# Modèles de traduction évolutifs

Frédéric Blain

Université du Maine

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

**Université du Maine**

# MODÈLES DE TRADUCTION ÉVOLUTIFS

# THÈSE

présentée et soutenue publiquement le 23 septembre 2013

pour l'obtention du

**Doctorat de l'Université du Maine**

**(spécialité informatique)**

par

FRÉDÉRIC BLAIN

**Composition du jury**

| | | | |
|---|---|---|---|
| *Rapporteurs :* | M. Marc Dymetman | Docteur, HDR | Xerox Research Centre Europe |
| | M. Laurent Besacier | Professeur des Universités | LIG, Université J. Fourier |
| *Examinateurs :* | M. Patrik Lambert | Docteur | Barcelona Media |
| | M. Yannick Estève | Professeur des Universités | LIUM, Université du Maine |
| *Directeur de thèse :* | M. Holger Schwenk | Professeur des Universités | LIUM, Université du Maine |
| *Co-encadrant de thèse :* | M. Jean Senellart | Docteur | SYSTRAN S.A. |

**Laboratoire d'Informatique de l'Université du Maine**

lium

# Résumé

Bien que la recherche ait fait progresser la traduction automatique depuis plusieurs années, la sortie d'un système automatisé ne peut être généralement publiée sans avoir été révisée humainement au préalable, et corrigée le cas échéant. Forts de ce constat, nous avons voulu exploiter ces retours utilisateurs issus du processus de révision pour adapter notre système statistique dans le temps, au moyen d'une approche incrémentale.

Dans le cadre de cette thèse Cifre-Défense, nous nous sommes donc intéressés à la post-édition, un des champs de recherche les plus actifs du moment, et qui plus est très utilisé dans l'industrie de la traduction et de la localisation.

L'intégration de retours utilisateurs n'est toutefois pas une tâche aussi évidente qu'il n'y paraît. D'une part, il faut être capable d'identifier l'information qui sera utile au système, parmi l'ensemble des modifications apportées par l'utilisateur. Pour répondre à cette problématique, nous avons introduit une nouvelle notion (les « Actions de Post-Édition »), et proposé une méthodologie d'analyse permettant l'identification automatique de cette information à partir de données post-éditées. D'autre part, concernant l'intégration continue des retours utilisateurs nous avons développé un algorithme d'adaptation incrémentale pour un système de traduction statistique, lequel obtient des performances supérieures à la procédure standard. Ceci est d'autant plus intéressant que le développement et l'optimisation d'un tel système de traduction est une tâche très coûteuse en ressources computationnelles, nécessitant parfois jusqu'à plusieurs jours de calcul.

Conduits conjointement au sein de l'entreprise SYSTRAN et du LIUM, les travaux de recherche de cette thèse s'inscrivent dans le cadre du projet ANR COSMAT [1]. En partenariat avec l'INRIA, ce projet avait pour objectif de fournir à la communauté scientifique un service collaboratif de traduction automatique de contenus scientifiques. Outre les problématiques liées à ce type de contenu (adaptation au domaine, reconnaissance d'entités scientifiques, etc.), c'est l'aspect collaboratif de ce service avec la possibilité donnée aux utilisateurs de réviser les traductions qui donne un cadre applicatif à nos travaux de recherche.

**Mots-clés:** Traduction automatique statistique, Post-édition, Retours utilisateurs, Actions de post-édition, Adaptation incrémentale

1. www.cosmat.fr

# Abstract

Although machine translation research achieved big progress for several years, the output of an automated system cannot be published without prior revision by human annotators. Based on this fact, we wanted to exploit the user feedbacks from the review process in order to incrementally adapt our statistical system over time.

As part of this thesis, we are therefore interested in the post-editing, one of the most active fields of research, and what is more widely used in the translation and localization industry.

However, the integration of user feedbacks is not an obvious task. On the one hand, we must be able to identify the information that will be useful for the system, among all changes made by the user. To address this problem, we introduced a new concept (the "Post-Editing Actions"), and proposed an analysis methodology for automatic identification of this information from post-edited data. On the other hand, for the continuous integration of user feedbacks, we have developed an algorithm for incremental adaptation of a statistical machine translation system, which gets higher performance than the standard procedure. This is even more interesting as both development and optimization of this type of translation system has a very computational cost, sometimes requiring several days of computing.

Conducted jointly with SYSTRAN and LIUM, the research work of this thesis is part of the French Government Research Agency project COSMAT [2]. This project aimed to provide a collaborative machine translation service for scientific content to the scientific community. The collaborative aspect of this service with the possibility for users to review the translations gives an application framework for our research.

**Keywords:** Statistical machine translation, Post-editing, User feedbacks, Post-editing actions, Incremental adaptation

2. www.cosmat.fr

# Table des matières

**Résumé** ii

**Abstract** iii

**Table des figures** ix

**Liste des tableaux** xi

**Introduction** 1

## Partie I Paradigmes fondamentaux 7

**Chapitre 1**
**Paradigmes fondamentaux** 9

1.1 Bref résumé historique . . . . . 11
1.2 Architectures linguistiques et computationnelles . . . . . 13
1.3 Approche experte . . . . . 14
1.3.1 Traduction automatique directe . . . . . 15
1.3.2 Traduction automatique par transfert . . . . . 15
1.3.3 Traduction automatique interlangua . . . . . 15
1.4 Approche empirique . . . . . 16
1.4.1 Modélisation du langage . . . . . 19

1.4.1.1 Modèle de type N-Gramme . . . . . 20
1.4.1.2 Lissage . . . . . 21
1.4.1.3 Modélisation dans un espace continu (*CSLM*) . . . . . 22
1.4.1.4 Évaluation d'un modèle de langage . . . . . 22
1.4.2 Modélisation de la traduction . . . . . 23
1.4.2.1 Alignement sous-phrastique . . . . . 23
1.4.2.2 Modèles IBM . . . . . 24
1.4.2.3 Modélisation basée sur les séquences de mots . . . . . 26
1.4.2.4 Extraction des séquences de mots . . . . . 27
1.4.2.5 Pondération des séquences de mots . . . . . 27
1.4.2.6 Modélisation log-linéaire . . . . . 28
1.4.2.7 Optimisation d'un système de TAS . . . . . 29
1.5 Approche hybride . . . . . 32
1.6 Évaluation de la traduction automatique . . . . . 33
1.6.1 Évaluation manuelle des traductions . . . . . 33
1.6.2 Évaluation automatique des traductions . . . . . 34
1.7 Post-Édition . . . . . 39
1.7.1 Motivation et principe . . . . . 39
1.7.2 Évaluer l'effort de post-édition . . . . . 40
1.7.3 Comment limiter ou réduire cet effort ? . . . . . 41

**Partie II Cadre applicatif** **45**

**Chapitre 2**
**COSMAT : Traduction Automatique de contenus scientifiques** **47**

2.1 Enjeux et problématiques . . . . . 48
2.2 Approches scientifiques et techniques . . . . . 50
2.2.1 Extraction de contenu structuré . . . . . 50
2.3 Intégration de connaissances linguistiques . . . . . 52

2.4 Adaptation en domaine d'un système de TAS . . . . . 54
2.4.1 Extraction de données bilingues du domaine . . . . . 54
2.4.2 Données d'apprentissage hors-domaine . . . . . 55
2.4.3 Sélection de données hors-domaine . . . . . 56
2.5 Interface de post-édition . . . . . 59
2.5.1 Campagnes d'évaluation . . . . . 60
2.5.1.1 Pertinence utilisateur . . . . . 61
2.6 Intégration dans HAL . . . . . 63
2.6.1 Architecture globale du service COSMAT . . . . . 64

**Partie III Contributions 67**

**Chapitre 3**
**Analyse qualitative et automatique de données post-éditées 69**

3.1 Analyser la post-édition . . . . . 72
3.1.1 Les Actions de Post-Édition (APE) . . . . . 72
3.1.2 Typologie des actions de post-édition . . . . . 76
3.2 Automatisation du processus d'analyse . . . . . 78
3.2.1 Protocole d'analyse en APE . . . . . 79
3.2.2 Règles linguistiques . . . . . 79
3.2.3 Disponibilité de l'outil « SmartDiff » . . . . . 80
3.3 Données expérimentales . . . . . 81
3.3.1 Annotation manuelle de référence . . . . . 81
3.3.2 Résultats de l'analyse automatique . . . . . 84
3.4 Conclusion . . . . . 86

**Chapitre 4**
**Adaptation incrémentale d'un système automatique statistique** **89**

4.1 Travaux connexes dans la littérature . . . . . 91
4.2 Protocole d'adaptation incrémentale . . . . . 93
4.2.1 Combinaison d'alignements mot-à-mot . . . . . 94
4.3 Évaluations expérimentales . . . . . 98
4.3.1 Données d'apprentissage . . . . . 98
4.3.2 Apprentissage du système de référence . . . . . 99
4.3.3 Temps de calcul *vs.* Qualité de traduction . . . . . 100
4.3.3.1 Protocole expérimental . . . . . 101
4.3.4 Combinaison des modèles de traduction . . . . . 103
4.4 Conclusion . . . . . 110

**Conclusions et perspectives futures** **111**

**Acronymes** **115**

**Bibliographie** **117**

**Annexes**

**Annexe A**
**Liste des publications** **125**

# Table des figures


1.1 Triangle de Vauquois. . . . . 14
1.2 Schéma de la traduction automatique statistique. . . . . 18
1.3 Exemple d'alignement des mots dans une paire de phrases français-anglais. . . 23
1.4 Traduction automatique statistique basée sur les séquences de mots. . . . . 27
1.5 Exemples de paires de séquences consistantes et non consistantes. . . . . 28

2.1 Exemple de description documentaire au format TEI. . . . . 51
2.2 Exemples sur l'impact de la reconnaissance d'entités scientifiques développée par SYSTRAN. . . . . 53
2.3 Interface de post-édition pour COSMAT développée par la société SYSTRAN. . . 60
2.4 Interface COSMAT de visualisation des traductions. Ici utilisée lors de la conférence LREC en 2012. . . . . 61
2.5 Architecture globale du service collaboratif COSMAT intégré dans HAL. Le protocole de communication entre HAL, le serveur GROBID et le serveur de traduction est basé sur une interface « RESTFUL ». . . . . 64

3.1 Distance d'éditions classique dite « mécanique » entre une hypothèse de traduction et sa version post-éditée. . . . . 73
3.2 Distance d'éditions basée sur l'analyse en APE, dite « logique », entre une hypothèse de traduction et sa version post-éditée. . . . . 74
3.3 Exemple d'annotations linguistiques pour une paire de phrases. . . . . 75
3.4 Architecture de notre outil d'analyse en APE. . . . . 78
3.5 Exemple d'annotations en APE. L'APE est représentée dans le noeud <pea> simultanément dans l'hypothèse de traduction (<target>) et sa version post-éditée (<pstedt>). . . . . 82

4.1 Protocole d'alignement séquentiel qui s'opère en trois temps. . . . . 93
4.2 Exemple d'alignement source-vers-référence utilisant l'hypothèse de traduction comme « pivot ». Ici est matéralisé ce que l'on veut apprendre : que « lattices » se traduit par « treillis » au lieu de « aspect algorithmique ». Tandis qu'en rouge, est matérialisé non pas une erreur de traduction du système de TA, mais ce que l'on considère comme étant un changement stylique de la part du post-éditeur. . 95
4.3 Le corpus « absINFO » du projet COSMAT est découpé en trois sous-corpus pour permettre la simulation d'un procéssus de post-édition et d'une adaptation d'un système de TAS. . . . . 99

4.4 Scores BLEU obtenus respectivement sur les corpus de développement et de test pour nos quatre systèmes : « Gizapp », « inc-Gizapp », « OnlineAdapt » et « inc-OnlineAdapt ». . . . . 102
4.5 Résultats pour l'utilisation de modèles par repli. La courbe « $+$ » représente notre système de TAS à séquences de mots utilisant un seul modèle de traduction. La courbe « $\chi$ » représente notre système de TAS utilisant deux modèles de traduction avec le modèle en domaine comme principal modèle et le modèle générique du système référence en repli. La courbe « $\Theta$ » représente une configuration similaire à la précédente avec simplement une inversion dans l'ordre des deux modèles de traduction. . . . . 104
4.6 Résultats comparatifs entre l'utilisation de deux modèles de traduction sans ($\Delta$) et avec repli ($\Theta$). La courbe « $+$ » représente notre système de TAS utilisant un seul modèle de traduction. . . . . 105
4.7 Scores BLEU pour des modèles de traduction sans repli pour des éditions de type « substitution » uniquement. . . . . 106
4.8 Scores BLEU obtenus en exploitant les deux meilleures hypothèses de traduction générées par les systèmes de TAS. . . . . 108
4.9 Scores BLEU pour une adaptation incrémentale sans tuning. . . . . 109

# Liste des tableaux


2.1 Statistiques sur les données d'apprentissage, de développement et de test extraites des résumés de thèses disponibles sur HAL, respectivement pour les domaines Informatique et Physique. (M pour million et k pour millier) . . . . . 55
2.2 Données d'apprentissage hors-domaine en nombre de phrases et de mots (après tokenisation). (M pour million et k pour millier) . . . . . 56
2.3 Scores BLEU sur le corpus de test en-domaine obtenus respectivement par les systèmes de TAS adaptés en-domaine avec et sans sélection de données monolingues pour le modèle de langage. . . . . 57

3.1 Typologie proposée pour la classification des APE pour le français. . . . . 77
3.2 Résultats de l'analyse manuelle sur 100 phrases post-éditées. 90% des éditions concernent un GN pour les deux systèmes. Les changements terminologiques étant la principale source des APE avec 59% pour le système RBMT et 62% pour le système de TAS. . . . . 82
3.3 Top-4 des APE les plus fréquentes identifiées pour le système RBMT. . . . . 83
3.4 Top-4 des APE les plus fréquentes identifiées pour le système de TAS. . . . . 83
3.5 Résultats de l'analyse automatique des APE sur le même jeu de 100 phrases post-éditées. La colonne #APE indique le nombre de APE identifiées, la colonne #Match indique le nombre d'APE bien reconnues, et les deux dernières colonnes indiquent la Précision et le Rappel, pour chaque APE actuellement implémentée dans SmartDiff. . . . . 85
3.6 Couverture des APE et des propagation observées pour les systèmes RBMT et de TAS. La première colonne montre le nombre d'éditions tandis que la seconde indique le taux de couverture. . . . . 85

# Introduction

Le processus de traduction se définit comme étant le passage en langue dite « cible », de ce qui a été énoncé en langue « source » en tentant de conserver le sens ainsi que le style. Une personne ayant pour motivation la traduction d'un texte est donc assujettie à la bonne maîtrise des langues dans lesquelles elle souhaite évoluer, mais pas seulement. Outre que cette aisance due aux langues est primordiale, une bonne connaissance du domaine dans lequel s'inscrit le texte à traduire l'est tout autant. Il paraît en effet difficile de vouloir traduire un texte dont on ne serait pas en mesure de lever toutes les ambiguïtés interlinguales : ambiguïtés lexicales, syntaxiques (ordre des mots dans la phrase) ou encore sémantiques (sens des unités linguistiques). La traduction, bien que simple dans son principe, est un processus intellectuel complexe qui nécessite un temps conséquent pour être qualitativement réalisé.

La traduction des langues naturelles par la machine, d'abord rêvée au XVIIe siècle, est devenue une réalité à la fin du XXe siècle. La traduction automatique (dorénavant TA) ne constitue pas un espace de recherche intellectuelle abstrait, mais l'application des sciences informatiques et linguistiques pour le développement de systèmes répondant à des besoins pratiques. Par TA, on désigne le fait de traduire un texte d'une langue source vers une langue cible, sans aucune intervention humaine, et c'est précisément dans ce domaine que s'inscrivent les travaux de recherche présentés dans ce manuscrit.

Bien qu'étant un domaine de recherche des plus récents, la TA permet aujourd'hui de disposer de systèmes aux performances satisfaisantes pour une utilisation courante dans un monde où la communication inter cultures et inter générations ne cesse de croître, avec de multiples connexions vers d'autres domaines tels que la Reconnaissance Automatique de la Parole (RAP), la Reconnaissance Optique de Caractère (*Optical Character Recognition* – OCR, en anglais), etc. Ce niveau de performance permet aujourd'hui à la TA d'être utilisée tout le temps, partout, sur le web et sur de multiples supports tels que les smartphones ou les tablettes, facilitant ainsi l'accès à des ressources uniquement disponibles en langues étrangères, mais également d'élargir les horizons de la communication mondialisée.

La TA touche également des domaines d'application plus discrets, présentant un intérêt « Défense ». Gouvernements et armées, américains notamment, restent parmi les acteurs principaux du financement de projets de recherche. Leur intérêt tient au fait, qu'ainsi, ils comprennent mieux le monde autour d'eux et voient leurs échanges facilités. On peut à ce sujet imaginer aisément l'utilité que peut avoir la TA, couplée avec un système de RAP par exemple, lorsque ceux-ci assurent une mission de maintien de la Paix et d'aide aux populations de pays étrangers.

La TA intéresse également beaucoup l'industrie de la traduction comme outil d'aide à la productivité. Malgré une nette amélioration de la qualité de traduction des systèmes automatiques ces dernières années, la TA ne permet pas encore de générer des traductions qui soient publiables en l'état. Ceci est d'autant plus vrai que les systèmes accusent une certaine inégalité qualitative entre les différentes paires de langues. Ces inégalités peuvent être dûes à un problème de ressources disponibles, ou bien de modélisation linguistique des langues concernées. Il est donc d'usage de procéder à la révision des hypothèses de traduction en sortie de système par l'intermédiaire d'annotateurs humains. Ce processus de révision suscite ces dernières années un vif intérêt de la part de la communauté scientifique. Cette dernière cherche ainsi non seulement à évaluer la qualité en sortie de la TA (mesurant entre autre l'effort nécessaire à la révision des traductions), mais aussi à utiliser ces nouvelles ressources d'informations pour adapter ses systèmes. C'est sur ce dernier point que portent les travaux de recherche présentés dans ce manuscrit.

Dans cette thèse, nous avons cherché à exploiter les données post-éditées résultantes d'un processus de révision, de sorte qu'un système de TA puisse apprendre continuellement de ses erreurs. Pour ce faire, nous nous sommes intéressés à modéliser l'intention du post-éditeur dans sa tâche par une analyse qualitative de la correction : concerne-t-elle la terminologie ? une correction grammaticale ? une réorganisation de la phrase ? Pour cette analyse, nous avons introduit une notion nouvelle d'« Actions de post-édition ». Nous sommes ainsi capables d'extraire la quintessence des informations que renferment les données post-éditées. Pour un système statistique par exemple (c'est le cas ici), l'adaptation peut être effectuée de façon continue par le biais d'une nouvelle technique permettant la mise à jour de ses paramètres, et ce sans avoir à effectuer un nouveau cycle d'apprentissage complet qui est une étape très chronophage.

## Cadre applicatif

Le travail de thèse présenté dans ce manuscrit s'incrit dans le cadre du projet ANR COSMAT [3]. Le principal objectif de ce projet est de proposer à la communauté scientifique un service collaboratif de TA de contenus scientifiques. Bien que soient abordées des problématiques liées au fait même de vouloir traduire des contenus scientifiques (tableau, références, etc.), c'est l'aspect collaboratif qui nous concerne ici plus particulièrement : donner la possibilité aux utilisateurs de réviser les traductions issues de ce service, c'est vouloir qu'à terme, les systèmes de TA qui le composent puissent être adaptés à partir des retours de ces utilisateurs. Ceci et d'autant plus justifié que ces scientifiques sont de fait considérés comme étant experts

3. www.cosmat.fr

des domaines dans lesquels ils publient. Ils ont donc la légitimité nécessaire pour réviser les traductions.

## Thèse Cifre-Défense

Cette thèse s'inscrit dans le cadre d'une collaboration entre la société SYSTRAN et le Laboratoire d'Informatique de l'Université du Maine (LIUM) au travers d'une convention CIFRE (Conventions Industrielles de Formation par la REcherche). Co-financés par la Délégation Générale pour l'Armement (DGA), ces travaux de recherche furent encadrés par le Pr Holger Schwenk, responsable du groupe traduction de l'équipe *Language and Speech Technologies* (LST) du LIUM, et par le Dr Jean Senellart, Directeur scientifique de SYSTRAN.

L'entreprise SYSTRAN, dont les bureaux sont à Paris, est le leader mondial des technologies de TA. Les logiciels SYSTRAN, historiquement fondés sur une approche à base de règles, permettent aujourd'hui de traduire instantanément en 52 paires de langues pour tous types de contenus. En 2009, la société SYSTRAN a mis sur le marché le premier moteur de traduction hybride, résultant de l'association de sa technologie à base de règles linguistiques et d'un post-traitement statistique, pour l'apprentissage automatique à partir de textes déjà traduits et validés.

Les solutions proposées par la société SYSTRAN sont aujourd'hui utilisées par des entreprises parmi lesquelles certaines possèdent des services de localisation [4]. L'objectif étant d'améliorer l'efficacité et la productivité de ces services en générant automatiquement une première traduction qui sera ensuite révisée. Cette thèse s'inscrit donc pleinement dans les besoins de la société SYSTRAN qui souhaite, à terme, être capable de proposer à ses clients de nouvelles solutions de TA « hyperspécialisées ».

## Organisation du manuscrit

Comme nous venons de le voir, le travail de thèse présenté dans ce manuscrit s'inscrit autour de deux problématiques : comment analyser les retours utilisateurs dans le cadre d'un processus de post-édition, et comment faire évoluer dans le temps un système de TA (par exemple statistique) à partir de ces retours ? Afin d'aborder ces problématiques dans leurs contextes, ce manuscrit est organisé comme suit :

4. Service ayant pour fonction d'adaptater un logiciel à un groupe linguistique ou culturel donné.

– La première partie est dédiée à une présentation des paradigmes fondamentaux de la TA avec un intérêt particulier pour l'approche empirique (également appelée « statistique »)de la TA fondée sur les corpus. Le concept de post-édition à travers la motivation d'une telle pratique, son évaluation ainsi que les approches proposées ces dernières années pour réduire son coût, y sont exposés.
– La deuxième partie est quant à elle dédiée à la présentation du cadre applicatif de nos travaux. Le chapitre 2 est consacré au projet ANR COSMAT. L'ensemble des problématiques de ce projet de recherche, dont certaines dépassent le cadre même de cette thèse, sont présentées.
– Les chapitres 3 et 4 constituent la troisième partie de ce manuscrit. Il s'agit des contributions apportées par ce travail de thèse. Nous y présentons tout d'abord notre réflexion quant à la modélisation de l'intention du post-éditeur, puis nous détaillons la procédure d'adaptation incrémentale pour permettre une intégration dans le temps de nouvelles connaissances dans un système de TA statistique. Chacun de ces chapitres présente les résultats expérimentaux obtenus ainsi qu'une analyse de ces derniers. L'objectif est de prouver qu'il est aujourd'hui possible d'apprendre continuellement du processus de post-édition.

Une clonclusion générale sur les travaux de recherche présentés ici accompagnée de plusieurs perspectives de travail envisagées viendront ensuite clore ce manuscrit.

# Première partie

# Paradigmes fondamentaux

# Chapitre 1

# Paradigmes fondamentaux

## Sommaire


**1.1 Bref résumé historique . . . . . 11**
**1.2 Architectures linguistiques et computationnelles . . . . . 13**
**1.3 Approche experte . . . . . 14**
1.3.1 Traduction automatique directe . . . . . 15
1.3.2 Traduction automatique par transfert . . . . . 15
1.3.3 Traduction automatique interlangua . . . . . 15
**1.4 Approche empirique . . . . . 16**
1.4.1 Modélisation du langage . . . . . 19
1.4.1.1 Modèle de type N-Gramme . . . . . 20
1.4.1.2 Lissage . . . . . 21
1.4.1.3 Modélisation dans un espace continu (*CSLM*) . . . . . 22
1.4.1.4 Évaluation d'un modèle de langage . . . . . 22
1.4.2 Modélisation de la traduction . . . . . 23
1.4.2.1 Alignement sous-phrastique . . . . . 23
1.4.2.2 Modèles IBM . . . . . 24
1.4.2.3 Modélisation basée sur les séquences de mots . . . . . 26
1.4.2.4 Extraction des séquences de mots . . . . . 27
1.4.2.5 Pondération des séquences de mots . . . . . 27
1.4.2.6 Modélisation log-linéaire . . . . . 28
1.4.2.7 Optimisation d'un système de TAS . . . . . 29
**1.5 Approche hybride . . . . . 32**
**1.6 Évaluation de la traduction automatique . . . . . 33**
1.6.1 Évaluation manuelle des traductions . . . . . 33

1.6.2 Évaluation automatique des traductions . . . . . 34
**1.7 Post-Édition . . . . . 39**
1.7.1 Motivation et principe . . . . . 39
1.7.2 Évaluer l'effort de post-édition . . . . . 40
1.7.3 Comment limiter ou réduire cet effort ? . . . . . 41

Les travaux de recherche présentés dans ce manuscrit s'appuient sur certains paradigmes fondamentaux de la traduction automatique (TA) qu'il convient d'introduire préalablement. C'est en ce sens que nous allons décrire les approches majeures qui la composent, dites respectivement « experte » et « empirique ». Nous aborderons par la suite la problématique de l'évaluation de la TA avant de nous intéresser à la post-édition, champ de recherche des plus actifs actuellement et dans lequel s'inscrivent les travaux présentés dans les chapitres suivants.

## 1.1 Bref résumé historique

La TA trouve ses origines dans la première moitié du 20-ième siècle avec l'avènement de l'ère informatique. Il faut en effet remonter jusqu'à la Seconde Guerre Mondiale et l'utilisation des premiers ordinateurs par les Anglais pour retrouver les prémices de ce que sera la TA d'aujourd'hui. Ces derniers cherchaient en effet à craquer les codes de communication de l'armée allemande. Perçue jusqu'alors comme le « simple » décodage d'une information bruitée, Waren Weaver aura en 1947 cette expression pour caractériser la TA dont certains aspects sont toujours d'actualité :

> When I look at an article in Russian, I say : "This is really written in English, but it has been coded in some strange symbols. I now proceed to decode."
>
> *Quand je regarde un article en russe, je dis : « C'est vraiment écrit en anglais, mais il a été codé dans certains symboles étranges. Je procède maintenant au décodage. »*
>
> [Weaver 1947, Weaver 1955]

Le 7 janvier 1954 est une date importante pour la TA puisqu'elle fait référence à la première démonstration publique d'un système de traduction, démonstration destinée à susciter l'intérêt du gouvernement et du grand public en vue d'obtenir des subventions. Celle que l'on nomme l'**expérience Georgetown-IBM**, fruit d'une collaboration entre la société IBM et l'université de Georgetown, consistait en un système de traduction du russe vers l'anglais conçu à partir de 6 règles de grammaire et de 250 mots de vocabulaire. Le succès fut total.

La recherche en TA a dès lors pu bénéficier de mécénats, notamment militaires pour les raisons que nous venons d'évoquer, et cela était d'autant plus motivé qu'à l'époque les scientifiques portaient haut les ambitions quant à l'élaboration rapide de systèmes pleinement opérationnels. Les prétentions étaient alors de fournir des systèmes de traduction automatisée aux performances égales ou supérieures aux traductions humainement produites.

La recherche s'est ainsi poursuivie pendant presque dix ans avant que l'enthousiasme généralisé finisse peu à peu par s'affaiblir. De fait, les problématiques de la linguistique finalement plus complexes qu'imaginées au départ, mais aussi et surtout les faibles ressources et capacités computationnelles de l'époque (comme on peut l'imaginer, très loin de ce que l'on connaît aujourd'hui) ont finit par peser sur le moral des scientifiques. C'est alors qu'un rapport viendra mettre un coup d'arrêt à la recherche en TA dans le milieu des années 60.

Publié en 1966 par l'ALPAC [5] (*Automatic Language Processing Advisory Committee*), ce rapport se montra très sceptique quant aux aboutissements, mais également envers des perspectives portées par des prétentions surévaluées. En conséquence, le gouvernement des États-Unis a réduit de manière drastique les financements jusqu'alors alloués pour les réorienter vers la recherche en linguistique computationnelle pure. Il faudra attendre la fin des années 1970 pour que les travaux en TA reprennent véritablement leur essor. À l'origine de ce renouveau, un besoin de plus en plus important en traduction automatique et une volonté de limiter les coûts, le recours à des traducteurs humains (rémunérés au mot) pouvant s'avérer très onéreux. C'est à cette même période que l'on voit apparaître les premiers systèmes commerciaux de traductions automatisées.

D'un point de vue technologique, plusieurs solutions ont été proposées, depuis les prémices de la TA, jusqu'aux systèmes les plus performants disponibles aujourd'hui. C'est ainsi que l'on distingue dans un premier temps ce que l'on nommera la « première génération » des systèmes de TA basée sur une traduction dite « directe ». Vint ensuite une seconde génération de systèmes avec des technologies de traduction par transfert ou basés sur une abstraction théorique langagière dite « interlangua ». Enfin, approche la plus récente caractérisant une troisième génération de systèmes de TA, celle basée sur l'exploitation probabiliste de données textuelles où l'on retrouve la Traduction Automatique Statistique (TAS) et la TA à base d'exemples.

Dans la suite de ce chapitre nous n'aborderons pas de façon détaillée l'ensemble des technologies existantes dans ce vaste domaine qu'est la TA, très actif et très diversifié. Nous aborderons, pour des raisons évidentes, les technologies les plus communes qui ont un rapport direct avec les travaux de recherche présentés en seconde partie de ce manuscrit.

5. ALPAC : comité composé de sept scientifiques créé deux ans plus tôt par le gouvernement des États-Unis, et chargé par ce dernier d'évaluer les progrès des travaux de recherche en linguistique computationnelle et en TA plus particulièrement

## 1.2 Architectures linguistiques et computationnelles

La traduction, qu'elle soit humaine ou automatisée, est un processus séquentiel. Le texte source à traduire doit tout d'abord être étudié. On en extrait ensuite un ensemble d'informations qui vont permettre d'assurer une certaine qualité à la traduction qui sera produite. En clair, il faut savoir de quoi on parle pour être en capacité de le traduire. On se situe donc dans une phase d'**analyse** du texte source. Alors qu'un traducteur humain va pouvoir déterminer le contexte, le domaine, la sémantique d'un texte à sa lecture (au niveau du document) dans le but de le traduire, un système de TA lui, va devoir procéder à un certain nombre de transformations (majoritairement au niveau de la phrase) pour en extraire les informations utiles par la suite pour produire une traduction.

Seconde étape du processus de traduction : le **transfert**. Cette étape correspond au moment où l'on transpose les informations issues de l'analyse en langue source vers la langue cible. Vient ensuite la troisième et dernière séquence du processus de traduction dite de **synthèse** (ou de génération) où la traduction en langue cible est produite.

Dans les étapes d'analyse et de synthèse, de nombreux systèmes de TA présentent des éléments clairement séparés impliquant différents niveaux de description linguistique : morphologie, syntaxe, sémantique. Par conséquent, l'analyse peut être divisée en analyse morphologique (identification de la fin des mots, composés de mots), l'analyse syntaxique (identification des structures de phrases, de dépendance, de subordination, etc.) ou encore en analyse sémantique (résolution d'ambiguïtés lexicales et structurelles). La phase de synthèse peut, elle, correspondre à une synthèse sémantique (sélection des formes lexicales et structurelles compatibles appropriées), une synthèse syntaxique (génération de phrase requise et des structures de phrases), ou encore en une synthèse morphologique (génération de formes correctes des mots).

Ces différents niveaux d'analyse et de synthèse, associés à des règles de transfert représentent les tracés du triangle de Vauquois présenté en figure 1.1 et composent l'**architecture linguistique** [Boitet 2008] d'un système de TA.

Chaque système de TA a une architecture linguistique qui le caractérise et le différencie des autres. Les règles de transformation qui permettent d'atteindre les différents niveaux de ces architectures, ainsi que les ressources nécessaires, représentent quant à elles l'**architecture computationnelle** [Boitet 2008]. Et de ces architectures computationnelles, on dénombre deux approches majeures : la première dite « experte » s'appuie sur des ressources spécialisées de professionnels de la linguistique, tandis que la seconde dite « empirique » s'appuie elle sur une exploitation probabiliste de données textuelles.
Chacune de ces deux approches sera abordée dans les sections suivantes.

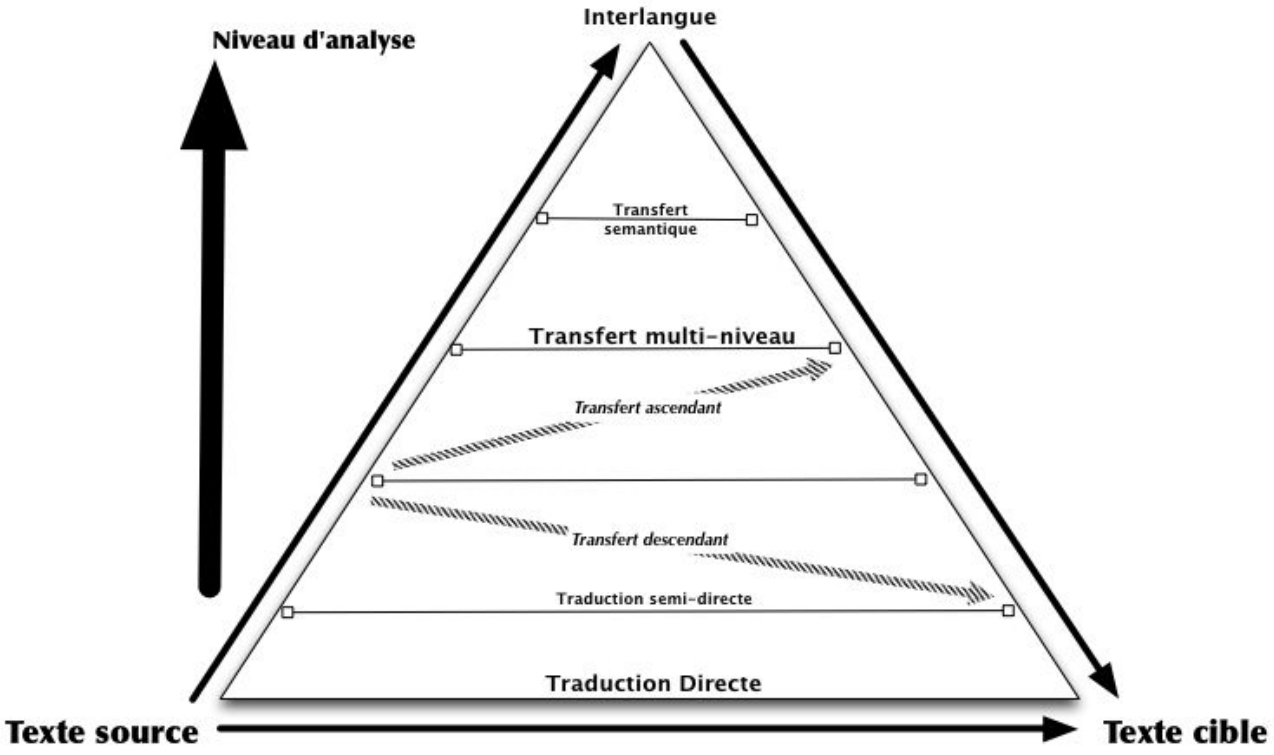


FIGURE 1.1 – Triangle de Vauquois.

## 1.3 Approche experte

Dans le cadre d'une approche experte, le développement d'un système repose sur l'utilisation de connaissances spécialisées de la part de professionnels humains. Ces connaissances sont nécessaires à chaque étape du processus de traduction. Lors de l'analyse du texte source à traduire tout d'abord. Ce dernier subira une suite de transformations plus ou moins importantes destinées à en extraire de plus petites unités : morphosyntaxiques, grammaticales ou sémantiques. Vient ensuite une phase de transfert où les unités sources précédemment citées sont transmuées en unités cibles par le biais de règles de transfert. Ces unités cibles sont alors utilisées pour générer une traduction en langue cible du texte source.

Chaque système de traduction basé sur cette approche possède un niveau d'analyse et des règles de transfert qui lui sont propres. En utilisant la représentation de Vauquois en figure 1.1, ces systèmes experts peuvent être répartis en trois composantes de cette approche et identifiées comme étant : la TA « directe », la TA à base de règles de transfert et la TA fondée sur l'interlangue.

### 1.3.1 Traduction automatique directe

La TA **directe** se situe à la base du triangle de Vauquois. On y retrouve des systèmes « bas-niveau » qui n'utilisent aucune représentation intermédiaire et reposent sur la simple consultation d'un dictionnaire bilingue. Concrètement, la traduction se fait mot-à-mot et de simples règles de réordonnancement sont appliquées sur les mots traduits afin de générer une traduction en langue cible. Cette approche est simpliste dans sa mise en application puisqu'elle ne requiert pas de connaissances expertes dans l'une ou l'autre des langues concernées. L'avantage est donc de pouvoir proposer une traduction, quand bien même la structure grammaticale de la phrase source est incorrecte. Toutefois, les systèmes basés sur cette approche sont limités à une paire de langues.

### 1.3.2 Traduction automatique par transfert

Les systèmes à base de **règles de transfert** requièrent une analyse plus approfondie du texte source, et se situent au centre du Triangle de Vauquois, au-dessus de la TA directe. La particularité de ces systèmes, c'est que le transfert peut s'effectuer à différents niveaux, que ce soit du côté de la langue source, ou de celui de la langue cible. On parle alors de transfert ascendant ou descendant. Cette méthodologie est plus facilement adaptable à plusieurs langues, et ce sans nécessiter une analyse poussée de chacune d'elles. Elle représente la majorité des systèmes actuels basés sur l'approche experte, dont notamment le moteur de traduction historique de la société SYSTRAN [Senellart 2001].

### 1.3.3 Traduction automatique interlangua

À l'opposé, les systèmes fondés sur l'**interlangue** s'appuient sur l'utilisation d'un métalangage résultant d'une analyse poussée de la langue source. Le processus de traduction se résume à la simple transformation du texte source et à la génération de sa traduction en langue cible. L'avantage de cette approche est que l'analyse réalisée pour permettre la transformation vers le métalangage peut-être appliqué à plusieurs paires de langues.

Cependant, ces architectures linguistiques qui renseignent sur les représentations intermédiaires utilisées par les systèmes de TA ne permettent pas de connaître comment ces mêmes systèmes passent d'une représentation à une autre.

## 1.4 Approche empirique

L'approche empirique de la TA tire ses origines du volume toujours croissant de corpus de données textuelles. Opposée sur le principe à l'approche experte précédemment évoquée, l'approche empirique ne requiert pas de posséder obligatoirement de connaissances spécifiques préalables pour l'une ou l'autre des deux langues. C'est en partie la raison pour laquelle cette approche est utilisée pour élaborer des systèmes sur des langues peu communes, même si cela signifie aussi en règle générale peu de données. On distingue deux types d'approche empirique : la TA statistique et la TA fondée sur les exemples. Bien que cette dernière ne soit pas une technologie utilisée dans le cadre des travaux de cette thèse, il est intéressant de savoir qu'elle existe. C'est pourquoi elle est évoquée brièvement dans la section suivante.

### Traduction automatique fondée sur les exemples

La TA basée sur les exemples (*Example-based Machine Translation* en anglais) fut initiée par [Nagao 1984]. Sa proposition tient dans l'observation du comportement du cerveau humain devant la tâche de traduction. Ce dernier ne va pas tenter une analyse linguistique profonde comme peuvent le faire certains types de systèmes que nous avons abordés précédemment, mais va plus simplement s'appuyer sur des exemples déjà rencontrés. En utilisant tout ou partie de ces exemples, il va les adapter en vue de produire une traduction de ce qu'il souhaite.

Cette approche nécessite de collecter et de stocker préalablement des paires de traductions (source et traduction équivalente en langue cible) qui lui serviront par la suite. Le processus de traduction se déroule alors en trois temps :

1. On va dans un premier temps extraire de notre base d'exemples les séquences sources qui se rapprochent le plus de ce que l'on cherche à traduire ;

Pour effectuer ce calcul de similarité entre séquences, plusieurs méthodes existent, mais s'opposent dans leurs approches. Certaines vont considérer la phrase source à traduire dans son ensemble, tandis que d'autres vont procéder à une segmentation de celle-ci. Par exemple, [Brown 1996] va segmenter la phrase source à traduire et rechercher pour chacun de ces segments, les segments similaires dans la base d'exemples. [Veale 1997] lui, propose de trouver la phrase source dans la base d'exemples la plus analogue à la phrase source à traduire. Ces deux approches ont ceci de commun qu'elles s'intéressent à la similarité au niveau des mots, alors que d'autres approches vont décomposer, et la phrase source à traduire, et les phrases de la base d'exemples [Nagao 1984, Deniz 2008]. La similarité entre phrases étant déterminée par rapport

aux résultats de cette analyse.

2. On extrait de notre base d'exemples des séquences cibles qui sont associées aux séquences sources que nous avons extraites à la phase 1 ;

Cette seconde phase peut-être considérée comme une phase d'alignement puisqu'on associe aux séquences en langue source, leurs traductions équivalentes en langue cible.

3. On tente de combiner les séquences cibles que nous venons d'extraire de sorte qu'on obtienne une traduction exploitable en langue cible.

Bien qu'au départ cette approche fondée sur les exemples fut proposée en complément d'une approche par règles, la TA basée sur les exemples s'est rapidement imposée comme une alternative viable et concurrente de cette dernière, et ce pour plusieurs raisons.

Tout d'abord, l'approche par analogie s'appuie comme nous venons de le souligner sur des traductions antérieures. En cela, il s'agit d'un historique déjà validé qui garantit de fait une certaine qualité de traduction. De plus, alors que l'approche experte nécessite le recourt à des ressources humaines expertes pour son évolution ([ré]écriture de nouvelles règles, adaptation à de nouvelles langues, coût de développement, complexité, etc.), l'approche par analogie n'a elle besoin que de collecter et conserver un historique de traductions déjà générées et validées pour évoluer.

## Traduction Automatique Statistique

Paradigme le plus récent dans le domaine de la TA, la Traduction Automatique Probabiliste (TAP) dite « Traduction Automatique Statistique » (TAS) s'appuie sur l'exploitation mathématique de corpus textuels dont le volume n'a cessé de croître ces dernières années. Concrètement, la TAS est une combinaison probabiliste dont le but est de fournir une traduction candidate $\hat{e}$ répondant aux principes fondamentaux de la traduction qui pour rappel sont : être juste par rapport à la phrase source $f$ et être vraisemblable dans la langue cible. Pour atteindre cet objectif, la TAS s'appuie sur un modèle de traduction qui va proposer une correspondance en langue cible du vocabulaire en langue source, et d'un modèle de langage qui lui doit valider grammaticalement la traduction en langue cible. Un décodeur statistique combinant ces deux modèles produira en sortie une traduction $\hat{e}$ la plus probable pour une phrase $f$ donnée, tel qu'illustré en figure 1.2 :

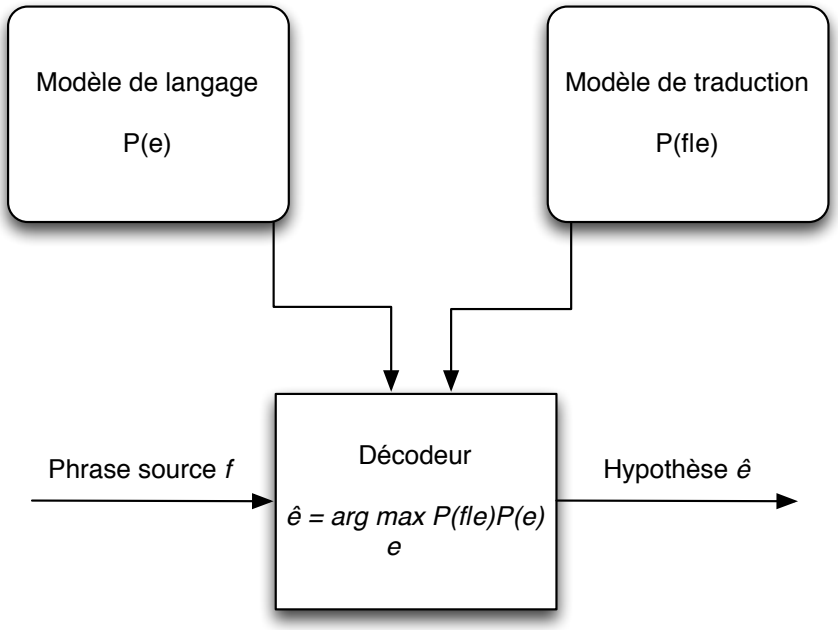


FIGURE 1.2 – Schéma de la traduction automatique statistique.

La définition mathématique de ce principe, inspirée de la recherche en Reconnaissance Automatique de la Parole (RAP), est présentée à l'équation 1.1 :

$$\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} \frac{P(f|e) \times P(e)}{P(f)} \tag{1.1}$$

$P(f)$ et $P(e)$ correspondent respectivement au modèle de langage en langue source et cible, tandis que $P(f|e)$ qui se lit « probabilité de la phrase $f$ sachant la phrase $e$ » correspond au modèle de traduction.

Le décodeur quant à lui est représenté par la fonction mathématique $argmax$ qui correspond à l'ensemble des arguments pour lesquels une expression atteint sa valeur maximale, ce que l'on a appelé précédemment la traduction candidate la plus probable, et qui sera notée $\hat{e}$.
Par ailleurs, la probabilité $P(f)$ de la phrase source $f$ n'ayant aucune influence sur le résultat du décodeur, elle peut être retirée de l'équation 1.1 qui se trouve alors simplifiée comme suit :

$$\hat{e} = \arg\max_{e} P(f|e) \times P(e) \tag{1.2}$$

On notera ici que le sens de lecture du modèle de traduction est inverse à son application factuelle. Les langues source et cible considérées se trouvent inversées en raison de l'application du théorème de Bayes dans la définition mathématique. Pour autant, et afin d'éviter toute confusion dans la suite de ce manuscrit, le modèle de traduction tel que nous le considérerons sera noté $P(e|f)$.

### 1.4.1 Modélisation du langage

Il est bien évident qu'en TA, l'objectif n'est pas de simplement produire en langue cible une suite de mots, mais bel et bien que cette suite de mots corresponde à une traduction lisible et exploitable, ce qui implique que les mots soient associés dans le bon ordre. Dans l'approche empirique de la TA, la modélisation statistique du langage cherche ainsi à garantir la vraisemblance grammaticale en langue cible des hypothèses de traduction qui pourront être proposées en sortie du décodeur. Ceci permettant *de facto* d'écarter un sous-ensemble de traductions candidates qui ne seraient pas valides.

Considérons $W = w_1 w_2 \dots w_k$, une séquence de $k$ mots dans une langue $L$. La définition mathématique de la modélisation du langage s'exprime alors :

$$P_L(W) = P_L(w_1 w_2, \cdots, w_k) = \prod_{i=1}^{k} P_L(w_i|w_1, \cdots, w_{i-1}) \tag{1.3}$$

L'équation 1.3 montre que pour estimer la vraisemblance grammaticale de $W$ dans $L$, le modèle de langage s'appuie sur $P_L(w_i|w_1, \cdots, w_{i-1})$, soit la probabilité d'apparition des mots de $W$ introduisant un historique $h_i$ tel que $h = w_1 \dots w_{i-1}$. L'équation 1.3 peut donc être simplifiée pour donner l'équation 1.4.

$$P_L(W) = \prod_{i=1}^{k} P_L(w_i|h_i) \tag{1.4}$$

Pour estimer la probabilité d'apparition $P_L(w_i|h_i)$, le modèle de langage est entraîné sur un corpus d'apprentissage monolingue. Quel que soit le volume de ce corpus d'apprentissage, il se doit avant tout d'être caractéristique de la langue $L$. La probabilité d'apparition étant calculée à partir du nombre d'occurrences de $h_i$ dans ce corpus en langue $L$ telle que :

$$P_L(w_i|h_i) = \frac{c(w_i h_i)}{c(h_i)} \tag{1.5}$$

#### 1.4.1.1 Modèle de type N-Gramme

Nous venons de le voir, la modélisation du langage en TAS repose sur l'historique d'occurrences d'une séquence dans un corpus d'apprentissage. Ceci n'est toutefois pas sans difficulté puisque la taille de l'historique à considérer dépend directement de la taille de la séquence que l'on cherche à déterminer. Bien que le corpus d'apprentissage soit d'un volume conséquent, il apparaît improbable qu'il contienne tous les historiques, pour tous les mots. Il devient donc impossible d'estimer toutes les probabilités $P(w_i|h_i)$. Comme alternative à cette problématique, nous utilisons la modélisation de type $n$-gramme qui limite l'historique d'un mot $w_i$ aux seuls $n-1$ mots qui le précèdent. À noter que cette approche représente de nos jours la méthodologie prédominante du domaine.

On distingue les modèles $n$-gramme par la profondeur de l'historique qu'ils conservent. Ainsi, un modèle unigramme (d'ordre $n = 1$) ne conservera aucun historique puisqu'il s'agit du mot lui-même. En revanche, les modèles bigrammes (d'ordre $n = 2$) et trigrammes (d'ordre $n = 3$) conservent un historique sur respectivement deux et trois mots. Et ainsi de suite.

Toutefois, compte tenu des contraintes citées précédemment, les modèles de langage de type $n$-gramme dépassent rarement le degré 5, même si l'on commence à entrevoir des modèles de degré $n = 7$, cela reste exceptionnel. La définition mathématique d'un modèle de langage de type $n$-gramme d'ordre $n \geq 2$ est décrite par l'équation 1.6 :

$$P(W) \simeq P(w_1) \prod_{i=n}^{k} P(w_i|w_{i-n+1}, \cdots, w_{i-1}) \tag{1.6}$$

Plusieurs approches existent pour estimer l'ensemble des probabilités à partir de notre corpus d'apprentissage [Federico 1998]. L'une d'entre elles consiste en l'**estimation par maximum de vraisemblance** de la distribution des probabilités du modèle sur le corpus d'apprentissage. Cette approche, la plus couramment utilisée, s'exprime comme suit :

$$P_{MV}(w_i|w_{i-n+1},\cdots,w_{i-1}) = \frac{c(w_{i-n+1},\cdots,w_i)}{c(w_{i-n+1},\cdots,w_{i-1})} \tag{1.7}$$

où $c(w_{i-n+1},\cdots,w_{i-1})$ représente la fréquence d'apparition du $n$-gramme $w_{i-n+1},\cdots,w_{i-1}$ dans le corpus d'apprentissage.

#### 1.4.1.2 Lissage

Bien que le corpus d'apprentissage se doit d'être représentatif de la langue qu'il va servir à modéliser, nous venons de voir qu'il est utopique de penser qu'il permettra de couvrir l'ensemble des $n$-grammes pouvant être observés durant le processus de traduction. C'est notamment le cas des langues peu communes où les ressources sont limitées. De fait, il est nécessaire d'ajuster les probabilités du modèle de langage en vue de réduire l'effet induit par le manque de données : c'est le principe du lissage (*smoothing* en anglais).

Pour se faire, il existe plusieurs techniques repertoriées par [Chen 1996], dont l'objectif est d'associer une probabilité non nulle aux $n$-grammes jusqu'alors non observés. Une de ces approches, proposée par [Katz 1987], consiste en un lissage par repli (*back-off* en anglais). Lorsqu'un n-gramme d'ordre $n$ et un historique donné n'ont pas été observés dans le corpus d'apprentissage, ce sont des $n$-grammes d'ordres inférieurs qui vont être utilisés et l'historique considéré pour ces $n$-grammes s'en trouve alors restreint. De fait, l'éventualité qu'ils aient été observés dans le corpus d'apprentissage et qu'une probabilité d'observation y soit associée augmente. Un coefficient de normalisation lui est alors attribué afin de garantir la bonne cohésion de la distribution des probabilités sur l'ensemble des $n$-grammes. À noter qu'il existe également une variante du lissage par repli proposée par Kneser-Ney dit « modifié > [Chen 1996] qui compte parmi les méthodes de lissage les plus utilisées actuellement.

D'autres techniques de lissage existent parmi lesquelles nous pouvons citer le lissage par interpolation linéaire [Jelinek 1980, Witten 1991]. Ce principe consiste à déterminer le modèle lissé d'ordre $n$ à partir d'une interpolation linéaire entre le modèle non lissé d'ordre $n$ et le modèle lissé d'ordre $n-1$.

#### 1.4.1.3 Modélisation dans un espace continu (*CSLM*)

Dans les modèles de langage de type $n$-gramme avec repli, les mots sont représentés dans un espace dit « discret » représenté par le vocabulaire. L'obtention d'une vraie interpolation des probabilités pour les $n$-grammes non observés s'en trouve limitée puisqu'un changement dans cet espace discret peut entraîner un changement arbitraire de la probabilité des $n$-grammes.

Une approche alternative s'appuie sur une représentation des mots dans un espace continu [Bengio 2003, Schwenk 2007]. L'intérêt de cette approche fut largement démontré ces dernières années que ce soit en TAS [Schwenk 2006, Schwenk 2007] mais également dans le domaine de la RAP [Schwenk 2002, Schwenk 2005].

#### 1.4.1.4 Évaluation d'un modèle de langage

L'objectif de modéliser le langage est, rappelons-le, de valider la bonne vraisemblance grammaticale de nos traductions dans la langue cible. Ainsi, nous garantissons une certaine qualité quant à la lisibilité de nos traductions. Il va de soit qu'il nous faut avant tout nous assurer de la bonne qualité du modèle en lui-même. On va pour se faire calculer la probabilité que notre modèle attribut à un corpus de développement que nous savons composé de phrases grammaticalement correctes. Probabilité que nous chercherons alors à maximiser.

La métrique d'évaluation d'un modèle de langage est communément appelée **perplexité** et se dénote : *ppl*. Sa définition mathématique s'appuie sur l'entropie croisée exprimée de la façon suivante :

$$\begin{aligned} H(P_{LM}) &= \frac{1}{n} \log P_{LM}(w_1, w_2, \cdots, w_n) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \log P_{LM}(w_i | w_1, \cdots, w_{i-1}) \end{aligned} \tag{1.8}$$

avec $P_{LM}$ la probabilité du modèle de langage pour une séquence de mots W, et $w_1, \cdots, w_{i-1}$ l'historique du $i$-ème mot. La perplexité étant quant à elle obtenue par simple transformation, telle que :

$$PPL(W) = 2^{H(P_{LM})} \tag{1.9}$$

### 1.4.2 Modélisation de la traduction

Le rôle du modèle de traduction est de déterminer statistiquement qu'une phrase source donnée se traduise en une phrase cible équivalente. L'apprentissage d'un tel modèle repose sur l'exploitation de corpus d'apprentissage parallèles (dits « alignés » ou « bilignues »). Ces corpus parallèles sont des corpus bilingues qui couplent deux ensembles de données textuelles alignés au niveau de la phrase et tels que l'un est la parfaite traduction de l'autre. De ce couple, on peut alors extraire des correspondances (ou alignements) entre deux langues comme l'illustre un exemple d'alignement français-anglais à la figure 1.3 :

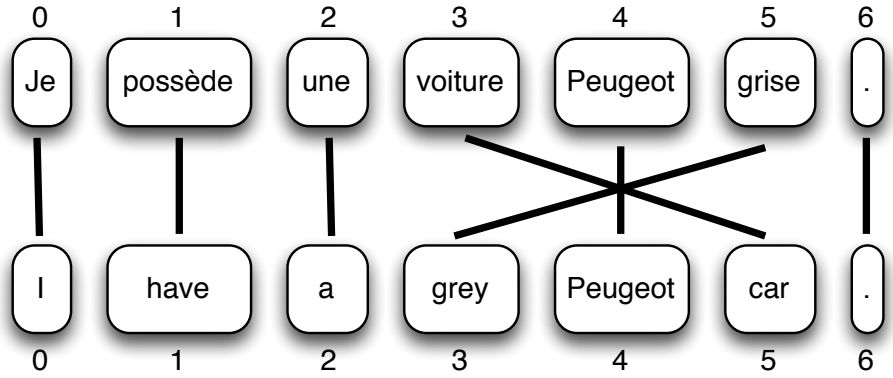


FIGURE 1.3 – Exemple d'alignement des mots dans une paire de phrases français-anglais.

La plus petite unité sur laquelle reposaient initialement les premiers modèles de traduction était le mot. Une telle modélisation s'est avérée être limitée, un mot dans une langue pouvant par exemple s'aligner à un ou plusieurs mots dans une autre. Ces modèles furent alors supplantés par une approche s'avérant plus adaptée, et donc permettant une modélisation plus efficace. Basée cette fois sur les séquences de mots, on parlera alors de *Phrase-based Machine Translation (PBMT) system*. Par définition, une séquence de mots est une suite de mots contigus, tandis qu'une paire de séquences (*phrase pair* en anglais) est un couple de séquences de sens équivalent entre deux langues.

L'apprentissage de modèles de traduction repose donc avant tout sur une notion d'alignement sous-phrastique que nous allons maintenant aborder.

#### 1.4.2.1 Alignement sous-phrastique

Pour bien appréhender la notion d'alignement, prenons un couple de phrases $(e, f)$ telles que l'une est la traduction de l'autre, avec une phrase source $f$ telle que $f = f_1 \cdots f_{l_f}$, et une phrase cible $e$ telle que $e = e_1 \cdots e_{l_e}$, où $l_f$ et $l_e$ sont respectivement la longueur de $f$ et $e$.

Considérant l'ensemble des alignements possibles, la vraisemblance $P(e|f)$ est alors définie par :

$$P(e|f) = \sum_{a} P(e, a|f) \tag{1.10}$$

où $a$ est une fonction d'alignement entre le $i$-ième mot de $e$ et le $j$-ième mot de $f$, telle que :

$$a : e_i \rightarrow f_j \tag{1.11}$$

Appliquée à notre exemple de la figure 1.3, la fonction d'alignement $a$ serait alors :

$$a : \{0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 5, 4 \rightarrow 4, 5 \rightarrow 3, 6 \rightarrow 6\} \tag{1.12}$$

À noter qu'il se peut toutefois qu'un mot ne s'aligne avec aucun autre de la langue opposée, il se trouve alors aligné avec le mot spécial « NULL ». Par ailleurs, l'aspect asymétrique de l'alignement n'autorise pas que plusieurs mots de la langue source puissent s'aligner avec un même mot de la langue cible. Les modèles IBM permettent d'y remédier en réalisant un alignement dans les deux directions : cible vers source et source vers cible. Ces alignements sont alors symétrisés par l'intersection des points d'alignement [Och 2003d].

#### 1.4.2.2 Modèles IBM

Un lexique se définit comme étant l'ensemble des mots d'une langue, son vocabulaire. Une traduction lexicale est alors par définition une traduction d'un mot par un mot. À noter qu'ici, on évoque « une » traduction et non « la » traduction, puisqu'il est fondamental de comprendre qu'il n'existe pas forcément une et une seule traduction équivalente d'un mot pour une langue cible donnée. Une raison à cela étant que dans un contexte, un mot peut avoir un sens différent, mais également, qu'une traduction équivalente de ce mot peut être plus couramment employée qu'une autre. Par exemple : le mot « bank » en anglais pouvant se traduire en français par le mot « banque », mais également par le mot « berge » ou « flanc » (de colline). Tout dépend du context dans lequel nous évoluons, ici : financier, aquatique ou topographique. Le fait qu'un mot soit traduisible par $n$ traductions équivalentes introduit la notion de **probabilité de traduction lexicale**.

Les modèles IBM tels qu'introduits par [Brown 1993] sont des modèles probabilistes qui s'appuient sur cette notion de probabilité de traduction lexicale pour modéliser statistiquement la traduction. Au nombre de cinq, les modèles IBM sont des modèles itératifs qui reposent sur le modèle précédent en y associant une propriété supplémentaire. Nous allons maintenant les voir plus en détails.

### Modèle IBM 1 – Traduction lexicale

La première hypothèse énoncée par [Brown 1993] est que pour une paire de phrases $(e, f)$, chaque mot $f_i$ de $f$ s'aligne de façon équiprobable avec chacun des mots $e_j$ de $e$, et ce quelque soit l'ordonnancement de ces alignements. Ainsi, le modèle IBM 1 ne s'appuie que sur la probabilité de traduction lexicale au niveau des mots.

Reprenons : pour exprimer la définition mathématique de la modélisation IBM 1, la paire de phrases $(e, f)$ telle que nous l'avons définit précédemment, et à laquelle nous associons une fonction d'alignement $a$ entre chaque mot $e_j$ de $e$ et $f_i$ de $f$. La modélisation IBM 1 s'exprime alors :

$$P(e, a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) \tag{1.13}$$

Ce premier modèle IBM résulte par la création d'une table de traduction qui n'est autre qu'un fichier texte qui associe à un mot source, un mot cible et une probabilité de traduction entre eux.

### Modèle IBM 2 – Réordonnancement

L'hypothèse du modèle IBM 1 sur l'équiprobabilité des alignements n'est plus considérée et le modèle IBM 2 associe désormais à la table de traduction introduite précédemment un modèle probabiliste d'alignement, tel que :

$$P(e, a|f) = \epsilon \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) a(a(j)|j, l_e, l_f) \tag{1.14}$$

où $a(i|j, l_e, l_f)$ représente la probabilité que le mot $e_i$ à la position $a_j = i$ et le mot $f_j$ à la position $j$ soient alignés.

### Modèle IBM 3 – Fertilité

Le modèle IBM 3 vient corriger deux insuffisances des modèles IBM 1 et IBM 2 que l'on vient de voir. La première est la possibilité qu'un mot source puisse ne pas être traduit par une équivalence en langue cible. Le modèle IBM 3 va donc considérer la probabilité de l'insertion nulle (« NULL » pour rappel). La seconde est qu'un mot source peut cette fois avoir une équivalence en langue cible, non pas avec un et un seul mot, mais plusieurs. C'est ce qu'on appelle la **fertilité**.

**Modèle IBM 4 – Distorsion relative**

Pour ce quatrième modèle, une nouvelle probabilité est introduite : la probabilité de **distorsion relative**. Celle-ci dépend de la position des mots alignés dans $e$ et $f$ ainsi que de leurs positions respectives, mais également de l'alignement potentiel d'autres mots de $f$ avec le mot considéré dans $e$.

**Modèle IBM 5 – Déficience**

Les résultats obtenus avec le modèle IBM 4 sont très bons, mais ce dernier n'est pas mathématiquement juste. De fait, le modèle de distorsion du modèle IBM 4 ne considère pas les positions cibles déjà sélectionnées pour un alignement et la masse des probabilités d'alignement s'en trouve tronquée. Le modèle IBM 5 vient corriger cette **déficience** du modèle IBM 4, en conservant un historique des positions toujours vacantes durant le processus d'alignement, elles seules pouvant être alors assujetties à un potentiel nouvel alignement.

#### 1.4.2.3 Modélisation basée sur les séquences de mots

Historiquement, les modèles de traduction étaient entraînés sur les seuls mots comme étant la plus petite unité textuelle d'apprentissage. De nos jours, ces modèles sont entraînés à partir de séquences de mots contigus (*phrases* en anglais). On parle alors d'un système de TA basé sur les séquences de mots, ou *Phrase-based Statistical Machine Translation (PBSMT) system* en anglais. À titre d'exemples, [Bertoldi 2006] et [Matusov 2006] sont des systèmes basés sur cette approche.

Ce changement de paradigme est destiné à contrer les lacunes d'une modélisation basée sur les mots uniquement. Il paraît en effet évident *a posteriori* qu'un mot en langue source ne se traduit pas toujours uniquement en un seul mot en langue cible. De même qu'une suite de mots en langue source peut également se traduire en un seul mot équivalent en langue cible, comme illustré en figure 1.4.

Mathématiquement, et en repartant de l'équation 1.1, la modélisation basée sur les séquences de mots peut être formalisée comme suit :

$$P(e|f) = \prod_{i=1}^{I} \phi(\bar{e}_i|\bar{f}_i) d(start_i - end_{i-1} - 1) \tag{1.15}$$

où $\phi(\bar{e}_i|\bar{f}_i)$ est le terme représentant la traduction modélisée entre les séquences $\bar{e}_i$ et $\bar{f}_i$, et $d(start_i - end_{i-1} - 1)$ le terme représentant le modèle de réordonnancement basé sur la distance relative. Ainsi, on constate que le réordonnancement d'une séquence est basé sur la séquence

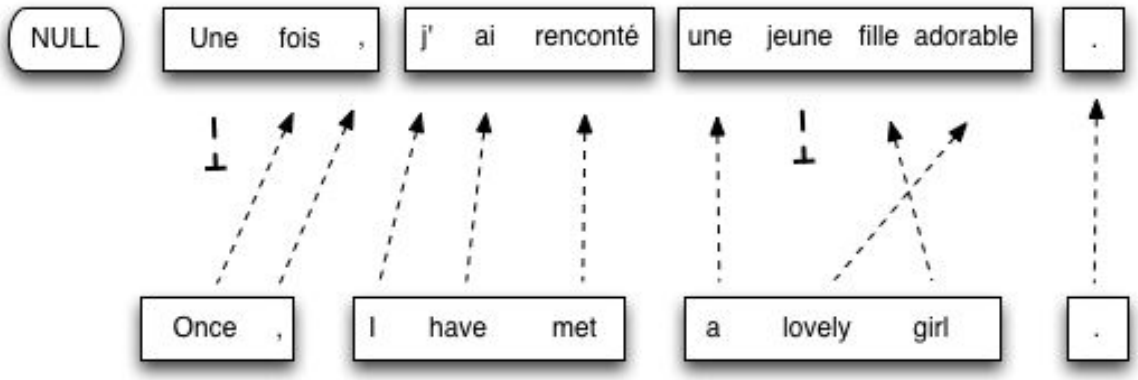


FIGURE 1.4 – Traduction automatique statistique basée sur les séquences de mots.

qui la précède avec $start_i$ et $end_i$, respectivement les points de départ et d'arrivée de la $i$-ème séquence.

#### 1.4.2.4 Extraction des séquences de mots

L'extraction des séquences de mots est un processus bilatéral à partir duquel on va générer ce qu'on appelle des **paires de séquences** (*phrase pairs* en anglais). Le critère à respecter pour l'extraction d'une paire de séquences est la **consistance** entre ces séquences. À ce titre, une paire de séquences $(\bar{e}|\bar{f})$ est dite consistante dans un alignement $a$, si tous les mots $f_1, \cdots f_n$ dans $\bar{f}$ qui ont un point d'alignement dans a, l'ont avec les mots $e_1, \cdots, e_n$ de $\bar{e}$, et inversement.

Cette définition de la consistance pour une paire de séquences est illustrée par la figure 1.5 [6]. Ainsi, on peut constater que l'exemple de gauche est consistant puisque tous les points d'alignement présents dans la paire de séquences matérialisée par le rectangle gris sont alignés les uns avec les autres. Inversement, l'exemple central lui ne respecte pas cette règle puisque le point d'alignement matérialisé par une croix blanche se trouve en dehors de la paire de séquences. Sur le troisième et dernier exemple à droite, c'est consistant avec toutefois la particularité d'inclure le mot (*is*) sans point d'alignement. Ceci n'est pas aberrant puisque n'ayant pas de point d'alignement, il ne transgresse pas, de fait, la règle de consistance que nous venons d'introduire.

#### 1.4.2.5 Pondération des séquences de mots

La probabilité de traduction pour une paire de séquences $\phi(\bar{e}|\bar{f})$ est estimée à partir de la fréquence relative de la séquence cible pour une séquence source donnée. Cette estimation

6. L'exemple donné est inspiré du chapitre 5 du livre « Statistical Machine Translation » de Philipp Koehn.

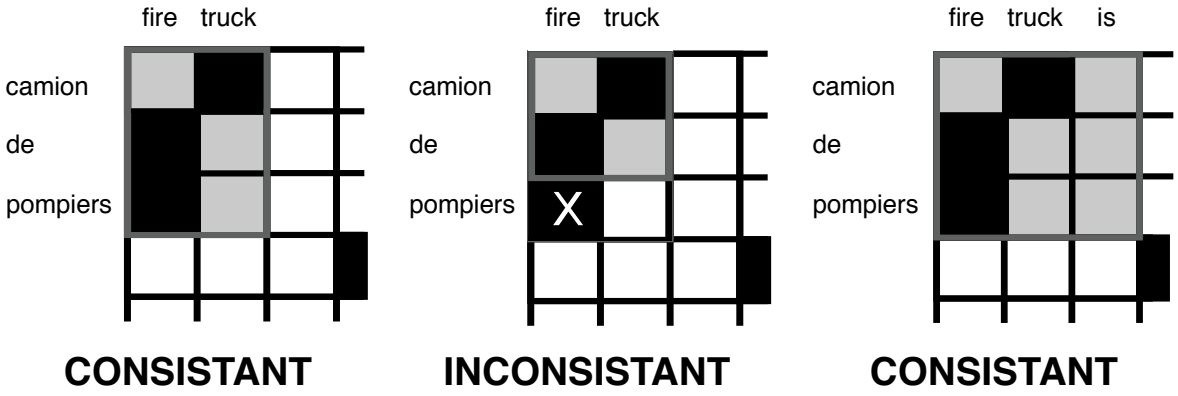


FIGURE 1.5 – Exemples de paires de séquences consistantes et non consistantes.

est réalisée en calculant le maximum de vraisemblance dont la définition mathématique est la suivante :

$$\phi(\bar{e}|\bar{f}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{e}_i} count(\bar{f}, \bar{e}_i)} \tag{1.16}$$

À noter qu'il existe des alternatives à la modélisation par séquences de mots telles que l'approche basée sur la syntaxe [Yamada 2001], l'approche hierarchique [Chiang 2005], très actuelle, ou encore l'utilisation de modèles de traduction factorisés [Koehn 2007a] pour les citer à titre d'exemples, sans les décrire davantage.

#### 1.4.2.6 Modélisation log-linéaire

Un système de TAS basé sur les séquences de mots est une association de modèles probabilistes combinant : un modèle de traduction $\phi(\bar{e}_i|\bar{f}_i)$, un modèle de réordonnancement $d$, et un modèle de langage $P_{LM}(e)$. Ainsi, l'équation 1.2 se joint à l'équation 1.15 pour donner :

$$\hat{e} = \arg\max_{e} \prod_{i=1}^{I} \phi(\bar{e}_i|\bar{f}_i) d(start_i - end_{i-1} - 1) P_{LM}(e) \tag{1.17}$$

qui une fois factorisée devient :

$$\hat{e} = \arg\max_{e} \prod_{i=1}^{I} h_i(x) \tag{1.18}$$

où

$$\begin{aligned} x &= variable\,(e,\ f,\ start,\ end) \\ h_1 &= \log \phi \\ h_2 &= \log d \\ h_3 &= \log P_{LM} \end{aligned}$$

De fait, la sortie du système résultant d'une combinaison homogène de modèles probabilistes, il peut être intéressant d'agir au niveau de chacun d'eux dans le but de favoriser leur distribution de probabilités. Pour se faire, on va leur associer respectivement un poids $\lambda_i$ :

$$\hat{e} = \arg\max_{e} \prod_{i=1}^{I} h_i(x)^{\lambda_i} \tag{1.19}$$

En agissant sur ce poids, il est possible d'avoir une influence sur la sortie d'un système, tels la « paramétrisation » du dit système (*features* en anglais).
Exprimée cette fois dans le domaine logarithmique, la pondération exprimée à l'équation 1.19 devient :

$$\hat{e} = \exp \sum_{i=1}^{n} \lambda_i h_i(x) \tag{1.20}$$

Et c'est ainsi que la **modélisation log-linéaire** de notre système de TAS basé sur les séquences de mots s'exprime sous la forme :

$$\begin{aligned} \hat{e} &= \exp[\lambda_\phi \sum_{i=1}^{I} \log \phi(\bar{e}_i|\bar{f}_i) \\ &+ \lambda_d \sum_{i=1}^{I} \log d(start_i - end_{i-1} - 1) \\ &+ \lambda_{LM} \sum_{i=1}^{I} \log P_{LM}(e)] \end{aligned} \tag{1.21}$$

#### 1.4.2.7 Optimisation d'un système de TAS

La modélisation log-linéaire que nous venons de voir à la section précédente associe à chaque modèle probabiliste un poids $\lambda_i$, tel qu'illustré par l'équation 1.21. Par l'intermédiaire de ces poids, nous allons pouvoir agir sur la combinaison de ces modèles et implicitement,

agir sur la qualité de traduction en sortie du système de TAS. La paramétrisation de ces poids doit par conséquent être optimum pour garantir la meilleure qualité de traduction, en tirant le meilleur parti des modèles probabilistes.

L'optimisation de ces poids (*tuning* en anglais) se fait à partir d'un corpus de développement et par itérations successives. Avec ce corpus de développement pour lequel on possède une ou plusieurs traductions de références, on va chercher à minimiser les erreurs de traduction jusqu'à obtenir un minimum local. Cette tâche reste toutefois complexe, notamment parce que l'optimisation est réalisée dans un espace pluridimensionnel et qu'un minimum local peut tout à fait ne pas correspondre à la configuration la plus optimale. De plus, l'aspect itératif de l'optimisation et sa complexité en font une tâche très coûteuse en temps de calcul.

#### Minimum Error Rate Training (MERT)

L'algorithme le plus communément utilisé pour l'entraînement et l'optimisation des poids des modèles est l'algorithme du *Minimum Error Rate Training* (MERT) proposé par [Och 2003c]. Ce dernier fixe des poids initiaux de façon aléatoire avant d'évoluer par itération (une douzaine en moyenne) à la recherche d'un minimum local. La meilleure combinaison des poids des modèles étant celle qui minimisera les erreurs de traduction en sortie du système. Pour évaluer la qualité des traductions produites par le système pour une combinaison de poids donnée, MERT s'appuie sur un score référence calculé automatiquement : le score BLEU (voir section 1.6.2).

Cependant, MERT est parfois décrié car il lui arrive occasionnellement de « s'égarer » lors de l'optimisation. Il est donc courant de procéder à plusieurs séries d'optimisations, pour ensuite faire une moyenne des résultats obtenus afin de prévenir toute dégradation des performances du système. Ce qui serait alors contraire à l'effet attendu d'une optimisation.

#### Margin Infused Relaxed Algorithm (MIRA)

L'algorithme MIRA [Crammer 2003] est une alternative à MERT de plus en plus appréciée par la communauté scientifique. Implémenté par [Hasler 2011], il tend en effet à résoudre certaines problématiques évoquées précédemment, notamment celle sur l'instabilité, mais également le fait que les poids finaux dépendent des poids initiaux de l'optimisation.

MIRA est conçu pour apprendre et optimiser un ensemble de paramètres (vecteur de poids) par traitement un à un de tous les exemples d'apprentissage donnés. Après chaque traduction d'une nouvelle phrase par le décodeur, les poids sont optimisés si l'algorithme se trompe avec une marge associée à une fonction de perte. En pratique, les poids sont légèrement réévalués si la différence en termes de scores du modèle entre une traduction de référence et une hypothèse

de traduction est au moins aussi importante que la perte entre elles (en termes de score BLEU par exemple).

## 1.5 Approche hybride

Nous venons de présenter dans les sections 1.3 et 1.4 deux approches en TA. L'une s'appuie sur un ensemble de connaissances expertes tandis que l'autre s'appuie sur une exploitation probabiliste de données textuelles. La frontière entre ces deux approches n'est toutefois pas étanche. De ce fait, on a pu constater ces dernières années l'émergence de systèmes de traduction tirant profit d'une combinaison des deux méthodes. De tels systèmes sont qualifiés de systèmes de traduction **hybrides**.

De par leurs natures, ces deux approches s'opposent l'une à l'autre et défendent chacune leurs atouts. Il est en effet souvent question de savoir quelle approche permet d'obtenir les meilleurs résultats. Il n'y a évidemment pas de réponse absolue à cette question, chacune de ses approches ayant ses forces et ces faiblesses selon le contexte dans lequel on va vouloir les utiliser.

Dans le cas de l'approche experte de la TA, et plus particulièrement pour l'approche à base de règles, il est plus aisé de contrôler la qualité en sortie du système de traduction de par l'utilisation de règles linguistiques sophistiquées. Cette qualité en sera d'autant plus adaptée pour les traductions généralistes, avec la connaissance notamment de règles grammaticales propres à la langue cible. Pour un système probabiliste, la qualité sera elle prévisible si on s'est assuré que les données d'entraînement sont de bonne qualité. De fait, ces systèmes étant dépendants de ces données, ils seront *a contrario* davantage efficaces sur des domaines spécifiques, et non généralistes. Mais à leur avantage, ils seront réputés plus fluides alors que les systèmes à base de règles seront réputés comme étant moins « humains » dans la construction de phrase.

On constate avec cette rapide comparaison que ces deux approches, la traduction à base de règles d'une part et la traduction statistique d'autre part, au-delà de s'opposer laisse à penser qu'elles peuvent en étant associées, se compléter. L'élaboration de systèmes de traduction hybrides peut donc être perçue comme étant l'union de ces deux mondes. Dans la littérature, plusieurs exemples de ces systèmes hybrides existent : un système fondé sur les exemples [Langlais 2006] qui, pour améliorer la qualité des traductions en sortie, utilise des règles linguistiques. Autre système hybride, le système MATREX [Tinsley 2008] qui associe cette fois l'approche statistique à l'approche fondée sur les exemples. [Simard 2007] propose lui un système hybride composé d'un système de TAS entraîné sur un corpus parallèle. La particularité de ce corpus bilingue, c'est qu'il est à la fois composé d'hypothèses de traduction issues d'un système à base de règles, mais également de références humaines. Ce système TAS est alors appliqué en sortie du premier système de TA à base de règles linguistiques pour venir corriger d'éventuelles erreurs initiales de traduction. Un système similaire fut développé par la société SYSTRAN [Dugast 2007] à partir de son propre système à base de règles. En 2009, la

mise au point par [Schwenk 2009a] d'un système hybride avec le traducteur développé par la société SYSTRAN a été classé premier sur la tâche de traduction de l'anglais vers le français lors de la campagne d'évaluation « Workshop on Statistical Machine Translation » (WMT).

## 1.6 Évaluation de la traduction automatique

Il est une tâche complexe que celle de réussir à générer automatiquement une traduction. Il en est une autre presque tout aussi difficile qui consiste à évaluer la qualité de cette dite traduction. Il existe d'ailleurs une manière simple de vérifier ces dires : demander à un groupe de traducteurs humains de vous traduire un texte, vous n'obtiendrez probablement pas deux fois la même traduction. Partant de ce constat, comment évaluer la sortie d'un système de TA ?

Rappelons dans un premier temps les deux critères majeurs [White 1994] permettant d'évaluer la qualité d'une traduction :

- *vraisemblance grammaticale* : la traduction est-elle lisible, fluide, intelligible dans la langue cible ?
- *vraisemblance sémantique* : le sens de la phrase à traduire est-il conservé dans sa traduction équivalente ?

À noter que l'intérêt d'évaluer la sortie d'un système de TA ne s'arrête pas à la simple mesure de sa qualité. Cela permet également d'évaluer l'évolution d'un système au fur et à mesure des modifications qu'on lui apporte, et bien sûr, de le comparer à d'autres systèmes lors de campagnes d'évaluation comme illustré précédemment.

### 1.6.1 Évaluation manuelle des traductions

Aux prémices de la TA, la qualité des traductions était évaluée exclusivement par des experts humains. Des traducteurs bilingues notamment qui ont compétence pour évaluer les deux critères que nous venons de citer. Chaque traducteur se voyait alors confier un ensemble de paires de traductions composées d'une phrase en langue source et d'une traduction équivalente en langue cible, qu'ils devaient annoter. Toutefois, l'évaluation manuelle bien que possiblement encadrée avec des consignes communes aux annotateurs est difficilement reproductible. De plus, les variances inter annotateurs rendaient difficile une classification objective des systèmes sur un même jeu de test, là encore lors de campagnes d'évaluation par exemple.

Seulement, le volume de traduction toujours plus important et le coût considérable que représente l'utilisation de traducteurs humains (sans compter le temps que cela nécessite) ont rapidement fait surgir le besoin de recourir à des procédures d'évaluation automatisées.

### 1.6.2 Évaluation automatique des traductions

Bien que l'évaluation humaine garantisse une évaluation qualitative respectant les deux critères citer ci-avant, le recours à une métrique automatique garantie de toujours obtenir le même résultat dans un même cas de figure, sans surcoût, et ce pour une rapidité de traitement de la tâche sans équivalence. On constate alors tout l'intérêt que peut avoir une telle approche lorsque l'on peut avoir besoin d'évaluer les sorties de son système plusieurs fois en une seule journée. Et c'est aussi la raison pour laquelle ces métriques automatiques sont utilisées pour l'optimisation des systèmes de TAS (comme évoqué à la section 1.4.2).

De nombreux travaux sur l'élaboration de métriques automatiques ont été menés depuis les années 2000 sans qu'encore aujourd'hui un compromis ait été trouvé pour désigner une métrique comme étant la métrique de référence dans le domaine. C'est pourquoi nous allons présenter un certain nombre d'entre elles, en précisant toutefois qu'il ne s'agit pas là d'une liste exhaustive, mais seulement les plus couramment utilisées que l'on pourra retrouver par la suite dans ce manuscrit.

#### F-Mesure

Le recourt à l'évaluation automatique nécessite de disposer d'un corpus de test pour lequel il existe une ou plusieurs traductions humaines dites **de références**. On parlera alors d'**hypothèse de traduction** lorsqu'on évoquera la traduction générée automatiquement en sortie du système.

L'idée principale est simple : une hypothèse de traduction est considérée comme étant correcte si elle ressemble à une traduction de référence pour la même phrase source, en langue cible.De fait, on va comparer mot à mot la traduction et sa référence afin de déterminer le ratio de mots corrects présents dans la traduction par rapport à sa longueur, c'est ce qu'on appelle la **précision** :

$$precision = \frac{nombre\ de\ mots\ corrects}{longueur\ hypothese\ de\ traduction} \tag{1.22}$$

Bien que ce ratio puisse être important, il ne garantit en rien la bonne qualité de la traduction générée automatiquement en comparaison de la traduction de référence, c'est pourquoi il peut être intéressant de comparer le ratio de mots corrects dans la première par rapport aux mots que le système aurait dû générer vis-à-vis de la référence, c'est ce qu'on nomme le **rappel** :

$$rappel = \frac{nombre\ de\ mots\ corrects}{longueur\ reference} \tag{1.23}$$

Ces deux métriques prises indépendamment peuvent ne pas être de grandes sources d'informations. En effet, si notre système génère une très courte traduction, mais dont tous les mots sont présents dans la référence, notre précision sera excellente sans que la traduction ne soit toutefois correcte. Inversement, si notre système génère cette fois une très longue traduction, nous aurions alors une plus grande couverture des mots vis-à-vis de la référence, là encore, sans que cette traduction ne soit pour autant valide.

En TA, nous sommes intéressés à la fois par la précision et le rappel, et un moyen pour concilier les deux est de mesurer ce qu'on appelle la **F-mesure** :

$$\begin{aligned} F-mesure &= \frac{precision \times rappel}{(precision + rappel)/2} \\ &= \frac{nombre\ de\ mots\ corrects}{(longueur\ hypothese + longueur\ reference)/2} \end{aligned} \tag{1.24}$$

### Word Error Rate

Le *Word Error Rate* (WER) (littéralement « *Taux d'Erreur de Mot* ») s'appuie sur la Distance de Levenshtein [Levenshtein 1966] qui permet de calculer la **distance d'édition** entre deux phrases. La distance d'édition en TA consiste à calculer le nombre minimal de mot à *insérer*, *supprimer* ou *substituer* pour transformer l'hypothèse de traduction dans le but d'atteindre la traduction de référence. Chacune de ces modifications étant pondérées individuellement, comme suit :

$$WER = \frac{nb\ insertion(s) + nb\ suppression(s) + nb\ substitution(s)}{longueur\ reference} \tag{1.25}$$

Première métrique utilisée en TA, le WER est emprunté à la RAP dont la problématique de réordonnancement n'est pas considérée, car absente. De fait, lorsqu'un mot de l'hypothèse de traduction est présent, mais pas à la bonne position dans celle-ci, vis-à-vis de la traduction de référence, le WER considérera qu'il s'agit à la fois d'une suppression du mot (à la mauvaise position) et d'une insertion de ce mot (cette fois à la position qui doit être la sienne) dans la référence. Le mauvais ordonnancement en TA est alors doublement sanctionné (coût suppression + coût insertion) et le score de la paire de phrases ainsi évaluée est biaisé.

### Position-independent Word Error Rate

Le *Position-independent Word Error Rate* (PER) de [Tillmann 1997] se propose de ne justement pas tenir compte de la position des mots dans la phrase, palliant ainsi la problématique évoquée précédemment pour le WER. La formule du PER est la suivante :

$$PER = \frac{d_{per}(reference, hypothse\ de\ traduction)}{longueur\ reference} \tag{1.26}$$

où $d_{per}(ref., hyp.)$ retourne la différence entre les occurrences des mots apparaissant dans l'hypothèse de traduction par rapport à la traduction de référence.

### Translation Error Rate

Le *Translation Error Rate* ou *Translation Edit Rate* (TER) (littéralement « *Taux d'Erreur de Traduction* ») proposé par [Snover 2006] est une extension du WER vu précédemment. Le TER reprend en effet la notion de distance d'édition du WER avec les types d'éditions qu'on lui connaît, tout en y ajoutant un quatrième : le mouvement (*shift* en anglais) correspondant au repositionnement d'un bloc de mots dans la phrase. Le TER se définit alors de la manière suivante :

$$TER = \frac{nb\ insertion(s) + nb\ suppression(s) + nb\ substitution(s) + nb\ mouvement(s)}{longueur\ reference} \tag{1.27}$$

Quand le WER sanctionne deux fois le mauvais ordonnancement d'un mot dans l'hypothèse, le TER lui associe simplement un coût de déplacement. Le meilleur déplacement étant déterminé à l'aide d'heuristiques.

Par ailleurs, le TER offre la possibilité d'optimiser les poids associés aux coûts d'édition. Ainsi, le score associé à la paire de phrases ainsi évaluée est plus affiné et est davantage représentatif de la distance entre l'hypothèse de traduction et la traduction de référence.

### TER-Plus

Le TER-Plus (ou TERp) est une extension du Translation Error Rate proposée par [Snover 2009]. Cette version améliorée du TER utilise notamment les stemmes, les synonymes et applique une contrainte sur la distance d'un déplacement.

### BLEU

Le score BLEU pour *BiLingual Evaluation Under study* proposé par [Papineni 2002] est la métrique la plus utilisée de nos jours pour évaluer et évoquer la qualité des systèmes de TA.

Basé sur la Précision introduite à la section 1.6.2, le score BLEU calcule la similitude entre l'hypothèse de traduction et un ensemble de traductions de référence au niveau des $n$-grammes. Mathématiquement, la formule du score BLEU est la suivante :

$$\text{BLEU} = BP \times \exp(\sum_{i=1}^{N} \lambda_i \log precision_i) \tag{1.28}$$

avec $\sum_{i=1}^{N} \lambda_i \log precision_i$ la moyenne géométrique des précisions pour les $n$-grammes d'ordre 1 jusqu'à $N$ et des poids $\lambda_i$ positifs. Quant à $BP$, c'est une pénalité calculée pour défavoriser les traductions automatiques courtes par rapport aux références, puisque nous l'avons vu en section 1.6.2, la précision ne pénalise pas les phrases courtes. Cette pénalité dite **pénalité de brièveté** est définie par :

$$BP = \begin{cases} 1 \; si \; c \geq r \\ e^{(1-\frac{r}{c})} \; si \; c \leq r \end{cases} \tag{1.29}$$

avec $c$ la longueur en nombre de mots de l'hypothèse de traduction et $r$ le nombre de mots de la traduction de référence la plus proche (en nombre de mots) de $c$.

### NIST

Variante du score BLEU pour laquelle un poids identique est attribué à chaque $n$-gramme, le score NIST lui va s'intéresser au $n$-gramme considéré. Si celui-ci est rare, il se verra attribuer un poids plus important [Doddington 2002].

### METEOR

METEOR pour *Metric for Evaluation of Translation with Explicit ORdering* de [Denkowski 2011], calcule la moyenne harmonique pondérée de la Précision et du Rappel des unigrammes, d'après [Banerjee 2005, Lavie 2007]. Il a été mis au point pour résoudre certains problèmes connus du score BLEU, notamment au niveau de la corrélation avec le jugement humain. Alors que ce dernier cherchera une corrélation au niveau du corpus, METEOR s'intéressera à une corrélation au niveau de la phrase ou du segment considéré.

Concrètement, le calcul du score METEOR se déroule en deux temps : trouver le meilleur alignement entre l'hypothèse de traduction et sa traduction de référence, et utiliser cet alignement pour établir un score. L'alignement se fait successivement sur les unigrammes de même

forme orthographique, puis les stemmes et les synonymes. À partir de cet alignement, le score est déterminé comme suit :

$$METEOR = (1 - Pen)\frac{P \times R}{\alpha P + (1 - \alpha)R} \tag{1.30}$$

où $P$ et $R$ sont respectivement la Précision et le Rappel, $\alpha$ un facteur de pondération pour favoriser $P$ ou $R$ dans le calcul de la moyenne, et $Pen$ un coefficient de pénalité pour les traductions n'ayant pas de correspondance d'ordre supérieur à l'unigramme.

**Discussion**

Nous venons de voir les principales métriques utilisées dans le domaine de la TA. Ces métriques sont utilisées aussi bien pour évaluer *a posteriori* les systèmes individuellement dans le cadre de leur développement (voir section 1.4.2), mais également en confrontations directes lors de campagnes d'évaluation. Répondant à la question suivante : « *mon système permet-il d'obtenir des traductions de bonne qualité ?* », elles permettent de mettre en exergue la proportion d'erreurs des systèmes. La réduction de cette proportion bénéficiant aujourd'hui d'un effort de recherche considérable de la part de la communauté scientifique.

La motivation première pour le développement de telles métriques fut de remplacer l'évaluation humaine trop coûteuse, lente et considérée comme peu fiable. Ce dernier aspect est un problème en soit puisqu'il apparaît difficile d'élaborer une métrique visant à reproduire un jugement humain considéré comme référence, et dont on sait qu'il n'est lui-même pas fiable. En effet, pour une même traduction, il peut arriver que deux traducteurs aient un jugement opposé [Turian 2003, Callison-Burch 2008]. Bien que des études aient démontré successivement que l'une ou l'autre de ces métriques était la plus adaptée, chaque fois dans un contexte particulier, la métrique de référence la plus utilisée dans le domaine de la TA reste de nos jours le score BLEU. Dans le contexte qui nous intéresse ici, nous avons donc choisi de l'utiliser comme métrique principale pour l'évaluation de nos systèmes. En revanche, lorsque nous évoquerons la distance d'édition entre deux phrases, nous utiliserons le TER car plus qu'une similarité entre celles-ci, c'est davantage la différence de la première vis-à-vis de la seconde qui nous intéresse.

Quoi qu'il en soit, la TA reste encore aujourd'hui un domaine de recherche des plus complexes de la linguistique et en Intelligence Artificielle. De fait, il n'existe pas de système de TA parfait et le recours à cette technologie ne permet pas la publication de données traduites sans que celles-ci n'aient été préalablement révisées. Ce concept de révision appelé aussi « **post-édition** » représente une part importante des travaux rapportés dans ce manuscrit, et c'est pourquoi nous allons l'aborder dans la section suivante.

## 1.7 Post-Édition

Dans cette dernière section de notre chapitre, nous allons aborder le concept de post-édition, au travers notamment d'un point de vue industriel. En montrant la motivation d'un tel concept et son application, nous voulons donner une idée précise du point de départ de notre réflexion pour nos travaux de recherche. Cela doit permettre de mieux comprendre en quoi les contributions de ces derniers sont novatrices et intéressantes pour la communauté scientifique, ainsi que pour les acteurs industriels sujets à son utilisation.

### 1.7.1 Motivation et principe

Avec l'expansion du marché de la traduction grâce notamment à l'approche empirique basée sur l'exploitation probabiliste de corpus bilingues, la TA est devenue un outil permettant l'accès aux individus à des ressources en langues étrangères, mais aussi et surtout un outil de productivité pour des sociétés qui possèdent notamment un service de localisation. La philosophie inhérente à la TA est alors différente de celle que l'on a pu avoir jusqu'à maintenant. Dans ce contexte de révision de la traduction, il ne va pas être question de savoir si la traduction est de la meilleure qualité qu'il soit, mais bien de savoir si elle est de qualité suffisante pour répondre à la question suivante : « *ma traduction est-elle publiable en l'état ?* »

Prenons l'exemple d'une société ayant des contraintes de temps et de coûts, et qui a besoin de localiser sa documentation pour laquelle elle a la possibilité de faire appel à des traducteurs humains. Idéalement, ces traducteurs doivent être bilingues afin de saisir parfaitement la sémantique du texte en langue source, pour être capables ensuite de la retranscrire en langue cible. Bien évidemment, de telles compétences ont un coût aussi bien financier que temporel. Ce coût se trouve en plus être multiplié par autant de traducteurs qu'il y aura de langues cibles dans lesquelles la société souhaite faire traduire sa documentation. On comprend alors très vite l'intérêt pour cette société d'utiliser la TA dans le cadre de ses activités de localisation.

À l'aide d'un système de TA adapté, la documentation est traduite automatiquement pour fournir une première version en langue cible. Un traducteur humain, appelé aussi « **post-éditeur** », va alors procèder à la révision de la version traduite et lui apporter les corrections nécessaires le cas échéant, dans le but de la rendre publiable. Il apparaît alors très important que la qualité de traduction soit suffisante afin de ne pas entraîner un nombre trop important de modifications de la part du post-éditeur. Ce qu'on assimilera alors à un **effort de post-édition** de la part du traducteur. Si la qualité du système n'est pas suffisante, l'effort de post-édition pourrait devenir plus important que le fait d'effectuer la traduction depuis le début (dite traduction **from-scratch**). Ce qui serait bien évidemment contre-productif.

### 1.7.2 Évaluer l'effort de post-édition

La mesure de l'effort de post-édition est importante d'un point de vue commercial, car elle établit la productivité des post-éditeurs et par la suite le potentiel de réduction des coûts supplémentaires. Les plupart des critères d'évaluation de cet effort reposent la mesure du temps de post-édition [Specia 2011] et la comparaison avec la traduction humaine [Plitt 2010]. Comme nous venons de le voir, si le post-éditeur passe plus de temps à corriger une hypothèse de traduction qu'à la réaliser entièrement, l'utilisation de la TA perd tout son sens. De même, si ce dernier réalise des modifications additionnelles qui n'ont pas de valeur ajoutée pour la traduction, et qu'elles n'influencent en rien son sens, on parlera de **corrections stylistiques**.

Afin d'encadrer au mieux cette pratique et de la rendre plus productive tout en limitant son coût, il est courant de définir ce qu'on appelle des **consignes de post-édition**, qui sont soumises aux traducteurs.

#### Consignes de post-édition

Il existe deux types (de consignes) de post-édition :
- **Post-édition légère** (*light post-editing* en anglais) pour obtenir une qualité « acceptable » ;
- **Post-édition complète** (*full post-editing* en anglais) pour une qualité comparable ou égale à une traduction humaine.

Ci-après un exemple de consignes de post-édition, telles qu'elles sont définies par TAUS [7], un groupe de réflexion composé d'experts pour l'industrie de la traduction :
- **Post-édition légère**
  - Viser une traduction correcte au niveau sémantique ;
  - Vérifier qu'aucune information n'a été accidentellement ajoutée ou oubliée ;
  - Exploiter au maximum le résultat brut de la traduction automatique ;
  - Appliquer les règles d'orthographe fondamentales ;
  - Inutile d'effectuer des corrections d'ordre uniquement stylistique.

- **Post-édition complète**
  - Viser une traduction correcte au niveau grammatical, syntaxique ;
  - Vérifier que la terminologie importante est correctement traduite et que les termes non traduits font partie de la liste des termes à ne pas traduire ;
  - Appliquer les règles fondamentales d'orthographe, de ponctuation et de coupure des mots.

7. Translation Automation User Society - www.taus.net

D'autres approches s'intéressent davantage aux « données d'activité de l'utilisateur » avec par exemple l'activité au clavier [Barrett 2001] (*keystroke* en anglais) ou encore la détection des mouvements oculaires [Doherty 2010] (*eye-tracking* en anglais).

Implicitement, l'estimation de l'effort de post-édition est le pilote pour établir de meilleurs indicateurs d'évaluation de la qualité, comme le HTER de [Snover 2006] par exemple, que nous allons voir maintenant.

#### Human-targeted Translation Error Rate

Le *Human-targeted Translation Error Rate* (HTER) est une évolution du TER développée par [Snover 2006] lui-même qui, selon lui, ne reflète pas entièrement l'acceptabilité d'une hypothèse de traduction. Souhaitant obtenir une mesure plus précise de la qualité de traduction, il met au point le HTER avec cette particularité que cette nouvelle métrique associe des post-éditeurs humains dans le processus (« *human-in-the-loop evaluation* »).

Le HTER consiste à demander à des traducteurs humains parlant couramment la langue cible, de générer une traduction de référence (dite **référence ciblée**), la plus proche possible de l'hypothèse de traduction et devant partager la même sémantique que des traductions de références présélectionnées au départ (dites **références non ciblées**). Les annotateurs ont le choix de repartir de l'hypothèse de traduction ou d'une des traductions de références présélectionnées. Cette référence ciblée est alors utilisée comme seule traduction de référence humaine pour évaluer le TER sur l'hypothèse de traduction.

Dans ses résultats expérimentaux, [Snover 2006] annonce une réduction significative de 33% du taux d'édition comparé avec le TER lorsqu'il utilise quatre références non ciblées. Bien que l'effort de post-édition semble donc réduit en nombre d'éditions, il semble cependant pénalisé par le coût en temps d'exécution requis par le HTER qui est évalué entre 3 à 7 minutes.

### 1.7.3 Comment limiter ou réduire cet effort ?

Au delà de l'évaluation même de l'effort de post-édition, se pose le problème général de comment celui-ci peut être réduit ? De multiples approches existent et quelques-unes répondant implicitement à cette question ont été énoncées précédemment lorsque nous avons abordé l'approche hybride pour la TA.

Pour rappel, [Simard 2007] et [Dugast 2007] proposent une approche où un système de TAS, appliqué en sortie d'un système à base de règles, est entraîné sur un corpus parallèle constitué de la sortie d'un système de TA et de références humaines. Ce procédé est appelé Post-Édition Statistique (PES), (*Statistical Post-Editing* en anglais). [Schwenk 2009b] reproduit ceci avec

un système SPE entraîné sur un très large corpus, transformant la traduction initiale en simple prétraitement. Ils montrent ainsi comment le système est capable de « corriger la sortie de traduction », le système de TAS étant transformé plus en correcteur statistique d'erreurs de traduction, qu'en réel traducteur. Celui-ci bénéficiant de plus hautes similitudes entre le texte pré-traduit et une référence, en comparaison d'une phrase source et de sa référence en deux langues différentes.

La Post-édition est considérée par plusieurs études comme une classification des erreurs liées à la TA afin de mieux rationaliser l'effort de post-édition, à l'instar de [Martinez 2003] :

> " familiarity with the pattern of errors produced by a particular MT system is an important factor in reducing post-editing time "
>
> *« la familiarité avec les motifs d'erreurs produites par un système de TA particulier est un facteur important dans la réduction du temps de post-édition »*

[Guzmán 2007] décrit lui une configuration où un ensemble de règles de post-édition est appliqué sur l'hypothèse de traduction afin de la « lisser » de façon à obtenir un système hautement personnalisé. Cependant, si la qualité finale de la traduction est supérieure, point commun à toutes ces approches, le système lui n'apprend pas de la post-édition. Nous entendons par là que dans la même situation, le système de TA refera les mêmes erreurs et qu'il faudra donc à nouveau les corriger. Or, il est possible d'apprendre du processus de post-édition et de faire en sorte que le système de TA puisse apprendre de ses erreurs, réduisant par la même, *a posteriori*, le coût de post-édition.

Prenons l'exemple d'un système hybride avec un système à base de règles linguistiques associé à un système de TAS pour une post-édition statistique, comme nous avons pu en citer précédemment. Dans cette configuration, le système de TAS peut-être utilisé pour renforcer les ressources linguistiques du système à base de règles comme suggéré par [Dugast 2007] en réalisant notamment une extraction automatique d'une terminologie bilingue [Daille 1994, Déjean 2002]. Cette terminologie pouvant alors être exploitée via un dictionnaire bilingue particulier.

Dans le cadre d'un système de TAS, il est tout à fait envisageables d'utiliser les données post-éditées comme données d'apprentissage de sorte que le système intègre de nouvelles connaissances. Cependant, cette approche soulève deux problématiques : doit-on se contenter d'exploiter les données post-éditées dans leur ensemble, au risque de bruiter les connaissances du système avec, par exemple, des corrections styliques du post-éditeur, voire des corrections erronées (involontairement ou non) ? Pour répondre à cette question, nous nous sommes intéressé à la nature même de la post-édition et avons proposé conséquemment une méthodologie

d'analyse des données post-éditées. Celle-ci s'appuie sur une nouvelle notion destinée à permettre de mieux identifier l'information primaire en sortie du processus de post-édition.

Autre problématique, celle du coût computationnel que représente l'apprentissage d'un nouveau système de TAS pour qu'il puisse intégrer et bénéficier de nouvelles ressources, que ce soit des données post-éditées ou de nouveaux corpus de données textuelles. Pour pallier cet aspect de la TAS dans un contexte d'apprentissage lié à un processus de post-édition, nous avons développé une procédure d'adaptation incrémentale d'un système de TAS permettant de bénéficier plus rapidement des retours utilisateurs, et ainsi à nouveau limiter l'effort de post-édition.

Les contributions de cette thèse en réponses à ces problématiques sont décrites en sections 3 et 4 de ce manuscrit. Mais avant de nous y intéresser, nous allons aborder en seconde partie le cadre applicatif de nos travaux de recherche.

# Deuxième partie

# Cadre applicatif

# Chapitre 2

# COSMAT : Traduction Automatique de contenus scientifiques

## Sommaire


**2.1 Enjeux et problématiques . . . . . 48**

**2.2 Approches scientifiques et techniques . . . . . 50**

2.2.1 Extraction de contenu structuré . . . . . 50

**2.3 Intégration de connaissances linguistiques . . . . . 52**

**2.4 Adaptation en domaine d'un système de TAS . . . . . 54**

2.4.1 Extraction de données bilingues du domaine . . . . . 54

2.4.2 Données d'apprentissage hors-domaine . . . . . 55

2.4.3 Sélection de données hors-domaine . . . . . 56

**2.5 Interface de post-édition . . . . . 59**

2.5.1 Campagnes d'évaluation . . . . . 60

2.5.1.1 Pertinence utilisateur . . . . . 61

**2.6 Intégration dans HAL . . . . . 63**

2.6.1 Architecture globale du service COSMAT . . . . . 64

Dans ce second chapitre, nous avons souhaité présenter dans son ensemble le projet de recherche ayant servi de cadre applicatif à nos travaux. Il s'agit du projet COSMAT.

COSMAT [8] est un projet de recherche industriel coordonné par Holger Schwenk (LIUM) de l'Université du Maine, et dont les partenaires sont la société SYSTRAN S.A. représentée par Jean Senellart, et l'INRIA [9] représentée par Laurent Romary. Lancé en octobre 2009 pour une durée de 42 mois et financé par l'Agence Nationale de la Recherche (ANR) (ANR-09-CORD-004), ce projet a pour objectif de proposer librement un service collaboratif de traduction automatique de contenus scientifiques. À destination de la communauté scientifique, il intègre un aspect collaboratif avec l'utilisateur autour d'une interface riche qui sera présentée dans ce chapitre. À terme, ce service devrait être mis en ligne via la plateforme HAL [10], une archive ouverte pluridisciplinaire en libre accès qui, créée en 2006, est destinée au dépôt et à la diffusion d'articles de niveau recherche de toute la communauté scientifique française.

Dans la suite de ce chapître, nous aborderons aussi bien les problématiques du projet dans lesquelles s'incrivent les contributions de cette thèse, que celles pour lesquelles nous avons peu ou pas été impliqué. Nous avons fait ce choix dans un souci de clareté.

## 2.1 Enjeux et problématiques

En raison de la mondialisation de la recherche, l'anglais est aujourd'hui la langue universelle pour la communication scientifique. En France, la réglementation impose toutefois l'utilisation de la langue française dans la rédaction de rapports, de thèses universitaires (comme celle-ci), de manuscrits, etc. Le français étant la langue d'enseignement officielle du pays. Cette situation oblige les chercheurs à traduire les publications de leurs travaux. En outre, la publication en langue étrangère est souvent un obstacle pour les étudiants et le grand public lorsqu'il de rechercher et de comprendre des articles publiés. C'est donc une réalité intéressante que celle d'une communauté de spécialistes qui doit régulièrement produire des traductions sur des domaines souvent très limités et très pointus. Parallèlement, une autre communauté d'utilisateurs a besoin de traductions pour ces mêmes documents. Sans outil approprié, l'expertise et le temps consacrés à l'activité de traduction par la première communauté sont peu ou pas exploités par la seconde.

8. www.cosmat.fr
9. www.inria.fr
10. hal.archives-ouvertes.fr

Inversement, la problématique peut être également vue sous un autre angle : des spécialistes étrangers n'envisagent peut-être tout simplement pas de consulter des articles français parce qu'ils manquent d'expertise de la langue française. De fait, il est possible d'énumérer trois types d'acteurs :

1. **Les scientifiques français qui écrivent dans leur langue maternelle, et qui possèdent un niveau suffisant en anglais** – *Ces scientifiques passent une partie de leur temps de recherche à traduire leurs propres publications vers l'anglais. Leurs effrots de traduction pourraient bénéficier à d'autres acteurs grâce à un outil de traduction approprié et librement accessible.*
2. **Un public francophone à la recherche de publications disponibles seulement en anglais** – *Ces acteurs ont souvent recours à l'utilisation d'outils de TA librement accessibles, mais qui ne sont pas adaptés à cette tâche. Le risque est alors que la qualité de traduction soit trop mauvaise pour permettre une bonne compréhension, voire une compréhension erronée du contenu scientifique.*
3. **Des scientifiques internationaux qui n'envisagent même pas de chercher des publications françaises, parce qu'elles ne sont pas disponibles dans leurs langues maternelles** – *Ils pourraient alors tout à fait utiliser un outil de TA collaboratif adapté aux domaines scientifiques. Ces acteurs étant par ailleurs experts dans leurs domaines respectifs, et possibles locuteurs anglophones, ils seraient à même d'évaluer la traduction ainsi générée et de la corriger si nécessaire.*

Jusqu'au lancement du projet COSMAT, ces problématiques restaient sans réponse, aucun projet répondant à ces enjeux n'existait, et aucune ressource n'était alors disponible pour permettre l'élaboration d'un système correspondant à cette tâche. Partant de ce constat, le défi pour le consortium fut de réfléchir conjointement au développement de solutions scientifiques et techniques pouvant permettre de proposer un service de traduction efficace. Une autre finalité du projet étant qu'à terme, ce service puisse être directement intégré dans HAL. Au départ, les partenaires avaient tous une expérience dans les domaines respectifs :

- Développement de systèmes de TA statistiques performants pour le LIUM ;
- Développement de systèmes de TA professionnels et réalisation de chaînes de traitements complexes pour la société SYSTRAN ;
- Développement et maintenance opérationnelle de HAL pour l'INRIA.

Le développement du projet COSMAT s'est donc déroulé autours des axes suivants : (i) élaboration de nouvelles techniques avancées pour les systèmes de TAS (LIUM), (ii) intégration de connaissances linguistiques pour le traitement d'entités particulières (SYSTRAN), (iii) développement d'une interface riche dédiée et intégration dans HAL (SYSTRAN+INRIA).

## 2.2 Approches scientifiques et techniques

Le partie traduction du projet COSMAT repose sur un moteur hybride pour des traductions de l'anglais vers le français (En→Fr) et du français vers l'anglais (Fr→En). Ce moteur est composé de modules linguistiques développés par la société SYSTRAN, associés à un traducteur statistique basé sur les séquences de mots développé par le LIUM.

Nous l'avons vu au chapitre précédent (voir section 1.4.2), le développement d'un système TAS nécessite l'utilisation d'un nombre conséquent de données bilingues et monolingues. Un large volume de données monolingues et bilingues est disponible pour entraîner un système entre le français et l'anglais, mais pas dans le domaine scientifique. Afin de renforcer les performances de notre système de TA sur cette tâche, nous avons procédé à l'extraction de données du domaine à partir de l'archive HAL. Toutefois, les documents stockés sur HAL sont pour la plupart au format PDF, ce qui n'est pas exploitable en l'état par le système de traduction. Une solution technique a donc dû être apportée pour extraire de ces documents PDF, le contenu dont nous avions besoin.

### 2.2.1 Extraction de contenu structuré

Bien que pour la rédaction des articles, les formats utilisés soient principalement un mélange de fichiers Microsoft Word et de documents LaTeX utilisant une variété de styles, le PDF reste le format de publication le plus couramment utilisé dans le monde par la communauté scientifique. Nos données provenant de la plateforme HAL qui héberge ces publications, le PDF constitue par conséquent le format d'entrée de notre système.

Jusqu'alors, le scénario classique pour la dépose d'une nouvelle ressource par un utilisateur sur HAL nécessitait de ne renseigner que certaines informations telles que le titre. Certaines informations comme la liste de ou des auteurs étaient automatiquement extraites. Le but de COSMAT est donc d'être capable d'extraire du contenu textuel à traduire, à partir d'un article, d'une thèse ou autre (une habilitation à diriger des recherches [11] par exemple). en conservant autant que possible les métas-informations (agencement, références, auteurs, etc.), pour pouvoir, à terme, générer un résultat final le plus proche possible du format source. Pour ce faire, l'INRIA a su faire évoluer de façon importante son outil d'extraction de contenu structuré : « GROBID ».

GROBID, pour « *GeneRation Of BIbliographic Data* » [Lopez 2009], est un outil open-source qui permet de décomposer et de convertir les articles scientifiques au format PDF, en un document conforme aux directives de la TEI (*Text Encoding Initiative*). Cette conversion est nécessaire pour que le contenu de l'article puisse être ensuite traité par le serveur de traduction

11. http ://fr.wikipedia.org/wiki/Habilitation_universitaire

dans un format qu'il supporte. On parlera alors de « format TEI » dont un exemple est donné ci-après en figure 2.1 :

```
<?xml version="1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mml="http://www.w3.org/1998/Math/MathM
  <teiHeader xml:lang="en">
    <fileDesc>
      <titleStmt>
        <title level="a" type="main">MAGNETO-PHOTOLUMINESCENCE AS A PROBE OF PHASES IN QUANTUM HALL MULTI-LAYERS</title>
      </titleStmt>
      <sourceDesc>
        <biblStruct>
          <analytic>
            <author>
              <persName>
                <forename type="first">Yu</forename>
                <forename type="middle">A</forename>
                <surname>Pusep</surname>
              </persName>
              <affiliation>
                <orgName type="department">Instituto de Fisica de São Carlos</orgName>
                <orgName type="institution">Universidade de São Paulo</orgName>
                <address>
                  <postCode>13560-, 970</postCode>
                  <settlement>São Carlos</settlement>
                  <region>SP</region>
                  <country>Brazil</country>
                </address>
              </affiliation>
              <email>yuri.pusep@gmail.com</email>
            </author>
```

FIGURE 2.1 – Exemple de description documentaire au format TEI.

### Valorisation

Développé par l'INRIA en collaboration avec le Centre pour la Communication Scientifique Directe (CCSD) (Unité Mixte de Service (UMS) du CNRS qui est opérateur de la plate-forme HAL), GROBID fut libéré durant le projet COSMAT et est désormais distribué sous licence Apache 2.0 [12] via la plateforme GITHUB [13].

12. http ://www.apache.org/licenses/LICENSE-2.0
13. https ://github.com/grobid/grobid

## 2.3 Intégration de connaissances linguistiques

Outre la problématique d'accessibilité aux ressources nécessaires à l'élaboration d'un système de TA, la nature même des données à traiter dans le cadre de COSMAT est un défi technique. L'utilisation de GROBID permet, à partir d'un article ou d'une thèse au format PDF, d'en extraire les méta-données sur son organisation, son agencement, ainsi que son contenu textuel comme on vient de le voir.

Cependant, la littérature scientifique a la particularité d'être abondante en formules, tableaux, références, annotations, etc. Ce qui complique considérablement la tâche du système de traduction. La principale raison à cette difficulté réside dans le fait que le système, lorsqu'il va analyser une phrase source contenant une formule par exemple, ne la verra pas en tant que telle. Alors qu'un traducteur humain saura parfaitement identifier une formule et la traiter comme une entité à part entière, le système de TA lui, n'y verra qu'une suite de symbôles qui vont potentiellement le perturber.

Afin de pallier à ces problèmes, il était essentiel de pouvoir identifier et gérer ces entités en amont du processus de traduction, et d'être ensuite capable, en aval cette fois, de réintroduire ces entités à leurs places respectives lors de la génération de l'hypothèse de traduction au bout de la chaine de traitement. Ce travail a été réalisé par la société SYSTRAN en intégrant des connaissances linguistiques dans le système de TA. Les efforts ont notamment porté sur les axes suivants :

1. Le développement d'un outil de détection et de reconnaissance d'entités nommées fréquemment observées dans des documents scientifiques ;
2. L'enrichissement des modèles de reconnaissance d'entités scientifiques, en particulier en l'étendant aux entités telles que les formules utilisées dans le domaine de la Physique ;
3. L'extraction des données de HAL pour en extraire de la terminologie scientifique ;
4. L'extraction d'une terminologie spécialisée à partir des corpus monolingues, et de dictionnaires bilingues à partir des données bilingues. Ces dictionnaires viennent en complément des ressources extraites à partir des glossaires spécialisés de SYSTRAN et de glossaires « TermSciences » obtenus par l'INRIA ;
5. L'adaptation de l'analyse syntaxique par apprentissage d'arbres de décision ;
6. L'entraînement de modèles de PES sur les données bilingues et monolingues collectées.

Une illustration de l'impact et de l'importance de ces analyses linguistiques en amont est donnée en figure 2.2. On peut constater que sans la reconnaissance d'entités scientifiques telles

| | | |
|---|---|---|
| Lemma 6 For any sub-σ-algebra B of B, there exists a sub-σ-algebra B determined by a factor and such that B = B modulo $\mu$. | Le lemme 6 pour n'importe quelle sous-σ-algèbre B de B, existe là une sous-σ-algèbre B déterminée par un facteur et tels que $\mu$ de modulo de B = B. | Le lemme 6 pour n'importe quelle sous-σ-algèbre B de B, existe là une sous-σ-algèbre B déterminée par un facteur et tels que $\mu$ de modulo de B = B. |
| Theorem 3 (Measurable Image Theorem) Suppose (Ω, B) and (Ω, B ) are standard Borel spaces and that f : Ω → Ω is measurable and 1 − 1, then f (A) ∈ B for every A ∈ B. | Le théorème 3 (théorème mesurable d'image) supposent (Ω, B) et (le Ω, B) sont les espaces standard de Borel et que le Ω de → de f : Ω est le ∈ B mesurable et de 1 − 1, puis f (a) pour chaque A ∈ B. | Le théorème 3 (théorème mesurable d'image) supposent que (Ω, B) et (Ω, B) sont les espaces standard de Borel et que f : Ω → Ω est mesurable et 1 − 1, puis f (A) ∈ B pour chaque A ∈ B. |
| The proof is given in appendix A. Using a positive integer representation of the prime field elements (integers between 0 and p − 1), the following corollary holds: | La preuve est donnée dans l'annexe A. utilisant une représentation positive de nombre entier des éléments de champ principal (les nombres entiers entre 0 et le p − 1), le corollaire suivant se tient : | La preuve est donnée dans l'annexe A. utilisant une représentation positive de nombre entier des éléments de champ principal (nombres entiers entre 0 et p − 1), que le corollaire suivant se tient : |
| Instead, using a balanced representation (integers between − p−1 2 and p−1 2 ), this bound can be improved: | Au lieu de cela, utilisant une représentation équilibrée (des nombres entiers entre le − p −1 2 et p−1 2), cette limite peuvent être améliorés : | Au lieu de cela, utilisant une représentation équilibrée (des nombres entiers entre le − p −1 2 et p−1 2), cette limite peuvent être améliorés : |
| Proof. Extending the previous notation MM (n), we denote by MM (m,k,n) the cost of multiplying a m × k by a k × n matrices. The cost function TRSM(m, n) satisfies the following equation: | Preuve. prolongeant la notation précédente le millimètre (n), nous dénotons par le millimètre (m, k, n) le coût de multiplier m × k par des matrices de k × n. la fonction de coût TRSM (m, n) satisfait l'équation suivante : | Preuve. prolongeant la notation précédente MM (n), nous dénotons par MM (m, k, n) le coût de multiplier a m × k par des matrices de a k × n. que la fonction de coût TRSM(m, n) satisfait l'équation suivante : |
| Let t = log 2 (m). Although the algorithm works for any n, we restrict the complexity analysis to the case where m ≤ n for the sake of simplicity. We then have: | Laissez le rondin de t = 2 (m). bien que l'algorithme fonctionne pour n'importe quel n, nous limitent l'analyse de complexité au cas où m ≤ n dans l'intérêt de simplicité. nous ont alors : | Laissez t = log 2 (m). bien que l'algorithme fonctionne pour n'importe quel n, nous limitent l'analyse de complexité au cas où m ≤ n dans l'intérêt de simplicité. nous ont alors : |

FIGURE 2.2 – Exemples sur l'impact de la reconnaissance d'entités scientifiques développée par SYSTRAN.

que les annotations mathématiques ou encore les formules, ces dernières se trouvent décomposées lors de l'analyse de la phrase source. Conséquemment, l'hypothèse de traduction correspondante et automatiquement générée s'en trouve fortement dégradée.

## 2.4 Adaptation en domaine d'un système de TAS

Une partie des contributions de cette thèse dans le projet COSMAT s'inscrit dans les travaux présentés ci-après.

La plateforme HAL référence plus de 220.000 documents avec texte intégral répartis dans une trentaine de domaines scientifiques. Pour le moment, nous nous sommes limités à deux de ces domaines : l'Informatique et la Physique. Ces deux domaines représentent respectivement 21% et 17,6% du volume total, soit plus du tiers de l'ensemble des domaines représentés. Pour chacun d'entre eux, nous avons respectivement développé un système de TAS adapté.

### 2.4.1 Extraction de données bilingues du domaine

Les documents de HAL sont presque exclusivement monolingues et en anglais. Toutefois, un sous-ensemble de ces documents est composé de thèses des universités françaises, qui elles sont en français et qui doivent inclure à la fois un résumé en français et en anglais. Bien que, dans certains cas, les deux résumés peuvent ne pas être strictement des traductions parallèles ou peuvent contenir des erreurs de traduction, nos expériences ont montré que ces résumés peuvent s'avérer être des données parallèles des plus utiles. Nous avons donc fait le choix de nous servir de ces résumés de thèse comme corpus bilingues pour adapter au domaine notre système de TAS. Cependant, pour éviter tout possible problème dans nos données, nous les avons filtrées avant utilisation.

Pour ce faire, nous avons préalablement aligné ces résumés de thèses au niveau de la phrase. Ensuite, pour extraire des données bilingues d'entraînement, de développement et de test, nous avons procédé comme suit : pour éviter d'inclure des paires de phrases mal alignées dans les données de développement et de test, nous avons effectué une sélection sur la base du coût du modèle IBM 1 [Brown 1993] (section 1.4.2.2) généré pour chaque paire de phrases. À partir de ces scores IBM 1, nous avons fixé un seuil qui servit de critère de sélection en dessous duquel, les paires de phrases pouvaient être considérées comme étant la traduction l'une de l'autre. De ces paires de phrases sélectionnées, les corpus de développement et de test ont ensuite été choisis aléatoirement. Environ 100k mots [14] pour chacun des domaines Informatique et Physique ont ainsi été sélectionnés. Le reste des données ayant été utilisé comme corpus d'entraînement. Les statistiques de ces ensembles de données parallèles sont résumées dans le tableau 2.1.

14. ici « k » signifie « millier »

| Corpus | Domaine | Lang. | #Phrases | #Mots | #Vocab. |
|---|---|---|---|---|---|
| *données bilingues en-domaine* | | | | | |
| App. | info+phys | En | 75.7 k | 1.98 M | 61.3 k |
| | | Fr | 75.7 k | 2.3 M | 64.2 k |
| Dev | info | En | 2053 | 50.3 k | 6.2 k |
| | | Fr | 2053 | 57.5 k | 6.9 k |
| | phys | En | 1958 | 49.8 k | 6.3 k |
| | | Fr | 1958 | 55.8 k | 6.9 k |
| Test | info | En | 2145 | 50.1 k | 6.0 k |
| | | Fr | 2145 | 56.5 k | 6.9 k |
| | phys | En | 2025 | 49.8 k | 6.5 k |
| | | Fr | 2025 | 55.8 k | 7.2 k |

TABLE 2.1 – Statistiques sur les données d'apprentissage, de développement et de test extraites des résumés de thèses disponibles sur HAL, respectivement pour les domaines Informatique et Physique. (M pour million et k pour millier)

**Valorisation**

Ce corpus que nous nommerons « Corpus COSMAT » ainsi constitué des domaines Informatique et Physique est aujourd'hui librement distribué [15] et semble la seule ressource de ce genre (composée de contenus scientifiques) à être disponible. Il a notamment déjà été utilisé en 2012 par le « JHU SMT workshop [16] » et par le projet européen « TransLectures [17] ».

### 2.4.2 Données d'apprentissage hors-domaine

Les données en domaine extraites de HAL ont été utilisées pour adapter au domaine le modèle de traduction du système de TAS correspondant. Ces modèles de traduction furent entraînés principalement sur des données parallèles génériques fournies lors de campagnes d'évaluations. Ces corpus bilingues hors-domaine utilisés étaient le corpus « EUROPARL7 » qui est composé de procédures du Parlement Européen, le corpus « NEWS COMMENTARY7 » qui est composé d'articles de commentateurs sur des nouvelles d'actualité, et le corpus « CCB2 » composé lui de données bilingues téléchargées automatiquement sur Internet. Les volumes respectifs de ces corpus en nombre de mots pour les langues anglais et français sont donnés dans le tableau 2.2 :

15. Contact : mailing-cosmat@cosmat.fr
16. http ://www.clsp.jhu.edu/workshops/archive/ws-12/groups/dasmt/
17. http ://www.translectures.eu/

| Corpus | #Phrases | #Mots | |
|---|---|---|---|
| | | Anglais | Français |
| *données bilingues tokenisées hors-domaine* | | | |
| Europarl7 | 2.0 M | 55.4 M | 61.2 M |
| News Commentary7 | 137 k | 3.3 M | 4.0 M |
| CCB2 | 7,4 M | 232 M | 266 M |

TABLE 2.2 – Données d'apprentissage hors-domaine en nombre de phrases et de mots (après tokenisation). (M pour million et k pour millier)

Pour l'apprentissage de notre modèle de langage, nous avons utilisé la version monolingue des corpus parallèles, ainsi que des données monolingues hors-domaine. Toutefois, ces données n'ont pas été exploitées dans leur globalité, mais seuls des sous-ensembles de celles-ci, identifiés comme étant les plus intéressants pour nous vis-à-vis des domaines concernés l'ont été. Pour ce faire, nous avons filtré nos données hors-domaine à l'aide de « XenC » [Rousseau 2013], un outil open-source [18] de sélection de données qui implémente les algorithmes de [Moore 2010, Axelrod 2011]. Dans notre cas, c'est en mode « sélection monolingue » qu'il fut utilisé.

### 2.4.3 Sélection de données hors-domaine

Nous l'avons évoqué au chapitre précédent, la modélisation du langage nécessite une quantité importante de données monolingues, et jusque là la volonté était d'utiliser le plus de données possible tel que le commentera Bob Mercer en 1985 :

> "There is no data like more data."
>
> [Jelinek 2004]

Idéalement les données utilisées sont du même domaine que la tâche, mais dans la pratique, la disponibilité de telles données est souvent limitée. Il convient alors d'ajouter des données supplémentaires qui sont cette fois hors domaine. Cependant, cette pratique n'est pas sans poser problème puisqu'elle revient à ajouter des données « inutiles » et potentiellement néfastes qui viendront de fait bruiter le modèle. Pour réduire au maximum ce bruit, nous avons réalisé une sélection de données sur ces corpus hors-domaine pour ne conserver qu'un sous-ensemble. Ce sous-ensemble correspond alors à ce qu'il y a de plus proche vis-à-vis de notre domaine, parmis nos données hors-domaine.

18. https ://github.com/rousseau-lium/XenC

Réalisé avec « XenC », cette sélection sur des données monolingues est effectuée en calculant l'entropie croisée (*cross-entropy* en anglais) pour chacune des phrases du corpus hors-domaine, par rapport à deux modèles de langage préalablement construits :

- Un premier modèle estimé à partir de l'ensemble des données considérées comme faisant partie du domaine ;
- Un second quant à lui estimé sur un sous-ensemble aléatoire des données parmis lesquelles on souhaite extraire les phrases les plus intéressantes, de taille similaire à l'ensemble des données faisant partie du domaine.

Cette pratique nous a permis de réduire sensiblement la taille des modèles en passant de 18Go pour la première version (utilisant l'ensemble des données), à 700Mo après réalisation d'une sélection des données. Soit une division par 25 du volume global. Les performances finales des systèmes de TAS en termes de qualité de traduction ne s'en trouvent pas pour autant diminuées. Comme le montre le tableau 2.3, elles s'en trouvent même améliorées :

| Sélection de données | | Aucune | XenC |
|---|---|---|---|
| *Système* | *Domaine* | | |
| En→Fr | info | *nc* | 31.03 |
| | phys | 35.13 | 36.01 |
| Fr→En | info | 30.00 | 30.75 |
| | phys | 35.13 | 36.83 |

TABLE 2.3 – Scores BLEU sur le corpus de test en-domaine obtenus respectivement par les systèmes de TAS adaptés en-domaine avec et sans sélection de données monolingues pour le modèle de langage.

À noter que cette approche, bien qu'efficace tant sur la réduction du volume des modèles que sur la qualité de traduction système, n'est pas la seule méthodologie existante. D'autre approches d'adaptation pour un système de TAS ayant en effet été proposées par le passé :

- L'apprentissage non supervisé [Lambert 2012b]. Cette technique permet d'adapter le modèle de traduction d'un système statistique en utilisant uniquement des données monolingues ;
- La pondération des corpus d'apprentissage selon leur importance pour la tâche [Shah 2011, Shah 2012]. Habituellement, les données d'apprentissage disponibles sont très hétérogènes par rapport à plusieurs facteurs (actualité, proximité au domaine, qualité des traductions, etc.). Cette connaissance pouvant être intégrée au système.

**Techniques avancées pour les systèmes de TAS**

Nous avons beaucoup travaillé sur des techniques de modélisation statistiques alternatives durant le projet COSMAT, en particulier sur une représentation en espace continu (voir section 1.4.1). Le LIUM comptant parmi les acteurs principaux dans ce domaine [Schwenk 2007]. L'ensemble des recherches effectuées par le LIUM dans le cadre de COSMAT est intégré dans un toolkit librement disponible incluant un support de cartes accélératrices GPU [Schwenk 2012b, Schwenk 2013]. Ainsi, il est possible de créer un modèle sur 4 milliards de mots en moins de 24 heures, alors que plus d'une semaine était nécessaire au début du projet. Des travaux ont également été menés pour une généralisation de cette approche aux modèles de traduction [Schwenk 2012a]. Celle-ci semble être très prometteuse et sera poursuivie au-delà du projet.

## 2.5 Interface de post-édition

Le projet COSMAT a également permis de travailler sur l'utilisation des corrections des utilisateurs pour améliorer notre système de TAS. L'idée étant que ce projet à destination de la communauté scientifique soit également porté par elle. L'aspect collaboratif du projet, que nous avons évoqué précédemment, prévoit que les utilisateurs puissent procéder à la révision de version préalablement traduite de la ressource qu'ils sont en train de soumettre sur HAL. De fait, ces retours utilisateurs peuvent être utilisés *a posteriori* pour une adaptation du système de TAS ayant généré cette traduction. Il s'agit pour être extact, d'adapter son modèle de traduction.

Cette adaptation repose en partie sur un nouvel algorithme d'alignement de mots élaboré pour permettre d'effectuer une mise à jour du modèle de traduction sans que la procédure d'alignement, chronophage, soit nécessaire [Blain 2012]. Ainsi, les modèles peuvent être mis à jour plus rapidement. Ces travaux, dont l'efficacité a été confirmée sur les données COSMAT, constituent une partie des contributions de cette thèse qui sont présentées en détail en seconde partie de ce manuscrit.

L'aspect collaboratif du service proposé avec COSMAT repose également sur un outil de post-édition intégré dans l'interface de visualisation des traductions. Une illustration de cette interface riche, développée par la société SYSTRAN, est donnée en figure 2.3. Les corrections sont enregistrées sous la forme de préférences terminologiques, et de phrases post-éditées. Réalisée en Javascript, cette interface de post-édition propose les fonctionnalités suivantes :

- **Un affichage dit « WYSIWYG** [19]**) de la phrase source et de sa traduction en langue cible (Zones 1+2)** – *La visualisation proposée à l'utilisateur est la même pour la partie source et la partie cible du document. De cette manière l'utilisateur voit immédiatement à quoi ressemblera la version traduite de son document, avec une mise en page scientifique conservée* ;
- **Un alignement au niveau de la phrase (Zone 3)** – *L'utilisateur peut vérifier de cette manière que la traduction respecte bien la phrase source* ;
- **Un espace de révision de la traduction (Zone 4)** avec un alignement des termes sources et cibles et **des références terminologiques (Zone 5)** ;
- **Des traductions alternatives (Zone 6)** – *L'utilisateur se voit proposer des traductions alternatives qu'il pourra utiliser ou à partir desquelles il pourra générer une traduction définitive s'il juge que la traduction qui lui est proposée est de trop mauvaise qualité. De cette manière, l'effort de post-édition tend à être réduit.*

19. « WYSIWYG – What You See Is What You Get » est une acronyme anglais signifiant « ce que vous voyez, est ce que vous obtenez »

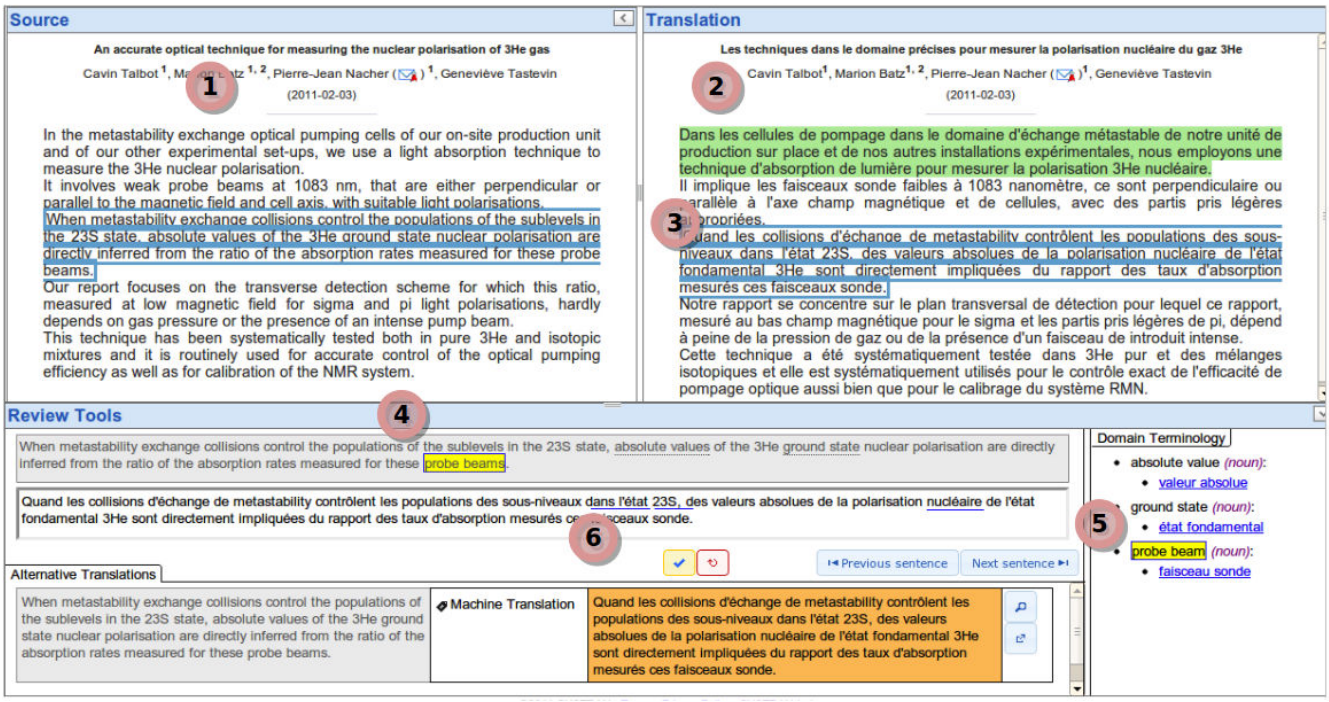


FIGURE 2.3 – Interface de post-édition pour COSMAT développée par la société SYSTRAN.

### 2.5.1 Campagnes d'évaluation

Cette interface de post-édition a pu être testée en conditions réelles d'utilisation par SYTRAN à l'occasion de deux conférences internationales :

- **La conférence en physique OECS12 (12-16 septembre 2011)** – *Cette première expérience a permis de raffiner les interactions utilisateurs et de produire une seconde version plus performante* ;
- **La conférence LREC (23-25 mai 2012)** – *Une copie d'écran de l'interface légèrement repensée suite aux retours obtenus d'OECS'12 est proposée en figure 2.4.*

A chacune de ces deux manifestations, nous avons proposé aux auteurs de réviser une traduction des résumés de leurs articles respectifs, dont la traduction était issue de nos systèmes de TA. Ces révisions devaient être réalisées en respectant un ensemble de consignes de post-édition communiquées aux auteurs. Destinées avant tout à faire connaître le projet COSMAT auprès de la communauté internationale, ces deux expériences nous ont donc permis d'évaluer notre interface, mais également de collecter un premier jeu réel de données humainement validées et correspondant à nos systems de TA. À partir de ces données, nous avons pu lancer des analyses sur l'activité des utilisateurs vis-à-vis la tâche de révision effectuée via notre interface dédiée.

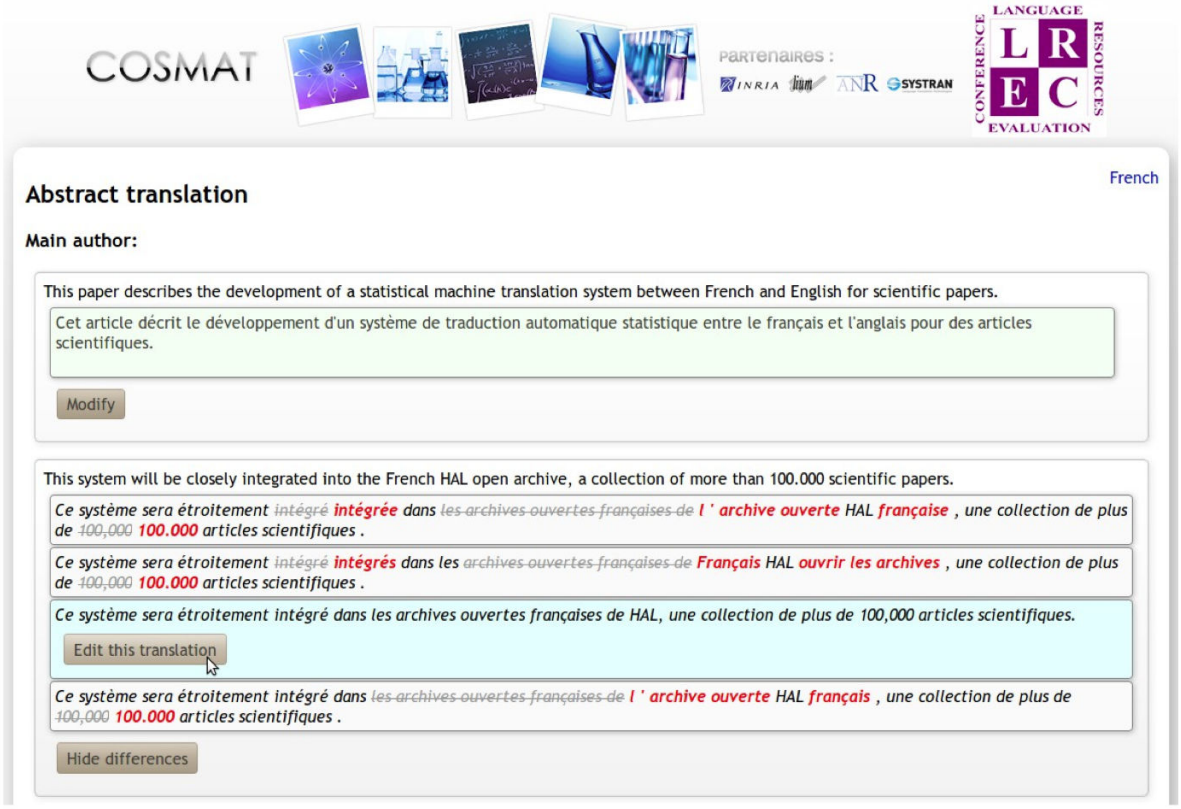


FIGURE 2.4 – Interface COSMAT de visualisation des traductions. Ici utilisée lors de la conférence LREC en 2012.

#### 2.5.1.1 Pertinence utilisateur

Parmis les analyses évoquées précédemment, nous nous sommes interéssé dans le cadre de cette thèse à ce qu'on a appelé la « pertinence de l'utilisateur ».

Lors du processus de révision, l'utilisateur est amené à faire des choix : il doit en effet choisir s'il conserve en l'état ou non la traduction automatique qui lui est proposée. De fait, cela induit un facteur humain dans la chaîne de révision qui, de notre avis, mérite d'être pris en compte lors de l'évaluation. Nous considérons en effet que si l'utilsateur venait à ne pas choisir la meilleure hypothèse de traduction qui lui est proposée (i.e. qu'il ne fait pas le bon choix), cela engendrerait un effort additionnel de post-édition et par conséquent, laisserait entendre que la traduction est de basse qualité (avec une majoration mécanique du score associé à cette traduction, que ce soit WER, TER ou BLEU). Ainsi, nous considérons que l'utilisateur fait le bon choix, respectivement le mauvais, lorsque la distance d'édition entre la version validée d'une traduction (que nous considérerons comme étant ici une référence) et son hypothèse de

départ est plus faible, respectivement plus élevé, qu'entre cette même référence et une hypothèse alternative.

**Discussion sur les données de la campagne OECS'12**

Partant du principe précédent, nous avions observé une très haute précision (99%) lorsque l'utilisateur sélectionnait la traduction alternative, et une précision assez élevée lorsque l'hypothèse par défaut était conservée.

Cette observation quoique logique, confirme que les post-éditeurs ont très majoritairement choisi l'hypothèse de traduction qui leur semblait la plus satisfaisante avant de potentiellement la post-éditer. Le premier point est que ce comportement respecte les consignes de post-édition qui leur avaient été données. Le second point est de bon augure pour nos perspectives de travail : en choisissant la meilleure hypothèse de traduction parmi celles qui leur avaient été proposées, les post-éditeurs ont généré une traduction finale ayant une distance d'édition potentiellement la plus faible possible vis-à-vis de l'hypothèse de départ. Ceci étant, nous sommes optimistes quant à l'extraction future d'informations en vue d'améliorer la qualité de traduction de notre système de TAS.

Afin de comprendre pourquoi certains post-éditeurs ont préféré retraduire entièrement certaines phrases plutôt que de choisir l'une des hypothèses de traduction proposées, nous avons calculé la distance d'édition (avec l'algorithme du TER) entre cette traduction *from-scratch* prise comme référence et respectivement chacune des hypothèses de traduction proposées. Les scores TER obtenus se sont révélés importants (supérieurs en moyenne à 50%), et compte tenu du ratio de mots par phrase constaté, ces résultats suggèrent nettement que l'effort de post-édition aurait été plus important qu'une retraduction humaine complète, due au nombre important d'éditions visiblement nécessaires. Ceci indique également qu'une fois encore, le comportement des post-éditeurs face à la tâche de post-édition tendait à respecter les consignes données.

De par ces constatations, nous sommes confiants quant à l'interface de post-édition développée qui, associée à des consignes de post-édition, tend à conforter le post-éditeur dans sa tâche. Cette information est d'autant plus importante que nous souhaitions exploiter ces retours utilisateurs pour adapter dans le temps nos systèmes de TAS comme indiqué jusqu'ici, et dont nous allons voir comment dans la troisième partie de ce manuscrit.

Avant cela, nous vous proposons de voir comment le service de traduction de COSMAT doit être intégré dans la plateforme HAL. De cette manière, nous aurons un aperçu complet du cadre applicatif de cette thèse.

## 2.6 Intégration dans HAL

Le service COSMAT, une fois pleinement intégré dans HAL, se veut totalement transparent pour l'utilisateur. Lorsque ce dernier dépose une nouvelle ressource scientifique sur la plateforme, une version traduite lui est proposée automatiquement, traduction qu'il peut ensuite modifier s'il estime que c'est nécessaire. D'un point de vue de l'utilisateur, le scénario d'usage est alors le suivant :

- **Une ressource scientifique au format PDF (par exemple un article) est uploadé sur la plateforme par un utilisateur** ;
- **Le document est pré-traité par l'outil open-source GROBID afin d'en extraire le contenu textuel** – *L'article ainsi extrait est alors structuré au format TEI où le titre, le(s) auteur(s), les références, les légendes, mais également les notes de bas de page, etc. sont identifiés* ;
- **Une reconnaissance d'entités est appliquée pour le balisage des entités propres au domaine concerné** – *telles que : formules chimiques pour les articles en Physique, les formules mathématiques, pseudo-codes et références objets pour les articles en Informatique, mais également les divers acronymes communément utilisés dans la communication scientifique* ;
- **La terminologie spécialisée est identifiée** – *en utilisant la base de données référence en terminologie dénommée* TERMSCIENCES [20]*, complétée avec de la terminologie extraite automatiquement du corpus d'entraînement. La traduction de l'article est alors réalisée en utilisant un modèle de traduction adapté* ;
- **Le processus de traduction génère un format TEI bilingue** – *Ce dernier préserve la structure source et qui intègre l'annotation des entités, les choix terminologies multiples quand ceux-ci existent, et l'alignement au niveau des mots entre les phrases sources et cibles* ;
- **La traduction issue du système de TA est proposée à l'utilisateur** – *Chaque paire de phrases (la source et sa traduction) est révisée par l'utilisateur via l'interface de post-édition de* COSMAT ;
- **La version définitive du document est générée et archivée au format TEI** – *Elle est désormais disponible à la consultation au travers d'un un affichage en HTML rendu possible par l'utilisation d'une feuille de style* XSLT.

20. http ://www.termsciences.fr

### 2.6.1 Architecture globale du service COSMAT

L'architecture globale, illustrée par la figure 2.5, s'articule de la façon suivante : d'un côté nous avons la plateforme HAL sur laquelle se connecte l'utilisateur via une interface dédiée, pour déposer ou consulter une nouvelle ressource. Cette interface va communiquer avec l'API [21] COSMAT pour obtenir une traduction de cette ressource. De l'autre côté, nous avons le Service COSMAT qui regroupe l'ensemble des technologies nécessaires pour le traitement de ressources provenant de HAL (i.e. un extracteur de contenu structuré et un serveur de traduction spécialisé et adapté à la tâche).

Le service de traduction est hébergé sur des serveurs de l'entreprise SYSTRAN dédiés au projet. Un protocole de communication « RESTfull » entre la plateforme HAL, le service d'extraction de contenu structuré GROBID et les serveurs de traduction a été défini et la communication est à ce jour opérationnelle.

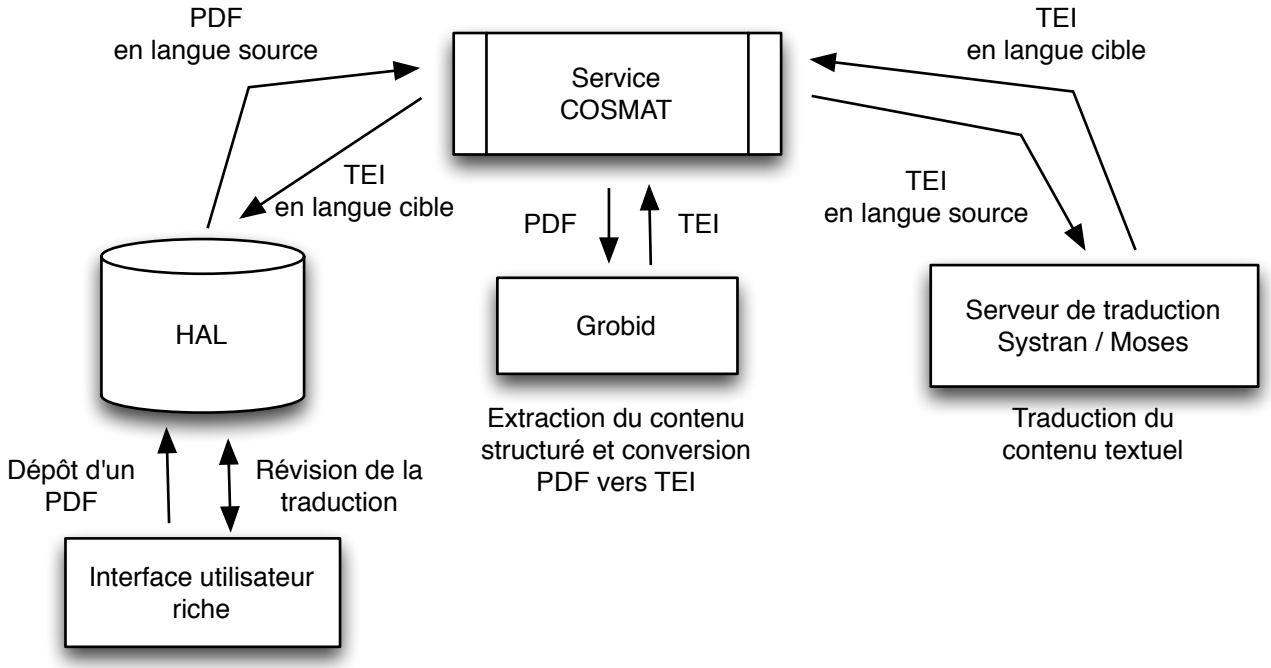


FIGURE 2.5 – Architecture globale du service collaboratif COSMAT intégré dans HAL. Le protocole de communication entre HAL, le serveur GROBID et le serveur de traduction est basé sur une interface « RESTFUL ».

21. API est un acronyme anglais pour « Application Programming Interface) »

Malheureusement, l'intégration opérationnelle du service dans la version courante de HAL n'a pu être faite dans le cadre du projet, car une nouvelle version de HAL a été planifiée. Celle-ci repose en grande partie sur l'intégration d'un moteur de recherche qui permettra de décliner des services plus performants, ce qui doit améliorer les services proposés aux utilisateurs. La procédure de dépôt reste elle inchangée, mais l'ergonomie évolue : l'ordre des étapes ne sera plus imposé. Ce nouveau workflow permettra ainsi d'intégrer la fonctionnalité de récupération des métadonnées à partir d'un fichier PDF, très attendue, ainsi que les services de traduction, notamment des résumés d'articles. La date de production de la version 3 de HAL est prévue pour novembre 2013, mais dès maintenant l'intégration des services a débuté dans le cadre d'un travail commun entre les équipes de l'INRIA et celles de SYSTRAN.

Le projet COSMAT est donc un servrice de TA porté pour et par la communauté scientifique. Premièrement, **pour la communauté** parce qu'il permettra lorsque son intégration dans la version 3 de la plateforme HAL sera effective, de traduire des ressources scientifiques et de permettre leur accessibilité de et vers l'anglais, langue référence de cette communauté scientifique internationale. Deuxièmement, **par la communauté** de par son aspect collaboratif, puisqu'il offrira également la possibilité à cette dernière comme nous venons de le voir, de le faire évoluer continuellement dans le temps, par la révision des traductions, à partir desquelles nous avons travaillé à extraire les informations utiles pour rendre les modèles traductions de nos systèmes de TAS évolutifs. C'est là tout le cadre applicatif dans lequel s'inscrivent les travaux de recherche de cette thèse CIFRE, que nous allons maintenant aborder en troisième partie de ce manuscrit.

Avant cela, il est intéressant de noter que l'ensemble de ce qui vient être décrit dans ce chapître peut être par la suite généralisable d'autres domaines scientifiques de HAL, ainsi qu'à d'autres paires de langues.

### Publication

Nos contributions dans le cadre du projet COSMAT ont donné lieu à la publication des articles [Lambert 2012a, Lambert 2012b] présentés respectivement à Language Resources and Evaluation (LREC) et à l'European Chapter of the Association for Computational Linguistics (EACL), toutes deux en 2012.

# Troisième partie

# Contributions

# Chapitre 3

# Analyse qualitative et automatique de données post-éditées

## Sommaire


**3.1 Analyser la post-édition . . . . . 72**
3.1.1 Les Actions de Post-Édition (APE) . . . . . 72
3.1.2 Typologie des actions de post-édition . . . . . 76
**3.2 Automatisation du processus d'analyse . . . . . 78**
3.2.1 Protocole d'analyse en APE . . . . . 79
3.2.2 Règles linguistiques . . . . . 79
3.2.3 Disponibilité de l'outil « SmartDiff » . . . . . 80
**3.3 Données expérimentales . . . . . 81**
3.3.1 Annotation manuelle de référence . . . . . 81
3.3.2 Résultats de l'analyse automatique . . . . . 84
**3.4 Conclusion . . . . . 86**

Au cours des dix dernières années, les professionnels de la traduction automatique, stimulés par des approches orientées corpus, ont renouvelé leurs offres et ont commencé à présenter des solutions de traductions « hautement personnalisées », et ce pour des domaines et usages spécifiques. Pour de la documentation technique et une assistance en ligne par exemple. Plusieurs rapports attestent de la réalité de cette activité et l'analyse du marché montre une tendance pour les fournisseurs de services linguistiques à offrir de la TA post-éditée à leurs clients (déjà 42% en 2010 selon [DePalma 2010]). De grandes sociétés comme Symantec, Autodesk ou encore Cisco, qui possèdent toutes un service de localisation, se sont également tournées vers la TA et la post-édition comme un moyen de réduire les coûts et le « *time-to-market* ». Dans ce contexte industriel d'utilisation massive, la gestion de l'effort correspondant est un élément important.

Au-delà de l'évaluation même de l'effort de post-édition, se pose le problème intrinsèque suivant : comment cet effort peut-il être réduit ? De multiples approches existent et quelques-unes répondant implicitement à cette question ont été énoncées précédemment lorsque nous avons abordé l'approche hybride pour la TA au chapitre 1 (voir section 1.5). Pour rappel, [Simard 2007] et [Dugast 2007] proposent une approche où un système de TAS, appliqué en sortie d'un système à base de règles, est entraîné sur un corpus parallèle constitué de la sortie d'un système de TA et de références humaines. Ce procédé est appelé Post-Édition Statistique (PES) (*Statistical Post-Editing* en anglais). [Schwenk 2009b] reproduit ceci avec un système PES entraîné sur un très large corpus, transformant la traduction initiale en « simple prétraitement ». Ils montrent ainsi comment le système est capable de « corriger la sortie de traduction ». Le système de TAS étant davantage un correcteur statistique d'erreur de traduction, qu'un traducteur au sens premier du terme. Celui-ci bénéficiant de plus hautes similitudes entre le texte pré-traduit et une référence, comparé à une source et sa référence.

Ces approches ont de commun qu'elles allègent la tâche de post-édition en améliorant la qualité finale de la traduction soumise aux post-éditeurs (sous réserve que ces derniers soient pertinents dans leurs choix). L'apprentissage de nouveaux modèles de correction d'erreurs de traduction par le système de TAS étant rendu possible avec ré-entrainement complet incluant de nouvelles données post-éditées. Il est alors possible d'en bénéficier par la suite.

Bien que ces approches soient efficaces pour préparer la tâche de post-édition, elles ne proposent aucune solution quant à sa **répétitivité**, une caractéristique malheureusement bien connue annotateurs humains. Lorsque l'on observe des données post-éditées, on constate très rapidement qu'une même édition peut être récurrente, et ce que ce soit dans un même document, un même paragraphe ou dans une même phrase. Ceci est somme toute logique puisque ce document aura été traduit, en une seule fois, avec un seul et même système de TA. Nous nous sommes donc intéressés dans le cadre de cette thèse à cette particularité pour finalement

proposer une méthodologie qui permet d'exploiter cette répétitivité pour la tourner à l'avantage du traducteur dans sa tâche. De cette manière, du moins théoriquement, l'effort de post-édition est réduit au fur et à mesure que l'annotateur progresse dans sa tâche.

Cette approche soulève deux problématiques : la première est celle du coût computationnel que représente l'apprentissage pour un système de TAS pour qu'il puisse intégrer rapidement ces nouvelles ressources. Pour pallier cet aspect de la TAS, dans notre contexte de post-édition nous avons développé une procédure d'adaptation incrémentale d'un système de TAS permettant de bénéficier plus rapidement des retours utilisateurs, et ainsi limiter l'effort de post-édition. Cette problématique ne sera pas abordée ici, mais dans le chapitre suivant.

Seconde problématique : doit-on se contenter d'exploiter les données post-éditées dans leur ensemble, au risque de bruiter les connaissances du système (avec par exemple des corrections styliques, ou possiblement erronées), ou tenter au contraire d'en extraire la quintessence ? Pour répondre à cette question, nous nous sommes intéressés à la nature même de la post-édition et avons proposé une méthodologie basée sur autre angle d'analyse de données post-éditées que les approches existantes. Celle-ci s'appuie sur une nouvelle notion destinée à permettre de mieux identifier l'information primaire en sortie du processus de post-édition. De fait, notre attention s'est portée sur l'intention du post-éditeur en vue, non pas de comprendre l'erreur, mais bel et bien de comprendre l'action ayant amené à la correction de cette erreur pour améliorer la traduction.

Dans la suite de notre chapitre, nous nous intéresserons dans un premier temps à la définition d'une nouvelle notion que nous avons introduite pour mieux souligner l'intention du post-éditeur durant sa tâche. Nous définirons ensuite une nouvelle typologie de la post-édition basée sur cette notion et dont le but est de permettre la classification de nos observations. Nous continuerons ensuite avec la présentation du protocole analytique qui nous permet d'automatiser cette nouvelle approche. Nous terminerons enfin par la présentation des résultats obtenus sur un jeu concret de données post-éditées, mis à disposition par des clients de la société SYSTRAN. Pour évaluer et valider notre approche, nous présenterons une analyse comparative des résultats de notre outil d'annotation avec ceux résultants d'une analyse manuelle.

## 3.1 Analyser la post-édition

La post-édition est considérée par plusieurs études comme une classification des erreurs liées à la TA afin de mieux rationaliser l'effort de post-édition :

> " familiarity with the pattern of errors produced by a particular MT system is an important factor in reducing post-editing time "
>
> *« la familiarité avec les motifs d'erreurs produites par un système de TA particulier est un facteur important dans la réduction du temps de post-édition »*
> [Martinez 2003]

[Guzmán 2007] décrit une configuration où un ensemble de règles de post-édition est appliqué sur l'hypothèse de traduction, dans le but de la « lisser » pour obtenir un système hautement personnalisé.

Dans cette thèse, nous soutenons que l'activité de post-édition peut être modélisée par l'intermédiaire d'un ensemble de règles, résultantes de la décomposition et d'une analyse qualitative des résultats de cette post-édition. Notre approche consiste en une extraction automatique d'**un ensemble d'éditions « minimales » et « logiques »**, que nous avons regroupé sous le terme d'« **Actions de Post-Édition** » (APE). Ces éditions logiques étant opposées par définition aux éditions que nous qualifions de « mécaniques », telles qu'on les connaît dans le calcul classique de la distance d'éditions entre deux phrases : l'insertion, la suppression, la substitution et le déplacement (voir section 1.6.2)

### 3.1.1 Les Actions de Post-Édition (APE)

Une APE est dite **minimale** dans le sens que nous ne pouvons pas trouver plus petite et indépendante édition. Une APE est dite **logique** si la transformation qu'elle décrit fait sens linguistiquement. Pour illustrer ce principe, prenons l'exemple de la traduction anglais→français post-éditée suivante :

SOURCE : " By default, the border is diplayed. "
TRADUCTION : « Par défaut, le bord est affiché. »
POST-ÉDITION « Par défaut, la bordure est affichée. » :

La distance d'édition classique telle que nous la connaissions, évaluée entre l'hypothèse de traduction et sa version post-éditée, serait alors de 3 mots substitués par 3 autres mots, comme l'illustre la figure 3.1. Dans ce cas de figure, si nous venions à évaluer la distance d'éditions avec l'algorithme du TER tel qu'il est proposé par Matthew Snover, en associant à chacun des

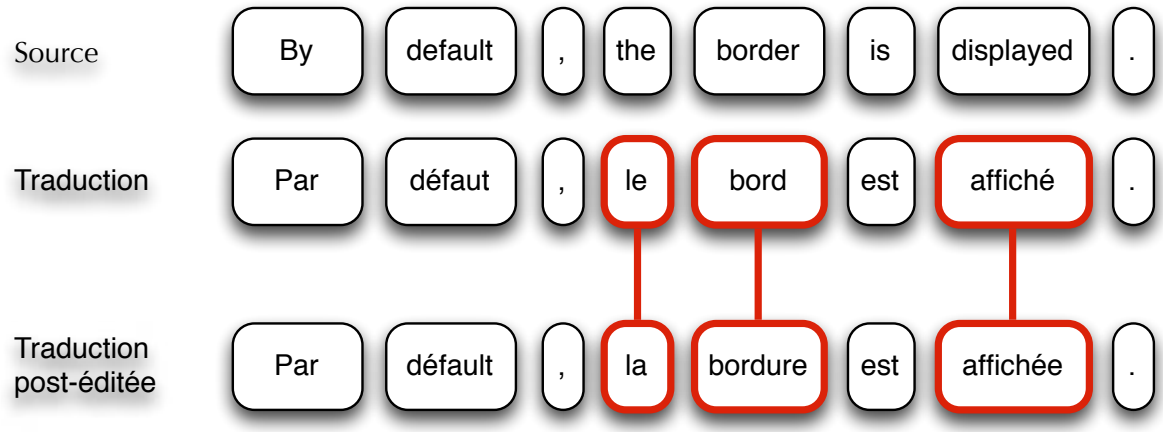


FIGURE 3.1 – Distance d'éditions classique dite « mécanique » entre une hypothèse de traduction et sa version post-éditée.

types d'éditions un poids équivalent, nous obtiendrions d'après la section 1.6.2, un score TER de 37,5 points.

Toutefois, si on y regarde de plus prêt, en effectuant une analyse plus linguistique entre ces deux phrases, on peut observer qu'il ne s'agit pas vraiment de 3 substitutions. En effet, si on considère la plus petite édition logique de notre exemple, on arrive à la conclusion qu'il n'y a en réalité qu'une seule **édition primaire** par la substitution du mot « bord » avec le mot « bordure », les deux étant une traduction valide dans ce contexte du mot anglais « border ». Cette substitution induit alors deux **éditions secondaires** que nous qualifierons de « propagations ». En effet, le changement de genre pour le nom principal va se propager au déterminant et à l'adjectif qui lui sont associés, comme l'illustre cette fois la figure 3.2. Ceci est dû, dans le cadre de notre exemple, à la nature de la langue française qui est morphologiquement riche : le déterminant et l'adjectif d'un nom s'accordent en genre et en nombre avec ce dernier. Désormais, si nous recalculons la même distance d'éditions que précédemment, mais en considérant les deux niveaux d'éditions que nous venons d'évoquer, nous obtenons un score TER en baisse significative de 33%, à 12,5 points.

Premièrement, nous ne comptabilisons pas les éditions secondaires dans notre évaluation puisque nous considérons qu'elles ne sont pas dues à une erreur directe du système. On obtient donc un score TER davantage représentatif de la réelle qualité, non pas de la traduction, mais du système.

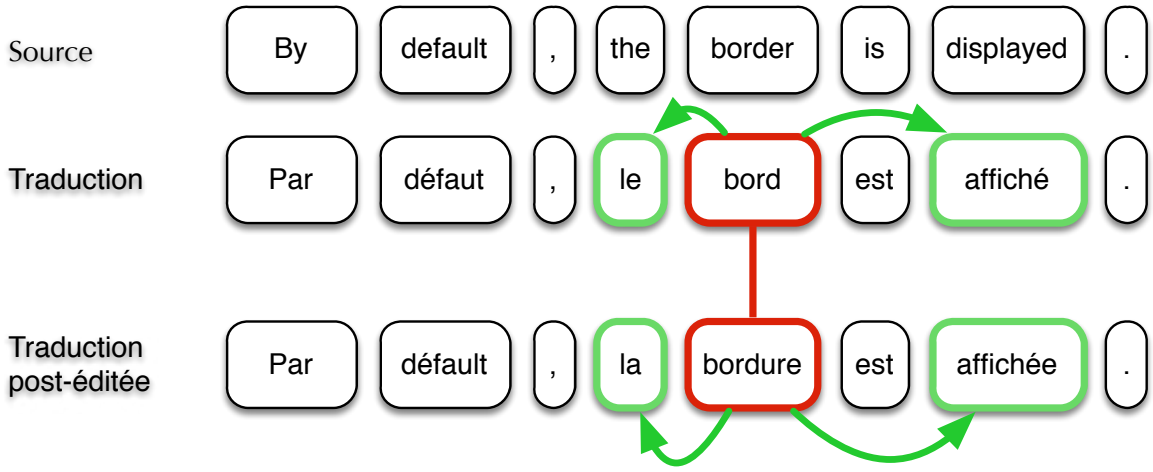


FIGURE 3.2 – Distance d'éditions basée sur l'analyse en APE, dite « logique », entre une hypothèse de traduction et sa version post-éditée.

Deuxièmement, en nous intéressant ainsi à l'intention du post-éditeur, nous avons pu identifier ce que nous considérons comme étant la véritable erreur de traduction du système pour cette phrase. On peut donc imaginer que si nous arrivions à rapidement intégrer cette connaissance à notre système de TA dans le cadre d'une traduction incrémentale d'un document, nous pourrions assûrément éviter la reproduction de cette erreur, et réduire ainsi l'effort de post-édition pour ce même document. Résumées en quelques mots, les différences majeures entre ce que nous considérons comme étant des éditions logiques en comparaison d'éditions mécaniques, nous dirions que les éditions logiques :

1. **Sont plus intuitives pour le post-éditeur** – *Ce dernier, lorsqu'il modifie une hypothèse de traduction, va d'abord corriger ce qu'il considère comme étant une erreur de traduction et effectuera ensuite les modifications nécessaires induites par cette correction. Rendant ainsi publiable la traduction* ;
2. **Permettent une évaluation plus approfondie de la traduction** – *Généralement, plusieurs éditions de mots peuvent être incluses dans une même APE. Dans l'exemple précédent, deux éditions correspondent à la même APE d'accord en genre avec le nom féminin « bordure »* ;
3. **Sont plus complexes à identifier** – *C'est la contre-partie de cette approche. Dans le cas d'éditions multiples et imbriquées par exemple, il sera très difficile de bien les différencier et ainsi de bien les identifier. Ceci est d'autant plus vrai si on veut le faire de manière totalement automatique.*

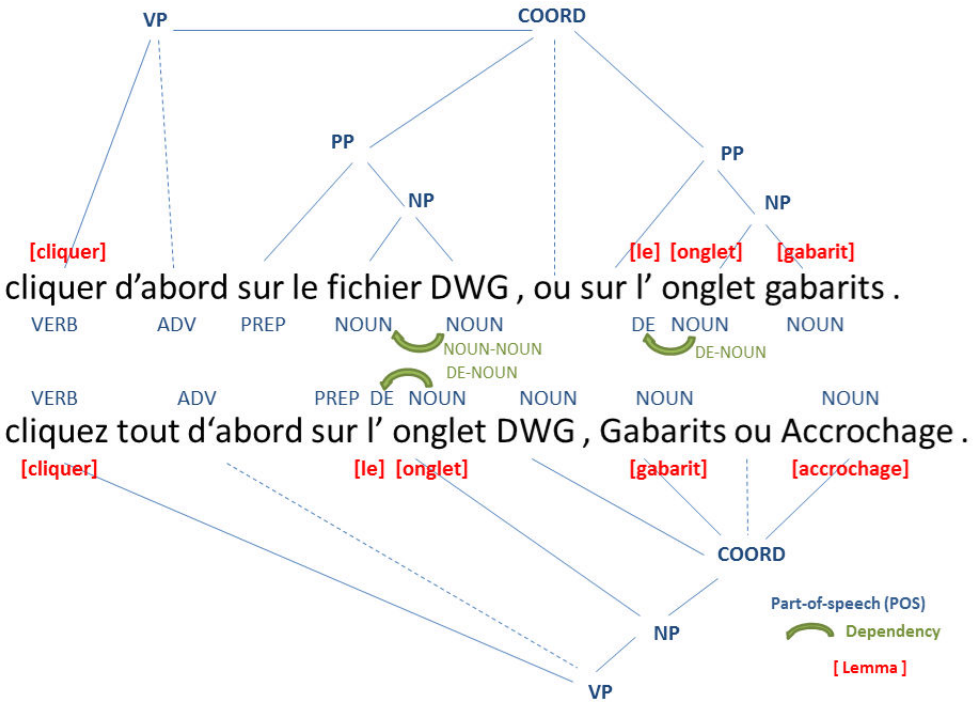


FIGURE 3.3 – Exemple d'annotations linguistiques pour une paire de phrases.

### Qualité de traduction et annotations linguistiques

Nous venons de l'évoquer : tenter de modéliser l'intention du post-éditeur est du domaine du possible dans la mesure où les éditions du processus de révision font sens. C'est-à-dire que le nombre d'édition reste limité et par conséquent, que les éditions logiques restent identifiables. Si le nombre d'édition est trop important, il sera plus difficile, voire impossible, de distinguer les éditions secondaires des éditions primaires.

Dans notre contexte, les post-éditeurs sont des traducteurs professionnels exerçant dans des services de localisation, ou bien des scientifiques experts du domaine dans lequel ils évoluent. Dans ces deux situations, les annotateurs sont soumis à des directives (ou consignes) très strictes pour effectuer une post-édition « légère » (voir section 1.7.2), ce qui est suffisant pour des traductions issues de systèmes de TA bien adaptés à la tâche. Ici, pour traiter de la documentation technique ou de la littérature scientifique. Cela crée d'une certaine manière des références humaines «ciblées », naturellement adaptées à l'évaluation HTER (voir section 1.7.2). Dans notre approche toutefois, le taux d'erreur de traduction basé sur l'énumération des modifications mécaniques n'est qu'une étape d'analyse intermédiaire en vue d'exposer *a posteriori* les éditions logiques.

Notre analyse en APE repose en effet sur un ensemble d'informations linguistiques, dont la figure 3.3 donne une illustration. Ces annotations sont couplées à un algorithme évaluant la distance d'édition et nous permettent de déterminer la nature des modifications apportées par le post-éditeur. Nous nous servons de la partie du discours (*part-of-speech* en anglais) qui nous indique si le mot modifié est un verbe, un nom ou encore un adverbe par exemple. Nous nous basons également sur la forme canonique des mots variables (genre et nombre d'un nom ou d'un adjectif, infinitif d'un verbe), et des dépendances entre les mots pour faciliter notamment l'identification des propagations.

Toutefois, certaines éditions ne correspondent pas à la définition d'une APE, soit parce que le texte d'origine ne faisait aucun sens linguistiquement pour être correctement analyser (ce que nous appellerons un « sac de mots »), soit parce que le post-éditeur introduit une erreur, ou bien encore parce la structure de la phrase après révision est radicalement changée, rendant alors impossible la décomposition en APE.

### 3.1.2 Typologie des actions de post-édition

Par opposition aux métriques automatiques que nous avons présenté au chapitre 1, nous souhaitions démontrer qu'en utilisant un ensemble d'informations linguistiques, nous pouvons modéliser les intentions des post-éditeurs.Pour se faire, à partir d'un jeu de données post-éditées sur lequel nous reviendrons plus loin dans ce chapitre, nous avons méticuleusement réalisé un nombre certain d'observations nous ayant permis de définir par la suite une typologie par classe d'APE telle que présentée en table 3.1. Nous y avons ajouté des exemples de manière à faciliter la compréhension de certaines classes. Pour arriver à ce résultat, nous nous sommes inspirés de classifications d'erreurs de traduction existantes proposées par [Font-Llitjós 2005], [Vilar 2006] et [Dugast 2007].

Bien que nous soyons convaincus du fort potentiel de cette approche, celle-ci n'en est encore qu'aux prémices de son développement et reste pour le moment dépendante de la langue cible dans laquelle on souhaite évoluer. Nous entendons par là que les règles d'identification des APE pour une langue ne sont pas forcément adpatées à une autre langue. Ici, nous nous sommes intéressés à l'analyse de données post-éditées pour des traductions de l'anglais vers le français (en→fr), et la typologie établie n'est donc pas la plus adaptée pour des traductions de l'anglais vers l'allemand (en→de) par exemple. Pour se faire, il nous faudrait établir une nouvelle typologie pour la langue allemande. Et cela vaut pour n'importe quelle langue.

| Classe | Sous-classe | Description (+ *exemple*) |
|---|---|---|
| **Groupe Nominal (GN)** | | Changements lexicaux |
| | Changement de déterminant | *Changement volontaire d'un déterminant* |
| | Changement de sens nominal | *Un nom est remplacé par un autre nom* ***avec changement du sens*** |
| | Changement nominal | *Un nom est remplacé par un synonyme* ***sans changement du sens*** |
| | Changement du nombre | Nombre grammaticale (singulier ou pluriel) |
| | Changement de la casse | Ajout/retrait de majuscule/minuscule |
| | Choix d'adjectif | *Un adjectif change pour un meilleur ajustement* ***avec le nom*** |
| | Changement multimots | *Changement d'expression multimots* ***avec changement de sens*** *(ex : carte bancaire → carte de crédit)* |
| | Changement de structure nominale | *La structure d'un GN change, mais le sens est préservé (ex : couleur de l'aperçu → couleur d'aperçu)* |
| **Groupe Verbal (GV)** | | Changements grammaticaux |
| | Changement de sens verbal | *Un verbe est remplacé par un autre nom* ***avec changement du sens*** |
| | Changement verbal | *Un verbe est remplacé par un synonyme* ***sans changement du sens*** |
| | Accord grammatical | *Correction de l'accord d'un verbe* |
| | Changement de structure verbale | *La structure d'un GV change, mais le sens est préservé (ex : elles s'affichent → elles sont affichées)* |
| **Changement de préposition** | | Une préposition est remplacée par une autre |
| **Changement de co-référence** | | Ajout/retrait d'un pronom ou changement possessif *(ex : la distance augmente → elle augmente)* |
| **Réordonnancement** | | Repositionnement d'un mot à un meilleur emplacement |
| **« Sac de mots »** | | Aucune explication linguistique ne fait sens |
| **Erreur de post-édition** | | Erreur du post-éditeur dans sa révision |
| **Changement stylistique** | | Modifications supplémentaires sans valeur ajoutée |
| **Divers** | | Toutes APE qu'on ne peut qualifier |

TABLE 3.1 – Typologie proposée pour la classification des APE pour le français.

## 3.2 Automatisation du processus d'analyse

Dans le but de pouvoir analyser en APE et de façon automatique des données post-éditées, nous avons développé un outil spécifique dénommé « SmartDiff », et dont l'architecture globale est représentée en figure 3.4 :

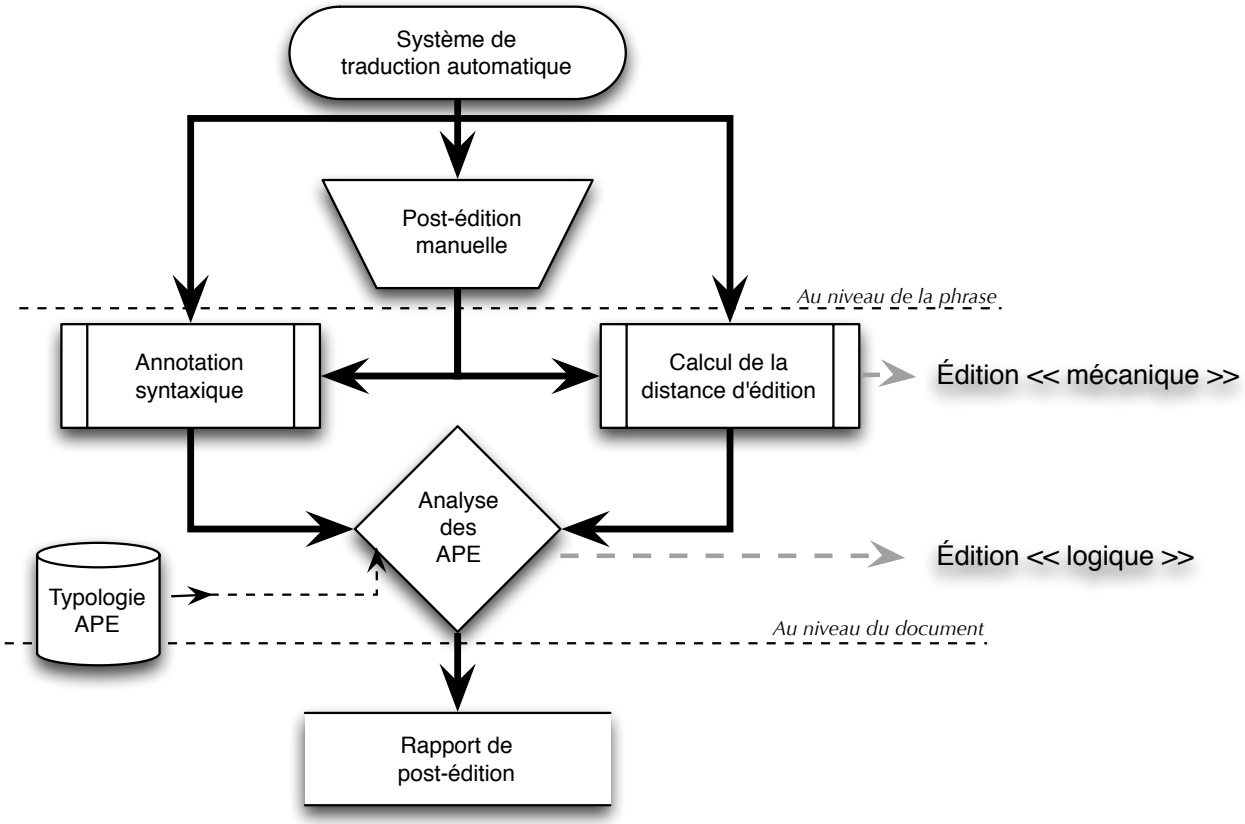


FIGURE 3.4 – Architecture de notre outil d'analyse en APE.

Notre système fonctionne à l'image d'un classifieur basé sur un ensemble de règles linguistiques représentant les APE que l'on veut identifier (ici pour le français). En entrée, il attend un ensemble de paires de phrases composées d'une hypothèse de traduction issue d'un système de TA, et de sa version finale post-éditée. Dans le cadre d'une analyse au niveau du document, notre outil nous permet d'éditer sur sa sortie un corpus annoté en APE à partir des traductions finales. Ce corpus peut-être utilisé par exemple pour extraire par la suite de nouvelles règles de traduction. De plus, nous pouvons également générer des statistiques sur l'analyse effectuée pour quantifier l'effort de post-édition et, implicitement, quantifier la réduction de l'effort de

post-édition envisageable si nous étions dans la possibilité d'utiliser rapidement ces informations. Chaque paire de phrases est donc analysée d'un point de vue « logique » en appliquant un processus en trois temps que nous allons maintenant aborder. Nous précisons à ce sujet que l'analyse est réalisée une paire de phrases à la fois, et indépendamment les unes des autres. Les résultats de cette analyse étant agrégés ensuite au niveau du document.

### 3.2.1 Protocole d'analyse en APE

Première étape de notre protocole : une annotation linguistique (classe grammaticale, partie du discours, lemmes et identification de structures élémentaires telles que les groupes nominaux ou verbaux) est réalisée sur l'hypothèse de traduction et sa version post-éditée. Cette annotation, réalisée avec l'analyseur syntactique [22] développé par la société SYSTRAN, est basée sur le format XLIFF (langage basé sur XML et utilisé à la standardisation de données localisées).

Dans un deuxième temps, l'hypothèse de traduction et sa version post-éditée sont alignées afin d'identifier tous les changements réalisés par le post-éditeur. Aux opérations d'éditions standards telles que l'insertion, la suppression, la substitution et le déplacement, nous avons introduit la notion d'une nouvelle opération appelée : *near* (« proche » en anglais), correspondant à une substitution d'un mot par un « cognat [23] ». Cette nouvelle opération est utile pour localiser les différences de morphologie potentielles sur un déterminant, nom, verbe ou un adjectif. La distance d'édition est donc estimée avec une version améliorée du TER de [Snover 2006].

La troisième et dernière étape de notre protocole analytique est consacrée à l'identification à proprement parlé des APE pour notre traduction post-éditée. Cette identification repose sur un ensemble de règles linguistiques prédéfinies comme évoqué précédemment, ces règles modélisant les classes de notre typologie pour le français.

### 3.2.2 Règles linguistiques

Dans le contexte de notre thèse Cifre, nous étions dans un cadre de prototypage destiné à valider notre approche. Pour cette raison, toutes les classes de notre typologie n'ont pas été modélisées par des règles linguistiques. Pour valider notre approche, nous nous sommes concentrés sur l'implémentation des classes les plus fréquentes (d'après nos observations). De fait, les classes implémentées dans la première version de SmartDiff furent les suivantes :

– **Changement dans un groupe nominal (GN)** – *changement de déterminant, changement de sens nominal, accord en nombre, modification de la casse, choix d'adjectif*

22. sous licence propriétaire, et bien qu'étant un élément majeur de SmartDiff, nous ne pouvons donner plus de détails sur cet outil.

23. se dit d'un mot ayant la même origine qu'un autre mot dans une autre langue

– **Changement dans un groupe verbal (GV)** – *accord grammatical, choix du sens verbal*
– **Changement de préposition**
– **Changement de co-référence**

Pour chaque classe implémentée (exceptée la classe « Divers » qui contient tous les cas non classés), nos règles sont définies en fonction des caractéristiques linguistiques de l'APE correspondante pour le français. Par exemple, une APE correspondant à un choix d'adjectif dans un GN est identifiée si :

1. il s'agit d'une substitution d'un mot par un autre mot ;
2. les deux mots sont des adjectifs ;
3. les deux lemmes sont différents ;
4. ils appartiennent tous deux au même GN.

### 3.2.3 Disponibilité de l'outil « SmartDiff »

Comme évoqué précédemment, notre outil d'analyse s'appuie sur l'analyseur syntactique et sur la description interne d'un document propre à la société SYSTRAN. Ces technologies étant propriétaires et par conséquent n'étant pas en libre accès, notre outil ne peut pas en l'état être mis à disposition de la communauté scientifique. Toutefois, plusieurs demandes ayant été formulées, nous envisageons de développer par la suite une version open-source de SmartDiff.

## 3.3 Données expérimentales

Le support de notre travail est un jeu de données provenant de processus de post-édition mis à notre disposition par les sociétés Autodesk [24] et Symantec [25], clients de la société SYSTRAN.

Le corpus pour lequel nous reportons nos résultats dans ce chapitre est un corpus de documentation technique de logiciel résultant d'un travail réel de révision : tout d'abord traduit automatiquement de l'anglais vers le français, il a été post-édité par quatre traducteurs professionnels différents, et francophones natifs [Plitt 2010]. Les post-éditeurs ont reçu des consignes de post-édition légère dont le but était donc de produire des traductions de qualité suffisant pour être publiables avec le moins d'édition possible, et interdisant des changements dus à des préférences stylistiques ou personnelles. Les post-éditeurs ont traité les hypothèses de traduction une par une, et dans le même ordre que celui dans lequel elles apparaissaient dans le document source d'origine. Aucune autre fonctionnalité asujettie à l'activité de post-édition, comme par exemple la recherche terminologique, n'a été effectuée.

Certaines hypothèses de traduction étaient issues d'un système de TAS basé sur le toolkit open-source Moses [26] [Koehn 2007b], et qui fut entraîné sur des données du domaine. Les autres hypothèses sont elles issues d'un système de TA à base de règles développé par SYSTRAN (que nous dénommerons par la suite « système RBMT »). À noter que les post-éditeurs n'ont pas été informés du type de système ayant fourni les traductions qu'ils ont révisées. Par ailleurs, et bien que notre objectif n'était pas de comparer le système RBMT et le système de TAS, il est intéressant de constater que notre approche s'applique à la fois sur l'une et l'autre des sorties de ces deux systèmes de technologies différente.

### 3.3.1 Annotation manuelle de référence

Notre travail a débuté par un l'annotation manuelle d'un sous-ensemble représentatif d'une centaine de phrases, en utilisant le format XML comme illustré en figure 3.5. L'objectif étant par la suite d'utuliser cette annotation manuelle comme référenciel pour évaluer notre annotations automatique avec SmartDiff.

La table 3.2 présente les résultats de notre analyse manuelle où pour chaque classe, les APE ont été annotées en fonction de notre typologie de la post-édition pour le français définie précédemment. La partie gauche correspond aux hypothèses de traduction en sortie du système RBMT, la partie droite *a contrario* correspond elle aux hypothèses de traduction en sortie du système de TAS. L'énumération et le ratio de représentation de chaque APE considérée sont

24. http ://www.autodesk.com
25. http ://www.symantec.com
26. http ://www.statmt.org/moses/

```
<annotations>
  <segment id="1">
    <source>
      Create factory layouts using the default system units.
    </source>
    <target>
      <pea type="verbStyle" id="1">Créez</pea>
      des présentations
      <edit type="agreement">de</edit>
      <pea type="nounMeaningChoice" id="2">famille</pea>
      avec
      <pea type="determinerChoice" id="3">des</pea>
      unités système par défaut.
    </target>
    <pstedt>
      <pea type="verbStyle" id="1">Créer</pea>
      des présentations
      <edit type="agreement">d'</edit>
      <pea type="nounMeaningChoice" id="2">usine</pea>
      avec
      <pea type="determinerChoice" id="3">les</pea>
      unités système par défaut.
    </pstedt>
  </segment>
</annotations>
```

FIGURE 3.5 – Exemple d'annotations en APE. L'APE est représentée dans le noeud <pea> simultanément dans l'hypothèse de traduction (<target>) et sa version post-éditée (<pstedt>).

| **Classe** | **Système RBMT** | | **Système de TAS** | |
|---|---|---|---|---|
| *Sous-classe* | #APE | %APE | #APE | %APE |
| **Groupe Nominal (GN)** | | | | |
| *Choix du déterminant* | 1 | 1.2% | 3 | 2.2% |
| *Choix du sens du mot* | 49 | 59% | 84 | 62% |
| *Accord en nombre* | 3 | 3.6% | 0 | 0% |
| *Changement de casse* | 19 | 23% | 37 | 27% |
| *Changement adjectif* | 2 | 2.4% | 1 | 0.7% |
| **Total** | **74** | **90%** | **125** | **92%** |
| **Groupe Verbal (GV)** | | | | |
| *Accord grammatical* | 3 | 3.6% | 2 | 1.5% |
| *Choix verbal* | 3 | 3.6% | 2 | 1.5% |
| **Total** | **6** | **7.2%** | **4** | **3%** |
| **Changement de préposition** | 1 | 1.2% | 0 | 0% |
| **Changement de co-reference** | 2 | 2.4% | 7 | 5% |
| **TOTAL** | **83** | **100%** | **136** | **100%** |

TABLE 3.2 – Résultats de l'analyse manuelle sur 100 phrases post-éditées. 90% des éditions concernent un GN pour les deux systèmes. Les changements terminologiques étant la principale source des APE avec 59% pour le système RBMT et 62% pour le système de TAS.

indiquées ainsi que leur couverture en nombre de mot. Nous observons ainsi que la principale catégorie d'APE identifiée est de type GN avec au moins 90% du total des annotations observées, et ce quel que soit le système de TA considéré. La sous-classe dominante de ces 90% et de loin, est celle des changements terminologiques avec 59% pour le système RBMT et 62% pour le système de TAS. Cette observation est intéressante puisque les groupe nominaux, et particulièrement dans le contexte d'une documentation technique, constituent un ensemble d'informations relativement facile à exploiter par la suite. En outre, nous avons comparé les résultats obtenus par le système RBMT avec ceux de l'analyse de concernant sa typologie de post-édition pour une PES. Dans [Dugast 2007], le type de modifications effectuées par un système SPE et la distribution déclarée est très similaire à notre analyse en APE. Cela montre que la couche SPE prépare le travail du post-éditeur, mais reste néanmoins limitée.

Autre résultat intéressant de cette analyse : celui sur la répétitivité des APE. Si on énumère combien de fois chaque APE est utilisée (une APE étant identifiée de manière unique par la modification qui est obtenue indépendamment du contexte), nous pouvons extraire les APE les plus fréquentes. Le résultat de cette observation est présenté dans les tableaux 3.3 et 3.4 :

| **Système RBMT** | | | |
|---|---|---|---|
| *avant* | après | **#occ.** | **%** |
| famille | usine | 96 | 20% |
| sol | atelier | 65 | 13% |
| plancher | sol | 11 | 2% |
| archive | actif | 9 | 2% |
| **Total (top-4)** | | **181** | **37%** |
| **TOTAL (toutes)** | | **488** | **100%** |

TABLE 3.3 – Top-4 des APE les plus fréquentes identifiées pour le système RBMT.

| **Système de TAS** | | | |
|---|---|---|---|
| *avant* | après | **#occ.** | **%** |
| archive | actif | 60 | 11% |
| superposition | calque | 39 | 7% |
| archive | ressource | 19 | 3% |
| sol | atelier | 13 | 2% |
| **Total (top-4)** | | **131** | **23%** |
| **TOTAL (toutes)** | | **558** | **100%** |

TABLE 3.4 – Top-4 des APE les plus fréquentes identifiées pour le système de TAS.

Dans ce contexte, une première réduction significative de l'effort de post-édition (de 37% à 23%) serait possible en ajoutant par exemple quatre entrées dans un dictionnaire pour le système RBMT, ou quatre entrées dans le modèle de traduction pour le système TAS. Cela montre que malgré une personnalisation du système de TA pour traiter de la documentation technique, il reste de façon évidente un certain écart terminologique. En effet, un système de TA, bien qu'entrainé et adapté au domaine, sera utilisé pour traduire de nouvelles données qui par définition n'étaient pas disponible auparavant. Plus important encore, cela donne aussi une idée du potentiel d'apprentissage à partir de ces nouvelles données.

### 3.3.2 Résultats de l'analyse automatique

Les résultats que nous avons obtenus avec SmartDiff, notre outil d'annotation automatisée en APE, sont présentés dans le tableau 3.5. La différence entre les résultats manuels et automatiques peut être expliquée, en dehors de potentielles erreurs d'analyse, par le fait que l'annotation humaine est effectuée sur un ou plusieurs mots, tandis que SmartDiff ne considère que le mot seul, qui analyse l'un après l'autre, en suivant le chemin d'édition préalablement calculé. En conséquence, certaines décisions sont prises trop tôt, surtout quand des propagations se produisent après la modification en cours (ce qui est le cas par exemple pour les déterminants où l'on observe une faible précision).

En s'intéressant aux changements terminologiques ayant un impact sur le sens, nous pouvons voir qu'une quantité importante de modifications terminologiques est détectée. Cela nous sera particulièrement utile pour adapter *a posteriori* nos systèmes de TA et ainsi éviter que ces erreurs apparaissent de nouveau.

Le tableau 3.6 montre la couverture en APE et les propagations observées sur notre corpus global. Sur les deux sorties des systèmes de TAS et RBMT, nous obtenons un taux de couverture d'environ 35% pour l'ensemble des classes traitées par notre typologie et les propagations implémentées dans cette première version de SmartDiff. Avec ces niveaux de Précision et de Rappel (voir section 1.6.2) pour notre analyse automatique, notre approche est sans nul doute améliorable et atteindre, à terme, un niveau d'analyse des plus utiles et intéressants.

| **Classe** *Sous-classe* | **Système RBMT** | | | | **Système de TAS** | | | |
|---|---|---|---|---|---|---|---|---|
| | #APE | #Match | %Prec. | %Rapp. | #APE | #Match | %Prec. | %Rapp. |
| **Grp Nominal (GN)** | | | | | | | | |
| *Choix du déterminant* | 15 | 1 | 7% | 100% | 16 | 1 | 6% | 33% |
| *Chgt. de sens nominal* | 89 | 35 | 40% | 71% | 97 | 69 | 71% | 82% |
| *Changement du nombre* | 3 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| *Changement de la casse* | 18 | 12 | 67% | 63% | 27 | 25 | 93% | 68% |
| *Choix d'adjectif* | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **Total** | **125** | **48** | – | – | **145** | **95** | – | – |
| **Grp Verbal (GV)** | | | | | | | | |
| *Accord grammatical* | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| *Chgt. de sens verbal* | 8 | 2 | 25% | 67% | 6 | 2 | 33% | 100% |
| **Total** | **9** | **2** | – | – | **8** | **2** | – | – |
| **Chgt. de préposition** | 34 | 0 | 0 | 0 | 53 | 0 | 0 | 0 |
| **Chgt. de co-reference** | 11 | 1 | 9% | 50% | 7 | 1 | 14% | 14% |

TABLE 3.5 – Résultats de l'analyse automatique des APE sur le même jeu de 100 phrases post-éditées. La colonne #APE indique le nombre de APE identifiées, la colonne #Match indique le nombre d'APE bien reconnues, et les deux dernières colonnes indiquent la Précision et le Rappel, pour chaque APE actuellement implémentée dans SmartDiff.

| | **Système RBMT** | | **Système de TAS** | |
|---|---|---|---|---|
| | #occ. | %couv. | #occ. | %couv. |
| **Nombre d'éditions** | 3231 | 100% | 3947 | 100% |
| **Nombre d'APE** | 1133 | 35% | 1340 | 34% |
| **Nombre de propagation** | 169 | 5,2% | 255 | 6,5% |
| **Nombre de déterminant** | 40 | 1,2% | 99 | 2,5% |
| **Nombre de préposition** | 102 | 3,2% | 97 | 2,5% |
| **Nombre de verbe** | 27 | 0,8% | 59 | 1,5% |

TABLE 3.6 – Couverture des APE et des propagation observées pour les systèmes RBMT et de TAS. La première colonne montre le nombre d'éditions tandis que la seconde indique le taux de couverture.

## 3.4 Conclusion

Dans ce troisième chapître, nous avons introduit et définit la notion d'« Actions de Post-Édition » comme étant des éditions « minimales » et « logiques », réalisées par le post-éditeur en opposition à des éditions dites « mécaniques ». Ces dernières correspondent aux éditions prises en compte par les métriques actuelles telles que comme BLEU (précision n-gramme), WER (insertion, suppression, substitution) et TER (WER + déplacement). Nous avons également proposé une typologie des APE suite à une série d'observations réalisées sur des données réelles mises à notre disposition par deux sociétés clientes de SYSTRAN. Ces observations sont le résultat d'une annotations manuelles des sorties de deux systèmes de TA qui venaient d'être post-éditées. Ces deux systèmes de TA étant par ailleurs de technologies différentes.

Dans un deuxième temps, nous avons proposé une procédure pour la détection automatique de ces APE, réalisée suivant les indications d'un chemin d'édition calculé entre une hypothèse de traduction et sa version post-éditée à l'aide de l'algorithme du TER. L'annotation automatique des APE actuellement implémentées, puisque nous nous sommes concentré sur les APE les plus fréquemment observées, permet d'obtenir un taux de couverture pour ces APE supérieur à 30%. Des taux de Rappel et de Précision intéressants évalués sur notre annotation manuelle ont également été observés. Ces résultats expérimentaux laissent entendre qu'une intégration de ces nouvelles connaissances permettrait théoriquement, dans le cas d'une adaptation incrémentale du système de TA, de réduire de façon significative l'effort de post-édition sur un même corpus.

### Le facteur humain

Notre approche s'applique aux flux de travail où la qualité de TA initiale est élevée et où les post-éditeurs sont invités à effectuer une post-édition légère. Cette situation ne s'applique pas aux tâches générales de traduction d'usage. À titre d'exemple, [Martinez 2003] donne les conseils suivants pour les post-éditeurs de brochures marketing :

> " it to look for synonyms [in order to] avoid the repetitive style caused by MT consistency, to simulate the performance of a human translator... "

> *« il faut chercher des synonymes [pour] éviter le style répétitif causé par l'uniformité de la TA, pour simuler les performances d'un traducteur humain »*

Ces instructions seraient certainement nuisibles à notre processus d'extraction automatique car ils réduiraient la facilité d'apprentissage.

À noter que, même dans ce contexte de post-édition dite « légère » sur des traductions de haute qualité, des commentaires informels de post-éditeurs montrent que l'apprentissage de

leurs retours est un élément clé pour les garder motivés.

Nous travaillons sur plusieurs améliorations : en particulier sur le raffinement des modèles utilisés pour détecter les APE et la capacité à faire face à des modifications de plusieurs mots. Notre prochain objectif est d'utiliser l'analyse en APE pour améliorer la qualité de traduction en tenant compte des retours utilisateurs, et plus particulièrement des éditions récurrentes. Cela devrait en toute logique permettre de réduire l'effort et ainsi le coût lié à la pratique de la post-édition, mais aussi la répétitivité de la tâche, qui sont des points importants dans un contexte industriel. L'adaptation incrémentale d'un système RBMT par exemple pourrait être réalisée en incluant un dictionnaire dynamique supplémentaire pour les nouvelles règles et la terminologie unique. Les systèmes de TAS quant à eux sont plus difficiles à adapter à la volée puisque leurs modèles sont généralement entraînés sur de grandes quantités de données impliquant un temps de calcul conséquent. Nous avons donc réfléchi à proposer une approche nous permettant de bénéficier de retours issus de la post-édition, et de manière incrémentale, rapide et efficace. C'est là tout le sujet du chapitre 4 que nous allons aborder.

**Publication**

Ces travaux sur l'analyse qualitative de données post-éditées ont donné lieu à la publication de l'article [Blain 2011] qui fut présenté oralement au Machine Translation Summit XIII organisé en 2011 à Xiamen, en Chine.

# Chapitre 4

# Adaptation incrémentale d'un système automatique statistique

## Sommaire


**4.1 Travaux connexes dans la littérature . . . . . 91**

**4.2 Protocole d'adaptation incrémentale . . . . . 93**

4.2.1 Combinaison d'alignements mot-à-mot . . . . . 94

**4.3 Évaluations expérimentales . . . . . 98**

4.3.1 Données d'apprentissage . . . . . 98

4.3.2 Apprentissage du système de référence . . . . . 99

4.3.3 Temps de calcul *vs.* Qualité de traduction . . . . . 100

4.3.3.1 Protocole expérimental . . . . . 101

4.3.4 Combinaison des modèles de traduction . . . . . 103

**4.4 Conclusion . . . . . 110**

Un système de traduction est assujetti à certaines limitations : en premier lieu, les ressources spécifiques disponibles à un instant *t*, et à partir desquelles il fut entraîné, peuvent s'avérer moins appropriées à l'instant *t+1*. Conséquemment, il est nécessaire de procéder régulièrement à une mise à niveau de ce système par l'intégration de nouvelles ressources. On parle alors d'adapter le système dans le temps avec de nouvelles ressources. Le mot « adaptation » ayant ici une signification d'évolution, on le fait évoluer en lui intégrant de nouvelles connaissances.

Or, dans un contexte de post-édition comme le notre, où nous avons vu au chapitre précédent qu'une façon de réduire l'effort de post-édition était notamment de s'intéresser à l'aspect répétitif de cette tâche, enrichir rapidement et efficacement de nouvelles connaissances apporterait le modèle de traduction d'un système apporterait une réponse très puissante. Nous avons également pu observer au chapitre 3 que les éditions les plus fréquentes concernaient des changements de type GN, et plus particulièrement de la terminologie. Si nous prenions l'exemple d'un système hybride comme cité précédemment : associtation d'un système à base de règles linguistiques et d'un système de TAS pour effectuer une PES. Dans cette configuration, le système de TAS peut-être utilisé pour renforcer les ressources linguistiques du système à base de règles comme suggéré par [Dugast 2007] en réalisant notamment une extraction automatique d'une terminologie bilingue [Daille 1994, Déjean 2002]. L'exploitation de nouvelles connaissances terminologiques se faisant généralement par l'utilisation d'un dictionnaire spécialisé. Ce qui est techniquement assez simple et donc rapide à mettre en oeuvre.

Dans le cadre cette fois d'un système de TAS, la façon la plus directe de l'adapter est encore de procéder à un réentraînement complet sur l'ensemble des données, ce qui est très coûteux en temps de calcul et en ressources matérielles. Nous nous sommes donc intéressés dans cette thèse à proposer une nouvelle méthode permettant d'adapter de façon incrémentale un système de TAS sans qu'il soit nécessaire de relancer tout la procédure d'entrainement au complet.

L'adaptation incrémentale est cependant une tâche complexe : comment adapter correctement un système ? Le résultat d'une bonne adaptation ne doit pas résulter en la dégradation des performances du système adapté ou la régression sur des données similaires. Cela signifie également que les connaissances qui sont ajoutées au système ne doivent pas venir supplanter celles qu'il possède déjà.

## 4.1 Travaux connexes dans la littérature

Une recherche dans la littérature nous a permis de faire ressortir deux antécédents.

Le premier d'entre eux, qui est approche la plus semblable à la notre, fut proposée par [Hardt 2010] par le biais d'un algorithme de réentraînement incrémental d'un système statistique (basé sur les séquences de mots), là aussi dans un contexte de post-édition. Ils proposent d'extraire de nouveaux syntagmes à partir d'« alignements approximatifs », lesquels sont obtenus en utilisant une version « modifiée » de l'outil Giza++ développé par [Och 2000, Och 2003b] (nommé « Gizapp » par la suite). À partir d'un alignement « 1-vers-1 », initialisé par défaut entre les mots aux mêmes positions pour une hypothèse de traduction et sa traduction de référence (*c.-à-d.* le mot à la position $i$ de l'hypothèse de traduction est aligné avec le mot à la position $i$ de la référence, et ainsi de suite), [Hardt 2010] effectue une mise à jour itérative de ces alignements tant que des améliorations sont observées. Dans la pratique, cette mise à jour est effectuée via un algorithme glouton [27] pour trouver l'alignement optimal local. Ainsi, toutes les positions de départ ayant une seule liaison sont essayées, et le changement de liens uniques qui produit l'augmentation de probabilité la plus forte en fonction du modèle IBM 4 de Giza-pp est maintenue. L'alignement résultant est amélioré avec deux simples étapes de post-traitement : (i) chaque mot inconnu du côté de la source est aligné avec le premier mot non-aligné inconnu du côté cible, (ii) les paires de positions non alignées, qui se trouvent entourées par des alignements correspondants, sont automatiquement alignées. De plus, [Hardt 2010] affirme que :

> " to be practical, incremental retraining must be performed in less than one second "
>
> *« pour être exploitable en pratique, un cycle de réapprentissage doit pouvoir être exécuté en moins d'une seconde »*

Dans la suite de ce chapitre nous présentons un algorithme d'alignement mot-à-mot efficace s'appuyant en partie sur l'algorithme du TER pour évaluer la distance d'édition entre deux phrases. Alors que [Hardt 2010] annonce n'avoir besoin que de quelques secondes pour aligner environ deux mille paires de phrases, nous verrons dans la seconde partie de ce chapitre consacré aux résultats expérimentaux que l'algorithme que nous avons développé est encore bien plus rapide. Pour ce faire, et comme faisant partie de nos expériences pour valider notre approche, nous avons comparé notre approche avec l'utilisation de l'outil librement disponible dénommé « *inc-Giza-pp* [28] », qui se veut une version incrémentale de *Gizapp*. Cet outil est précisément destiné à l'injection de nouvelles données dans un système statistique sans qu'il ne soit nécessaire de

27. se dit d'un algorithme itératif dont le principe est de faire choix optimal local, dans le but d'obtenir un résultat optimum local(source : Wikipedia)

28. http ://code.google.com/p/inc-giza-pp/

relancer entièrement la procédure d'alignement de mots. À notre connaissance, cette méthode était à l'état de l'art dans le domaine au moment où nous avons réalisé nos expériences. Nous l'avons donc utilisé comme référentiel pour l'évaluation de nos résultats.

Le deuxième antécédent bibliographique, proposé par [Levenberg 2010], présente un processus d'adaptation incrémentale fondé sur une version en ligne de l'algorithme Espérance-Maximisation (EM). Cette approche adaptée pour de grandes quantités de données ne l'est pas vraiment pour le contexte particulier de la post-édition dans lequel nous évoluons. Tout comme [Hardt 2010], nous proposons un processus d'adaptation progressive qui est plus orienté vers le traitement en temps réel, et donc sur de petites quantités de données. Nous n'avons donc pas comparé les résultats expérimentaux de notre approche avec celle de Levenberg.

Dans la suite de ce chapitre, nous abordons en détails le protocole expérimental que nous avons élaborer, et destiné à extraire des données post-éditées les informations nécessaires à l'adaptation d'un système de TAS pour un domaine scientifique particulier. Nous présenterons également les données d'apprentissage que nous avons utilisé avant de présenter nos résultats expérimentaux.

## 4.2 Protocole d'adaptation incrémentale

Nous avons développé un protocole d'alignement séquentiel au niveau des mots qui s'opère en trois temps. Ces trois étapes sont liées entre elles par un algorithme d'alignement mot-à-mot qui nous permet d'aligner une phrase source et sa traduction de référence, en utilisant une hypothèse de traduction générée automatiquement par un système PBMT. De ce fait, nous sommes en mesure d'extraire de nouvelles paires de séquences de mots, lesquelles correspondent aux erreurs du système, et à partir desquelles nous allons l'adapter tel qu'illustré par la figure 4.1 :

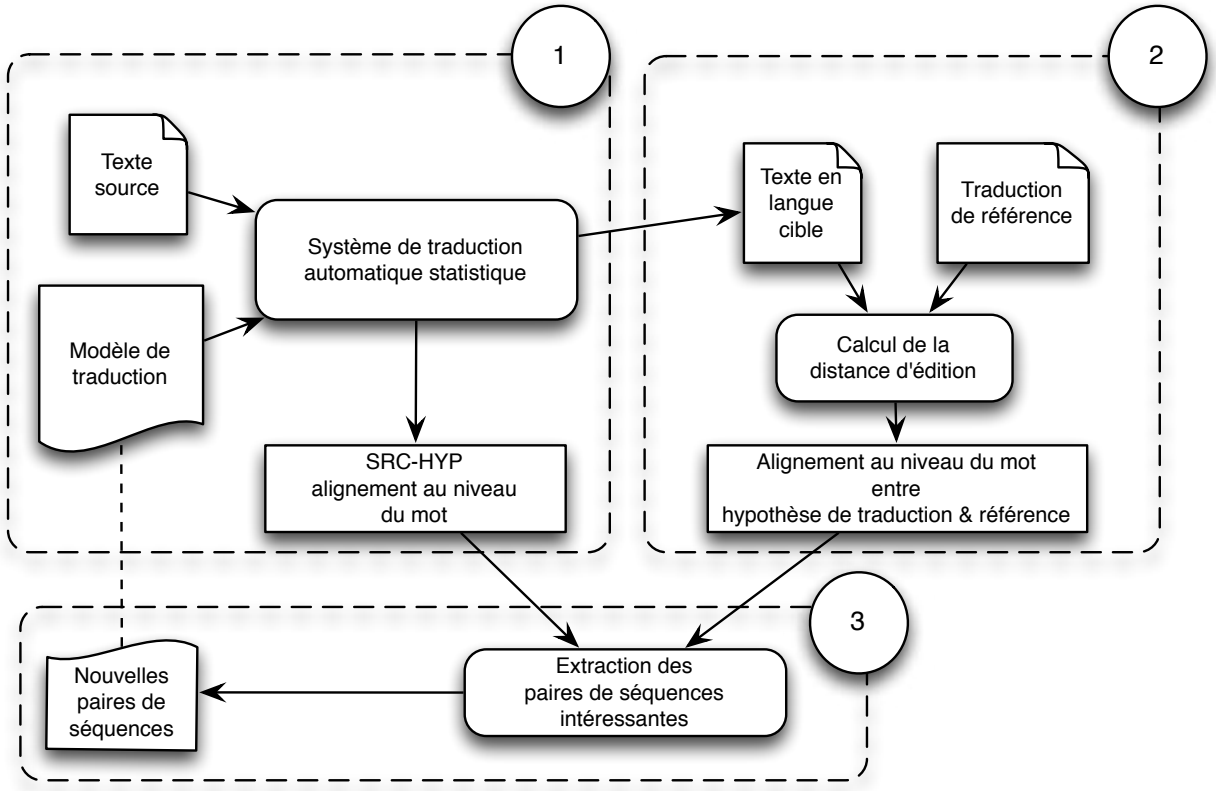


FIGURE 4.1 – Protocole d'alignement séquentiel qui s'opère en trois temps.

En trois étapes que nous dénommerons « traduction », « post-édition » et « adaptation », telles que :

1. **Traduction** – *Chaque phrase source est traduite en langue cible avec un alignement bilingue mot-à-mot source-vers-traduction* ;

2. **Post-édition** – *La distance d'éditions entre l'hypothèse de traduction et sa traduction de référence est calculée, et un alignement monolingue traduction-vers-référence est déduit du chemin d'édition* ;
3. **Adaptation** – *Les deux alignements précédents sont combinés afin d'en déduire un alignement bilingue source-vers-référence. De cet alignement sont extraites les informations qui serviront à l'adaptation du système de traduction*.

À l'instar de [Hardt 2010], c'est par le biais de l'association « traduction + post-édition » que nous identifions les informations bilingues qui sont ensuite extraites et à partir desquelles nous ferons évoluer notre système :

### 4.2.1 Combinaison d'alignements mot-à-mot

Nous allons maintenant revenir plus en détails sur les séquences d'alignement qui nous guident vers l'extraction des informations à partir desquelles nous faisons évoluer notre système.

#### Traduction : alignement source-vers-traduction

L'élément central de notre approche est la génération automatique d'une hypothèse de traduction à partir d'une phrase source en entrée de notre processus. À partir de cette hypothèse de traduction, nous pouvons générer un alignement entre notre source et une référence de traduction afin de mettre en exergue les lacunes de notre système. Le principe de cette idée est illustré par la Figure 4.2. Pour se faire, nous avons entraîné un système de TAS basé sur le toolkit Moses, lequel nous permet par la suite de générer les alignements mot-à-mot entre la phrase source et l'hypothèse de traduction en sortie du décodage. Cette information d'alignement représente la première étape de notre combinaison d'alignement.

#### Analyse : alignement traduction-vers-référence

Une telle approche nécessite d'avoir à disposition une interface de post-édition ainsi qu'une équipe de traducteurs humains dont l'objectif serait, au travers de cette interface, de corriger les hypothèses de traduction générées automatiquement. N'ayant malheureusement pas eu la possibilité de bénéficier de telles ressources, nous avons opté pour une simulation de la phase de post-édition (comme [Hardt 2010]) afin de nous focaliser sur l'extraction des informations qui nous intéressent. Pour ce faire, l'idée est toute simple : à partir de corpus bilingues alignés, nous exploitons le côté cible de notre corpus parallèle comme post-éditions des hypothèses de

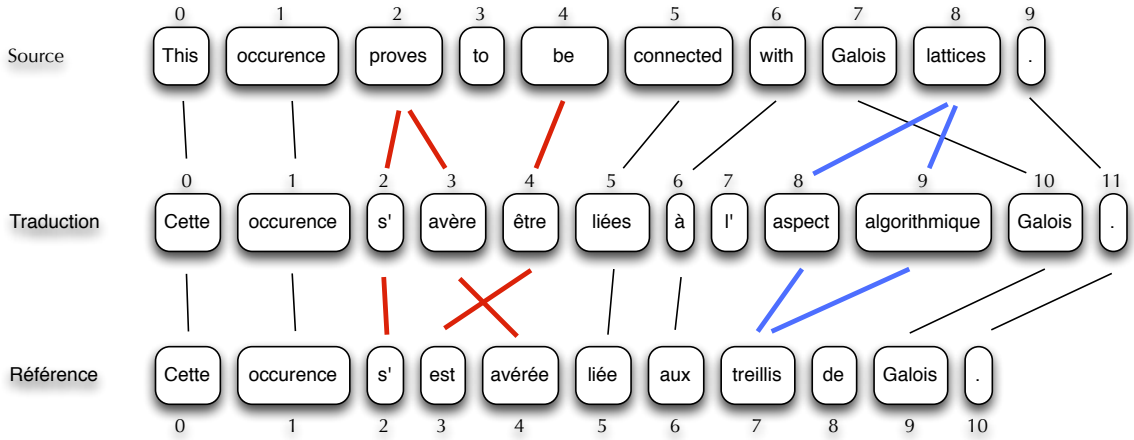


FIGURE 4.2 – Exemple d'alignement source-vers-référence utilisant l'hypothèse de traduction comme « pivot ». Ici est matéralisé ce que l'on veut apprendre : que « lattices » se traduit par « treillis » au lieu de « aspect algorithmique ». Tandis qu'en rouge, est matérialisé non pas une erreur de traduction du système de TA, mais ce que l'on considère comme étant un changement stylique de la part du post-éditeur.

traduction en sortie de notre système de TA, et ainsi en extraire un certain nombre d'informations, qui serviront ensuite à adpater notre systèmes en retour. Les TA considérées ici comme étant nos versions post-éditées ne sont pas nécessairement les références qu'un post-éditeur humain aurait produites, et peuvent s'avérer plus « éloignées » de celles-ci, impliquant ainsi un biais potentiel avvec une distance et une difficulté d'apprentissage plus importantes. Pour palier à ce problème, nous aurions apprécié avoir la possibilité de travailler sur des cas concrets en collaboration avec des traducteurs humains, mais comme évoqué précédemment, ce ne fut pas le cas. Toutefois, en procédant de cette manière, nous avons pu développer et valider notre approche ce qui en soit était notre principal objectif.

Pour obtenir un alignement mot-à-mot entre notre hypothèse de traduction et notre traduction de référence, nous avons utilisons l'algorithme du TER [Snover 2006] et plus particulièrement le chemin d'éditions de celui-ci. De ce chemin d'édition, nous déduisons un alignement traduction-vers-référence, mais nous ne le considérons pas dans son ensemble. Nous n'explorons en effet qu'une partie représentative de celui-ci en nous concentrant sur ce que nous considérons comme étant les erreurs commises par notre système de traduction. Comme annoncé précédemment, nous soutenons que ce qui est identifié comme étant « aligné » en sortie du TER correspond aux connaissances de notre système. Inversement, nous considérons que les éditions correspondantes à des substitutions représentent les lacunes de notre système. C'est donc à partir de ces substitutions que nous allons par la suite extraire nos informations.

Cette méthodologie peut-être bien évidemment être étendue à l'utilisation du TER-Plus [Snover 2009], cette extension du TER qui utilise les paraphrases, les stemms et les synonymes afin d'obtenir un meilleur alignement mot-à-mot (voir section 1.6.2).

**Adaptation : alignement source-vers-référence**

Considérant l'hypothèse de traduction comme un « pivot » pour aligner la phrase source et sa référence, nous avons conçu un algorithme (voir Algorithme 1) d'alignement mot-à-mot. Dénommé « Bitext-tokaligner [29] », il combine les alignements source-vers-traduction et traduction-vers-référence présentés précédemment, et en déduit un chemin d'alignement source-vers-traduction-vers-référence qu'on résumera alors en un **alignement source-vers-référence**.

**Extraction des séquences de mots pour la création du modèle de traduction**

Les modèles de traduction sont créés par extraction et pondération de séquences de mots extrait de corpus bilingues, comme évoqué au chapitre 1, section 1.4.2.4. L'extraction de ces séquences de mots repose sur les alignements mot-à-mot réalisés en amont. La procédure standard d'entrainement du toolkit Moses, que nous utilisons, repose en principe sur des alignements générés avec Giza, qui implémente les modèles IBM (voir section. 1.4.2.2).
Sans modifier le reste de la procédure d'entrainement [30] de Moses, nous substituons les alignements Giza par les alignements issus de notre approche. De cette manière, nous sommes en mesure de comparer l'efficacité de deux méthodes.

29. Implémentation open-source de « Bitext-toaligner » : https ://github.com/fredblain/bitext-tokaligner
30. Voir script « train-model.perl ».

**Data**: Alignements mot-à-mot ($Al_{src-trad}$) pour chaque paire (source,traduction)
**Data**: Chemins d'édition ($C_{ed}$)pour chaque paire (traduction, référence)
**Result**: Alignements mot-à-mot pour chaque paire (source-référence)
**while** *pour chaque paires [(source,traduction) ;(traduction,référence)] à traiter* **do**
  Aligner(source,traduction,$Al_{src-trad}$) ;
  Aligner(traduction, référence) ; // tout à 1 par défaut.
  // on applique l'ensemble des mouvements identifiés (*shift* du TER)
  // que ce soit pour 1 mot ou un bloc de mots.
  **foreach** *déplacement ($D$) dans le chemin d'édition ($C_{ed}$)* **do**
    Appliquer($D$) ;
    MAJ-indices($C_{ed}$) ;
  **end**
  $ind_{trad} = 0$ ; // itérateur positionné sur le 1er mot de la traduction
  $ind_{ref} = 0$ ; // itérateur positionné sur le 1er mot de la traduction post-éditée
  **foreach** *édition ($E$) du chemin d'édition ($C_{ed}$)* **do**
    **if** *$E$ est un alignement ou une substitution* **then**
      Aligner($ind_{trad}$, $ind_{ref}$) ;
      $ind_{trad} + 1$ ;
      $ind_{ref} + 1$ ;
    **end**
    **if** *$E$ est une insertion* **then**
      Aligner($ind_{trad}$, $ind_{ref}$) ;
      $ind_{trad} + 1$ ;
    **end**
    **if** *$E$ est une suppression* **then**
      $ind_{ref} + 1$ ;
    **end**
  **end**
  **foreach** *mot de la référence ($M_{ref}$)* **do**
    **foreach** *mot de la traduction ($M_{trad}$) aligné avec ($M_{ref}$)* **do**
      **if** *($M_{trad}$) est aligné avec un mot de la source ($M_{src}$)* **then**
        Ajouter($M_{src}$,$M_{ref}$) dans $Al_{src-ref}$ ;
      **end**
    **end**
  **end**
  Afficher($Al_{src-ref}$) ;
**end**

**Algorithm 1:** Algorithme d'alignement pivot, mot-à-mot, entre une prhase source et sa version de référence (traduction post-éditée). Cet algorithme exploite les alignements Moses entre la source et une hypothèse de traduction, et un chemin d'édition généré par TERcpp entre l'hypothèse de traduction et sa version post-éditée.

## 4.3 Évaluations expérimentales

Dans cette partie du chapitre, nous présentons les expériences réalisées dans le but de valider notre méthodologie. Pour rappel, une partie de ces expériences passe par une comparaison à l'utilisation du toolkit *inc-Giza-pp*, lequel est considéré comme étant l'état de l'art pour l'entraînement incrémental de systèmes statistiques.

Dans nos premières expériences, chaque système utilise un seul modèle de traduction qui est mis à jour incrémentalement, c'est-à-dire de re-entraîné complètement après chaque itération. Pour les résultats que nous présentons ci-après, « *inc-Giza-pp* » désignera le système s'appuyant sur l'outil du même nom, tandis que « OnlineAdapt » désignera le système conçu à partir de notre approche.

### 4.3.1 Données d'apprentissage

Les expériences ont été réalisées à partir de données rendues disponibles grâce au projet COSMAT. Un des objectifs de ce projet concerne l'utilisation de données post-éditées l'adaptation en continue du système de TA. À noter que pour les expériences que nous présentons ci-après, nous avons considéré des traductions de l'anglais vers le français.

Nous disposions de trois corpus parallèles pour entraîner notre modèle de traduction : deux corpus génériques et un corpus du domaine pour l'adaptation. Les deux premiers corpus sont le corpus EUROPARL et le corpus NEWS COMMENTARY, dont les statistiques sont donnés en table 2.2. Ceux-ci ont été utilisés pour entraîner notre système de TAS de référence.
Le troisième corpus, que nous nommerons « absINFO », contient cinq cents mille mots extraits à partir des résumés d'articles scientifiques identifiés comme appartenant au domaine « Informatique ». Les informations sur les sous-domaines, également disponibles (réseaux, intelligence artificielle, base de données, informatique théorique, . . .), n'ont pas été exploitées.

Ce corpus en domaine a été découpé en trois sous-corpus comme illustré en figure 4.3 :

- **absINFO.corr.train** est composé de 350k mots et est utilisé pour simuler la post-édition utilisateur ;
- **absINFO.dev** est un ensemble de 75k mots utilisé pour l'optimisation du modèle de traduction ;
- **absINFO.test** un autre ensemble de 75k mots utilisé comme corpus de test pour surveiller la non-régression des performances du système sur des données similaires, mais pas utilisées pendant le processus d'adaptation.

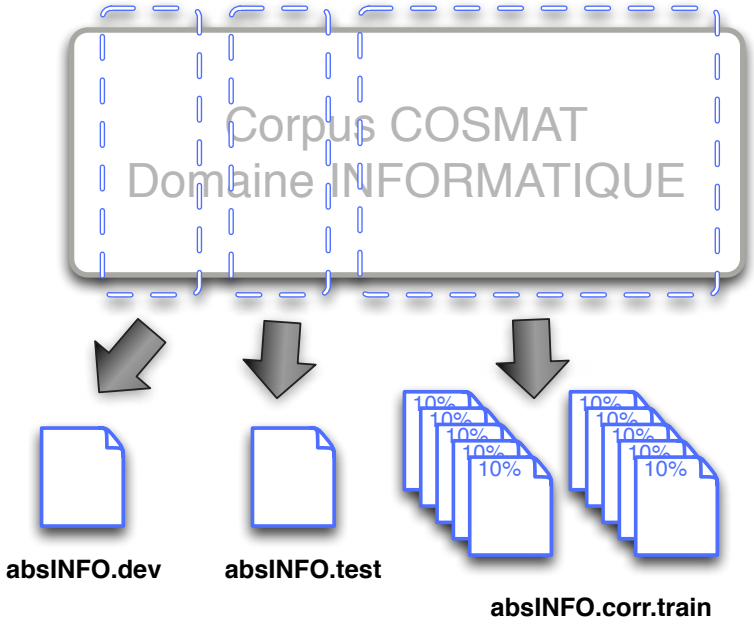


FIGURE 4.3 – Le corpus « absINFO » du projet COSMAT est découpé en trois sous-corpus pour permettre la simulation d'un procéssus de post-édition et d'une adaptation d'un système de TAS.

Par ailleurs, et ce afin de simuler au mieux un processus de post-édition séquentiel, le corpus absINFO.corr.train a été découpé en 10 sous-corpus (environ 1.5k phrases avec 35k mots chacun). Cela correspond assez bien à la mise à jour d'un système de traduction après la post-édition d'un document entier.

### 4.3.2 Apprentissage du système de référence

Les systèmes de TAS de référence ont été construits en suivant la procédure standard et Giza-pp pour l'alignement mot-à-mot. Afin d'utiliser plus tard *inc-Giza-pp*, la version incrémentale de Giza-pp, nous devions entraîner un système de référence spécifique utilisant l'option « Modèle de Markov caché » (« Hidden Markov Model » (HMM) en anglais) pour le modèle d'alignement de mots. De plus, pour faire une comparaison équitable entre les deux techniques d'adaptation, les systèmes de référence et les systèmes suivants ont été entraînés sur les mêmes données et ajustés (tuning) avec l'algorithme MERT [Och 2003a] (voir section 1.4.2.7), et ce à partir de la même paramétrisation.

Les systèmes de TAS de référence ont obtenu sur le corpus de développement un score BLEU de 35,27 points pour le système incrémantal sans *Gizapp* (notre approche) et 35,32 points pour le sytème incrémental avec *Gizapp*. Sur le corpus de test, ils ont obtenu un score BLEU

de 31,89 points pour le premier et 32,27 points pour le second. Bien que la procédure eut été différente pour les raisons que nous venons d'évoquer, ces systèmes de référence sont de qualité comparable.

### 4.3.3 Temps de calcul *vs.* Qualité de traduction

Par la suite, nous continuerons à comparer ces deux méthodes d'adaptation incrémentale par rapport à leurs scores BLEU respectifs obtenus en ajoutant les données extraites supplémentaires. Nous rapporterons également le temps de calcul nécessaire pour effectuer ces alignements mot-à-mot.

#### Méthode standard

Tout d'abord, il faut prétraiter les données pour une utilisation par le toolkit Giza-pp. Il s'agit de mettre à jour les fichiers de vocabulaire, de convertir les phrases dans le format *snt* de Giza-pp, puis de mettre à jour le fichier de cooccurrences. Ensuite, Giza-pp est exécuté pour mettre à jour et calculer les alignements sur les nouvelles données. Cette opération est réalisée dans les deux sens, source-vers-traduction et traduction-vers-source. Pour chaque itération de notre expérience, nous avons calculé que ce processus prend 14 minutes environ, le protocole d'alignement étant lui composé de plusieurs séquences (pour plus de détails, se reporter au paragraphe « Incremental Training » de la section « Advanced Features » dans le manuel utilisateur de Moses [31]).

#### Notre approche

Pour notre système *OnlineAdapt*, le temps requis pour réaliser l'alignement source-vers-traduction est considéré comme nul, puisque réalisé durant le processus de traduction par le système de TA. Le chemin d'édition entre l'hypothèse de traduction et sa version de référence est calculé en utilisant une implémentation C++ du TER, rapide et librement accessible [32]. Cet outil peut aligner 35k mots en 3 secondes environ (correspondant au 1,5k phrases des sous-corpus de absINFO.corr.train). La combinaison d'alignements entre la traduction source et sa référence, décrite à l'algorithme 1, nécessite elle moins d'une seconde. Soit seulement quelques secondes pour générer les alignements source-vers-référence pour 35k mots. Ce qui est, comparé à l'approche standard, une réduction très significative.

31. Disponible sur le web : http ://www.statmt.org
32. http ://sourceforge.net/projects/tercpp/

#### 4.3.3.1 Protocole expérimental

Pour l'ensemble de nos expériences, quatre systèmes ont donc été construits :

1. « Gizapp » – pour chaque sous-corpus de absINFO.corr.train (10%, 20%, 30%... 100%), toutes les données d'apprentissage disponibles sont concaténées et le processus d'entraînement complet est effectué, ce qui comprend un nouvel alignement mot-à-mot sur toutes les données d'apprentissage. Nous considérons cela comme la limite supérieure de la performance que nous pourrions atteindre avec une procédure d'adaptation incrémentale. Toutefois, cette procédure est très coûteuse en temps de calcul.

2. « inc-Giza-pp » – les sous-corpus des données d'apprentissage absINFO.corr.train sont ajoutés en utilisant la version incrémentale de Giza-pp. Cela s'est traduit dans nos expériences par une légère diminution du score BLEU sur les données d'apprentissage et une performance tout à fait instable sur les données de test.

3. « OnlineAdapt » – apprentissage incrémental basé sur la notre approche présentée dans ce chapitre. Nous avons uniquement utilisé le système de référence pour traduire les données d'adaptation que nous ajoutions. C'est à dire qu'avec notre système, nous traduisions 10%, puis 20%, puis 30% des données, et ainsi de suite.

4. « inc-OnlineAdapt » – similaire à *OnlineAdapt*, mais nous utilisons le système adapté à l'itération précédente pour traduire les données additionnelles. C'est à dire que le système résultant d'une adaptation avec 10% des données sert à traduire les 10% suivants, et ainsi de suite. La différence avec *OnlineAdapt* se situe dans le fait que nous utilisins le système qui venait d'être adapté pour la suite, alors que pour *OnlineAdapt*, le système de traduction reste le même. De cette façcon, nous pouvons déterminer l'impact réel de l'ajout des données du domaine.

Nous avons dans un premier temps procédé à l'entraînement d'un système pour chaque sous-corpus que nous avions. Partant de nos systèmes de référence, nous avons entraîné ces systèmes sur respectivement 10%, 20%, 30%... 100% de absINFO.corr.train, notre corpus en domaine. Grâce à ces systèmes partiellement adaptés au domaine, nous avions une sorte de « limite haute », d'oracle, pour l'adaptation incrémentale. Les résultats de cette expérience sont représentés graphiquement à la Figure 4.4. Ensuite, nos systèmes furent entrainés de façon itérative et incrémentale. Là encore, partant des systèmes de référence, nous avons ajouté de façon incrémentale 10% de notre corpus absINFO.corr.train. Les résultats en score BLEU résultant de

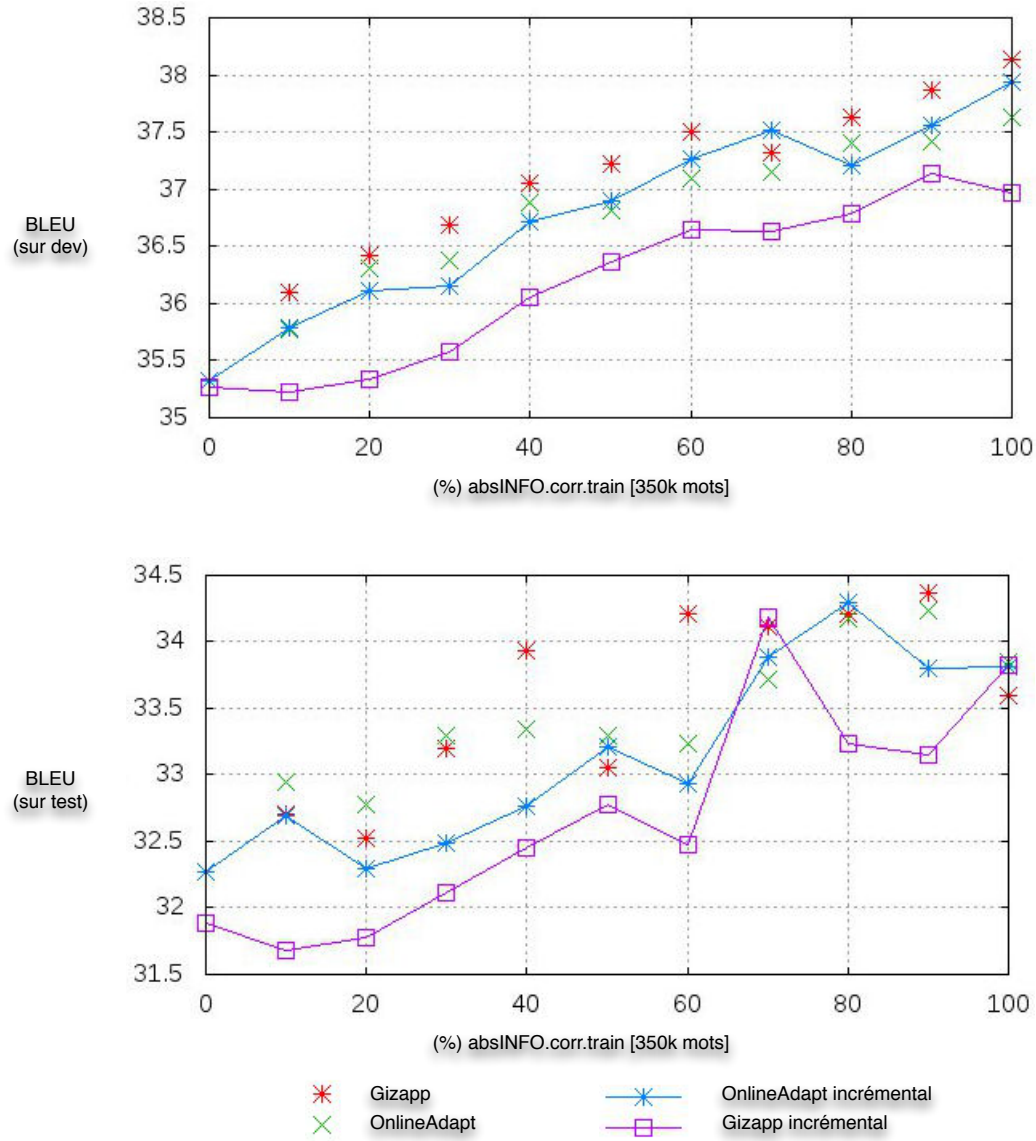


FIGURE 4.4 – Scores BLEU obtenus respectivement sur les corpus de développement et de test pour nos quatre systèmes : « Gizapp », « inc-Gizapp », « OnlineAdapt » et « inc-OnlineAdapt ».

cette approche sont représentés graphiquement à la Figure 4.4 (courbes dénommées « inc-Gizapp » et « (incremental) OnlineAdapt »).

### Discussion

L'approche proposée pour obtenir des alignements mot-à-mot permet d'obtenir des scores BLEU légèrement supérieurs, à la fois sur le développement et le corpus de test, mais surtout, permet de les obtenir de façon beaucoup plus rapide.

Les larges variations pour toutes les approches sur le corpus de test peuvent s'expliquer par deux raisons potentielles. D'une part cela pourrait être due à la caractéristique même du corpus absINFO.corr.train corpus. Pour rappel, il fut créé à partir de résumés d'articles scientifiques du domaine Informatique qui furent sélectionnés de façon aléatoire. Conséquemment, un sous-corpus représenté de façon prédominante dans un sous-corpus de absINFO pourrait ne pas être pas représenté dans le corpus de test.

D'autre part, cela pourrait être due à l'utilisation d'un seul et unique modèle de traduction. Comme expliqué précédemment, ce modèle de traduction est mis à jour à partir de nouvelles paires de séquences de mots extraites à chaque itération. Parce que nous nous sommes seulement intéressés par les types d'édition correspondant à « aligné » et « substitué » durant l'analyse de la distance d'édition, les paires de séquences de mots extraites peuvent être génériques ou en domaine. Ajoutés à toutes les entrées déjà présentes dans le modèle de traduction, ces nouvelles séquences de mots perturbent la distribution de probabilités. Cela peut également expliquer que nos systèmes incrémentaux soient moins performants que nos systèmes non évolutifs (ceux que l'on a appelés « oracles »), pour lesquels la distribution de probabilités est mieux évaluée.

Une autre possibilité pourrait être alors d'utiliser deux modèles de traduction, à l'instar de ce qu'a proposé [Hardt 2010]. De cette façon, nous pouvons rapidement créer un modèle de traduction à partir des alignements de mots pour les données additionnelles. Nous obtenons alors un modèle de traduction que l'on pourrait qualifier de « générique » et un second modèle en domaine hyper-spécialisé, composé uniquement de données du domaine.

### 4.3.4 Combinaison des modèles de traduction

Dans cette section, nous présentons les résultats expérimentaux obtenus en combinant plusieurs modèles de traduction. Les techniques décrites dans les sections précédentes peuvent considérablement accélérer le processus alignement de mots, par rapport à l'exécution progressive Giza-pp, mais nous avons encore besoin de créer un nouveau modèle de traduction sur toutes les données. Ceci peut demander un temps de calcul conséquent, c'est pourquoi nous proposons de ne créer qu'un nouveau modèle de traduction sur les seules données nouvellement ajoutées, et de l'associer à un modèle de traduction original (modèle « générique »).

#### Modèles avec repli (« back-off »)

Le toolkit Moses supporte plusieurs modes permettant l'utilisation de multiples tables de traductions. Nous explorons dans un premier temps le mode repli (voir 1.4.1.2) qui favorise le modèle de traduction principal : le second modèle de traduction est uniquement considéré si le

mot ou la séquence de mots ne sont pas trouvés dans le premier modèle. Les résultats obtenus sont présentés en Figure 4.5.

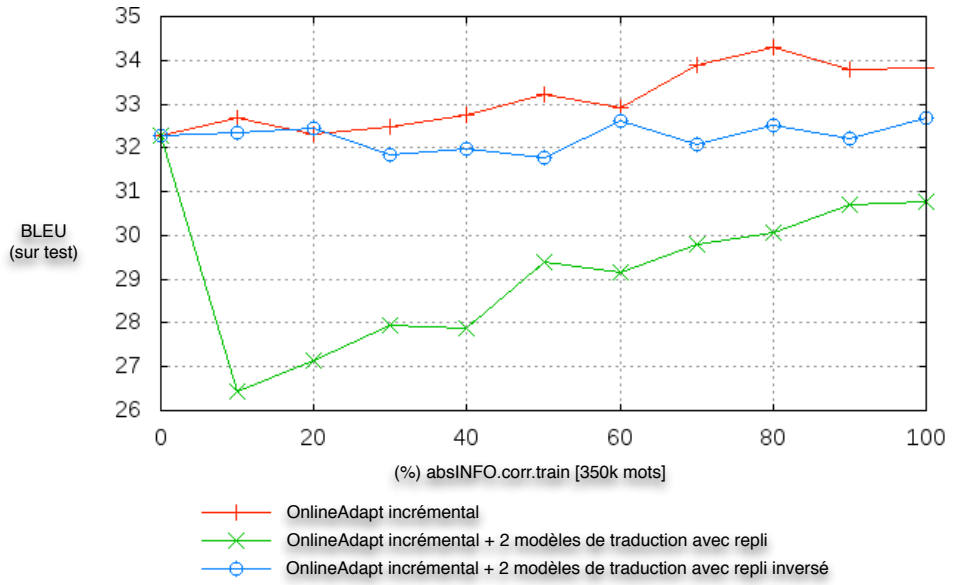


FIGURE 4.5 – Résultats pour l'utilisation de modèles par repli. La courbe « + » représente notre système de TAS à séquences de mots utilisant un seul modèle de traduction. La courbe « $\chi$ » représente notre système de TAS utilisant deux modèles de traduction avec le modèle en domaine comme principal modèle et le modèle générique du système référence en repli. La courbe « $\Theta$ » représente une configuration similaire à la précédente avec simplement une inversion dans l'ordre des deux modèles de traduction.

### Discussion

Comme nous pouvons le voir, nous obtenons des résultats vraiment différents selon quel modèle de traduction est utilisé en premier, mais cela s'explique facilement par la nature même des modèles back-off. La courbe « $\chi$ » représente notre système de TAS utilisant nos deux modèles de traduction, et priorité est donnée au modèle du domaine, le modèle générique étant lui utilisé en repli. Notre modèle de traduction du domaine est construit incrémentalement avec des données ajoutées à chaque itération, c'est-à-dire à partir d'une quantité très faible quantité de données à chaque fois. Bien que ce modèle puisse atteindre, à terme, une taille conséquente, il n'en pas moins « vide » lors des premières itérations. Surtout, ces données très restreintes ne couvrent pas de connaissances plus génériques et de fait, fait s'effondrer les performances du systèmes de TAS.

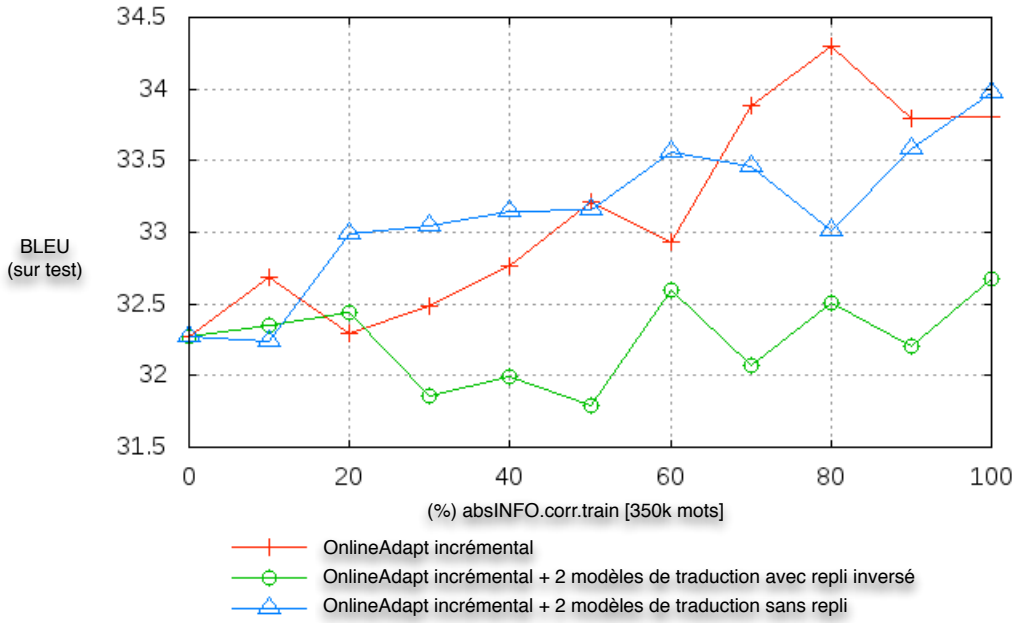


FIGURE 4.6 – Résultats comparatifs entre l'utilisation de deux modèles de traduction sans ($\Delta$) et avec repli ($\Theta$). La courbe « + » représente notre système de TAS utilisant un seul modèle de traduction.

La Figure 4.6 présente les résultats obtenus lorsque les deux modèles de traduction sont utilisés de façon conjointe. Dans cette configuration, il n'y a pas de modèle prioritaire, des options de traductions séparées sont créées pour chaque occurrence, les scores étant combinés si la même option de traduction est trouvée dans les deux modèles de traduction. Comparé à l'utilisation d'un seul modèle de traduction, nous pouvons observer une dégradation significative aux environs de 80% des données d'adaptation injectées, avant de finalement obtenir un score BLEU similaire (supérieur de 0,2 point) comparé à *inc-Giza-pp* et *OnlineAdapt*.

### Discussion

Une fois encore, nous pensons que la nature de notre corpus « absINFO » permet d'expliquer l'évolution de notre score. Lorsque nos systèmes de TAS doivent traduire plus de phrases génériques, il est probable que les options de traduction proviennent de notre modèle de traduction générique plutôt que de notre modèle du domaine.
À partir de cette observation, nous avons essayé de limiter l'analyse de la distance d'édition aux éditions du type « substitution » seulement.

### Filtrage par type d'édition

La Figure 4.7 montre les résultats obtenus avec un modèle de traduction en domaine entraîné seulement à partir des substitutions détectées durant l'analyse de la post-édition. Comme soutenu auparavant, nous considérons que lorsqu'une édition de type « substitution » est détectée, cela correspond à ce que le système de TA ne connaît pas puisqu'il a été nécessaire de corriger la traduction sur sa sortie.

### Discussion

Comme nous pouvons le constater, la dégradation précédemment observée est moins importante. Dans l'ensemble, l'évolution du score BLEU est plus stable que pour les approches testées jusque là. En ne nous intéressant qu'aux paires de séquences de mots ne correspondant qu'aux seules substitutions (dans le chemin d'édition), nous avons aussi limité les séquences de mots contextuelless dans notre modèle de traduction du domaine. Il convient également de prendre en compte les erreurs d'alignement qui auraient un impact plus important dans cette configuration sur la qualité du modèle de traduction.

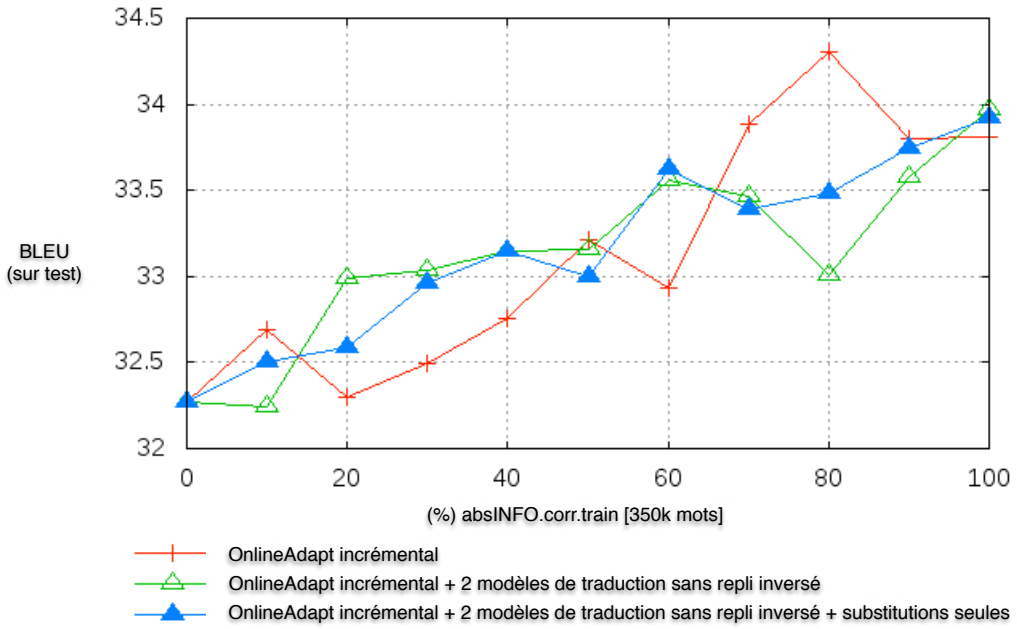


FIGURE 4.7 – Scores BLEU pour des modèles de traduction sans repli pour des éditions de type « substitution » uniquement.

### Génération des *n*-bests à la traduction

Un des points clés présentés dans ce chapitre est l'utilisation des hypothèses de traduction pour générer les liens d'alignements entre la phrase source et son hypothèse de traduction issue du système de TAS. Par défaut, ce système retourne la meilleure traduction candidate en sortie du décodage. Cela signifie que cette hypothèse de traduction a obtenu le meilleur score de décodage (on parle de « 1-best »), mais cela ne signifie pas nécessairement que l'alignement généré soit, lui, le meilleur alignement possible. Dans le manuel utilisateur du toolkit Moses, il est justement précisé que :

> " the phrase table can include different word-to-word alignments for the source-to-target and target-to-source directions, at least in principle. Hence, the two alignments can differ. "
>
> *« la table de traduction, ou modèle de traduction, peut inclure des alignements mot-à-mot différents pour les directions source-to-target et target-to-source, du moins en principe. Par conséquent, les deux alignements peuvent différer. »*

À partir de cette observation, nous avons exploré les $n$ plus probables hypothèses de traduction (*i.e.* générer une liste de « $n$-best »). En effet, une phrase source pourrait être traduite en la même hypothèse de traduction en ayant toutefois une segmentation en séquences de mots qui elles, seraient différentes. Par l'intermédiaire de notre approche, pour la même paire (source,traduction), si nous arrivons à avoir plusieurs alignements candidats, nous pouvons générer plus d'alignements source-vers-référence, et ainsi, renforcer potentiellement notre modèle de traduction du domaine. Cela a toutefois un effet négatif sur la rapidité globale du procédé puisque plus d'alignements sont à traiter. Nous avons donc décidé de nous limiter dans un premier temps sur la génération des $n$-bests de façon à entrevoir une répercussion sur la qualité du modèle de traduction, tout en limitant le temps de calcul nécessaire. C'est ainsi qu'en utilisant seulement les deux meilleurs hypothèses de traduction du système de TA, non distinctes, nous avons obtenu les résultats présentés en Figure 4.8.

### Discussion

La courbe étoilée représente les résultats de notre système pour lequel nous avons utilisé les 2-best traductions candidates pour extraire les paires de séquences de motss, tandis que la seconde courbe représente le même système, mais seul le 1-best candidat est utilisé. Malheureusement comme on peut le voir, les résultats obtenus sont moins bons que ce à quoi nous nous attendions. Nous restons persuadés que l'utilisation des $n$-meilleurs candidats peut apporter

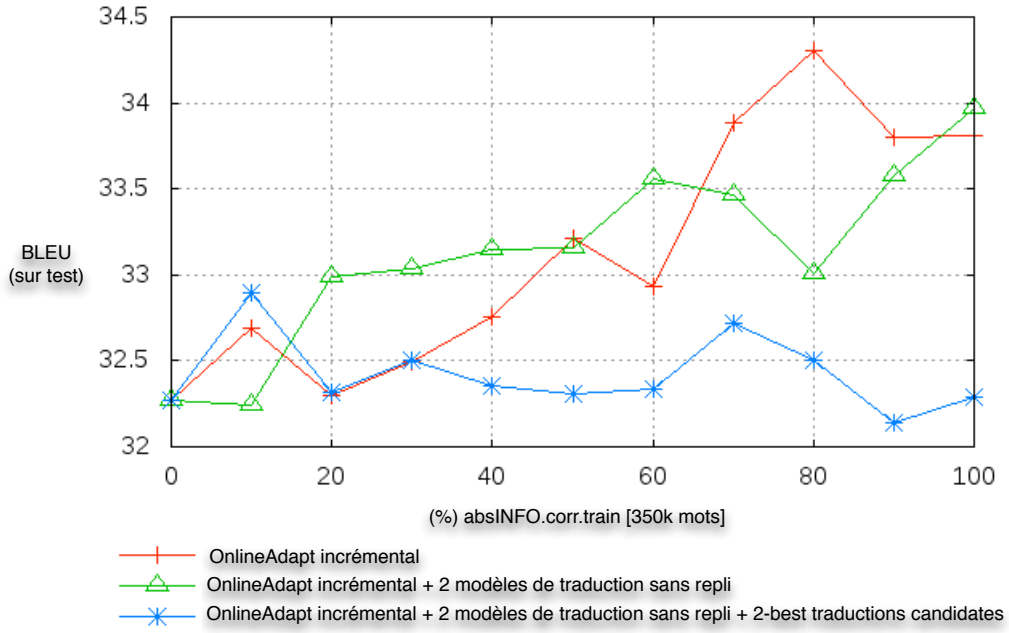


FIGURE 4.8 – Scores BLEU obtenus en exploitant les deux meilleures hypothèses de traduction générées par les systèmes de TAS.

d'avantage que ce que les résultats le montrent, ce qui fera de cette approche une future piste de travail à explorer.

### Suppression du tuning

Jusqu'alors, à chaque itération, nous procédions au réglage (tuning) des poids des modèles de traduction de nos systèmes. Nous présentons ci-après les résultats obtenus pour une adaptation incrémentale d'un système de TAS basé sur les séquences de mots, sans que soit réalisée une phase de tuning après chaque itération. Nous soutenons que nous n'avons vraiment pas besoin d'optimiser nos modèles après chaque itération, car l'adaptation ne fait qu'ajouter de petites quantités d'informations. Le tuning est uniquement appliqué à la création du modèle, et les paramètres résultants sont maintenus pendant le processus d'adaptation. De cette façon, notre procédure est beaucoup moins chronophage, tout en restant stable comme l'illustrent les résultats présentés par la Figure 4.9.

La courbe aux carreaux représente notre système avec un processus de tuning basé sur MERT effectué à chaque itération du processus d'adaptation. La courbe avec pointillés représente quant à elle le même système pour lequel les poids obtenus après un tuning effectué après la première itération (10% de données injectées) sont conservés pour les itérations suivantes. Nous avons

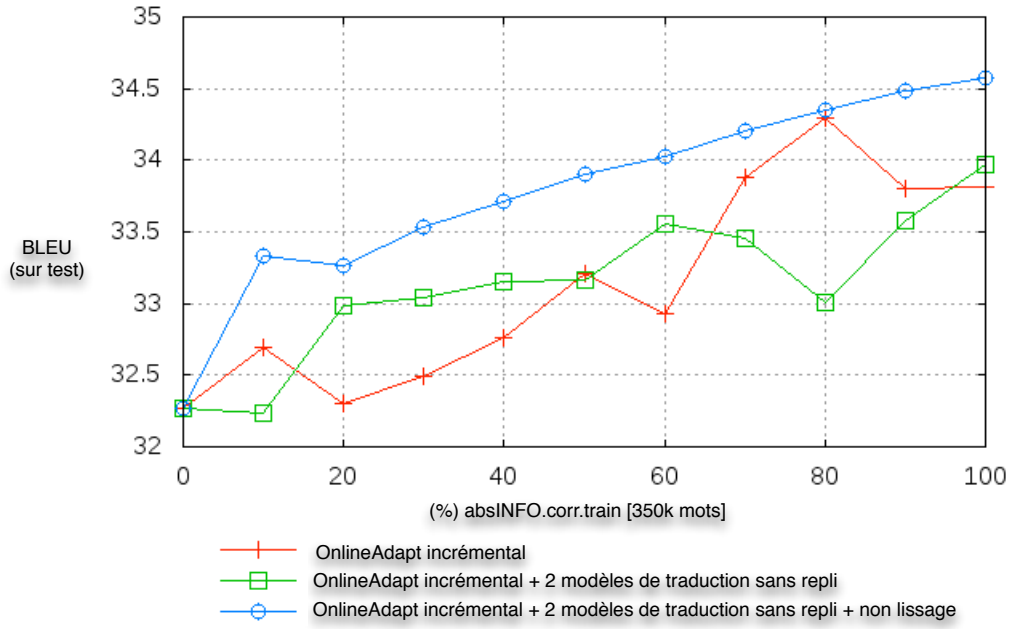


FIGURE 4.9 – Scores BLEU pour une adaptation incrémentale sans tuning.

conservé les poids suite à la première itération puisque pour rappel, nos systèmes de référence ont été entraînés avec un seul modèle de traduction. Il nous fallait donc effectuer un tuning pour la première utilisation d'un système à doubles modèles de traduction.

### Discussion

Premièrement, nous pouvons observer une nette différence entre les courbes « avec repli » et « avec repli + non lissage » à 10% de données d'adaptation injectée, alors qu'elles résultent de la même approche. Cela est dû à notre système de référence que nous avons utilisé : par défaut, notre système de TAS utilise un seul modèle de traduction. Il nous a donc été nécessaire de retuner les poids de notre système à 10% afin de pondérer notre modèle de traduction du domaine. Ce système à 10% devient en quelque sorte notre « nouveau système de référence » qui cette fois utilise deux modèles de traduction. Ces poids ayant donc été conservés par la suite, tout au long de l'adaptation.

Deuxièmement, la courbe résultante est plutôt lisse, ce qui donne une bonne idée sur l'instabilité qui peut résulter du processus de tuning. À noter que nous n'avons fait qu'une itération de tuning. Peut-être l'instabilité aurait été moindre si nous avions réalisé plusieurs tuning avant de faire une sorte de moyenne.

Pour résumer, en appliquant notre approche d'adaptation incrémentale, nous obtenons une nette amélioration en terme de score BLEU de 0,5 point, et ce, sans que nous procédions à un tuning à chaque itération. Le tuning étant toutefois nécessaire, il peut être réalisé à intervalle de temps plus important, par exemple – dans un contexte de post-édition en milieu industriel – il peut être procédé au tuning du système chaque nuit, ou lorsqu'une quantité conséquente de nouvelles données est injectée dans le système.

## 4.4 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthodologie d'alignement mot-à-mot pour l'adaptation incrémentale utilisant un système de TAS basé sur les séquences de mots. Cette méthode utilise les informations générées pendant la phase de traduction et s'appuie ensuite sur une analyse de la post-édition (ici simulée) pour déduire un alignement source-vers-référence au niveau du mot.

Comparée à la version incrémentale de *Gizapp*, la méthode standard actuellement utilisée dans le domaine, la première partie des expériences montre que notre approche nous permet d'obtenir des performances similaires en terme de score BLEU, ainsi qu'une amélioration significative du temps de calcul requis. Alors que la version incrémentale de *Gizapp* nécessite aux alentours de 14 minutes pour aligner deux corpus d'environ 35k mots, l'approche que nous proposons ne requiert que quelques secondes.

Dans la seconde partie de ce chapitre, nous avons présenté une série d'expériences sur l'utilisation et la combinaison de modèles de traduction. Ces expériences montrent que nous pouvons obtenir de meilleurs résultats avec une utilisation conjointe de deux modèles plutôt qu'avec un seul. Ceci est plutôt intéressant dans la mesure où nous pouvons avoir un modèle générique et lui associer un modèle de traduction additionnel lorsque l'on souhaite apporter des connaissances spécifiques pour adapter le système à un domaine donné. De plus, l'utilisation de deux modèles de traduction tend à favoriser la rapidité d'adaptation d'un système de TA pour une approche incrémentale telle que nous venons de le présenter dans ce chapitre. Rapidité qui peut être augmenté en supprimant une étape de tuning à chaque itération lorsque la quantité de données injectée n'est pas très importante.

### Publication

Cette même approche et les résutlats expérimentaux associés que nous venons de présenter ont été publiés [Blain 2012] au *International Workshop on Spoken Language Translation* organisé en 2012 à Hong-Kong, en Chine.

# Conclusions et perspectives futures

# Conclusions

Le travail de thèse présenté dans ce manuscrit s'inscrit dans le cadre d'une convention Cifre-Défense entre la société SYSTRAN, leader mondial des technologies de TA, et le LIUM, spécialisé dans la TA fondée sur les corpus.

Dans la première partie de cette thèse, nous nous sommes attachés à présenter les paradigmes fondamentaux du domaine dans lequel nous avons évolué. Plus particulièrement, l'approche empirique de la TA fondée sur les corpus ainsi que le concept de post-édition furent présentés.

En deuxième partie, nous avons présenté le projet ANR COSMAT, cadre applicatif dans lequel s'inscrivent nos travaux de recherche. L'objectif principal de ce projet était de mettre à disposition de la communauté scientifique par le biais d'une intégration à la plateforme en ligne HAL, un service de TA de contenus scientifiques. Bien que cette intégration ne soit pas encore effective, le projet COSMAT a d'ores et déjà permis le développement d'une interface riche de post-édition, ainsi que la création et la libre distribution d'un corpus scientifique composé des domaines Physique et Informatique extraits de HAL. Ce corpus a déjà été utilisé par le « JHU SMT workshop » en 2012 et par le projet européen « TransLectures ».

Dans la troisième partie de cette thèse, nous présentons les contributions quant aux problématiques que nous avons traitées. Nous avons présenté dans un premier temps nos travaux sur l'analyse qualitative de données post-éditées. Pour modéliser l'intention du post-éditeur durant sa tâche, nous avons proposé l'introduction d'une notion nouvelle d'« Actions de Post-Édition ». Ces APE définissent les éditions minimales et logiques réalisées par le post-éditeur en opposition aux éditions que nous avons qualifiées de mécaniques et qui sont actuellement considérées dans les métriques d'évaluation standards. Pour formaliser cette notion, nous avons proposé une typologie pour le français en nous inspirant de classifications d'erreurs de traduction. Cette typologie résulte d'un ensemble d'observations réalisées sur un jeu de données post-éditées mis à la disposition de SYSTRAN par certains de ses clients. De cette notion d'APE et de la typologie associée, nous avons développé un prototype de classifieur permettant d'annoter automatiquement en APE un corpus de données post-éditées. Cette annotation automatique nous permet actuellement d'identifier les informations importantes parmis ces données avec un taux de rappel et de précision intéressant pour un taux de couverture globale supérieur à 30% (pour les classes implémentées). Ces résultats sont très prometteurs quant au potentiel que peut avoir une adaptation en conséquence d'un système de TA .

Dans un second temps, nous présentons une procédure d'adaptation continue d'un système de TAS fondée sur une nouvelle méthodologie d'alignement au niveau du mot. En simulant un processus de post-édition, nous déduisons un alignement source-vers-référence en utilisant l'hypothèse de traduction comme pivot. Les performances que nous obtenons sont semblables à une adaptation incrémentale utilisant l'outil d'alignement standard *inc-Giza-pp*. Toutefois, notre approche obtient de bien meilleurs résultats quant au temps nécessaire à l'alignement et à l'extraction des informations de post-édition, de l'ordre d'une quinzaine de minutes pour la procédure standard à quelques secondes pour notre approche. Des résultats expérimentaux nous ont montré également comment il est possible de réduire davantage le temps nécessaire à l'adaptation d'un système en utilisant notamment la combinaison de plusieurs modèles de traduction.

# Perspectives futures

En l'état actuel des travaux de recherche présentés dans ce manuscrit, plusieurs perspectives de travail sont envisagées.

## Analyse qualitative et automatique de données post-éditées

Nous avons montré au cours du chapitre 3 qu'une analyse plus fine de la post-édition, en s'intéressant à l'intention du post-éditeur à travers différents niveaux d'éditions, permettait également de scorer différemment la qualité de l'hypothèse de traduction générée par le système. Cette évaluation tient compte de la différence entre ce que nous considérons être les éditions primaires, liées à une erreur du système de TA, et les éditions secondaires induites par les éditions primaires. Partant de cette observation, il peut être intéressant de travailler à une nouvelle métrique dédiée à l'évaluation de données post-éditée et qui se voudrait une version améliorée du (H)TER actuel. Ce dernier ne tenant pas compte des différents niveaux d'éditions que nous venons d'aborder.

Par ailleurs, et suite à plusieurs demandes de la part de membres de la communauté, il est envisagé de développer un analyseur de données post-éditées sous licence open source, afin de pouvoir le mettre à disposition de la communauté. La version actuellement développée repose sur des technologies qui sont propriétés de la société SYSTRAN et qui n'ont pas vocation à être librement distribuées. Un certain nombre de librairies linguistiques sont aujourd'hui disponibles et utilisables gratuitement et sur lesquelles nous pourrions nous appuyer.

## Adaptation incrémentale d'un système de TAS

Notre procédure d'adaptation incrémentale d'un système de TAS repose sur un algorithme d'alignement source-vers-référence au niveau des mots qui aujourd'hui n'exploite pas nos travaux sur l'analyse qualitative en APE présentée au chapitre précédent.

À l'occasion du lancement prochain du service COSMAT, nous prévoyons de renforcer cet alignement via l'utilisation de l'interface de post-édition développée par la société SYSTRAN. Grâce à cette interface, nous pourrons en effet exploiter l'historique des modifications du post-éditeur pour créer un alignement traduction-vers-référence plus intelligent et représensatif de la tâche. Cet alignement pouvant cette fois être réalisé non plus au niveau des mots seuls, mais au niveau des groupes de mots, facilitant par la même l'analyse en APE.

Notre technique d'adaptation incrémentale est aussi très intéressante pour le projet Matecat [33] dont l'objectif est d'intégrer de manière efficace et ergonomique la TA dans un workflow de traduction humaine. Notre approche y sera pleinement intégrée pour compléter un processus de traduction assistée par ordinateur (TAO) évolutive, où le système de TA doit s'adapter rapidement aux traductions effectuées chaque jour, voire en temps réel pour éviter l'aspect répétitif de la post-édition comme nous avons pu le voir au chapitre 3.

33. http ://www.matecat.com/matecat/the-project/

# Acronymes

| | |
|---|---|
| **ANR** | *Agence Nationale de la Recherche* |
| **APE** | *Action de Post-Édition* |
| **CSLM** | *Continuous Space Language Model* |
| **DGA** | *Délégation Générale pour l'Armement* |
| **GN** | *Groupe Nominale* |
| **GV** | *Groupe Verbale* |
| **HTER** | *Human-targeted Translation Error Rate* |
| **IWSLT** | *International Workshop on Spoken Language Translation* |
| **MERT** | *Minimum Error Rate Training* |
| **MIRA** | *Margin Infused Relaxed Algorithm* |
| **ML** | *Modèle(s) de langage* |
| **NIST** | *National Institute of Standards and Technology* |
| **PES** | *Post-Édition Statistique* |
| **TA** | *Traduction Automatique* |
| **TAS** | *Traduction Automatique Statistique* |
| **TER** | *Translation Error Rate* |
| **WER** | Taux d'Erreur/Mot (*Word Error Rate*) |
| **WMT** | *Workshop on statistical Machine Translation* |
| **WYSIWYG** | *What You See Is What You Get* |

# Bibliographie


[Axelrod 2011] Axelrod A., He X. et Gao J., Domain adaptation via pseudo in-domain data selection, dans *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362, Juillet 2011.

[Banerjee 2005] Banerjee S. et Lavie A., Meteor : An automatic metric for mt evalua- tion with improved correlation with human judgments., *In Proceedings of the ACL Workshop on Intrinic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

[Barrett 2001] Barrett R. C., Maglio P. P. et Underwood G. M., User-centered push methods and system, Mai 23 2001, uS Patent App. 20,020/054,057.

[Bengio 2003] Bengio Y., Ducharme R., Vincent P. et Jauvin C., A neural probabilistic language model, dans *Journal of Machine Learning Research*, volume 3, pages 1137–1155, Février 2003.

[Bertoldi 2006] Bertoldi N., Cettolo M., Cattoni R., Chen B. et Federico M., ITC-IRST at the 2006 TC-STAR SLT evaluation campaign, dans *TC-STAR Workshop on Speech-to-Speech Translation*, pages 19–24, 2006.

[Blain 2012] Blain F., Schwenk H. et Senellart J., Incremental adaptation using translation informations and post-editing analysis, *International Workshop on Spoken Language Porcessing (IWLST)*, pages 234–241, 2012.

[Blain 2011] Blain F., Senellart J., Schwenk H., Plitt M. et Roturier J., Qualitative analysis of post-editing for high quality machine translation, dans for Machine Translation (AAMT) A.-P. A., rédacteur, *Machine Translation Summit XIII*, Xiamen (China), 19-23 sept. 2011.

[Boitet 2008] Boitet C., Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes, *TALN-08*, 2008.

[Brown 1993] Brown P. F., Pietra S. A. D., Pietra V. J. D. et Mercer R. L., The mathematics of statistical machine translation, dans *Computational Linguistics*, volume 19, pages 263–311, Juin 1993.

[Brown 1996] Brown R. D., xample-based machine translation in the pangloss system., *In Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

[Callison-Burch 2008] Callison-Burch C., Fordyce C., Koehn P., Monz C. et Schroeder J., Further meta-evaluation of machine translation, *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Juin 2008.

[Chen 1996] Chen S. F. et Goodman J. T., An empirical study of smoothing techniques for language modeling, dans *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Juin 1996.

[Chiang 2005] Chiang D., A hierarchical phrase-based model for statistical machine translation, dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, 2005.

[Crammer 2003] Crammer K. et Singer Y., Ultraconservative online algorithms for multiclass problems, dans *Journal of Machine Learning Research*, volume 3, pages 951–991, Janvier 2003.

[Daille 1994] Daille B., Gaussier É. et Langé J.-M., Towards automatic extraction of monolingual and bilingual terminology, *COLING '94 Proceedings of the 15th conference on Computational linguistics*, 1 :515–521, 1994.

[Déjean 2002] Déjean H., Gaussier É. et Sadat F., Bilingual terminology extraction : an approach based on a multilingual thesaurus applicable to comparable corpora, *In Proceedings of the 19th International Conference on Computational Linguistics COLING*, pages 218–224, 2002.

[Deniz 2008] Deniz N. et Turhan C., English to turkish example-based machine trans- lation with synchronous sstc, *In Proceedings of the Fifth International Conference on Information Technology : New Generations*, pages 674–679, 2008.

[Denkowski 2011] Denkowski M. et Lavie A., Meteor 1.3 : Automatic metric for reliable optimization and evaluation of machine translation systems, dans *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Juillet 2011.

[DePalma 2010] DePalma D. A. et Hegde V., The market for mt post-editing, page 4, November 2010.

[Doddington 2002] Doddington G., Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, dans *roceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

[Doherty 2010] Doherty S., O'Brien S. et Carl M., Eye tracking as an mt evaluation technique, *Machine translation*, pages 1–13, 2010.

[Dugast 2007] Dugast L., Senellart J. et Koehn P., Statistical post-editing on systran's rule-based translation system, dans *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Association for Computational Linguistics, 2007.

[Federico 1998] Federico M. et De Mori R., Language modelling, dans *Spoken Dialogues with Computers*, pages 204–210, Avril 1998.

[Font-Llitjós 2005] Font-Llitjós A., Carbonell J. G. et Lavie A., A framework for interactive and automatic refinement of transfer-based machine translation, dans *European Association of Machine Translation (EAMT) 10th Annual Conference. Budapest, Hungary*, Citeseer, 2005.

[Guzmán 2007] Guzmán R., Automating mt post-editing using regular expressions, 2007.

[Hardt 2010] Hardt D. et Elming J., *Incremental Re-training for Post-editing SMT.*, 2010.

[Hasler 2011] Hasler E., Haddow B. et Koehn P., Margin infused relaxed algorithm for moses, dans *The Prague Bulletin of Mathematical Linguistics*, numéro 96, pages 69–78, Octobre 2011.

[Jelinek 2004] Jelinek F., Some of my best friends are linguists, dans *Proceedings of LREC 2004*, http ://www.lrec-conf.org/lrec2004/doc/jelinek.pdf, Mai 2004.

[Jelinek 1980] Jelinek F. et Mercer R. L., *Pattern recognition in practice*, chapitre Inter- polated Estimation of Markov Source Parameters from Sparse Data, pages 381–397, 1980.

[Katz 1987] Katz S. M., Estimation of probabilities from sparse data for the language model component of a speech recognizer, dans *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 35, pages 400–401, Avril 1987.

[Koehn 2007a] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A. et Herbst E., Moses : Open source toolkit for statistical machine translation, dans *Meet- ing of the Association for Computational Linguistics*, pages 177–180, 2007a.

[Koehn 2007b] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R. et al., Moses : Open source toolkit for statistical machine translation, dans *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Association for Computational Linguistics, 2007b.

[Lambert 2012a] Lambert P., Schwenk H. et Blain F., Automatic translation of scientific documents in the hal archive, dans *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages p.3933–3936, European Language Resources Association (ELRA), Istanbul, Turkey, may 2012a.

[Lambert 2012b] Lambert P., Senellart J., Romary L., Schwenk H., Zipser F., Lopez P. et Blain F., Collaborative machine translation service for scientific texts, *n Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 11–15, Avril 2012b.

[Langlais 2006] Langlais P. et Gotti F., Ebmt by tree-phrasing, *Journal of Machine Translation*, 20(1) :1–23, 2006.

[Lavie 2007] Lavie A. et Agarwal A., Meteor : An automatic metric for mt evaluation with high levels of correlation with human judgments, dans *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, 2007.

[Levenberg 2010] Levenberg A., Callison-Burch C. et Osborne M., Stream-based translation models for statistical machine translation, dans *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 394–402, Association for Computational Linguistics, 2010.

[Levenshtein 1966] Levenshtein V. I., Binary codes capable of correcting deletions, insertions and reversals., *Soviet Physics Doklady*, 10(8) :707–710, Février 1966.

[Lopez 2009] Lopez P., Grobid : Combining automatic bibliographic data recognition and term extraction for scholarship publications, *Proceedings of the 13th European Conference on Digital Library (ECDL), Corfu, Greece*, 2009.

[Martinez 2003] Martinez L. G., *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output*, Thèse de doctorat, Citeseer, 2003.

[Matusov 2006] Matusov E., Zens R., Vilar D., Mauser A., Popovic M., Hasan S. et Ney H., The RWTH machine translation system, dans *TC-STAR Workshop on Speech-to-Speech Translation*, pages 31–36, 2006.

[Moore 2010] Moore R. C. et Lewis W., Intelligent selection of language model training data, dans *Proceedings of the ACL Conference Short Papers*, pages 220–224, Juillet 2010.

[Nagao 1984] Nagao M., Artificial and human intelligence, chapitre a framework of a mechanical translation between japanese and english by analogy principle., 1984.

[Och 2003a] Och F., Minimum error rate training in statistical machine translation, dans *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 160–167, Association for Computational Linguistics, 2003a.

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

[Och 2003c] Och F. J., Minimum error rate training in statistical machine translation, dans *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 160–167, Juillet 2003c.

[Och 2000] Och F. J. et Ney H., Giza++ : Training of statistical translation models, 2000.

[Och 2003d] Och F. J. et Ney H., A systematic comparison of various statistical alignment models, dans *Computational Linguistics*, volume 29, pages 19–51, Mars 2003d.

[Papineni 2002] Papineni K., Roukos S., Ward T. et Zhu W.-J., BLEU : a method for automatic evaluation of machine translation, dans *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Juillet 2002.

[Plitt 2010] Plitt M. et Masselot F., A productivity test of statistical machine translation post-editing in a typical localisation context, *The Prague Bulletin of Mathematical Linguistics*, 93(-1) :7–16, 2010.

[Rousseau 2013] Rousseau A., Xenc : an open-source tool for data selection in natural language processing, *Prague Bulletin of Mathematical Linguistics*, 100 :73–82, 2013.

[Schwenk 2007] Schwenk H., Continuous space language models, dans *Computer Speech and Language*, volume 21, pages 492–518, Janvier 2007.

[Schwenk 2012a] Schwenk H., Continuous space translation models for phrase-based statistical machine translation, *Coling*, pages 1071–1080, 2012a.

[Schwenk 2013] Schwenk H., Cslm - a modular open-source continuous space language modeling toolkit, *Interspeech*, 2013.

[Schwenk 2009a] Schwenk H., Abdul-Rauf S., Barrault L. et Senellart J., Smt and spe machine translation systems for wmt'09, *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134, 2009a.

[Schwenk 2009b] Schwenk H., Abdul-Rauf S., Barrault L. et Senellart J., Smt and spe machine translation systems for wmt'09, dans *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134, Citeseer, 2009b.

[Schwenk 2006] Schwenk H., Costa-Jussà M. R. et Fonollosa J. A. R., Continuous space language models for the iwslt 2006 task, dans *Proceedings of International Workshop on Spoken Language Translation*, pages 166–173, Novembre 2006.

[Schwenk 2002] Schwenk H. et Gauvain J.-L., Connectionist language modeling for large vocabulary continuous speech recognition, dans *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 765–768, Mai 2002.

[Schwenk 2005] Schwenk H. et Gauvain J.-L., Training neural network language models on very large corpora, dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 201–208, Octobre 2005.

[Schwenk 2012b] Schwenk H., Rousseau A. et Attik M., Large, pruned or continuous space language models on a GPU for statistical machine translation, dans *NAACL Workshop on the Future of Language Modeling*, pages 11–19, Juin 2012b.

[Senellart 2001] Senellart J., Dienes P. et Váradi T., New generation systran translation system, *In MT Summit VIII*, Septembre 2001.

[Shah 2011] Shah K., Barrault L. et Schwenk H., Parametric weighting of parallel data for statistical machine translation, *A General Framework to Weight Heterogeneous Parallel Data for Model Adaptation in Statistical Machine Translation*, pages 1323–1331, Novembre 2011.

[Shah 2012] Shah K., Barrault L. et Schwenk H., A general framework to weight heterogeneous parallel data for model adaptation in statistical machine translation, *MT Summit*, Octobre 2012.

[Simard 2007] Simard M., Ueffing N., Isabelle P. et Kuhn R., Rule-based translation with statistical phrase-based post-editing, dans *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206, Association for Computational Linguistics, 2007.

[Snover 2006] Snover M., Dorr B., Schwartz R., Micciulla L. et Makhoul J., A study of translation edit rate with targeted human annotation, dans *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, 2006.

[Snover 2009] Snover M., Madnani N., Dorr B. et Schwartz R., Fluency, adequacy, or HTER ? exploring different human judgments with a tunable MT metric, dans *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Mars 2009.

[Specia 2011] Specia L., Exploiting objective annotations for measuring translation post-editing effort, dans *15th Annual Conference of the European Association for Machine Translation, EAMT*, volume 11, 2011.

[Tillmann 1997] Tillmann C., Vogel S., Ney H., Zubiaga A. et Sawaf H., Accelerated dp based search for statistical translation, *In Fifth European Conf. on Speech Com- munication ans Technology*, pages 2667–2670, Septembre 1997.

[Tinsley 2008] Tinsley J., Ma Y., Ozdowska S. et Way A., M a t r e x : the dcu mt system for wmt 2008, *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 171–174, Juin 2008.

[Turian 2003] Turian J. P., Shen L. et Melamed I. D., Evaluation of machine translation and its evaluation, *In Proceedings of MT-Summit IX*, 2003.

[Veale 1997] Veale T. et Way A., Gaijin : A template-driven bootstrapping approach to example-based machine translation., *In Proceedings of NeMNLP97*, 1997.

[Vilar 2006] Vilar D., Xu J., D'Haro L. F. et Ney H., Error analysis of statistical machine translation output, dans *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, Citeseer, 2006.

[Weaver 1947] Weaver W., 1947, letter to Norbert Wiener.

[Weaver 1955] Weaver W., Translation, *Machine Translation of Languages*, 14 :15–23, 1955.

[White 1994] White J. S., The ARPA MT evaluation methodologies : Evolution, lessons, and further approaches, dans *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, pages 193–205, 1994.

[Witten 1991] Witten I. H. et Bell T. C., The zero-frequency problem : estimating the probabilities of novel events in adaptive text compression, dans *IEEE Transactions on Information Theory*, volume 37, pages 1085–1094, Juillet 1991.

[Yamada 2001] Yamada K. et Knight K., A syntax-based statistical translation model, dans *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530, Juillet 2001.

# Annexes

# Annexe A

# Liste des publications

– Blain Frédéric, Senellart Jean, Schwenk Holger, Plitt Mirko, Roturier Johann. « *Qualitative Analysis of Post-Editing for High Quality Machine Translation* ». Proceedings of the 13th Machine Translation Summit, Xiamen(China), Septembre 2011, 8 pages.

– Lambert Patrik, Senellart Jean, Romary Laurent, Schwenk Holger, Zipser Floren, Lopez Patrice, Blain Frédéric. « *Collaborative Machine Translation Service for Scientific texts* ». Proceedings of the demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Avignon(France), Avril 2012.

– Lambert Patrik, Schwenk Holger, Blain Frédéric. « *Automatic Translation of Scientific Documents in the HAL Archive* ». Proceedings of the Eight International Conference on LREC, Istanbul(Turkey), Mai 2012.

– Blain Fréderic, Schwenk Holger, Senellart Jean. « Incremental Adaptation Using Translation Information and Post-Editing Analysis ». International Workshop on Spoken Language Translation, Hong-Kong(China), Décembre 2012, 8 pages.