'ascenseur horizontal en cas de phrase trop longue

Une interface adaptée est importante pour des systèmes recourant à des utilisateurs [76] [86]. Mais si l'affichage horizontal règle le problème de la lisibilité, les grands chassés-croisés restent gênant. C'est pourquoi l'élément graphique possède deux barres de scrolling horizontal indépendantes pour les parties source et cible de la biphrase. Les liens s'adaptent dynamiquement à un scrolling asymétrique (voir figure 3.9).

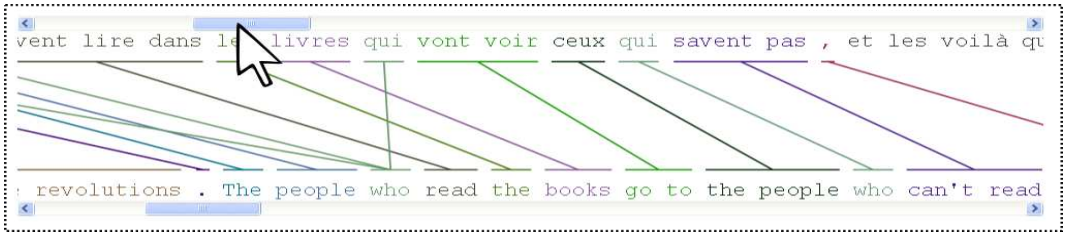


Figure 3.9 – Il est possible de décaler une des deux phrases toute seule

L'affichage des biphrases est géré par un élément graphique intéractif : les mots de la biphrase sont cliquables et il est possible de les lier intuitivement à la souris et au clavier. Nous décrivons ici le fonctionnement de l'interface d'annotation. Premièrement, placer un lien simple entre deux mots se fait à la souris : on selectionne les mots que l'on souhaite lier en cliquant dessus (ils se mettent alors en surbrillance). Un clic droit (ou la touche *Entrée*) validera alors le lien qui sera tracé comme le montre la figure 3.10.

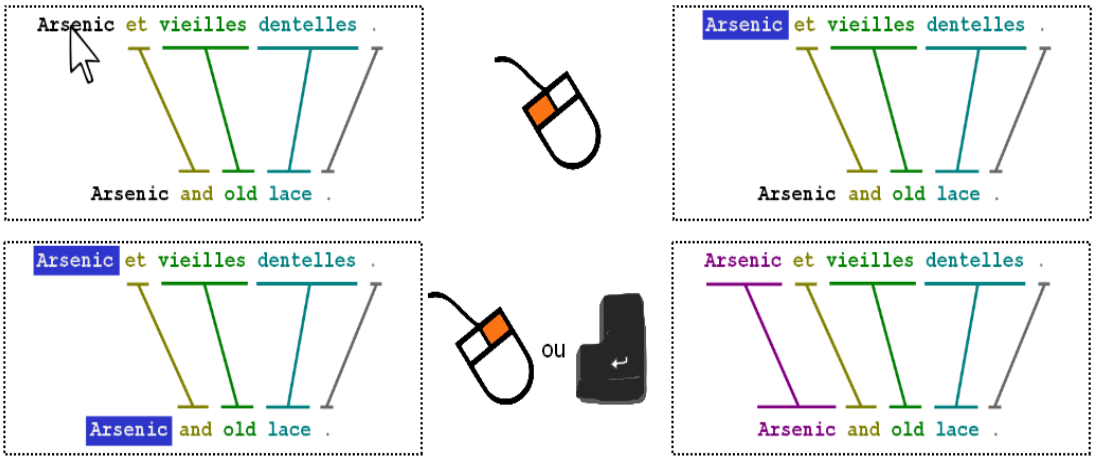


Figure 3.10 – Créer un lien simple

Tout alignement est faisable avec ces deux seules interactions. Toutefois, nous avons ajouté quelques possibilités supplémentaires afin de rendre l'interface plus ergonomique, car comme le montre l'expérience du Blinker une interaction difficile est aussi cause d'erreurs d'alignement. Notamment, l'utilisateur, lorsqu'il souhaite aligner de grands groupes de mots ensemble ne sera pas obligé de cliquer sur chacun des mots de manière répétée, mais pourra faire un *clic gauche appuyé* et survoler en une seule fois les mots à sélectionner (voir figure 3.11). Lier l'ensemble se fait alors aussi par un clic droit ou la touche entrée.

Figure 3.11 – Sélectionner un groupe entier par un clic appuyé et le survol des mots

Ce n'est pas tout de créer des liens, mais en situation de post-édition, il faut être capable de les modifier simplement. L'utilisation du *double-clic* s'est avérée utile pour sélectionner des ensembles de mots déjà tous liés ensemble pour, par exemple, les lier à un autre ensemble de mots (figure 3.12).

Figure 3.12 – Un double clic sera plus rapide si les mots sont liés ensembles

Enfin, le plus important pour la post-édition est de pouvoir *supprimer* des liens faux pour les corriger.

C'est possible en utilisant le *clic droit* sur l'un des mots *"mal liés"* (figure 3.13).

Figure 3.13 – Le clic droit sert aussi à effacer des liens

Mais de la même manière qu'il peut être fastidieux de sélectionner clic par clic un gros groupe de mots, ça l'est tout autant d'effacer des liens un par un avec cette technique. Nous ajoutons donc la possibilité d'effacer un ensemble de liens en une fois. Pour cela, il faut d'abord sélectionner certains des mots *"mal liés"* comme dans le premier exemple de la figure 3.14 et appuyer sur la touche *"Suppr"*. Parfois, l'alignement à corriger est trop mauvais et il peut être préférable d'effacer tous les liens. On peut le faire en appuyant sur *"Echap"*.

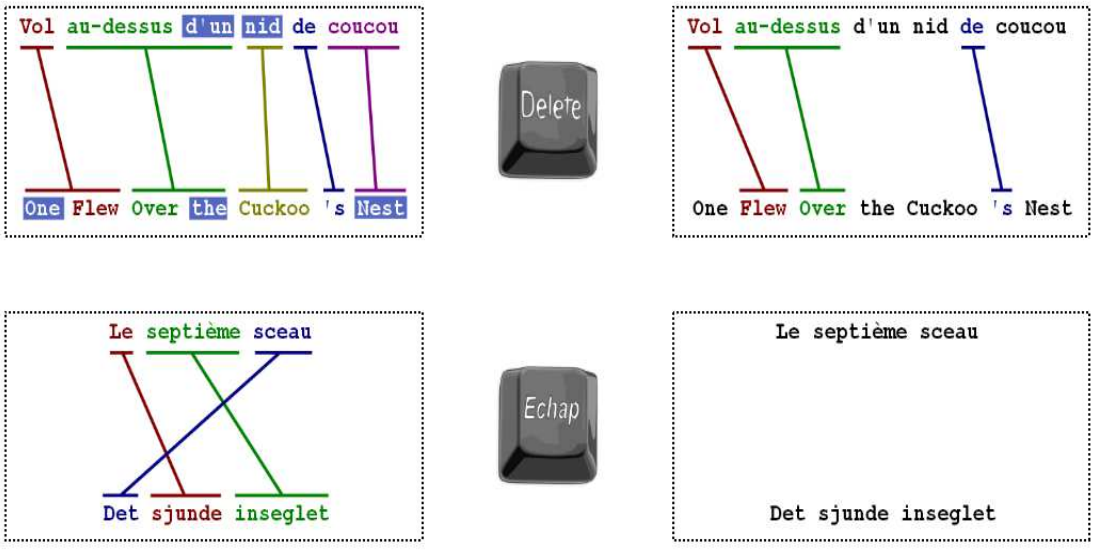


Figure 3.14 – On peut effacer plus de liens grâce aux touches *Suppr* et *Echap*

Remarque 6. *Les transformations élémentaires décrites au chapitre 2 pour introduire les distance d'édition sont bien sûr toutes réalisables par l'interface d'annotation, y compris l'opération de transfert étendu (figure 2.21 page 65). Pour effectuer chacune de ces transformations, il faut d'abord sélectionner les mots concernés, ce qui se fait en une seule opération grâce au clic appuyé. L'exécution de l'opération élémentaire demande alors un clic gauche ou la pression de la touche Suppr pour la suppression. Une opération élémentaire demande donc au minimum deux clics pour être effectuée. Finalement, le nombre de clics minimum pour post-éditer un alignement (de* ℓ_{in} *vers* ℓ_{out} *) grâce à l'interface graphique de Align*^{It} *est majoré par la distance d'édition* $2 * t_2(\ell_{in}, \ell_{out})$ *(section 2.3.2.3). De même, le nombre minimum de clics pour créer un alignement* ℓ *à partir d'une biphrase vierge est majoré par* $t_2(\ell_\emptyset, \ell)$*. Il y a majoration stricte en général car la distance d'édition ne compte pas les opérations de transfert étendu qui comptent pour deux opérations élémentaires. Cette majoration donne une bonne idée de l'optimalité de l'interface graphique en terme d'opérations possibles. On rappelle que la distance d'édition unitaire comptait un nombre minimal d'opérations déplaçant un seul mot à la fois. Si l'on retire la possibilité de sélectionner les groupes de mots par un clic gauche glissé, l'interface ne permet de réaliser que des transformations unitaires et le nombre minimal de clics pour créer un alignement* ℓ *devient donc majoré par* $t_1(\ell_\emptyset, \ell)$*. C'était par exemple le cas pour l'interface du Blinker.*

3.1.2.5 Naviguer dans le corpus

Dans cette partie nous détaillons le fonctionnement de la navigation au travers d'un corpus, c'est-à-dire les manières de passer d'une paire de phrases à une autre. Les biphrases du corpus sont affichées en partant de la première et l'utilisateur peut passer aux suivantes grâce à plusieurs options de navigation (figure 6.1).

Le bouton le plus important de la barre est  . Il permet à la suite d'un alignement de le sauvegarder et de passer à la biphrase suivante. Par défaut les modifications ou les nouveaux alignements ne seront pas sauvegardés tous seuls[14]. On constate qu'il y

[14] ça peut être amené à changer, on a constaté que quelques annotateurs oublient souvent de sauvegarder à la fin d'un alignement

a quatre options de navigation qui correspondent à des comportements différents des flèches. Avant de les décrire, on précise que le numéro de la biphrase indiqué à droite de la barre de la navigation est un champ modifiable dans lequel il est possible d'indiquer le numéro de la biphrase que l'on souhaite afficher. C'est la première manière de naviguer.

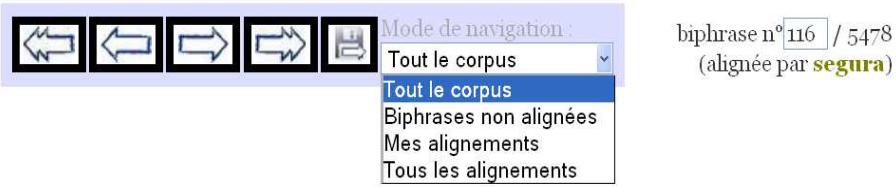


Figure 3.15 – Les options de navigation dans l'espace de travail

Naviguer sur "tout le corpus" donnera aux flèches simples les fonctions intuitives, à savoir que la flèche droite permet de passer à la biphrase suivante ($n+1$) et la flèche gauche à la précédente ($n-1$). Les doubles flèches permettront un saut au travers du corpus : la double flèche gauche permet d'aller à la première biphrase ($n=1$) du corpus et la double flèche droite à la dernière ($n =$*longueur du corpus*).

Naviguer sur "biphrases non alignées" restreindra l'espace de navigation aux biphrases jamais alignées. Les icônes ⇦ et ⇨ permettent d'afficher la précédente et la suivante paire non alignée. Les doubles flèches ⇦ et ⇨ permettent de sauter un grand nombre de biphrases. La double flèche droite permet de se placer juste avant la prochaine paire de phrases alignées, c'est-à-dire de sauter le paquet de biphrases non alignées dans lequel on se trouve. Ce mode de navigation est utile à ceux qui souhaitent aligner directement des phrases vierges sans considérer le travail d'un autre.

Naviguer sur "mes alignements" restreindra l'espace de navigation aux alignements de l'utilisateur seul. Les flèches ⇦ et ⇨ permettent de passer à la précédente et à la suivante paire de phrases alignées par l'utilisateur. Les doubles flèches ⇦ et ⇨ permettent de se déplacer plus vite. A titre d'exemple, si l'utilisateur a aligné les biphrases de 1 à 10 et de 20 à 30 en ignorant les biphrases de 11 à 19, la flèche simple

droite permettra à l'utilisateur de passer de la biphrase 10 à la biphrase 20 directement. De même, la flèche simple gauche ⇦ lui permettra de passer de la biphrase 20 à la biphrase 10. Si l'utilisateur est en train de visualiser la biphrase 7 par exemple, la flèche de droite lui permettra d'aller à la biphrase 8 comme dans le mode précédent. Les doubles flèches permettent de faire des sauts de manière plus utiles que dans le mode précédent. En observant le même exemple, si l'utilisateur est en train de visualiser la biphrase 2 et qu'il utilise la double flèche droite ⇨, celle-ci lui permettra de sauter directement à la biphrase 10. Un nouveau clic de la double flèche droite le fera sauter à la biphrase 20, puis 30. Ce mode de navigation un peu particulier permet aussi de vérifier les oublis.

Naviguer sur "tous les alignements" restreindra l'espace de navigation au travail de tous les utilisateurs. Par défaut, les biphrases affichées seront toujours celles de l'utilisateur lorsque c'est possible. Les flèches ont le même fonctionnement que pour le mode précédent, hormis que les biphrases visualisées seront choisies parmi tous les utilisateur. Dans l'avenir une quatrième option permettra de choisir l'affichage des alignements parmi un groupe d'utilisateurs choisis. Cela permettra un travail de groupe pour la récupération d'un sous-corpus aligné.

3.2 L'alignement sous-phrastique à base d'exemples

3.2.1 Avant-propos

La construction d'une ressource alignée manuellement représente une tâche colossale. Nous insistons donc dans la partie précédente sur l'importance d'utiliser des outils automatiques au moins comme pré-alignement pour se diriger vers un travail de post-édition. Pourtant il est clair que ces outils automatiques figés ne suffiront pas à eux seuls à faire de l'alignement manuel autre chose que ce qu'il est actuellement, c'est-à-dire une tâche fastidieuse et extrêmement coûteuse en temps. En reprenant l'argument du "*Translator's Amanuensis*", nous insistons sur l'accompagnement des annotateurs par des outils automatiques évolutifs. Nous souhaitons proposer une méthode s'appuyant sur les alignements déjà constitués et capable de s'en "*inspirer*" pour aligner en retour.

Cette alternative changerait la donne si elle se montrait capable, dans un cercle vertueux, d'alléger l'effort humain tout en augmentant la qualité de l'outil. Les outils automatiques seront "human-aided" tandis que les annotateurs seront "computer-guided". En reposant sur la mémoire d'alignements de Align^{It}, nous proposons des outils possédant les caractéristiques de la traduction à base d'exemples et y feront référence comme d'un outil d'*alignement à base d'exemples* (ABE) (enrobant un outil d'annotation collaboratif). La motivation de nos travaux est donc moins de constituer une ressource ayant la qualité d'alignements manuels, mais plus certainement le développement d'un outil capable de les produire, ou du moins de définir un tel outil. Nous consacrons cette partie à la description d'une architecture d'ABE (section 3.2.2), ses particularités et les solutions proposées (section 3.2.3).

3.2.2 Architecture générale

L'idée d'une approche à base d'exemples pour la traduction automatique est généralement attribuée à M. Nagao qui en énonce les principes dans un article de 1981 publié trois ans plus tard [100]. Il part du principe que personne ne traduit en effectuant au préalable des analyses linguistiques profondes, mais plutôt en observant la phrase source sur des parties locales reconnaissables, en les traduisant, puis en les recombinant proprement pour former la phrase cible :

> *Man does not translate a simple sentence by doing deep linguistic analysis, rather, Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into case frame units), then, by translating these fragmental phrases into other langage phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference.*

Nous pensons que cet énoncé, qui peut légitimement être discuté, peut tout aussi bien décrire le comportement d'un annotateur investi dans un projet de constitution de ressources alignées manuellement. On se place dans le cadre de l'interface décrite dans la

section 3.1.2. Le travail de traduction étant irrémédiablement lié à une compréhension progressive et une rédaction "dans le sens de lecture" [29], le commentaire concernant la décomposition de la phrase en fragments à traduire séparemment s'adapterait même plus favorablement à un travail d'alignement en remplaçant "*phrase*" par "*biphrase*" et "*traduire*" par "*aligner*". Bien sûr, il peut y avoir différentes manières d'aligner au même titre qu'il existe différents profils de traducteurs, mais un comportement possible est le suivant : reconnaître et lier des paires de mots ou expressions les plus rapidement identifiables puis, autour de ces "*liens d'ancrage*", rechercher des correspondances plus complexes entre les mots libres restants. En ceci, l'alignement peut s'envisager par touches locales indépendantes.

Dans son article fondateur, Makato Nagao définissait les trois étapes qui caractérise en général les approches à base d'exemples. Premièrement, la *collecte de fragments compatibles* provenant d'exemples déjà rencontrés et validés, puis l'*identification* d'équivalents de traduction correspondants et enfin la phase de *recombinaison*. Ces trois étapes sont plus ou moins respectées selon les approches de traduction à base d'exemples. L'architecture de l'outil d'alignement que nous proposons se sépare en quatre traitements principaux, suffisemment proches des trois étapes principales de la TABE pour que nous nous permettions d'employer cette expression d'*alignement sous-phrastique à base d'exemples* :

1. L'acquisition d'exemples manuels

2. Le traitement de la base d'exemples en vue de former les fragments

3. La sélection des fragments compatibles

4. La synthèse du résultat

Tout résultat pourra éventuellement être corrigé manuellement par l'annotateur pour enrichir la mémoire de nouveaux fragments. Le schéma 3.16 représente l'architecture générale de l'outil proposé et les transitions entre ses quatre étapes. Chacune d'elles est spécifiquement décrite dans les paragraphes suivants.

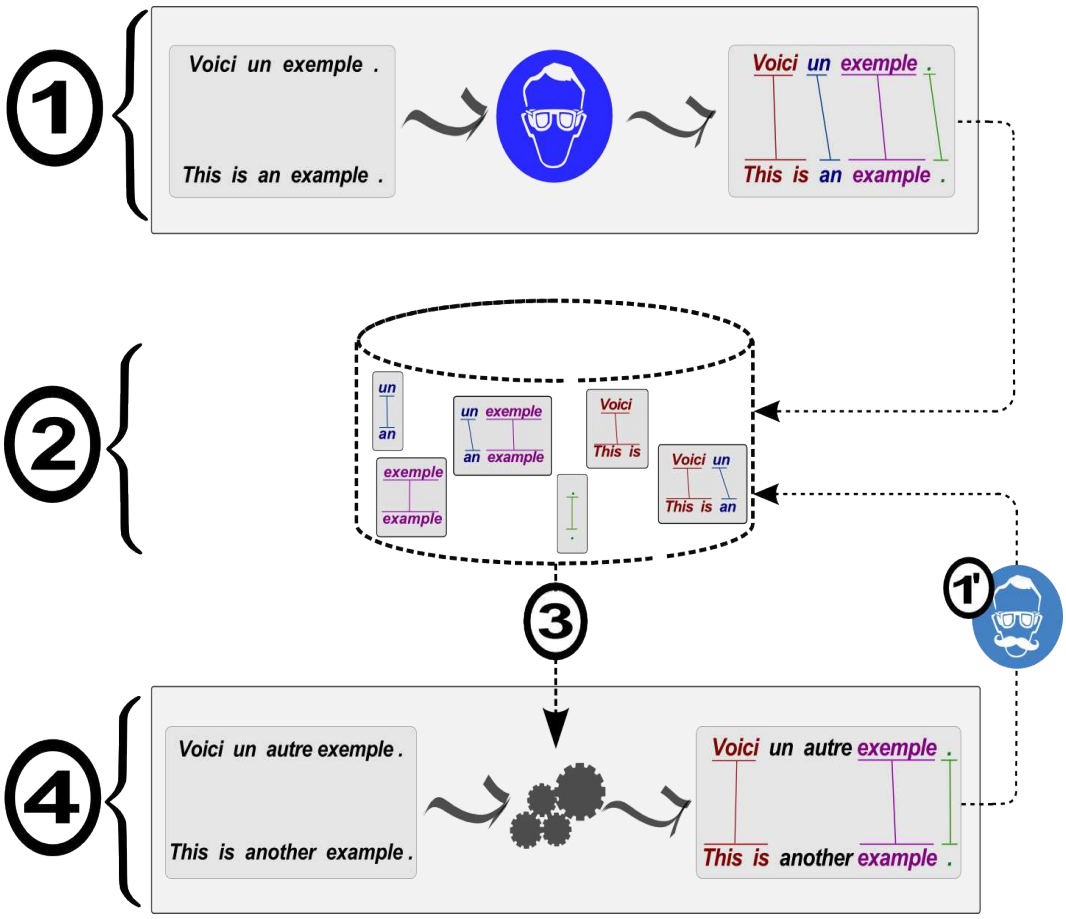


Figure 3.16 – schéma d'une approche d'alignement à base d'exemples

On pourrait penser que l'approche décrite ressemble par bien des points à un simple *outil d'aide à l'alignement* qui se différencie d'une approche dite *à base d'exemples* par le fait que l'utilisateur intervienne [135] aux étapes ① et ①'. Ce n'est pas tout à fait le cas ici puisque la partie automatique du processus ne nécessite aucune intervention humaine : l'étape de pré-alignement ① du schéma est facultative. Pour l'étape ①', les annotateurs se contentent de corriger le résultat et de le sauvegarder dans cette même mémoire que l'outil utilise comme référence.

3.2.2.1 Étape 1 : l'acquisition des exemples

Cette partie concerne l'étape ① du schéma, à savoir la manière de constituer les exemples. Cette partie a en fait déjà été décrite dans la section précédente 3.1 puisque les exemples proviennent simplement de l'outil Align^{It}. Nous rappelons donc que cette étape est plus complexe que ne le laisse présager le schéma 3.16 page 103 puisque l'outil permet une interopérabilité des annotateurs. Une situation plus correcte est schématisée à la figure 3.17.

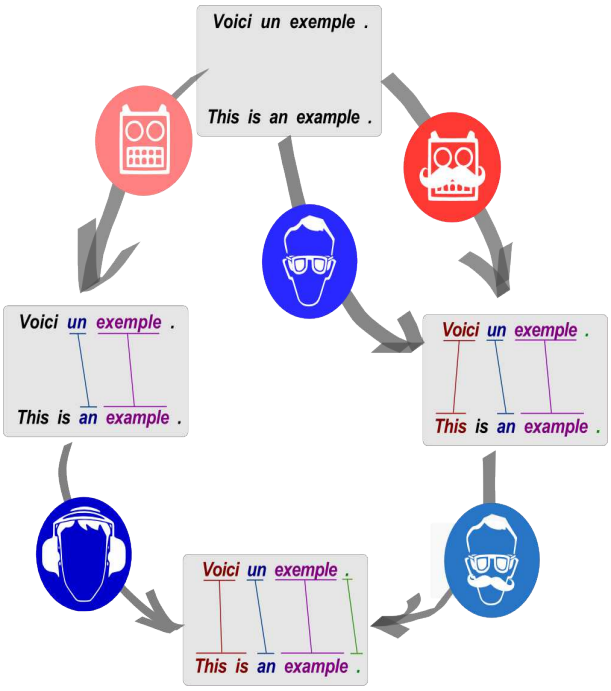


Figure 3.17 – Un alignement manuel collaboratif ouvert

Comme il est suggéré, des "robots" peuvent se mêler aux utilisateurs humains. Il peut s'agir de résultats provenant d'outils automatiques existants[15], mais leurs résultats ne seront pas mélangés à la mémoire d'alignements des utilisateurs qui sous-tend le mécanisme à base d'exemples. En effet, il en résulterait ultimement un aligneur singeant le comportement d'un autre aligneur. Un robot imitant un aligneur statistique ou à base de règles, bref , un clone dégénéré.

[15]Nous avons mentionné avant l'utilisation de *Giza++*

Nous avons insisté à de nombreuses reprises sur la nécessité pour l'outil proposé de pouvoir s'améliorer sous la supervision des annotateurs afin d'évoluer vers une forme entièrement automatique. Dans le schéma général, une transition empêche la réalisation de cet objectif : il s'agit de la boucle de rétroaction contrôlée par un utilisateur en étape ①'. Il est bien sûr nécessaire d'évaluer la qualité des alignements avant de les utiliser pour enrichir la mémoire d'exemples. Les outils automatiques utilisés en amorce sur la figure 3.17 représentant l'étape ① peuvent donc aussi bien provenir du système à base d'exemples que nous sommes en train de décrire. La figure représente donc tout aussi bien l'étape ①'. L'émancipation de l'outil automatique correspondrait à remplacer la boucle ①' par une transition non supervisée. Nous discuterons au chapitre 5 de l'éventualité d'un relâchement semi-supervisé de cette transition.

3.2.2.2 Étape 2 : Le traitement des exemples

À la manière de la TABE, l'ABE repose sur une *mémoire d'exemples* qu'il est nécessaire de traiter avant de les réutiliser. Nous reprenons l'approche classique qui consiste à *fragmenter* les exemples pour former la mémoire. Pour désigner les fragments représentés à l'étape ②, nous parlons de **bi-fragments alignés** ou de **fragments d'alignements** (une définition formelle sera établie en section 5.1.2). En l'absence d'ambiguïté, nous n'hésiterons pas à parler tout simplement de **fragments**.

L'importance de la fragmentation est justifiée notamment par l'idée suivante : un même exemple long ne se rencontre pas souvent, et n'est donc pas largement réutilisable [104]. Il convient de le préciser car toutes les machines de traduction à base d'exemples ne fragmentent pas leurs phrases, notamment celles qui utilisent l'analogie proportionnelle (voir [88]).

Nous verrons en section 5.1 une définition formelle ainsi que plusieurs propositions de fragmentation découlant sur différentes résolutions adaptées. On constatera que le type de fragmentation choisi aura des conséquences sur les deux étapes suivantes ③ et ④.

Un fragment de phrase correspond simplement à un groupe de mots extraits d'une phrase (ordre conservé). Par contre, un *fragment d'alignement* ne correspondra pas sim-

plement à deux fragments des phrases source et cible. Il faudra également que ce fragment conserve l'information des liens présents dans l'alignement sans en casser la structure. On peut voir à la figure 3.18 un exemple où l'on propose une bonne fragmentation et une deuxième qui nous semble mauvaise.

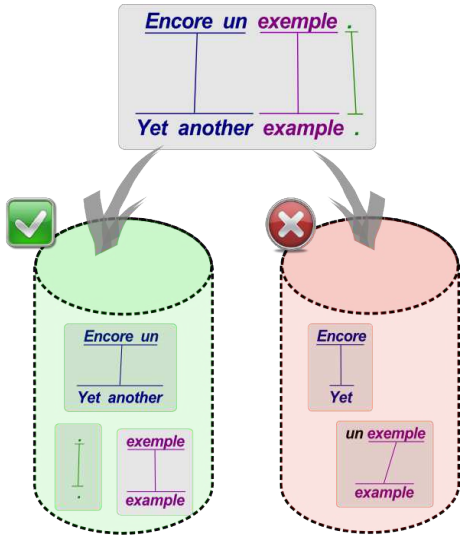


Figure 3.18 – Des fragments d'alignements

On remarque que notre mémoire d'alignements peut donner lieu à une mémoire de traduction mais pas le contraire. Les mémoires constituées seront multiples et pourront contenir des fragments d'alignement issus de différentes langues, de différentes longueurs, avec des informations lexicales, morphosyntaxiques ou sémantiques. En partant de corpus analysés syntaxiquement, nous utiliserons ce que l'on appelera des *patrons syntaxiques bilingues alignés* (nous parlerons abusivement de *patrons syntaxique* car il ne sera jamais question ici de patrons syntaxiques monolingues classiques) dans les méthodes proposées, c'est-à-dire des bi-fragments expurgés d'informations lexicales. Ils formeront des fragments plus ambigus mais aussi beaucoup plus génériques (voir la figure 3.19)

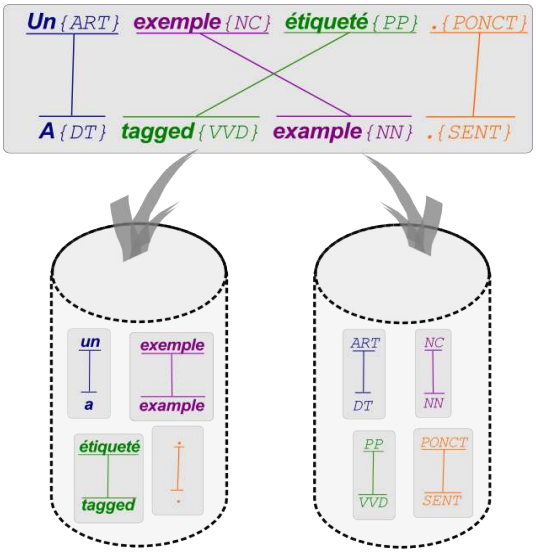


Figure 3.19 – Un étiquetage permettra aussi de former des patrons syntaxiques

3.2.2.3 Étape 3 : Compatibilité des fragments

L'étape ③ de du schéma page 103 correspond à l'extraction de fragments utiles à l'outil pour proposer l'alignement d'une certaine biphrase. Un fragment de la mémoire pourra s'avérer utile si les informations de ses parties source et cible apparaissent dans les parties source et cible de la biphrase. Dans le cas d'un fragment normal (ne contenant que des informations lexicales) la compatibilité se résume à une égalité de chaîne tandis que pour des patrons syntaxiques, la compatibilité se vérifiera en se référant aux étiquettes morphosyntaxique de la biphrase considérée. La base de patrons syntaxiques est dépendante des jeux d'étiquettes utilisés. Nous reprendrons donc les étiquettes des outils automatiques utilisés pour l'analyse des différentes langues (les analyseurs sont décrits partie 6.1.2). Nous n'avons pas abordé la possibilité de proposer des fragments proches selon des considérations de "*fuzzy matching*".

Il est important que la taille de cet ensemble reste raisonnable sans quoi le coût algorithmique s'en ressentirait fortement. Dans la section 5.1.3, nous étudierons quantitativement les ensembles de fragments compatibles retournés par cette étape.

3.2.2.4 Étape 4 : Synthèse du résultat

Chaque fragment retenu par l'étape ③ propose un alignement local potentiel pour la biphrase en cours. Bien sûr, ces fragments ne sont pas forcément compatibles deux à deux et il ne suffit pas d'en collecter beaucoup pour obtenir un bon alignement. L'étape ④ sert à sélectionner un sous-ensemble des meilleurs fragments compatibles afin de former un bon alignement selon certains critères. L'approche directe, similaire à certaines approches de traduction à base de règles, est l'application successive de fragments ordonnés par fréquences croissantes. Cette approche révèlera ses limites et nous lui préfèrerons une solution que nous reprenons de la TABE, qui est d'envisager une résolution du problème selon des critères de couverture (voir par exemple [43]). Ce type d'approche permet de considérer une solution dans son ensemble et s'avérera plus appropriée. La section 5.2 se consacrera à proposer des solutions algorithmiques pour cette étape.

3.2.3 L'analyse syntaxique en renfort

Nous présentons ici quelques difficultés inhérentes aux différents problèmes que sont la **nécessité d'étendre la mémoire**, l'**ambiguïté** et la **taille des fragments**. Au travers de celles-ci, nous verrons en quoi l'utilisation d'informations syntaxiques peut apporter des solutions intéressantes pour certaines des étapes du processus.

3.2.3.1 Quelle taille pour les fragments ?

Nombreuses sont les approches à se poser cette question en abordant la TABE au travers de la fragmentation d'exemples. Dans la littérature sur la TABE un compromis (que l'on souhaite appliquer à l'ABE) doit s'installer entre des fragments de taille trop petite qui, bien que génériques, soulevent le problème de l'ambiguïté, et d'une taille trop grande, plus sûrs mais rarement réutilisables [104]. Un consensus se fait donc autour d'une granularité sous phrastique, quelque part, donc, entre la phrase et le mot.

Les outils automatiques d'analyse morphosyntaxique permettent de décomposer les phrases en sous-parties, ce qui offre un solution au problème de la granularité des fragments pour des approches syntaxiques de TABE [4] [79]. Les outils existants pourront

fournir des analyses de différentes natures. Nous appelons *chunker* des outils proposant un découpage en *chunk*, c'est-à-dire des séquences non récursives formées de mots contigus (*TreeTagger* [123] est un exemple de chunker). On parlera plus généralement de *syntagmes* pour des découpages plus intriqués, potentiellement récursifs et formés de séquences de mots pas nécessairement contigus (on parlera plutôt de syntagmes pour un analyseur comme *Sygfran* [35]). Les chunk et les syntagmes possèdent une tête lexicale qui peut être un nom ou un verbe, plus rarement un adverbe, un adjectif ou un pronom. Un *chunk* est aussi appelé *syntagme non récursif* [142].

Il existe également des travaux qui ne s'encombrent pas de considérations linguistiques pour segmenter en unités plus longues, préférant par exemple considérer toutes les sous-séquences de deux mots ou plus comme dans [61]. C'est souvent le cas lorsque se pose le problème de la disponibilité des outils automatiques dans des langues peu dotées. Cependant, l'alignement possède l'avantage par rapport à la traduction de pouvoir s'appuyer sur une langue pour laquelle existe des outils d'analyse pour le bénéfice d'une langue moins bien dotée. En effet, nous verrons en partie 5.1.2 comment l'analyse syntaxique de la phrase source peut structurer, même grossièrement, la phrase en langue cible via les liens bilingues. Cette idée est commune aux travaux sur les grammaires bilingues de Dekai Wu [154] ou à ceux de Tang Enya Kong [5] qui construit une structure syntaxique pour une phrase cible à partir d'une structure source en la faisant "traverser" des liens d'alignement (nous reparlerons de ces structures alignées dans la section 4.1.2).

Les arguments en faveur des outils d'analyse syntaxique sont nombreux dans les approches à base d'exemples. Une fragmentation linguistiquement motivée augmenterait à la fois la généricité et la qualité de la phase finale de recombinaison [62] en réduisant les incompatibilités de fragments qui propagent des erreurs (d'où, également, la nécessité d'un critère global plutôt qu'un traitement séquentiel). De plus, des fragments dans chaque langue avec un statut syntaxique identifié (e.g. groupe nominal, groupe verbal, etc...) seront plus facilement mis en correspondance pour former des fragments d'alignement usuels et génériques [86]. Cette dernière remarque rejoint l'*hypothèse de cohésion* défendue dans [56] selon laquelle les syntagmes dans une langue ont tendance à être peu affectés par le reordonancement provoqué par la traduction.

Au final, nous souhaitons proposer une méthode capable de tirer avantage d'informations syntaxiques founies par des outils automatiques de différentes qualités, sans pour autant que cela n'exclut de notre approche des langues moins bien dotées et n'ayant que des outils d'analyse plus rudimentaires, comme des *étiqueteurs* morphosyntaxiques.

3.2.3.2 Extension de la mémoire

L'approche que nous souhaitons proposer repose sur la nécessité pour l'outil d'être capable de progresser et donc de diminuer progressivement l'effort manuel sans quoi il ne peut rester qu'une interface d'annotation [76]. Il nous faut donc imaginer une mémoire d'exemples permettant, d'une part, une *réutilisation* rapide de nombreux cas dans de nombreuses situations, mais aussi de *s'étendre* à des cas nouveaux. Nous pensons qu'une mémoire de *patrons syntaxiques* peut apporter des solutions intéressantes. Imaginons un instant que notre approche repose uniquement sur des fragments se limitant à des informations lexicales. Considérons l'exemple court qui suit :

Les fourberies de Scapin
Scapin's Deceits

On peut facilement imaginer que la mémoire de fragments ne contienne pas la correspondance *"fourberie"-"deceit"* tant les termes ne sont pas usuels. À partir de celle-ci, nous ne produirions qu'un alignement partiel comme l'illustre la figure 3.20.

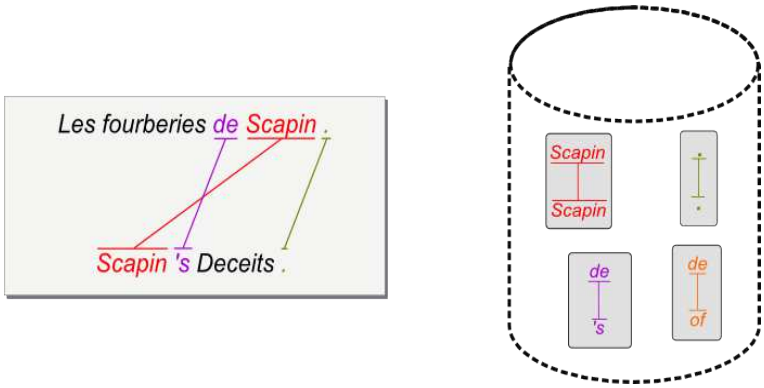


Figure 3.20 – Mémoire partielle, alignement partiel

La mémoire, qui est purement lexicale, propose des fragments très pertinents, mais faiblement réutilisables. Construire une mémoire d'alignements dans ces conditions n'est tout simplement pas envisageable. Une solution envisagée par certains travaux en TABE est la *"généralisation"* de leurs fragments pour accélérer le processus et augmenter la généricité. Cela se traduit par la possibilité de remplacer, dans un fragment, un mot par une information générale le caractérisant. À titre d'exemple, l'approche de R.D. Brown [27] permet par exemple de généraliser par des informations sémantiques des *couleurs*, des *lieux* ou encore des *prénoms*, mais aussi par des informations morphosyntaxiques des *noms communs*, des *adjectifs* ou des *verbes*. D'une manière analogue, des approches purement statistiques tentent de palier la rareté des fragments plus longs en utilisant des classes de mots déterminées automatiquement (voir [105]). En ce qui nous concerne, nous n'avons travaillé qu'avec un ensemble de catégories morphosyntaxiques provenant d'analyseurs (décrits en section 6.1.2), mais l'approche reste adaptée à l'utilisation d'informations plus générales.

Dans des situations où l'information lexicale manque, force est de constater qu'un outil d'analyse syntaxique (ou simplement un étiqueteur), permet d'obtenir des informations supplémentaires, même partielles, succeptibles d'aider l'alignement. En effet, un analyseur pourra éventuellement :

1. reconnaître le mot et lui attribuer une catégorie morphosyntaxique

2. ne pas reconnaître le mot, mais proposer une catégorie par analyse morphologique

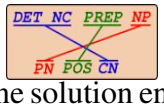
3. ne pas reconnaître le mot, mais proposer une catégorie par rapport au contexte

4. ne pas le reconnaître et donner une catégorie de type INCONNU

Dans tous les cas on a une information, même ténue, sur les zones non connues de notre biphrase à aligner. Même le cas d'une étiquette INCONNU contenue dans un fragment plus long pourra s'avérer intéressant.

Dans cet exemple, le transfert se fait suivant deux transformations classiques qui sont en fait des cas de divergence syntaxique (voir section 1.2), à savoir la disparition

de l'article et l'inversion en cas de possessif (respectivement au couple de langues français/anglais).

En introduisant le patron syntaxique bilingue 

dans notre mémoire, on peut tenir compte de cette divergence et proposer une solution en superposant le fragment à notre exemple aligné partiellement (voir figure 3.21). Le fragment introduit lie deux syntagmes nominaux dont les étiquettes sont : *NC* pour *nom commun*, *NP* pour *nom propre*, *PREP* pour *préposition*, *CN* pour *common noun*, *PN* pour *proper noun* et *POS* pour *possessive*. On remarque que si les étiquettes morphosyntaxiques semblent parfois manquer de précision (*PREP* par exemple), l'élément lié peut apporter une précision (on sait que *PREP* marque la possession grâce au lien avec *POS*).

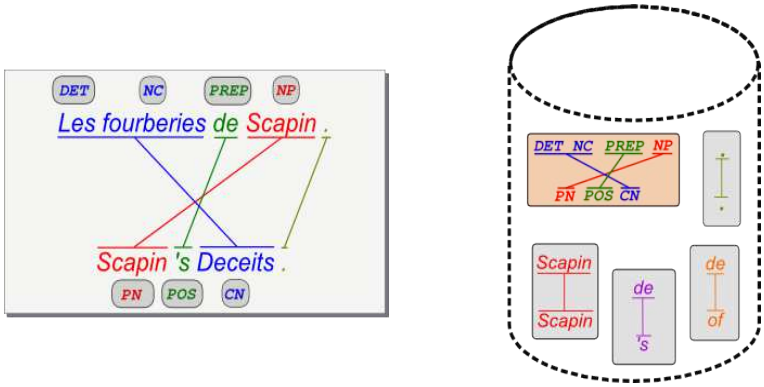


Figure 3.21 – Un patron syntaxique bilingue pour lier les "*fourberies*"

La correspondance *"fourberies"-"deceits"* a ainsi pu être trouvée. Le fragment nécessaire qu'il a fallu ajouter est assez répandu et donc rapide à obtenir. La majorité des exemples de divergence de la section 1.2 peuvent être alignés de manière analogue. Nous verrons en partie 6.2.1 qu'à un stade précoce de son développement la mémoire contiendra déjà des patrons syntaxiques suffisament génériques pour assurer de bonnes performances aux outil automatiques.

3.2.3.3 Ambiguïté et alignement

L'ambiguïté est un problème courant en TABE, lié à la dernière des étapes (la synthèse de la phrase cible) qui est la plus difficile. L'ABE souffrira aussi de problèmes

liés à l'ambiguïté pour lesquels les patrons syntaxiques, utilisés pour propager des liens, peuvent apporter leur solution (la propagation de liens est une technique efficace qui a fait ses preuves dans [107]).

Mais nous allons voir qu'en ce qui concerne l'alignement, la notion d'ambiguïté se révèle assez différente de celle rencontrée en traduction, pour des approches à base d'exemples. En effet, en traduction, l'ambiguïté est affaire de **choix** et gravite autour de la polysémie. Prenons un exemple classique d'ambiguïté.

L'avocat est véreux.

Dans un but de traduction, il convient en premier lieu de se poser la question du sens du mot *avocat*. Comme on le constate ici, la réponse n'est pas forcément tranchée. Ajoutons à cela qu'une mémoire pourra proposer de subtiles variations (voir figure 3.22 vue comme une mémoire de traduction).

Le processus devra alors séparer les sens possibles, exclure les mauvaises combinaisons et choisir les meilleures tournures en s'aidant d'indices tels que la fréquence, le contexte, la syntaxe, etc...

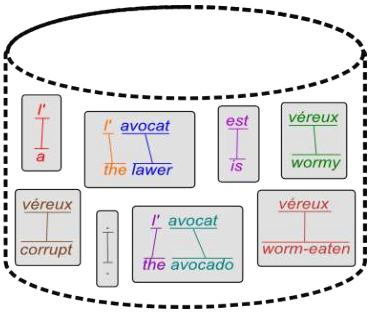


Figure 3.22 – Une mémoire de traduction/d'alignement simplifiée

Mais en se plaçant dans une problématique d'alignement, la situation est différente car la donnée initiale est une biphrase. Adaptons notre exemple de l'avocat :

L'avocat est véreux.
The avocado is wormy.

Des phrases dont la traduction dans une autre langue préserve l'ambiguïté sont très rares. Généralement, le sens devient clairement identifiable dès lors que l'on confronte une même phrase traduite dans plusieurs langues. Si l'on souhaite donc aligner cette biphrase en prenant comme mémoire d'alignements la même figure 3.22, on se rend compte que tout problème d'ambiguïté a disparu et que la résolution est directe pour obtenir l'alignement souhaité en figure 3.23.

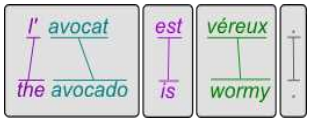
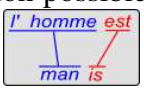


Figure 3.23 – L'alignement est direct par juxtaposition des fragments disponibles

N'en concluons pas pour autant que les problèmes d'ambiguïté en ABE sont inexistants ou extrêmement simplifiés. Ils sont en fait plus contraints et nécessitent des approches de résolution totalement différentes.
Quels sont alors les cas d'ambiguité pouvant être problématiques lors d'un choix de fragments d'alignement ? Ils ne sont pas liés aux cas rencontrés en traduction, il faut plutôt regarder du côté des répétitions dans une des phrases (voire même des deux). Les répétitions posent un réel problème de positionnement des liens. Les cas les plus fréquents apparaissent avec les déterminants et les prépositions, mais pas seulement. Prenons l'exemple suivant :

L'homme est un loup pour l'homme.
Man is wolf to man.

La résolution via une mémoire basée sur des informations lexicales peut ne pas fonctionner à cause des répététions de "*homme*" et "*man*" (fig. 3.24). Pourtant, la traduction dans un sens ou dans l'autre, en utilisant la même mémoire de fragments, ne pose aucune difficulté liée à l'ambiguïté, ni au transfert. La figure suggère une solution possible via des fragments plus longs. En effet, en ajoutant simplement le fragment 

à la mémoire, l'alignement devient possible.

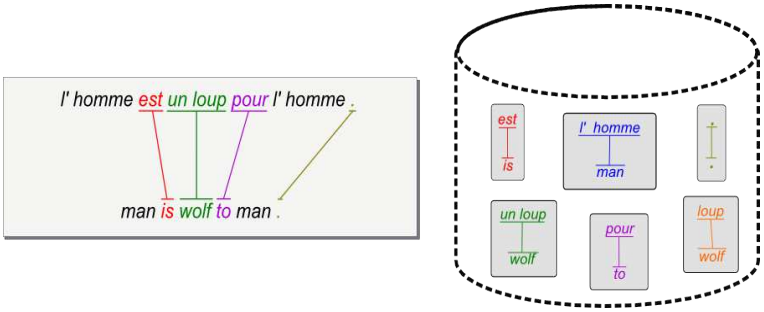
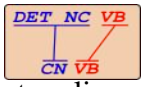


Figure 3.24 – Un cas d'ambiguité

Or, les fragments longs sont plus rares et donc plus difficiles à obtenir en général (à plus forte raison dans une approche recourant à des annotateurs). En ne retenant de ce fragment que le squelette syntaxique, on obtient ce cas transfert classique 

qui sera obtenu dès les premiers alignements annotés. En le superposant à notre alignement partiel, encore une fois, nous étendons l'alignement pour obtenir la configuration souhaitée (fig. 3.25).

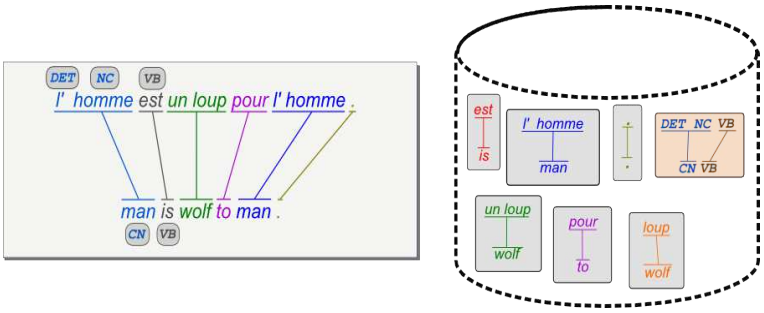


Figure 3.25 – L'insertion de "patrons syntaxiques" permet de résoudre l'ambiguité

Les outils automatiques proposés en section 5.2 se baseront exclusivement sur des mémoires de patrons syntaxiques. Les pré-alignements que nous pourront faire proviendront soit d'outils automatiques (*Giza++* surtout), soit d'heuristiques utilisant des dictionnaires ou des mesures de similarité visant à détecter des cognats. La constitution d'une mémoire de fragments purement lexicaux ou hybrides est envisageable, mais hors du cadre de notre étude .

CHAPITRE 4

MODÈLE DE REPRÉSENTATION THÉORIQUE POUR DES ALIGNEMENTS STRUCTURÉS

Le présent chapitre propose de mettre en place un cadre formel pour la structure globale des éléments traités par l'outil Align$^{\text{It}}$, à savoir des biphrases alignées munies de deux analyses syntaxiques. L'intrication des différentes composantes nécessite différents niveaux de "synchronisation". Tout d'abord, une correspondance monolingue entre phrase et structure sera prise en charge par le modèle de représentation des SSTC, présentée en section 4.1.1. Les S-SSTC, pendant bilingue des SSTC, nous permettront de structurer une correspondance bilingue aux deux niveaux lexical et syntaxique (section 4.1.2). Nous adapterons les formalismes à notre situation en présentant des raffinements de la définition de S-SSTC et présenterons quelques résultats théoriques. Nous détaillerons les éléments manipulés par l'outil Align$^{\text{It}}$ (section 4.2) nécessitant l'expressivité des S-SSTC.

4.1 Une structure expressive pour décrire les exemples

Nous rappelons ici les notions de SSTC (*Structured String-Tree Correspondance*) introduite dans [18] et de S-SSTC (*Synchronous SSTC*) [2]. Pour des raisons de cohérence et de lisibilité, certaines notations pourront être légèrement modifiées tout en restant reconnaissables. Quand de nouvelles propriétés seront énoncées, nous le rappellerons.

4.1.1 Une correspondance entre l'arbre et la phrase : la SSTC

Nous avons vu dans la partie précédente que nous souhaitions utiliser des analyseurs syntaxiques afin de doter les phrases d'un étiquetage morphosyntaxique ainsi que d'une structure d'arbre syntagmatique et/ou en dépendance. Nous incluons aussi le cas où nous ne disposons que d'un étiqueteur en créant artificiellement un arbre d'analyse rudimen-

taire, *projectif*[1] et de profondeur 1. On souhaite aussi tenir compte du fait que certains analyseurs peuvent proposer des représentations en arbre non projectif. C'est notamment pour répondre à ce dernier problème, dans le cadre d'éditeurs de structures de textes, qu'ont été introduites les SSTC. La représentation permet de maintenir une correspondance entre les mots d'une phrase et les nœuds de la structure (syntaxique, dans notre cas) associée. Cette correspondance permettra l'accès aux informations morpho-syntaxiques des mots affichés dans notre interface d'acquisition de Align$^{\text{It}}$. On rappelle la définition générale :

Définition 17. *Soient $n \in \mathbb{N}$ la longueur de la chaîne st décomposée en mots contigus : $st = (st_1, \ldots, st_n)$, $tr = (E, V)$ l'arbre associé où E et V sont respectivement les arêtes et les sommets. On pose $V = (v_1, \ldots, v_m)$. On se donne un couple de fonctions $co = (snode, stree)$ supposées représenter les correspondances entre la structure et la chaîne :*

- *$snode : V \longmapsto \mathscr{P}(\{st_1, \ldots, st_n\})$. L'ensemble $snode(v_i)$ est l'ensemble des mots correspondant au nœud v_i dans l'arbre d'analyse tr.*
- *$stree : V \longmapsto \mathscr{P}(\{st_1, \ldots, st_n\})$. L'ensemble $stree(v_i)$ est l'ensemble des mots correspondant au sous-arbre de tr enraciné en v_i.*

Le triplet (st, tr, co) est une ***Structured String-Tree Correspondance*** *ou* ***SSTC*** *si les conditions d'homogénéité suivantes sont vérifiées :*

1. *$snode(v_i) \subset stree(v_i)$, pour tout $i \in \{1, \ldots, m\}$*
2. *Soient v_i un nœud de tr et $v_{i_1}, \ldots, v_{i_m}$ ses nœuds fils, alors :*

$$stree(v_i) \supseteq stree(v_{i_1}) \cup \cdots \cup stree(v_{i_m}) \cup snode(v_i)$$

On remarque qu'un nœud ou un sous-arbre peut être associé à l'ensemble vide $\emptyset$, ce qui signifie qu'il ne représente aucun groupe de mot.

[1] un arbre est projectif si la phrase dont il dérive peut être obtenue en lisant les nœuds dans un ordre bas-haut, gauche-droite

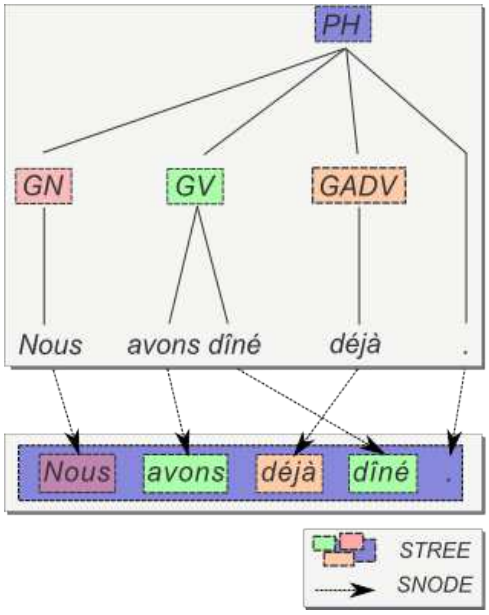


Figure 4.1 – Exemple de SSTC

La figure 4.1 décrit une SSTC sur un exemple non projectif. On peut observer que les liens *snode* sont positionnés sur des nœuds terminaux de l'arbre syntaxique. Pourtant la définition n'interdit pas une correspondance *snode* entre des nœuds internes et des mots, mais cela se révèle utile pour des arbres d'analyse en dépendance. Les arbres syntagmatiques n'ont que rarement des nœuds internes correspondant à des mots du texte (sauf sous-analyses d'expressions composées par exemple). N'ayant travaillé qu'avec des analyseurs formant des arbres syntagmatiques, les exemples de SSTC donnés ici ne seront jamais des représentations en dépendance.

Remarque 7. *Nous précisons que nous avons omis de la définition d'origine une troisième correspondance ssubt entre des sous-arbres de l'arbre tr et des mots de la chaîne. Cette fonction est structurellement assez différente des deux autres puisqu'elle est définie de l'ensemble des sous-arbres de tr dans l'ensemble des parties des mots. On perd donc, par rapport à stree, la représentation d'un arbre par son unique racine. Nous n'utiliserons dans notre approche que les correspondances snode et stree. Nous nous sommes donc permis d'alléger la définition.*

Nous introduisons des notations commodes de descendant et d'ascendant car il en sera souvent question dans cette partie :

Notation. *Soient $v, v' \in V$ les nœuds d'un arbre tr. On utilisera les notations suivantes :*

- $v' \downarrow v$ *si* v' *est un descendant de* v *ou* $v = v'$.
- $v' \uparrow v$ *si* v' *est un ascendant de* v *ou* $v = v'$.

Les définitions suivantes sont données dans l'article original et décrivent des SSTC ayant des propriétés de "bonne formation". Une SSTC sera dite *complète* si la correspondance arbre-chaîne n'omet aucun mot (ce qui est généralement attendu d'une SSTC dans des cas pratiques). La propriété de *non-chevauchement* sera vérifiée pour des SSTC dont les correspondances lexicales sont dominées par les correspondances structurelles dans l'arbre d'analyse (cette propriété concerne plutôt les structure en dépendance puisque les arbres de constituants n'ont que rarement des nœuds internes associés à un mot).

Définition 18. *On définit les caractéristiques :*

- *La SSTC est* ***complète*** *si tous les mots st_i sont associés à un nœud par snode.*
- *La SSTC est* ***non chevauchante*** *si les deux conditions sont remplies :*
  1. $snode(v_1) \cap stree(v_2) = \emptyset$ *pour tous* $v_1, v_2 \in V$ *tels que* v_2 *est un descendant de* v_1.
  2. $stree(v_1) \cap stree(v_2) = \emptyset$ *pour tous* $v_1, v_2 \in V$ *indépendants (i.e. l'un n'est pas le descendant de l'autre).*

En s'apputant sur ces définitions, nous avons pu énoncer la propriété 3, exprimant le fait que sous l'hypothèse d'une bonne formation, la correspondance structurelle *stree* se calcule entièrement à partir des correspondances *snode* par remontée au travers de l'arbre. Cette proposition est intéressante du point de vue de l'interface d'acquisition de Align$^{\text{It}}$. En effet, nous souhaitons que l'interaction avec l'utilisateur se concentre essentiellement sur la partie alignement. Il est donc préférable de ne pas compliquer outre mesure l'interface en y faisant apparaître une structure syntaxique. Les éléments

graphiques affichés ne seront donc que les mots de la biphrase traitée et les structures seront "cachées". Être capable de calculer exactement la correspondance *stree* à partir de *snode* permet de gérer automatiquement la partie structurelle de la SSTC et justifie la minimalité de l'interface.

Propriété 3. *Si la SSTC est complète et non chevauchante, alors stree est entièrement déterminée par snode. Soit* $v \in V$, $stree(v)$ *se calcule par induction :*

- *Si v est une feuille, alors* $stree(v) = snode(v)$
- *Soient* $v_{i_1}, \ldots, v_{i_k} \in V$ *les fils de* $v \in V$ *dans tr,*

$$stree(v) = stree(v_{i_1}) \cup \cdots \cup stree(v_{i_k}) \cup snode(v)$$

Preuve. D'après la définition, on a déjà l'inclusion :

$$stree(v) \supseteq stree(v_{i_1}) \cup \cdots \cup stree(v_{i_m}) \cup snode(v)$$

Par l'absurde, on suppose l'inclusion stricte.
On se donne $st_{i_0} \in stree(v) \backslash stree(v_{i_1}) \cup \cdots \cup stree(v_{i_m}) \cup snode(v)$. La SSTC étant complète, il y a un nœud $w \in V$ tel que $st_{i_0} \in snode(w)$ et donc $st_{i_0} \in stree(w)$. Il y a trois positions relatives possibles entre w et v que nous allons écarter :

1. *w est un descendant de v* : exclu, sinon w est le descendant (non strict) d'un des v_{i_k} et on aurait $st_{i_0} \in stree(v_{i_k})$, ce qui contredit l'hypothèse de départ.
2. *w est un ascendant de v* : de même, par remontée, on aurait $st_{i_0} \in snode(w) \cap stree(v)$ ce qui contredit la première hypothèse de non-chevauchement.
3. *w et v sont indépendants* : dans ce cas, $st_{i_0} \in stree(w) \cap stree(v)$, ce qui contredit la deuxième hypothèse de non-chevauchement.

□

Remarque 8. *Cette propriété s'appliquera aux structures sur lesquelles notre approche repose car elles seront toujours complètes et non chevauchantes.*

Nous ajoutons une définition supplémentaire n'étant pas dans l'article original, celle d'une **SSTC terminale**. Cela nous permettra d'exclure des structures peu pertinentes dans lesquelles des branches ne portent aucune information, ou du moins aucune information relative à la phrase représentée.

Définition 19. *Une SSTC est **terminale** si pour toute feuille $v \in V$, on a $snode(v) \neq \emptyset$*

Une conséquence directe de cette définition découlant du deuxième axiome d'homogénéité de la définition 17 :

Propriété 4. *Dans une SSTC terminale, pour tout $v \in V$, on a $stree(v) \neq \emptyset$*

Notation. *Les fonction snode et stree sont définies sur les sommets de l'arbre structure une SSTC. On étend leur utilisation à un ensemble de sommets en notant simplement pour $V' \subset V$ (ce qui allègera les notations dans la partie suivante) :*

$$SNODE(V') = \bigcup_{v \in V'} snode(v) \qquad et \qquad STREE(V') = \bigcup_{v \in V'} stree(v)$$

4.1.2 Une correspondance bilingue entre SSTC : la S-SSTC

Le modèle de représentation présenté ici décrit des structures bilingues en relation de traduction. Elle s'apparente à d'autres modèles commes les dérivations de S-TAG (*Synchronous Tree-Adjoining Grammars*) [132] ou de ITG [154]. Cette structure est en fait double puisqu'elle propose deux alignements entre des SSTC : l'un lexical et l'autre structurel. Il s'agit des S-SSTC (*Synchronous SSTC*) introduites par Tang Enya Kong dans le but de construire une BKB (Bilingual Knowledge Bank) [1] pour le couple de langues malais-anglais en vue de créer une machine de traduction à base d'exemples. Cette structure était, à l'instar des SSTC, associée à un éditeur de structures (bilingues cette fois), où il était possible d'afficher une biphrase avec deux structures d'analyse et d'en modifier les paramètres (arbres, liens entre mots et liens entre arbres). Cette structure très adaptée aux approches de TA à base d'exemples utilisant l'analyse syntaxique a fait son apparition dans notre approche de manière assez naturelle. Nous observons à la figure 4.2, page 124, un schéma expressif de ces structures en S-SSTC qui forment notre mémoire d'exemples et sur lesquelles nous travaillerons.

Nous donnons ici la définition d'une S-SSTC, qui pourra sembler légèrement différente de celle d'origine aux lecteurs familiers de la structure. Nous donnons après celle-ci les raisons de ces différences qui ne seront en fait qu'apparentes.

Définition 20. *On se donne $S = (st^S, tr^S, co^S)$ et $C = (st^C, tr^C, co^C)$ deux SSTC. L'ensemble des nœuds de l'arbre source tr^S sera noté V et l'ensemble des nœuds de l'arbre cible tr^C sera noté W. On se donne également $(\ell_{node}, \ell_{tree})$ un couple d'**alignements** sur (V,W) donnant deux correspondances interlingues pour lesquelles on notera $\ell_{node} = (\mathscr{V}_{node}, \mathscr{W}_{node}, \sigma_{node})$ et $\ell_{tree} = (\mathscr{V}_{tree}, \mathscr{W}_{tree}, \sigma_{tree})$.*

- *ℓ_{node} est une correspondance entre V et W les nœuds des deux arbres source et cible. On parlera de **correspondace lexicale**. On notera pour les partitions $\mathscr{V}_{node} = \{V^1_{node}, ..., V^p_{node}\}$ et $\mathscr{W}_{node} = \{W^1_{node}, ..., W^p_{node}\}$.*
  Pour $i = 1, ..., p$, on pose $\sigma_{node}(V^i_{node}) = W^i_{node}$ (quittes à réindexer une partition).

- *ℓ_{tree} est une correspondance entre V et W (vu comme les racines de sous-arbres inclus dans les deux arbres source et cible). On parlera de **correspondance structurelle**. On notera les partitions $\mathscr{V}_{tree} = \{V^1_{tree}, ..., V^q_{tree}\}$ et $\mathscr{W}_{tree} = \{W^1_{tree}, ..., W^q_{tree}\}$.*
  Pour $i = 1, ..., q$, on pose $\sigma_{tree}(V^i_{tree}) = W^i_{tree}$ (quittes à réindexer une partition)

*Le quadruplet $(S, T, \ell_{node}, \ell_{tree})$ est une **Synchronous Structured String-Tree Correspondance** (S-SSTC) si les conditions d'homogénéité sont vérifiées :*

1. ***appartenance** : Soient $i \in \{1, ..., p\}$ et $j \in \{1, ..., q\}$:*

$$SNODE^S(V^i_{node}) \subset STREE^S(V^j_{tree}) \Leftrightarrow SNODE^C(W^i_{node}) \subset STREE^C(W^j_{tree})$$

   Les correspondances lexicales s'inscrivent dans les correspondances structurelles.

2. ***inclusion** : Soient $i, j \in \{1, ..., q\}$:*

$$STREE^S(V^i_{tree}) \subset STREE^S(V^j_{tree}) \Leftrightarrow STREE^C(W^i_{tree}) \subset STREE^C(W^j_{tree})$$

   Les correspondances structurelles respectent les paternités des deux structures.

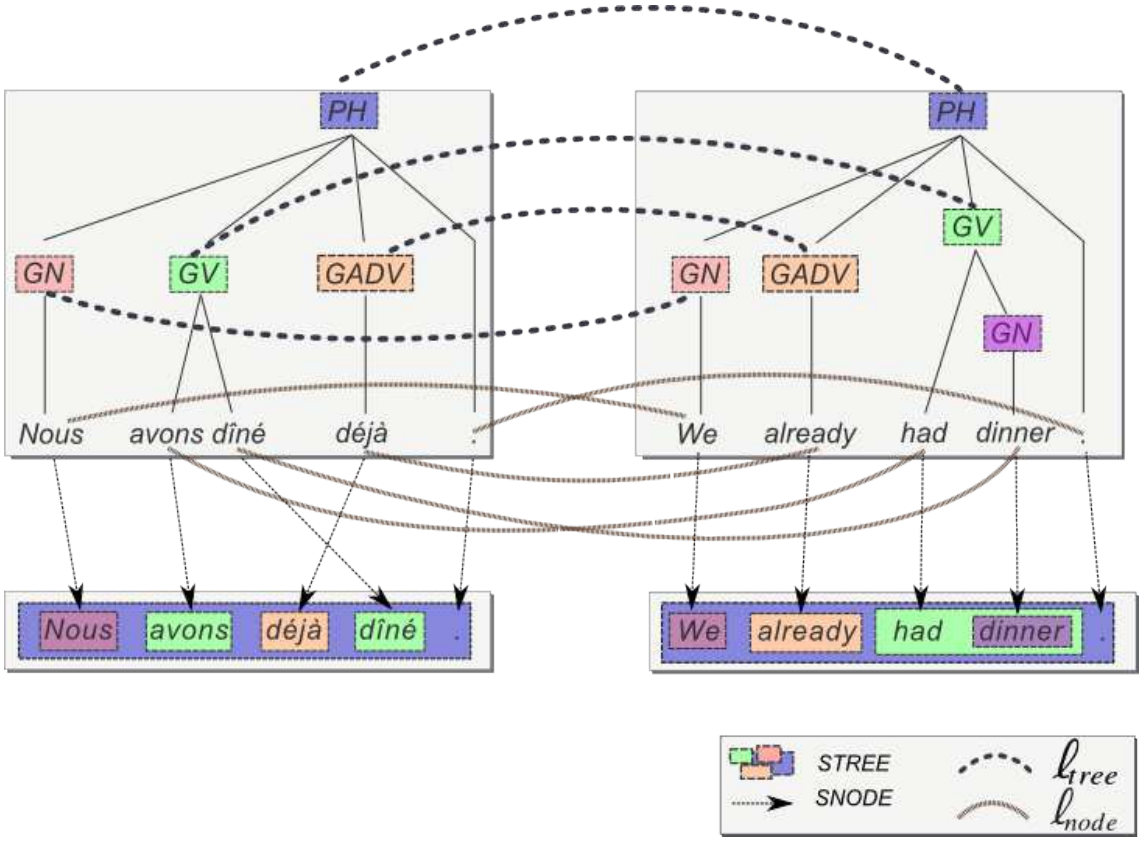


Figure 4.2 – Exemple de S-SSTC

3. ***globalité*** *: Les deux nœuds racines source et cible sont liés par* ℓ_{tree}.

Remarque 9. *On remarque que cette définition a été modifiée par rapport à la définition d'origine car nous avons retiré deux contraintes de la définition originale : la contrainte d'**unicité** n'apparaît plus. Elle exprimait le fait que chaque unité ne doive intervenir que dans une seul correspondance. Cette condition est en réalité cachée par le fait que nous ayons utilisé des **alignements** (selon définition en section 2.2.2) au lieu de **groupes de liens** pour définir les correspondances interlingues. Le fait que les alignements offrent une correspondance entre des sous-partitions garantit le fait qu'un nœud n'interviendra que dans un groupement de liens. Nous avons également ommis une correspondance entre sous-arbres, que nous appelleront* $\ell_{subtree}$ *et qui est structurellement assez différente des deux correspondances ici puisqu'elle est définie de l'ensemble des sous-arbres*

de tr^S dans l'ensemble des sous-arbres tr^C (le terme "sous-arbre" désignait un type particulier de sous-arbre : un arbre auquel on aurait ôté toute la descendance d'un nœud). On perd donc la représentation d'un arbre par sa racine et la condition d'unicité. La définition d'origine devait vérifier une contrainte de domination pour $\ell_{subtree}$ similaire à la contrainte d'inclusion pour ℓ_{tree}. Dans notre approche nous n'utiliserons pas de correspondance entre sous-arbres, estimant qu'une correspondance entre "ensembles de nœuds" est suffisamment expressive.

Les S-SSTC, par leur correspondance entre structures ℓ_{tree}, permettent de lier des ensembles de nœuds. Il nous sera souvent utile, dans cette partie, de parler des ascendants/descendants d'un ensemble de nœuds internes d'un arbre. Nous introduisons donc les notations suivantes qui généralisent les notions d'ascendant/descendant d'un nœud.

Notation. *Soient V, V' deux ensembles de nœuds d'un même arbre, on définit les notions suivantes :*

- *V et V' sont indépendants si pour tous $v \in V, v' \in V'$, v et v' sont indépendants*
- *$V' \Downarrow V$ si pour tous $v \in V, v' \in V'$, $v' \downarrow v$ ou v indépendant de v' et il existe deux nœuds en relation de descendance.*
- *$V' \Uparrow V$ si pour tous $v \in V, v' \in V'$, $v' \uparrow v$ ou v indépendant de v' et il existe deux nœuds en relation d'ascendance.*

Nous proposons d'ajouter deux notions supplémentaires sur les S-SSTC, à savoir de *S-SSTC propre* et de *S-SSTC dominante*. À l'instar des notions de *complétude* et de *dominance* pour les SSTC, il s'agira de propriétés de bonne formation attendues des cas pratiques et que la définition générale ne suffit pas à assurer. Elles pourraient sembler être des contraintes qui font perdre en généralité la définition, mais en fait, seulement sur des structures linguistiquement peu pertinentes (voir figures 4.3, 4.4).

Les correspondances structurelles d'une S-SSTC *propre* incluent les correspondances lexicales. Nous estimons que cette contrainte supplémentaire sert le modèle et ajoute en pertinence. La notion de *S-SSTC dominante* est proche des considérations sur les arbres

partiels dans la définition d'origine ; elle est toutefois un peu plus permissive, car elle autorise une mise en correspondance de listes de sous-arbres. Elle permet d'empêcher des formes dégénérées telles que celle obervée en figure 4.4, qui présentent peu d'intérêt mais sont permises par la définition générale.

Définition 21. *On dira que la S-SSTC est* ***propre*** *si pour tous* $i \in [\![1,p]\!]$ *et* $j \in [\![1,q]\!]$ *:*

$$\begin{cases} SNODE^S(V^i_{node}) \cap STREE^S(V^j_{tree}) \neq \emptyset & \Rightarrow \quad SNODE^S(V^i node) \subset STREE^S(V^j_{tree}) \\ SNODE^C(W^i_{node}) \cap STREE^C(W^j_{tree}) \neq \emptyset & \Rightarrow \quad SNODE^C(W^i node) \subset STREE^C(W^j_{tree}) \end{cases}$$

La figure 4.3 représente une S-SSTC correcte mais impropre. On y observe des correspondances lexicales ℓ_{node} qui ne sont pas dominées par les correspondances lexicales ℓ_{tree}, mais seulement en "décalage".

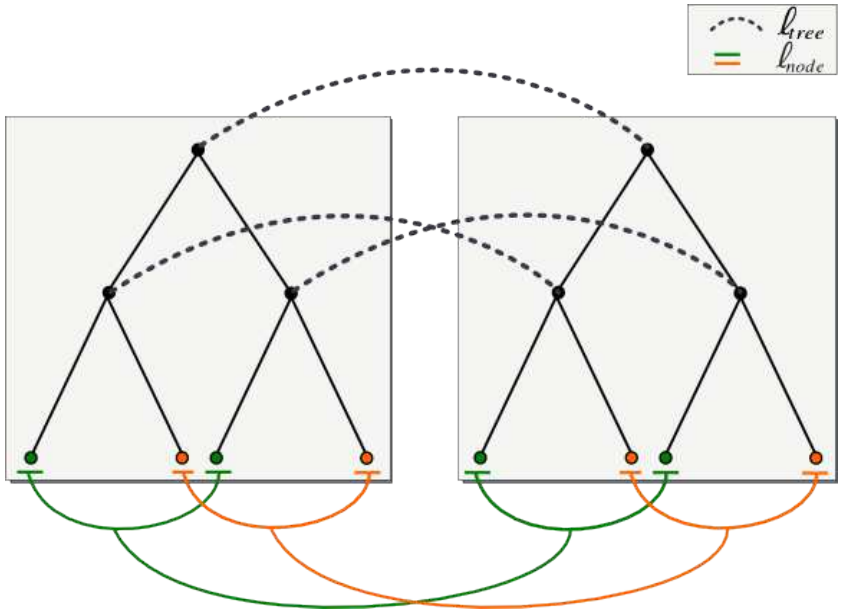


Figure 4.3 – Un exemple de S-SSTC impropre

Définition 22. *On dira que la S-SSTC est* ***dominante*** *si pour tous* $i, j \in \{1,...,q\}$ *:*

$$\begin{cases} V^i_{tree} \text{ et } V^j_{tree} \text{ indépendants} & ou \quad V^i_{tree} \Downarrow V^j_{tree} & ou \quad V^i_{tree} \Uparrow V^j_{tree} \\ W^i_{tree} \text{ et } W^j_{tree} \text{ indépendants} & ou \quad W^i_{tree} \Downarrow W^j_{tree} & ou \quad W^i_{tree} \Uparrow W^j_{tree} \end{cases}$$

La figure 4.4 représente sur un exemple une S-SSTC qui est correcte selon la définition, mais n'est en aucun cas dominante. On y observe un entrelacement des correspondances structurelles qui n'a guère de signification linguistique.

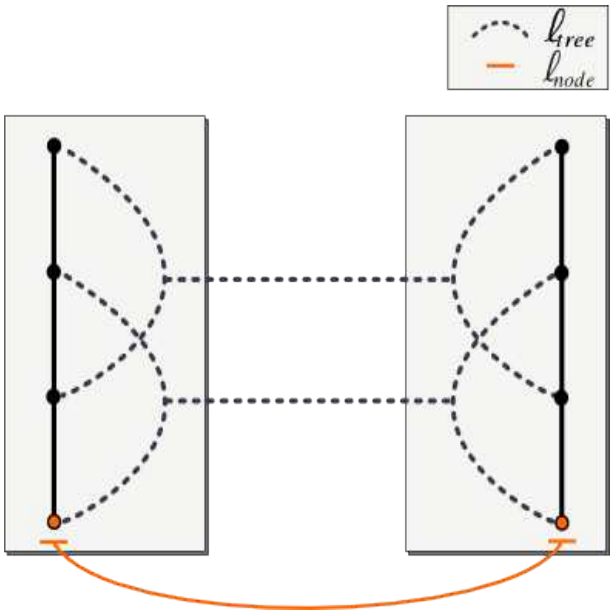


Figure 4.4 – Un exemple de S-SSTC non dominante

Nous proposons pour les deux contre-exemples des figures 4.3 et 4.4, des alternatives respectivement propre et dominante à la figure 4.5

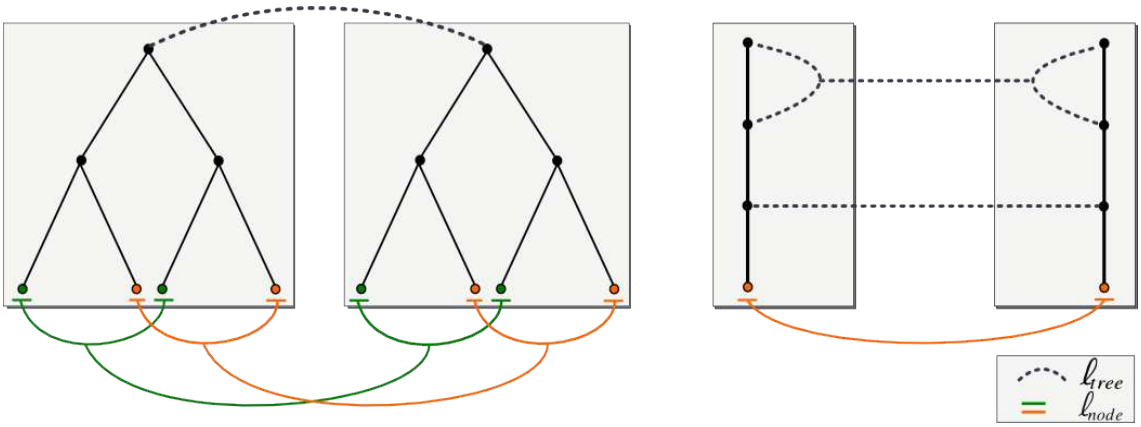


Figure 4.5 – Alternatives propre et dominante

Une S-SSTC vérifiant ces deux contraintes vérifiera les conditions de la propriété 5 qui impose une structuration hiérarchique forte des liens de la correspondance structurelle ℓ_{tree}.

Propriété 5. *Si la S-SSTC est propre et dominante et lie deux SSTC terminales et non chevauchantes, alors pour tous $V, V' \in \mathcal{V}_{tree}$, en posant $W = \sigma_{tree}(V), W' = \sigma_{tree}(V')$,*

$$\begin{cases} V \Downarrow V' & \Leftrightarrow & W \Downarrow W' \\ V \Uparrow V' & \Leftrightarrow & W \Uparrow W' \\ V \text{ et } V' \text{ indépendants} & \Leftrightarrow & W \text{ et } W' \text{ indépendants} \end{cases}$$

On définit formellement l'*inversion de dominance* introduite dans [2] qui traduit un phénomène complexe attestant de la grande expressivité des S-SSTC. Ce phénomène est propre à des arbres de représentation en dépendance où les mots de la phrase peuvent correspondre à des nœuds internes de l'arbre. Le phénomène se produit lorsqu'un terme dominant dans l'arbre source devient le dominé dans l'arbre cible. On peut l'observer sur un exemple à la figure 4.6 pour la biphrase *"Il monte la rue en courant"/"He runs up the street"* où les termes *"monte", "en courant", "runs"* et *"up"* participent à une inversion de dominance.

Définition 23. *Soient s et $s' \in st^S$ deux mots de la phrase source. On définit la relation binaire $\propto_s$ sur st^S entre s et s'. On dira que s et s'* ***participent à une inversion de dominance****, noté $s \propto_s s'$, si il existe $v, v' \in V^S$ et $w, w' \in V^C$ tels que :*

$$\begin{cases} (v \downarrow v' \text{ et } w \uparrow w') \text{ ou } (v \uparrow v' \text{ et } w \downarrow w') \\ v \text{ et } w \text{ sont liés par } \ell_{node} \\ v' \text{ et } w' \text{ sont liés par } \ell_{node} \\ s \in snode^s(v) \text{ et } s' \in snode^s(v') \\ snode^c(w) \neq \emptyset \text{ et } snode^c(w') \neq \emptyset \end{cases}$$

On pourra définir de même $\propto_c$ sur st^C.

Remarque 10. *Les deux relations binaires d'inversion de dominance $\propto_s$ et $\propto_c$ ne sont* ***pas*** *des relations d'équivalence sur leur ensemble respectif.*

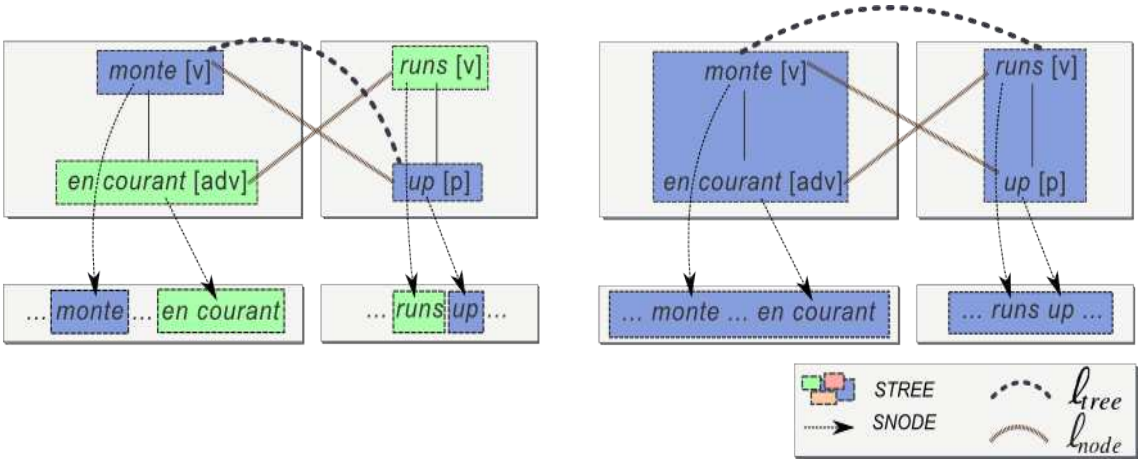


Figure 4.6 – Cas d'inversion de dominance

On observe à la figure 4.6 deux S-SSTC valides, mais différentes, représentant le même phénomène d'inversion de dominance. Dans celle de gauche, on constate que la correspondance structurelle ℓ_{tree} n'exprime que partiellement l'inversion de dominance puisqu'elle ignore la correspondance *"en courant"-"runs"*. Celle de gauche est plus proche de ce que l'on attend de la structure, c'est-à-dire inclure l'inversion de dominance dans la correspondance structurelle ℓ_{tree}. Les axiomes d'une S-SSTC sont trop permissifs et admettent des structures non souhaitées. Il n'y a donc pas de représentation canonique. En formulant une hypothèse supplémentaire de bonne formation, nous pouvons contraindre les S-SSTC à respecter le phénomène d'inversion de dominance. C'est ce qu'exprime la propriété suivante que nous prouvons ensuite.

Propriété 6. *On suppose la S-SSTC propre et dominante liant deux SSTC non chevauchantes. Soient $s, s' \in st^S$ tels que $s \propto_s s'$. On définit $v, v' \in V^S$ et $w, w' \in V^C$ les nœuds associés selon la définition 23. On a alors l'implication :*

$$\forall V, V' \in \mathscr{V}_{tree}, v \in V \text{ et } v' \in V' \Rightarrow V = V'$$

Démonstration. Soient $V, V' \in \mathscr{V}_{tree}$ et $W, W' \in \mathscr{W}_{tree}$ tels que $v \in V$, $v' \in V'$, $w \in W$ et $w' \in W'$. Les paires de nœuds v, w et v', w' étant respectivement liées par ℓ_{node}, on pose

$V_0, V_0' \in \mathscr{V}_{node}$ et $W_0, W_0' \in \mathscr{W}_{node}$ tels que $v \in V_0$, $v' \in V_0'$, $w \in W_0$ et $w' \in W_0'$.
Si $v = v'$, on conclue.
Sinon, on suppose, quitte à inverser, $v \downarrow v'$. Comme $w' \downarrow w$, alors on a :

$$
\begin{array}{lrcll}
 & snode^c(w') & \subset & stree^c(w) & \\
\text{de plus,} & snode^c(w') & \neq & \emptyset & (\text{définition } \propto_s) \\
\Rightarrow & SNODE^C(W_0') & \cap & STREE^C(W) \neq \emptyset & \\
\Rightarrow & SNODE^C(W_0') & \subset & STREE^C(W) & (\text{S-SSTC propre}) \\
\Rightarrow & SNODE^S(V_0') & \subset & STREE^S(V) & (\text{appartenance}) \\
\Rightarrow & snode^s(v') & \subset & STREE^S(V) & (\text{simple inclusion})
\end{array}
$$

En utilisant l'hypothèse de non-chevauchement pour la SSTC source, il existe $\tilde{v} \in V$ ascendant de v', ce qui contredit l'hypothèse de dominance car $\underbrace{\tilde{v}}_{\in V} \uparrow \underbrace{v'}_{\in V'} \uparrow \underbrace{v}_{\in V}$ □

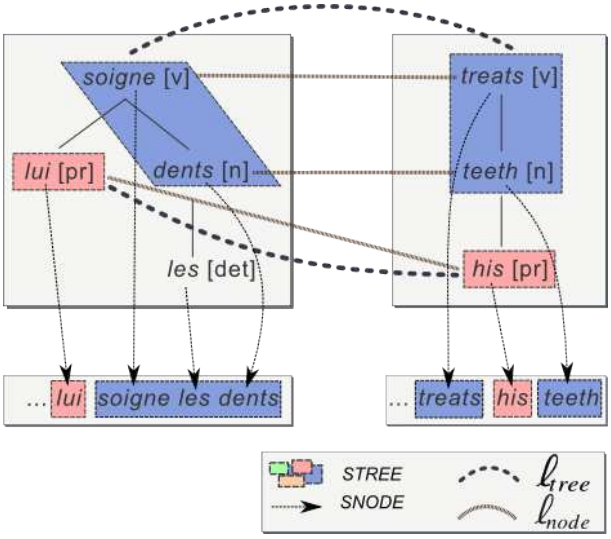


Figure 4.7 – L'inversion de dominance dans *"Le docteur lui soigne les dents"/"The doctor treats his teeth"*

Remarque 11. *Dans [2], l'inversion de dominance est introduite avec l'**élimination de dominance**. Ce second phénomène considère des mots indépendants dans une structure source qui sont liés à des mots en relation d'ascendance dans la structure cible. Nous*

reprenons l'exemple l'introduisant sur la biphrase "Le docteur lui soigne les dents"/"The doctor treats his teeth" à la figure 4.7. Ce deuxième phénomène, moins divergent, ne vérifiera pas nécessairement une propriété analogue : sur l'exemple de la figure 4.7 on constate que les mots "lui" et "dents" qui sont indépendants dans l'arbre français, sont alignés via ℓ_{node} à "his" et "teeth" et participent bien à une élimination de dominance. Pourtant, ils ne sont pas pour autant réunis dans une correspondance structurelle ℓ_{tree} et l'on pourra vérifier que la structure définit bien une S-SSTC dominante et propre.

4.2 Les structures bilingues dans Align^{It}

Nous avons souhaité faire de Align^{It} un outil multilingue. Une langue, pour y être intégrée, doit disposer au minimum d'un étiqueteur morphosyntaxique, idéalement d'un analyseur syntaxique plus profond [2]. Nous allons voir qu'au travers de quelques traitements automatisés, la structure sous-jacente à l'interface d'acquisition est équivalente à une *S-SSTC normale*.

Certains analyseurs n'offraient pas toujours une représentation projective et le cas échéant, une correspondance entre les positions des mots de la phrase et les nœuds de l'arbre n'est pas toujours donnée. Or, elle était nécessaire à notre représentation par une SSTC afin d'encoder *snode*. L'ajout d'un analyseur non projectif était accompagné par l'introduction d'heuristiques visant à calculer cette correspondance. Il est heureusement possible de vérifier automatiquement si une correspondance *snode* proposée est correcte et forme bien une SSTC complète. Les phrases analysées pour lesquelles les heuristiques ne fonctionnaient pas ont été écartées des corpus, mais il s'agissait de moins de $0,01\%$ des cas[3]. Lorsque l'analyseur était projectif, la correspondance *snode* était trivale. Dans chaque langue, pour chaque couple phrase/analyse, la propriété 3 nous fournit un calcul explicite de la correspondance $stree$ à partie de *snode*. Ainsi, les corpus découpés en phrases ont été analysés et stockés sous forme de *SSTC complètes non chevauchantes*.

[2]Les analyses sont nécessaires pour la partie automatique. Une segmentation par défaut, utilisant le blanc typographique, suffit pour l'intégration à l'interface d'alignement manuel.

[3]Des cas de SSTC non complètes ont été détectés avec notre analyseur non projectif sur français utilisé sur plus de 65000 phrases.

Reste alors la représentation bilingue des biphrases alignées. La structure d'arbre n'apparaît donc pas directement et les liens lexicaux représentés par ℓ_{node} ne sont pas éditables. Nous rappelons que les correspondances ℓ_{node} des S-SSTC relient les nœuds de deux structures, tandis que l'interface d'acquisition des alignements d'Align$^{\text{It}}$ propose de placer des liens entre les mots des phrases selon les segmentations données par ces analyses. Les alignements récoltés sont donc adaptés pour fournir aux doubles structures une correspondance ℓ_{node} correcte. Comme on le voit sur la figure 4.2, le calcul se fait au travers des correspondances chaîne/arbre *snode* des deux SSTC source et cible. La structure résultante est déjà une S-SSTC mais pour laquelle la synchronisation structurelle ℓ_{tree} serait l'alignement grossier $\ell_{\mathbb{O}}$[4].

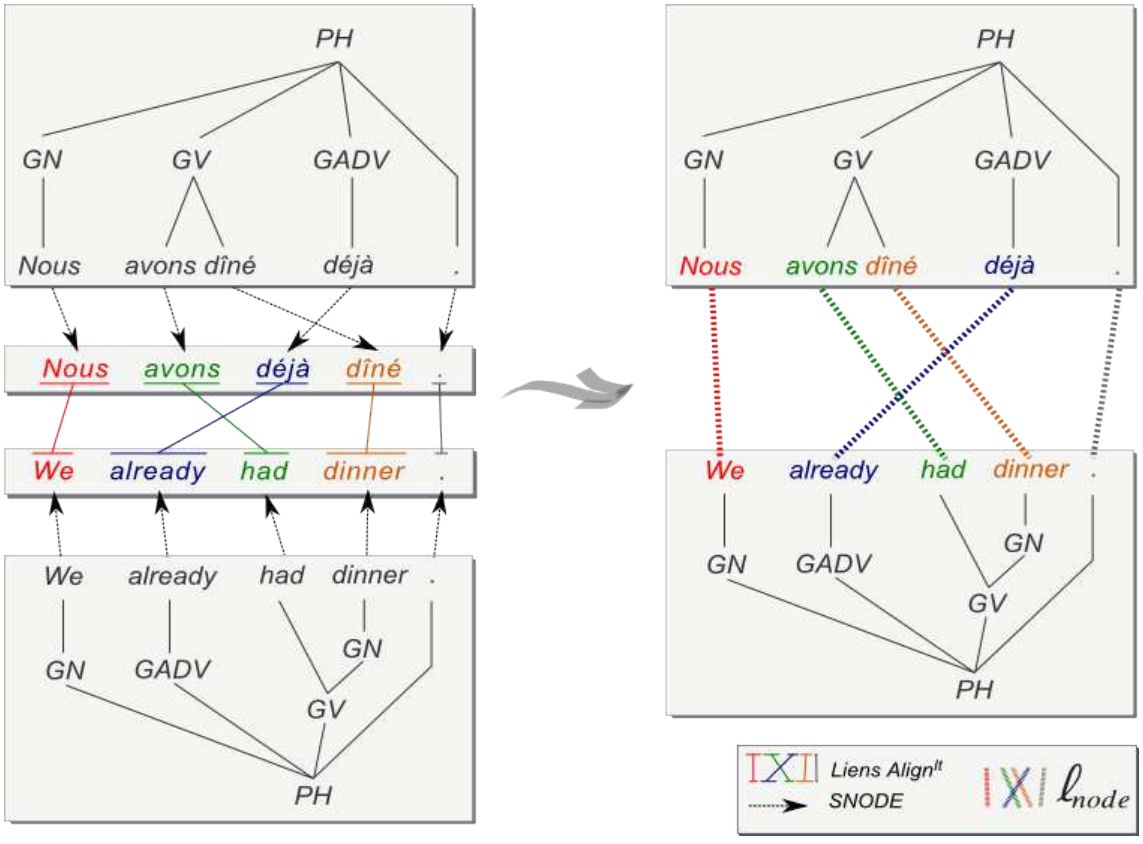


Figure 4.8 – Construction des liens ℓ_{node}

[4]Pour vérifier l'hypothèse de globalité, il conviendrait plutôt de supposer un alignement grossier $\ell_{\mathbb{1}}$ qui n'apporte cependant pas plus d'information

Nous ne souhaitons pas alourdir l'interface d'acquisition par un affichage de structures d'arbre (notamment lorsque les phrases sont longues) et éditables et afin de déterminer une correspondance structurelle ℓ_{tree}. En pratique, nous avons eu recours à un calcul permettant d'estimer une correspondance ℓ_{tree} par remontée au travers des arbres source et cible, pour segmenter par les syntagmes (voir section 5.1.2). Il faut noter qu'il s'agissait d'un cadre simplifié car l'une des deux structures opposées était plate (voir les structures représentées à la figure 6.6 page 183 en partie 6.1.2, présentant les analyseurs). La correspondance ℓ_{tree} était donc facilement déductible de ℓ_{node} et de l'arbre profond. En raisonnant par "remontée", il est clair que la S-SSTC obtenue sera *propre* et *dominante*. Étendre la recherche d'une correspondance ℓ_{tree} en situation générale est une problématique intéressante qui mériterait d'être approfondie. La déduction d'un alignement structurel à partir d'un alignement lexical semble apparenté au problème de la construction d'*arbre de concensus* en bio-informatique [64].

La représentation en S-SSTC a été introduite dans le cadre d'une problématique de parsing à base d'exemples [3] dans laquelle l'auteur construit des structures syntaxiques cible en partant d'une analyse de la phrase source et d'un alignement entre les deux phrases. Les liens proviennent d'outils d'alignement automatiques. Dans la solution apportée, l'arbre produit est en quelque sorte l'arbre source transformé en passant au travers des liens bilingues. Il fait l'hypothèse d'une intrication forte entre les trois éléments principaux qui constituent le modèle des S-SSTC : la structure source, les liens bilingues et la structure cible. Notre approche se base sur cette hypothèse pour faire une utilisation des S-SSTC qui pourrait sembler à contre-emploi : en partant de deux structures syntaxiques provenant d'outils automatiques, nous souhaitons déduire un ensemble de liens.

CHAPITRE 5

UNE MÉMOIRE DE FRAGMENTS POUR L'ALIGNEMENT À BASE D'EXEMPLES

Cet chapitre présente en détail deux composantes principales d'un outil automatique d'alignement à base d'exemples, à savoir une mémoire de fragments et des algorithmes de résolution. Nous commençons par aborder le thème de la fragmentation en partie 5.1, étape primordiale, qui permet de constituer la mémoire d'alignements. On y trouvera une présentation formelle des fragments (section 5.1.1) ainsi que les différentes opérations les concernant. Nous verrons en section 5.1.2, qu'il n'y a pas une manière canonique pour fragmenter nos exemples et présentons les différentes solutions envisagées. Nous étudierons quantitativement les conséquences d'une fragmentation trop "générale" en section 5.1.3. La deuxième partie 5.2 modélisera le problème de l'alignement à partir des bases de fragments et exposera les conséquences des choix de fragmentation en terme de complexité (section 5.2.2).

5.1 Constitution d'une mémoire d'exemples

5.1.1 Fragments formels

On se donne deux *phrases* (deux listes de symboles[1]), S de longueur n et C de longueur m. Nous rappelons qu'un alignement ℓ sur une biphrase $B = (S,C)$ est la donnée de 3 briques de base : $\ell = (\mathcal{V}, \mathcal{W}, \sigma)$ où $\mathcal{V}$ est une sous-partition de $\{1,\ldots,n\}$, $\mathcal{W}$ est une sous-partition de $\{1,\ldots,m\}$, et σ est une permutation de $\mathcal{V}$ dans $\mathcal{W}$. Si nous souhaitons définir une fragmentation, il est important que cette notion respecte les interdépendances de ces trois éléments. Elle ne devrait donc ni séparer des groupements de mots faits par les 2 sous-partitions, ni être en contradiction avec les correspondances données par σ. On note $\mathcal{A}(S,C)$ (où $\mathcal{A}(n,m)$) l'ensemble des alignements de la biphrase B.

[1] Les symboles pouvant être de différentes natures. Dans notre cadre applicatif, il s'agira d'informations morphosyntaxiques issues d'outils automatiques d'analyse de différentes langues.

Dans cette partie, nous reprendrons les notations suivantes sans le rappeler systématiquement :

$$\begin{array}{rcll}
\mathscr{l} & = & (\mathscr{V},\mathscr{W},\sigma) & \text{est de longueur } (n,m) \text{ sur } B=(S,C)\\
\mathscr{l}' & = & (\mathscr{V}',\mathscr{W}',\sigma') & \text{est de longueur } (n',m') \text{ sur } B'=(S',C')\\
\mathscr{l}'' & = & (\mathscr{V}'',\mathscr{W}'',\sigma'') & \text{est de longueur } (n'',m'') \text{ sur } B''=(S'',C'')\\
\hat{\mathscr{l}} & = & (\hat{\mathscr{V}},\hat{\mathscr{W}},\hat{\sigma}) & \text{est de longueur } (\hat{n},\hat{m}) \text{ sur } \hat{B}=(\hat{S},\hat{C})\\
\bar{\mathscr{l}} & = & (\bar{\mathscr{V}},\bar{\mathscr{W}},\bar{\sigma}) & \text{est de longueur } (\bar{n},\bar{m}) \text{ sur } \bar{B}=(\bar{S},\bar{C})\\
\tilde{\mathscr{l}} & = & (\tilde{\mathscr{V}},\tilde{\mathscr{W}},\tilde{\sigma}) & \text{est de longueur } (\tilde{n},\tilde{m}) \text{ sur } \tilde{B}=(\tilde{S},\tilde{C})
\end{array}$$

Et de même avec les indices numériques.

Tout d'abord, on définit la notion de *sous-biphrase* qui correspond à une biphrase s'obtenant, à partie d'une autre, par suppression de mots source et cible :

Définition 24. *Soient $B'=(S',C')$ et $B=(S,C)$ deux biphrases. B' sera une **sous-biphrase** de B si il existe deux injections croissantes $\varphi:[\![1,n']\!]\mapsto[\![1,n]\!]$ et $\psi:[\![1,m']\!]\mapsto[\![1,m]\!]$ telles que :*

$$\begin{cases} S'=\left(S_{\varphi(k)}\right)_{k\in[\![1,n']\!]} \\ C'=\left(C_{\psi(k)}\right)_{k\in[\![1,m']\!]} \end{cases}$$

*La condition exprime le fait que S' et C' sont formées de sous-suites de mots de S et C. Nous ferons référence aux fonctions φ et ψ comme des **positionnements** source et cible de B' dans B.*

Remarque 12. *Il n'y a pas forcément unicité des fonctions φ et ψ. Les cas multiples correspondent à des situations d'ambiguïté, évoquées en section 3.2.3.3.*

Pour notre approche à base d'exemples, nous souhaitons réutiliser des alignements "plus petits" dans de nouvelles situations. Ce sont en fait des alignements sur des sous-biphrases qui nous intéressent, pouvant d'être *plongées* dans une nouvelle biphrase plus longue. Nous définissons pour cela les deux opérations naturelles de *plongement* et de *projection*.

Définition 25. *On suppose que B' est une sous-biphrase B positionnée par φ et ψ. Soit ℓ' un alignement sur B'. On définit le* ***plongement*** *de ℓ' dans B selon φ et ψ, que l'on note :*

$$\ell = \begin{bmatrix} \varphi \\ \psi \end{bmatrix} (\ell')$$

Le plongement $\begin{bmatrix} \varphi \\ \psi \end{bmatrix}$ *est défini de $\mathscr{A}(B')$ dans $\mathscr{A}(B)$ et se calcule de la manière suivante :*

$$\begin{cases} \mathscr{V} & = & \varphi(\mathscr{V}') \\ \mathscr{W} & = & \psi(\mathscr{W}') \\ \sigma(\varphi(V')) & = & \sigma'(V'), \textit{pour tout } V' \in \mathscr{V}' \end{cases}$$

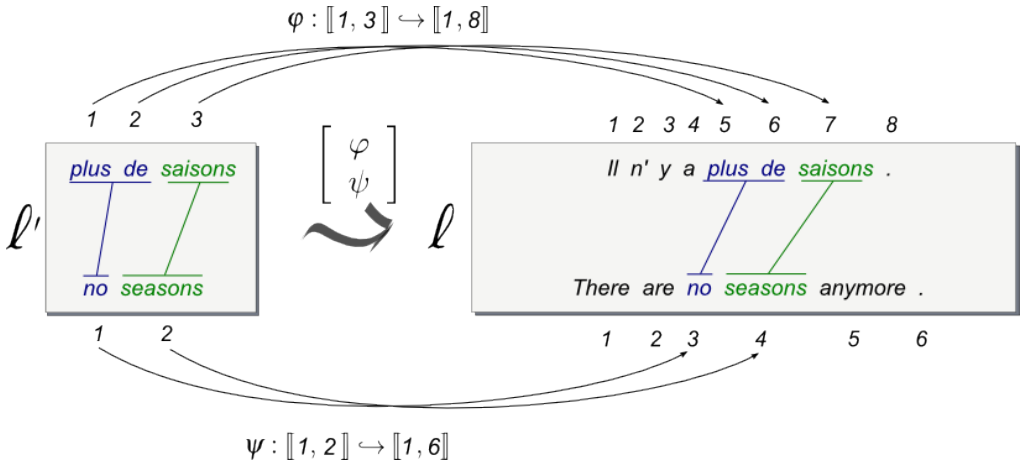


Figure 5.1 – Le plongement de ℓ' dans B forme ℓ

L'opération de plongement permet d'aligner une biphrase vierge en y *plongeant* une sous-biphrase alignée. Préciser les positionnement φ et ψ (voir figure 5.1) permet de rendre le calcul explicite. De plus, il n'est pas exclu qu'une sous-biphrase alignée admette différents plongements relativement à une biphrase plus longue. Ce sera le cas lorsqu'il y a répétition du motif de la sous-biphrase dans la biphrase considérée.

Définition 26. *On reprend les mêmes notations avec B' une sous-biphrase de B. Soit ℓ un alignement sur B. On définit le* ***projeté*** *de ℓ dans B' selon φ et ψ, que l'on note :*

$$\ell' = \Pi_{\varphi,\psi}\left(\ell\right)$$

La projection $\Pi_{\varphi,\psi}$ est définie de $\mathscr{A}(B)$ dans $\mathscr{A}(B')$ et se calcule de la manière suivante :

$$\begin{cases} \mathscr{V}' & = & \left\{\varphi^{-1}(V) \neq \emptyset \text{ tel que } \psi^{-1}(\sigma(V)) \neq \emptyset\right\} \\ \mathscr{W}' & = & \left\{\psi^{-1}(W) \neq \emptyset \text{ tel que } \sigma'(\varphi^{-1}(W)) \neq \emptyset\right\} \\ \sigma'(\varphi^{-1}(V)) & = & \psi^{-1}(\sigma(V)), \text{ pour tout } V \in \mathscr{V} \text{ tel que } \varphi^{-1}(V) \in \mathscr{V}' \end{cases}$$

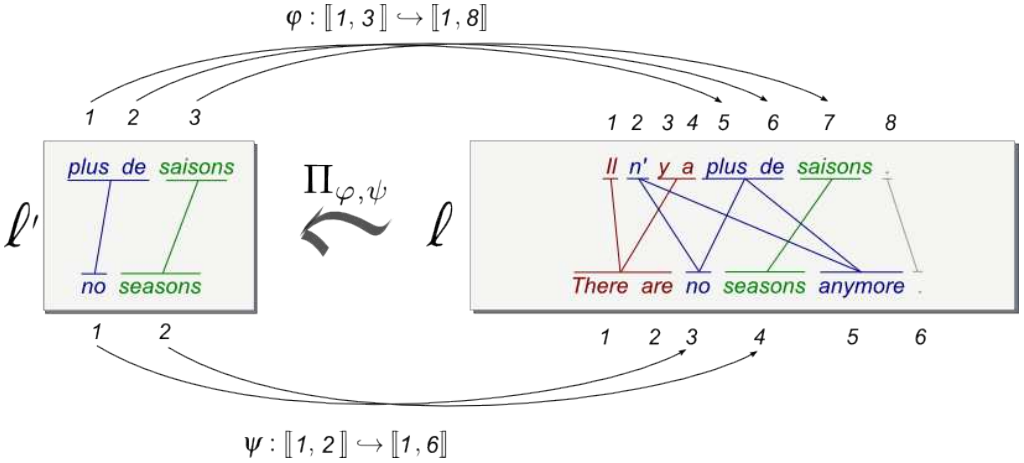


Figure 5.2 – La projection de ℓ dans B' forme ℓ'

Nous définissons la notion de *sous-alignement* qui à l'instar de la sous-biphrase, s'obtient par suppression de mots sur un alignement de départ (voir figure 5.3). On le définit à l'aide de la relation de finesse définie en partie (voir 2.2.2.2) :

Définition 27. *On dira que ℓ' est un sous-alignement de ℓ, positionné par φ et ψ, si les conditions suivantes sont vérifiées :*

- *B' est une sous-biphrase de B positionnée par φ et ψ*
- $\left[\begin{array}{c}\varphi\\ \psi\end{array}\right]\left(\ell'\right) \prec \ell$

Nous noterons $\mathscr{A}_\ell$ l'ensemble des sous-alignements de ℓ.

Notation. *Il sera utile de manipuler des sous-alignements de ℓ conjointement à leur positionnement dans la biphrase B, c'est-à-dire un triplet (ℓ', φ, ψ) que l'on notera aussi $\ell'^{\varphi,\psi}$. Nous appellerons ces éléments des* ***alignements positionnés*** *et noterons $\mathscr{A}_\ell{}^{\Phi\Psi}$ l'****ensemble des alignements positionnés*** *dans ℓ.*

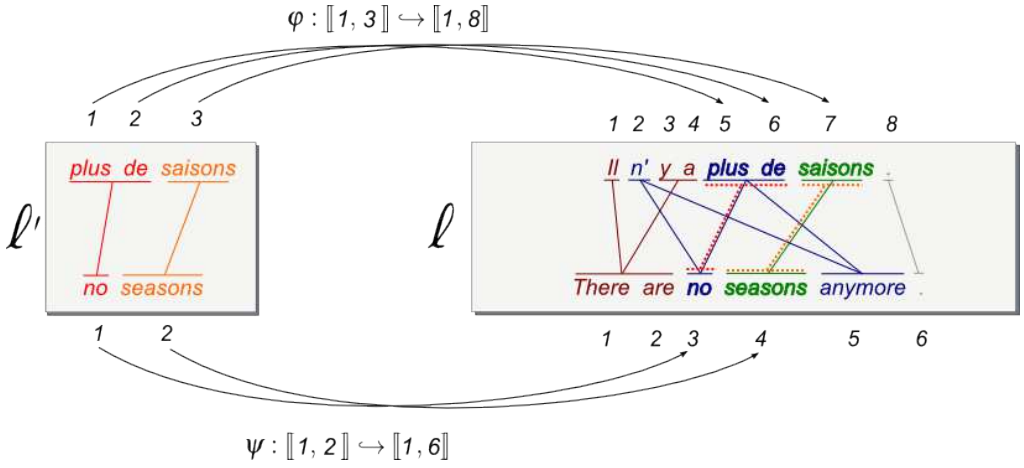


Figure 5.3 – ℓ' est un sous-alignement de ℓ

L'étape ② du processus décrit en partie 3.2.2 correspond à la fragmentation d'un exemple validé par l'annotateur. La notion de sous-alignement semble correspondre, mais reste un petit peu trop générale. En effet, il vaut mieux éviter de séparer les éléments d'un groupe lié. Les *fragments* d'un alignement formeront un sous-ensemble strict des sous-alignements $\mathscr{A}_\ell$ (voir figure 5.4 page 140).

Définition 28. *On dira que ℓ' est un* ***fragment*** *de l'alignement ℓ (noté $\ell' \triangleleft \ell$) si B′ est une sous-biphrase de B positionnée par φ et ψ vérifiant de plus :*

- $\varphi(\mathscr{V}') \subset \mathscr{V}$
- $\psi(\mathscr{W}') \subset \mathscr{W}$
- $\forall V' \in \mathscr{V}', \psi \circ \sigma'(V') = \sigma \circ \varphi(V')$

La première et la deuxième conditions expriment le fait que les fragments source et cible de ℓ' sont inclus dans ceux de ℓ (respect des frontières). La troisième impose que les liens interlingues soient respectés. Nous noterons $\mathfrak{A}_\ell$ l'ensemble des fragments de ℓ

La mémoire d'exemples formée par l'étape ② (schéma 3.16 page 103) contiendra des alignements qui seront en fait des *fragments* d'alignements validés par les annotateurs. Envisager une mémoire constituée de *sous-alignements* est une idée que nous avons écarté d'une part pour une simple raison d'économie car les sous-alignements sont bien plus nombreux que les fragments. D'autre part car un sous-alignement, en excluant des termes des groupes source et cible liés, est plus à même de créer des alignements partiels et potentiellement mauvais.

Propriété 7. *Nous avons les résultats suivant :*

- *Les fragments de ℓ sont des sous-alignements de ℓ, i.e. $\mathfrak{A}_\ell \subset \mathscr{A}_\ell$*
- *La relation "$\triangleleft$" signifiant "**être un fragment de**" est une relation d'ordre partiel sur l'ensemble des alignements $\mathscr{A}(S,C)$. Ses éléments maximaux sont les alignements couvrants. L'alignement vide $\ell_\emptyset$ est l'élément minimum.*

Notation. *Nous noterons de même que pour les alignements $\mathfrak{A}_\ell^{\Phi\Psi}$ les fragments positionnés dans ℓ.*

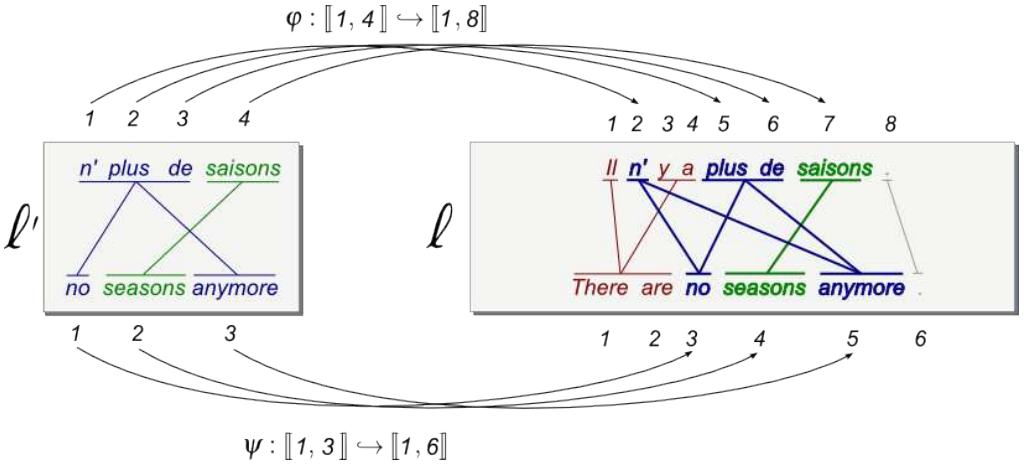


Figure 5.4 – ℓ' est un fragment de l'alignement ℓ ($\ell' \triangleleft \ell$)

L'étape ③ du schéma d'architecture 3.16 (page 103) consiste en deux étapes : extraire de la mémoire d'exemples des *fragments compatibles*, avec la nouvelle biphrase à aligner (éventuellement pré-alignée), puis les positionner dans celle-ci (de multiples

positionnements sont possibles en cas d'ambiguïté). Nous définissons donc ce que l'on entend par *fragment compatible* :

Définition 29. *On se donne deux alignements ℓ et ℓ'. L'alignement ℓ correspond à l'état de départ, et ℓ' un alignement déjà en mémoire. On dira que ℓ' est un* ***fragment compatible*** *avec ℓ si les conditions suivantes sont vérifiées :*

- *B' est une sous-biphrase de B positionnée par φ et ψ*
- $\Pi_{\varphi,\psi}\left(\ell\right) \prec \ell'$

*Nous noterons $\mathfrak{a}_\ell$ l'**ensemble des fragments compatibles** avec ℓ. Pour l'alignement vide $\ell_\emptyset$, nous parlerons de l'**ensemble des fragments compatibles avec la biphrase** $B = (S,C)$ que nous noterons de manière indifférenciée $\mathfrak{a}_{\ell_\emptyset}$, $\mathfrak{a}(B)$ ou $\mathfrak{a}(S,C)$.*

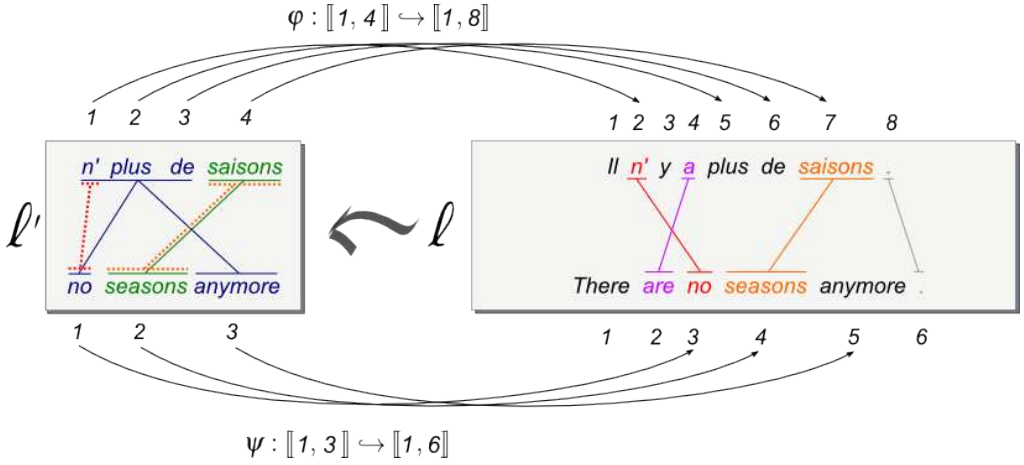


Figure 5.5 – ℓ' est un fragment compatible avec ℓ

Un fragment ℓ', sera dit *compatible* avec ℓ si localement, il étend cet alignement, comme on peut le voir sur l'exemple de la figure 5.5. On exprime cette localité en projetant ℓ dans la biphrase courte B'. Le fait d'"étendre" l'alignement est exprimé par le fait que l'alignement ℓ' est plus grossier (par rapport à la relation de finesse "$\prec$").

Notation. *Nous noterons $\mathfrak{a}_\ell{}^{\Phi\Psi}$ l'ensemble des fragments compatibles avec ℓ munis de leurs positionnements.*

Les fragments positionnés dans la nouvelle biphrase ont pour vocation d'être *composés* (étape ④ du schéma page 103) afin de créer des sous-alignements plus longs via les algorithmes de sélection décrits au chapitre suivant. Encore faut-il que les informations portées par chacun ne se contredisent pas pour être *composables*.

Définition 30. *Soient* $(\ell_1, \varphi_1, \psi_1)$ *et* $(\ell_2, \varphi_2, \psi_2) \in \mathfrak{A}_{\ell}^{\Phi\Psi}$ *(des fragments compatibles avec* ℓ*). Ils seront* ***composables*** *si les conditions suivantes sont vérifiées :*

- *Pour* $V_1 \in \mathscr{V}_1$ *et* $V_2 \in \mathscr{V}_2$, $\varphi_1(V_1)$ *et* $\varphi_2(V_2)$ *sont disjoints ou égaux.*
- *Pour* $W_1 \in \mathscr{W}_1$ *et* $W_2 \in \mathscr{W}_2$, $\psi_1(W_1)$ *et* $\psi_2(W_2)$ *sont disjoints ou égaux.*
- *Pour* $W_1 \in \mathscr{W}_1$ *et* $W_2 \in \mathscr{W}_2$, $\varphi_1(V_1) = \varphi_2(V_2)$ *ssi* $\psi_1(\sigma_1(V_1)) = \psi_2(\sigma_2(V_2))$

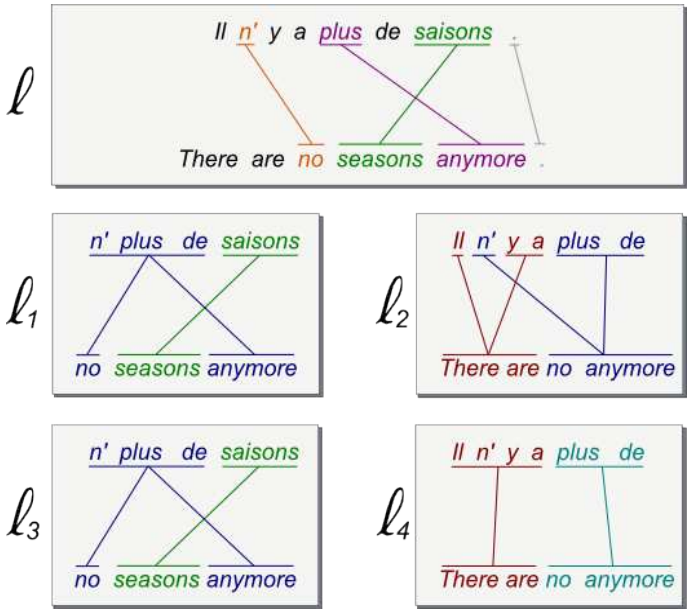


Figure 5.6 – ℓ_1 et ℓ_2 sont composables, ℓ_3 et ℓ_4 ne le sont pas

Autrement formulé, les fragments sont composables si en les plongeant séparément dans la biphrase neuve, les zones se chevauchant respectent les mêmes groupes de mots et les mêmes liens. À la figure 5.6, ℓ_1,ℓ_2,ℓ_3 et ℓ_4 sont des segments compatibles avec la biphrase pré-alignée ℓ. Seuls les fragments ℓ_1 et ℓ_2 sont composables car leur partie commune *"n '... plus de" / "no ... anymore"* est entière.

Définition 31. *Avec les mêmes notations,* $(\ell_1, \varphi_1, \psi_1)$ *et* $(\ell_2, \varphi_2, \psi_2)$ *seront* ***disjoints*** *dans* ℓ *si les conditions suivantes sont vérifiées :*

- *Pour* $V_1 \in \mathcal{V}_1$ *et* $V_2 \in \mathcal{V}_2$, $\varphi_1(V_1)$ *et* $\varphi_2(V_2)$ *sont disjoints.*
- *Pour* $W_1 \in \mathcal{W}_1$ *et* $W_2 \in \mathcal{W}_2$, $\psi_1(W_1)$ *et* $\psi_2(W_2)$ *sont disjoints.*

Propriété 8. *Des fragments disjoints sont composables.*

Une autre approche aurait pu consister à composer des fragments avec des zones d'intersection ne coïncidant pas exactement grâce à un compromis de *grossissement* (l'opération $\vee$ de la partie 2.2.2.2). Mais l"étape (4) s'adresse au problème de l'ambiguïté là où l'opération "$\vee$" l'entretiendrait en proposant une zone de flou. On définit l'opération de *composition* entre fragments composables :

Définition 32. *Soient* $(\ell_1, \varphi_1, \psi_1)$ *et* $(\ell_2, \varphi_2, \psi_2) \in \mathfrak{A}_{\ell}^{\Phi\Psi}$ *des fragments* ***composables****. On définit leur* ***composition*** *que l'on note :*

$$(\tilde{\ell}, \tilde{\varphi}, \tilde{\psi}) = (\ell_1, \varphi_1, \psi_1) \oplus (\ell_2, \varphi_2, \psi_2)$$

Tout d'abord définissons $\tilde{\ell}$ *:*

$$\begin{cases} \tilde{\mathcal{V}} & = & \mathcal{V}_1 \cup \mathcal{V}_2 \\ \tilde{\mathcal{W}} & = & \mathcal{W}_1 \cup \mathcal{W}_2 \\ \tilde{\sigma}(\tilde{V}) & = & \begin{cases} \sigma_1(\tilde{V}) \text{ si } \tilde{V} \in \mathcal{V}_1 \\ \sigma_2(\tilde{V}) \text{ si } \tilde{V} \in \mathcal{V}_2 \end{cases} \end{cases}$$

Les positionnements de $\tilde{\ell}$, $\tilde{\varphi}$ *et* $\tilde{\psi}$ *sont définis ci-dessous par récurrence forte :*

$$\begin{cases} \tilde{\varphi}(1) & = & \min\left(\min(\varphi_1), \min(\varphi_2)\right) & \text{si } \tilde{n} \geq 1 \\ \tilde{\psi}(1) & = & \min\left(\min(\psi_1), \min(\psi_2)\right) & \text{si } \tilde{m} \geq 1 \\ \tilde{\varphi}(k+1) & = & \min\left(\bigcup\limits_{1 \leq i \leq n_1} \{\varphi_1(i) > \tilde{\varphi}(k)\} \cup \bigcup\limits_{1 \leq j \leq n_2} \{\varphi_2(j) > \tilde{\varphi}(k)\} \right) & \forall k < \tilde{n} \\ \tilde{\psi}(l+1) & = & \min\left(\bigcup\limits_{1 \leq i \leq m_1} \{\psi_1(i) > \tilde{\psi}(l)\} \cup \bigcup\limits_{1 \leq j \leq m_2} \{\psi_2(j) > \tilde{\psi}(l)\} \right) & \forall k < \tilde{m} \end{cases}$$

Le nouveau fragment positionné $(\tilde{\ell}, \tilde{\varphi}, \tilde{\psi})$ se calcule en un nombre d'opérations linéaires en les tailles de $(\ell_1, \varphi_1, \psi_1)$ et $(\ell_2, \varphi_2, \psi_2)$. On définit également l'opération "duale" à la *composition* que nous appelons la *partie commune* et qui présente un intérêt pratique moindre.

Définition 33. *On définit la* ***partie commune*** *entre fragments composables :*

$$(\hat{\ell}, \hat{\varphi}, \hat{\psi}) = (\ell_1, \varphi_1, \psi_1) \odot (\ell_2, \varphi_2, \psi_2)$$

Tout d'abord définissons $\hat{\ell}$ *:*

$$\begin{cases} \mathscr{V}\hat{} & = & \mathscr{V}_1 \cap \mathscr{V}_2 \\ \mathscr{W}\hat{} & = & \mathscr{W}_1 \cap \mathscr{W}_2 \\ \hat{\sigma}(\hat{V}) & = & \sigma_1(\hat{V}) = \sigma_2(\hat{V}) \end{cases}$$

Les positionnements de $\hat{\ell}$*,* $\hat{\varphi}$ *et* $\hat{\psi}$ *sont définis ci-dessous par récurrence forte :*

$$\begin{cases} \hat{\varphi}(1) & = & \min\limits_{\substack{1\le i\le n_1 \\ 1\le j\le n_2}} \{\varphi_1(i) = \varphi_2(j)\} & si\ \hat{n} \ge 1 \\ \hat{\psi}(1) & = & \min\limits_{\substack{1\le i\le n_1 \\ 1\le j\le n_2}} \{\psi_1(i) = \psi_2(j)\} & si\ \hat{m} \ge 1 \\ \hat{\varphi}(k+1) & = & \min\limits_{\substack{1\le i\le n_1 \\ 1\le j\le n_2}} \{\varphi_1(i) = \varphi_2(j) > \hat{\varphi}(k)\} & \forall k < \hat{n} \\ \hat{\psi}(k+1) & = & \min\limits_{\substack{1\le i\le n_1 \\ 1\le j\le n_2}} \{\psi_1(i) = \psi_2(j) > \hat{\psi}(k)\} & \forall k < \hat{m} \end{cases}$$

La partie commune se calcule également en un nombre d'opérations linéaire en les tailles de $(\ell_1, \varphi_1, \psi_1)$ et $(\ell_2, \varphi_2, \psi_2)$.

L'ensemble des fragments produits à partir d'un alignement ℓ et munis de leurs positionnements forme l'ensemble $\mathfrak{A}_\ell{}^{\Phi\Psi}$ évoqué avant. Les fragments de cet ensemble sont naturellement compatibles avec ℓ et deux à deux composables. La composition nous permet naturellement d'obtenir ℓ :

$$\bigoplus_{\ell' \in \mathfrak{A}_\ell{}^{\Phi\Psi}} \ell' = \ell$$

Ce résultat est une conséquence directe de la fragmentation et permet d'insister sur une idée clé de l'architecture : comme nombres d'approches à base d'exemples, la notre se caractérise par cette étape consistant à découper des exemples de manière à ce qu'ils puissent être reconstruits, ce qui permettra éventuellement d'en construire d'autres. Nous avons en fait, les propriétés suivantes :

Propriété 9. *$\oplus$ et $\odot$ sont bien définies sur $\mathfrak{A}_{\ell}^{\Phi\Psi}$ et :*

- *$\oplus$ et $\odot$ sont des lois de composition interne sur $\mathfrak{A}_{\ell}^{\Phi\Psi}$ dont les éléments neutres sont respectivement $\ell_{\emptyset}$ et ℓ.*
- *$(\mathfrak{A}_{\ell}{}^{\Phi\Psi}, \oplus, \odot, \lhd)$ est un treillis borné de plus petit élément $\ell_{\emptyset}$ l'alignement vide et de plus grand élément ℓ.*
- *C'est un treillis complémenté. c'est-à-dire que pour tout sous-alignement ℓ', il existe un sous-alignement noté $\bar{\ell}'$ tel que $\ell' \odot \bar{\ell}' = \ell_{\emptyset}$ et $\ell' \oplus \bar{\ell}' = \ell$. On parle de l'alignement complémentaire de ℓ' selon ℓ.*

Nous nous sommes intéressés à la contiguïté et aux croisements dans les problèmes liés à la divergence (en partie 1.2). Définissons maintenant formellement les notions de fragments *contigus* et *croisants* :

Définition 34. *Soit $(\ell', \varphi, \psi) \in \mathscr{A}_{\ell}{}^{\Phi\Psi}$ un sous-alignement de B. On dit que ℓ' est **contigu** dans la biphrase B si :*

$$\begin{cases} \bigcup\limits_{V' \in \mathscr{V}'} \varphi(V') \text{ est un intervalle d'entiers} \\ \bigcup\limits_{W' \in \mathscr{W}'} \psi(W') \text{ est un intervalle d'entiers} \end{cases}$$

Nous pourrions définir la notion de croisement de manière générale entre deux sous-alignements, comme pour la contiguïté, mais il n'y aura concrètement peu d'intérêt à comparer deux fragments non composables. Nous proposons donc une définition entre deux fragments compatibles dans un alignement ℓ. Le critère s'adapte à deux sous-alignements d'une biphrase, et en cas non-composabilité, un programme répondrait certainement qu'ils sont *croisants*.

Définition 35. *Soient $\ell_1, \ell_2 \in \mathfrak{A}_\ell{}^{\Phi\Psi}$ deux alignements composables dans ℓ. On dit que ℓ_1 et ℓ_2 **se croisent** si il existe $i_1 \in \{1,\ldots,n_1\}, i_2 \in \{1,\ldots,n_2\}, j_1 \in \{1,\ldots,m_1\}, j_2 \in \{1,\ldots,m_2\}$ tels que :*

$$\begin{cases} \varphi_1(i_1) < \varphi_2(i_2) & et \quad \psi_1(j_1) > \psi_2(j_2) \\ & ou \\ \varphi_1(i_1) > \varphi_2(i_2) & et \quad \psi_1(j_1) < \psi_2(j_2) \end{cases}$$

*On dira que ℓ_1 et ℓ_2 **se croisent strictement** si la même propriété est vraie pour tous $i_1 \in \{1,\ldots,n_1\}, i_2 \in \{1,\ldots,n_2\}, j_1 \in \{1,\ldots,m_1\}, j_2 \in \{1,\ldots,m_2\}$.*

Un ensemble de fragments sera dit **non croisant** s'il contient des fragments deux à deux non croisants. Il sera dit **contigu** s'il ne contient que des fragments contigus.

Définition 36. *Soient $\ell_1, \ell_2 \in \mathfrak{A}_\ell{}^{\Phi\Psi}$ deux alignements compatibles avec ℓ. On dit que ℓ_1 et ℓ_2 sont **successifs** dans ℓ (noté $\ell_1 <_\ell \ell_2$) si pour tous $i_1 \in \{1,\ldots,n_1\}, i_2 \in \{1,\ldots,n_2\}, j_1 \in \{1,\ldots,m_1\}, j_2 \in \{1,\ldots,m_2\}$, on a :*

$$\varphi_1(i_1) < \varphi_2(i_2) \;\; et \;\; \psi_1(j_1) < \psi_2(j_2)$$

Propriété 10. *Deux fragments disjoints non croisants $\ell_1, \ell_2 \in \mathfrak{A}_\ell{}^{\Phi\Psi}$ sont successifs.*

Remarque 13. *D'un point de vue informatique, toutes les notions écrites ici sont représentables simplement et économiquement. Par exemple, les positionnements φ et ψ d'un fragment de longueur (n,m) pourront prendre la forme de deux listes, l'une de longueur n, l'autre de longueur m, correspondant aux valeurs respectives que prennent les fonctions sur $[\![1,n]\!]$ et $[\![1,m]\!]$. Les propriétés de compatibilité, compositionnalité, contiguïté et croisement seront vérifiables par un parcours linéaire des instances. De même que pour les opérations de composition et le calcul d'une partie commune.*

Les notions formalisées dans cette partie correspondent à différents éléments transitant dans le modèle décrit par le schéma d'architecture global (figure 3.16 page 103 au chapitre 3). À la figure 5.7, nous rappelons les différentes étapes de notre modèle en y incorporant les différents éléments formels introduits ici.

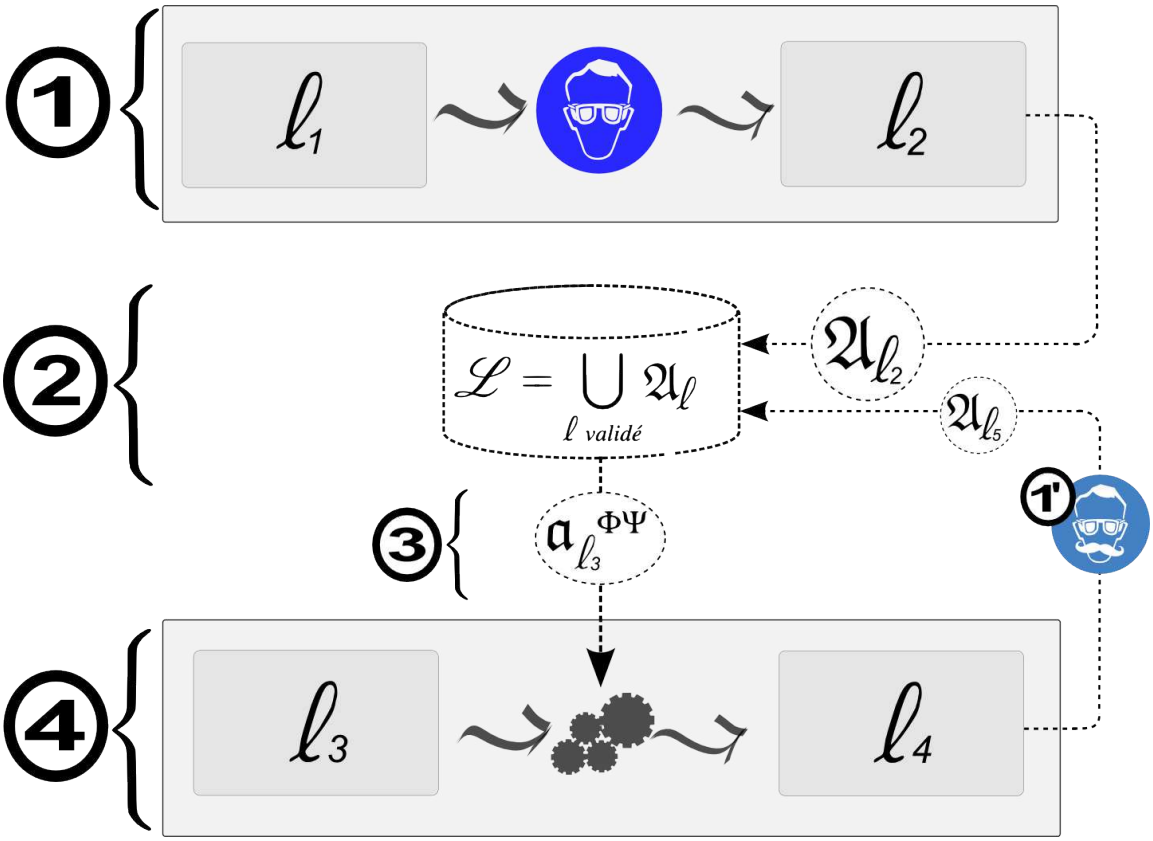


Figure 5.7 – Les fragments réutilisés sont positionnés selon l'alignement en cours

On rappelle que lorsqu'une biphrase ℓ_2 est fragmentée à l'étape ②, des *fragments* de $\mathfrak{A}_{\ell_2}$ sont produits et ajoutés à la mémoire $\mathscr{L}$. Nous verrons à la section 5.1.2 que nous ne mémorisons pas $\mathfrak{A}_{\ell_2}$ entier mais une sous partie selon différents critères pour former plusieurs mémoires. Le positionnement dans leur biphrase d'origine n'est pas conservé. Les segments sont donc mémorisés comme des alignements indépendants car sur une future biphrase, les positionnements changeront. L'étape ③, observe une biphrase vierge ou pré-alignée selon ℓ_3 (éventuellement vide), puis fouille la mémoire pour en extraire des fragments compatibles avec ℓ_3, dont les positionnements par rapport à celle-ci sont calculés. Les éléments extraits sont donc inclus dans $\mathfrak{a}_{\ell_3}^{\Phi\Psi}$ (Nous évaluons le poids de cette extraction dans la partie 5.1.3). Enfin, les fragments sont composés à l'étape ④ via l'opération $\oplus$.

5.1.2 Les mémoires d'alignements

Nous utilisons le terme *fragmentation* pour faire référence au processus de constitution des fragments, mais aussi à l'ensemble de fragments produits (la notation courante pour un ensemble de fragments sera $\mathscr{L}$). En pratique, nous préfèrerons retenir certains fragments plutôt que l'ensemble exhaustif des sous-fragments. Sélectionner un sous-ensemble de fragments est courant dans les travaux existant de TABE. En général, les arguments avancés sont, la réduction de la masse des informations à traiter ou une restriction à des fragments de qualité (voir [62], par exemple). Ici, les fragments sont bidimensionnels et il y a donc une composante géométrique supplémentaire à prendre en compte.

Il n'y a pas une unique manière de fragmenter un alignement, et donc pas une unique manière de construire une mémoire d'alignements. Notre approche étant expérimentale, nous avons abordé plusieurs types de fragmentation en prenant garde toutefois que les fragments retenus, pour un alignement ℓ validé, suffisent à le reconstruire par la composition $\oplus$. Ainsi, nous verrons que la mémoire, notée $\mathscr{L}$ dans la partie précédente, est en fait divisée en sept mémoires provenant de sept types de fragmentation :

$$\mathscr{L} = \mathscr{L}_B \cup \mathscr{L}_W \cup \mathscr{L}_X \cup \mathscr{L}_{W1} \cup \mathscr{L}_{X1} \cup \mathscr{L}_{W2} \cup \mathscr{L}_{X2}$$

Nous les décrivons dans la partie suivante. Dans chaque cas, en partant d'un alignement de départ ℓ, la fragmentation produira un ensemble de fragments à partir desquels il sera possible de reconstruire ℓ. la fragmentation imposera une contrainte spécifique pouvant être de nature géométrique ou syntaxique. Nous suivront l'exemple de l'alignement en figure 5.8 pour représenter les différentes fragmentations possibles.

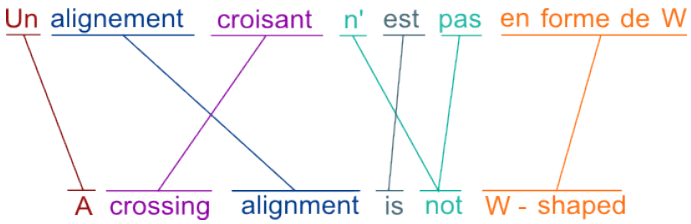


Figure 5.8 – Exemple ℓ à fragmenter

Fragments B

Les **fragments de type B** constituent la première mémoire $\mathscr{L}_B$ que nous avons formé. Cette mémoire, parmi toutes, est celle qui présente les fragments les plus petits et les plus génériques possibles. Le processus de fragmentation ne retient que les sous-alignements minimaux non nuls de $\mathfrak{A}_\ell$ (au sens de $\lhd$). On peut observer ces *fragments B* repérés à la figure 5.9. Nous verrons que par rapport à cette première, les autres mémoires sembleront présenter un défaut de puissance (au sens de la généricité).

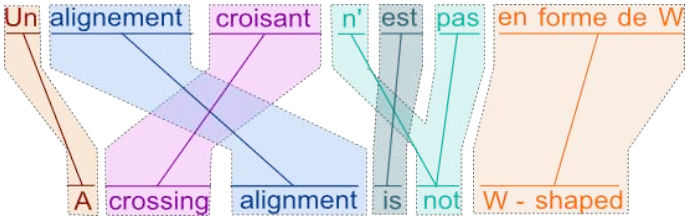


Figure 5.9 – Fragmentation B

Fragments W

Les **fragments de type W** constituent une seconde mémoire $\mathscr{L}_W$ dans laquelle la fragmentation englobe les phénomènes non contigus. Pour celle-ci, nous retiendrons les fragments de $\mathfrak{A}_\ell$, contigus et minimaux pour la relation '$\lhd$'. La configuration en *W* observée en figure 5.10 lui vaut cet appellation (les liens entre *"n'est pas"* et *"is not"*).

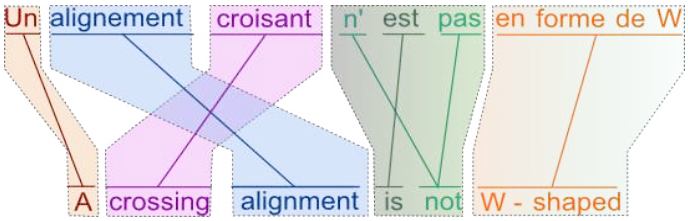


Figure 5.10 – Fragmentation W

Fragments X

Les **fragments de type X** constituent une mémoire $\mathscr{L}_X$ dans laquelle la fragmentation englobe les croisements (d'où la lettre X). $\mathscr{L}_X$ sera la plus grande collection de fragments deux à deux disjoints et non croisants.

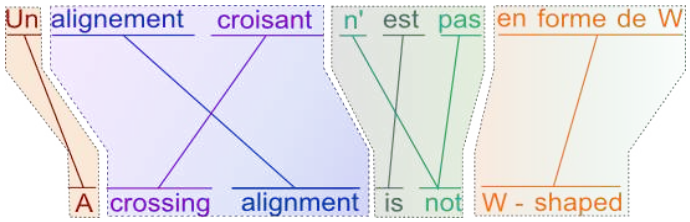


Figure 5.11 – Fragmentation X

Fragmentation unilingue par les syntagmes

Dans l'article [104], sont énoncés deux principes d'une importance cruciale dans toute approche à base d'exemples : les fragments courts ont plus de chance d'être réutilisables, mais les fragments très courts présentent un risque d'ambiguïté. Il s'agit donc de trouver un compromis intéressant pour la taille des fragments. En utilisant l'arbre syntagmatique pour découper en fragments plus longs, nous proposons de créer des fragments un peu moins génériques, mais plus sûrs. De plus, en respectant un découpage linguistiquement motivé, nous espérons moins de "friction" entre les fragments compatibles retenus.

La fragmentation proposée ici, que nous appelons la *fragmentation unilingue par les syntagmes*, impose une dimension syntaxique asymétrique en s'appuyant sur un seul arbre d'analyse pour former des fragments plus longs. Nous faisons l'hypothèse selon laquelle le groupement de mots cible induit par l'arbre source présente une certaine cohérence. Il ne sera pas rare dans des cas de divergence modérée, que les fragments forment de véritables paires de syntagmes. De plus, cette fragmentation asymétrique aura l'intérêt pratique, lorsqu'une des deux langues ne dispose pas d'un analyseur syntaxique générant des structures profondes, de s'appuyer sur les outils disponibles de l'autre.

En pratique, la fragmentation unilingue par les syntagmes sera utilisée en conjonction avec les contraintes de contiguïté (type W) et de croisement (type X). Dans les deux cas, les fragmentations produites devront respecter une double contrainte à la fois syntaxique (découpage syntagmatique asymétrique) et géométrique (contiguïté des fragments ou non-croisement). Nous notons les deux mémoires $\mathscr{L}_{W1}$ et $\mathscr{L}_{X1}$. On peut observer en figure 5.12 une fragmentation correspondant autant à $\mathscr{L}_{W1}$ qu'à $\mathscr{L}_{X1}$. Nous parlerons aussi de *fragments de types* $W1$ ou $X1$.

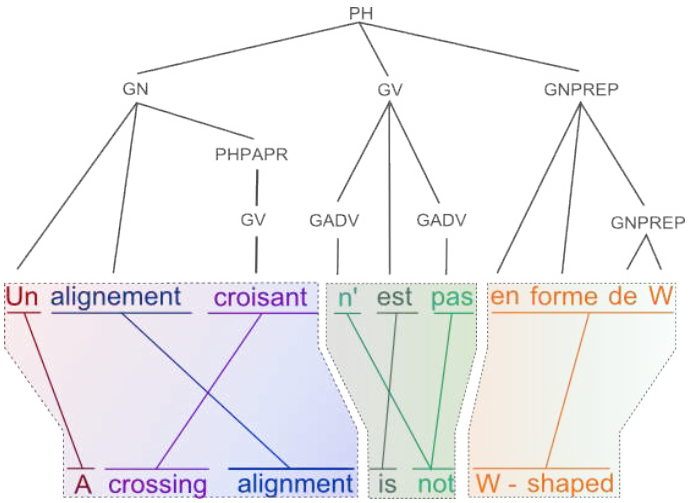


Figure 5.12 – Fragmentation unilingue par les syntagmes, de type $X1$ et $W1$

Le schéma ne représente qu'une étape de fragmentation puisque les fragments observés descendent des nœuds internes situés en bas de l'arbre. Par remontée au travers de l'arbre, on collecte des fragments plus longs et on finit par rencontrer le nœud racine qui formera le fragment le plus long correspondant à la biphrase tout entière. Un tel fragment se sera pas très réutilisable, mais à le mérite de contribuer à ce qu'un alignement déjà rencontré soit toujours correctement reformé par le processus. Il permet aussi de justifier la justesse de la définition de fragmentation unilingue qui sera toujours capable de produire des fragments, même triviaux.

Formellement, la structure sous-jacente est une S-SSTC $(S, C, \ell_{node}, \ell_{tree})$ (voir le chapitre 4) dont l'arbre d'analyse cible tr^C est plat. Nous nous reposons donc sur l'arbre tr^S pour fragmenter l'alignement ℓ_{node}. Donnons nous un fragments $\ell' = (\mathscr{V}', \mathscr{W}', \sigma')$

compatible avec l'alignement $\ell_{node} = (\mathcal{V}_{node}, \mathcal{W}_{node}, \sigma_{node})$ issu d'une fragmentation unilingue par les syntagmes. La contrainte syntaxique devant être respectée par ℓ' est :

$$\text{il existe } \mathcal{V}'_{tree} \subset \mathcal{V}_{tree} \text{ tel que } \bigcup_{V' \in \mathcal{V}'} STREE^S(V') = \bigcup_{V_{tr} \in \mathcal{V}'_{tree}} STREE^S(V_{tr})$$

Autrement dit, les mots source couverts par $\mathcal{V}'$ correspondent à l'union de syntagmes de l'arbre tr^S. Les fragments s'obtiennent par remontée au travers de l'arbre source. Le nombre d'opérations nécessaires est linéaire en la taille de la structure source.

Alimenter la mémoire de ces fragments plus longs permettra aussi au système de prendre des raccourcis dans le processus d'alignement sans avoir à revenir systématiquement aux briques de bases. Notamment, une biphrase déjà rencontrée sera reconnue en tant que tel et l'alignement sera donné sans devoir être recalculé.

Fragmentation bilingue par les syntagmes

Enfin, nous proposons une fragmentation respectant les deux structures au niveau du découpage en syntagmes. La condition nécessaire pour utiliser cette fragmentation est de travailler sur une paire de langues dotées d'analyseurs syntaxiques en structures profondes et similaires. Aligner une structure en dépendance et une structure en constituants aura ici pour effet de produire des fragments très longs et très peu génériques. Le même problème se produira si l'on applique cette fragmentation à une paire de langues très divergentes, le consensus ne saura produire que quelques fragments très longs.

De la même manière que pour la fragmentation unilingue, deux contraintes devront être respectées : une contrainte syntaxique et une contrainte géométrique concernant la contiguïté ou les croisements. Nous reprenons les notations de la partie précédente en prenant une S-SSTC $(S, C, \ell_{node}, \ell_{tree})$ et un fragment $\ell' = (\mathcal{V}', \mathcal{W}', \sigma')$ compatible avec l'alignement $\ell_{node} = (\mathcal{V}_{node}, \mathcal{W}_{node}, \sigma_{node})$. La contrainte syntaxique que doit respecter ℓ' est maintenant double et correspond à l'intersection de deux contraintes unilingues, à savoir l'existance de $\mathcal{V}'_{tree} \subset \mathcal{V}_{tree}$ et $\mathcal{W}'_{tree} \subset \mathcal{W}_{tree}$ tels que :

$$\begin{cases} \bigcup_{V' \in \mathcal{V}'} STREE^S(V') & = & \bigcup_{V_{tr} \in \mathcal{V}'_{tree}} STREE^S(V_{tr}) \\ \bigcup_{W' \in \mathcal{V}'} STREE^C(W') & = & \bigcup_{W_{tr} \in \mathcal{W}'_{tree}} STREE^S(W_{tr}) \end{cases}$$

Il faut noter qu'en plus de produire des fragments a priori longs (donc peu expressifs mais stables), cette fragmentation nécessitera des traitements plus importants. La mémoire produite sous la contrainte de contiguïté sera notée $\mathscr{L}_{W2}$ et celle sous la contrainte de non-croisement, $\mathscr{L}_{X2}$. Nous parlerons de *fragments de type* $W2$ et $X2$. Nous pouvons observer des fragments de ce type en figure 5.13. De même que pour la fragmentation unilingue, des fragments plus longs peuvent être constitués par remontée au travers des arbres.

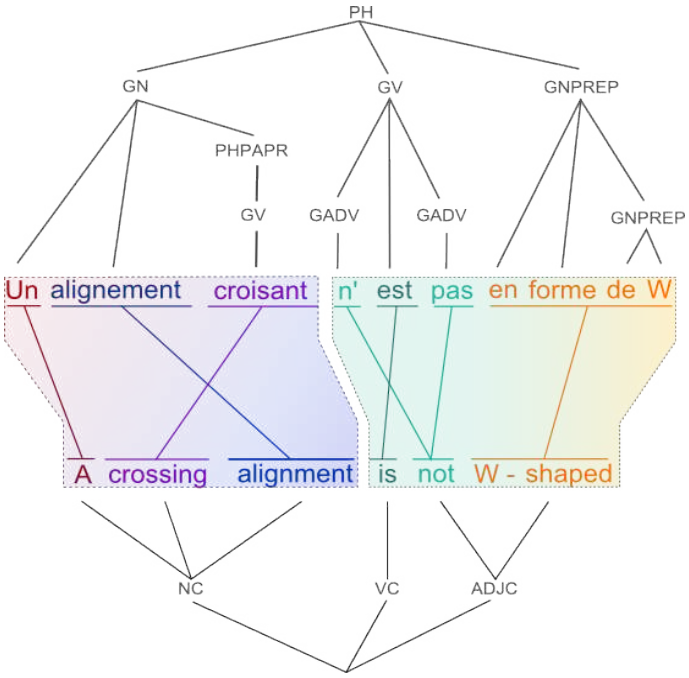


Figure 5.13 – Fragmentation bilingue par les syntagmes

Quelques remarques

Les différentes fragmentations sont présentées séparément, car elles possèdent des propriétés structurelles différentes, mais il n'y aura qu'une seule mémoire de fragments qui seront typés selon la fragmentation dont ils sont issus. Le type des fragments utilisés aura un impact sur l'étape de résolution ④ (figure 5.7 page 147). Notamment, les processus utilisés seront différents et par exemple, reconstruire un alignement à partir de ses

fragments de type X pourra se faire par une composition sans croisement. Mais les différentes fragmentations ne sont pas pour autant à séparer complètement. En effet, on sait que peut importe le type, un fragment est toujours la composition de ses B-fragments. De plus, un X-fragment peut-être obtenu comme la composition de W-fragments. On note également qu'un fragment/syntagme bilingue est composé de fragments/syntagmes unilingues. Une hiérarchie entre les types de fragments est représentée en figure 5.14, allant du plus générique tout en bas, aux plus spécifiques tout en haut. Les fragmentations B, W et X requièrent seulement un étiqueteur morphosyntaxique pour former les patrons syntaxiques, tandis que la partie haute impose la présence d'outils d'analyse profonds. La fragmentation bilingue par les syntagmes ne sera pas utilisée dans notre approche.

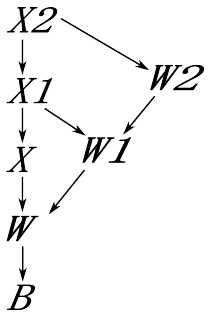


Figure 5.14 – Hiérarchie des différentes fragmentations

Ainsi, un processus d'alignement adapté à des W-fragments pourra tout à fait utiliser des fragments de types supérieurs X, $W1$, $W2$, $X1$ ou $X2$ comme le suggère la hierarchie (les fragments plus longs sont un raccourci pour l'étape ④). En général, compléter par des fragments de types supérieurs et moins génériques l'ensemble de fragments compatibles collectés par l'étape ③ ne pourra que se montrer avantageux en terme de qualité, peut importe le type de départ.

On pourrait alors supposer que le contraire soit une erreur : mêler des fragments de types inférieurs, plus courts et linguistiquement moins pertinents, à de bons fragments compatibles risquerait d'ajouter du bruit et de nuire à la reconstruction. Ce n'est pas toujours vrai, et nous pouvons constater dans certains cas que cette hiérarchie reste assez artificielle. Par exemple, un fragment non contigu tel que la négation française *"**ne**...**pas**"*

rencontré dans "*il **ne** faut **pas** nourrir les animaux*", peut tout à fait se retrouver employé de manière contigües dans la forme "***ne pas** nourrir les animaux*". Un fragment de type *B* est ici réutilisé en fragment de type supérieur.

On peut également donner l'exemple de l'inversion adjectif-nom entre l'anglais et le français qui, bien que majoritaire, n'est pas toujours vérifiée. On l'observe en confrontant les deux phrases : *"Elle préfère les **hommes grands**"/"She likes better **tall men**"* et *"C'est l'époque des **grands hommes** de petite vertu"/"These are the times of **tall men**, and short character"*. Une fragmentation de type *X* du premier alignement ne donne pas de fragments suffisemment génériques pour résoudre le cas de non-inversion. Par contre, une *B*-fragmentation aurait permis d'en tenir compte.

5.1.3 La taille (potentielle) de l'ensemble des fragments compatibles

Les approches à base d'exemples doivent généralement faire face à la question du nombre d'exemples récoltés par l'étape ③ et qui forment l'*ensemble des fragments compatibles*. Les travaux en TABE utilisant des mémoires de fragments se méfient généralement du nombre de fragments que peut générer l'étape de récolte, pouvant devenir insoutenable comme il est remarqué dans [80] :

> *The example number can be exponentially large [...], the impracticality of EBMT might arise.*

Nous différencions ici les deux notions que sont l'*ensemble de fragments potentiellement compatibles* et l'*ensemble de fragments effectivement compatibles*. Le premier, auquel nous nous intéresserons plus particulièrement, est l'ensemble théorique des fragments compatibles $\mathfrak{A}(S,C)$ avec une biphrase de départ vierge que pourrait retourner à terme une mémoire totale. Le deuxième est l'ensemble des fragments compatibles retournés par la mémoire $\mathscr{L} \cap \mathfrak{A}(S,C)$. Il dépend de l'état courant de la mémoire $\mathscr{L}$ et est amené à évoluer au court du temps.

Nous étudions ici la corrélation entre le type des fragments utilisés et la quantité de ces exemples potentiellement problématiques. Nous nous plaçons dans la situation de

départ d'une biphrase (S,C) de longueur (n,m) à aligner. Nous recherchons des fragments compatibles dans la mémoire, candidats pour participer à l'alignement. Il faut alors choisir un type de fragment voulu : B, W ou X.

Des fragments non contigus

Pour une biphrase (S,C) vierge, le nombre théorique de **fragments de type *B*** compatibles peut s'avérer être assez élevé à cause des cas de non contiguïté. Chaque B-fragment possède une partie source et une partie cible. Le nombre de parties source (potentiellement non contigües) de longueur $k \leq n$ peut se compter par le nombre d'injections de $[\![1,k]\!]$ dans $[\![1,n]\!]$, c'est-à-dire le nombre d'arrangements A_n^k. Il en est de même pour le nombre de parties cible. Ainsi, le nombre de B-fragments compatibles de longueur $(k,l) \leq (n,m)$ vaut $A_n^k \cdot A_m^l$. Finalement, le nombre de B-fragments compatibles vaut :

$$\sum_{\substack{1\leq k\leq n\\ 1\leq l\leq m}} A_n^k \cdot A_m^l \underset{n,m\to+\infty}{\sim} e^2 \cdot n! \cdot m!$$

La quantité de B-fragments potentiellement compatibles explose et n'est donc pas naturellement contrôlable.

Des fragments contigus

L'ensemble des fragments de types W ou X qui sont *potentiellement* compatibles avec la biphrase à aligner sont les mêmes : il s'agit de l'ensemble des fragments liant côté source et cible des fragments contigus ou intervalles de mots. Le nombre de parties source contigües de longueur k vaut $n-k+1$ et donc, le nombre de fragments contigus potentiellement compatibles de longueur $(k,l) \leq (n,m)$ vaut :

$$\sum_{\substack{1\leq k\leq n\\ 1\leq l\leq m}} (n-k+1)\cdot(m-l+1) \underset{n,m\to+\infty}{\sim} \frac{n^2\cdot m^2}{4}$$

La quantité est plus raisonnable que pour les fragments non contigus. Il s'agit ici aussi d'une majoration, le nombre de fragments compatibles présents en mémoire sera en

pratique plus réduit. Le nombre effectif de fragments retournés sera plus important si l'on souhaite fouiller les deux mémoires X et W plutôt qu'une seule. Si ces deux mémoires sont peu différenciées ici, ça ne sera pas le cas dans l'étape ④ de reconstruction (fig. 5.7 page 147), notamment au niveau des stratégies mises en place (décrite en section 5.2).

Conséquence du pré-alignement

L'architecture proposée doit être capable de prendre comme point de départ, un biphrase partiellement alignée. En imposant cette contrainte de pré-alignement, le nombre de fragments compatibles diminue, ce qui aura pour conséquence de réduire le coût des traitement pour les processus ultérieurs. Nous évaluons ici l'économie apportée par un pré-alignement *mot à mot*. On se donne donc un pré-alignement ℓ_{init} et supposons qu'il recouvre N mots de la partie source et M mots de la partie cible. Cette fois, les fragments potentiellement compatibles ont la contrainte supplémentaire d'avoir ℓ_{init} comme sous-alignement.

Dans le cas de B-fragments qui peuvent admettre un nombre arbitraire de défauts de contiguïté, l'espace de recherche est très peu affecté. En effet, le nouveau problème se ramène simplement à rechercher un ensemble de B-fragment compatibles sur la sous-biphrase de longueur $(n-N, m-M)$ formée des mots qui ne sont pas recouverts par ℓ_{init}. On applique la même formule et le nombre de B-fragments potentiellement compatibles devient :

$$e^2 \cdot (n-N)! \cdot (m-M)! \underset{n,m\to+\infty}{\sim} e^2 \cdot n! \cdot m!$$

L'équivalent asymptotique reste un carré de factorielles. On peut dire que, d'une certaine manière, la géométrie des B-fragments est trop libre pour véritablement être contrainte par un pré-alignement.

Dans le cas de fragments contigus, la contrainte d'un pré-alignement est plus importante et le gain est meilleur. Par commodité, nous raisonnerons avec un pré-alignement ℓ_{init} à liens simples (on peut toujours se ramener à ce cas en considérant un sous-alignement de ℓ_{init}). On pose L le nombre de liens composant ℓ_{init}. On suppose que le pré-alignement sépare la phrase source en $L+1$ parties de longueurs maximum p et

la phrase cible en $L+1$ parties de longueurs maximum q (voir figure 5.15). Nous allons pouvoir paramétrer le nombre de fragments potentiellement compatibles par p et q de sorte que pour p et q petits, le nombre de fragments soit plus réduit.

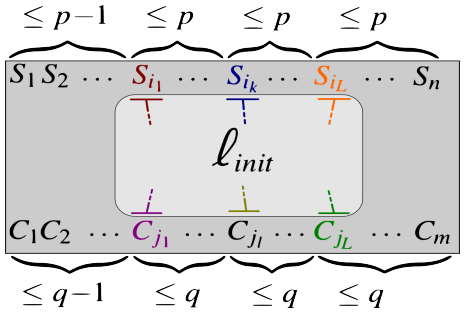


Figure 5.15 – Pré-alignement découpant la biphrase en $L+1$

On note N_k^S (resp. N_k^C) le nombre d'intervalles source (resp. cible) chevauchant exactement k liens côté source (resp. cible). Pour $k=0$, les intervalles doivent être bloqués entre deux liens. On a donc :

$$\begin{cases} N_0^S & \leq & \frac{L+1}{2} \cdot p \cdot (p-1) \\ N_0^C & \leq & \frac{L+1}{2} \cdot q \cdot (q-1) \end{cases}$$

Le nombre de bi-fragments n'étendant aucun lien de ℓ_{init} est majoré par :

$$\frac{(L+1)^2}{4} \cdot p \cdot (p-1) \cdot q \cdot (q-1)$$

On prend $k \geq 0$, c.à.d. un fragment source recouvrant k mots liés. Le fragment étant contigu, la partie entre le premier mot lié et le k-ième est fixe. Il y a au plus $p-1$ choix à gauche et au plus $p-1$ à droite. De plus, il y a $L-k+1$ manières de choisir les k mots liés, ce qui nous donne :

$$\begin{cases} N_k^S & \leq & (p-1)^2 \cdot (L-k+1) \\ N_k^C & \leq & (q-1)^2 \cdot (L-k+1) \end{cases}$$

Le cas où le nombre de bi-fragments étendant ℓ_{init} sera le plus élevé sera quand les liens du pré-alignement seront parallèles. En effet, si ça n'est pas le cas, un fragment

source chevauchant k liens de ℓ_{init} ne pourra pas toujours être mis en relation avec un fragment cible chevauchant k liens. Dans le cas de liens parallèles, chaque fragment source (chevauchant des liens) peut être lié au plus à $(q-1)^2$ fragments cible.

Ainsi, le nombre de fragments contigus étendant ℓ_{init} est majoré par :

$$\frac{(L+1)^2}{4} \cdot p \cdot (p-1) \cdot q \cdot (q-1) + \sum_{k=1}^{L} (p-1)^2 \cdot (q-1)^2 \cdot (L-k+1)$$

Et en utilisant les majorations $(L+1) \cdot p \leq n-1$ et $(L+1) \cdot q \leq m-1$, on obtient la majoration suivante pour le nombre de fragments compatibles :

$$\frac{3}{4} \cdot (p-1) \cdot (q-1) \cdot (n-1) \cdot (m-1) \cdot$$

Le nombre de fragments contigus potentiellement compatibles n'est plus quartique, mais devient quadratique avec un facteur multiplicatif $\frac{3}{4} \cdot (p-1) \cdot (q-1)$.

Remarque 14. *Si on choisit d'ignorer les fragments ne rencontrant aucun lien de ℓ_{init}, le facteur multiplicatif descend à $\frac{1}{2} \cdot (p-1) \cdot (q-1)$, ce qui n'est pas un gain substentiel, mais est tout de même envisageable pour une approche par propagation de liens sûrs.*

La quantité peut devenir linéaire en $\max(n,m)$ si l'on se limite aux fragments n'étendant qu'un seul lien de ℓ_{init}. Le facteur multiplicatif constant devient $\max(p-1, q-1)$. Au risque de perdre des fragments importants, cette limitation permet d'aligner par propagation locale.

Commentaires

Les mémoires constituées auront des tailles allant en décroissant selon l'ordre des types : B, W puis X. C'est naturel car les unités produites par les fragmentations sont de tailles croissantes et la fragmentation X produira donc moins d'éléments que la fragmentation W qui elle-même en produira moins que la fragmentation B. Nous le constatons naturellement dans le dénombrement théorique des B-fragments potentiellement compatibles par rapport à ceux de type W ou X.

En abordant l'alignement par des fragmentations d'exemples, nous proposons un modèle d'expressivité totale : potentiellement, toute configuration d'alignement peut être atteinte car la mémoire accepte tout alignement et le fragmente de manière à pouvoir être "reconstruit". Pourtant on constate que les fragmentations ne sont pas égales entre elles. Si on compare les *B*-fragments aux fragments contigus, on voit que s'installe un compromis naturel : la mémoire de *B*-fragments est plus rapide à construire étant donné qu'elle se base sur une fragmentation plus fine que les fragmentations X et W. L'avantage est de proposer un ensemble de fragments d'une plus grande généricité, là où les fragments contigus peuvent avoir une certaine difficulté à "capturer" un phénomène non contigu. Par exemple, la négation "ne...pas" en français, sera exprimée totalement par un *B*-fragment là où il faudra autant de fragments contigus que de manière d'utiliser et de traduire la négation de manière courante (*"ne + Verbe + pas", "ne + Adverbe + Verbe + pas",* etc...). Un effort plus conséquent doit être donc fait pour construire une base de fragments contigus génériques.
D'un autre côté, bien que très générique, la *B*-fragmentation présente un désavantage clair : le nombre de fragments potentiellement compatibles est borné par une quantité virtuellement infinie (plus de 185 millions pour des biphrases de longueur $(7,7)$). Cela ne signifie pas que le nombre de *B*-fragments présents dans la mémoire explose effectivement car cela dépendra de nombreux facteurs dont celui de la paire de langue considérée (le japonais-français présentera probablement plus de phénomènes non contigus que le français-espagnol) ; cela signifie seulement que la quantité est difficilement contrôlable, ce qui pose un sérieux problème de calculabilité. Soudain les fragments contigus, moins génériques, paraissent plus acceptables, ce que confirmeront dans la section 5.2 des considérations de complexité du problème d'alignement relativement au type de fragmentation.

5.2 Reconstruction à base de fragments

L'idée même de fragmentation est irrémédiablement liée au problème inverse de la reconstruction. Une stratégie de fragmentation produira des sous-alignements qu'il sera

possible de combiner pour obtenir l'alignement initial. On verra que le problème de reconstruction est étroitement lié à la stratégie initiale de fragmentation.

5.2.1 Modélisation du problème

Nous nous plaçons maintenant à l'étape ④, dite de **synthèse du résultat** à partir d'un ensemble de fragments retournés par l'étape ③ (revoir les figures 3.16 et 5.7 pages 103 et 147). Nous rappelons que les fragments produits à partir d'un alignement ℓ et munis de leurs positionnements sont notés $\mathfrak{A}_{\ell}^{\Phi\Psi}$ (nous utiliserons la même notation lorsqu'une fragmentation n'est pas exhaustive). L'ensemble des *fragments positionnés compatibles* avec ℓ est noté $\mathfrak{a}_{\ell}^{\Phi\Psi}$. Il correspond à l'ensemble des fragments compatibles avec ℓ que l'étape ③ peut potentiellement extraire de la mémoire $\mathscr{L}$. Pour alléger les notations, un fragment positionné $(\ell', \varphi, \psi) \in \mathfrak{a}_{\ell}^{\Phi\Psi}$ sera abusivement noté $\ell' \in \mathfrak{a}_{\ell}^{\Phi\Psi}$. Nous noterons $\mathfrak{L}$ l'ensemble des fragments compatibles avec ℓ **effectivement** retournés par l'étape ③, puisqu'il est clair que la mémoire $\mathscr{L}$, en pratique, est partielle et en évolution. Nous avons les inclusions suivantes :

$$\mathfrak{A}_{\ell}^{\Phi\Psi} \subset \mathfrak{a}_{\ell}^{\Phi\Psi} \quad et \quad \mathfrak{L} \subset \mathfrak{a}_{\ell}^{\Phi\Psi}$$

Lorsque la biphrase à aligner à déjà été rencontrée et alignée par le passé, les fragments nécessaires à son "bon alignement" sont déjà présents en mémoire $\mathscr{L}$ (étape ②). Dans ce cas l'étape ④ sera plutôt une **étape de reconstruction** et nous avons les inclusions successives :

$$\mathfrak{A}_{\ell}^{\Phi\Psi} \subset \mathfrak{L} \subset \mathfrak{a}_{\ell}^{\Phi\Psi}$$

Ce cadre favorable serait systématiquement vérifié pour une mémoire idéale pour laquelle $\mathfrak{L} = \mathfrak{a}_{\ell}^{\Phi\Psi}$. Bien sûr la situation favorable dans laquelle un ensemble de fragments nécessaires est retourné peut se produire, en pratique, sur des nouvelles biphrases. Pour être viable, la *synthèse du résultat* doit être capable de reconstruire un alignement ℓ sur une biphrase B déjà fragmentée par le passé. Si la mémoire $\mathscr{L}$ de fragments fouillée est celle d'une fragmentation par les syntagmes ($\mathscr{L}_{W1}$, $\mathscr{L}_{X1}$, $\mathscr{L}_{W2}$ ou $\mathscr{L}_{X2}$), l'aligne-

ment pourra être reconstitué puisqu'un fragment total (ℓ lui-même), correspondant au(x) nœud(s) racine, aura été sauvegardé.

Pour les mémoires non exhaustives $\mathscr{L}_B$, $\mathscr{L}_W$ et $\mathscr{L}_X$ un fragment total ne sera en général pas disponible. Seuls des alignements compatibles plus courts issus d'un découpage en fragments disjoints seront présents. Intuitivement, il est toujours possible de combiner ces fragments compatibles pour reformer l'alignement ℓ initial, car comme nous l'avons remarqué précédemment (section 5.1.1 page 144), la composition $\oplus$ des fragments de $\mathfrak{A}_\ell^{\Phi\Psi}$ donne ℓ. L'étape ③ renvoie un ensemble de *fragments positionnés* $\mathfrak{L}$ qui, en général, contient non seulement l'ensemble des fragments nécessaires à la reconstruction $\mathfrak{A}_\ell^{\Phi\Psi}$, mais aussi des fragments inutiles et/ou erronés provenant de fragmentations d'autres biphrases que ℓ. L'étape ④ de notre architecture consiste donc en un processus de filtrage, c'est-à-dire sélectionner le meilleur sous-ensemble de $\mathfrak{L}$ selon un critère adapté. Dans le cas de la reconstruction de ℓ, le sous-ensemble attendu est $\mathfrak{A}_\ell^{\Phi\Psi} \subset \mathfrak{L}$.

On remarque que si ℓ recouvre entièrement la biphrase S, le sous-ensemble de fragments $\mathfrak{A}_\ell^{\Phi\Psi}$ pourra être retrouvé parmi les sous-ensembles de $\mathfrak{L}$ dont les fragments sont **deux à deux composables** et dont **la composée a une couverture maximum**. Cela revient à considérer l'étape ④ comme la résolution du problème suivant :

Problème 1. *À partir d'un ensemble de fragments positionnés $\mathfrak{L}$, sélectionner un sous-ensemble $\mathfrak{L}' \subset \mathfrak{L}$ tel que :*

$$\begin{cases} \textit{Les fragments de } \mathfrak{L}' \textit{ sont deux à deux composables} \\ \textit{couv}\left(\bigoplus_{\ell' \in \mathfrak{L}'} \ell'\right) \textit{ réalise son maximum sur } \mathfrak{L} \end{cases}$$

Mais le calcul de la couverture d'un ensemble de fragments deux à deux composables nécessite le calcul de la composée, et complexifie la modélisation du problème. En effet, nous allons voir à la partie 5.2.2 que des versions faibles de ce problème ont des complexités déjà élevées. Nous faisons une hypothèse simplificatrice en recherchant un sous-ensemble de fragments non pas deux à deux composables, mais **deux à deux**

disjoints. Il est important de se rendre compte que même si il y a un affaiblissement du problème, l'expressivité des solutions n'en est nullement entamée et que la reconstruction de l'alignement ℓ est toujours possible. C'est le cas car nous avons pris soin de ne définir en section 5.1.2 que des fragmentations produisant des fragments deux à deux disjoints ou en relation d'inclusion. Par conséquent, la reconstruction de ℓ à partir des fragments $\mathfrak{A}_{\ell}^{\Phi\Psi}$ sera possible en composant uniquement des fragments disjoints. On remarque cependant, comme le représente la figure 5.16, que le problème ainsi posé ne tient pas compte des liens internes, mais seulement des positionnements des fragments.

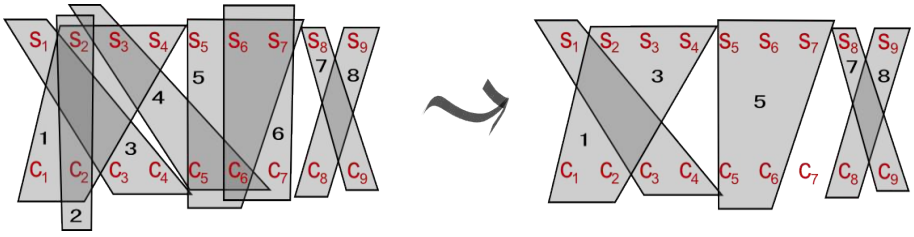


Figure 5.16 – Le problème 2 sur un ensemble de fragments contigus (les liens internes ne sont pas pris en compte)

Cette hypothèse permet de simplifier considérablement le problème car la couverture de la composée de fragments deux à deux disjoints est simplement la somme de la couverture des fragments. On voit que pour le problème simplifié, défini ci-dessous, il n'est plus nécessaire de calculer la composée des fragments dans la deuxième condition :

Problème 2. *À partir d'un ensemble de fragments positionnés $\mathfrak{L}$, sélectionner un sous-ensemble $\mathfrak{L}' \subset \mathfrak{L}$ tel que :*

$$\begin{cases} \textit{Les fragments de } \mathfrak{L}' \textit{ sont deux à deux disjoints} \\ \sum_{\ell' \in \mathfrak{L}'} \textit{couv}(\ell') \textit{ réalise son maximum sur } \mathfrak{L}' \end{cases}$$

Nous pouvons remarquer à ce stade que le critère de couverture maximum peut subir deux critiques essentielles. Premièrement, les alignements manuels ne sont pas toujours couvrants. Il arrive effectivement que certains mots ne soient pas liés. Mais nous remarquons empiriquement, de même que dans le projet Blinker [95], que les alignements ma-

nuels ont plutôt tendance à être très couvrants, même dans des cas de grande divergence. De plus les fragmentations contigües proposées, X et W, ont tendance à "capturer" des mots non liés au sein d'un fragment, ce qui compte dans l'estimation de la couverture comme un mot liés avec un "*mot vide*". Enfin, la mémoire étant partielle, les alignements reconstruits ne seront en général pas toujours couvrants. Une deuxième critique peut être faite au sujet de la non-unicité d'une solution aux problèmes 1 et 2. Par conséquent, un ensemble solution $\mathfrak{L}'$ ne correspondra pas nécessairement à $\mathfrak{A}_{\ell}^{\Phi\Psi}$ dans le cas du problème de reconstruction. Il existe un cas où $\mathfrak{L}'$ n'est pas à proprement parler une erreur, mais constitue plutôt une solution alternative. Il s'agit d'une situation inhérente à l'annotation manuelle : les **choix contradictoires**. Parfois un annotateur fait des choix différents dans des situations identiques et le phénomène est encore plus visible lorsqu'il y a plusieurs annotateurs. Nous pouvons voir en figure 5.17 un exemple courant pour lequel des annotateurs proposent des fragments différents bien que positionnés sur les mêmes parties de la biphrase.

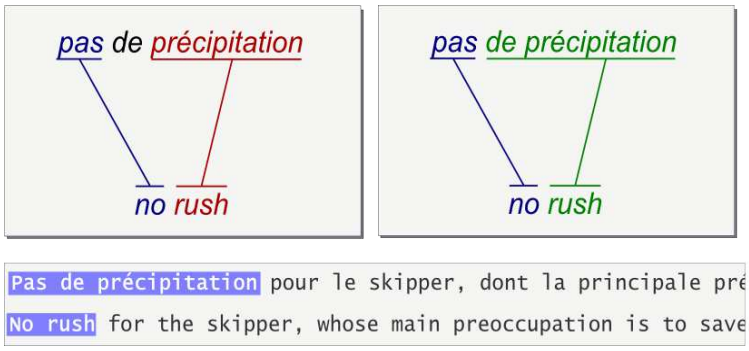


Figure 5.17 – Deux fragments différents compatibles avec la biphrase mais identiquement positionnés

La problème 1, ne tenant pas compte des liens internes aux fragments sélectionnés, pourra proposer autant de solutions alternatives qu'il y a de combinaisons de fragments aux positionnements identiques. Une résolution du problème posé en tant que tel ne pourra faire un choix qu'arbitraire entre ces fragments, il est donc plus judicieux de filtrer au préalable $\mathfrak{L}$ de sorte à ce qu'il ne contienne qu'un seul fragment pour un positionnement dans ℓ donné.

Ce choix peut être géré en amont par l'étape ③ : plutôt que de se contenter de prendre le premièr représentant venu, on a alors la possibilité de faire intervenir différents critères parmi lesquels nous avons retenu la **fréquence** et l'**utilisateur courant**. Dans la mémoire d'alignements, les fragments sont représentés, de même que dans la définition formelle, par trois éléments : le support source, le support cible et les liens internes. Le champ des liens internes est, entre autres, attaché à un compteur d'utilisation (incrémenté à chaque sauvegarde par un utilisateur) et un champ utilisateur[2]. Parmi les listes de liens possibles, sera donc retenue en priorité celle utilisée le plus fréquemment par l'utilisateur en cours. Si celui-ci n'a jamais aligné ce fragment, seul le critère de fréquence compte. Ce qui était en apparence une faiblesse dans le modèle de résolution constitue en fait un avantage pour cette approche collaborative en nous permettant de tenir compte de choix individuels. Pré-filtrer les fragments de même support présenterait l'avantage, pour un utilisateur, de se voir rappeler ses choix les plus fréquents et donc de consolider son style.

5.2.2 Complexités

Le cas général

Nous allons voir que la complexité du **problème de reconstruction** est liée au type de la fragmentation choisie. Intuitivement, plus fine sera la fragmentation plus il sera difficile de "recoller les morceaux". Nous traduisons pour cela le problème 2 sous la forme d'un problème de graphe. Les complexités seront données par rapport au nombre N d'un ensemble de fragments positionnés $\mathfrak{L}$ retourné par l'étape ③.

Nous faisons un bref rappel sur les quelques notions abordées ici : un graphe G est la donnée d'un ensemble de nœuds traditionnellement noté V (comme "*vertex*", c'est-à-dire *"sommet"* en anglais) et d'un ensemble d'arêtes, traditionnellement noté E (comme "*edge*" signifiant "*arête*" en anglais), ainsi $G = (V, E)$. Le graphe G est dit non orienté si les arêtes de E sont des **paires** de sommets : si i et j sont des sommets de V, la paire $\{i, j\}$ peut être une arête de E. Il sera dit orienté si les arêtes de E sont des **couples** :

[2] ces données nous permettant par ailleurs de repérer les patrons syntaxiques les plus utilisés

ici, pour deux sommets i et j, les paires (i,j) et (j,i) peuvent être deux arêtes différents. Une fonction ρ de V dans $\mathbb{R}$ sera appelée une pondération des sommets, et le triplet $G = (V,E,\rho)$ constituera un graphe pondéré. On appelle **clique** de G un ensemble de sommets deux à deux adjacents dans G et **stable** de G un ensemble de sommets deux à deux non adjacents. Le *poids* d'un stable ou d'une clique dans un graphe pondéré sera la somme des poids des sommets concernés. La recherche de clique maximum et de stable maximum (en nombre de sommets), d'une clique de poids maximum ou d'un stable de poids maximum sont des problèmes NP-difficiles classiques en théorie des graphes. De nombreux problèmes concrets s'y ramènent, et ce sera notre cas, en introduisant la notion de *graphe d'incidence* d'un ensemble de fragments positionnés :

Définition 37. *Soit $\mathfrak{L} = \{\ell_1, \cdots, \ell_k\}$ un ensemble de fragments positionnés dans une biphrase B pré-alignée par ℓ. On définit le* ***graphe d'incidence*** *$G_{\mathfrak{L}} = (V,E,\rho)$, où V est l'ensemble des sommets, E l'ensemble des arêtes et ρ une fonction de pondération des sommets de $V \mapsto \mathbb{N}$:*

$$\begin{cases} V = [\![1,k]\!] \text{ (représentant les k fragments)} \\ \{i,j\} \in E \text{ si } \ell_i \text{ et } \ell_j \text{ ne sont pas disjoints} \\ \rho(i) = |supp_S(\ell_i)| + |supp_C(\ell_i)| \text{ (nombre de mots concernés par le i-ème fragment)} \end{cases}$$

Le problème 2 est équivalent au problème de graphe :

Problème 3. *À partir du graphe d'incidence $G_{\mathfrak{L}} = (V,E,\rho)$ associé à $\mathfrak{L}$, sélectionner un stable $V' \subset V$ de poids maximum $\sum_{i \in V'} \rho(i)$*

La recherche d'un stable de poids maximum est NP-difficile pour un graphe général. Si nous exploitons une base de **fragments contigus**, le graphe d'incidence ne correspond pas au cas le plus général, et le problème est en fait équivalent à celui de la recherche d'un *ensemble de rectangles parallèles indépendants de poids maximum*[3] (noté *IR*) :

Un rectangle $R = [a,b] \times [c,d]$ est le produit cartésien de deux intervalles finis de $\mathbb{R}$ (représentable dans le plan euclien par un rectangle dont les côtés sont parallèles aux

[3] *Weighted Independent Sets of Axis Parallel Rectangles*

axes). Deux rectangles R et R' sont dits *indépendants* si sur chaque axe la projection de l'un n'intersecte pas la projection de l'autre. De plus, à chaque rectangle est associé un poids. Le problème *IR* consiste, à partir d'un ensemble de N rectangles, à extraire un sous-ensemble de rectangles indépendants dont le poids est maximum. Le problème 2 est équivalent au problème *IR* dont la fonction de poids est la somme des côtés (voir figure 5.18).

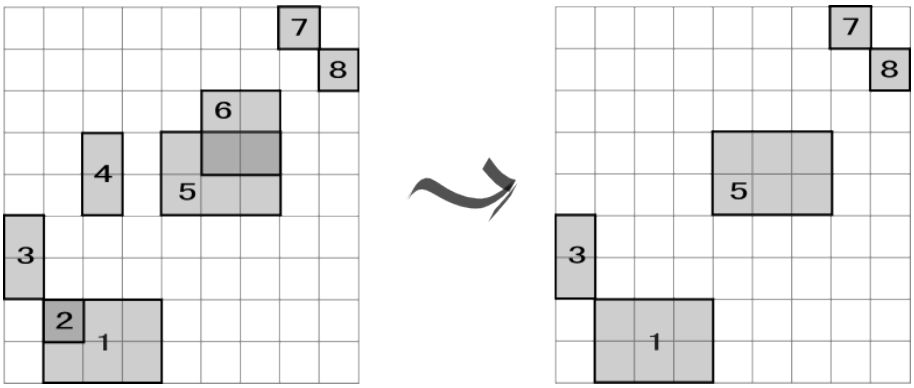


Figure 5.18 – Le même problème sous sa forme "rectangle"

Le problème *IR* reste NP-difficile [11], mais il est **approximable** (classe de complexité *APX*). Un algorithme d'approximation est une heuristique garantissant, à la qualité de la solution fournie, un rapport inférieur à une constante, par-rapport à la qualité optimale d'une solution, pour toutes les instances possibles du problème.

Le *ratio d'approximation* d'un algorithme A approchant *IR*, noté r_A, se définit comme la borne supérieure, sur les ensembles finis de rectangles $\mathscr{R}$, des rapports entre le poids total des rectangles dans une solution optimale $opt(\mathscr{R})$ et le poids total dans la solution approchée $A(\mathscr{R})$.

$$r_A = \sup_{\mathscr{R}} \frac{\rho(opt(\mathscr{R}))}{\rho(A(\mathscr{R}))}$$

Le ratio r_A donne la qualité de l'approximation puisqu'il majore la pire situation. Dans [16], un algorithme d'approximation en $O(N\ log\ N)$ est donné avec un ratio de 3. Il est prouvé dans [101] qu'il n'existe pas d'approximation polynomiale avec un ratio inférieure à $\frac{8668}{8665}$ (sauf si $P = NP$).

Reconstruction par les X-fragments

La fragmentation initiale influe sur l'effort qu'il sera nécessaire de fournir pour **reconstruire** ℓ. Notamment, la reconstruction d'un alignement ℓ à partir de ses fragments de type X (en mémoire $\mathscr{L}_X$) est possible par application *monotone* de fragments **successifs** (voir définition 10). Ainsi, la **reconstruction** de ℓ par des X-fragments reste possible en résolvant une version affaiblie du problème 2 :

Problème 4. *À partir d'un ensemble de fragments positionnés $\mathfrak{L}$, sélectionner un sous-ensemble $\mathfrak{L}' \subset \mathfrak{L}$ tel que :*

$$\begin{cases} \textit{Les fragments de } \mathfrak{L}' \textit{ sont deux à deux } \textbf{\textit{non croisants}} \\ \sum\limits_{\ell' \in \mathfrak{L}'} \textit{couv}(\ell') \textit{ réalise son maximum sur } \mathfrak{L}' \end{cases}$$

Le problème 4 peut aussi se ramener à un problème de graphe connu. À partir d'une collection de fragments $\mathfrak{L}$, nous construisons le *graphe d'incomparabilité* $G'_{\mathfrak{L}}$ qui représente la relation d'ordre *succession* "$<_{\ell}$".

Définition 38. *Soit $\mathfrak{L} = \{\ell_1, \cdots, \ell_k\}$ un ensemble de fragments positionnés dans une biphrase B pré-alignée par ℓ. On définit le* ***graphe d'incomparabilité*** *$G'_{\mathfrak{L}} = (V, E, \rho)$, où V est l'ensemble des sommets, E l'ensemble des arêtes et ρ une fonction de pondération des sommets de $V \mapsto \mathbb{N}$:*

$$\begin{cases} V = [\![1, k]\!] \textit{ (représentant les k fragments)} \\ \{i, j\} \in E \textit{ si } \ell_i \textit{ et } \ell_j \textit{ ne sont pas comparables par } <_B \\ \rho(i) = |supp_S(\ell_i)| + |supp_C(\ell_i)| \end{cases}$$

On remarque que $G_{\mathfrak{L}}$ est un sous-graphe de $G'_{\mathfrak{L}}$.

Le problème non croisant 4 est donc équivalent au problème de graphe :

Problème 5. *À partir du graphe d'incomparabilité $G'_{\mathfrak{L}} = (V, E, \rho)$ associé à $\mathfrak{L}$, sélectionner un stable $V' \subset V$ de poids maximum $\sum\limits_{i \in V'} \rho(i)$*

Un stable de poids maximum peut être trouvé en temps polynomial sur des graphes d'incomparabilité (pour toute relation d'ordre partiel). Une solution naïve consiste à explorer l'ensemble des solutions du problème en parcourant les chemins de longueur maximum dans le graphe complémentaire détransitivé ce qui mène à une complexité cubique. Ici, $G'_{\mathscr{L}}$ est moins général car il s'agit d'un *graphe trapézoïdal* :

Un graphe est **trapézoïdal** si il existe un ensemble de *trapèzoïdes entre deux axes parallèles* correspondants aux sommets tel que deux sommets du graphes sont liés si et seulement si les trapézoïdes correspondants s'intersectent (voir figure 5.19)

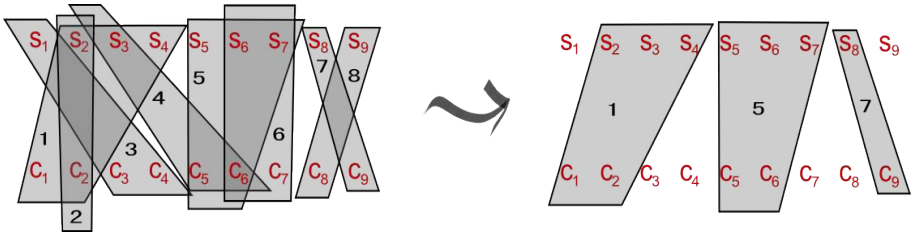


Figure 5.19 – Le problème 4 monotone sur un ensemble de fragments contigus

Un stable de poids maximum dans un graphe trapézoïdal peut être trouvé par un algorithme en $O(N\ log\ N)$ (voir [55]). la reconstruction monotone par les X-fragments est donc assez légère. Nous précisons cependant que les expériences menées en section 6.2.1 emploient une méthode de résolution équivalente non optimisée.

5.2.3 Bilan

La résolution **monotone** basée sur une fragmentation de type X sera exacte pour un alignement déjà rencontré (ou dont la structure syntaxique est identique), c'est-à-dire dans le cadre d'un problème de reconstruction. Les solutions, sont de plus, calculables avec une complexité faiblement polynomiale, ce qui est de bon augure pour l'intégration à l'interface Align$^{\text{It}}$ comme outil automatique en temps réel. De plus, l'alignement monotone étant central en alignement phrastique, l'algorithme pourrait s'y adapter à peu de frais pour le problème d'alignement phrastique. Une limitation évidente apparaît lors de grandes divergences avec les chassés-croisés importants.

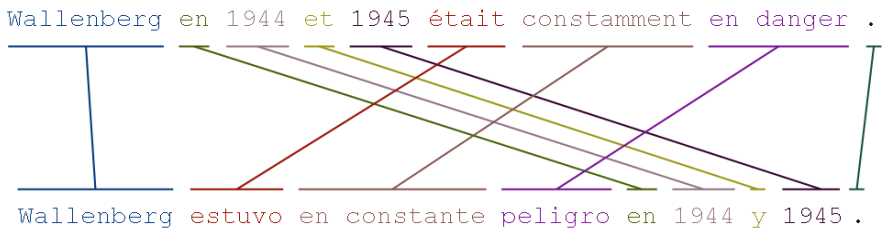


Figure 5.20 – Un cas de chassé-croisé respectant fortement l'hypothèse de cohésion

Les alignements dont des blocs entiers sont permutés tels que celui observé en figure 5.20 ou encore dans les formes passives (respectant l'hypothèse de cohésion de Fox [56]) ne peuvent pas être résolus par un traitement monotone (voir section 6.2.1). En effet, l'alignement résultant ne peut être que partiel, puisque la méthode sera dans l'incapacité de rendre compte du croisement. Nous avons toutefois remarqué que des biphrases analogues, présentant un chassé-croisé clairement marqué, étaient en général alignées avec succès par deux applications successives d'un traitement monotone (au-delà de deux, le résultat est souvent nul ou dégradé).

Nous avons évoqué une résolution naïve plus coûteuse nous permettant de parcourir l'ensemble des solutions au problème 4. Celle-ci a été utilisée en section 6.2.1, nous permettant notamment de retenir parmi les solutions de couverture maximum, celle qui est composée du plus petit nombre de fragments. Ce deuxième critère permet de proposer des solutions moins recomposées et vraisemblablement meilleures.

Il est clair que l'approche monotone ne présentera un intérêt pratique que dans des situations où les langues décrivent des structures relativement proches. Il est toutefois envisageable de composer cette approche avec des techniques de monotonisation par la syntaxe [158] pour des paires plus divergentes, ce qui n'a pas été abordé durant ces travaux.

Par ailleurs, le cas **non monotone** du problème ne présente pas de méthode de résolution exacte et polynomiale. Des solutions envisageables se trouvent du côté des approximations ou des résolutions exactes exponentielles. Des approximations faiblement polynomiales du problème 3 des *rectangles indépendants* existent et peuvent, au même titre que la résolution monotone, être intégrées à l'interface comme des outils d'aide à

l'alignement en temps réel. Ce faisant, il serait intéressant d'évaluer un ratio de performance moyen empirique à comparer au ratio théorique annoncé, relativement aux paires de langues. Des résolutions exactes par des solveurs optimisés ont déjà été utilisées pour l'alignement de fragments [48], mais n'ont pas été envisagées ici. Ce type de résolution ne pourrait évidemment pas être intégré aux outils d'aide à l'alignement, mais pourrait éventuellement revêtir la forme d'un *R-utilisateur*.

CHAPITRE 6

CADRE EXPÉRIMENTAL

Nous présentons dans ce chapitre les corpus (section 6.1.1) ainsi que les outils automatiques d'analyse utilisés pour nos travaux (section 6.1.2). Nous partagerons ensuite des résultats obtenus au cours de quelques expérience. Ils y sont présentés dans l'ordre chronologique : nous avons commencé par évaluer des techniques de reconstruction sur les fragmentations X et W (section 6.2.1) puis tenté d'évaluer la qualité d'un alignement non-expert en en utilisant les différentes métriques introduites (section 6.2.2).

6.1 Ressources utilisées

6.1.1 Des corpus parallèles

Cette partie aborde les ressources parallèles multilingues utlisées pour entraîner les machines de traduction utilisant des données externes, telles que les machines statistiques ou à base d'exemples. Les approches statistiques requièrent des corpus massifs de plusieurs millions de mots. Notre approche pour l'alignement préfèrera travailler sur des quantités plus réduites. Nous présentons ici les corpus que nous avons utilisés en évoquant différentes ressources accessibles en ligne.

6.1.1.1 Les corpus classiques

Les corpus multilingues historiquement les plus utilisés sont généralement des documents juridiques nécessitant une traduction en plusieurs langues comme le corpus bilingue *Hansard* (anglais-français) qui est le journal des débats à la chambre des communes du parlement canadien[1] ou provenant d'organismes internationaux comme pour le *Acquis Communautaire*[2] regroupant les textes de loi de l'UE dans 22 langues. Les

[1] `http://www.parl.gc.ca/HouseChamberBusiness/ChamberSittings.aspx?View=H&langage=F`

[2] `http://langtech.jrc.ec.europa.eu/JRC-Acquis.html`

traductions sont, par nécessité, rigoureuses et de bonne qualité. Ces corpus sont aussi disponibles sur le site *StatMT*[3] dédié à la traduction automatique statistique. On peut également y trouver la première ressource parallèle que nous avons utilisé, le corpus journalistique *News Commentary* disponible en cinq langues provenant du troisième *Workshop on Machine Translation* en 2008 (un extrait peut être observé en annexe I.1). Il est plus petit que le Hansard et le AC (environ 2 millions de mots) mais sa taille est suffisante pour notre approche qui ne requière pas autant de données qu'un aligneur statistique. De plus le thème étant plus général le travail d'annotation ne s'en trouvera que moins rébarbatif. On trouvera de plus amples informations sur ces corpus de qualité ainsi que de nombreuses références par exemple dans [83], dont le corpus littéraire de récits de voyages Carmel [117] en quatre langues.

6.1.1.2 La collaboration en ligne

Ces dernières années, de nombreuses autres ressources en ligne sont venues élargir la quantité de ressources parallèles dans de nombreuses langues, ainsi que leur thématique. De nombreux projets de traduction collaborative, de plus ou moins grande ampleur apportent indirectement leur pierre à l'édifice. On peut saluer à ce titre la mouvance du logiciel libre qui requière pour son expansion la traduction de nombreux documents et fichiers de manière simple et libre. La collection *Opus - "the open parallel corpus"* [4] propose une collection de textes parallèles issus de données librement diffusées sur internet, parmi lesquels nous avons utilisé la documentation KDE4 disponible en plus de 70 langues (nous avons retenu 8 de ces langues). On peut également citer le corpus de sous-titres *OpenSubtitles* dans 30 langues provenant de l'effort massif d'internautes traduisant des sous-titres de films et de séries. Il forme une ressource riche en langage courant et en dialogues avec peu de narration. Il faut noter que les ressources issues de la collection *Opus* ont été alignées (phrase à phrase) par des outils automatiques et ne font l'objet d'aucune correction manuelle. Elles nécessitent généralement un pré-traitement ou un filtrage selon l'utilisation souhaitée.

[3] `http://www.statmt.org`
[4] `http://opus.lingfil.uu.se/`

L'apparition de nombreux fichiers de sous-titres traduits massivement dans plusieurs langues n'est pas passé inaperçu dans le monde de la traduction automatique avec des approches statistique [146] mais aussi à base d'exemples [9]). Les différents formats de fichiers de sous-titrage (srt, sub, ssa, etc...) expriment invariablement pour chaque segment le temps du début de l'affichage et de la fin (un segment du découpage temporel est appelé un *timestamp*). En plus de constituer une ressource massive elle a l'avantage certain d'être structurellement plus à même d'être correctement alignée pour former des corpus parallèles. Il est démontré expérimentalement dans [138] que l'indice supplémentaire du timing permet, par des traitements simples, d'améliorer les résultats des approches traditionnelles basées sur la longueur des phrases. Dans son papier [145], Volk montre qu'il est possible d'augmenter grandement la qualité de l'alignement phrastique en ne retenant que des paires de fichiers sous-titres partageant le même découpage temporel.

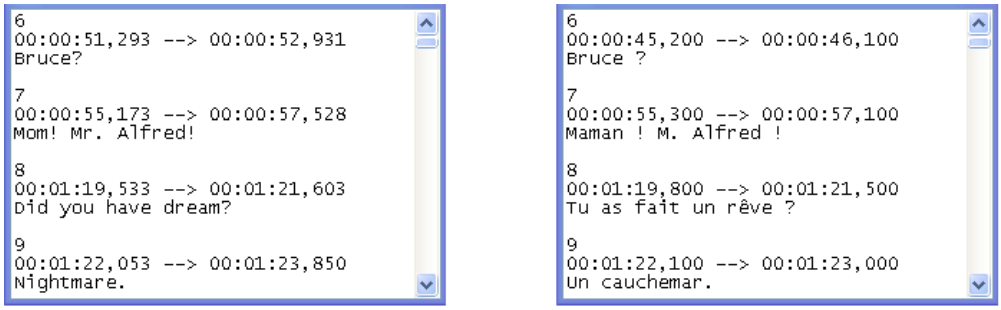


Figure 6.1 – Pour deux fichiers *.srt* ayant le même timing, l'alignement est naturel

6.1.1.3 Les textes dans le jeu vidéo

Nous souhaitons évoquer l'apparition d'une ressource moins connue, provenant également d'un effort collaboratif en ligne et qui partage les bonnes propriétés structurelles des sous-titres : les informations textuelles contenues dans un jeu vidéo (on parlera parfois de *script dump*). Ci et là, des ressources apparaissent, issues du travail désintéressé de traducteurs amateurs, disponibles en ligne et provenant de jeux PC autant que de jeux consoles. L'extraction d'un jeu de console, sous forme de fichier, est couramment appelé

ROM, pour *Read Only Memory* ou *mémoire morte* en référence au support cartouche des débuts du jeux vidéo sur consoles de salon. La récupération et/ou la modification de données des graphismes, des dialogues, des niveaux, du gameplay d'une image *ROM* d'un jeu vidéo se revendique du *ROM-hacking*. Le monde du hack, comme souvent très prolifique, a créé et mis a disposition des outils permettant d'éditer et de modifier les éléments graphiques, textuels et structurels de nombreuses *ROMs* sous différents supports. Les *éditeurs de tuiles* (unités de données graphiques 8x8 pixels) permettent l'édition et la modification graphique de jeux sur les plus anciennes consoles (voir figure 6.2).

Figure 6.2 – La table graphique des caractères d'un jeu *NES* dans un éditeur de tuiles

En ce qui nous concerne plus directement, des *éditeurs hexadécimaux* adaptés en font de même avec les données textuelles. En général, les vieilles *ROMs* ne codent par leurs textes en *ASCII*, mais lorsqu'il n'est pas compressé, une lettre correspond à un octet. Cette correspondance est décrite dans une table qui associe généralement aux lettres de l'alphabet des valeurs hexadécimales contigües. La table est utilisée par l'éditeur pour visualiser ou modifier la *ROM* (6.3).

Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00A23200	26	02	00	00	4A	02	00	00	85	02	00	00	E7	02	00	00
00A23210	6A	03	00	00	8B	03	00	00	F4	03	00	00	55	04	00	00
00A23220	86	04	00	00	AF	04	00	00	B0	04	00	00	B1	04	00	00
00A23230	B2	04	00	00	B3	04	00	00	B4	04	00	00	B5	04	00	00
00A23240	B6	04	00	00	B7	04	00	00	B8	04	00	00	B9	04	00	00
00A23250	BA	04	00	00	BB	04	00	00	BC	04	00	00	00	4C	65	20
00A23260	02	01	4E	65	63	74	61	72	20	52	6F	75	67	65	02	00
00A23270	20	70	65	72	6D	65	74	20	64	65	20	74	72	6F	75	76
00A23280	65	72	0A	70	6C	75	73	20	64	65	20	02	01	43	9C	75
00A23290	72	73	02	00	21	0A	07	2B	0F	00	4C	65	20	02	01	4E
00A232A0	65	63	74	61	72	20	42	6C	61	6E	63	02	00	20	70	65
00A232B0	72	6D	65	74	20	64	65	20	74	72	6F	75	76	65	72	0A
00A232C0	70	6C	75	73	20	64	65	20	02	01	46	72	61	67	6D	65

&	.	.	.	J	.	.	.	…	.	.	.	Ç	.	.	.	
j	.	.	.	‹	.	.	.	ô	.	.	.	U	.	.	.	
†	.	.	.	¯	.	.	.	°	.	.	.	±	.	.	.	
²	.	.	.	³	.	.	.	´	.	.	.	µ	.	.	.	
¶	.	.	.	·	.	.	.	¸	.	.	.	¹	.	.	.	
°	.	.	.	»	.	.	.	¼	.	.	.	L	e			
.	.	N	e	c	t	a	r		R	o	u	g	e	.	.	
	p	e	r	m	e	t		d	e		t	r	o	u	v	
e	r	.	p	l	u	s		d	e		.	.	C	œ	u	
r	s	.	.	!	.	.	+	.	.	L	e		.	.	N	
e	c	t	a	r		B	l	a	n	c	.	.		p	e	
r	m	e	t		d	e		t	r	o	u	v	e	r	.	
p	l	u	s		d	e		.	.	F	r	a	g	m	e	

Figure 6.3 – Le code hexadécimal à gauche est converti à droite via la table

Les outils (parfois spécifiques à certains jeux) ont vu le jour en même temps que les émulateurs, ainsi que de nombreux projets de traduction de *ROMs*. parmi les raisons motivant les équipes de traducteurs, la plus couramment avancée est que le jeu n'a jamais été traduit en une certaine langue. C'était souvent le cas dans les années 80-90 en France quand les jeux avaient tendance à ne sortir qu'en version anglaise, les éditeurs estimant que le public de joueurs (souvent jeunes) s'en contenterait. Cela contribue sans doute de nos jours à l'émergence de ces équipes de passionnés créant de nombreux patchs de traduction français que l'on peut trouver entre autres sur le site de la *TRAF*[5] (Traduction de *ROMs* Anglais Français). Les traductions d'anciens jeux forment de petits projets car les vieux supports étaient limités en mémoire (40Ko pour une cartouche standard de *NES* par exemple), mais après l'avènement du support CD les données textuelles sont devenues beaucoup plus conséquentes. On peut donner l'exemple de l'impressionant travail de traduction effectué sur *Shenmue II* et qui a mobilisé une équipe[6] d'une dizaine de personnes pendant 4 ans pour produire bénévolement une traduction française pour le jeu qui n'était sorti qu'en japonais et en anglais. On peut aussi citer les traductions depuis le japonais vers l'anglais et le portugais de *Snatcher*[7], possédant une importante narration. Par ailleurs, certains titres jamais édités dans certains pays et ayant tout de même rencontré un succès international ont motivé des équipes de traducteurs bénévoles comme ça a été le cas pour le projet "The *Mother 3* fan translation"[8] qui représente deux ans de travail. Il en résulte chaque fois des ressources multilingues, pas forcément alignées phrase à phrase, mais possédant de nombreux marqueurs susceptibles de faciliter l'alignement phrastique, au même titre que les marqueurs temporels pour les sous-titres.

Hormis les ressources issues de projets collaboratifs de traduction, de nombreux jeux ont tout de même été édités dans plusieurs langues avec des tables de dialogues suffisament similaires pour les aligner facilement. Notamment de nombreux marqueurs, correspondant à des déclencheurs dynamiques tels que le déroulement des boîtes de dialogues ou des mimiques de l'avatar sont autant d'indices pour aligner les segments de

[5] `http://traf.romhack.org/`
[6] `http://shenmuemaster.fr/`
[7] `http://junkerhq.net`
[8] `http://mother3.fobby.net/`

textes. C'est ce qui nous a permis d'inclure le petit corpus bilingue tiré des dialogues de *Discworld 2* en anglais et en français (figure 6.4).

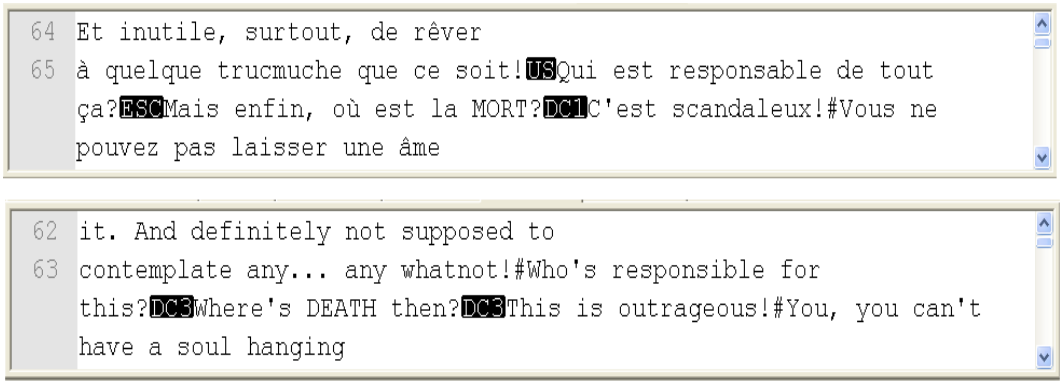


Figure 6.4 – Des marqueurs permettant de séparer les segments, aident à les aligner

Il faut savoir qu'il s'agit d'un jeu très narratif inspiré des oeuvres de l'écrivain Terry Pratchett. La traduction depuis l'anglais est très fidèle et les dialogues d'un style plutôt soutenu sont généralement moins ambigus que ceux des sous-titres du corpus *OpenSubtitles*. Cela tient certainement des situations plus simples que dans un film car moins dynamiques, faisant intervenir moins d'interlocuteurs par scène et avec un contexte plus pauvre. Les dialogues sont aussi sans doute plus clairs que dans un film pour compenser un manque d'expressivité visuelle. Un extrait est donné en annexe I.2. Il y a aussi le cas des jeux multilingues qui permet aux éditeurs de distribuer une seule version dans toute l'Europe par exemple. C'est généralement l'occasion de récupérer une ressource très proche d'un corpus multilingue parallèle, les différents segments de textes étant d'une granularité de l'ordre de la phrase et partageant un index commun entre les différentes langues (ou simplement entrelacés). Le site *Zelda Legends*[9], entretenu par des aficionados de l'univers Zelda, met à disposition de nombreux *dumps* dont ceux des versions européenne et japonaise de *The Wind Waker*. Il en résulte un corpus parallèle en six langues (japonais, français, anglais, allemand, italien, espagnol). Malheureusement, les biphrases sont souvent assez divergentes, car il semblerait que l'effort de traduction ait été fait du japonais vers chacune des cinq autres langues.

[9] `http://www.zeldalegends.net`

6.1.1.4 Traitements effectués sur les corpus

Pour différentes raisons, nous avons préféré corriger manuellement les corpus utilisés. Tout d'abord, il a fallu choisir un encodage commun pour les corpus, sachant que toutes les paires de langues étaient suceptibles d'être affichées dans le navigateur. Nous avons donc converti les corpus en UTF-8 pour une plus grande praticité. Un caractère n'a pas pu être converti depuis l'encodage *Shift JIS* d'un corpus japonais. Il s'agissait d'une sorte de tilde "∼" utilisée comme un tiret entre deux nombres pour désigner un intervalle et nous l'avons remplacé par la première tilde compatible venue. Il faut également noté que certains corpus contenaient de nombreuses erreurs d'encodage résultant de paragraphes copiés-collés sans soin dans un encodage "diffÃ©rent".

Le problème qui aura nécessité le plus d'attention concerne les erreurs de parallélisme : tous les corpus alignés automatiquement souffraient en général d'erreurs d'alignement (seul le corpus Zelda était parfaitement aligné). Nous avons développé une collection d'heuristiques afin de détecter ces erreurs et de les réparer à la main. Il arrivait notamment que deux phrases en langue source soient traduites en langue cible par une seule, étant en fait l'articulation des deux par un connecteur logique, une conjonction de coordination, voire simplement un signe de ponctuation autre que le point. Dans ce cas, les outils automatiques avaient généralement choisi d'aligner la première phrase source et de laisser la deuxième non alignée. Un autre problème intervenant dans le corpus journalistique, qui est une agglomération de nombreux articles de journaux, était l'insertion dans certaines langues de paragraphes ne correspondant à aucun dans les autres langues. Cette erreur locale générait généralement des trous et un décalage se propageant sur quelques paragraphes. On note aussi l'apparition de parties en russe dans le corpus allemand qui provoquaient le même genre de déformation. Une erreur surprenante intervenait également dans ce corpus chaque fois qu'un nombre ordinal ou le mot "*numéro*" étaient représentés avec un petit "o" en exposant (*"n^o"*) : le reste de la phrase disparaissait, probablement à cause d'un outil automatique utilisé en amont ne supportant pas ce caractère. C'est donc en repérant des alignement vides et phrases liées avec des longueurs inappropriées que nous avons détecté et corrigé nombres de ces erreurs.

Le corpus *KDE4* a nécessité un nettoyage important en raison des nombreux marqueurs contenus dans des messages pointant vers des valeurs variable comme dans cet exemple (en français) :

%1 & #160 ; heures %2 & #160 ; minutes %3 & #160 ; secondesMedia time description

Nous avons donc calculé, pour chaque phrase, un ratio entre les caractères alphabétiques (pour les langues latines) ou les kanas/kanji (pour le japonais) et les caractères indésirables comme "%#&" pour exclure celles qui comme dans l'exemple ci-dessus seraient finalement plutôt simples à aligner. À titre d'exemple, le corpus français du bitexte français-anglais passe de 210173 à 121409 phrases après filtrage. Disposant d'une ressource tout de même conséquente pour notre approche et de neuf langues[10], nous espérions obtenir une ressource parallèle multilingue importante, mais il y avait un hic : toutes les phrases du corpus KDE4 ne sont jamais traduites dans toutes les autres langues. En fait l'intersection entre chaque langue diminue singulièrement la quantité de données finales résultant en un corpus novemlingue parallèle de seulement 12618 phrases. Bien que le corpus KDE4 existe en 70 langues, il ne s'agit pas d'un corpus parallèle de 70 langues. Il n'est d'ailleurs pas possible de le télécharger pour certaines paires de langues.

Le corpus *Discworld 2* (anglais/français) a été séparé en segments grâce aux marqueurs observés avant, pour former deux corpus de tailles proches mais pas identiques. Nous avons alors utilisé un aligneur disponible en ligne sur le site *terminotix* [11] puis post édité le résultat, repérant les erreurs comme décrit plus haut en comparant les longueurs de phrases et les blancs. Un problème supplémentaire a dû être pris en compte, les segments coupaient parfois des phrases en plein milieu. Nous avons également tâché de détecter cela par des heuristiques, notamment sur la ponctuation et les majuscules. Le résultat est plutôt bon, mais il faut préciser qu'il s'agit d'un petit corpus de 5418 biphrases plus facilement contrôlable et vérifiable.

En travaillant sur des corpus de tailles raisonnables comme nous le permet notre approche, nous avons essayé de limiter le plus possible des erreurs dues aux outils automatiques et à l'encodage. Nous rappelons que le but étant de proposer des alignements

[10] anglais, allemand, espagnol, italien, français, grec, russe, thai et japonais

[11] `http://terminotix.com`

sous-phrastiques de qualité, il aurait été dommage que les biphrases considérées soient déviantes dès le départ. Nous avons parfois dû modifier légérement le découpage ou la ponctuation de fin de phrase pour corriger les erreurs d'alignement mais aussi remplacer des caractères mal supportés, ou des erreurs d'encodage, ce qui était simple pour le français, mais parfois moins pour le polonais ou le grec. On ne prétend donc pas avoir détecté et corrigé toutes les erreurs, mais grandement amélioré la qualité de certains corpus. Nous donnons un tableau récapitulatif 6.5 des 4 corpus utilisés durant nos travaux.

	# multiphrases	langues
news commentary	42911	**de**, fr, en, es
KDE4	12618	**en**, de, es, it, fr, gr, ru, th, jp
Discworld 2	5418	**en**, fr
Zelda TWW	4282	**jp**, fr, en, es, it, de

Figure 6.5 – Les quatres corpus utilisés. Les langues originales supposées sont en gras

6.1.2 Les analyseurs syntaxiques

Les analyseurs utilisés pour chaque langue, proviennent de différents outils et nous fournissent des structures d'arbres syntaxiques que nous adjoignons aux biphrases alignées via le modèle des *S-SSTC* (section 4.1.2). Celles-ci offrent une grande souplesse en permettant de lier des structures parfois très asymétriques. Nous présentons ici les différents analyseurs utilisés pour les différentes langues, en précisant toutefois qu'elles n'ont pas toutes été au centre des expériences menées. Leur présence est généralement liée aux ressources disponibles et aux outils conseillés au cours d'échanges scientifiques.

Pour le **français**, nous avons utilisé l'analyseur *Sygfran* de Jacques Chauché [35]. Cet analyseur syntaxique (et sémantique) en constituants et en dépendances est basé sur une grammaire de règles de transduction d'arbres. L'arbre produit est un arbre en constituants, pas forcément projectif, tenant compte d'ambiguïtés lexicales et syntaxiques, et fournissant un étiquetage morphosyntaxique assez détaillé. Il est l'analyseur le plus expressif que nous avons utilisé et justifie dès le départ l'utilisation d'une représentation en *SSTC* à cause de certaines structures non projectives.

L'utilisation initiale de *TreeTagger*[123] pour l'**anglais** nous a également permis d'étendre les langues traitées à l'**allemand**, l'**espagnol**, l'**italien** et le **russe**. TreeTagger fourni une lemmatisation et une étiquette morphosyntaxique, de plus, pour l'anglais et l'allemand, les arbres peuvent être d'une plus grande profondeur puisqu'il propose un découpage en "chunks[12]". Dans les deux cas, l'arbre résultant des analyses est toujours projectif. Il d'agit d'une approche probabiliste utilisant des arbres de décision nécessitant à la base un corpus annoté sur lequel TreeTagger doit être entraîné.

Nous avons inclu le **japonais** grâce à l'analyseur *KNP* [84] qui repose lui même sur l'étiqueteur morphosyntaxique *JUMAN*. Le japonais porte la difficulté supplémentaire de ne pas être une langue utilisant des séparateurs graphique tels que l'espace. L'étiqueteur *JUMAN* propose, à l'aide d'un dictionnaire, de passer par différentes segmentations possibles et de marquer la meilleure selon des considérations de fréquence. *KNP* utilise un ensemble de règles propres à traiter le résultat de l'analyse de *JUMAN* pour former des *bunsetsu* [13] dont il tâche ensuite de construire les dépendances. Selon plusieurs options le résultat peut produire un arbre de constituants ou un arbre en dépendances et proposer une lemmatisation, un étiquetage morphosyntaxique ainsi que des informations d'ordre sémantique. En cas d'ambiguïté lexicale, *KNP* pourra retenir comme meilleur candidat, un autre que celui choisi par *JUMAN*. Pour une présentation plus complète des outils automatiques existant pour le japonais ainsi que des notions touchant à sa grammaire, il est possible de se référer à [78].

Le corpus **polonais** a été analysé par *TaKIPI* [111] qui repose lui même sur l'analyseur morphologique *Morfeusz* [152]. Il propose un arbre d'analyse projectif en chunks, une lemmatisation, un étiquetage précis et conserve les ambiguïtés lexicales.

Nous incluons également le **thai** grâce à l'étiqueteur morphosyntaxique *SWATH* (*Smart Word Analysis for Thai*) [94] (en thai) entraîné sur le corpus Orchid [33]. Il fourni une segmentation de la phrase (comme le japonais, le thai n'utilise pas de caractère typographique de séparation) ainsi qu'un étiquetage simple. La structure résultante est donc un arbre plat.

[12] séquence formée de mots contigus et qui n'est pas récursive

[13] notion proche de syntagme

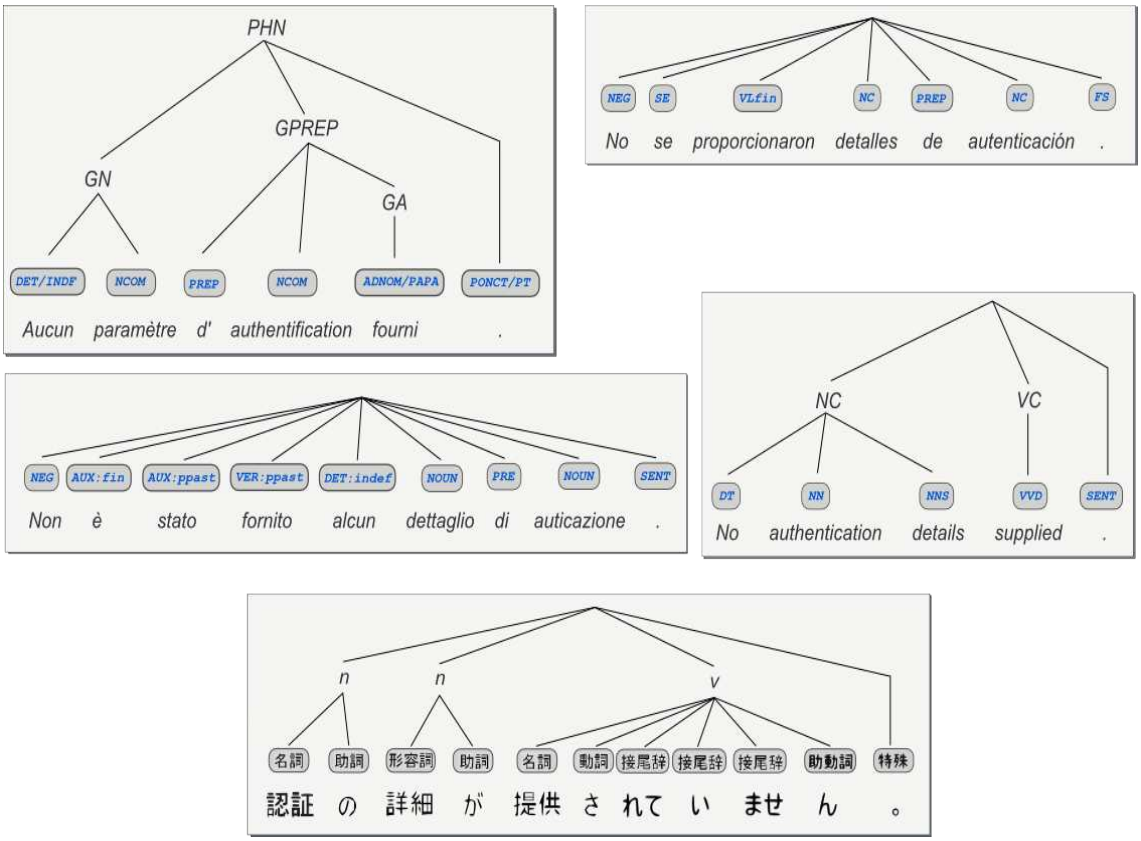


Figure 6.6 – Quelques structures issues des analyseurs évoqués

Ainsi, on voit que les arbres d'analyse mis en correspondance peuvent être assez différents (figure 6.6), et pour certains plus ou moins expressifs. Les arbres plats correspondent à l'utilisation d'un étiqueteur morphosyntaxique. Les *S-SSTC* que nous observons dans notre approche ont cette particularité de ne pas toujours présenter des structures proches (contrairement aux structures observées dans des travaux de parsing bilingue).

Nous avons vu en section 5.1.2 que notre approche permettait de se reposer de manière asymétrique sur une seule analyse (côté source) lorsque la deuxième (côté cible) manque de profondeur, jouant d'une certaine manière le rôle de structure globale de la biphrase. Les langues sans analyseurs de structures syntaxiques peuvent ainsi bénéficier du soutien de l'analyseur de la langue source à la manière des représentation en *Transla-*

tion Corresponding Tree (voir [153]). La différence principale intervenant avec les *TCT* est la présence d'une segmentation et d'informations morphosyntaxique pour la phrase cible.

6.2 Quelques expériences

Nous présentons chronologiquement des expériences et des mesures faites durant nos travaux. La paire de langues principalement abordée est le français/anglais pour la raison simple qu'il s'agissait des langues parlées par les annotateurs bénévoles. Nous verrons toutefois qu'une petite ressource dans la paire français/espagnol a pu être constituée.

6.2.1 La reconstruction par des fragments courts

Nous avons mené nos premières expériences dans le but d'évaluer les processus de **reconstruction** et de **synthèse** (section 5.2.1) sur des fragmentations de types X et W, c'est-à-dire sur des fragments courts et purement syntaxiques [124]. La méthode utilisant la mémoire de fragments X qui a été utilisée correspond à une reconstruction monotone (section 5.2.2). L'algorithme utilisé explore l'ensemble des solutions du problème 4 (maximisation de la couverture) pour sélectionner une famille de fragments aussi réduite que possible. On rappelle que l'algorithme est polynomial et exact.

La méthode utilisant la mémoire de fragments W est une heuristique qui se ramène à l'algorithme monotone. Les fragments W sont comparés et lorsque la composition de plusieurs d'entre eux forme un fragment contigu plus long, il est ajouté à la mémoire. Le nombre de fragments ajoutés est limité arbitrairement par un seuil (200 dans ce cas). Il ne s'agit pas d'une méthode éprouvée et les résultats pourraient être grandement dégradés dans le cas d'une paire de langues plus divergente que le couple français/anglais.

Les premiers résultats (voir table 6.7) ont été obtenus à partir d'une version prototypale de l'outil actuel[14] non accessible en ligne. Ils ont renforcé notre conviction qu'une

[14]La version actuelle de Align$^{\text{It}}$ est toujours en développement. L'interface d'annotation, la structure sous-jacente des alignements et la partie administrant la mémoire d'exemples sont stables. Les algorithmes utilisés pour la première version, n'étant pas optimisés, ne sont pas intégrés à la version actuelle en ligne.

approche d'alignement à base d'exemples syntaxiquement motivés pouvait atteindre une expressivité manuelle pour le problème de reconstruction.

Alignement	**P %**	**R %**	**F**
Reconstruction, fragments W (biphrases $1 \rightarrow 100$)	84	82	0.83
Recons. avec cognats, fragments W (biphrases $1 \rightarrow 100$)	92	86	0.89
Reconstruction, fragments X (biphrases $1 \rightarrow 100$)	98.7	97.2	0.98
Recons. avec cognats, fragments W (biphrases $1 \rightarrow 100$)	99.5	97.7	0.99
Reconstruction, fragments W (biphrases $1 \rightarrow 200$)	85	83	0.84
Recons. avec cognats, fragments W (biphrases $1 \rightarrow 200$)	91	88	0.89
Reconstruction, fragments X (biphrases $1 \rightarrow 200$)	97.9	97.1	0.98
Recons. avec cognats, fragments X (biphrases $1 \rightarrow 200$)	98.9	97.8	0.98
Alignement, fragments W (nouvelles biphrases $101 \rightarrow 200$)	77	52	0.62
Alignement, fragments X (nouvelles biphrases $101 \rightarrow 200$)	82.3	49.9	0.62
IBM4 $f \rightarrow e$ (entraîné sur 67941 biphrases)	75	60	0.67

Figure 6.7 – Les résultats des premières expériences menées

Le corpus journalistique *news commentary* a été manuellement aligné sur 200 biphrases (un extrait est donné en annexe I.1). Le protocole expérimental était le suivant : les 100 premières biphrases ont été manuellement alignées par un seul expert pour former un corpus de test ainsi que deux mémoires de fragments via les fragmentations de types W et X. Une première expérience consistait à reconstruire à partir des bases de fragments ces 100 mêmes biphrases. L'objectif était dé vérifier si l'étape ④ (du schéma 3.16 page 103) sélectionne les bons fragments lorsque tous ceux nécessaires à l'alignement sont effectivement en mémoire et donc retournés par l'étape ③, ce qui n'est pas le cas en général.

Les fragments étant uniquement composés d'informations syntaxiques, l'ambiguïté peut être assez forte. Nous proposons une deuxième expérience avec un léger pré-alignement de cognats. Nous appelons cognats des mots apparentés qui ont une origine commune entre deux langues et qui sont par conséquent identiques ou très proches (e.g. : *"musharraf"-"moucharraf", "judges"-"juges", "unpopularity"-"impopularité"*,...). En nous inspirant de [107], nous utilisons comme indice de similarité, la distance de Levenshtein [89] normalisée par rapport à la longueur des mots (les mots de moins de 4 lettres sont écartés).

Il en résulte un nombre moyen de 4 liens par biphrase[15]. Nous avons vérifié que ce pré-alignement léger ne contenait aucune erreur. La reconstruction est quasi-optimale à partir des fragments X, ce qui est attendu, puisqu'ils sont "taillés sur mesure". Nous avons remarqué que les erreurs provenaient de **choix contradictoires** de la part de l'annotateur sur des liens internes pour des mêmes fragments (voir l'exemple à la figure 5.17 page 164). Ainsi, les alignements produits étaient non pas faux, mais discutables. Ce problème, discuté en section 5.2.1, pourrait être atténué en tenant compte des fréquences d'utilisation des fragments de chaque utilisateur.

La troisième expérience consiste à aligner 100 nouvelles biphrases en utilisant les mémoires constituées par les 100 premières afin d'évaluer la généricité des fragments. Les données sont plutôt réduites et il est difficile d'en tirer des conclusions précises, mais l'on constate d'ores et déjà que le manque de généricité attendu pour les fragments X n'est pas rédhibitoire et que les résultats sont relativement encourageants pour une approche reposant sur un corpus d'entraînement aussi réduit.

Les deux résolutions obtiennent un même score *AER* (voir section 2.3.1.1), mais pour des raisons différentes car l'alignement de fragments X atteint une meilleure précision pour un rappel plutôt faible et celui des fragments W, au contraire, un rappel plus élevé mais une moins bonne précision. L'alignement monotone a tendance à laisser des "trous" lorsqu'un fragment nécessaire manque, tandis que la résolution approchée qui admet des configurations croisantes propose (parfois à tort) des fragments très courts pour combler les vides.

Enfin, la correction manuelle de ces 100 biphrases nous a permis de réitérer les deux premières expériences de reconstruction sur un corpus de 200 biphrases, avec les fragmentations W ou X, un pré-alignement par les cognats, ou pas.

À titre de comparaison, nous donnons les résultats d'un alignement par *Giza++* sur les mêmes biphrases (entraîné avec le modèle IBM 4 sur les 67941 biphrases du corpus français/anglais avant l'intersection multilingue qui le réduit à 42911). Le type d'alignement (asymétrique) est assez différent et il serait pertinent d'envisager une comparaison

[15] Sur le corpus *news commentary*, les phrases ont une longueur moyenne de 27 mots en français et de 23 en anglais.

avec d'autres approches, notamment avec des approches alignant des intervalles de mots. Les résultats sont donnés ici en terme de précision, de rappel et de F-score (analogue à une mesure AER, sauf que les liens n'ont pas été différenciés en deux types *probable* et *sûr*). Il conviendrait également de généraliser ces prémières expériences en utilisant les distances introduites en section 2.3 et des mesures plus adaptées aux alignements manuels, tels que WAA ou CPER.

6.2.2 Accords et désaccords d'une approche non experte

Les distances introduites en section 2.3 sont utilisées pour observer les accords et les désaccords entre utilisateurs de l'interface d'alignements manuels de Align$^{\text{It}}$. Nous avons tâché d'évaluer la qualité potentielle d'une ressource recourant à la participation de bénévoles non experts. Pour cela nous avons différencié deux types d'utilisateurs : les utilisateurs auxquels des instructions similaires à celles du guide Blinker [96] ont été données pour aligner (les annotateurs u_1,u_2 et u_4) et les autres. Les alignements effectués par les différents annotateurs ont été faits sur les corpus *DW2* (français/anglais) et *news commentary* (français/anglais et français/espagnol). Deux exercices différents pouvaient être proposés : aligner une biphrase vierge ou post-éditer un alignement existant. Nous ajoutons de plus un annotateur "singulier" noté $u_\cup$ qui est en fait l'union de plusieurs utilisateurs sporadiques de l'outil Align$^{\text{It}}$, n'ayant pas été "formés".

Nous avons effectué trois expériences : tout d'abord les deux annotateurs u_1 et u_2 ont aligné **en parallèle** 244 biphrases sur le corpus *DW2*. L'annotateur u_1 aligne indépendamment 144 biphrases déjà alignées par des utilisateurs $u_\cup$ [16] sur le corpus *news commentary* français/anglais, tandis que l'annotateur u_2 post-édite 162 alignements de $u_\cup$ [17] sur ce même corpus. Le but de cette deuxième expérience est d'évaluer la qualité d'une approche non experte par un travail indépendant ou par la post-édition. Nous observerons que la post-édition aurait tendance à orienter l'annotateur vers des corrections spécifiques. Enfin, nous avons comparé 208 alignements sur la version franco-espagnole

[16] Ici, $u_\cup$ est constitué de 7 utilisateurs occasionnels dont la quantité de biphrases alignées est respectivement $54, 40, 24, 16, 6, 3$ et 1 (pour un total de 144 biphrases)

[17] Ici, $u_\cup$ est composé de 12 utilisateurs occationnels dont les nombres de biphrases alignées respectifs sont $35, 29, 24, 23, 17, 12, 9, 5, 3, 3, 1$ et 1 (pour un total de 162 biphrases)

du corpus *news commentary* dans un travail de post-édition où l'annotateur u_3 corrigeait l'annotateur u_4 (ce dernier n'ayant pas été accompagné par un guide introductif).

Nous avons utilisé les trois distances introduites pour interpréter les résultats, mais aussi les mesures *WAA* et $CPAR = 1 - CPER$ (que nous adaptons pour former un "*Consistency Phrase Agreement Rate*"). Afin de pouvoir être comparées avec ces mesures à valeurs dans $[0,1]$, nous normalisons nos distances en faisant les remarques suivantes. Prenons deux alignements ℓ_1 et ℓ_2 sur une biphrase avec *n* mots source et *m* mots cible. Pour la *distance d'édition unitaire* et la *distance des divisions*, la plus grande valeur pouvant être atteinte est $n+m-1$. La plus grande valeur pour la *distance d'édition* est $max(n,m)$. Nous définissons trois indices associés à nos distances transformationelles t_1, t_2 et d :

$$I_{t_1}(\ell_1,\ell_2) = 1 - \frac{t_1(\ell_1,\ell_2)}{n+m-1} \qquad I_{t_2}(\ell_1,\ell_2) = 1 - \frac{t_2(\ell_1,\ell_2)}{max(n,m)} \qquad I_d(\ell_1,\ell_2) = 1 - \frac{d(\ell_1,\ell_2)}{n+m-1}$$

On remarquera que nos indices transformationnels donnent des valeurs légèrement plus élevées que les le *WWA* dans le tableau récapitulatif 6.8.

	langues	#	#words	users	type	WAA	CPAR	I_{t_1}	I_{t_2}	I_d
DW2	fr-en	244	4161	u_1,u_2	indpdt	0.85	0.79	0.89	0.89	0.91
news	fr-en	144	6016	u_1,$u_\cup$	indpdt	0.85	0.80	0.90	0.89	0.90
news	fr-en	162	7728	u_2,$u_\cup$	post-ed.	0.88	0.84	0.91	0.90	0.92
news	fr-es	203	11372	u_3,u_4	post-ed.	0.84	0.80	0.87	0.88	0.91

Figure 6.8 – Quatres expériences comparant différents annotateurs

Nous pouvons visualiser, pour ces quatre expériences, les désaccords en nombre minimum d'opérations élémentaires cumulées (voir figures 6.9 et 6.10). Ces courbes théorisent un effort minimal cumulé pour transformer les alignements d'un annotateur en ceux d'un autre suivant certains types de transformation. La distance t_1 compte des transformations déplaçant les mots individuellement (transformations unitaires). La distance t_2 compte des opérations "plus directes" sur des groupes entiers (sauf transfert étendu). La distance d se place entre les deux en comptant des divisions et des agrégations de groupes de mots, mais seulement des insertions et des suppressions unitaires.

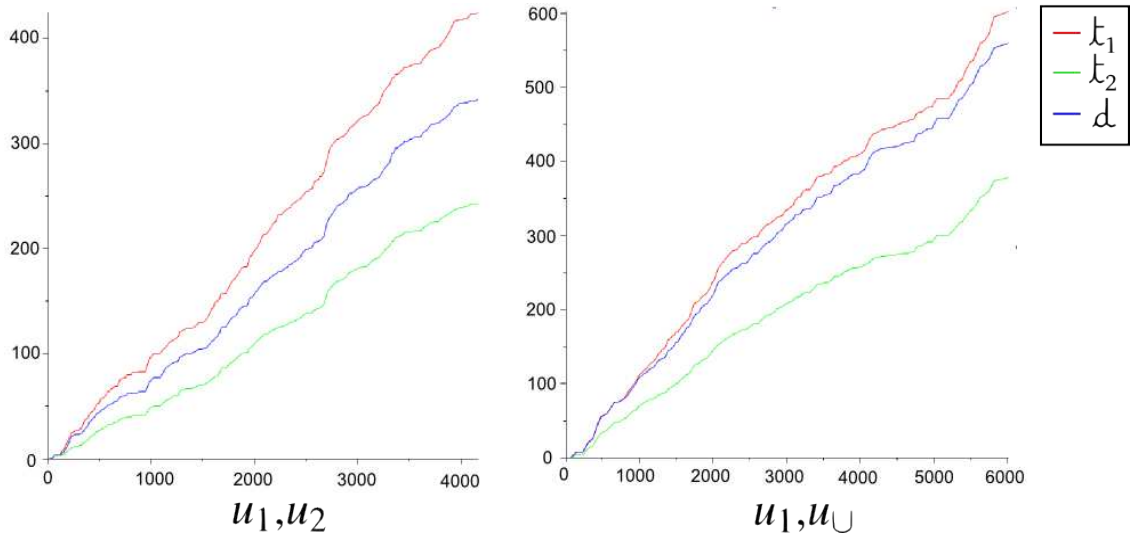


Figure 6.9 – Les distances cumulées exprimées par rapport au nombre de mots parcourus en situation d'alignements indépendants

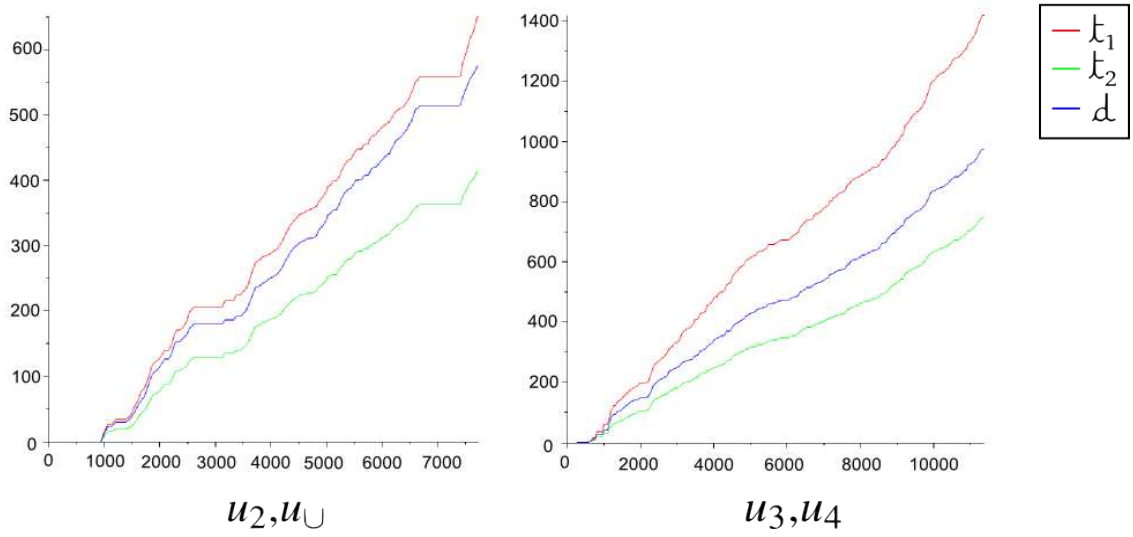


Figure 6.10 – Les distances cumulées exprimées par rapport au nombre de mots parcourus en situation de post-édition

Nous observons que les distances sont très corrélées mais notons que les courbes représentant t_2 sont plus lisses (figures 6.9 et 6.10), car la distance ne compte pas les déplacements de manière individuelle. Les zones stationnaires correspondent à des séries

de biphrases consécutives alignées avec un parfait accord entre les utilisateurs. On peut remarquer qu'elles sont plus marquées dans des travaux de post-édition.

Il y a d'ailleurs, généralement, un accord moyen moindre lors de travaux indépendants. On remarque que la courbe 3 correspondant à la post-édition de $u_{\cup}$ par u_2 est la seule à présenter des paliers aussi marqués. Il y avait un accord très fort entre un utilisateur de $u_{\cup}$ et u_2 du fait de "recopiages". Par contre, la dernière courbe représentant la post-édition de u_3 par u_4 ne présente pas de palier. Le post-éditeur était, en moyenne, en accord fort avec les alignements de u_3, même si il corrige systématiquement les alignements qui lui sont proposés.

La dernière figure montre trois courbes très séparées. Le post-éditeur a avoué trouver que l'annotateur u_4 formait des groupes trop longs et que la grande majorité de ses corrections étaient des raffinements, c'est-à-dire des opérations de transfert/division, ce qui explique la proximité des courbes t_2 et d. A contrario, sur les courbes 2 (deuxième graphique de la figure 6.9) et 3 (premier graphique de la figure 6.10), les courbes les plus rapprochées sont t_1 et d, indiquant une plus grande proportion d'opérations de suppression/insertion, ce qui est symptomatique d'un désaccord aux frontières des groupes formés. On peut interpréter de la même manière la proximité des indices I_d des expériences 3 et 4 tandis que les deux autres indices sont assez différents.

La post-édition serait donc susceptible de provoquer des désaccords internes lorsque le post-éditeur se sent investi d'une mission de raffinement tandis que des problèmes de frontières sont plutôt mis en évidence en confrontant des alignements faits indépendamment. Les deux exercices ont donc leur intérêt pour la construction d'une base de fragments intéressants. Il est possible, en observant le rapport qu'entretiennent nos trois indices I_{t_1}, I_{t_2} et I_d, de distinguer un désaccord majoritairement interne aux fragments d'un désaccord concernant principalement les frontières des groupes liés.

Nous constatons de plus que des annotateurs bénévoles non experts, par une participation même très ponctuelle, peuvent contribuer à produire une ressource de qualité, sans imposer nécessairement la lecture d'un guide introductif tel que celui du projet Blinker [96], ce qui encourage la construction de manière collaborative de ressources de qualités par des non-experts bénévoles.

CHAPITRE 7

CONCLUSION

7.1 Synthèse

L'alignement sous-phrastique pose le problème de reconnaître et de mettre en relation des éléments en relation de traduction entre la partie source et la partie cible de biphrases. En dressant un état de l'art de l'alignement sous-phrastique, principalement traité d'une manière statistique, nous avons constaté que, bien souvent, les modèles de représentation résultaient le plus souvent d'un compromis visant à réduire la complexité des algorithmes de résolution. Il est en effet connu que **les** problèmes d'alignement ne présentent de solutions calculables en temps polynomial que dans les cas particuliers que sont l'alignement monotone et l'alignement de mots. En abordant conjointement les outils d'alignements existants et les différents travaux concernant les phénomènes de divergence, nous souhaitons mettre en évidence l'écart qui existe entre les phénomènes interlingues exprimables et ceux que l'on souhaiterait exprimer. Plus concrètement, cette différence peut être estimée en comparant les espaces d'alignement de différents modèles. L'importance de ce manque d'expressivité est la motivation principale des travaux effectués durant cette thèse et qui s'adressent au problème de l'alignement en s'efforçant de prendre en compte des phénomènes de divergence dont on sait qu'ils forment un ensemble ouvert (voir partie 1.2), ou du moins mal circonscrit.

Le chapitre 2 s'intéresse à définir une notion d'alignement suffisamment expressive pour représenter les différents modèles connus ainsi que les idiosyncrasies d'un alignement manuel. La structuration apportée par ce cadre formel fait de l'espace des alignements décrit un treillis métrique dont les opérations canoniques et les distances transformationnelles associées permettent d'accompagner l'ensemble du propos. Les opérations élémentaires d'affinement et d'élargissement nous permettent, à l'instar des heuristiques d'intersection et *grow-diag-final* [82], de proposer un consensus en cas de désaccord dans le cas général des alignements manuels. Enfin, cet espace est muni de trois distances

inspirées de mesures utilisées en psychométrie et en classification. Elles présentent l'intérêt de constituer des métriques transformationnelles très adaptées à des alignements manuels car interprétables en terme de nombre minimal requis de transformations élémentaires pour passer d'un alignement à un autre. Chacune dépend d'un type de transformation particulier pour exprimer la proximité de deux alignements, permettant de plus de mettre en évidence de subtiles différences structurelles.

En remarquant que l'expressivité la plus générale se trouve dans les alignements manuels, nous orientons l'approche vers une architecture à base d'exemples amorcée par une participation bénévole de non-experts via un outil d'annotation collaboratif en ligne. Le chapitre 3 argumente en faveur d'une telle approche et relativise la notion de qualité défendue par les guides d'annotation pour l'alignement manuel grâce la relation de finesse introduite au chapitre 2 appliquée à des exemples caractéristiques. L'interface d'alignement manuel Align$^{\text{It}}$ propose une intéraction intuitive et favorise une utilisation transparente et collaborative. Il est de plus adapté à des langues sans séparateur typographique (tels que le japonais ou le thai) car les éléments cliquables y sont définis relativement à des analyses provenant d'outils de qualités diverses[1]. L'outil nous a de plus permis de constater expérimentalement la bonne qualité d'alignements non experts entre le français, l'anglais et l'espagnol via deux protocoles d'évaluation décrits au chapitre 6. Les alignements non experts sont comparés à deux références produites par des experts : un ensemble d'alignements produits indépendamment et un autre par post-édition.

Une architecture générale est fixée en seconde partie de ce chapitre 3 qui propose un traitement des alignements manuels pour nourrir une mémoire d'alignements fragmentés réutilisables par des outils automatiques. Cela permet aussi d'adresser une réponse partielle à la question : "*des outils automatiques peuvent-ils produirent des alignements de qualité manuelle ?*" [130]. Nous pensons que c'est possible à condition que les processus développés permettent une intéraction entre utilisateurs humains et "robots". Jusqu'ici, l'alignement automatique a pu aider l'alignement manuel pour accélérer des travaux de post-édition mais, à notre connaissance, jamais les alignement manuels n'ont réellement

[1] L'analyseur du japonais utilisait donnait en plus d'une segmentation, une structure syntaxique complexe et un étiquettage multiple. L'analyseur du thai, plus modeste, offrait tout de même une segmentation et une étiquette

pu aider les outils automatiques à améliorer leur qualité ni leur expressivité, notamment en ce qui concerne le transfert syntaxique.

La structure sous-jacente en S-SSTC est présentée au chapitre 4 où les différentes notations et les définitions sont remaniées autour de la notion d'alignement formalisée au chapitre 2. On y remarque que certaines configurations admises par la définition initiale n'ont que peu d'intérêt linguistique et peuvent être écartées via deux contraintes supplémentaires de bonne formation analogues à la *complétude* et au *non-chevauchement* pour les SSTC. Ces propriétés sont, selon nous, nécessaires à la déduction de correspondances structurelles à partir d'une correspondance lexicale.

Les étapes de fragmentation et d'alignement introduites précédemment par le schéma d'architecture général 3.16 à la page 103 (étape ③ et ④) sont explicitées au chapitre 5. Nous détaillons le modèle permettant d'envisager un processus d'alignement automatique compatible avec un pré-alignement des biphrases via la notion de finesse. Différentes fragmentations sont proposées afin de construire des mémoires de fragments préservant l'expressivité des alignements manuels. Nous constatons qu'une fragmentation trop générique telle que la *B-fragmentation* peut donner lieu à un espace de résolution trop vaste pour une utilisation sans poser de limites arbitraires, contrairement aux fragmentations contigües, plus .

La mémoire évoluant, elle contiendra de nombreux fragments syntaxiques très génériques et réutilisables localement dans différentes biphrases. Pourtant, il n'est pas clair qu'un processus automatique sache faire la distinction entre les fragments utiles et les autres, sources d'erreurs. Nous nous posons donc le problème de la viabilité de l'étape d'alignement en posant la question suivante : "*une biphrase déjà alignée par le passé pourra-t-elle être reconstruite par cette étape ?*". En ramenant l'étape d'alignement à des problèmes connus de type graphe, nous apportons une réponse mitigée. La fragmentation X, par ailleurs assez peu générique pour des paires de langues divergentes, permet une reconstruction en temps faiblement polynomial par une sélection de fragments consécutifs. On ne pourra pas en dire autant des fragmentations plus génériques pour lesquelles le problème est NP-difficile. Ce constat est tout de même nuancé par le fait que le problème de reconstruction à partir des fragments W est approximable avec

un ratio constant.

Nous présentons au chapitre 6 les analyseurs et les corpus concernés par les expériences et disponibles sous forme de biphrases "intéractives" en ligne. Des corpus parallèles "classiques" ont été utilisés, mais pas seulement car nous introduisons une ressource qui, à notre connaissance, n'est pas utilisée dans les domaines de la traduction automatique ni de l'alignement. Il s'agit des "*dump text*" issus de jeux vidéos qui ont pour caractéristique commune avec les fichiers de sous-titres de films, d'être des bitextes parallèles présentant des marqueurs extérieurs forts, bien que moins explicites que la synchronisation temporelle.

Les travaux de recherche effectués durant ces trois années de thèse se sont articulés autour de différents axes. La mise en place d'une architecture générale pour un alignement à base d'exemples. La mise en place d'un cadre théorique propre et adapté aux différents éléments de l'approche. Le développement d'une interface d'annotation collaborative en ligne. L'apport expérimental reste secondaire, mais tout de même présent en ce qui concerne l'évaluation de la qualité de la ressource créée et les processus d'alignement relativement aux fragmentations X et W.

Une première expérience a été menée assez tôt pour évaluer la réutilisation des fragments syntaxiques de types X et W lors de la phase d'alignement évoquée au chapitre 5. Les résultats montrent une réutilisation presque immédiate entre le français et l'anglais et une phase de reconstruction quasi-optimale pour les X-fragments. Une approximation non optimale pour l'alignement de fragments W est proposée pour laquelle un pré-alignement semble tout de même nécessaire.

Le choix de construire une ressource de manière collaborative via la participation d'utilisateurs non experts peut sembler discutable, c'est pourquoi nous nous sommes attelés à évaluer la qualité des alignements obtenus d'une part par des indices classiques d'accord inter-annotateurs, mais aussi grâce à des métriques transformationnelles originales introduites au chapitre 2. Ce faisant, nous avons pu confirmer la bonne qualité de la ressource créée via deux protocoles d'évaluation mais aussi comparer nos métriques aux standards existants. Il en ressort que, ramenées à des indices d'accord, nos trois distances sont corrélées aux mesures connues mais présentent l'avantage supplémentaire d'une inter-

prétation qualitative fine.

Nous avons finalement pu mettre en place une architecture à base d'exemples pour un outil complet d'alignement automatique capable de préserver l'expressivité d'alignements manuels, en déplaçant le compromis qui existait généralement entre complexité et expressivité vers un nouveau compromis, cette fois entre complexité et généricité de la fragmentation. L'outil développé constitue un terrain fertile pour développer de futures expériences translingues. L'approche originale et le cadre formel ouvrent des perspectives intéressantes pour la constitution de ressources de qualité ainsi que pour l'alignement sous-phrastique en général.

7.2 Perspectives

7.2.1 Apports techniques souhaitables

Tout d'abord, évoquons des apports techniques souhaitables pour améliorer l'interface de Align$^{\text{It}}$ sur les plans de l'ergonomie, de la visibilité de la ressource et du fonctionnement collaboratif. L'interface est actuellement fonctionnelle et intuitive, mais pourrait être complétée de différentes manières. Une meilleure visibilité des corpus et des biphrases alignées pourra être apportée grâce à des indicateurs d'avancement et des rendus visuels pour les granularités importantes (paragraphes, ensemble du texte). De plus, même si l'interaction est relativement simplifiée, des raccourcis clavier supplémentaires mériteraient d'être ajoutés, par exemple pour aligner de manière monotone une suite de mots en un seul clic. L'introduction d'un triple clic permettrait de sélectionner des fragments plus longs, plus rapidement. Favoriser plus fortement une intéraction entre utilisateurs pourrait également s'avérer utile pour l'annotation collaborative. Permettre, par exemple, la discussion des alignements in situ via des commentaires permettrait aux utilisateurs de régler des points de désaccord. La possibilité de sélectionner une liste de collaborateurs permettrait de former des groupes d'annotateurs travaillant de concert et ayant une visibilité globale de l'évolution de la ressource commune.

7.2.2 R-utilisateurs trilingues

La forme ouverte et accessible qui caractérise Align$^{\text{It}}$ permet, comme exposé en section 3.2, la définition de "R-utilisateurs", autrement dit, des robots pouvant être associés à des outils. Il pourrait être intéressant de leur confier des missions bien précises comme par exemple la correction automatique de la ponctuation (nombreux sont les annotateurs à ne pas en tenir compte). Des robots d'extension de la mémoire pourraient se voir assigner la tâche d'étendre les alignements à d'autres langues via un fonctionnement triangulaire sur les corpus multilingues où le nombre de langues est supérieur à deux [133]. Aligner via une langue pivot serait pertinent dans le cadre multilingue de Align$^{\text{It}}$. Nous souhaitons à l'avenir implémenter une opération translingue Δ qui permette, lorsque deux alignements ℓ_{FA} et ℓ_{FE} existent respectivement entre deux biphrases (F,A) et (F,E), de proposer un alignement $\ell_{AE} = \ell_{FA} \,\Delta\, \ell_{FE}$ sur la biphrase (A,E). Il en résultera des alignements manquant de finesse mais structurellement corrects si tant est que les deux alignements sur (F,A) et (A,E) le soient. Cela permettrait d'initier la construction d'une mémoire pour une nouvelle paire de langues et d'alléger une partie du travail des annotateurs, pouvant alors post-éditer.

7.2.3 Fragmentations

Bien que les différentes fragmentations proposées préservent l'expressivité, nous avons vu au chapitre 5 que la contrepartie pour des solutions algorithmiques légères se payait au niveau de la généricité. À ce titre, certains phénomènes de divergence, pouvant avoir une influence négative sur la fragmentation, nécessiteraient parfois le relâchement de quelques contraintes.

Pour ce qui est des expériences, nous avons écarté l'utilisation des *B*-fragments, jugés trop peu contrôlables en section 5.1.3. Certaines approches envisagent de traiter les phénomènes non contigus en utilisant des fragments contigus à jokers [134], ce qui limite grandement l'espace de recherche en diminuant l'expressivité dans une mesure linguistiquement acceptable (aucune langue ne parsème l'ensemble de ses phrases de fragments non contigus hautement complexes). Il s'agirait d'un bon compromis qui permettrait

l'utilisation des *B*-fragments tout en les rapprochant d'une géométrie de fragments contigus pouvant bénéficier des résolution déjà proposées. En l'état actuel, les *B*-fragments sont mémorisés en tenant compte du défaut de contiguïté constaté, mais aucune méthode de résolution proposée ici n'en tient réellement compte.

L'expressivité à tout prix peut également se révêler handicapante lorsque, parfois, des fragments non contigus de grandes envergures apparaîssent à cause d'anaphores grammaticales ou d'omissions d'une langue à l'autre (voir section 1.1.3.5). Cela provoque des fragmentations *W* et *X* très grossières et peu génériques. Ce type d'alignement relève, selon nous, d'une problématique autre que l'alignement de fragments et est un cas hors-limite de l'*hypothèse de cohésion* de Fox (section 1.1.3.7). Elle mérite en tant que tel une étude spécifique. Il pourrait être utile d'admettre une exception dans ce cas afin de produire un fragment sous-optimal additionnel qui ne tienne pas compte du défaut de contiguïté, mais permette une réutilisation plus générique (les alignement ℓ_4 et ℓ'_4 à la figure 3.2 page 83 sont un bon exemple).

En ce qui concerne les divergences liées à la *distorsion*, une piste intéressante se dessine du côté des techniques de monotonisation [75] dont pourraient tirer partie les structures S-SSTC de notre modèle. Bien que requérant des outils d'analyse puissants, une piste pour des travaux futurs pourrait consister à coupler cette technique avec notre alignement de fragments monotones itéré sur une paire de langues comme le français/japonais.

En chapitre 1, nous incluions les erreurs d'analyse comme un type de divergence extrême (voir section 1.2). Il est vrai que lorsque deux structures syntaxiques profondes sont opposées, des erreurs d'analyse peuvent entraîner une mauvaise correspondance au niveau des syntagmes de base. Dans ce cas, les fragmentations contigües bilingues (voir section 5.1.2) produiront des fragments très longs et peu expressifs, ce qui n'entraîne donc pas vraiment une grave pollution de la mémoire de fragments, mais réduira la généricité. Des travaux concernant l'harmonisation des analyses [70] mériteraient d'être envisagés ainsi qu'une étude concernant les conséquences des erreurs d'analyse sur la qualité des mémoires d'alignement.

L'analyseur utilisé pour le français représente l'ambiguïté syntaxique de manière

disjonctive par une structure d'arbre double. Nous n'avons pas utilisé cette double structure en général, mais avons choisi arbitrairement la première. Tenir compte d'ambiguïtés importantes pourrait également participer à la réduction des erreurs d'alignement dues aux mauvaises analyses. On note qu'à l'inverse, des travaux de désambiguïsation translingues existent via l'alignement sous-phrastique [93].

Enfin, une des pistes que nous aurions souhaité explorer plus en profondeur est celle de l'impact à différent niveaux d'un pré-alignement. Nous avons pu voir dans la partie 5.1.3 qu'un pré-alignement présentait notamment l'avantage de réduire avantageusement l'ensemble des fragments compatibles, potentiellement volumineux. Or, en l'état prototypal de l'outil, la base de données n'est pas structurée pour traiter de manière optimale ce genre de requête et se limite à filtrer les fragments compatibles pour ne retenir que les fragments dits de propagation (i.e. respectant le pré-alignement), ce qui constitue un gain nul en vitesse d'exécution. Des travaux d'optimisation devraient être menés en ce sens ainsi qu'une étude plus accomplie de techniques de pré-alignement qui sauraient être un angle d'approche adapté aux divergences, améliorer la qualité et éventuellement réduire la complexité algorithmique.

7.2.4 Des solutions du côté de la bioinformatique

De nombreux problèmes auxquels s'adressent les bioinformaticiens partagent des caractéristiques très communes à des problèmes de TAL et notamment en ce qui nous concerne, l'alignement sous-phrastique. Certains travaux mériteraient à ce titre une étude plus minitieuse et plus profonde que ça n'a été le cas dans ces travaux et en général dans le domaine de la traduction automatique. Nous avons notamment constaté des similarités fortes entre les représentations en S-SSTC de type constituants utilisées ici et les problèmes d'*arbre de concensus* abordés en phylogénie où les arbres étiquetés représentent l'histoire évolutive d'un ensemble d'espèces [64].

Par ailleurs, le problème des rectangles parallèles indépendants (problème 3) énonçant formellement le problème d'alignement par des fragments contigus, est connu en bioinformatique sous le nom du problème des *alignements localement non chevauchants* (*Nonoverlapping Local Alignements*). Il a été grandement étudié et de nombreuses solu-

tions d'approximation ont été proposées. On peut citer par exemple [31] qui, pour une instance de n rectangles, propose une solution randomisée donnant une approximation du problème 3 avec un ratio non constant en $log(log(n))$. Il existe dans ce domaine des travaux exploitant des "ancres" pour réduires la complexité du problème. Pour nous, ces ancres correspondent à une situation de pré-alignement. Nous avons vu qu'il était possible de réduire naturellement la complexité en diminuant le nombre de fragments compatibles issus de l'étape ③ du schéma 3.16 d'architecture générale page 103 grâce à un pré-alignement mais nous ne savons pas si un pré-alignement réduit structurellement le problème de synthèse ni si il induit des méthodes de résolution exactes plus intéressantes que pour le problème général.

7.2.5 Le mot de la fin

Les pistes proposées dans le cadre de cette approche collaborative à base d'exemples ne correspondent qu'à une étape initiale de fonctionnement. L'émancipation souhaitée de l'outil automatique est probablement une chimère comme le fait remarquer Martin Kay [76]. En effet, si il est clair que pour l'annotateur, la phase d'amorçage présente une grande valeur d'utilité perçue (différence entre *bénéfice perçu* et *sacrifice perçu*), en s'améliorant, le delta risque fort de diminuer. L'effort devenant considérable pour un faible gain, comme souvent en traitement automatique des langues quand il s'agit de s'attaquer à la longue traîne. Ce sera alors le moment d'envisager des alternatives automatiques à la validation manuelle dans la boucle de rétroaction ①' afin de ne plus avoir recours à l'annotateur que pour des cas difficiles.

BIBLIOGRAPHIE

[1] Mosleh H Al Adhaileh and Enya Kong Tang. The construction of bilingual knowledge bank based on the synchronous sstc annotation schema.

[2] M. H. Al-Adhaileh and E. K. Tang. Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema. In *Machine Translation Summit VII*, pages 244–249, 1999.

[3] Mosleh Hmoud Al-Adhaileh and Tang Enya Kong. A Flexible Example-Based Parser Based on the SSTC. In *COLING-ACL*, pages 687–693, 1998.

[4] Mosleh Hmoud Al-Adhaileh and Enya Kong Tang. A Synchronization Structure of SSTC and Its Applications in Machine Translation. In *The COLING 2002 Post-Conference Workshop on Machine Translation in Asia, Taipei, Taiwan*, pages 1–8, 2002.

[5] Mosleh Hmoud Al-Adhaileh and Enya Kong Tang. Synchronous Structured String-Tree Correspondence (S-SSTC). In *IASTED02, Innsbruck, Austria*, pages 279–275, 2002.

[6] Yaser Al-Onaizan and Kishore Papineni. Distortion Models for Statistical Machine Translation. In *ACL*, 2006.

[7] Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. In *COLING-ACL*, pages 41–47, 1998.

[8] P. Arabie and S. A. Boorman. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10 :148–203, 1973.

[9] Stephen Armstrong, Andy Way, Colm Caffrey, Marian Flanagan, Dorothy Kenny, and Minako O Hagan. Improving the quality of automated DVD subtitles via example-based machine translation. In *Proc. of Translating and the Computer*, volume 28, 2006.

[10] Necip Fazil Ayan and Bonnie J. Dorr. Going Beyond AER : An Extensive Analysis of Word Alignments and Their Impact on MT. In *ACL*, 2006.

[11] Vineet Bafna, Babu O. Narayanan, and R. Ravi. Nonoverlapping Local Alignments (weighted Independent Sets of Axis-parallel Rectangles). *Discrete Applied Mathematics*, 71(1-3) :41–53, 1996.

[12] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 5(2) :179–190, 1983.

[13] Mohit Bansal, Chris Quirk, and Robert C. Moore. Gappy Phrasal Alignment By Agreement. In *ACL*, pages 1308–1317, 2011.

[14] Y. Bar-Hillel. The Present State of Research in Machine Translation. *Massachusetts Institute of Technology Mimeograph*, 1952.

[15] Richard Beaufort et al. Une approche hybride traduction/correction pour la normalisation des SMS. In *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10)*, 2010.

[16] Piotr Berman, Bhaskar DasGupta, and S. Muthukrishnan. Simple approximation algorithm for nonoverlapping local alignments. In *SODA*, pages 677–678, 2002.

[17] Y. Bey, C. Boitet, and K. Kageura. The TRANSBey Prototype : An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators. In *Proc. 3rd International Workshop on Language Resources for Translation Work, Research and Trainning (LR4Trans-III)*, volume 1, pages 49–54, 2006.

[18] Christian Boitet and Yusoff Zaharin. Representation trees and string-tree correspondences. In *COLING*, pages 59–64, 1988.

[19] G. Bonnin and V. Prince. Emphasizing Syntax for French to German Machine Translation. In *Proceedings of the Seventh Symposium on Natural Language Processing, Pattaya, Thailand*, pages 12–20, 2007.

[20] S.A. Boorman and D.C. Olivier. Metrics on spaces of finite trees. *Journal of Mathematical Psychology*, 10 :26–59, 1973.

[21] D. Bourigault. *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'habilitation à diriger les recherches, 2007.

[22] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2) :79–85, 1990.

[23] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning Sentences in Parallel Corpora. In *ACL*, pages 169–176, 1991.

[24] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, 19(2) :263–311, 1993.

[25] Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4) :467–479, 1992.

[26] Ralf D. Brown. Example-Based Machine Translation in the Pangloss System. In *COLING*, pages 169–174, 1996.

[27] Ralf D. Brown. Brown-Adding linguistic knowledge to a lexical example-based translation system. *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1999.

[28] Jaime G. Carbonell, Katharina Probst, Erik Peterson, Christian Monson, Alon Lavie, Ralf D. Brown, and Lori S. Levin. Automatic Rule Learning for Resource-Limited MT. In *AMTA*, pages 1–10, 2002.

[29] Michael Carl. A Computational Framework for a Cognitive Model of Human Translation. In *Translating and the Computer Conference 2010*, 2010.

[30] E. Cazal, C. Serp, M. Roche, and L. Laurent. Extraction de terminologie pour l'ancien français : la quête du graal. In *Proceedings of atelier FDC07 (Fouille de Données Complexes dans un processus d'extraction des connaissances) à la conférence EGC2007*, pages 11–20, 2007.

[31] Parinya Chalermsook. Coloring and Maximum Independent Set of Rectangles. In *APPROX-RANDOM*, pages 123–134, 2011.

[32] J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase Detectives - A Web-based Collaborative Annotation Game. In *In Proceedings of I-Semantics*, 2008.

[33] T. Charoenporn, V. Sornlertlamvanich, and H. Isahara. Building A Large Thai Text Corpus - Part-Of-Speech Tagged Corpus : ORCHID. In *Proceedings of NLPRS'97*, 1997.

[34] I. Charon, L. Denoeud, O. Hudry, et al. Maximum de la distance de transfert à une partition donnée. *Mathematiques et Sciences humaines*, pages 45–86, 2007.

[35] J. Chauché. Un outil multidimensionnel de l'analyse du discours. In *Proceedings of the 10th international conference on Computational linguistics*, COLING '84, pages 11–15, Stroudsburg, PA, USA, 1984. Association for Computational Linguistics.

[36] B. Chen, M. Haddara, O. Kraif, and G. M. de Montcheuil. Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale. In *Actes de TALN'05)*, volume 1, pages 415–418, 2005.

[37] C. Chenon. *Vers une meilleure utilisabilité des mémoires de traduction, fondée sur un alignement sous-phrastique*. 2005.

[38] Colin Cherry and Dekang Lin. A probability model to improve word alignment. In *41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–95, July 2003.

[39] Colin Cherry and Dekang Lin. Inversion transduction grammar for joint phrasal translation modeling. In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics Workshop on Syntax and Structure in Statistical Translation*, 2007.

[40] David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *ACL*, 2005.

[41] Marta R. Costa-Jussà, José A. R. Fonollosa, and Enric Monte. Recursive alignment block classification technique for word reordering in statistical machine translation. *Language Resources and Evaluation*, 45(2) :165–179, 2011.

[42] Louise-Amélie Cougnon and Richard Beaufort. SSLD : a French SMS to standard language dictionary. In *Proc. eLexicography in the 21st century : New applications, new challenges (eLEX 2009)*, pages 11–20, 2009.

[43] Lambros Cranias, Harris Papageorgiou, and Stelios Piperidis. A Matching Technique In Example-Based Machine Translation. In *COLING*, pages 100–104, 1994.

[44] Josep Crego and François Yvon. Gappy translation units under left-to-right SMT decoding. In *Proc. of EAMT*, 2009.

[45] Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 515–521, 1994.

[46] P.C. Davis. *Stone Soup Translation : The Linked Automata Model*. Ohio State University, 2002.

[47] William H. E. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, 1(3) :269–287, 1981.

[48] John DeNero and Dan Klein. The Complexity of Phrase Alignment Problems. In *ACL (Short Papers)*, pages 25–28, 2008.

[49] Lee Raymond Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3) :297–302, July 1945.

[50] H. Déjean and E. Gaussier. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22, 2002.

[51] Bonnie J. Dorr. Parameterization of the Interlingua in Machine Translation. In *COLING*, pages 624–630, 1992.

[52] Bonnie J. Dorr. Machine translation divergences : a formal description and proposed solution. *Comput. Linguist.*, 20(4) :597–633, December 1994.

[53] Abdessamad Echihabi and Daniel Marcu. A Noisy-Channel Approach to Question Answering. In *ACL*, pages 16–23, 2003.

[54] C. Fairon, J.R. Klein, and S. Paumier. *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête Faites don de vos SMS à la science*. Cahiers du Cental. Presses universitaires de Louvain, 2006.

[55] Stefan Felsner, Rudolf Müller, and L. Wernisch. Trapezoid Graphs and Generalizations, Geometry and Algorithms. *DISCRETE APPLIED MATHEMATICS*, 74 :13–32, 1993.

[56] Heidi Fox. Phrasal cohesion and statistical machine translation. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2002.

[57] William A. Gale and Kenneth Ward Church. A Program for Aligning Sentences in Bilingual Corpora. In *ACL*, pages 177–184, 1991.

[58] William A. Gale and Kenneth Ward Church. Identifying Word Correspondences in Parallel Texts. In *HLT*, 1991.

[59] William A. Gale and Kenneth Ward Church. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1) :75–102, 1993.

[60] Éric Gaussier. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *COLING-ACL*, pages 444–450, 1998.

[61] Fabrizio Gotti, Philipphe Langlais, and Claude Coulombe. Vers l'intégration du contexte dans une mémoire de traduction sous-phrastique : détection du domaine de traduction. In *TALN*, 2006.

[62] Fabrizio Gotti, Philipphe Langlais, Elliott Macklovitch, Didier Bourigault, Benoit Robichaud, and Claude Coulombe. 3gtm : A third-generation translation memory. In *3rd computational Linguistics in the North-East (CLiNE) Workshop*, 2005.

[63] Declan Groves, Mary Hearne, and Andy Way. Robust sub-sentential alignment of phrase-structure trees. In *COLING*, 2004.

[64] Sylvain Guillemot. *Approches combinatoires pour le consensus d'arbres et de séquences*. Thèse de doctorat, 2008.

[65] Jin Guo. Critical Tokenization and its Properties. *Computational Linguistics*, 23(4) :569–596, 1997.

[66] Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. Improving alignment for SMT by reordering and augmenting the training corpus. 2009.

[67] S. Huet, B. Julien, and P. Langlais. Intégration de l'alignement de mots dans le concordancier bilingueTransSearch. In *Actes de la 16è Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, 2009.

[68] W.J. Hutchins. *Machine translation : past, present, future*. 1986.

[69] Kenji Imamura. A Hierarchical Phrase Alignment from English and Japanese Bilingual Text. In *CICLing*, pages 206–207, 2001.

[70] Kenji Imamura. Hierarchical Phrase Alignment Harmonized with Parsing. In *NLPRS*, pages 377–384, 2001.

[71] R. Jackendoff. *Semantic Cog*. Current Studies in Linguistics Series. Mit Press, 1983.

[72] R. Jackendoff. *Semantic Structures*. Current Studies in Linguistics Series. M. I. T. P., 1992.

[73] Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. Learning Translation Templates From Bilingual Text. In *COLING*, pages 672–678, 1992.

[74] Megumi Kameyama, Ryo Ochitani, and Stanley Peters. Resolving Translation Mismatches with Information Flow. In *ACL*, pages 193–200, 1991.

[75] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. Novel reordering approaches in phrase-based statistical machine translation. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 167–174, 2005.

[76] Martin Kay. The proper place of men and machines in language translation. Technical Report CSL-80-11, Xerox Res. Cent., Palo Alto, CA, Oct 1980. A version of this paper to appear in Stat. Meth. Linguistics.

[77] Martin Kay and Martin Röscheisen. Text-Translation Alignment. *Computational Linguistics*, 19(1) :121–142, 1993.

[78] Jae Dong Kim. *Analyse syntaxique du japonais*. Mémoire de d.e.a., 2003.

[79] Jae Dong Kim. *Chunk alignment for Corpus-Based Machine Translation*. Thèse de doctorat, 2007.

[80] Chunyu Kit, Haihua Pan, and Jonathan J. Webster. Example-Based Machine Translation : A New Paradigm. In *Translation and Information Technology*, pages 57–78. Chinese University of HK Press, 2000.

[81] Kevin Knight. Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, 25(4) :607–615, 1999.

[82] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. In *HLT-NAACL*, 2003.

[83] O. Kraif. Qu'attendre de l'alignement de corpus multilingues ? *Revue Traduire, 4e Journée de la traduction professionnelle, Société Française des Traducteur*, (210) :17–37, 2006.

[84] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Comput. Linguist.*, 20(4) :507–534, December 1994.

[85] M. Lafourcade. Making People Play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, 2007.

[86] Philippe Langlais and Michel Simard. Récupération de segments sous-phrastiques dans une mémoire de traduction. In *Actes de TALN 2001*, 2001.

[87] Adrien Lardilleux. *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. PhD thesis, Université de Caen, 2010.

[88] Yves Lepage and Étienne Denoual. Purest ever example-based machine translation : Detailed presentation and assessment. *Machine Translation*, 29 :251–282, 2005.

[89] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 :707, 1966.

[90] Lieve Macken. An Annotation Scheme and Gold Standard for Dutch-English Word Alignment. In *LREC*, 2010.

[91] D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on EmpiricalMethods in Natural Language Processing, EMNLP*, 2002.

[92] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2) :313–330, 1993.

[93] Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. Structural matching of parallel texts. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 23–30, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.

[94] Surapant Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijsirikul. Feature-based Thai Word Segmentation, 1997.

[95] Dan Melamed. Manual annotation of translational equivalence : The blinker project. Technical report, Institute for Research in Cognitive Science, Philadelphia, 1998.

[96] I. Dan Melamed. Annotation Style Guide for the Blinker Project. *CoRR*, cmp-lg/9805004, 1998.

[97] I. Dan Melamed. Models of Translational Equivalence among Words. *Computational Linguistics*, 26(2) :221–249, 2000.

[98] Arul Menezes and Stephen D. Richardson. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *ACL*, 2001.

[99] Magnus Merkel. Annotation Style Guide for the PLUG Link Annotator. Technical report, Linköping University, 1999.

[100] Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proc. of the international NATO symposium on Artificial and human intelligence*, pages 173–180, New York, NY, USA, 1984. Elsevier North-Holland, Inc.

[101] Hiroyuki Nagashima and Koichi Yamazaki. Hardness of approximation for non-overlapping local alignments. *Discrete Appl. Math.*, 137(3) :293–309, March 2004.

[102] N. Nassr and M. Boughanem. Croisement de langues en recherche d'information : traduction et désambiguisation de requêtes par phrases alignées. In *INFORSID 2002, XXème Congrès INFORSID IRIN, Polytech' Nantes*, pages 135–142, 2002.

[103] P. Nesi and H. Howarth. The teaching of collocations in eap. Technical report, University of Leeds, 1996.

[104] S. Nirenburg, C. Domashnev, and D. J. Grannes. Two Approaches to Matching in Example-Based Machine Translation. In *TMI*, pages 47–57, 1993.

[105] Franz Josef Och. An efficient method for determining bilingual word classes. In *EACL'99 : Ninth Conference of the EuropeanChapter of the Association for Computational Linguistics*, pages 71–76, 1999.

[106] Franz Josef Och and Hermann Ney. Improved Statistical Alignment Models. In *ACL*, 2000.

[107] Sylwia Ozdowska. *ALIBI, un système d'ALIgnement BIlingue à base de règles de propapagation syntaxique*. Thèse de doctorat, 2006.

[108] Sylwia Ozdowska. Trois expériences d'évaluation dans le cadre du développement d'un systèmed'alignement sous-phrastique. In *ACL*, pages 93–114, 2007.

[109] Sylwia Ozdowska and Vincent Claveau. Inferring Syntactic Rules for Word Alignment through Inductive Logic Programming. In *LREC*, 2010.

[110] Rachel Panckhurst, Catherine Détrie, Bertrand Verine, Claudine Moïse, Mathieu Roche, and Cédric Lopez. sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée., 2011.

[111] Maciej Piasecki. Polish Tagger TaKIPI : Rule Based Construction and Optimisation. *Task Quarterly*, 11 :151–167, 2007.

[112] D. Porumbel, J-K. Hao, and P. Kuntz. An Effcient Algorithm for Computing the Distance between Close Partitions. *Discrete Applied Mathematics*, 2011.

[113] V. Prince and J. Chauché. Emphasizing Syntax for French to German Machine Translation. In *Proceedings of the 8th WSEAS International Conference on Applied Computer and Applied Computational Science*, 2009.

[114] Chris Quirk, Chris Brockett, and William B. Dolan. Monolingual Machine Translation for Paraphrase Generation. In *EMNLP*, pages 142–149, 2004.

[115] Christopher Quirk, Arul Menezes, and Colin Cherry. Dependency Treelet Translation : Syntactically Informed Phrasal SMT. In *ACL*, 2005.

[116] S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *Mathématiques et Sciences Humaines*, 82 :13–29, 1983.

[117] Claude Richard, Régis Meyer, and Marc El-Bèze. Projet CARMEL : récits de voyages. In *Actes de TALN'06)*, 2006.

[118] Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu O. Mittal, and Yi Liu. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *ACL*, 2007.

[119] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[120] Fatiha Sadat and Alexandre Terrasa. Exploitation de Wikipédia pour l'Enrichissement et la Construction des Ressources Linguistiques. In *TALN 2010, Montréal*, 2010.

[121] Markus Saers and Dekai Wu. Improving phrase-based translation via word alignments from Stochastic Inversion Transduction Grammars. In *Proceedings of SSST-3, Third Workshop on Syntax and Structure in Statistical Translation (at NAACL HLT 2009)*, pages 28–36, 2009.

[122] K. P. Scannell. Machine translation for closely related language pairs. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC06*, pages 103–107, 2006.

[123] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees, 1994.

[124] Johan Segura and Violaine Prince. Alignment memories : A useful tool to handle phrase alignment bottleneck. In *CLA'2011 : Computational Linguistics-Applications*, 2011.

[125] Johan Segura and Violaine Prince. Two memory-based methods for phrase alignment. In *LTC'11 : Language and Technology Conference*, 2011.

[126] Johan Segura and Violaine Prince. Using alignment to detect associated multiword expressions in bilingual corpora. *Tralogy - En ligne*, 2011.

[127] Johan Segura and Violaine Prince. Aligning through divergence. In *SNLP-AOS 2011 : The Joint International Symposium on Natural Language Processing and Agricultural Ontology*, 2012.

[128] Violeta Seretan. *Collocation extraction based on syntactic parsing*. PhD thesis, University of Geneva, 2008.

[129] Violeta Seretan. Extraction de collocations et leurs équivalents de traduction à partir de corpus parallèles. *TAL*, 50(1) :305–332, 2009.

[130] Anders Søgaard. Can inversion transduction grammars generate hand alignments ? In *EAMT Workshop*, 2010.

[131] Louis W. Shapiro and A. B. Stephens. Bootstrap Percolation, the Schröder Numbers, and the N-Kings Problem. *SIAM J. Discrete Math.*, 4(2) :275–280, 1991.

[132] Stuart M. Shieber and Yves Schabes. Synchronous Tree-Adjoining Grammars. In *COLING*, pages 253–258, 1990.

[133] M. Simard. Text-translation Alignment : Three Languages Are Better Than Two. In *Proceedings of EMNLP/VLC-99, College Park, MD*, 1999.

[134] M. Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, P. Langlais, A. Mauser, and K. Yamada. Traduction automatique statistique avec des segments discontinus. In *Proceedings of TALN 05*, 2005.

[135] Harold Somers and Gabriela Fernandez Diaz. Translation Memory vs. Example-based MT : What is the difference ? *International Journal of Translation*, 16(2) :5–33, 2004.

[136] Benjamin Taskar, Simon Lacoste-Julien, and Dan Klein. A Discriminative Matching Approach to Word Alignment. In *HLT/EMNLP*, 2005.

[137] Jörg Tiedemann. Word alignment - step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics NODALIDA99*, pages 216–227, 1999.

[138] Jörg Tiedemann. Improved sentence alignment for movie subtitles. In *Proceedings of RANLP, Borovets, Bulgaria*, pages 515–521, 2007.

[139] Christoph Tillman. A unigram orientation model for statistical machine translation. In *HLT-NAACL 2004 : Short Papers*, pages 101–104, 2004.

[140] John Tinsley, Ventsislav Zhechev, Mary Hearne, and Andy Way. Robust Language Pair-Independent Sub-Tree Alignment. In *Proc. of Machine Translation Summit XI*, pages 467–474, 2007.

[141] Pim van der Eijk. Automating the Acquisition of Bilingual Terminology. In *EACL*, pages 113–119, 1993.

[142] Jacques Vergne and Emmanuel Giguet. Regards théoriques sur le tagging. *Actes de Traitement Automatique des Langues Naturelles (TALNŠ98)*, pages 22–31, 1998.

[143] David Vilar, Jan-Thorsten Peter, and Hermann Ney. Can we translate letters ? In *In Proceedings of the ACL Workshop on Statistical Machine Translation, Prague, Czech Republic*, 2007.

[144] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-Based Word Alignment in Statistical Translation. In *COLING*, pages 836–841, 1996.

[145] M. Volk, R. Sennrich, C. Hardmeier, and F. Tidström. Machine translation of TV subtitles for large scale production. In *Proceedings of the 2nd Joint EM+/CNGL Workshop on "Bringing MT to the User"*, pages 53–62, 2010.

[146] Martin Volk. The Automatic Translation of Film Subtitles. A Machine Translation Success Story ? *JLCL*, 24(3) :115–128, 2009.

[147] T. Watanabe, K. Imamura, and E. Sumita. Statistical machine translation based on hierarchical phrase alignment. In *TMI*, pages 188–198, 2002.

[148] T. Watanabe, E. Sumita, and H. Okuno. Chunk-based statistical translation. In *ACL*, pages 303–310, 2003.

[149] W. Weaver. Translation (1949). *Machine Translation of Languages, MIT Press*, 1955.

[150] Eric Wehrli, Violeta Seretan, and Luka Nerima. Sentence Analysis and Collocation Identification. In *Proceedings of the Workshop on Multiword Expressions : from Theory to Applications (MWE 2010)*, pages 27–35, Beijing, China, August 2010. Association for Computational Linguistics.

[151] Julian West. Generating trees and the Catalan and Schröder numbers. *Discrete Mathematics*, 146(1-3) :247–262, 1995.

[152] M. Wolinski. Morfeusz - a practical tool for the morphological analysis of Polish. In *Intelligent Information Processing and Web Mining, Advances in Soft Computing*, pages 551–520, 2006.

[153] Fai Wong, Dong-Cheng Hu, Yu-Hang Mao, and Ming-Chui Dong. A Flexible Example Annotation Schema : Translation Corresponding Tree Representation. In *COLING*, 2004.

[154] Dekai Wu. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3) :377–403, 1997.

[155] Dekai Wu. *CRC Handbook of Natural Language Processing (chapter Alignment)*. CRC Press, 2010.

[156] Dekai Wu and Xuanyin Xia. Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation*, 9(3-4) :285–313, 1994.

[157] F. Xia and M. McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling*, 2004.

[158] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. Using a dependency parser to improve SMT for subject-object-verb languages. In *NAACL 2009 : Proceedings of Human Language Technologies, The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253.

[159] Kenji Yamada and Kevin Knight. A Syntax-based Statistical Translation Model. In *ACL*, pages 523–530, 2001.

[160] David Yarowsky. One Sense Per Collocation. In *Proceedings, ARPA. Human Language Technology Workshop*, pages 266–271, 1993.

[161] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Human Language Technology Conference*, pages 109–116, 2001.

[162] V. Zampa and M. Lafourcade. PtiClic et PtiClic-Kids : jeux avec les mots permettant une acquisition lexicale par le joueur et par la machine. *STICEF*, 18 :135–157, 2012.

[163] Richard Zens and Hermann Ney. A Comparative Study on Reordering Constraints in Statistical Machine Translation. In *ACL*, pages 144–151, 2003.

[164] Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-Based Statistical Machine Translation. In *KI*, pages 18–32, 2002.

[165] Ying Zhang, Ralf D. Brown, Robert E. Frederking, and Alon Lavie. Preprocessing of Bilingual Corpora for Mandarin-English EBMT. In *In Proceedings of MT Summit VIII, Santiago de Compostela, Spain*, 2001.

Annexe I

Extraits de corpus

I.1 Corpus *news commentary*

Le dernier numéro de Moucharraf ?	Musharraf's Last Act ?	¿ El último acto de Musharraf ?
Le Général Moucharraf est apparu sur la scène nationale le 12 octobre 1999, lorsqu'il a forcé le gouvernement élu à démissionner et annoncé son projet ambitieux de « construction d'une nation ».	General Musharraf appeared on the national scene on October 12, 1999, when he ousted an elected government and announced an ambitious "nation-building" project.	El general Musharraf apareció en el escenario nacional el 12 de octubre de 1999, cuando derrocó a un gobierno electo y anunció un ambicioso proyecto de "construcción nacional".
Bon nombre de Pakistanais – qui avaient perdu toute illusion sur la classe politique du pays – sont restés silencieux, pensant qu'il tiendrait ses promesses.	Many Pakistanis, disillusioned with Pakistan's political class, remained mute, thinking that he might deliver.	Muchos pakistaníes, decepcionados de la clase política de su país, permanecieron callados creyendo que podría cumplir sus promesas.
Le 11 septembre 2001, les attaques terroristes sur l'Amérique ont placé Moucharraf sur le devant de la scène internationale, alors qu'il décidait de soutenir les Etats-Unis dans la guerre contre la terreur, plutôt que les Talibans.	The September 11, 2001, terrorist attacks on America brought Musharraf into the international limelight as he agreed to ditch the Taliban and support the United States-led war on terror.	Los ataques terroristas del 11 de septiembre de 2001 contra Estados Unidos pusieron a Musharraf en los reflectores internacionales, ya que acordó abandonar a los talibanes y apoyar la guerra contra el terrorismo encabezada por EU.

Moucharraf a pris des mesures contre des militants religieux au Pakistan et contre ceux qui luttent contre les forces indiennes au Cachemire.

Musharraf clamped down on some religious militants operating inside Pakistan and also on those fighting Indian forces in Kashmir.

Musharraf aplicó mano dura contra algunos militantes religiosos que operaban en Pakistán y también contra los que luchaban contra las fuerzas indias en Cachemira.

Le Pakistan a donc été récompensé par l'assistance et les armes des États-Unis.

As a result, Pakistan was rewarded with American financial assistance and arms.

Como premio, Pakistán recibió asistencia financiera y armas de Estados Unidos.

Pour mieux redistribuer ses cartes, Moucharraf a envoyé l'armée pakistanaise dans les zones ethniques qui longent l'Afghanistan, pour la première fois depuis l'indépendance du Pakistan.

In furtherance of his realignment, Musharraf sent the Pakistani army into the tribal areas bordering Afghanistan for the first time since Pakistan's independence.

Para confirmar su alineamiento, Musharraf envió al ejército pakistaní a las zonas tribales en la frontera con Afganistán por primera vez desde la independencia del país.

Les opérations contre les forces des Talibans et d'Al-Qaeda ont obtenu des résultats mitigés.

Operations there against Taliban and al-Qaeda forces brought mixed results.

Las operaciones que se llevaron a cabo ahí contra los talibanes y las fuerzas de al-Qaeda tuvieron resultados variados.

Si les Etats-Unis voient Moucharraf comme un agent de changement, ce dernier n'est jamais parvenu à avoir une légitimité dans son propre pays, où ses politiques ont toujours été considérées comme un tissu de contradictions.

Although the US viewed Musharraf as an agent of change, he has never achieved domestic political legitimacy, and his policies were seen as rife with contradictions.

Aunque Estados Unidos creía que Musharraf era un agente del cambio, él nunca ha logrado la legitimidad política interna y se considera que sus políticas están llenas de contradicciones.

Par exemple, il a conclu des alliances avec des forces politiques islamistes (qui en 2004 ont voté pour des changements constitutionnels légitimant sa position et ses actions) ; parallèlement, il remplaçait les dirigeants des partis politiques traditionnels modérés, tout en affirmant défendre la « modération éclairée ».

For example, he made alliances with Islamist political forces (who in 2004 voted for constitutional changes legitimizing his position and actions). At the same time, he sidelined moderate, mainstream political leaders while claiming that he stood for "enlightened moderation."

Por ejemplo, se alió con fuerzas políticas islamistas (que en 2004 votaron a favor de los cambios constitucionales que legitimaban la posición y las acciones del presidente). Al mismo tiempo, hizo a un lado a los principales líderes políticos moderados con el argumento de que él representaba la "moderación ilustrada".

Une série d'opérations militaires mal planifiées dans les zones ethniques ont davantage compliqué la situation dans une région frontalière instable.

A series of ill-planned military operations in the tribal areas further complicated the situation in the volatile border region.

Una serie de operaciones militares mal planeadas en las zonas tribales complicó aún más la situación en la volátil región fronteriza.

En mars dernier, Moucharraf a fait l'avancée la plus éhontée en destituant Iftikhar Chaudhry, président de la Cour suprême.

Last March, Musharraf took his boldest step, removing the Chief Justice of the Supreme Court, Iftikhar Chaudhry.

En marzo, Musharraf tomó su medida más audaz al despedir al presidente de la Suprema Corte, Iftikhar Chaudhry.

A la surprise générale, la communauté judiciaire du pays a organisé un mouvement à l'échelle nationale pour que le président réintègre son poste.

To the surprise of many, the country's legal community organized a nation-wide movement to restore the Chief Justice to his post.

Para sorpresa de muchos, la comunidad jurídica del país organizó un movimiento nacional para restablecer en su puesto al presidente de la Suprema Corte.

Des centaines de milliers de gens ordinaires ont exigé l'État de droit et la suprématie de la constitution, enhardissant le corps judiciaire et changeant la dynamique politique du pays.

Dans un arrêt historique que Moucharraf n'a pas eu d'autre choix que d'accepter, la Cour suprême a elle-même réintégré son président en juillet.

Ragaillardis, les magistrats ont donc continué à donner tort aux décisions du gouvernement et à embarrasser celui-ci – en particulier ses services des renseignements.

Les fonctionnaires ont été tenus pour responsables d'actions hors-la-loi, allant du passage à tabac de journalistes, à des détentions illégales pour des raisons de « sécurité nationale ».

Moucharraf – et ses alliés politiques – a tenté de s'adapter à cette nouvelle réalité.

Hundreds of thousands of ordinary people demanded the rule of law and the supremacy of the constitution, emboldening the judiciary and changing the country's political dynamic.

In a historic ruling that Musharraf had little choice but to accept, the Supreme Court itself reinstated the Chief Justice in July.

Subsequently, the energized judiciary continued ruling against government decisions, embarrassing the government – especially its intelligence agencies.

Government officials were held accountable for actions that were usually beyond the reach of the law, ranging from brutal beatings of journalists, to illegal confinement for "national security."

Musharraf and his political allies tried to adjust to this new reality.

Cientos de miles de ciudadanos comunes exigieron el Estado de derecho y la supremacía de la constitución, lo que dio valor al poder judicial y cambió la dinámica política del país.

En una decisión histórica, que a Musharraf no le quedó más remedio que aceptar, en julio la propia Suprema Corte restableció a su presidente en su cargo.

Posteriormente, el poder judicial revitalizado siguió emitiendo fallos en contra de decisiones del gobierno, humillándolo – sobre todo a sus organismos de inteligencia.

Se hizo que los funcionarios de gobierno rindieran cuentas por actos que generalmente estaban fuera del alcance de la ley, desde golpizas salvajes a periodistas hasta arrestos ilegales por "seguridad nacional".

Musharraf y sus aliados políticos intentaron ajustarse a esta nueva realidad.

I.2 Corpus *DW2*

Chers collègues - Messieurs et vous confrères sorciers... !	Colleagues - gentlemen and fellow wizards... !
On s'en jette un p'tit à la bonne vôtre !	Here's looking at your bottom !
Cul sec !	Up your eye !
Tire sur l'autre, il y a des petits trucs tordus dessus !	Huh, pull the other one, it's got strange knobbly bits on !
Bonne fête du Père Porcher !	Happy Hog's watch day !
Merci à tous !	Thank you, huh !
Chers collègues - nous voici réunis aujourd'hui pour la cérémonie de "départ définitif" de notre très cher futur disparu, le Sorcier Windle Poons !	Colleagues - we are gathered here today, for the "final departure" party of our dear soon-to-be-departed comrade, the Wizard Windle Poons !
Sacré vieux Windle !	Good old Windle !
Envoie-nous tes Mémoires d'Outre-Tombe !	Don't forget to ghost-write ! *wheeze*
TROIS !	THREE !
DEUX !	TWO !
UN !	ONE !
ZEEEE-ROOOOO !	ZEEEE-ROOOOO !
Très bien, tout le monde. Cérémonie funéraire à deux heures et demie, puis boissons et sandwiches au jambon à trois heures dans le hall principal. Hè, là !	Right everybody - errr... funeral at two thirty, then drinks and ham rolls in the main hall at three.
Qu'est-ce qui se passe ?	Here - wha...what's happening ?
Vous appelez ça du service, vous ?	Call this service, do you ?
Je suis mort, moi !	I'm dead, I am !
J'exige d'être emporté vers une vie meilleure, comme il est stipulé dans le contrat !	I demand to be taken away to a better life, as per contract !
De mon temps, ça ne se passait pas comme ça !	Things were different in my day !

Au moins, on mourrait correctement, et non pas comme ce qu'on vous propose aujourd'hui !	You died properly, not like the deaths you get nowadays !
Il - il dit qu'il n'est pas mort !	Err he, he umm - he, he says he's not dead !
Je "suis" mort..	I *am* dead.
Mais je tiens fichtrement bien sur mes pattes !	But I'm still bloody ambulatory !
Mais non, voyons.	No, you're not.
Tu ne trompes personne d'autre que toi-même, tu sais.	You're fooling no one but yourself, you know.
Mmmmmm...	Mmmmmm
Ma foi, il a l'air mort !	Well, ee, uh - he looks dead !
En tous cas, il le sent.	Smells dead.
Cela dit, il a toujours senti comme ça.	Course, he always did though.
Et bien entendu, j'imagine que je n'ai pas mon mot à dire dans tout ça, hein ?	And I suppose my word doesn't count for anything around here ?
Je ne peux pas être mort si je parle ? Qu'est-ce que vous dites de ça ?	I can't be dead if I'm still talking, now can I ?
Ecoute, mon vieux - c'est tout bien considéré que nous, professionels, te déclarons sorcier disparu.	Look old chap - it's our considered professional view that you are an extinct wizard.
Ton opinion ne compte pas parce que tu es mort.	Your opinion doesn't count. Because you're dead.
Oh.	Oh, yeah.
Très juste.	Good point.
Bon, j'imagine que je n'ai plus qu'à rester assis là, n'est-ce pas ?	Well I suppose I'll just sit here then, shall I ?
Je suppose que ça va prendre un petit moment...	I suppose it takes a while...
Alors - heu - c'est comment la mort, au fait ?	So - umm - how is death actually.
Aucune lumière blanche éblouissante ?	See any, um, white lights ?

Pas de tunnel ?
De jolies dames qui jouent de la harpe ?
Oh oui, s'il vous plaît, j'en prendrai deux.
Non !
Mmmmm - si c'est ça, le paradis, je regrette tous les trucs affreux que j'ai pas fait quand j'étais vivant.
Qu'est-ce qui m'arrive ?
Manifestement, ton corps est bien mort mais ton âme n'a pas - hum - déménagé.

Mais enfin, je ne vais pas traîner comme ça ici le reste de mon après-vie !
J'ai eu une vie difficile, Archichancelier !
J'ai droit à mon petit coin de paradis.
Je sais comment c'est, j'ai lu les prospectus.
Des jolies filles, du vin, du trucmuche...
Ecoute - ta vie est terminée !
Inutile de te lamenter là-dessus.
Et inutile, surtout, de rêver à quelque trucmuche que ce soit !
Qui est responsable de tout ça ?
Mais enfin, où est la MORT ?
C'est scandaleux !
Vous ne pouvez pas laisser une âme traîner comme ça dans un corps mort !
Et pourquoi pas ?
C'est contraire à l'hygiène !
Regardez - il y a de la nourriture, ici !
On peut pas le laisser près des amuse-gueules !

You know, Tunnels ?
Girls with harps ?
Ooh, yes please, I'll take two.
No !
Mmmmm - if this is heaven I wish I'd done wicked things when I was alive.

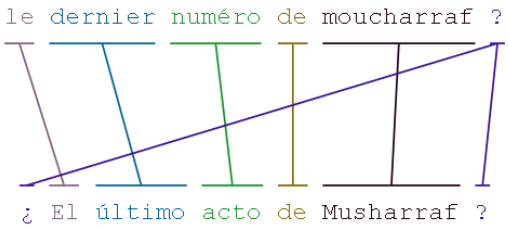
What's happening to me ?
Errr - well it... it seems that your body's dead, but your soul's still in - euh - well, in residence.
Well I'm not hanging about here for the rest of my afterlife !
I've had a hard life, Archchancellor !
I'm entitled to a bit of paradise.
I've read about it.
Young women and wine and whatnot...
Look - your life's over !
You're not supposed to moan about it.
And definitely not supposed to contemplate any... any whatnot !
Who's responsible for this ?
Where's DEATH then ?
This is outrageous !
You, you can't have a soul hanging about a deceased body like that !
Why not ?
It's unhygienic !
Ere - there's... there's food laid out.
We can't have him near the nibbles !

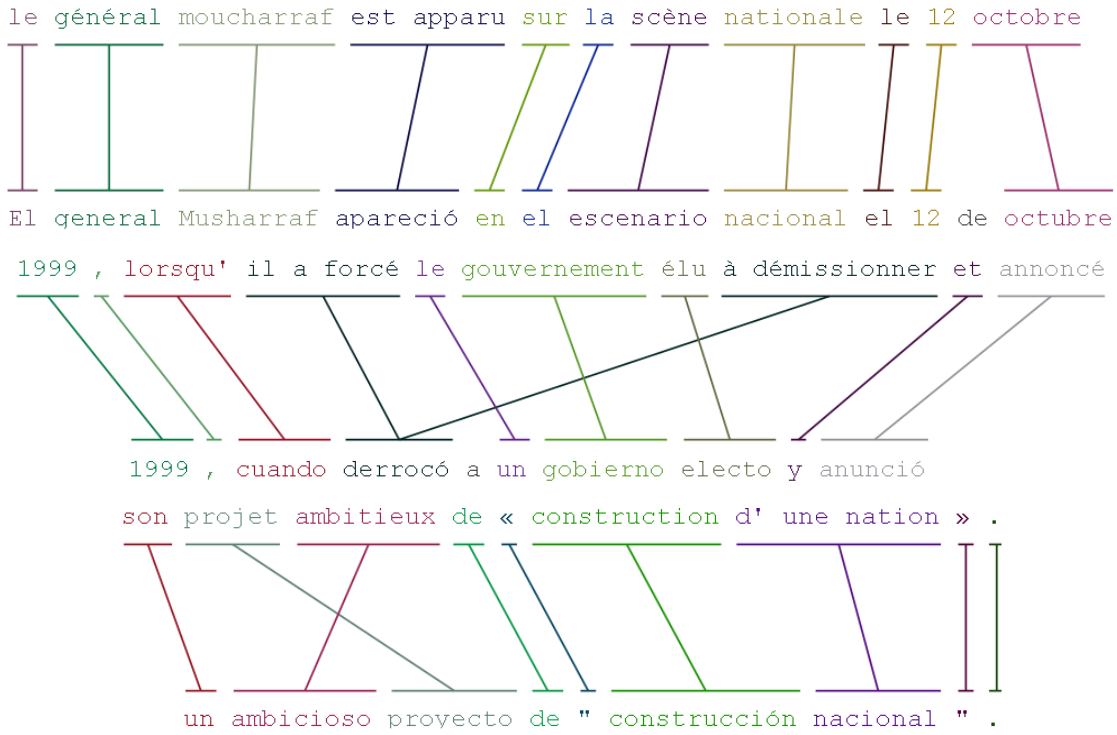
L'inspection de l'hygiène va nous tomber dessus !	The health inspectors will be onto us !
Tout à fait exact.	Good point.
Fais bonne figure, Windle.	Now compose yourself, Windle.
Tu ne peux pas te décomposer ici.	You can't decompose here.
Je te demande maintenant de circuler.	I shall have to ask you to move along.
Où va le monde si on ne peut même plus tomber raide mort en paix.	It comes to something when a man can't even drop stone dead in peace.
Le repos éternel ?	Eternal rest ?
Vous dîtes bien le repos éternel ?	Eternal rest, is it ?
Et bien, je ne vais pas accepter ça les bras croisés !	Well, I'm not going to take this lying down !
Je vais de ce pas me trouver une jolie petite tombe pas trop profonde...	I'm off to find myself a nice shallow grave...
Ce genre d'incident s'est produit bien trop souvent ces derniers temps.	There's been too much of this sort of thing lately.
Rincevent !	Rincewind !
Ah, Rincevent, te voilà enfin !	Ah, Rincewind. There you are !
Bon, comme tu le sais, de drôles de choses se sont passées dans notre cité, dernièrement.	Now, as you're aware, there have been some very odd goings on in this city of late.
Je fais bien sûr allusion à la soudaine disparition de la MORT.	I am referring, of course, to the sudden disappearance of DEATH.
Personne ne meurt ?	No one's dying ?
Oh si, ils meurent.	Oh, they're dying.
Mais leur âme ne part pas avec eux.	But their souls aren't being taken away.
Ils sont à la fois morts et vivants. Et voilà que ça vient d'arriver à ce pauvre Windle.	They're dead and alive at the same time, and now it's happened to poor Windle.
La Mort a disparu, et nous devons absolument la faire revenir - alors à toi de jouer !	Death's gone, and we need to summon him back - so, err, here you go !

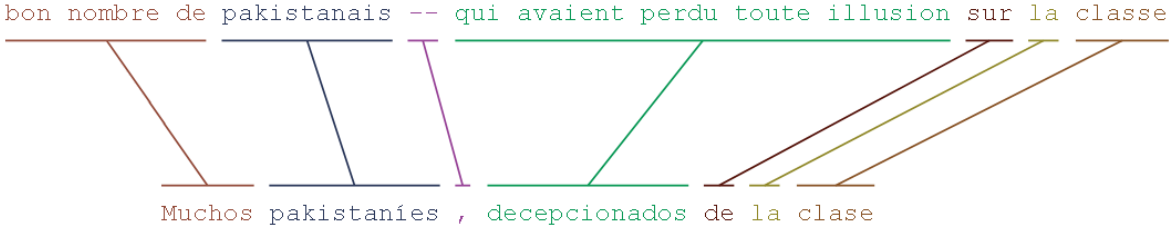
Annexe II

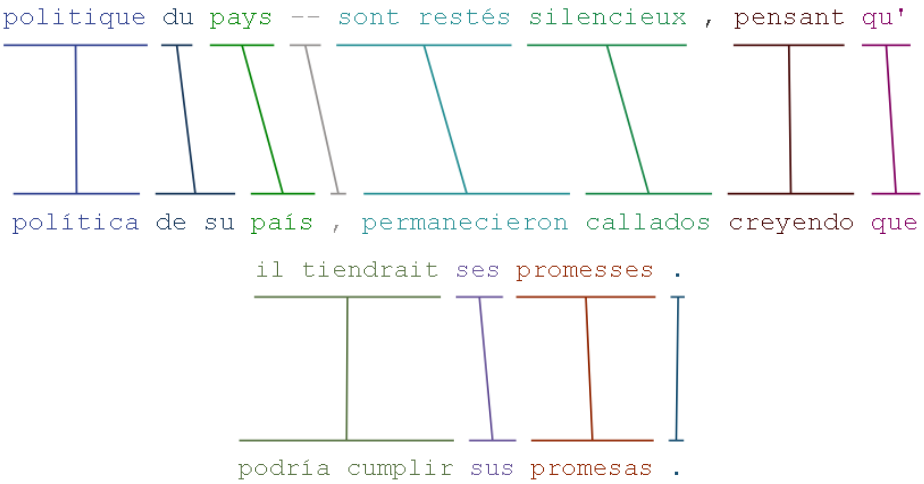
Alignements détaillés

II.1 Corpus *news commentary*

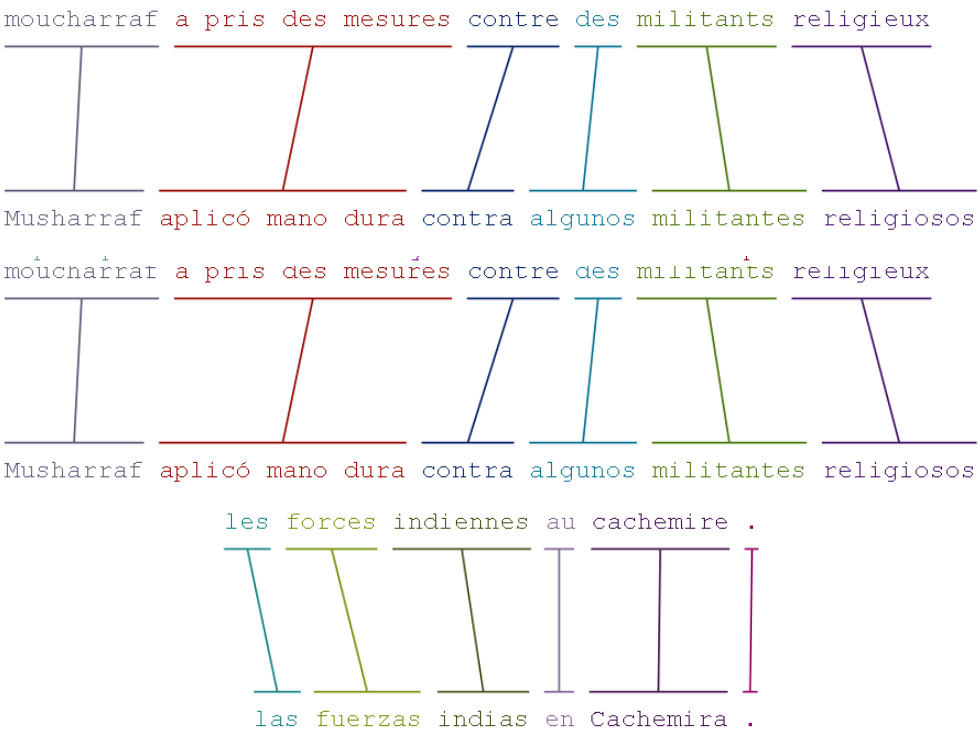


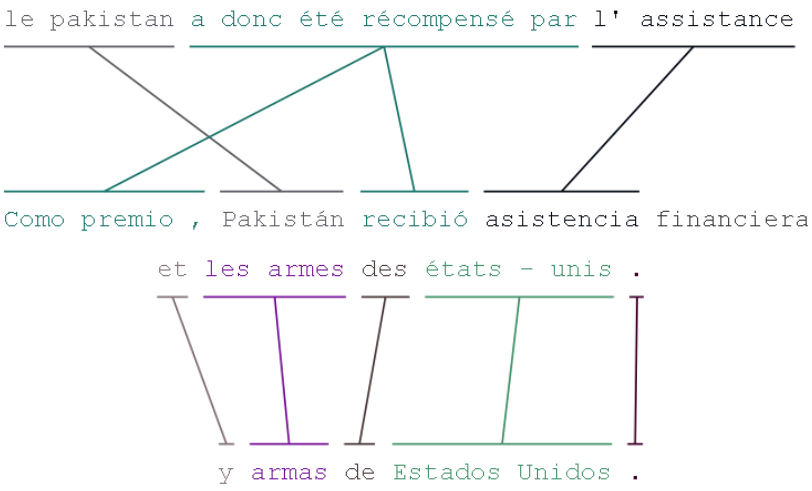


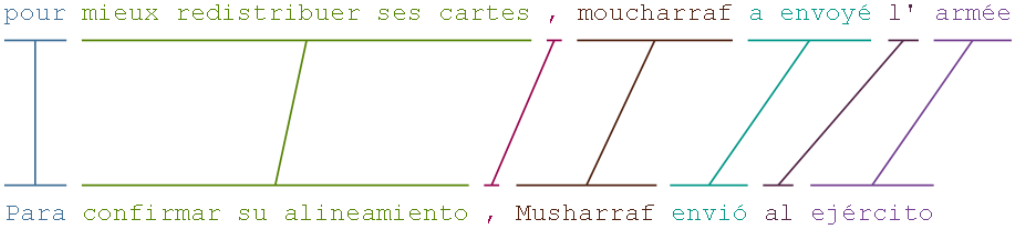



politique du pays -- sont restés silencieux , pensant qu'
política de su país , permanecieron callados creyendo que
il tiendrait ses promesses .
podría cumplir sus promesas .

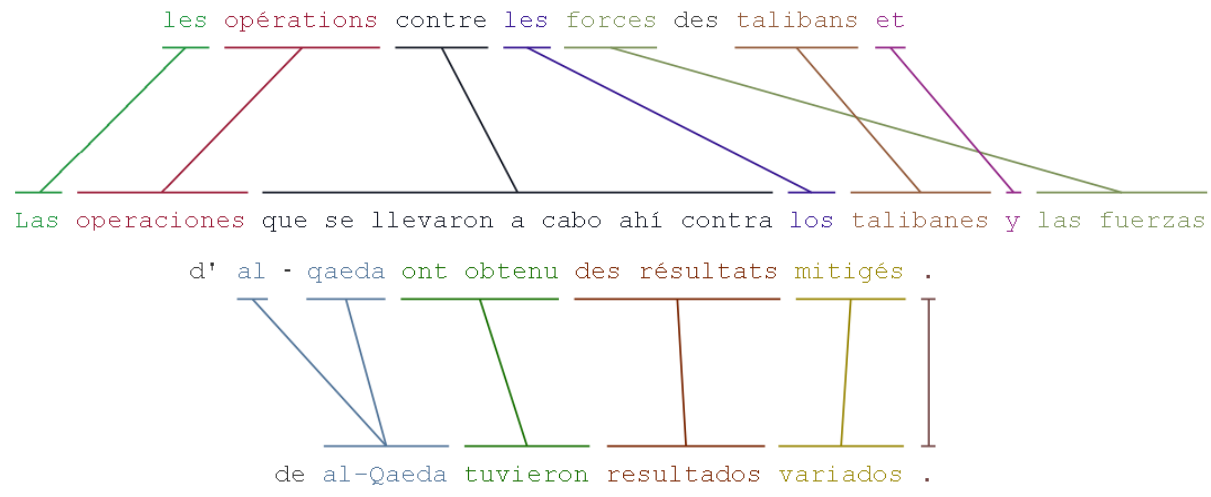

le 11 septembre 2001 , les attaques terroristes sur l' amérique
Los ataques terroristas del 11 de septiembre de 2001 contra Estados Unidos
ont placé moucharraf sur le devant de la scène
pusieron a Musharraf en los reflectores
internationale , alors qu' il décidait de
internacionales , ya que acordó
soutenir les etats - unis dans la guerre contre la terreur , plutôt que les talibans
abandonar a los talibanes y apoyar la guerra contra el terrorismo encabezada por EU

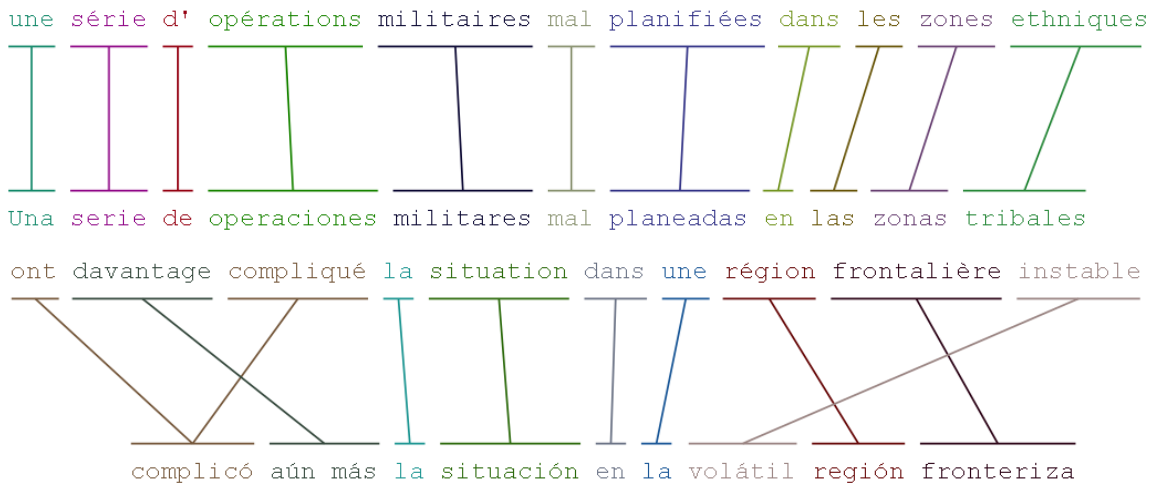

moucharraf a pris des mesures contre des militants religieux
Musharraf aplicó mano dura contra algunos militantes religiosos
moucharraf a pris des mesures contre des militants religieux
Musharraf aplicó mano dura contra algunos militantes religiosos
les forces indiennes au cachemire .
las fuerzas indias en Cachemira .


le pakistan a donc été récompensé par l' assistance
Como premio , Pakistán recibió asistencia financiera
et les armes des états - unis .
y armas de Estados Unidos .

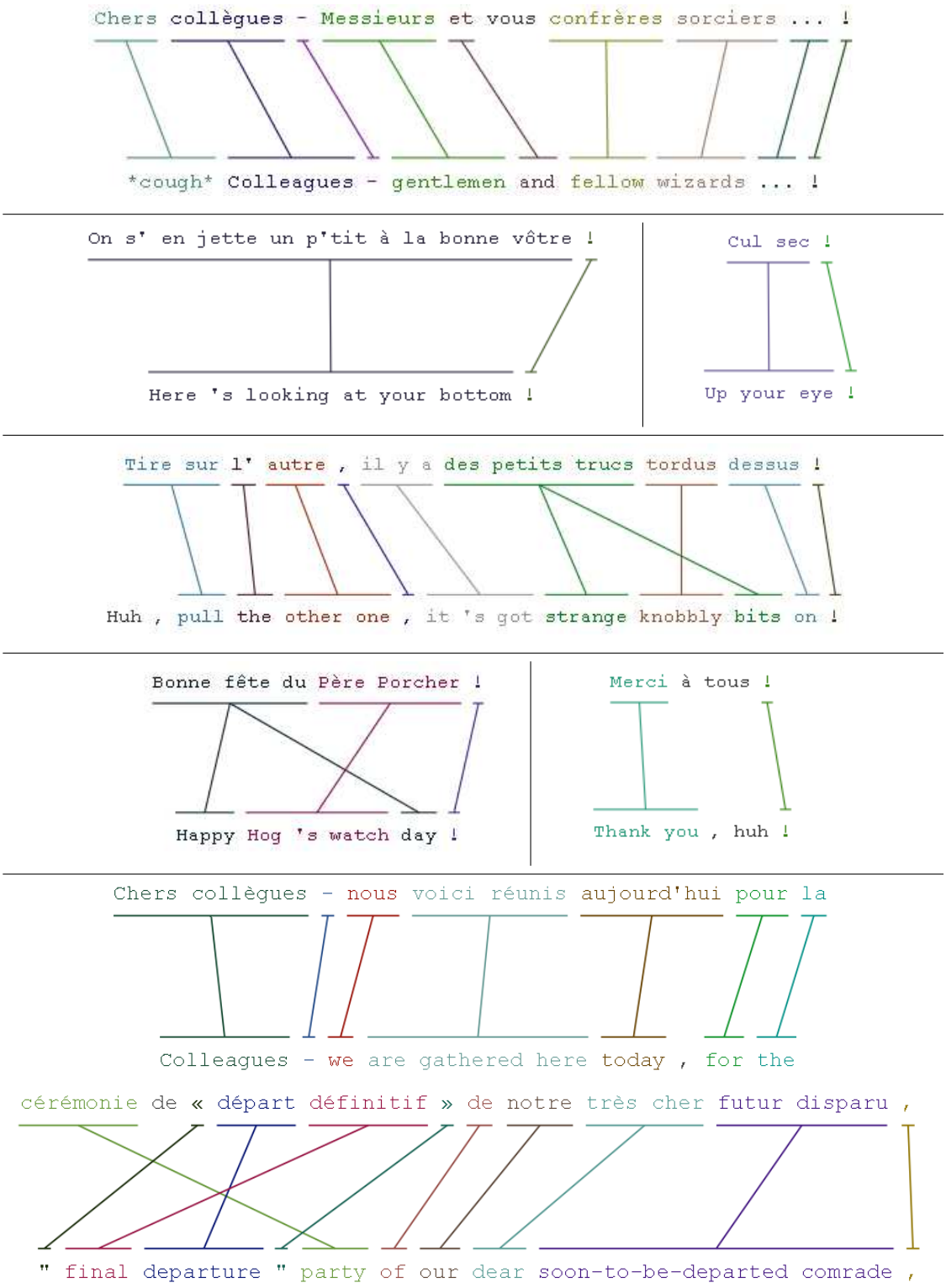

pour mieux redistribuer ses cartes , moucharraf a envoyé l' armée
Para confirmar su alineamiento , Musharraf envió al ejército

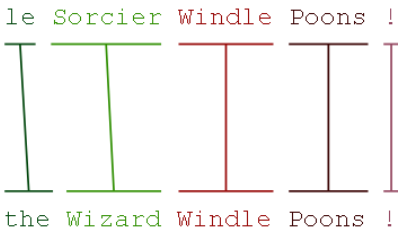

pakistanaise dans les zones ethniques qui longent l' afghanistan ,
pakistaní a las zonas tribales en la frontera con Afganistán
pour la première fois depuis l' indépendance du pakistan
por primera vez desde la independencia del país

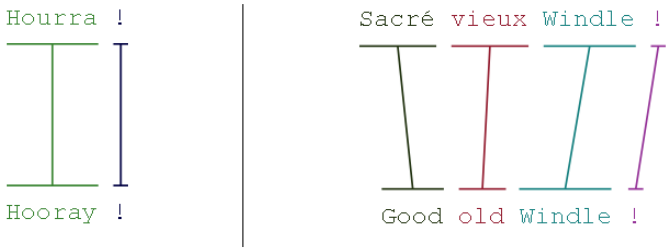

les opérations contre les forces des talibans et
Las operaciones que se llevaron a cabo ahí contra los talibanes y las fuerzas
d' al - qaeda ont obtenu des résultats mitigés .
de al-Qaeda tuvieron resultados variados .

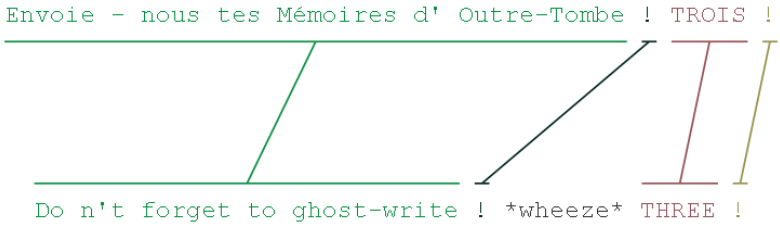

une série d' opérations militaires mal planifiées dans les zones ethniques
Una serie de operaciones militares mal planeadas en las zonas tribales
ont davantage compliqué la situation dans une région frontalière instable
complicó aún más la situación en la volátil región fronteriza

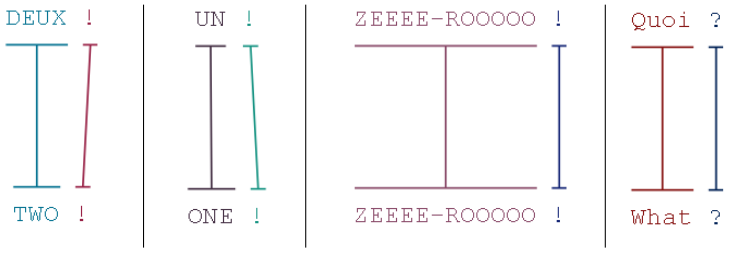
II.2 Corpus *DW2*

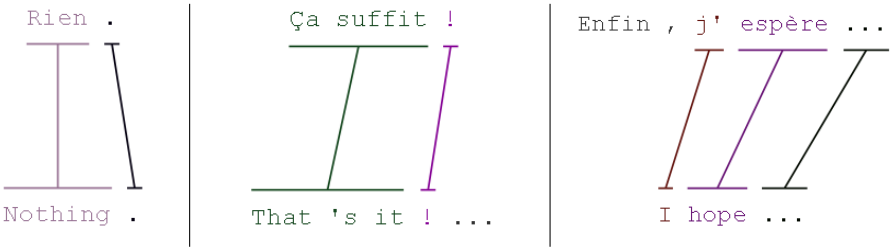


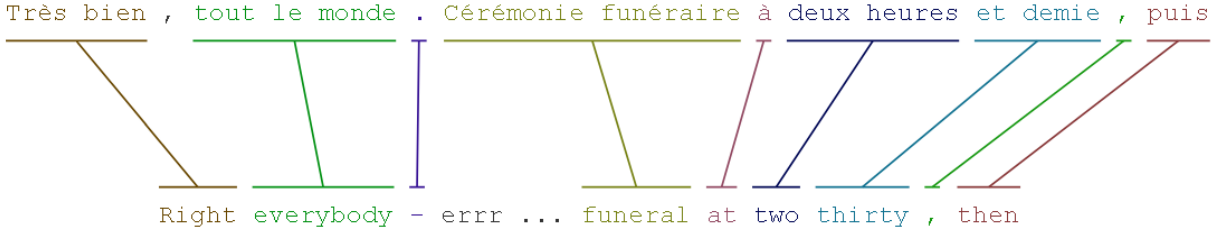

le Sorcier Windle Poons !
the Wizard Windle Poons !

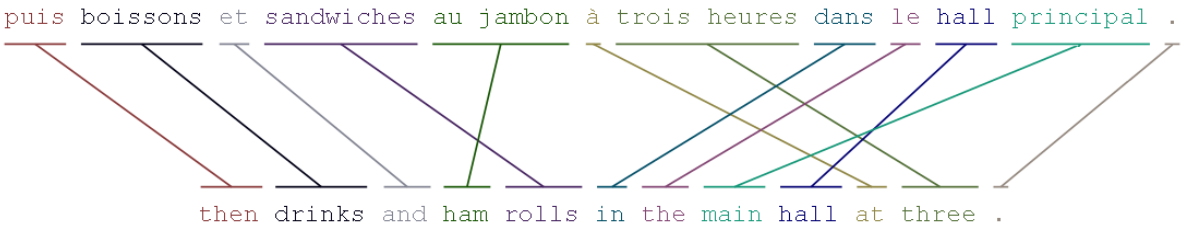

Hourra !
Hooray !
Sacré vieux Windle !
Good old Windle !

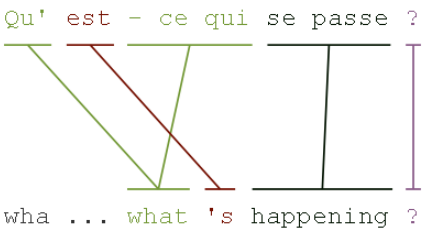

Envoie - nous tes Mémoires d' Outre-Tombe ! TROIS !
Do n't forget to ghost-write ! *wheeze* THREE !

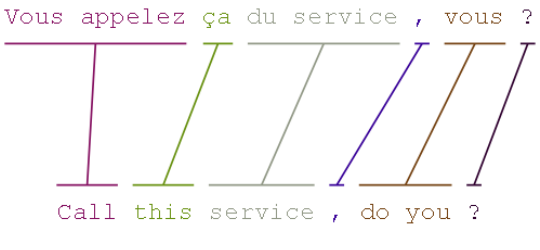

DEUX !
TWO !
UN !
ONE !
ZEEEE-ROOOOO !
ZEEEE-ROOOOO !
Quoi ?
What ?

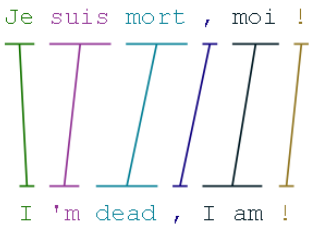

Rien .
Nothing .
Ça suffit !
That 's it ! ...
Enfin , j' espère ...
I hope ...

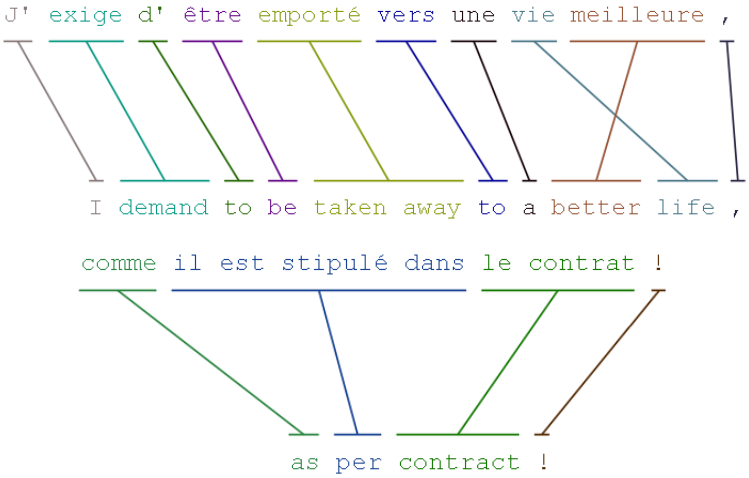

Très bien , tout le monde . Cérémonie funéraire à deux heures et demie , puis
Right everybody - errr ... funeral at two thirty , then

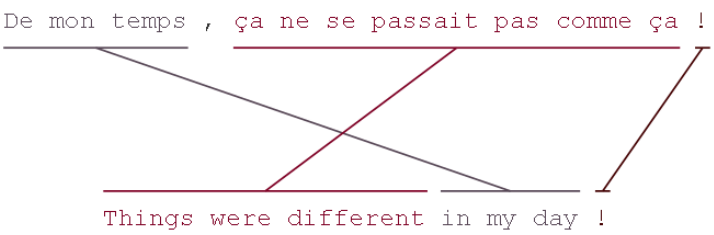

puis boissons et sandwiches au jambon à trois heures dans le hall principal .
then drinks and ham rolls in the main hall at three .


Qu' est - ce qui se passe ?
wha ... what 's happening ?

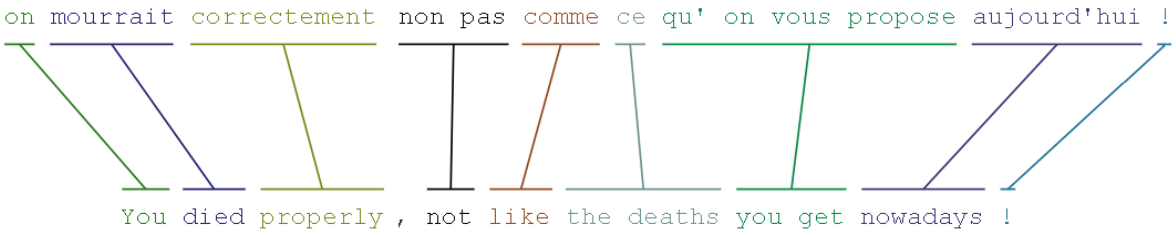

Vous appelez ça du service , vous ?
Call this service , do you ?

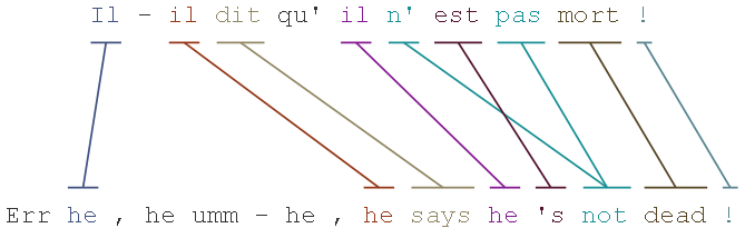

Je suis mort , moi !
I 'm dead , I am !

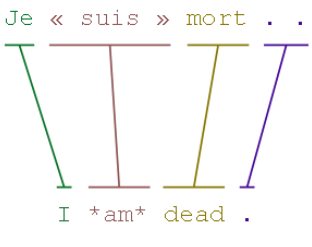

J' exige d' être emporté vers une vie meilleure ,
I demand to be taken away to a better life ,
comme il est stipulé dans le contrat !
as per contract !

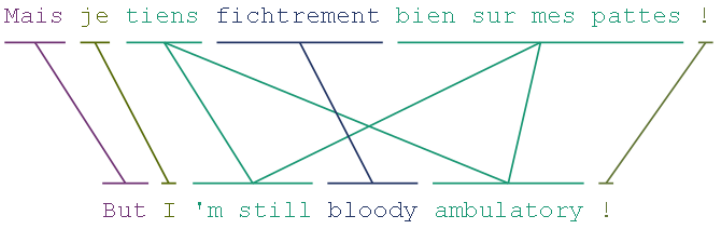

De mon temps , ça ne se passait pas comme ça !
Things were different in my day !

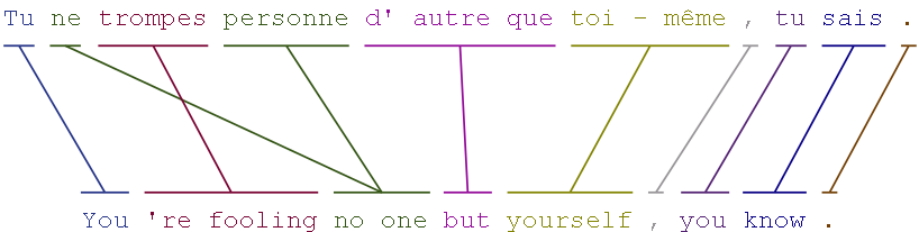
xxx

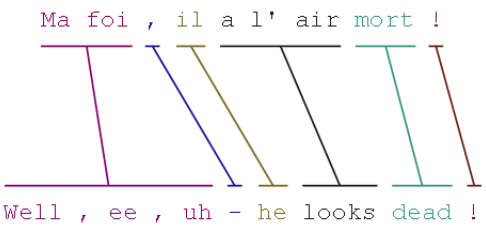

on mourrait correctement non pas comme ce qu' on vous propose aujourd'hui !
You died properly , not like the deaths you get nowadays !

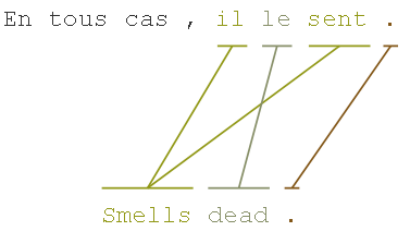

Il - il dit qu' il n' est pas mort !
Err he , he umm - he , he says he 's not dead !


Je « suis » mort . .
I *am* dead .


Mais je tiens fichtrement bien sur mes pattes !
But I 'm still bloody ambulatory !


Tu ne trompes personne d' autre que toi - même , tu sais .
You 're fooling no one but yourself , you know .


Ma foi , il a l' air mort !
Well , ee , uh - he looks dead !


En tous cas , il le sent .
Smells dead .

RÉSUMÉ

Cette thèse s'inscrit dans le cadre du traitement automatique du langage naturel, et traite plus précisément de l'alignement sous-phrastique bilingue classiquement lié à la traduction automatique statistique. Les travaux exposés s'en distinguent en proposant un fonctionnement évolutif à base d'exemples initialisé par des annotateurs non-experts via une interface adaptée. L'approche est principalement motivée par la recherche d'une expressivité comparable à celle observée dans les alignements manuels. Une partie importante de ce travail consiste à définir un cadre formel sous-tendant une architecture originale à base d'exemples alignés. Plusieurs mémoires d'alignements ont été constituées en tirant parti d'informations provenant d'analyseurs syntaxiques automatiques, en plaçant les prérequis technologiques à un niveau raisonnablement peu élevé. Deux nouvelles méthodes d'alignement sont comparées à des références connues via des mesures d'accord classiques, et trois distances transformationnelles sont introduites.

Mots-clés : *Alignement sous-phrastique, corpus de référence , divergence, expressivité, annotation, interface homme-machine*

ABSTRACT

This research belongs to the Natural Language Processing (NLP) field and more specifically focuses on the topic of Sub-sentential Alignments which is closely related to Machine Translation. The originality of this work consists in an example-based approach bootstrapped by the participation of non-expert annotators through an appropriate interface. The quest for a greater expressivity, such as observed in manual alignments, mainly motivates the whole approach. An important effort has been made to define a formal environment for this original architecture based on aligned examples. Several memories have been created, using syntactic informations from parsers outputs with reasonnable low-tech requirements. Two new alignment methods were compared with state-of-the-art measures and three transformational metrics were introduced.

Keywords : *Sub-sentential Alignement, golden corpus, divergence, expressivity, annotation, human-machine interface*