Intégration des expressions polylexicales dans un système de traduction statistique

Zied Elloumi

Intégration des Expressions Polylexicales dans un Système de Traduction Statistique

Nom: Elloumi
Prénom : Zied

UFR LLASIC

Mémoire de Master 2 Recherche - 30 crédits – Mention Sciences du Langage

Spécialité: Industries de la langue - Parcours : TALEP

Sous la direction de M. Olivier Kraif et M. Laurent Besacier

Année universitaire 2013-2014

Déclaration anti-plagiat

DECLARATION

Ce travail est le fruit d'un travail personnel et constitue un document original.

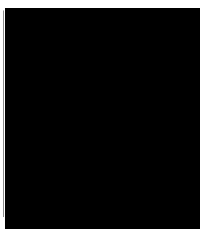
Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.

Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.

4. Les propos repris mot à mot à d'autres figurent entre guillemets (citations).

5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

Nom : **ELLOUMI** Prenom : **ZIED**

Date : **25/06/2014** Signature : 

Remerciements

Je tiens à remercier mes deux encadrants M. Olivier KRAIF et M. Laurent BESACIER pour toute l'attention, les conseils enrichissants et les aides qu'ils m'ont apportés durant la réalisation de ce mémoire.

Qu'ils trouvent ici, les marques de ma reconnaissance et de mon plus profond respect.

Je tiens à remercier sincèrement les membres du jury M. Georges ANTONIADIS, le responsable du Master IDL et Mme Agnès TUTIN qui me font le grand honneur d'évaluer ce travail.

Mes remerciements s'adressent également à tous mes enseignants, pour leur collaboration durant ces deux années de Master.

Un grand merci à tous mes amis qui m'ont encouragé et soutenu tout au long de ce travail.

Zied

Table des matières

Liste des figures

Liste des tableaux

Mots clés : Traitement automatique des langues naturelles, expressions polylexicales, corpus spécifiques, traduction automatique statistique, modèles de traduction.

Résumé

Les expressions polylexicales (EPL) constituent un champ de recherche intéressant dans le domaine de la linguistique computationnelle. Elles peuvent être considérées comme un vrai défi pour les systèmes de traduction automatique statistique (TAS). Dans ce mémoire, nous avons enrichi une ressource textuelle littéraire existante en vue d'extraire des corpus spécifiques pour évaluer le problème des expressions polylexicales en traduction automatique statistique. Par ailleurs, dans le but d'améliorer les performances d'un système de traduction automatique nous avons testé une stratégie d'intégration des EPL en anglais. Les résultats obtenus au niveau du score BLEU sont encourageants.

Keywords: Natural language processing, multiword expression, specific corpus, statistical machine translation, translation models.

Abstract

Multiword Expressions constitute an interesting field of research in Computational Linguistics. They can be a real challenge for Statistical Machine Translation systems (SMT). In this paper, we have improved an existing literary lexical resource in order to extract a corpus in order to assess the problem of Multiword Expressions in SMT. Moreover, in order to improve the performance of an automatic translation system, we integrated English EPL. As regards the BLEU score, the results were encouraging.

1. Introduction générale

Ce mémoire de recherche s'inscrit dans le cadre de la formation de deuxième année de Master Sciences du Langage, spécialité Industries de la langue, de l'université Stendhal Grenoble 3. Les organismes accueillants sont : le LIDILEM (Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles) de l'Université Stendhal Grenoble 3, dont les activités s'organisent autour des axes de recherche principaux suivants : « description linguistique, sociolinguistique, acquisition, constitution et exploitation de corpus, didactique des langues, traitement automatique des langues, étude des formes nouvelles d'interaction suscitées par les Technologies de l'Information et de la Communication » [LIDILEM][1];ainsi que l'équipe GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) du laboratoire d'informatique de Grenoble (LIG[2]). Il s'agit d'un groupe pluridisciplinaire qui comprend des informaticiens, des linguistes, des phonéticiens et spécialistes du traitement des signaux, et encore des traducteurs dont « l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale) » GETALP[3].

Les expressions polylexicales constituent un champ de recherche intéressant en linguistique computationnelle. En effet ce sont des unités formées de plus d'un seul mot et qui regroupent un nombre important de phénomènes phraséologiques : collocations, constructions à verbe support, expressions idiomatiques, etc. Toutefois, ces expressions posent un défi pour les systèmes de traduction automatique statistique, parce qu'elles ne peuvent pas se traduire, le plus souvent, de façon compositionnelle.

Les locuteurs natifs utilisent souvent des expressions polylexicales dans diverses situations de la vie courante comme *bonne journée*, *c'est moi*, *au fur et à mesure*, *figue de barbarie*, *pomme de terre* … sans se rendre compte de leur composition, ni même de leur opacité sémantique.

En outre les expressions polylexicales (EPLs ou MWEs en anglais, pour *multiword expressions*) sont difficiles à définir, « car il n'y a même pas de consensus sur la définition d'un mot mot » (Ramisch, 2012:p183), et parce qu'elles décrivent tout un continuum de degrés de figement.

---

[1]http://lidilem.u-grenoble3.fr/

[2]http://www.liglab.fr/

[3]http://getalp.imag.fr/

Les EPLs sont des termes formés à plus d'un seul mot graphique. Elles concernent des phénomènes phraséologiques divers :

- les collocations qui désignent une association habituelle d'une lexie à une autre au sein de l'énoncé (ex. *pluie torrentielle, essuyer des échecs, grièvement blessé, etc.*).

- les expressions figées qui sont des successions de mots n'admettant aucune modification, et dont le sens est fréquemment figuré (ex. *pied-à-terre, au fur et à mesure, va-et-vient*).

- les expressions idiomatiques qui peuvent être identifiées comme des unités lexicales non combinées de manière libre, non prévisibles, dont le sens est également figuré, et qui sont spécifiques à une langue (ex. *il tombe des cordes, tiré à quatre épingle, sur la même longueur d'onde*).

- les expressions nominales qui désignent les termes polylexicaux (ex. *domaine de liaison à l'ADN*), les noms propres (ex. *Louis XIV, New York*) et les noms composés (ex. *grands-parents, sage-femme*).

- les expressions verbales contenant les verbes à particules (ex. *cut off, look up, put out*), les locutions verbales (*prendre en compte, mettre en oeuvre*) et les constructions à verbe support (ex. *prendre une douche, donner une conférence*).

- enfin, les locutions adverbiales/adjectivales qui comportent des expressions comme : *à tête reposée, tête nue, la tête en bas, en ce moment,* etc.

Les EPLs comprennent non seulement ces différentes typologies mais aussi d'autres cas de figure moins faciles à classer.

Les expressions polylexicales sont généralement considérées comme un vrai problème dans plusieurs domaines du Traitement automatique des langues (désormais nous abrégerons : TAL) comme la traduction automatique (TA ou MT en anglais *Machine Translation*), l'extraction de connaissances, etc.

Les capacités de compréhension de l'homme ainsi que ses connaissances du monde lui permettent généralement d'interpréter des phrases contenant à la fois des mots simples et des mots composés avec aisance - ce qui n'est pas toujours à la portée de la machine.

De leur coté, les systèmes de traduction automatique ont été inventés afin de traduire un texte source dans une autre langue sans l'intervention d'un traducteur humain. Il est vrai que les performances des systèmes de traduction automatique ont été améliorées pendant ces dernières années mais la qualité de traduction reste souvent médiocre.

Assez fréquemment, cette mauvaise qualité est due à une mauvaise identification des EPL. Par exemple, on trouve des traductions de ce type, du français vers l'anglais : « *Voici votre pied-à-terre, dans une résidence de standing* » traduite par « *This is your base from the ground, in a luxury residence* » (Google Traduction[4]) : ici « *pied-à-terre* » devrait désigner un logement, pas le niveau du sol**.**

De même pour la phrase« *Même si je suis très soupe au lait, je ne me suis pas mis en colère* » traduite par « *Although I am very milk soup, I am not angry* ». D'après le contexte, le sens de « *soupe au lait* » devrait être « *un individu de caractère vif* ».

Comme le notent Carpuat et Diab (2010 : page 242) : « Les expériences montrent que l'inclusion d'EPL dans les systèmes de TA améliore la qualité de la traduction. »

Peu de systèmes, pourtant, intègrent ce niveau. Dans la perspective d'intégrer le traitement des EPL dans un système de TA, la première étape consiste à se doter d'un corpus spécifique en vue d'évaluer les problèmes de traduction posés par les EPLs. Dans un deuxième temps, nous chercherons à proposer des solutions pour améliorer le système grâce à une meilleure prise en compte de ces problèmes.

Cette étude est organisée en 4 parties, numérotées de 2 à 5 :

La partie 2 présente un état de l'art de la question. Nous examinerons d'abord la notion d'expression polylexicale d'un point de vue linguistique. Ensuite nous présenterons le domaine de la traduction automatique statistique. Nous passerons en revue quelques problèmes de traduction posés à Google-TR au niveau des EPL. Enfin, nous examinerons les travaux portant sur l'intégration des EPL dans un système de traduction automatique statistique.

La partie 3 est consacrée à la description de la chaine de traitement réalisée en vue de la préparation d'un corpus spécifique, riche en EPL, qui nous sera utile à des fins d'évaluation. Nous verrons que cette chaine de traitement passe par l'enrichissement des corpus du projet Emolex qui sont de type littéraire, en vue d'avoir une ressource lexicale riche et multi-domaine. Ensuite, nous allons utiliser les outils de recherche d'EmoConc (développés dans le cadre du projet Emolex) qui permettent d'exprimer des contraintes syntaxiques et d'obtenir des biconcordances sur des corpus alignés.

La partie 4 présente le processus semi-automatique d'extraction des listes des EPL que nous avons mis en œuvre pour l'extraction des corpus EN-FR et FR-EN spécifiques à notre tâche.

---

[4] http://translate.google.fr

La partie 5 enfin décrit une méthode d'intégration des expressions polylexicales en anglais (pour les verbes à particules ou *Phrasal Verbs* en anglais, et les expressions idiomatiques ou *idioms* en anglais) dans un système de traduction automatique que nous avons adapté.

2. État de l’art

Dans cette section, nous allons d’abord citer quelques types d’expressions polylexicales et quelques ressources lexicales ; par la suite nous passerons à une description des étapes du traitement des EPLs dans les applications du TAL, et mentionnerons quelques outils statistiques disponibles pour les identifier. Nous passerons ensuite à la description de l'approche statistique de la traduction automatique ; De plus, nous montrerons certains problèmes de traduction posés au niveau des EPLs ainsi que quelques stratégies d'intégration des EPLs dans un système de traduction. Enfin, nous allons décrire notre approche et citer quelques pistes pour traiter les EPLs dans les systèmes de traduction.

2.1. Les expressions polylexicales

« Les expressions polylexicales regroupent les expressions figées et semi-figées, les collocations, les entités nommées, les verbes à particule, les constructions à verbe support, les mots composés, les termes, etc. » (Sag et al,2002).

- Expressions totalement figées : ces expressions ont généralement un sens figuré. Par exemple : *soupe au lait, ventre à terre, en avoir ras le bol*, etc.

- Expressions semi-figées (ou collocations) : ce sont des expressions « ni complètement figées, ni complètement libres », (Tutin, 2005), qui posent des problèmes sur le plan syntaxique et sémantique, comme elle présentent des variations (lexicales, morphologiques, etc) très importantes**.**

Exemple de collocations : *peur bleue, fort comme un turc, mort de fatigue, etc.*

-Les verbes à particule (*phrasal verbs* en anglais) **:** ce sont des constructions récurrentes fréquentes en anglais, par exemple *break down, give up, cut out*, etc.. Elles sont quasi-inexistantes en français (on trouve *faire avec*). Ces expressions sont composées d’un verbe principal et d’une ou plusieurs particules (adverbiales ou prépositionnelles) qui complètent ou modifient le sens.

Les *phrasal verbs* en anglais peuvent être définis comme suit :

- **verbe+préposition :** *look like, look after , look at,* etc.
- **verbe+adverbe :** *put on, get up, take off,* etc.
- **Verbe+Adverbe+Préposition :***keep up with, get up to, get away with,*etc.

2.2. Lexiques d'expressions polylexicales

Il existe différentes ressources pour l'extraction des expressions polylexicales : à partir des dictionnaires électroniques, des vidéos d'apprentissage des langues, des tutoriels etc. En effet, certains dictionnaires peuvent être utiles pour l'identification des EPL, voire pour la traduction car ils répertorient des expressions polylexicales dans une langue source ainsi que leurs équivalents, avec éventuellement leur définition, dans une langue cible.

Mais, ceux-ci ne couvrent qu'un petit sous ensemble de phénomènes, la plupart des EPL qu'on trouve dans les corpus (les collocations) formant un ensemble plus vaste, et ouvert.

Parmi les ressources lexicales contenant des mots composés, il y a : DELAC, LAF, PRE, TLFi, DC, etc. En effet, « Les dictionnaires spécialisés (DC et LAF) contiennent dans l'ensemble d'avantage de collocations que les dictionnaires de langue. Parmi ceux-ci, c'est le *LAF* qui en recense le plus. » (Tutin,2005).

2.2.1. DELAC

Il s'agit d'un dictionnaire électronique contenant des mots composés ainsi que leur morphologie (des indications de traits sémantiques et des codes précisant leurs variations de formes).

Exemples : *Pluie torrentielle, N+NA+Conc+E01+z1:fs*

2.2.2. LAF

Il se base sur les principes du dictionnaire de collocations DC, le plus connu. Comme lui, il est fondé sur deux phénomènes lexicaux que sont les collocations et les dérivations sémantiques.

2.2.3. DC

Le *Dictionnaire des cooccurrents* (DC) de Beauchesne (2001), « contrairement au LAF, ne consigne que des informations de cooccurrence, organisées comme préconisé par Hausmann (1989) à partir de la base, parfois sommairement désambiguïsée » (Tutin, 2005).

- Les collocatifs adjectivaux sont organisés par ordre alphabétique.

- Sans informations sémantiques, les collocatifs verbaux sont rangés par type de construction.

2.3. Traitements des expressions polylexicales

Selon Anastasiou et al. (2009), l'intégration des expressions polylexicales dans une application de TAL nécessite l'enchainement des étapes suivantes :

- Acquisition : En raison de la variabilité des expressions polylexicales, l'identification des ces expression est très difficile.

- Interprétation : interprétation sémantique et syntaxique des expressions polylexicales.

- Désambiguïsation : La plupart des EPLs sont ambiguës de diverses manières. Il est important de déterminer si une EPL est utilisée au sens figuré ou littéralement dans son contexte.

- Applications : l'intégration des EPLs dans un système de traduction est très important pour l'identification de la syntaxique et de la sémantique des EPLs.

2.3.1. Outils pour l'acquisition des expressions polylexicales

Pour évaluer le taux d'occurrence des expressions polylexicales dans un corpus, il est nécessaire de posséder des outils permettant d'accomplir cette tâche.

Dans cette section, nous allons présenter quelques outils servant à l'identification des expressions polylexicales.

2.3.1.1. Le *Lexicoscope*

Le *Lexicoscope* est un outil d'exploration de la combinatoire lexico-syntaxique développé par Kraif et Diwersy(2012) dans le cadre du projet *Emolex*[5]. Cet outil est fondé sur un modèle de cooccurrence flexible et un langage de définition des expressions complexes permettant à l'utilisateur de préciser le contexte de ses pivots.

Le *Lexicoscope* sert à extraire des expressions polylexicales qui sont statiquement significatives dans les corpus ; autrement dit, les cooccurrences dépassent un certain seuil (t-score, z-score ou log-likelihood) de significativité, mais nous ne savons pas s'ils s'agit vraiment d'EPLs, de collocations, ou encore de routines lexicales.

[5]http://www.emolex.eu/

2.3.1.2. MWETOOLKIT

La boite à outils *MWETOOLKIT*[6]est un logiciel libre développé par Ramsich et al. (2012). Elle peut être définie comme un système hybride fondé sur des règles et des mesures d'association statistiques, dans le but d'identifier les expressions polylexicales à partir d'un corpus monolingue prétraité. De fait, *MWETOOLKIT* permet d'extraire des verbes à particules pour l'anglais, des noms composés pour le grec et des expressions verbales pour le portugais. Cette boite prend en charge des outils comme *TreeTagger*[7]pour l 'étiquetage morphosyntaxique et *Rasp Parser*[8]qui permet d'annoter un corpus textuel brut en constituants syntaxiques, au format XML.

Les motifs (ou patrons) choisis par l'utilisateur sont définis dans un format XML, et chargés à partir d'un script, pour faire la correspondance entre les motifs et les phrases du corpus.

A titre d'exemple, pour trouver une expression comportant un adjectif (A) suivi d'un nom (N), nous pouvons utiliser le modèle suivant :

<pat>

<w pos="A" />

<w pos="N" />

</pat>

Par conséquent, l'outil *MWETOOLKIT* peut être utile pour identifier les *phrasal verbs* en anglais dans un corpus monolingue, également l'outil *Lexicoscope* peut nous aider à créer une liste des EPL (collocations, *phrasal verbs,* etc.) bien organisée.

2.4. Traduction automatique statistique (TAS)

Contrairement aux approches expertes qui se basent essentiellement sur des règles et des connaissances réalisées par des experts linguistiques, l'approche statistique de la traduction automatique est fondée sur un modèle mathématique de distribution et d'estimation probabiliste développée par les chercheurs d'IBM. Les principaux composants d'un système de traduction automatique statistique sont : le modèle de traduction, le modèle de langage et le décodeur. En outre, la construction d'un système TAS nécessite l'existence des corpus parallèles volumineux pour l'apprentissage (Besacier, Cours).

---

[6]http://mwetoolkit.sourceforge.net/

[7]http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[8]http://www.sussex.ac.uk/Users/johnca/rasp/

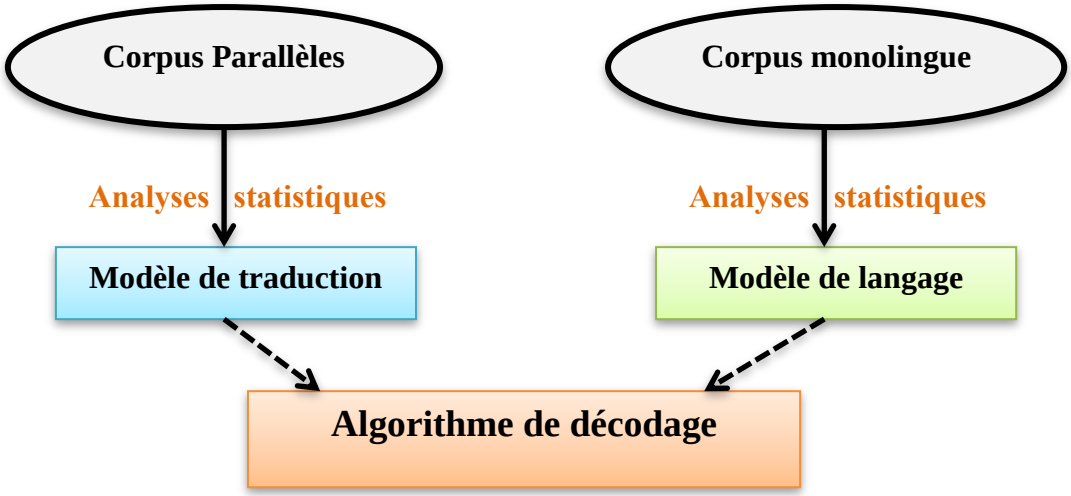


Figure1: processus de la traduction automatique statistique

2.4.1. Equation fondamentale

Le but de la traduction automatique statistique est de trouver la phrase cible *t* la plus probable sachant la phrase source *s*.

$$P(t|s)=\frac{P(s|t)\times P(t)}{P(s)}$$

P(s) ne dépend pas de la phrase cible *t*, il n'a donc aucune influence sur le calcul de la fonction *argmax*. P(s) peut donc être négligée, et l'on obtient par suite l'équation fondamentale suivante:

$$t^*=\operatorname{argmax}_t P(t|s) = \operatorname{argmax}_t P(s|t)\times P(t)$$

Dans cette formule, P(t) désigne le modèle de langage qui est construit à partir d'un corpus d'apprentissage monolingue et P(s|t) le modèle de traduction qui repose sur un corpus d'apprentissage parallèle.

2.4.2. Modèle de langage

Parmi les modèles de langages utilisés dans les systèmes de TAS les principaux sont le modèle n-gramme, le modèle Cache (Kuhn et al.1990) et le modèle Trigger (Lau, 1993). Le modèle Cache repose sur les dépendances des mots non contigus. Quant à lui, le modèle Trigger consiste à déterminer le couple de mots (*X*, *Y*) où la présence de *X* dans l'historique déclenche l'apparition de *Y*.

Toutefois, le modèle n-gramme ($1 \leq n \leq 5$) reste le plus utilisé dans les systèmes de traduction actuels et plus précisément le modèle trigramme (3-gramme) pour le traitement des langues européennes. De fait, le modèle n-gramme permet d'estimer la vraisemblance d'une suite de mots en lui attribuant une probabilité.

Soit $t = w_1 w_2 \ldots w_k$ une séquence de k mots dans une langue donnée et n la taille maximale des n-gramme ($1 \leq n \leq 5$) , la formule de p(t) est exprimée en :

$$P(t) = \prod_{i=1}^{k} (w_i | w_{i-1} w_{i-2} \cdots w_{i-n+1})$$

2.4.3. Modèle de traduction

Le modèle de traduction est représenté par la probabilité **P(t|s)** dans l'équation fondamentale de la traduction automatique statistique, où *t* est la traduction de la phrase source *s*. ces phrases sont extraites à partir d'un corpus parallèle aligné au niveau des phrases. Autrement dit, chaque phrase rédigée dans une langue source correspond à une phrase dans la langue cible. Ces deux dernières doivent être segmentées en plus petites unités, afin que les unités de la phrase cible soient mises en correspondance avec les unités de la phrase source. On utilise pour cela une technique d'alignement essentielle pour produire un modèle de traduction.

2.4.3.1. Modèles de traduction à base de mots

Cinq modèles de traduction à base de mots ont été proposés par IBM et présenté (Brown et al.1991), (Brown et al.1993), etc. Le but de ces modèles est d'aligner le corpus bilingue au niveau des mots.

Comme le montre la figure ci-dessous, la méthode d'alignement permet à un mot source d'être aligné avec un ou plusieurs mot(s) cible(s) (n-à-1), mais pas le contraire (1-a-n).

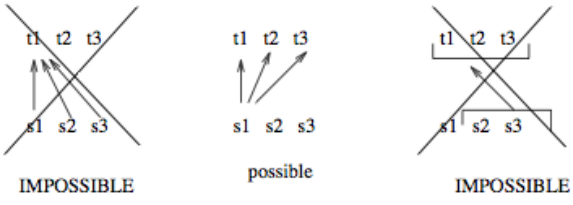


Figure 2 : Limite de l'alignement à base de mots

2.4.3.2. Modèles de traduction à base de segments

La construction d'un modèle de traduction à base segments nécessite la succession de ces trois étapes :

- découper la phrase en séquences de mots.
- traduire les séquences de mots en se basant sur la table de traduction.
- Réordonner les séquences à l'aide d'un modèle de distorsion.

Vu qu'il y a divers problèmes posés par la méthode d'alignement à base de mots (*Och* et *Ney,* 2002) présentent une méthode d'alignement en segments de phrases. Le segment devient alors l'unité sur laquelle se fonde la traduction.

La figure ci-dessous représente un exemple d'alignement à base de segments :

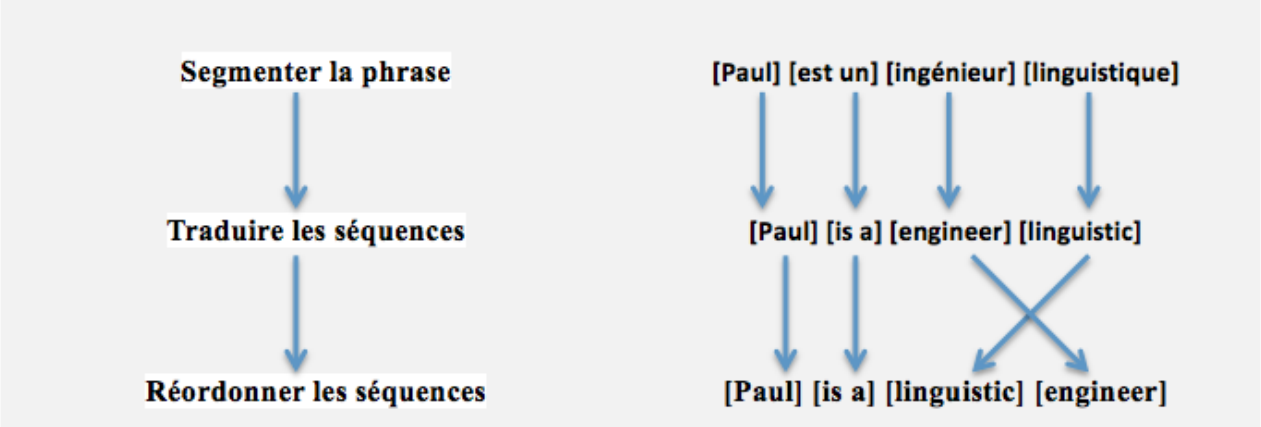


Figure 3: Exemple d'alignement à base de segments

Les principaux composants de l'équation des modèles de traduction basée sur les segments sont :

- modèle de langage P (t)
- modèle de traduction P(s,t)
- modèle de distorsion Ω (s|t)

Avec $P(t|s) = P(t) \times P(s,t) \times \Omega(s|t)$

2.4.4. Modèle de distorsion

Le modèle de distorsion ou de ré-ordonnancement permet de capturer le ré-ordonnancement des hypothèses de traduction. Il sert ainsi à gérer l'alignement des segments produits par la traduction. En effet, les séquences de mots dans la phrase cible ne doivent pas être ordonnées de la même façon que celles de la phrase source.

2.4.5. Décodeur

Le décodeur représente la partie la plus importante des systèmes de traduction statistiques, c'est le processus qui va implémenter la fonction *argmax* de l'équation fondamentale de TAS, pour extraire la ou les traductions les plus probables à l'aide de ses paramètres, qui combinent le modèle de traduction, le modèle de langage, et le modèle de distorsion.

2.4.6. Evaluation automatique

Il est indispensable de posséder une ou plusieurs traductions qui seront considérées comme référence lors de l'évaluation automatique d'une hypothèse de traduction de la phrase source. De ce fait, les mesures automatiques permettent d'estimer le niveau de similarité entre la traduction produite par le système et celle(s) de référence. Ainsi la qualité de la traduction de référence est un facteur important pour assurer une bonne évaluation automatique. Dans ce mémoire de recherche, nous allons définir la qualité de la traduction des systèmes en terme de score BLEU, la mesure la plus souvent utilisée dans la communauté de la traduction automatique.

2.4.6.1. Score BLEU

Le score BLEU (en anglais : Bilingual Evaluation Understudy) a initialement été proposé par Papineni et al. en 2001.C'est un algorithme utilisé en vue d'évaluer la qualité des hypothèses de sortie produites par un système de traduction automatique.

En effet, le concept est fondé sur l'idée de comparer l'hypothèse de traduction avec une ou plusieurs références au niveau des mots, des bigrammes, trigrammes etc.

Le score BLEU est normalisé entre 0 et 1, et il est exprimé généralement en pourcentage. Notons qu'une traduction humaine peut parfois obtenir un mauvais score BLEU, si elle s'écarte de la référence.

2.4.7. Moses

Moses est une boite à outils libre disponible gratuitement sur le site officiel[9]. Cette boite à outils est basée sur l'approche statistique de la traduction automatique. L'avantage d'utiliser Moses est que nous pouvons ajuster et manipuler notre système de traduction en fonction de nos besoins.

Comme il est décrit dans le manuel de Moses (Koehn, 2014), cette boite à outils est constituée principalement de deux composants:

- **Processus d'entrainement:** consistant à construire un modèle de langage à partir des données parallèles à l'aide d'une collection de codes écrits généralement en langages C++ et Perl.

- **Décodeur :** c'est une application C++ permettant de déterminer la traduction la plus probable d'une phrase source selon le modèle de traduction.

En outre, pour construire un système de traduction statistique, Moses prend en charge plusieurs outils comme :

-MGIZA++: c'est un outil *multithread* basé sur GIZA++ qui implémente les modèles IBM sans consommer beaucoup de mémoire en éliminant les tables répétées (plus efficace donc que GIZA++[10]).

- **SRILM :** c'est une boîte à outils dédiée pour la construction des modèles de langage statistiques.

2.5. La TAS et les expressions polylexicales

Les expressions polylexicales peuvent être vues comme un problème clé de la technologie actuelle en TAL. Examinons maintenant plus en détail quelles sont les difficultés et les problèmes posés par la polylexicalité pour la traduction automatique statistique, et quelles solutions ont été proposées dans la littérature.

2.5.1. Limites

Dans cette section, nous allons montrer quelques traductions erronées au niveau des EPL (*phrasal verbs* en anglais, expressions idiomatiques et collocations en français) produites par le système de traduction statistique *Google Traduction.* Ces expressions ont été identifiées à partir

[9] www.statmt.org

[10] http://www.kyloo.net/software/doku.php/mgiza:overview

des dictionnaires électroniques en ligne. Les phrases ont été extraites à partir de la ressource textuelle d'*EmoConc*.

2.5.1.1. Les *phrasal verbs*

Exemple 1 :

La phrase source : « He wanted to bring that soup up. »

La traduction de *Google TR*: « Il voulait apporter cette soupe jusqu'à. »

La traduction correcte : « Il avait envie de rendre cette soupe. »

Exemple 2 :

La phrase source : « Surely they must call the operation off now? »

La traduction de *Google TR* : « Certes, ils doivent appeler l'opération maintenant? »

La traduction correcte : « Sans doute veulent-ils annuler l'opération maintenant ?»

=>*Google Traduction* a traduit les *phrasal verbs* ci-dessus mot-à-mot. Il n'a pas réussi à détecter ces *phrasal verbs* pour les traduire correctement.

2.5.1.2. Les expressions totalement figées

Exemple 1 :

La phrase source : « Même si je suis très soupe au lait, je ne me suis pas mis en colère. »

La traduction de *Google TR* : « Although I am very milk soup, I am not angry. »

=>L'expression *soupe au lait* dans la phrase source signifie que l'agent a un caractère vif, par contre, cette expression est traduite en anglais par *milk soup* qui n'a pas aucun sens dans ce contexte.

Exemple 2 :

La phrase source :« Ils étaient tombés dans les pommes sur un lit. »

La traduction de *Google TR*: « They had fallen in apples on a bed. »

La traduction correcte : «They fainted on a bed »

=>*fallen in apples* n'est pas la traduction pertinente de l'expression « tombés dans les pommes » qui signifie que l'individu s'évanouit.

Exemple 3 :

La phrase source : « Il ne veut *pas* avoir travaillé pour le roi de Prusse. »

La traduction de *Google TR* : « he does not want to have worked for the King of Prusse. »

La traduction correcte : «He hates having profited nothing. »

=>L'expression idiomatique soulignée dans la phrase source désigne que la personne ne veut pas travailler pour rien, mais elle est traduite par *Google Translate*.

Google Translate n'a pas pu traduire ces expressions polylexicales figées, répertoriées dans les dictionnaires, qui ne posent guère de problème d'identification ou d'ambiguïté.

2.5.1.3. Les collocations (semi-figées)

Exemple 1 :

La phrase source : « L'honorable parlementaire me fait trop d'honneur.»

La traduction de *Google TR*: « The honorable member made me too much honor.»

La traduction correcte : « The honorable member is being too kind to me.»

Exemple 2 :

La phrase source : «Dois je vraiment vous dresser un catalogue de la criminalité itinérante?»

La traduction de *Google TR* « Should I really get a catalog of shifting crime?»

La traduction correcte : «Should I really make a catalogue of shifting crime ? »

=>La traduction mot-à-mot des collocations aboutit à une traduction incorrecte de ces dernières.

2.5.2. Stratégies d'intégrations des EPLs dans un système de TA

2.5.2.1. Stratégie statique

La méthode statique *Forced* (proposé par le décodeur *Moses*) consiste à identifier les expressions polylexicales en amont et à les transformer en une seule unité (sous format XML) dans le corpus, permettant au décodeur Moses d'utiliser les traductions du modèle ou bien la traduction XML spécifiée (on parle alors de « traduction forcée »).

Exemple :<EPL translate="*flares up easily*" probabilité="0.8">*soupe au lai*t</EPL>

2.5.2.2. Stratégies dynamique

Bouamor (2013) a présenté quelques méthodes pour l'intégration dynamique des EPLs dans les systèmes de TAS.

Il s'agit de trois méthodes dynamiques :

- *Nouveau modèle de traduction (Train)*: cette méthode consiste à considérer les EPLs traduites comme des couples de phrases parallèles et à les intégrer au corpus d'entrainement pour construire un nouveau modèle de traduction.

- *Trait* : cette méthode consiste à définir des traits binaires (0 et 1) en indiquant pour chaque entrée de la table de traduction s'il s'agit d'une EPL ou pas.

- *Extension de la table de traduction (Table) :* dans cette méthode Bouamor (2013) consiste à étendre la table de traduction avec des EPLs acquises. Ainsi, elle utilise un score de confiance qui est calculé sur la base de mesure de l'indice de Jaccard pour chaque paire d'EPLs, en estimant une probabilité de traduction pour chaque paire d'EPL et en attribuant une probabilité lexicale égale à 1 pour chaque EPL, afin que le décodeur prenne en charge les EPLs bilingues lors de la recherche de segments candidats.

Dans son travail, elle extrait des EPLs bilingues à partir d'un corpus parallèle Français/Anglais en identifiant chaque EPL dans le corpus. Ensuite, elle propose un algorithme d'alignement à base de segments qui prend en charge les correspondances bilingues des expressions polylexicales. Pour l'intégration de ces expressions dans le système de TAS, elle exploite la méthode statique (décodeur en mode forcé proposé par Moses) et les trois autres méthodes dynamiques citées précédemment. Ses résultats montrent que la méthode « Trait » peut améliorer significativement la qualité de traduction de *MOSES* (+0,23 point BLEU).

2.5.3. Travail existant sur l'examination de la TA des *Phrasal verbs*

Dans le cadre de son mémoire de recherche, *Kobzar (2013*) examine la qualité de la traduction des verbes à particule (PV), en comparant deux systèmes de traduction automatique statistiques basées sur les séquences de mots et les séquences de mots hiérarchiques dans la traduction des expressions polylexicales, afin de déterminer les facteurs influençant la qualité de la traduction.

Dans son travail, A. Kobzar se concentre sur les verbes à particule en anglais (PV) avec certaines contraintes (verbe + objet + particule sans incorporation d'un verbe au milieu) extraites à l'aide de l'outil *MWETOOLKIT* à partir d'un corpus anglais, afin de s'assurer que la particule dépende syntaxiquement du verbe. L'analyseur syntaxique *RASP PARSER* de l'outil *MWETOOLKIT* pose des problèmes pour les verbes à particules qui possèdent un verbe au milieu.

Exemple d'identification erronée du PV *set up* :

The Committee on Legal Affairs **set** about taking **up** the challenge.

Dans cet exemple la particule *up* dépend du verbe *take* et non pas du verbe *set*. Ce type d'identification doit être ignoré dans l'intégration des PV dans le système TAS.

A. Kobzar a évalué son système de traduction d'une manière automatique et manuelle (avec 9 évaluateurs volontaires). Il en déduit que la fréquence des verbes à particule influence la qualité de traduction. Par ailleurs, les paramètres d'évaluation automatique ne donnent aucune indication sur la nature de l'erreur présente dans la traduction. Ainsi, le système à base de séquences (*Phrase-Based*) est pluspertinent que le système à base de séquences hiérarchiques (*Hierarchical Phrase-Based*) pour la traduction des *phrasal verbs*.

2.6. Notre approche

Dans ce mémoire de recherche, nous allons aborder les questions principales suivantes: Quelles EPL posent le plus de problèmes ? Comment intégrer leur traitement à un système de TAS ? Et puis, comment évaluer les améliorations liées à ces traitements ?

2.6.1. Méthode

Avec une traduction « mot-à-mot », un système de traduction statistique ne peut pas produire une traduction correcte des expressions polylexicales. Par la suite, nous allons utiliser le modèle de traduction à base de segments et le modèle de langage n-gramme pour élaborer notre système à l'aide de l'outil Moses. En outre, la fréquence d'existence des EPLs dans le corpus de test est importante pour évaluer notre système.

Grâce à la grammaire complexe de l'outil Lexicoscope qui nous permet de chercher une collection de phrases contenant quelques types de phénomènes, nous allons construire un corpus spécifique pour le test qui soit riche en expressions polylexicales. En outre, nous sélectionnerons manuellement les expressions qui sont traduites par le système Google Translate.

2.6.2. Pistes pour traiter les expressions idiomatiques

Pour traiter les expressions idiomatiques, nous proposons deux méthodes, la première sert à post-éditer le corpus de test en identifiant ces expressions a partir d'un dictionnaire bilingue, afin de spécifier la bonne traduction de chaque expression idiomatique dans le but d'appliquer la méthode de décodage forcée proposé par la boite à outils *Moses.* La deuxième méthode consiste à considérer le couple d'expression idiomatique (expression et sa traduction) du dictionnaire bilingue comme une paire de phrases, et à les ajouter dans les données d'apprentissage pour entrainer de nouveau le système de traduction et avoir un nouveau modèle de traduction.

2.6.3. Pistes pour traiter les collocations

Tout d'abord, nous allons identifier les expressions semi-figées (collocations) dans les corpus d'apprentissages et le corpus de test français-anglais. En effet, l'identification des collocations nécessite une analyse syntaxique pour avoir les informations linguistiques appropriées pour chaque forme (relations de dépendances, lemme, catégorie, etc). En outre, nous allons utiliser l'outil *MGIZA++* pour aligner les expressions identifiées. Si une expression est alignée avec plusieurs traductions dans la langue cible, alors nous allons garder l'expression ayant le plus grand score d'alignement affecté par *MGIZA++*. Enfin, nous allons créer une liste des collocations bilingues pour traiter ces expressions par la méthode *Forced* (méthode statique) proposée par le décodeur de *Moses*.

Par ailleurs, nous allons faire le même traitement en utilisant une méthode dynamique (*Merged*) qui sert à analyser syntaxiquement les corpus d'apprentissage et le corpus de test pour identifier les collocations contenant certaines contraintes linguistiques. Cette méthode repose sur des traits affectés aux composantes de la collocation. Par exemple, pour une collocation composée de 2 mots, nous allons affecter pour chaque mot sa catégorie correspondante ainsi qu'un identifiant (chiffre) : *mot1 [id1,C1] mot2 [id2,C2]*. Ensuite, nous allons ajouter une information dans une balise *XML* pour dire que les mots *de id1+id2* forment en soi une collocation. Avec cette méthode nous allons forcer l'alignement, pour considérer le *mot1 mot2* comme un seul segment. Enfin, nous allons traduire le corpus de test afin d'évaluer la fiabilité de cette méthode.

3. Préparation des corpus (Enrichissement d'EmoConc)

Pour commencer, nous avons besoin de construire nos propres corpus de test riches en EPLs, nécessaires pour évaluer et comparer différents systèmes de TA par rapport à ce problème particulier. EmoConc peut nous permettre d'élaborer des requêtes complexes afin d'identifier des expressions polylexicales non contigües.

Mais le corpus interrogeable via cet outil nécessite d'être complété. En effet, les corpus parallèles du projet *Emolex* se basent principalement sur des ressources littéraires.

Corpus	Nombre de mots	Nombre de textes
Français-Anglais	1 097 667	7
Anglais-Français	16 808 073	95

Tableau 1 : Description du corpus EN-FR du projet *Emolex*

Pour enrichir et diversifier nos données, nous allons intégrer de nouveaux corpus : *Europarl*, *TED* et *News Commentary*.

Dans cette section, nous allons présenter en premier lieu les nouveaux corpus, ensuite nous détaillerons les processus de prétraitement et les annotations effectuées en vue d'intégrer les corpus dans *Emolex*, pour avoir finalement une ressource lexicale riche et multi-domaine.

3.1. Présentation des corpus

3.1.1. Europarl

Le corpus *Europarl*[11] est un corpus parallèle Français-Anglais. Il s'agit d'une retranscription des sessions plénières du parlement européen. Il a été développé par un groupe de chercheurs dirigé par Philipp Koehn à l'Université d'Edimbourg. Initialement, ce corpus a été conçu à des fins de recherche en traduction automatique statistique (Koehn, 2005). Il comprend des versions dans 21 langues européennes (danois, néerlandais, anglais, finnois, français, allemand, grec, italien, portugais, espagnol et suédois, etc.).

Dans ce mémoire de recherche, nous avons utilisé le corpus Français-Anglais de la version 7 qui est la plus récentes du corpus *Europarl*.

[11]http://www.statmt.org/europarl

Europarl V7	Nombre de mots	Nombre de paires de phrases
Français-Anglais	51 388 643	2 007 723
Anglais-Français	50 196 035	

Tableau 2 : description du Corpus Europarl V7

Dans sa version 7, Europarl propose deux versions de corpus, une sous format texte brut et une autre version segmentée et alignée sous la forme de fichiers XML.

3.1.2. News Commentary

Le corpus *News commentary*[12](nous abrégerons désormais en *News*) est un corpus parallèle aligné au niveau des phrases. Ce corpus contient des extraits de diverses publications de presse et de commentaires du projet Syndicate[13]. Il existe des versions en anglais, français, espagnol, allemand, et tchèque.

News	Nombre de mots	Nombre de paires de phrases
Français-Anglais	4 086 635	157 168
Anglais-Français	3 535 313	

Tableau 3 : description du Corpus News

3.1.3. TED Talks

TED Talks[14] est un ensemble de transcriptions des conférences en anglais présentés sous format vidéo sur le site officiel de TED. Ces transcriptions ont été traduites par les bénévoles pour plus de 70 autres langues (arabe, français, italien, coréen, portugais, etc.).

[12]http://www.statmt.org/wmt07/shared-task.html

[13]http://www.project-syndicate.org/

[14]11www.ted.com

TED	Nombre de mots	Nombre de paires de phrases
Français-Anglais	2 725 851	155 389
Anglais-Français	2 599740	

Tableau4 : description du Corpus TED

3.2. Prétraitement des corpus

Les corpus *News, TED* et *Europarl* ont été créés et post-édités par des humains, mais ils comportent encore des données bruitées qui doivent être nettoyées et préparées avant leur exploitation. De ce fait, nous avons appliqué les modifications suivantes afin d'améliorer la qualité des données :

- Conversion des entités HTML: nous avons commencé par la conversion des entités HTML tel que : - qui sera remplacé par un tiret '-'. Cette conversion a été effectuée en utilisant des expressions régulières.

- Élimination des symboles : nous avons supprimé les caractères spéciaux tels que ♫ *et* ♪.

- Filtrage des phrases : nous avons développé un script *Perl* pour supprimer les lignes vides ou qui contiennent que des ponctuations.

- Correction de l'encodage: comme le corpus est destiné à subir une chaîne de traitement complexe impliquant étiquetage morphosyntaxique et analyse syntaxique, il est important de corriger dès les départ d'éventuels problèmes d'encodage. Un problème d'encodage pour quelques caractères, pourrait compromettre le processus d'annotation et d'intégration. Nous nous sommes appliqué à corriger l'encodage de certains caractères accentués pour chaque corpus.

Exemple : **Ã©**ducation =>**é**ducation.

- Découpage des corpus: étant donné que les nouveaux corpus sont volumineux et qu'il y a des opérations d'indexation couteuses dans la chaine d'intégration, nous avons découpé les corpus TED et News en plusieurs fichiers textes, contenant chacun 1 000 lignes. Par ailleurs, nous avons utilisé la version XML segmentée du corpus Europarl qui contient 9 442 paires de fichiers parallèles.

- Ajout des entêtes XML: nous avons ajouté des entêtes au format XML indiquant le nom du fichier, la langue source et la langue cible, afin d'avoir les informations nécessaires dans chaque fichier, et de se conformer au schéma TEI requis par le Lexicoscope.

```
<header>
        <fileDesc>
                <titleStmt>
                        <title value="Erp.14.xml"/>
                        <author value="Erp"/>
                    <translation source_language="fr" source_title="" translator="" date="" />
                </titleStmt>
                <publicationStmt>
                        <publisher value="" />
                        <pubPlace value="" />
                        <pubDate value="" />
                        <pubURL value="" />
                        <pubNumeric value="" />
                </publicationStmt>
                <formatSource value="" />
        </fileDesc>
        <profileDesc>
                <langUsage>
                        <language ident="en" />
                </langUsage>
                <textDesc type="" sub_genre="" thema="" />
                <annotation value=""/>
                <wordsNumber value=""/>
        </profileDesc>
</header>
```

Figure 4 : Exemple d'entête XML

3.3. Annotation des corpus

Dans le cadre du projet Emolex, des annotations ont été produites grâce à l'analyseur XIP[15]de *Xerox*, qui a été utilisé pour annoter les corpus parallèles et monolingues en anglais.

Le parseur XIP, en anglais « *Xerox Incremental Parser* », est un analyseur qui prend en entrée un texte et qui fournit en sortie des informations linguistiques relatives à chaque élément du texte. XIP permet d'enrichir les entrées lexicales de traits morphosyntaxiques, d'identifier des *chunks* et d'autres types de constituants, ou encore d'identifier des relations de dépendance entre les mots.

Nous avons annoté nos 3 corpus, Europarl, TED et News. Mais, lors de l'annotation, nous avons rencontré des problèmes pour conserver la segmentation originale des corpus qui sont déjà alignés au niveau des phrases. Donc, nous avons ajouté des marques de segmentation qui sont récupérées ensuite en sortie de *XIP*, puis nous avons converti les sorties en XML, en utilisant le même type de schéma que pour le corpus Emolex.

Comme on le voit dans l'exemple de sortie post-traité de XIP ci-dessous, les informations produites pour chaque forme sont : l'identifiant, la catégorie, le lemme, les traits ainsi qu'une information d'espacement. En outre, XIP fournit des relations de dépendances syntaxiques, que nous avons encodées grâce aux balises <d> représentant chacune une relation entre une tête (attribut *h*) et une forme dépendante (attribut *d*).

[15]open.xerox.com/Services/XIPParser

```
<p id="3">
        <s id="3">
                <tc>
                        <t id="1" c="PREP" l="in" f="CR1+Adv+notly+ADV" e="">in</t>
                        <t id="2" c="QUANT" l="less" f="Quant+Comp+QUANTCMP" e=" ">less</t>
                        <t id="3" c="CONJ" l="than" f="Conj+Coord+COTHAN" e=" ">than</t>
                        <t id="4" c="NUM" l="two" f="Num+Card+CARD" e=" ">two</t>
                        <t id="5" c="NOUN" l="hour" f="Noun+countable+Pl+NOUN" e=" ">hours</t>
                        <t id="6" c="PUNCT" l="," f="Punct+Comma+CM" e=" ">,</t>
                        <t id="7" c="PRON" l="they" f="Pron+Pers+Nom+3P+Pl+PRONPERS" e=" ">they</t>
                        <t id="8" c="AUX" l="would" f="Aux+Elid+VAUX" e=" ">'d</t>
                        <t id="9" c="VERB" l="cut" f="Verb+s_sc_pacross+s_sc_pthrough+s_sc_pon+s_sc_pabout+s_p_up+s
">cut</t>
                        <t id="10" c="PRON" l="something" f="Pron+NomObl+3P+Sg+PRON" e=" ">something</t>
                        <t id="11" c="ADV" l="else" f="Adv+notly+ADV" e=" ">else</t>
                        <t id="12" c="ADV" l="off" f="Adv+notly+ADV" e=" ">off</t>
                        <t id="13" c="PUNCT" l="." f="Punct+Sent+SENT" e=" ">.</t>
                </tc>
                <dc>
                        <d t="MOD_PRE_TEMP" h="9" d="5"/>
                        <d t="MOD_POST" h="9" d="11"/>
                        <d t="MOD_POST" h="9" d="12"/>
                        <d t="MOD_PRE" h="4" d="2"/>
                        <d t="QUANTD" h="5" d="4"/>
                        <d t="SUBJ_PRE" h="8" d="7"/>
                        <d t="NUCL_VLINK_MODAL" h="8" d="9"/>
                        <d t="OBJ_POST" h="9" d="10"/>
                        <d t="MAIN_MODAL" h="9" d=""/>
                </dc>
        </s>
</p>
```

Figure 5 : Exemple de sortie post-traitée de XIP

3.4. Alignement

Afin de finaliser le processus de prétraitement des corpus, il est nécessaire de produire les fichiers d'alignement qui servent à déterminer les liens entre chaque phrase source et sa traduction.

Au cours de l'annotation nous avons trouvé une difficulté de conserver la segmentation originale pour certaine corpus qui sont déjà alignés, ce qui pouvait produire, pour un couple de fichiers, un nombre de phrases différent. Pour éviter ce problème, nous avons ajouté des marques de segmentation à la fin de chaque phrase qui sont récupérées ensuite en sortie de *XIP*. Mais, il reste encore du bruit lors de l'annotation, car *XIP* peut diviser une phrase en plusieurs à cause d'un certain type de ponctuations qui ce trouvent au milieu de la phrase.

Par conséquent, nous avons établi des fichiers d'alignements pour les corpus qui posent des problèmes de segmentations en utilisant l'outil *Alinea*[16] développé par (Kraif, 2005). *Alinea* est un programme dédié à la constitution et l'édition de corpus bilingues alignés. Quand l'alignement d'origine était correct, nous avons créé nous-même les fichiers d'alignement avec une relation (i;i) avec i id des phrases source et cible .

[16]olivier.kraif.u-grenoble3.fr/

Information d'alignement	<link xtargets="1 ; 1" />	
Source	**Id=1**	It can be a very complicated thing, the ocean.
Cible	**Id=1**	Ca peut être très compliqué, l'océan.

Tableau 5: Exemple de relation (i;i)

Information d'alignement	<link xtargets="6 7 ; 6" />	
Source	**Id=6**	And I'm going to start with this one:
Cible	**Id=7**	If momma ain't happy, ain't nobody happy.
	Id=6	Et je vais commencer par celui-ci: Si maman n'est pas contente, personne n'est content.

Tableau 6 : Exemple de sortie de l'outil *Alinea*

3.5. Intégration

Après avoir terminé le prétraitement des corpus, l'annotation et la création des fichiers d'alignements, nous avons suivi la chaine de traitement de (Kraif, 2013), qui a été développé dans le cadre du projet Emolex.

Le tableau ci-dessous illustre l'enrichissement (*4) de la taille des corpus anglais-français et français-anglais interrogeables avec EmoConc.

	Avant l'intégration		**Après l'intégration**	
	EN-FR	**FR-EN**	**EN-FR**	**FR-EN**
Taille du corpus EmoConc	16 808 073 mots	1 097 667 mots	55 868 296 mots	16 767 151 mots
Total	18 846 627 mots		73 576 334 mots	

Tableau 7 : Evolution de la taille du corpus EmoConc

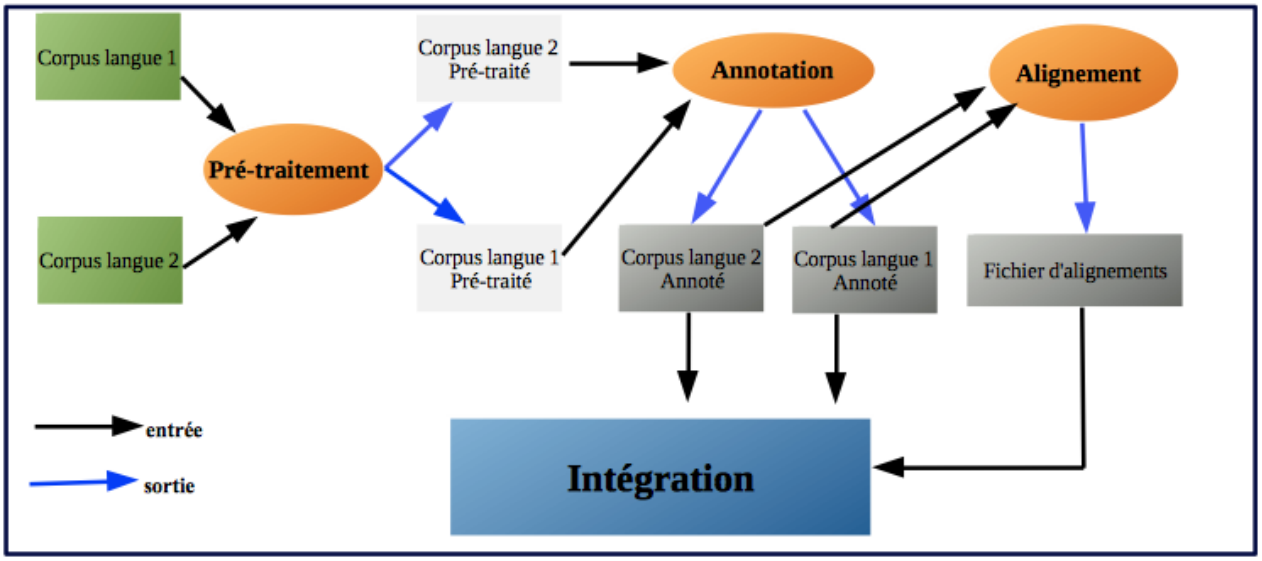


Figure 6 : Processus d'intégration des corpus dans Emolex

Dans cette partie, nous avons présenté les différents corpus utilisés, ainsi que le processus de prétraitement de ces derniers. Ensuite, nous détaillé les étapes de l'annotation et de la création des fichiers d'alignement, nécessaires pour l'intégration des corpus dans EmoConc. Nous avons désormais une nouvelle ressource couvrant des genres variés : textes littéraires, débats, conférences, articles de presse.

Nous disposons donc de données suffisantes pour la construction de nos corpus de test, qui doivent présenter de véritables difficultés sur le plan de la traduction des unités phraséologiques, afin de constituer un *benchmark* utile pour évaluer et faire progresser les systèmes de traduction automatique statistique dans ce domaine.

4. Construction des Corpus spécifiques pour l'évaluation de la TAS

Comme il est mentionné dans la partie 2, les expressions polylexicales constituent un véritable problème pour plusieurs applications de TAL, notamment pour la traduction automatique statistique. D'autre part, l'identification des EPL dans les corpus nécessite une validation humaine.

Dans ce chapitre nous allons décrire la méthodologie que nous avons suivie pour sélectionner les EPLs afin de construire nos corpus de tests (Français-Anglais et Anglais-Français) qui comportent des phrases posant des problèmes pour les systèmes de traduction statistiques au niveau des expressions polylexicales. Evidemment, ces corpus ont été extraits à partir de la nouvelle ressource lexicale interrogeable sous EmoConc.

4.1. Choix des expressions polylexicales

En vue de constituer notre corpus de test Anglais-Français nous avons créé une liste contenant deux types d'expressions polylexicales en anglais qui sont les « *phrasal verbs* » (nous noterons désormais PV) ou verbe à particule en français, et les « *idioms* » ou expressions idiomatiques en français (totalement figées).

En outre, pour construire le corpus de test Français-Anglais, nous avons élaboré une autre liste des EPLs en français comportant :

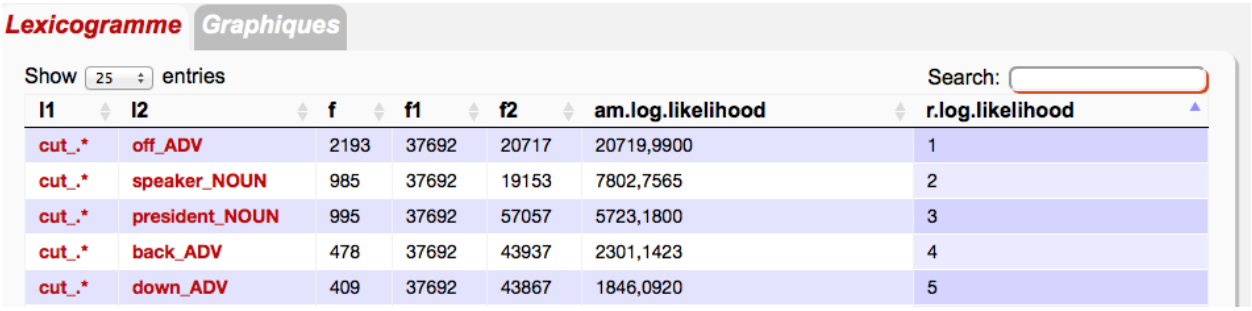
- des verbes support qui ont un sens plutôt vide (*avoir+peur*, *faire+guerre,* etc.), où le verbe sert de support à un prédicat nominal.

- des collocations de type Verbe+Nom où le nom est un objet (*forcer+admiration, annoncer+nouvelle, régler+conflit,* etc.)

- des collocations de type Nom+ Verbe où le nom est en position de sujet (*tension+monter, chance+tourner, jour+se lever, nuit+tomber* ,etc.).

- des expressions idiomatiques (*avoir la haine, accuser le coup, tourner sa langue,* etc.).

Ces listes ont été élaborées grâce à des dictionnaires électroniques disponibles sur Internet, des forums, etc. Par ailleurs, nous avons les lexicogrammes calculés par EmoConc pour extraire des collocations (méthode semi-automatique), en précisant un pivot et éventuellement une catégorie grammaticale ainsi que des relations syntaxiques. Par exemple pour extraire un PV qui se base sur le verbe *cut*, nous n'avons retenu comme collocatif que des particules de catégorie adverbe et préposition.


Lexicogramme Graphiques

Show 25 entries Search:

l1	l2	f	f1	f2	am.log.likelihood	r.log.likelihood
cut_.*	off_ADV	2193	37692	20717	20719,9900	1
cut_.*	speaker_NOUN	985	37692	19153	7802,7565	2
cut_.*	president_NOUN	995	37692	57057	5723,1800	3
cut_.*	back_ADV	478	37692	43937	2301,1423	4
cut_.*	down_ADV	409	37692	43867	1846,0920	5

Figure 7 : Extraction des *phrasal verbs* (*Lexicogramme*)

4.2. Extraction du Corpus de Test

4.2.1. Méthode d'extraction

Après l'intégration des nouveaux Corpus Europarl, News et TED, nous passons à l'étape de la recherche avec EmoConc des EPLs sélectionnées[17], suivant des distances variées entre 0 et 5 pour les Phrasal Verbs en anglais, et des distances de 0 à 3 pour les collocations en français. Dans cette étape de recherche, nous avons lancé nos requêtes directement sur le site *EmoConc*, car il était plus simple de copier/coller les phrases à partir de la sortie HTML (plutôt que d'extraire les résultats en XML à partir des scripts disponibles sur le serveur).

Ainsi, pour chaque requête lancée, nous vérifions tout d'abord si le nombre d'occurrences est supérieur à 5, si c'est le cas, nous choisissons manuellement des phrases contenant des vraies EPLs. Ensuite nous passons à l'étape de traduction automatique en utilisant *Google Traduction* ; Si l'expression polylexicale est mal traduite par ce système, nous gardons la phrase, et ainsi de suite. Sachant que, nous obtenons à la fin entre 5 et 10 phrases comportant une même EPL.

Par exemple pour chercher des phrases contenant le PV *set up***,** il suffit de lancer la requête suivante :

<l=set,#2><>{0,5}<l=up,#1>::(.*,2,1)

[17]http://emolex.u-grenoble3.fr/emoConc/index.php

Contexte	
Distance = 0	He was ***setting up*** an alliance with the Uzbeks and the Russians .
Distance = 1	She probably felt like Wiggin had poured it on ,deliberately ***setting*** her ***up*** for humiliation .
Distance = 2	It had taken some time to***set*** it all ***up*** properly ,but for a dozen years now ,every scrap of dubious information had been followed up ,and ,every once in a while ,something seemed sufficiently questionable to be referred to Pitt .
Distance = 3	You ***set*** all of it ***up*** and paid for it ,I said as I became more convinced and my incredulity grew .
Distance = 4	Notehow the balconies are **setback** in steps right**up** to the ring of spires .
Distance = 5	Jeb **set** the flashlight **down** ,bulb **up**.

Tableau 8 : Exemple de résultat de la concordance pour *set up*

Comme le montre le tableau ci-dessus, les résultats trouvés ne contiennent pas tous de vraies PV (cf. le cas où la distance=4 et 5 dans le tableau 7). Evidemment, plus la distance est grande, plus la probabilité d'avoir un vrai PV devient faible.

Au final, nous avons construit deux corpus de test Français-Anglais et Anglais-Français qui sont cohérents et exploitables, chaque corpus étant constitué de 500 paires de phrases parallèles issues des corpus *EMOLEX*, *TED*, *NEWS* et *EUROPARL*. Ces dernières comportent des expressions polylexicales (des verbes à particules, des expressions figées, des collocations etc.) et posent des problèmes de traduction à *« Google Traduction »* au niveau sémantique.

4.2.2. Corpus anglais-français

Le corpus de test anglais-français (Tst-EPL) présente 500 paires de phrases contenant deux types d'EPLs extraites des différents corpus (*Emolex, Europarl, TED et News*) :

- 40 PV qui représentent 73.2 % du corpus (nommé pv366), avec des distances variant de 0 à 5.
- Des expressions idiomatiques constituant 26,8% du corpus (134 phrases) parmi lesquelles il y a 74 expressions totalement figées.

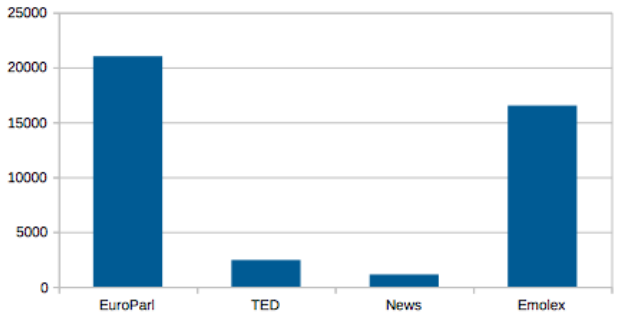


Figure 8 : Occurrences des 40 *phrasal Verbs*

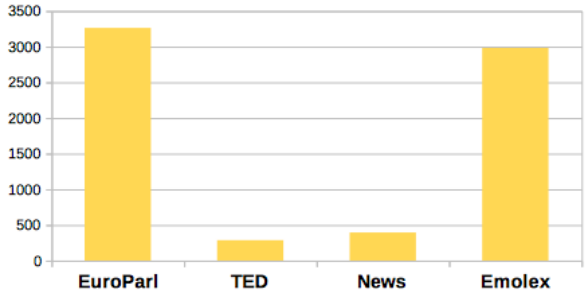


Figure 9 : Occurrences des *idioms*

Étant donné que les *phrasal verbs* sont assez fréquents dans les corpus *Europarl* et *Emolex*, notre corpus de test Anglais-Français se base essentiellement sur ces deux derniers, ce qui est décrit dans le tableau suivant :

Origine	Détail du corpus Tst-EPL (phrases)	Détail des « Phrasal Verbs » dans le corpus Tst-EPL(phrases)	Détail des « idioms » dans le corpus Tst-EPL (phrases)
Emolex	272	189	83
Europarl	213	163	50
News	15	14	1
TOTAL	**500**	**366**	**134**

Tableau 9 : Distribution du corpus de teste Anglais-Français

4.2.3. Corpus français-anglais

Le corpus français-anglais contient de même 500 paires de phrases parallèles extraites à partir des corpus *Emolex*, *Europarl* et *News*. Il contient divers types de collocations :

- des constructions à verbe support (*avoir+peur*, *faire+guerre*, *occasion+se présenter*, etc)

- des collocations Verbe+Nom où le nom est l'objet (*forcer+admiration, annoncer+nouvelle, régler+conflit*, etc.)

- des collocations Verbe+Nom où le nom est le sujet (*ton+monter*, *temps+écouler*, *colère+gronder*, etc.)

En outre, ce corpus de test français-anglais comporte en soi 80% de phrases (400 phrases) contenant des collocations, ainsi que 20% de phrases contenant des expressions idiomatiques (100 phrases).

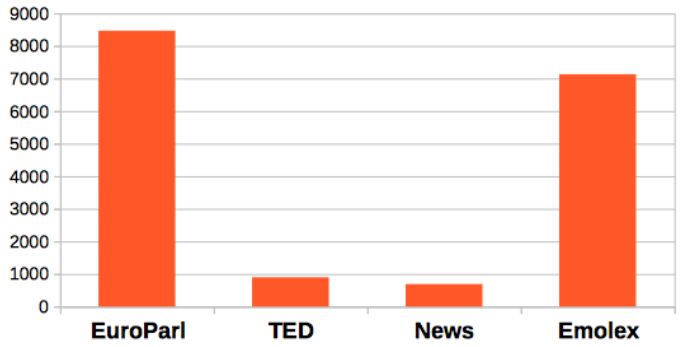


Figure 10 : Occurrences des collocations

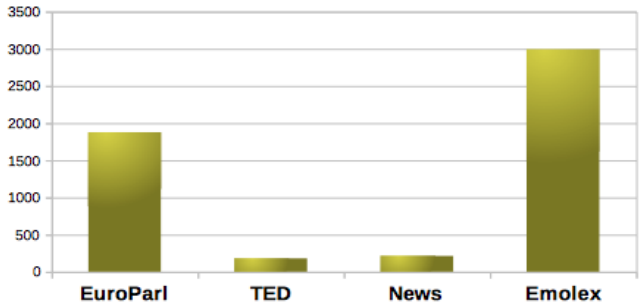


Figure 11 : Occurrences des idiomes

Origine	Corpus FR-EN	Collocations (phrase)	Expressions idiomatiques (phrases)
Emolex	253	201	52
Europarl	235	191	44
News	12	8	4
TOTAL	500	400	100

Tableau 10 : Distribution du corpus de test Français-Anglais

Dans cette partie, nous avons décrit en premier lieu la méthode que nous avons suivie pour la création des listes des EPLs, ensuite nous avons détaillé la méthodologie d'extraction et de

constitution des corpus de test Anglais-Français et Français-Anglais qui présentent des difficultés pour le système *Google Traduction*. Dans la prochaine partie, nous allons étudier une nouvelle méthode pour améliorer les performances d'un système de traduction.

5. Intégration des EPL dans un système de TA

Dans cette partie, nous cherchons à intégrer l'identification des expressions polylexicales en anglais (PV et *idioms*) dans un système de traduction de l'anglais vers le français, que nous allons adapter. Ensuite, nous allons proposer une méthode pour la prise en compte des PV dans le modèle de traduction. Finalement, nous allons utiliser une approche statique pour le traitement d'une liste des expressions idiomatiques.

5.1. Description de System de *Moses-LIG*

Le système *Moses-LIG* (Besacier et al., 2012) est un système à base de segments fondé sur la boite à outils Moses (Hieu et al., 2007), composé de trois modèles de traductions qui sont construits à partir des corpus : *News, Ted, Europarl, Eu-const, Dgt-tm, Un, Pct* et une extraction de 5M de phrases du corpus *Gigaword*, ainsi que d'un modèle de langue 5-gramme appris en utilisant la partie en français des mêmes corpus, tout en ajoutant le corpus *News-shuffle*. Le modèle de langue 5-gramme est appris séparément sur chaque corpus à l'aide de la boîte à outils SRILM Stolcke (2002) avec un modèle de repli Kneser-Ney modifié, puis les modèles sont interpolés en optimisant la perplexité sur le corpus *dev2010* de la campagne *IWSLT*. Le système Moses-LIG n'est pas optimal pour traduire des textes littéraires (Besacier et al., 2012).

5.2. Création du système de base

Pour réaliser nos expérimentations, nous avons créé un nouveau modèle de traduction pour le système *Moses-LIG*, mais en gardant le même modèle de langage. Ce nouveau modèle de traduction repose principalement sur les corpus *Europarl, TED* et *News*.

Afin de rendre les données d'apprentissage exploitables et cohérentes pour le système de traduction, nous les avons segmentées en mots. En effet MGIZA++ s'appuie sur une étape préalable de tokenization. Par exemple, la séquence *it's* doit être considérées comme deux unités lexicales « it » et « 's ». Cette opération a été effectuée à l'aide du script *tokenizer.perl* fournit par Moses.

Afin de comprendre l'impact de notre corpus spécifique (Tst-EPL) sur les systèmes de traduction, nous avons construit un corpus témoin de 500 phrases tirées aléatoirement qui contient la même répartition que le corpus de test spécifique.

Vu que, 45,6 % des phrases de notre corpus de test et du corpus témoin viennent d'*Europarl* et de *News* et que les données d'apprentissage se basent sur ces derniers, nous avons développé un script Perl pour éliminer ces phrases du corpus d'apprentissage.

État	**Initial**		**Après le filtrage**	
Langue	**Anglais**	**Français**	**Anglais**	**Français**
Mots	45 883 533	50 139 301	45 869 512	50 124 025
Phrases	2 006 954		2 006 374	

Tableau 11 : Taille des donnés d'apprentissage

5.3. Évaluation des système TA (Moses-LIG Vs Google-TR)

Nous nous sommes appuyé sur notre corpus de test Anglais-Français ainsi que sur le corpus témoin préparé, pour l'évaluation de la qualité de traduction des deux systèmes : *Moses-LIG* (en utilisant le nouveau modèle de traduction que nous avons créé) et *Google-TR*.

Le tableau ci-dessous, présente une évaluation détaillée de ces deux systèmes en terme de score Bleu :

Origine	**Corpus de 500 phrases**		***phrasal verbs* par origine**				***idioms* dans Tst-EPL**
	Témoin	**Tst-EPL**	**PV 366**	**Emolex**	**Europarl**	**News**	
System de base	24.87%	20.83%	22.72%	12.59%	29.33%	21.94%	15.21%
Google-TR	19,27%	19,81 %	18.67%	10,9%	21,82%	12,00%	19.75%

Tableau 12 : évaluation des qualités de traduction (Moses-LIG vs Google)

Ces résultats montrent que notre système de base est globalement meilleur que celui de Google-TR, notamment au niveau de la traduction des PV. Entre le corpus témoin et le corpus Tst-EPL, nous avons observons une dégradation au niveau des scores Bleu du système de base contrairement à *Google-TR* qui reste stable, avec un score assez bas. Cependant le système Google-TR traduit mieux les *idioms*, sans doute parce qu'il met en œuvre des dictionnaires spécifiques d'expressions idiomatiques.

D’autre part, le score Bleu du corpus témoin est plus élevé (24,87%), que celui du corpus de test (20,83%) : cela confirme que le corpus de test que nous avons créé pose vraiment des problèmes à notre système de base.

En outre, nous avons constaté que les deux systèmes de traduction traduisent mal les données littéraires d’Emolex.

Par suite, nous proposons dans les prochaines sections des méthodes pour améliorer la qualité de traduction de notre système de base au niveau des expressions idiomatiques et des PV.

5.4. Traitement des *phrasal verbs*

5.4.1. Détection des *phrasal verbs*

Afin d'identifier d’une manière automatique les PV que nous avons utilisés dans la constitution de notre corpus de test, nous avons annoté syntaxiquement tout d’abord le corpus de test en utilisant *XIP*. Ensuite, nous avons développé un outil en langage *Perl* pour reformer la sortie de *XIP*, et avoir par suite une nouvelle version *XML* pour notre corpus de test qui est une version adaptée pour l’entrée de *Moses.* Ainsi nous fournissons pour chaque forme toutes les informations linguistiques correspondantes.

A l’aide de cet outil, nous avons réussi à identifier tous les PV en nous basant sur les catégories des formes ainsi que sur les relations de dépendances *NUCL_PARTICLE_* et *MOD_POST* comme suit :

- La relation NUCL_PARTICLE (X, Y) signifie que X Y représente un PV, et elle est produite généralement pour les *phrasal verbs* de distance 0.

- La relation MOD_POST (X, Y) signifie que Y est situé à droite de X. En effet, pour que le PV soit valide, il ne faut pas avoir un verbe entre X et Y et que la distance entre X et Y doit être inférieur ou égale à 5.

Lors de l’identification, cet outil nous a permis d’ajouter une information de lien *EPL* (entre le verbe et la particule) dans la balise XML du verbe, ce qui rend la manipulation du fichier plus facile.

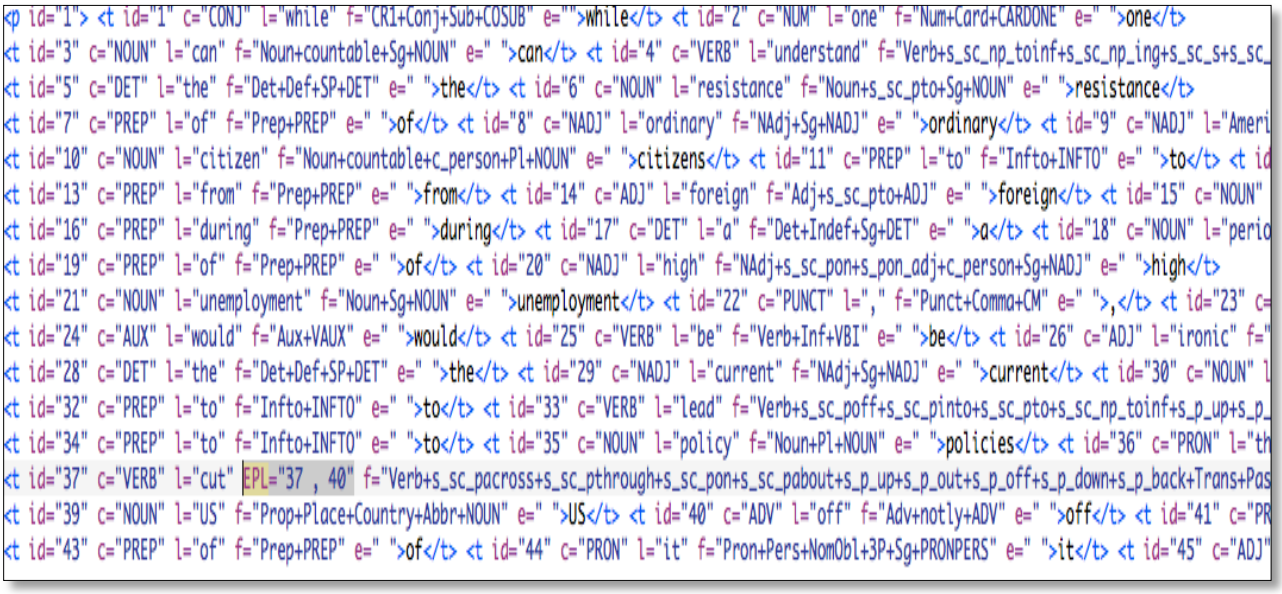


Figure 12 : Exemple de phrase du corpus de test XML

5.4.2. Méthode d'intégration

La source principale de connaissance du décodeur est sa table de traduction. En effet, le décodeur consulte cette table dans le but de décider comment traduire une phrase source en langue cible. Toutefois, à cause des erreurs d'alignement automatique de certains mots, des segments extraits peuvent ne pas se correspondre. Ainsi, pour améliorer l'alignement, nous avons considéré chaque PV détecté comme une seule unité lexicale, en vue de forcer la segmentation lors de l'alignement.

En utilisant la méthode de détection mentionnée précédemment, nous avons accroché le verbe à sa particule (sous la forme d'un composé *verbe-particule*). Cette approche a été appliquée sur les données d'apprentissage et sur les données de notre corpus de test.

Par exemple *keepyour voicedown, Hermione begged him.*

Keep-downyour voice, Hermione begged him.

En outre, nous avons ré-entrainé le système pour créer un nouveau modèle de traduction, tout en utilisant les nouvelles données d'apprentissage. Ce système est noté « *Exp-EPL* » dans la suite de ce mémoire.

Sentence pair (117035) source length 18 target length 19 alignment **score :1.39038e-35**
and I believe we can turn this whole story around to one of celebration and one of hope .

NULL ({ 16 }) et ({ 1 }) je ({ 2 }) crois ({ 3 }) qu' ({ }) on ({ 4 }) peut ({ 5 }) changer ({ 6 }) cette ({ 7 }) histoire ({ 8 9 10 }) en ({ 11 }) une ({ 12 }) histoire ({ }) de ({ 13 }) joie ({ 14 }) et ({ 15 }) d' ({ 17 }) espoir ({ 18 }) . ({ 19 })

Figure 13 : Exemple d'alignement pour le système de base

Lors de l'alignement, avant que les PV ne soit soudés, le système n'a pas réussi à traiter le « Phrasal Verb » :***'turn around'*** correctement. Il a considéré le verbe *'turn'* comme étant un seul segment aligné avec le verbe *'changer'*. En outre, la particule *'around'* fait partie d'un autre segment *('whole story around'*) et qui a été aligné avec le mot *'histoire'*. Ce type de problème d'alignement influence la qualité de traduction du système.

Sentence pair (117035) source length 18 target length 18 alignment **score :3.97955e-30**
and I believe we can turn-around this whole story to one of celebration and one of hope .

NULL ({ 15 }) et ({ 1 }) je ({ 2 }) crois ({ 3 }) qu' ({ }) on ({ 4 }) peut ({ 5 }) changer ({ 6 }) cette ({ 7 }) histoire ({ 8 9 }) en ({ 10 }) une ({ 11 }) histoire ({ }) de ({ 12 }) joie ({ 13 }) et ({ 14 }) d' ({ 16 }) espoir ({ 17 }) . ({ 18 })

Figure 14 : Exemple d'alignement pour le système Exp-EPL

D'après la figure ci-dessus, le « Phrasal Verb » :*'turn around'* a été analysé correctement comme un seul segment aligné avec le verbe *'changer'*.

Globalement, nous avons constaté une amélioration au niveau de score d'alignement qui varie de 1.39038e-35 à 3.97955e-30.

5.4.3. Evaluation

Afin d'évaluer notre méthode d'intégration, nous avons calculé les scores Bleu des sorties de deux systèmes : système de base ainsi que celui Exp-EPL. En outre, afin de maximiser le score Bleu, nous avons utilisé le programme *MERT* (*Minimum Error Rate Training*) (Och and Ney, 2003) qui permet d'ajuster les poids du modèle de distorsion, modèle de langage et le modèle de traduction .

Origine	Corpus de 500 phrases		Détail des *phrasal verbs* par origine				*Idioms* dans Tst-EPL
	Témoin	Tst-EPL	PV 366	Emolex	Europarl	News	
System de base	24.87%	20.83%	22.72%	12.59%	29.33%	21.94%	15.08%
Exp-EPL	23.81%	21.19%	23.18%	14.10%	29.38%	19.80%	14 .79%
System de base optimisé	26.46%	23.14%	23.47%	14.18%	29.20%	20.96%	22.03%
Exp-EPLoptimisé	26.28%	23.68%	24.16%	14.67%	29.96%	19.89%	21.83 %

Tableau 13 : Evaluation générale

Les résultats ci-dessus montrent une dégradation en terme du score Bleu entre les scores de l'évaluation de traduction du corpus de Tst-EPL et du corpus Témoin, ce qui confirme que notre corpus de test pose vraiment un défi aux systèmes de traduction, et que la fréquence des EPL influence sur la qualité de traduction.

Par ailleurs, nous remarquons une dégradation du score bleu (-1.06 point), pour les systèmes non optimisés en utilisant le corpus témoin, causée par la nouvelle segmentation effectuée au niveau de l'alignement. Ce qui n'est pas le cas pour les systèmes optimisés (une dégradation légère du score Bleu de -0.18). En outre, nous avons obtenu une amélioration observée (+0.54 point Bleu) pour la qualité de traduction du corpus de test en appliquant notre méthode d'intégration (système Exp-EPL). D'autre part, nous avons remarqué que le corpus littéraire Emolex est mal traduit par les systèmes ; malgré ça, nous avons constaté une augmentation observée de +1.51 point Bleu pour les systèmes non optimisés et de +0.49 point pour les systèmes optimisés.

Exemple de traduction :

Source	surely they must **call** the operation **off** now ?
Référence	maintenant , ils doivent sûrement **annuler** l' opération .
Hypothèse (système de base)	ils doivent ***appeler*** l' opération maintenant ?
Source V2	surely they must **call-off** the operation now ?
Hypothèse (Exp-EPL)	ils doivent ***annuler*** le fonctionnement maintenant ?

5.5. Traitement des expressions idiomatiques

Dans cette section, nous allons traiter les expressions de notre corpus spécifique avec la méthode de décodage forcé implantée par le décodeur de Moses. Tout d'abord, nous avons crée un dictionnaire des expressions idiomatiques dont le contenu est présenté sous forme de couple

(expression et sa traduction) comportant les *idioms* que nous avons utilisés pour construire notre corpus de test. Ensuite, nous avons développé un outil en Perl pour identifier les expressions idiomatiques dans notre corpus de test Anglais-Français afin de mettre la bonne traduction de chaque expression entre des balises XML compréhensibles par le décodeur.

Exemple : <idiom translation="***facile***">***piece of cake***</idiom>

Après avoir identifié les EPL, nous avons traduit les données post-éditées de notre corpus de test pour voir l'impact de notre méthode sur le score Bleu.

Exemple de traduction avant l'identification des « idioms » :

Source avant l'identification	I can not ***go dutch***.
Cible	je ne peux pas ***aller néerlandais*** .

Exemple de traduction après l'identification des « idioms » :

Source après l'identification	I can not <idiom translation="payer ma place">**go dutch**</idiom> .
Cible	je ne peux pas ***payer ma place*** .

	Score Bleu	
	idioms	**Tst-EPL**
Système de base	15.21%	20.83%
Méthode forcé	30.71%	24.77%

Tableau 14 : Evaluation de la méthode forcée

Le tableau ci-dessus, montre une augmentation de 15% du score Bleu en exploitant la méthode forcée proposée par le décodeur Moses. Ainsi, nous avons gagné 4 points pour la totalité de notre corpus de test.

Dans ce chapitre nous avons adapté tout d'abord, le système *Moses-LIG* en construisant un nouveau modèle de traduction avec de nouvelles données d'apprentissage. Ensuite, nous avons évalué la qualité de traduction de notre système de base qui est meilleur que celui de *Google-TR*. Par la suite, nous avons proposé une méthode d'intégration des *phrasal verbs* que nous avons exploité dans la construction du système Exp-EPL, ce qui a donné des résultats relativement significatifs. Finalement, nous avons utilisé la méthode de décodage forcé proposée par le décodeur Moses, pour traiter les expressions idiomatiques, produisant une nette amélioration des résultats.

6. Conclusion générale et perspectives

6.1. Conclusion

Dans ce travail de recherche, nous nous sommes intéressés à la prise en compte des expressions polylexicales dans les systèmes de traduction automatique statistique, qui constituent un vrai défi pour ces derniers.

Dans la première partie de ce manuscrit, nous avons donné une définition générale des expressions polylexicales, ainsi que de l'approche statistique de la traduction automatique. Ensuite, nous avons montré quelques problèmes de traduction de *Google-TR* au niveau des EPLs (*phrasal verbs*, expressions idiomatiques et collocation). D'autre part, nous avons cité quelques ouvrages réalisés en vue d'intégrer les EPLs dans un système de TA.

Dans la deuxième partie, nous avons préparé des corpus dans le but d'enrichir la ressource lexicale du projet Emolex en suivant une chaine de traitement réalisée dans le cadre de ce projet. Lors de l'annotation syntaxique des corpus, nous avons trouvé des difficultés à conserver la segmentation originale des corpus qui sont déjà alignés en phrase. Pour cela, nous avons ajouté des marques de segmentation qui sont récupérées ensuite en sortie de XIP. En outre, pour éviter le problème de surcharge du mémoire sur le serveur au cours de cette chaine de traitement, nous avons découpé les corpus (News, Europarl et TED) en corpus plus petits (1000 phrases par corpus).

Dans la troisième partie, nous avons proposé un processus semi-automatique d'extraction des EPLs dans le but de construire des corpus EN-FR et FR-EN spécifiques posant des problèmes de traduction au niveau des EPLs, à l'aide de *EmoConc*. De ce fait, nous avons préparé une liste de centaine d'EPLs afin de lancer des requêtes, et garder que les phrases qui posent des problèmes de traduction au niveau des EPLs. Ces corpus sont indispensables pour intégrer les EPLs dans les systèmes de traduction.

Dans la dernière partie de ce mémoire, nous nous sommes concentré à intégrer les expressions polylexicales en Anglais (*phrasal verbs* et *idioms)* dans un système de traduction que nous avons adapté. Grâce aux sorties de *XIP,* nous avons produit un nouveau format *XML* pour les corpus tout en gardant les informations linguistiques de chaque forme. Ce format *XML* (compatible avec l'entrée de Moses), nous a facilité la tâche d'identification des *phrasal verbs* que nous avons utilisés pour construire notre corpus de test. Nous avons proposé une méthode pour la prise en compte de ces *phrasal verbs* dans le modèle de traduction afin d'adapter le système *Moses-LIG*. Ainsi, nous avons obtenu quelques améliorations de traduction observées (gain de

0,54 point) en termes de score Bleu. Enfin, nous avons utilisé la méthode de décodage forcé implantée par le décodeur Moses pour le traitement d'une liste des expressions idiomatiques bilingue identifiées. Grâce à cette méthode nous avons réussi à améliorer la qualité de traduction (+4 points du score Bleu) de notre corpus spécifique EN-FR. Toutefois, cette méthode montre ses limites, si une expression a plusieurs sens (expression ambigüe) ou n'apparaît pas dans le dictionnaire (problème d'identification).

6.2. Perspectives

Nous avons commencé un premier travail pour l'évaluation humaine en produisant une interface graphique (annexe 5) pour faciliter la tache des annotateurs. D'autre part, il est envisageable d'utiliser d'autres méthodes pour l'intégration des *phrasal verbs* et des expressions idiomatiques en anglais dans le système *MOSES-LIG*.

En outre, nous pourrions améliorer le système *MOSES-LIG* français-anglais en intégrant les collocations et les expressions idiomatiques en français. Il faut développer des outils pour identifier les EPLs dans les deux langues (français et anglais) afin de produire un dictionnaire d'EPLs bilingue.

Par ailleurs, il faudrait continuer à enrichir les corpus spécifiques français-anglais et anglais-français pour avoir une plus grande quantité de données, nécessaire pour l'évaluation de la qualité de traduction. Nous pourrions aussi appliquer notre méthode d'intégration des PV sur les collocations, en soudant les verbes aux noms ou les noms aux verbes (pour les collocations de type Nom+Verbe ou Verbe+Nom) . Les outils d'identification et les corpus spécifiques sont indispensables pour traiter ces expressions dans les systèmes de traduction automatique statistique.

ANNEXES

Annexe 1 : Les scripts réalisés

Nom du script	Rôle
clean-corpus.pl :	Prétraitement du corpus, supprimer les lignes vide, convertir les entités HTML,etc .
extractCollocations.pl	Interroger la nouvelle ressource lexicale pour extraire touts les contextes d'une liste de collocations (en Français) préparée à l'avance avec des distances de 0 à 3.
extractPhrasalverbs.pl	Interroger la nouvelle ressource lexicale pour extraire touts les contextes d'une liste de « Phrasal Verbs » avec des distances de 0 à 5.
extractIdoms.pl	Extraire les contextes d'une liste des idiomes.
extractCsvFR.pl	Extraire les fréquences des EPLfrançais sous formats Csv .
extractCsvEN.pl	Extraire les fréquences des EPL anglais sous formats Csv .
extractTags.pl	Déterminer le corpus source d'une phrase quelconque .
reformerOutXIP.pl	Reformer la sortie de l'outil XIP pour mettre chaque phrase dans une ligne .
extract-corpus-Dist.pl	Extraire grand corpus
XML2HTML.pl	Créer une interface graphique pour l'évaluation humaine de la traduction.

Annexe 2 : Extrait du Corpus de test EN-FR:

this Parliament firmly believes that intellectual property rights must be protected , but not by giving private companies sweeping rights to monitor indiscriminately every citizen 's activities on the Internet something that we refuse to allow even our police to do when fighting terrorism and certainly not through the disproportionate penalty of **cutting** whole households **off** from the Internet .

by playing with international borders Yugoslavia 's since 1991 , and Macedonia 's and Albania 's in the future by disregarding the lessons of the past , by pretending to ignore Russia 's role in the region , by seeking a military knockout instead of a political solution , and by trying , in particular , to **cut** the Serbian people **off** from Europe , which they have done so much for , we will undoubtedly provoke the very conflagration we claim we want to avoid .

we shall now continue with religious freedom in Vietnam , and the presidency of this sitting would like to apologise wholeheartedly for having to **cut** Mr Belder **off** .

everyone will make their own **mind up** about the political consistency of such a reversal .

you really could not **make** some of this stuff **up** .

Annexe 3 : Extrait du Corpus de test FR-EN

Lorsqu'il eut terminé, la **nuit** était presque **tombée**.

Hier, le commissaire Rehn a **annoncé** diverses **nouvelles** importantes venant de la Commission.

Il est de notre responsabilité d' éviter de **déclencher** un nouveau **conflit** gazier.

Après cela, nous avons **ravalé** notre **colère**.

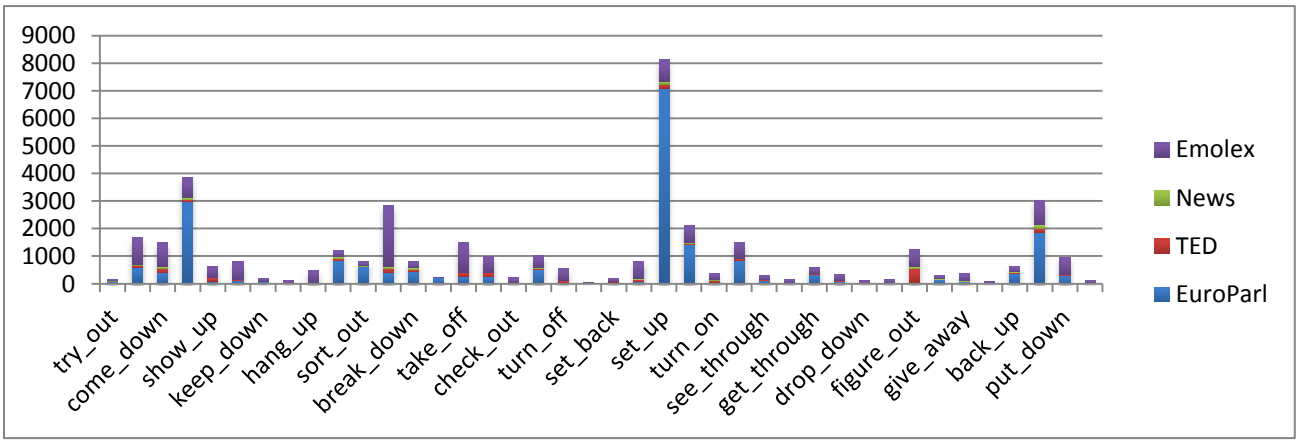
Ces derniers temps, les dragons ont toutefois **essuyé des échecs** cuisants.

De nombreux citoyens n' ont pas été expulsés qui, s' ils l' avaient été, n' auraient peut-être pas **commis de crimes** en Italie.

Son premier sermon avait d'ailleurs **annoncé la couleur**.

Millenium avait l'intention de passer l'histoire dans le numéro de juin mais Mikael Blomkvist a **arrêté les frais**.

Annexe 4 : Occurrence des « Phrasal Verbs »



Annexe 5:Extrait de l'interface HTML de l'évaluation humaine

Pour des travaux ultérieurs, et dans le but de faciliter la tache d'évaluation manuel de la traduction, nous avons pour produit une interface graphique HTML. Nous avons développé le script XML2HTML.pl qui prend en entré un fichier XML (en identifiant les EPL) et une sortie segmenté par le système de traduction. En suite, nous avons utilisé les expressions régulières et des fonctions en JavaScript pour colorer les EPL Anglais en jaune et le segment qui contient sa traduction en rouge.

Exemple de segmentation: M. Schulz **|0-1|** , prenez votre **|2-4|** propre signature **|5-6|** ! **|7-8|**

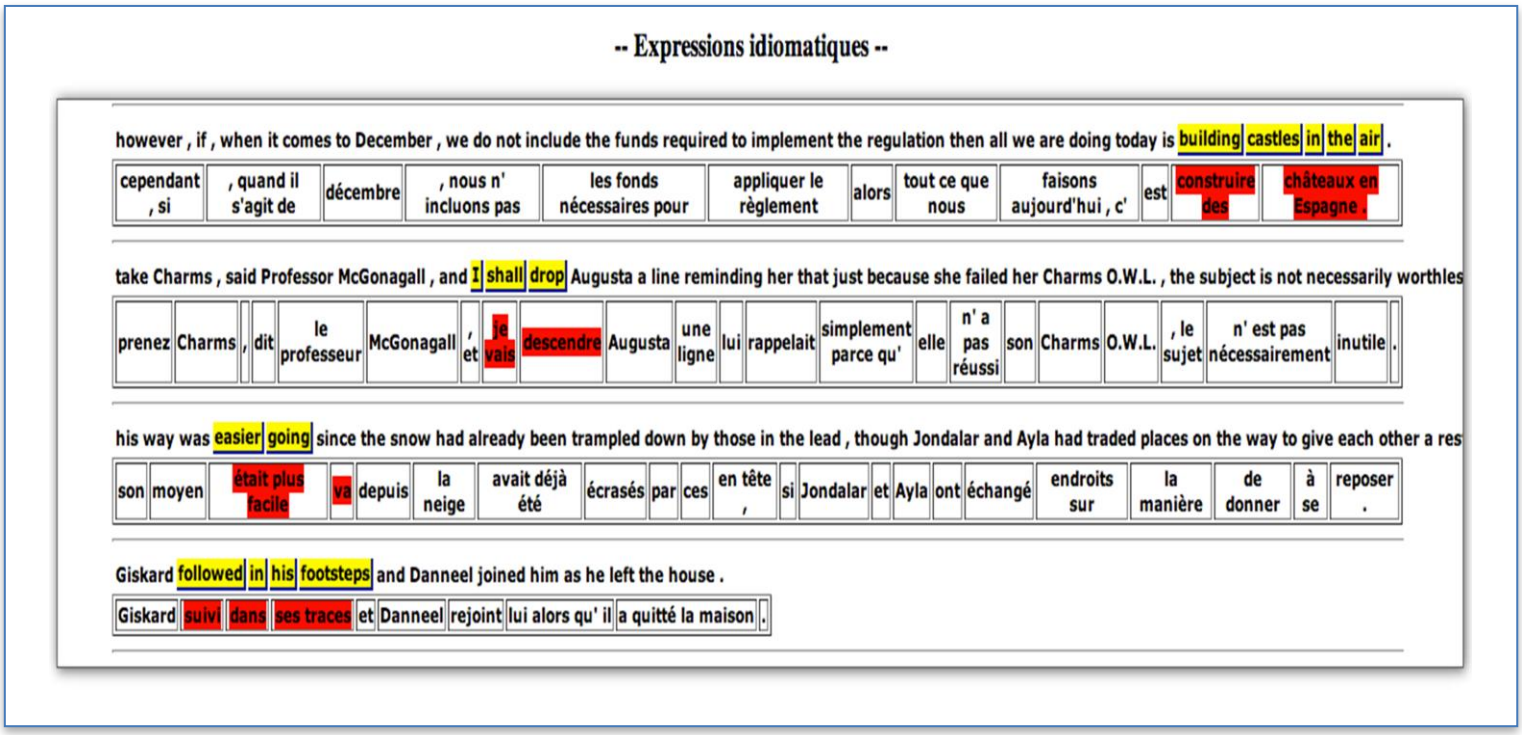
Exemple de paire de phrases :

and you thought my job was a	piece	of	cake	!
et vous vous êtes dit	mon travail était	une sinécure	!	

Pour afficher le segment rouge en langue source, il se suffit de placer le curseur de la souris sur ce dernier.

Exemple :

and you thought my job was a	piece	of	cake	!
et vous vous êtes dit	mon travail était	a-piece-of-cake	!	



Bibliographie

- Anastasiou et al. (2009),"***2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications***", Aout 2009, Suntec, Singapore.

- Besacier (cours), "***Introduction à la traduction automatique statistique***", Université Joseph Fourier, Grenoble France.*http://www.clips.imag.fr/geod/User/laurent.besacier/traduction.pdf*

- Bouamor (2013), "***Acquisition de lexique bilingue d'expressions polylexicales: une application à la traduction automatique statistique***", TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne, Université Paris sud, Paris France.

- Courtois (1994-1995), "***Buts et méthodes de l'élaboration des dictionnaires électroniques du LADL***", dans Ca-hiers CIEL 1994-1995, Université Paris 7, Paris France.

- Kobzar (2013), "***How Hard Is It to Translate Multi-Word Expressions?*** ", mémoire de recherche,juin 2013, Université Joseph Fourier France.

- Koehn (2005), ***"Europarl: a parallel corpus for statistical machine translation, Machine Translation",***sommet TA,2005,Université d'Edimbourg,Scotland.

- Koehn (2014), " ***Statistical Machine Translation, System User Manual and Code Guide***", modification la plus récente aout 2014, Université de Edinburgh Royaume-Uni..

- Kraif et Diwersy (2012), ***"Le Lexicoscope: un outil pour l'étude de profls combinatoires et l'extraction de constructions lexico-syntaxiques***",p 399—406Actes de la conférence TALN juin2012, Grenoble France.

- Kuhn et al. (1990), ***"A cache-based natural language model for speech recognition"***. TransactionIEEEsur l'analyse des formes et de l'intelligence artificielle. Vol. 12,No.6,juin 1990, pages 570–582.

- Lau et al. (1993) , ***"Trigger-based language models: a maximum entropy approach",*** *acte de la conférence 27-30 avril 1993, page 45-48, Volume 2 , Minneapolis, MN, USA.*

- Tutin (2005),"***Le dictionnaire de collocations est-il indispensable ?***", *Revue Française de Linguistique Appliquée*. Dictionnaires : nouvelles approches, nouveaux modèles (Th. Fontennelle ed.). Volume X-2. Décembre 2005 31-48,Université Stendhal Grenoble-France.

- Tutin (2010), "***Sens et combinatoire lexicale : de la langue au discours***", volume 1, Dossier en vue de l'habilitation à diriger des recherches, janvier 2010 Université Stendhal, Grenoble France.

- Potet (2013), "***Vers l'intégration de post-éditions d'utilisateurs pour améliorer les systèmes de traduction automa- tiques probabilistes***", Nouvelle thèse, avril 2013, Université de Grenoble France.

- Ramsich et al. (2012), "***A generic and open framework for multiword expressions treatment : from acquisition to applications***", Nouvelle thèse, sep 2012, Université de Grenoble (France) et Université fédérale de Rio Grande do Sul (Brazil).