La traduction automatique statistique factorisée : une application à la paire de langues français - roumain

Elena-Mirabela Navlea Laporte

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE 520 « Humanités »

U.R. 1339 LiLPa (*Linguistique, Langues, Parole*)

THÈSE présentée par :

Elena-Mirabela NAVLEA

soutenue le : **13 juin 2014**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : Sciences du Langage / Linguistique et Informatique

La traduction automatique statistique factorisée :

une application à la paire de langues français - roumain

THÈSE co-dirigée par :

M. GRASS Thierry	Professeur des Universités, Université de Strasbourg
Mme TODIRAȘCU Amalia	Maître de conférences HDR, Université de Strasbourg

RAPPORTEURS :

M. HEID Ulrich	Professeur Dr., Universität Hildesheim
M. TUFIȘ Dan	Directeur de recherche, Académie Roumaine de Bucarest

---

AUTRES MEMBRES DU JURY :

M. GROSSMANN Francis	Professeur des Universités, Université Stendhal - Grenoble 3

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE 520 « Humanités »

U.R. 1339 LiLPa (*Linguistique, Langues, Parole*)

THÈSE présentée par :

Elena-Mirabela NAVLEA

soutenue le : **13 juin 2014**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : Sciences du Langage / Linguistique et Informatique

La traduction automatique statistique factorisée :

une application à la paire de langues français - roumain

MEMBRES DU JURY :

Président	**M. Francis GROSSMANN**, Professeur, Université Stendhal - Grenoble 3 (France)
Directeur	**M. Thierry GRASS**, Professeur, Université de Strasbourg (France)
Co-directrice	**Mme Amalia TODIRAȘCU**, Maître de conférences, Université de Strasbourg (France)
Rapporteur	**M. Ulrich HEID**, Professeur Dr., Universität Hildesheim (Allemagne)
Rapporteur	**M. Dan TUFIȘ**, Directeur de recherche, Académie Roumaine de Bucarest (Roumanie)
Examinateur	**M. Francis GROSSMANN**, Professeur, Université Stendhal - Grenoble 3 (France)

À mes chers,

ma mère,

mon père,

Corneliu,

Frédéric.

Remerciements

Tout d'abord, je remercie mon directeur de thèse, M. Thierry Grass, et ma co-directrice, Mme Amalia Todirașcu, pour l'intérêt qu'ils ont porté à cette thèse ainsi que pour leur confiance et leurs encouragements.

Ensuite, j'exprime ma gratitude envers M. Francis Grossmann, M. Ulrich Heid et M. Dan Tufiș d'avoir accepté de faire partie du jury de cette thèse.

J'adresse également mes remerciements aux chercheurs de l'*Institut de Recherches en Intelligence Artificielle « Mihai Drăgănescu »* de l'Académie Roumaine de Bucarest, M. Dan Tufiș, M. Alexandru Ceaușu, M. Radu Ion, M. Dan Ștefănescu et M. Ștefan Dumitrescu pour les échanges fructueux sur la traduction automatique et le partage généreux des ressources et des outils qui m'ont été très utiles dans le cadre de ce travail.

Un grand merci à toute l'équipe *LiLPa* pour son accueil chaleureux et, tout particulièrement, à M. Rudolph Sock, le directeur de *LiLPa*, et à Mme Maryvonne Boisseau, la directrice de l'équipe *Fonctionnements discursifs et Traduction.*

Je remercie vivement Mme Catherine Schnedecker, l'ancienne directrice de *LiLPa*, et Mme Marie-Carmen Ramirez, la secrétaire de l'École Doctorale des Humanités, pour la confiance qui m'a été accordée pendant les cours d'informatique et de spécialité que j'ai enseignés à l'École Doctorale. Outre l'expérience enrichissante d'enseignement, ces cours m'ont aussi permis de financer partiellement ce travail de thèse, hormis les nombreuses autres vacations et une année d'*ATER* à temps complet au Département d'Informatique de l'*UFR LSHA/UDS*.

Je tiens à remercier également les membres du Conseil de *LiLPa* de m'avoir accordé le financement d'un stage de 6 mois dans le cadre de cette thèse ainsi que le financement de mes conférences nationales et internationales.

Je remercie aussi les deux masterands stagiaires encadrés Goncharova Yuliya (Université de Strasbourg) et Havași Sebastian-Flaviu (Université Babeș-Bolyai de Cluj-Napoca, Roumanie) qui ont participé à des tâches précises de ce travail, coûteuses en temps et en ressources humaines.

Un merci chaleureux à mes collègues docteurs et doctorants Constanze Armbrecht, Martha Makassikis, Mihaela Ruşitoru, Julien Rentz et Olga Turcan pour leurs encouragements précieux et leurs bons conseils.

Enfin, je remercie ma chère famille, ma mère, mon père, mon frère et tout spécialement mon mari, pour leur soutien, leur patience, leur ouverture et leur indicible compréhension.

Table des matières

1. Introduction

Cette thèse porte sur la traduction automatique statistique factorisée entre deux langues latines, riches du point de vue morphologique : le français et le roumain.

La traduction automatique (*TA*) désigne la traduction effectuée entièrement de manière automatique, d'un texte d'une langue de départ (langue source) vers une langue d'arrivée (langue cible). Lorsque le traducteur humain intervient dans le processus de traduction, il s'agit alors de la traduction assistée par ordinateur (*TAO*).

Ce domaine de recherche dont l'intérêt remonte vers le milieu du XXème siècle se trouve à l'origine du développement du traitement automatique des langues (*TAL*) le long du temps, et représente aujourd'hui un point d'intersection de nombreuses applications de *TAL*, comme l'analyse linguistique automatique de divers niveaux, l'extraction de terminologie, la constitution de ressources langagières pour différentes paires de langues (corpus, dictionnaires, grammaires), etc. Si, à ses débuts, la traduction automatique avait des buts militaires, elle est devenue actuellement une recherche de pointe de la linguistique computationnelle, étant utile pour plusieurs domaines d'applications qui seront exemplifiés ci-dessous.

L'essor économique, d'une part, mais aussi la mise à disposition et le stockage de textes multilingues en quantité considérable, par le développement du Web et d'ordinateurs puissants, d'autre part, ont augmenté la demande en traductions automatiques pour différentes paires de langues, au sein des entreprises et des administrations, à l'échelle mondiale. Mais ce besoin est ressenti également par les utilisateurs individuels désirant apprendre des langues étrangères, par exemple, et avoir ainsi accès à la culture d'un pays en particulier. De ce fait, de nombreuses recherches ont été menées pour la mise en place des logiciels de traduction automatique nécessaires pour des domaines comme la veille économique, la recherche d'information multilingue, la rédaction technique dans plusieurs langues, la génération automatique de contenu Web, etc.

Le développement rapide d'applications multilingues en ligne nécessitant des techniques de traduction automatique suscite l'intérêt d'adapter les systèmes de traduction automatique pour

différentes paires de langues, surtout que la plupart des logiciels actuels considèrent l'anglais comme langue source ou cible. L'anglais possède une morphologie simple par rapport à d'autres langues ayant une morphologie flexionnelle complexe, comme les langues latines, slaves et balkaniques, par exemple. Pour ces langues, la constitution de ressources linguistiques (dictionnaires, grammaires, bases de données terminologiques) nécessite du temps et des efforts humains et matériels importants. Si l'anglais dominait initialement le Web, d'autres langues s'y font aujourd'hui de plus en plus présentes et donc la nécessité d'adapter les logiciels de traduction automatique pour ces langues devient imminente.

Dans le domaine de la traduction automatique se sont développées deux grandes approches : l'experte et l'empirique. L'approche experte (le transfert, l'interlangue) utilise des dictionnaires et des règles linguistiques complexes (lexicales, d'analyse syntaxique, etc.), coûteuses en temps et en ressources humaines et matérielles. Notons que la traduction directe, faisant aussi partie de cette approche et caractérisant plutôt les débuts de la traduction automatique, utilise des règles linguistiques basiques. À la différence de l'approche experte, l'empirique (les méthodes à base d'exemples et statistiques) extrait ses connaissances à partir de grosses collections de textes (dénommées corpus parallèles, c'est-à-dire des textes bilingues ou multilingues qui sont des traductions humaines réciproques), par le biais des calculs statistiques. L'avantage de ces systèmes est qu'ils peuvent facilement s'améliorer de manière automatique, dès que de nouveaux corpus parallèles sont disponibles, en s'avérant ainsi moins coûteux en temps et en personnels par rapport aux méthodes à base de règles.

Les premiers modèles de traduction automatique statistique ayant mis les bases des développements ultérieurs dans le domaine ont été décrits dans les travaux de référence de Brown *et al*. (1990) et Brown *et al.* (1993). Ces modèles fonctionnent par apprentissage automatique des équivalents de traduction à partir de corpus parallèles alignés au niveau lexical (autrement dit les mots sont mis en correspondance à l'intérieur des phrases parallèles). Ceux-ci utilisent ainsi dans le processus de traduction le seul mot, montrant rapidement les limites d'une traduction de type mot-à-mot (par exemple des séquences plus ou moins compositionnelles non traduites, l'ordre erroné des mots cibles, etc.).

Afin de diminuer le taux de ces erreurs de traduction, se sont développées les approches à base de séquences (Och *et al.*, 1999 ; Zens *et al.*, 2002 ; Zhang *et al.*, 2003 ; Koehn *et al.*, 2003). Ces systèmes n'utilisent plus le mot mais plutôt la séquence de mots dans le processus

de traduction, qui est censée, par exemple, mieux gérer le ré-ordonnancement local au niveau de la phrase cible.

Si les systèmes linguistiques *(*p. ex. *Systran*[1]) donnent des résultats de traduction performants de par la modélisation de divers phénomènes linguistiques caractérisant les langues naturelles, les méthodes statistiques fournissent des résultats comparables en s'appuyant sur des techniques factorisées (p. ex. le projet *EuroMatrix*[2]). Celles-ci exploitent des corpus parallèles lemmatisés, étiquetés et alignés au niveau propositionnel et lexical et s'avèrent aussi moins coûteuses en temps et en ressources humaines.

Les systèmes factorisés (Koehn et Hoang, 2007 ; Avramidis et Koehn, 2008 ; Tufiș *et al.*, 2008b ; Ceaușu et Tufiș, 2011) prennent en compte divers facteurs linguistiques associés aux unités lexicales (lemmes, propriétés morphosyntaxiques, informations syntaxiques, etc.), afin d'améliorer les résultats des systèmes statistiques à base de séquences standard. Les méthodes factorisées impliquent ainsi plusieurs étapes de traduction de facteurs linguistiques et de génération de formes fléchies cibles, en fonction des facteurs traduis au préalable. En effet, à partir des traductions des lemmes et des propriétés morphosyntaxiques, par exemple, le système peut générer en sortie la variante morphologique correcte d'un mot cible. Par conséquent, le système peut traduire des mots supplémentaires inconnus ou moins représentatifs dans les corpus d'apprentissage (Koehn et Hoang, 2007). Ces systèmes sont donc utiles pour les langues moins dotées en ressources langagières et riches du point de vue morphologique (présentant des paradigmes flexionnelles importantes), comme nous le verrons aussi dans le chapitre 2 de ce travail.

Certaines approches (Koehn et Hoang, 2007 ; Tufiș *et al.*, 2008b ; Ceaușu et Tufiș, 2011) montrent que l'intégration d'informations morphologiques riches améliore nettement les résultats de la traduction, par rapport à un système de base purement statistique. D'autres méthodes (Birch *et al*., 2007 ; Avramidis et Koehn, 2008) obtiennent des résultats prometteurs en exploitant l'information syntaxique dans les systèmes factorisés. Les résultats de ces systèmes dépendent néanmoins des paires de langues considérées, du volume des données d'apprentissage mais aussi du sens de la traduction.

---

[1] *Systran* (http://www.systransoft.com/) était un système purement linguistique basé sur la méthode du transfert jusqu'en 2009. Depuis, celui-ci est devenu hybride par l'intégration de techniques statistiques.

[2] http://www.euromatrix.net/

Notre étude est motivée, d'une part, par le manque de systèmes de traduction automatique adaptés pour la paire de langues étudiées et, d'autre part, par le nombre important d'erreurs qui sont générées néanmoins par les systèmes de traduction automatique actuels.

Les ressources linguistiques monolingues ou multilingues et les outils disponibles pour le roumain sont peu nombreux (une description détaillée des outils développés figure dans (Tufiș *et al.*, 2008a)). En outre, le roumain est de plus en plus présent sur le Web, étant donné que le marché de l'Internet en Roumanie augmente progressivement (Trandabăț *et al.*, 2012). Par exemple, en 2012 le domaine *.ro* enregistrait 0,4% des pages Web existantes, par comparaison avec le domaine *.eu* (Trandabăț *et al.*, 2012). De plus, le roumain est l'une des langues officielles de l'Union Européenne depuis 2007.

Ainsi, le besoin en traductions automatiques de et vers le roumain s'avère aujourd'hui important. Notons que la plupart des systèmes incluant le roumain disposent des ressources pour la paire de langues anglais - roumain (Marcu et Munteanu, 2005 ; Tufiș *et al.*, 2008b ; Irimia, 2008 ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011 ; Dumitrescu *et al.*, 2012 ; Tufiș et Dumitrescu, 2012 ; Dumitrescu *et al.*, 2013 ; Tufiș *et al.*, 2013ab ; Boroș *et al.*, 2013). D'autres systèmes se concentrent sur la paire de langues allemand - roumain (Gavrilă, 2009, 2012 ; Vertan et Gavrilă, 2010). Ces systèmes ont adopté soit des techniques statistiques pures (Marcu et Munteanu, 2005) ou factorisées (Tufiș *et al.*, 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011, etc.), soit à base d'exemples (Irimia, 2008 ; Gavrilă, 2012). Certains logiciels libres purement statistiques comme *Google Translate*[3] ou basés sur l'approche directe de la traduction comme *Intertran*[4] incorporent aussi la paire de langues français - roumain. À notre connaissance, aucun système de traduction automatique n'a pas été conçu spécialement pour le français et le roumain.

Malgré les avancées actuelles considérables dans la matière, les systèmes de traduction automatique commettent encore un nombre important d'erreurs de traduction dues, entre autres, au manque de ressources linguistiques performantes pour différentes paires de langues. La constitution de ces ressources est une tâche difficile conditionnée par de nombreux phénomènes linguistiques, spécifiques aux langues naturelles (ambigüités, manque d'équivalence de traduction d'une langue à l'autre, paraphrase, etc.).

[3] http://translate.google.fr/
[4] www.tranexp.com:2000/Translate/result.shtml

Grass (2010)[5] relève treize catégories d'erreurs fréquentes fournies par les outils de traduction automatique basés sur des méthodes linguistiques comme le transfert (*Systran V6 Premium Translator*) et statistiques pures (*Google Translate*). Ces catégories d'erreurs seront énumérées et illustrées ici par des exemples concrets français - anglais ou français - allemand (Grass, 2010). Précisons que cette étude comparative a été réalisée en utilisant la version commerciale purement linguistique de *Systran* (*Systran V6 Premium Translator)* et *Google Translate* (2008). Notons aussi que cette classification des erreurs de traduction automatique est relevante pour notre approche, car elle a été effectuée en 2008, l'année même où ce projet de thèse a débuté, fournissant donc, par des exemples concrets de traduction, un aperçu des problèmes rencontrés à cette époque par des outils performants comme *Systran* et *Google Translate*.

Les catégories d'erreurs relevées sont les suivantes (Grass, 2010) :

1) la polysémie et l'homonymie ;

    a. Exemple d'erreur concernant la polysémie (l'adjectif polysémique français *« léger »* traduit comme *« light »* en anglais) : *Le directeur est **léger** dans son travail*. vs. *The director is **light** in his work.* [*Google Translate*] ;

    b. Exemple d'erreur liée à l'homonymie (l'homonyme *avocat* du français - homme de loi et fruit - traduit seulement par *« lawyer »* en anglais par les deux systèmes) : *Julia aime son **avocat**. / Julia aime les **avocats**.* vs. *Julia loves its **lawyer**. / Julia loves **lawyers**.*

2) l'ambiguïté syntaxique (p. ex. *to fly gliders* et **to clean fluids*) ; Ce problème est dû au fait que certaines structures syntaxiques sont ambigües sans connaissance du monde.

    a. ***Cleaning fluids** can be dangerous (« cleaning fluids » / *to clean fluids)* vs. **__To clean fluids__ can be dangerous*. (Dans ce cas, *to clean* n'est pas compatible avec *fluids*) ;

    b. ***Flying gliders** can be dangerous. (« flying gliders » / « to fly gliders »)* vs. *To fly gliders can be dangerous.* (*to fly* requiert comme objet un « objet volant »).

[5] http://www.cahiersdugepe.fr/index1367.php

3) l'ambiguïté référentielle (p. ex. *Paul a heurté le vase du pied et **l'**a cassé.*) ; Concernant le pronom *le* et son référent, il peut s'agir du vase ou du pied.

   a. *Paul ran up against **the vase of the foot** and broke **it**.* [*Systran*] ;

   b. *Paul struck **the foot of the vase** and broke.* [*Google Translate*].

4) les expressions floues (*Fuzzy hedges*) ; Ces expressions sont « des mots ou des groupes de mots au caractère idiomatique marqué, donc très dépendants de l'organisation sémantique de la langue source, qui sont difficiles à traduire et dont le rôle est d'exprimer une approximation » (Grass, 2010)[6] - « *words whose job it is to make things more or less fuzzy* » (Lakoff, 1972 : 183, *cit. in* Grass (2010)). Dans ce cas, pour l'expression française *« en un sens »* les traductions proposées en anglais sont *« in a direction »* [*Systran*] et *« in a sense »* [*Google Translate*].

5) les idiotismes et les métaphores ; Par exemple, l'expression du français *à couteaux tirés* (*at daggers drawn*) est traduite en anglais **«** *on with drawn knives »* [*Systran*] et « *at loggerheads »* [*Google Translate*].

6) la néologie ; Les logiciels ne sont pas toujours mis à jour suite à l'évolution de la langue. Par conséquent, les néologismes français « *internautes »* et *« une Web star »* sont traduits en anglais « *Net surfers »* et ***«*** *A Web star »* par *Systran*, mais « *to the Internet »* et « *Web is a star »* par *Google Translate*.

7) les noms propres ; Ceux-ci représentent l'un des problèmes les plus ardus pour un logiciel de traduction automatique, car ils sont difficiles à recenser de par leur nombre très élevé. De plus, leur orthographe est souvent différente d'une langue à l'autre (par exemple, *Vladimir Poutine* est traduit en anglais *Vladimir Poutine (Putin)* [*Systran*] et *Vladimir Putin* [*Google Translate*]).

8) les mots d'origine étrangère et les emprunts ; À titre d'illustration, l'expression idiomatique anglaise *« nothing in the pipeline »* (l'expression équivalente en français « *rien dans les tuyaux »* est empruntée de l'anglais) est traduite en français « *rien dans la canalisation »* [*Systran*] et « *rien en dehors de la canalisation »* [*Google Translate*].

[6] http://www.cahiersdugepe.fr/index1367.php

9) les séparateurs (les signes de ponctuation, certaines abréviations) ; Dans ce cas, « *ab dem 16.* » de l'allemand où *16.* est un numéral ordinal (« à partir de la 16ème année » en français) est traduit « *16 à partir de cela* » par *Systran*, mais « *à partir du 16 Ans* » par *Google Translate*.

10) les sigles et les acronymes ; Même si les deux systèmes résolvent généralement ce type de problème, la traduction fournie par le système par transfert n'est pas toujours correcte. Par exemple, la dénomination d'une entreprise « *Total SA* » est traduite en allemand par « *Bezeichnung Gesamtzahl AG* ».

11) les synonymes ; La synonymie est un problème crucial pour un logiciel de traduction automatique dû à la richesse lexicale d'une langue. En effet, à partir de plusieurs variantes de traduction d'un mot dans ses différents contextes, le logiciel ne choisit pas toujours le synonyme approprié. Par exemple, dans la séquence anglaise « *after unfurling a Tibetan flag and banner* », la traduction proposée par les deux systèmes en français concernant le mot « *banner* » (« *banderole* ») est « *bannière* ».

12) la transposition (le changement de catégorie grammaticale) ; Il s'agit d'un procédé de traduction fréquemment utilisée en traduction humaine, qui pose problème aux logiciels. À titre d'illustration, la transposition de « *house for sale* » (anglais) en « *maison à vendre* » (français) est rendue comme « *maisons de course vers le bas à vendre* » par *Systran*, mais comme « *maisons en vente* » par *Google Translate*.

13) l'orthographe ; Dans les textes en provenance du Web surtout, les mots présentent souvent des fautes d'orthographe (p. ex. « *traductteur* », « *cpable* » vs. « *traducteur* », « *capable* ») et les logiciels les traitent ainsi comme des mots inconnus et ne les traduisent pas.

En outre, Ramisch *et al.* (2013) montrent que même si les systèmes de traduction automatique statistiques sont devenus actuellement performants, les constructions flexibles comme les verbes à particule de l'anglais ne sont pas traitées de façon appropriée. Dans leurs expériences de traduction de l'anglais vers le français, ils prouvent que plus de la moitié des traductions ont des problèmes d'adéquation et/ou de fluence. Par exemple, pour la construction anglaise *think through* (*repenser, réfléchir*), le système propose la traduction « *penser à travers* ».

Au vu de ces problèmes de la traduction automatique, force est de constater que les logiciels actuels, qu'ils soient par transfert ou statistiques, rencontrent encore de nombreuses difficultés. Toutefois, la traduction statistique semble parfois mieux fonctionner que celle par transfert du fait qu'elle utilise des ensembles très larges de traductions déjà effectuées par des humains (Grass, 2010), appelés mémoires de traduction ou corpus parallèles. Si les problèmes comme les ambigüités ou la transposition sont dus au facteur de la connaissance du monde, la majorité des problèmes apparaît à cause du codage insuffisant des dictionnaires. En effet, ceux-ci ne prennent pas en compte tous les progrès de la linguistique en ce qui concerne les systèmes par transfert (Grass, 2010). De ce fait, les méthodes statistiques qui s'améliorent facilement dès que de nouvelles données d'apprentissage deviennent disponibles, semblent plus prometteurs.

Cependant, pour une traduction de qualité élevée, les techniques de traduction automatique ne sont pas suffisantes actuellement et une post-édition des résultats (autrement dit la correction par les traducteurs humains des ébauches obtenues de façon automatique) reste encore nécessaire. En outre, si le but de la traduction n'est pas sa qualité élevée, mais plutôt la compréhension du message ou l'obtention d'une ébauche dans la langue cible, la traduction automatique s'avère très efficace aujourd'hui pour des traductions rapides, en temps réel, sur le Web ou au sein des entreprises et des administrations, à travers le monde.

Dans notre projet de thèse, nous avons adopté l'approche statistique factorisée de la traduction automatique car, comme il a déjà été précisé auparavant, il s'agit de l'approche appropriée pour des paires de langues moins dotées en ressources langagières et riches du point de vue morphologique, comme le sont le français et le roumain.

Notre étude s'inscrit à la lignée du projet *SEE-ERA.net* (Tufiș *et al.*, 2008b) ayant comme objectif la construction de systèmes de traduction automatique statistique factorisée pour des langues slaves et balkaniques (bulgare, grec, roumain, serbe, slovène), de et vers l'anglais. Ces systèmes utilisent des corpus parallèles lemmatisés, étiquetés et alignés aux niveaux propositionnel et lexical et une configuration appropriée de facteurs linguistiques en fonction des paires de langues considérées. Nous avons ainsi adapté, pour la paire de langues français - roumain, un système factorisé anglais - roumain développé dans la même optique (Ceaușu,

2009) à l'*Institut de Recherches en Intelligence Artificielle « Mihai Drăgănescu »* (*ICIA*)[7].de l'Académie Roumaine de Bucarest.

Pour aligner lexicalement le corpus parallèle d'apprentissage, nous sommes partis de l'idée qu'un alignement performant combinant des techniques statistiques et des informations linguistiques (Tiedemann, 2003 ; Cherry et Lin, 2003 ; Tufiș *et al.*, 2006 ; Schrader, 2006 ; Hermjakob, 2009 ; Cendejas *et al*., 2009 ; Pal *et al*., 2013) améliore les résultats des systèmes d'alignement statistiques purs (Brown *et al.*, 1993 ; Och et Ney, 2000, 2003) et, par conséquent, la qualité de la traduction automatique. En même temps, nous avons pris en considération le constat qu'une appréciation significative de l'alignement ne mène qu'à une faible amélioration de la traduction (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007).

Afin de vérifier ces deux constats pour le français et le roumain, nous avons mis en place un système d'alignement lexical adapté à la paire de langues traitées. Celui-ci applique, dans un premier temps, l'aligneur statistique *GIZA++* (Och et Ney, 2000, 2003) et exploite ensuite deux modules supplémentaires : un module à base de cognats (des mots bilingues similaires au niveau orthographique et/ou phonétique, susceptibles de partager un sens commun) et un autre utilisant des règles d'alignement lexical linguistiques. Ces modules ont été développés suite à l'étude linguistique des erreurs d'alignement lexical fournies par *GIZA++*. Cette étude a révélé principalement des erreurs apparues au niveau des cognats, des différences morphosyntaxiques entre les deux langues étudiées, des contraintes stylistiques en traduction humaine et des collocations (autrement dit des expressions poly-lexicales dont les mots entretiennent une relation lexico-syntaxique (Todirașcu *et al*., 2008)).

Dans sa version initiale, le système intègre aussi un dictionnaire de collocations (Todirașcu *et al*., 2008). Cependant, ce dictionnaire n'apparaît plus dans la version finale du système à cause de son impact négligeable sur les résultats d'alignement finaux.

Pendant nos expériences de traduction automatique effectuées à l'aide du décodeur *MOSES* (Koehn *et al*., 2007), nous avons vérifié initialement si l'exploitation de facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction (Koehn et Hoang, 2007 ; Avramidis et Koehn, 2008 ; Tufiș *et al.*, 2008b) apprécie les résultats de la traduction automatique, par rapport à un système de base statistique standard (Koehn *et al*., 2003).

[7] http://www.racai.ro/en/about-us/about-racai/

Puis, nous avons évalué, d'un côté, l'influence de l'alignement lexical mis en place sur les résultats des systèmes de traduction factorisés construits dans les deux directions du processus de traduction. D'un autre côté, nous avons étudié l'influence de chaque module supplémentaire d'alignement lexical développé sur les sorties des systèmes factorisés afin de tester leur efficacité individuelle.

Hormis l'ensemble des ressources linguistiques français - roumain ayant résulté à la suite de ce travail (voir l'Annexe 8), notre principale contribution est la mise en place d'une méthode d'alignement lexical adaptée à la paire de langues étudiées qui permet, d'une part, l'amélioration des résultats d'un système de base purement statistique. D'autre part, elle permet l'appréciation des résultats des systèmes de traduction automatique factorisés, construits dans les deux sens du processus de traduction.

Le premier objectif de notre thèse consiste ainsi dans la constitution de ressources linguistiques pour un système de traduction automatique statistique factorisée français - roumain. Notre second objectif réside dans l'étude de l'influence des informations linguistiques utilisées (lemmes et propriétés morphosyntaxiques), tant sur les résultats de l'alignement lexical que sur la qualité des traductions fournies par les systèmes de traduction automatique factorisés développés.

Suite à cette introduction, cette thèse est organisée ainsi :

Le chapitre 2 est dédié à l'état de l'art des systèmes de traduction automatique. Ce chapitre comprend deux grandes sections. La première présente les approches de la traduction automatique dans leurs contextes historiques, ainsi que leurs avantages et leurs inconvénients. La deuxième section présente en détail l'approche de la traduction automatique statistique, car c'est celle-ci qui fait l'objet de notre étude.

Le chapitre 3 décrit l'architecture du système de traduction factorisé français - roumain. Sont présentés les outils utilisés (le décodeur *MOSES* (Koehn *et al*., 2007) et l'aligneur statistique *GIZA++* (Och et Ney, 2000, 2003)), mais aussi les corpus bilingues parallèles et monolingues disponibles pour la paire de langues étudiées, ainsi que leur prétraitement.

Le chapitre 4 est consacré au système d'alignement lexical français - roumain développé.

Tout d'abord, est présenté et évalué le système d'alignement lexical de base purement statistique, obtenu par l'application de *GIZA++* (Och et Ney, 2000, 2003). Une analyse linguistique des résultats d'alignement, suivie d'une étude du corpus bilingue parallèle au niveau de la traduction humaine sont effectuées, afin de repérer les erreurs d'alignement et de proposer d'éventuelles solutions de correction. Suite à cette étape, une base de règles heuristiques linguistiques de correction, pour certaines catégories d'erreurs identifiées, est constituée.

Ensuite, le module d'identification et d'alignement automatique de cognats mis en place afin de corriger les erreurs apparus au niveau de ces mots est décrit en détail et évalué. Une comparaison des résultats obtenus avec ceux fournis par des méthodes traditionnelles de détection automatique de cognats y figure aussi.

Puis, le module implémentant la base de règles heuristiques linguistiques définie est également décrit et évalué.

La dernière étape du système d'alignement initial et notamment l'intégration d'un dictionnaire de collocations ainsi que son évaluation y sont présentées aussi.

Enfin, l'évaluation globale du système d'alignement lexical et une comparaison des résultats finaux avec ceux fournis par d'autres systèmes existants et incluant aussi le roumain, figurent également dans ce chapitre.

Le chapitre 5 contient la description des corpus utilisés et des systèmes de traduction automatique statistiques purs et factorisés mis en place et évalués dans les deux sens du processus de traduction.

Finalement, le chapitre 6 comprend les conclusions et les perspectives de notre travail de thèse.

2. État de l'art des systèmes de traduction automatique

Tout d'abord, nous avons fait un inventaire des approches existantes dans le domaine de la traduction automatique, dans leurs contextes historiques, afin de voir des systèmes disponibles pour différentes paires de langues ainsi que leurs avantages et leurs inconvénients.

Dans les sections suivantes, seront présentés l'historique de la traduction automatique (section 2.1.), ainsi que l'état de l'art effectué sur la traduction automatique statistique (section 3.2.).

2.1. Historique de la traduction automatique

Comme le remarque Koehn (2010 : 14) :

> *« The history of machine translation is one of great hopes and disappointments. We seem currently to be riding another wave of excitement, so it is worth keeping in mind the lessons of the past ».*

La recherche en traduction automatique a connu, dès ses origines, vers le milieu du XXème siècle, des hauts et des bas, pour culminer aujourd'hui comme une tâche de pointe de la linguistique computationnelle. Cela est dû, non seulement au fait qu'actuellement, au temps de la mondialisation, il existe un besoin réel de traduire automatiquement des documents dans toutes les langues du globe, mais aussi au fait qu'elle représente un noyau de convergence pour de nombreuses autres applications de traitement automatique de langues (extraction de terminologie, segmentation lexicale ou propositionnelle, analyse linguistique de divers niveaux - syntaxique, morphologique, sémantique, etc.).

L'historique réalisé dans cette section est basé principalement sur les travaux de Hutchins et Somers (1992), Léon (2001, 2002), Hutchins (1993, 1994), Loffler-Laurian (1996), Iancu (2007), Boitet (2008) et Koehn (2010).

Les premières idées matérialisées concernant la traduction automatique ont vu le jour en France et en Russie. En effet, en 1933, le Français d'origine arménienne George Artsrouni avait breveté un dispositif de stockage sur bande de papier, pouvant être utilisé afin de trouver l'équivalent de traduction de tout mot dans une autre langue. En outre, dans la même année, le

Russe Petr Smirnov-Trojanskij avait envisagé la traduction mécanique en trois pas (Hutchins et Somers, 1992) :

- Premièrement, un éditeur connaissant seulement la langue source effectuait l'analyse « logique » des mots sources dans leurs formes de base et fonctions syntaxiques ;

- Ensuite, une machine transformait des séquences de ces formes de mots et fonctions dans leurs séquences correspondantes cibles ;

- Enfin, un deuxième éditeur connaissant seulement la langue cible convertissait les sorties de l'étape précédente dans les formes correctes de la langue cible.

Trojanskij avait breveté la machine utilisée pendant la deuxième étape du processus décrit ci-dessus mais il pensait que la première étape, c'est-à-dire l'analyse logique des mots sources, pouvait être également mécanisée.

L'intérêt pour développer des systèmes de traduction automatique se concrétise avec l'arrivée des ordinateurs. C'est le mathématicien britannique Alan Turing qui construit le premier ordinateur. Il décode ainsi, à l'aide des ordinateurs, les messages cryptés produits par la machine *Enigma* créée en 1920 par les Allemands et utilisée pendant la deuxième guerre mondiale dans un but militaire. En effet, *Enigma* était une machine électromécanique pour chiffrer et déchiffrer l'information. Ainsi, le concept de « décodage » de langues, utilisé par la suite dans le domaine de la traduction automatique, « semblait comme une métaphore apte » (Koehn, 2010 : 15) pour ce domaine.

Il y a plusieurs périodes qui ont marqué la recherche en traduction automatique et qui seront abordées tout au long de cette section (Léon, 2002)[8] :

1) 1948-1960 : la période des idées et des expérimentations dans le domaine ;

2) 1960-1966 : l'époque où l'analyse syntaxique est mise en avant ;

3) 1966-1980 : cette période est marquée par l'arrêt des recherches en traduction automatique comme conséquence du rapport *ALPAC* (*Automatic Language*

[8] http://histoire-cnrs.revues.org/3461

Processing Advisory Committee)[9], étant caractérisée toutefois par la survie et la « force brute » des grands systèmes ;

4) 1980-1990 : pendant cet intervalle ont lieu le tournant japonais et l'automatisation de la communication ; C'est l'époque de la commercialisation des produits au niveau mondial et la traduction devient donc une condition indispensable pour les échanges commerciaux. Apparaissent aussi les systèmes d'aide à la traduction (cf. sous-section 2.1.2.) ;

5) Depuis 1990, les méthodes empiriques sont de retour.

Concernant la recherche en traduction automatique, deux approches importantes se sont développées dans le domaine : les méthodes dites *expertes*, qui nécessitent des connaissances linguistiques apportées par des experts humains (cf. sous-section 2.1.1.), et les méthodes dites *empiriques* qui puisent les connaissances à partir de données textuelles brutes très volumineuses (cf. sous-section 2.1.3.).

D'une part, l'approche experte comprend trois méthodes principales :

- la méthode directe de traduction automatique ; il s'agit d'une traduction de type mot-à-mot qui utilise un dictionnaire bilingue et des règles de traduction basiques.

- le transfert ; cette méthode utilise des dictionnaires et des règles linguistiques monolingues et bilingues qui assurent la traduction d'un texte de la langue source dans un texte de la langue cible.

- l'interlangue ou la traduction par langue pivot ; cette méthode fait appel à une langue intermédiaire afin de traduire un texte source dans un texte cible.

D'autre part, l'approche empirique concerne les méthodes basées sur l'exemple et les méthodes statistiques. Celles-ci exploitent des corpus bilingues parallèles très volumineux afin d'en extraire, par le biais des statistiques, des exemples de traduction ou des séquences bilingues de traduction qui sont utilisés, par la suite, dans le processus traductionnel d'un texte de la langue source dans un texte de la langue cible.

[9] Les conclusions et les conséquences du rapport *ALPAC* seront présentées plus loin.

Le processus de traduction de chacune de ces approches comprend principalement trois étapes successives : l'analyse, le transfert et la génération (ou la synthèse). Ainsi, la première étape consiste à analyser divers niveaux du texte source. L'idée est de voir comment ce texte peut être structuré ou découpé pour pouvoir être traduit. Ensuite, dans un deuxième temps, se fait le transfert à partir de la langue source vers la langue cible. Finalement, la troisième phase génère la traduction du texte source de départ à partir des résultats obtenus antérieurement. Les trois étapes mentionnées sont appliquées différemment en fonction de l'approche utilisée. Celles-ci sont représentées schématiquement par le bien connu triangle de Bernard Vauquois (cf. Figure 1 suivante).

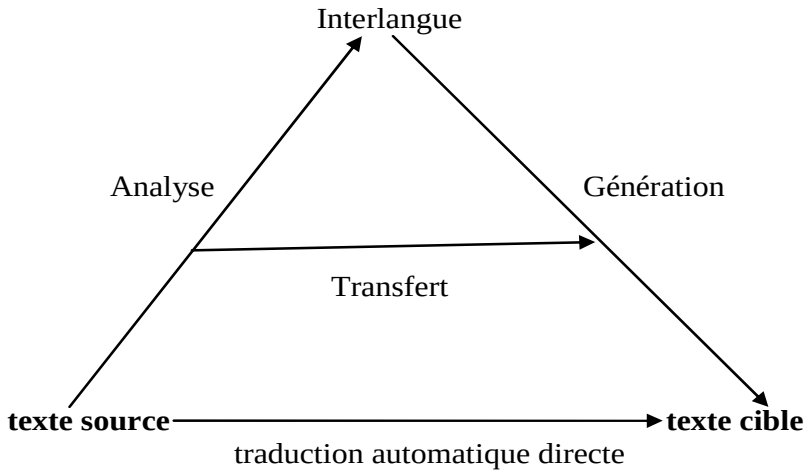


Figure 1. Le triangle de Vauquois

Les deux approches de la traduction automatique - experte et empirique - sont présentées ci-dessous.

2.1.1. L'approche experte

Dans cette sous-section, seront décrits le fonctionnement des trois principales méthodes de l'approche experte - la traduction automatique directe, le transfert et l'interlangue - ainsi que leur contexte historique.

Notons que les premiers pas de la recherche en traduction automatique ont été plutôt caractérisés par l'approche directe de la traduction automatique. Cette approche fera l'objet de la sous-section suivante.

2.1.1.1. Les débuts : la traduction automatique directe

L'intervalle 1948-1960 est une époque florissante d'idées et d'expérimentations pour la traduction automatique (Léon, 2002). Pendant cette période, des idées riches sont apparues telles que les méthodes statistiques et empiriques utilisant des corpus, les méthodes sémantiques par langues intermédiaires, les analyseurs syntaxiques. À cette même époque, les premières ressources pour le traitement automatique de langues telles les dictionnaires électroniques voient le jour. La plupart des idées (théoriques ou méthodologiques) de cette période ont mené au développement du traitement automatique des langues (Léon, 2002).

Ce sont l'américain Warren Weaver de la Fondation Rockefeller et le cristallographe britannique Andrew D. Booth qui ont abordé, pour la première fois, l'idée d'utiliser des ordinateurs pour traduire. Booth étudie alors la mécanisation d'un dictionnaire bilingue et collabore avec Richard H. Richens de Cambridge qui, de son côté, avait utilisé des cartes perforées pour effectuer une traduction mot-à-mot des résumés scientifiques.

En juillet 1949, c'est Warren Weaver qui avance l'idée de la traduction automatique et les méthodes à mettre en place par un *Mémorandum*. Ainsi, ont été proposées : les techniques de cryptographie utilisées pendant la guerre, les méthodes statistiques, la théorie de l'information de Shannon (1948), la logique sous-jacente et les caractéristiques universelles du langage (Hutchins et Somers, 1992).

La recherche en traduction automatique démarre alors pendant quelques années dans certains centres américains et, en 1951, Yehoshua Bar-Hillel devient le premier chercheur à temps plein dans le domaine, au *MIT*[10]. En 1952, il organise une conférence sur la traduction automatique où il y est entre autres reconnu le besoin des traducteurs humains pour la pré- ou la post-édition et où il y est proposé le travail sur la syntaxe.

En janvier 1954 a lieu la première démonstration publique, à New York, d'un système de traduction automatique issu d'un projet collaboratif de Leon Dostert de l'Université Georgetown et *IBM*. Ce système traduisait en anglais un ensemble de 49 phrases sources russes soigneusement choisies. Il utilisait, dans le processus de traduction, un vocabulaire très limité contenant 250 mots et six règles de syntaxe. Malgré ses limites évidentes, le système avait impressionné au point de rendre possible le financement considérable de plusieurs

[10] Massachusetts Institute of Technology

projets de traduction automatique aux États-Unis, en Grande-Bretagne, dans l'Union Soviétique puis successivement au Japon (1956), en Tchécoslovaquie (1957), en Chine (1958-1959), en Italie et en France (1959), au Mexique (1960), en Belgique (1961). Toutefois, en République Fédérale d'Allemagne, en Suède et en Finlande les travaux en traduction automatique n'étaient guère développés. Dans le contexte de la guerre froide, le russe était principalement concerné par ces recherches aussi bien dans des buts militaires et politiques que scientifiques (Léon, 2002).

L'époque suivante dans le déroulement des recherches concernant la traduction automatique, 1960-1966, promeut l'analyse syntaxique « comme la seule voie de recherche possible pour faire avancer la traduction automatique » (Léon, 2002)[11], ce qui mènera à donner la priorité à la recherche en linguistique computationnelle au détriment de la traduction automatique. Ainsi, l'analyse syntaxique s'impose exclusivement grâce au développement des grammaires formelles facilement applicables par des analyseurs syntaxiques, tels que la grammaire catégorielle de Yehoshua Bar-Hillel (1953) et la grammaire générative de Noam Chomsky (1955). Ce fait est appuyé également par la disparition des méthodes empiriques et statistiques et l'éloignement des méthodes sémantiques par langues intermédiaires.

En 1966, sur ce terrain propice à l'arrêt des recherches en traduction automatique, apparaît le rapport *Language and Machines, Computers in Translation and Linguistics* de l'*ALPAC* qui mettait en avant la linguistique computationnelle et stoppait le financement des recherches en traduction automatique aux États-Unis. Les conclusions et les conséquences de ce rapport seront présentées ci-dessous.

Les premières expériences (la traduction russe - anglais fournie par le système de l'Université Georgetown et *IBM* en 1954) promettaient que le problème de la traduction automatique serait résolu assez vite et donnaient un grand optimisme aux recherches dans le domaine. Il y avait aussi ceux qui prônaient le scepticisme en avançant l'idée que certains problèmes, comme ceux liés à la désambigüisation sémantique, ne pouvaient pas être résolus de manière automatique.

[11] http://histoire-cnrs.revues.org/3461#tocto1n1

Avec l'apparition du rapport *ALPAC*, l'enthousiasme et les recherches en traduction automatique ont toutefois pris fin. En effet, les agences gouvernementales des États-Unis, ayant financé ces recherches, ont chargé la commission *ALPAC* d'effectuer une étude des perspectives dans le domaine. La traduction automatique est jugée en termes de vitesse, précision et coût par rapport à la traduction humaine. Le rapport *ALPAC* concluait, entre autres (Hutchins et Somers, 1992 ; Koehn, 2010) :

- la traduction automatique est plus lente, moins précise et deux fois plus chère que la traduction effectuée par les humains ;

- la post-édition des sorties n'est pas moins chère ou plus rapide que les traductions humaines ;

- il existe très peu de littérature scientifique russe importante à traduire et les traducteurs ne manquent pas ;

- il n'y a donc pas de perspectives pour une traduction automatique utile.

Vu ces inconvénients, le rapport *ALPAC* suggère qu'il n'y a plus aucun besoin pour des financements futurs dans le domaine de la traduction automatique et la recherche est stoppée ainsi complètement pour une bonne décennie aux États-Unis. L'image de la traduction automatique sera aussi marquée négativement pour plusieurs années. Le rapport *ALPAC* recommande néanmoins la mise en place d'outils d'aide à la traduction tels que les dictionnaires automatiques. La recherche en linguistique computationnelle est également encouragée.

Ce rapport a été largement critiqué : il condamnait à tort la traduction automatique pour le fait que ses sorties nécessitaient une post-édition et il évaluait mal les facteurs économiques en jeu. Malgré les nombreuses critiques, celui-ci a détourné l'intérêt pour la traduction automatique pour une bonne période de temps, surtout aux États-Unis. Koehn (2010) observe que même si le rapport *ALPAC* a peut-être injustement considéré seulement l'objectif d'obtenir une traduction de haute qualité, l'expérience montre les dangers des promesses excessives concernant les capacités des systèmes de traduction automatique *: « While the ALPAC report may have unfairly considered only the goal of high-quality translation, the experience shows the dangers of over-promising the abilities of machine translation systems. »* (Koehn, 2010 : 16).

Les systèmes développés jusqu'au rapport *ALPAC* ont été généralement basés sur la méthode directe de traduction (Hutchins, 1994). Dans le triangle de Vauquois (cf. Figure 1, section 2.1.), la méthode directe de la traduction automatique se situe à la base de ce triangle. Cette méthode effectue la traduction d'un texte source de départ mot par mot par l'utilisation d'un dictionnaire bilingue contenant des règles permettant la traduction de chaque mot source. L'analyse du texte source n'est pas profonde, étant seulement une analyse morphologique. Après la traduction des mots sources (étape du transfert lexical), les mots de chaque phrase sont réordonnés en fonction de règles simples de ré-ordonnancement local. Finalement, la traduction est générée en fonction des critères morphologiques.

Cette approche directe de la traduction automatique a l'avantage d'être facilement implémentable car elle utilise peu de ressources et les règles définies sont basiques. Cette méthode peut être efficace pour des phrases où les traductions sont plutôt littérales et dans des domaines restreints comme, par exemple, les informations liées aux produits au sein d'une entreprise ou la météorologie. D'ailleurs, l'un des systèmes utilisant cette approche, *TAUM-METEO* (Université de Montréal), montre son efficacité dans la traduction des bulletins météorologiques pour la paire de langues anglais - français au Canada. Ce système est devenu opérationnel en 1976 et a continué d'exister jusqu'à nos jours. Un autre système actuel utilisant l'approche directe est *Intertran*[12] qui traduit des mails, des mémos, des pages Web, etc. dans plusieurs paires de langues. Notons que ce système comprend aussi la combinaison français - roumain.

Cependant, l'approche directe rencontre des difficultés importantes dans le cas de la traduction oblique où le transfert lexical se fait plutôt par groupes de mots. Ainsi, cette approche n'aboutit pas à traduire les séquences de mots plus ou moins compositionnelles (noms composés, expressions, locutions) si celles-ci ne se trouvent pas dans le dictionnaire utilisé. Également, le ré-ordonnancement des mots cibles n'est pas toujours correct en l'absence de connaissances de grammaire explicites liées à la langue cible. À cause de ces inconvénients majeurs, cette approche de la traduction automatique n'est plus guère implémentée actuellement.

Si la traduction automatique directe était la plus utilisée jusqu'au rapport *ALPAC*, l'approche dominante depuis ce rapport jusqu'à la fin des années quatre-vingts a été celle basée sur des

[12] www.tranexp.com:2000/Translate/result.shtml

règles linguistiques, telles que : règles d'analyse syntaxique, lexicales, de désambiguïsation, de transformation des arbres syntaxiques, de transfert lexical, règles de génération syntaxiques et morphologiques, etc. (Hutchins, 1994). Cette approche, appelée aussi *indirecte*, fait l'objet de la sous-section qui suit.

2.1.1.2. La traduction automatique indirecte

L'approche indirecte de la traduction automatique concerne les systèmes par transfert (cf. sous-section 2.1.1.2.1.) et les systèmes par langue pivot (ou l'interlangue) (cf. 2.1.1.2.2.).

2.1.1.2.1. Le transfert

Dans les sous-sections suivantes, seront présentés l'architecture des systèmes par transfert (cf. sous-section 2.1.1.2.1.1.), ainsi que des exemples de systèmes basés sur cette approche de la traduction automatique (cf. sous-section 2.1.1.2.1.2).

2.1.1.2.1.1. L'architecture des systèmes par transfert

Le processus de traduction des systèmes par transfert se déroule selon trois étapes principales :

1) Premièrement, un module procède à l'analyse linguistique du texte d'entrée source. Pour ce faire, le module utilise des dictionnaires et des règles de grammaire de la langue source. Le résultat de cette étape consiste dans la représentation du texte source sous forme d'arbres étiquetés. L'analyse du texte source est donc plus poussée par comparaison à l'approche directe de la traduction automatique.

2) Ensuite, un module de transfert transforme la représentation du texte de la langue source dans la représentation de ce texte en langue cible. En effet, il s'agit de la transformation des arbres sources dans les arbres correspondants dans la langue cible. Pour ce faire, ce module utilise des dictionnaires bilingues (langue source - langue cible) et des règles de transfert (syntaxiques, lexicales ou sémantiques) bilingues. Ces règles définissent les correspondances linguistiques et structurales entre la langue source et la langue cible et sont censées s'appliquer dans les deux sens du processus de traduction. Celles-ci décrivent le bon ordre des mots dans la langue cible, suppriment ou ajoutent des mots cibles manquants, etc.

3) Finalement, un module de génération (ou de synthèse) fournit le texte de sortie cible à partir de la représentation cible obtenue antérieurement. Ce module utilise des dictionnaires et des règles de grammaire de la langue cible.

L'architecture du système de traduction automatique par transfert est donnée dans la Figure 2 ci-dessous.

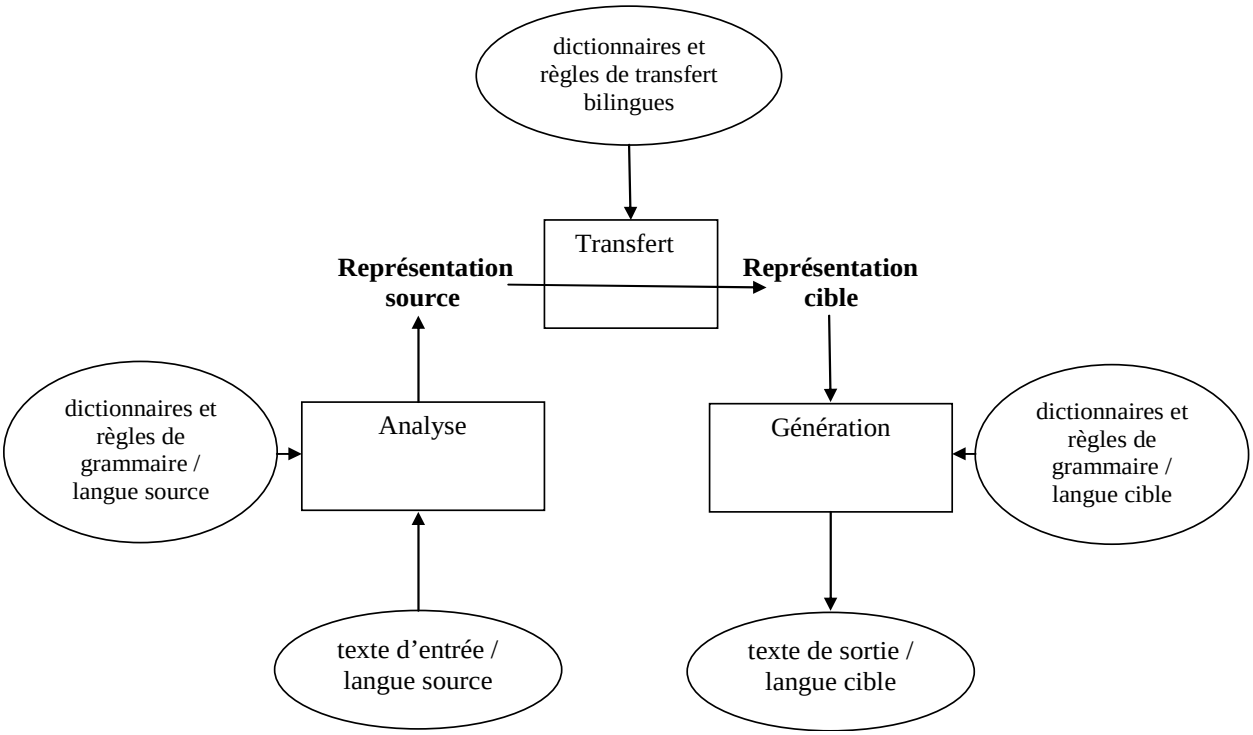
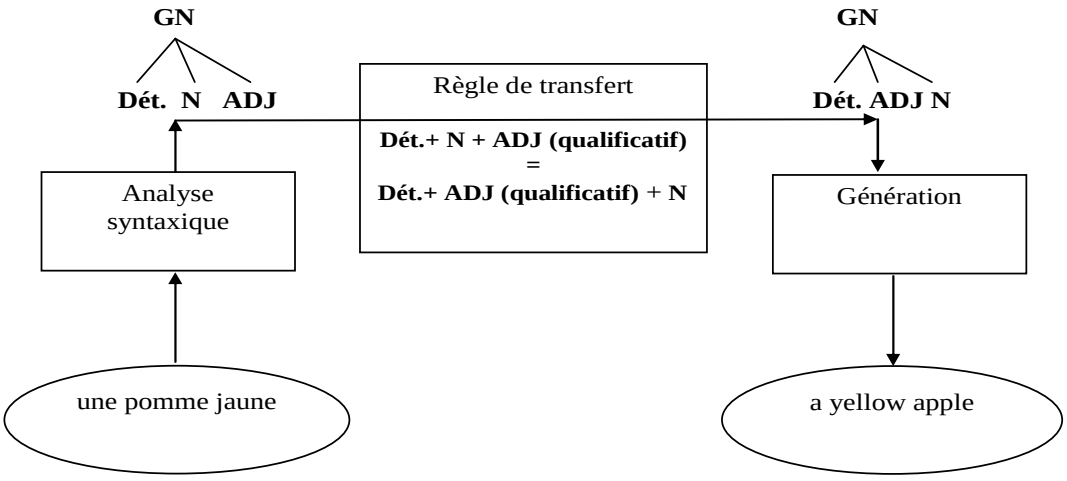


Figure 2. L'architecture du système de traduction automatique par transfert

Un exemple de traduction par transfert du français vers l'anglais apparaît dans la Figure 3 suivante. Cet exemple concerne le couple de séquences français - anglais *une pomme jaune* vs. *a yellow apple.*



GN : groupe nominal ; Dét. : déterminant ; ADJ : adjectif ; N : nom ;
« + » : « suivi de » ; « = » : « équivalent à »

Figure 3. Exemple de traduction par transfert du français vers l'anglais

Dans cet exemple, la séquence source française *une pomme jaune* est analysée seulement syntaxiquement. Mais, en fonction des couples de langues traitées, cette analyse peut être plus profonde (sémantique, par exemple). Ensuite, le module de transfert applique une règle de transfert syntaxique pour transformer l'arbre syntaxique source dans son correspondant cible. Cette règle décrit la position de l'adjectif qualificatif par rapport au nom qu'il détermine pour les deux langues : l'adjectif qualificatif précède le nom en anglais tandis qu'en français celui-ci suit généralement le nom. Enfin, la séquence cible anglaise *a yellow apple* est générée en sortie à partir de l'arbre cible obtenu antérieurement.

Les méthodes de traduction automatique par transfert ont l'avantage d'être assez facilement adaptables pour d'autres paires de langues car chaque module de leur architecture est indépendant. De ce fait, pour ajouter une nouvelle langue, il suffit de construire une grammaire pour cette nouvelle langue et de définir un ensemble de règles de transfert entre celle-ci et les autres langues du système. Ainsi, par l'utilisation de connaissances linguistiques avancées, les systèmes par transfert (comme, par exemple, *Systran*[13]) donnent des résultats de traduction performants pour les différentes paires de langues incorporées. Notons que *Systran* à ces débuts utilisait l'approche directe de la traduction automatique, tandis que depuis 2009

[13] http://www.systran.fr/systran

celui-ci est devenu hybride par l'intégration des méthodes statistiques également (cf. sous-section suivante où *Systran* est décrit plus en détail).

Il est donc évident que ces systèmes ont l'inconvénient d'être très coûteux en termes de temps et de ressources humaines nécessaires pour leur implémentation. En effet, pour construire des grammaires pour chaque langue et aussi des grammaires de transfert bilingues entre les paires de langues traitées, il faut faire appel à des spécialistes pour chaque couple de langues considérées. De plus, la description d'une grammaire exhaustive pour chaque langue s'avère être une tâche assez fastidieuse requérant un temps considérable. Des exemples de systèmes utilisant la méthode du transfert et leur contexte historique seront donnés dans la sous-section suivante.

2.1.1.2.1.2. Les systèmes par transfert

Marquée par l'arrêt des recherches en traduction automatique comme conséquence du rapport *ALPAC*, la période 1966-1980 est néanmoins caractérisée par la survie et la « force brute » des grands systèmes (Léon, 2002)[14].

Une première catégorie des systèmes ayant survécu à la crise est constituée par les systèmes industrialisés qui « marchent » (Léon, 2002)[15] tels que : *Systran* (créé en 1970[16] par la société *Systran*[17] aux États-Unis), *TAUM-METEO* (créé en 1976 à l'Université de Montréal), *Logos* (créé en 1980 par la société *Logos* aux États-Unis), *METAL*[18] (créé en 1980 à l'Université de Texas), etc. Ces systèmes fournissent des traductions qui ne sont pas destinées à la publication car ils se plient à un besoin interne des entreprises. De ce fait, il est accepté que les traductions obtenues puissent être de moins bonne qualité.

Les systèmes de cette catégorie sont basés soit sur l'approche directe de la traduction (comme *TAUM-METEO*), soit sur la méthode de transfert syntaxique (comme *Systran*, *Logos*, *METAL*). Ceux-ci exploitent des dictionnaires bilingues riches pour les domaines des textes utilisés, l'analyse linguistique effectuée n'est pas abstraite ou très profonde, l'analyse

[14] http://histoire-cnrs.revues.org/3461#tocto1n1
[15] http://histoire-cnrs.revues.org/3461#tocto1n1
[16] Les années données entre parenthèses pour chaque système sont les années à partir desquelles ces systèmes sont devenus opérationnels.
[17] *System Translation*
[18] *Mechanical Translation and Analysis of Languages*

sémantique manque presque entièrement et l'exploitation des connaissances non-linguistiques est carrément absente (Hutchins, 1994).

Le système le plus poussé a été *Systran*. Il est lui-même issu d'un système mis en œuvre dès les années cinquante à l'Université Georgetown à Washington D. C. et constitue l'un des rares systèmes à avoir franchi le niveau d'expérimentation pour donner des traductions brutes exploitables. Les traductions fournies appartiennent à un domaine très spécialisé et répondent aux besoins internes d'une entreprise. Ces caractéristiques lui ont donné le surnom de « force brute » (Léon, 2002)[19].

La société *Systran* a été fondée en 1968 aux États-Unis. Depuis 1970, le système russe - anglais a été utilisé par la *US Air Force*. Ensuite, en 1976, une version français - anglais du système a été achetée par la Commission Européenne (Koehn, 2010). À l'heure actuelle, *Systran* est disponible pour 52 paires de langues, étant considéré comme le leader mondial des technologies de traduction automatique. À ces origines, rappelons-le, *Systran* utilisait l'approche directe de la traduction automatique. Depuis 2009, le système combine l'approche à base de règles linguistiques et les statistiques[20], étant le premier moteur de traduction hybride mis sur le marché. Ainsi, grâce aux techniques d'apprentissage automatique à partir de grands corpus, *Systran* est facilement adaptable à un nouveau domaine et peut fournir des traductions de qualité élevée à des coûts réduits pour différentes paires de langues. La société *Systran* a désormais son siège social en France, à Paris. Celle-ci a aussi une filiale aux États-Unis, à San Diego en Californie[21].

Une deuxième catégorie de systèmes, n'étant pas parvenu au stade commercial, a survécu à la crise provoquée par les conclusions du rapport *ALPAC*, comme *Ariane* (Boitet *et al.*, 1982), mis en œuvre par le *GETA*[22] à partir de 1971 à Grenoble, ou *SUSY*[23] (Maas, 1977) développé à l'Université de Sarrebruck, à partir de 1972. Elle dut sa survie par l'appui de grands centres de recherche, comme le *CNRS*[24], qui n'étaient principalement pas préoccupés par les enjeux économiques de la traduction automatique. Parmi d'autres projets faisant également partie des rescapés de l'histoire de la traduction automatique, on relèvera le projet *Eurotra* (King, 1981)

[19] http://histoire-cnrs.revues.org/3461#tocto1n1
[20] Les méthodes statistiques de traduction automatique seront détaillées plus loin dans la section 2.2.
[21] http://www.systran.fr/systran
[22] *Groupe d'Étude pour la Traduction Automatique*
[23] *Saarbrücker Übersetzungssystem*
[24] *Centre National de la Recherche Scientifique*

développé pour la communauté européenne et soutenu pour des raisons politiques (Léon, 2002)[25].

Les systèmes mentionnés ci-dessus possèdent les traits typiques de l'approche par transfert (Hutchins, 1994) :

- l'application des trois phases d'analyse, de transfert et de génération ;
- les étapes d'analyse et de génération suivent plusieurs niveaux distincts de morphologie, de syntaxe et de sémantique ;
- des représentations d'interface assez abstraites sous la forme d'arbres étiquetés ;
- des règles de transformation des arbres d'une étape à l'autre ;
- des traitements par lots supposant une post-édition et sans implication humaine pendant la traduction ;
- l'information pragmatique et textuelle manque presque entièrement.

Le projet *Eurotra* (1977-1994) constitue un projet d'envergure ayant initié et promu des développements linguistiques et computationnels de base importants pour le développement d'un système de traduction automatique multilingue (Hutchins et Somers, 1992).

Les objectifs de ce projet sont les suivants :

1) la mise en place d'un « prototype préindustriel » (Hutchins et Somers, 1992 : 240) pour les langues de la communauté européenne, pouvant servir de base pour un possible développement industriel, ultérieur au projet ;
2) l'encouragement de la recherche en traduction automatique en Europe.

Le prototype en cause doit intégrer neuf langues des pays de la communauté européenne et notamment le danois, le néerlandais, l'anglais, le français, l'allemand, le grec, l'italien, le portugais et l'espagnol. Ces langues forment 72 paires de langue à traiter. Le système est limité aux textes techniques appartenant au domaine de la technologie de l'information. De

[25] http://histoire-cnrs.revues.org/3461#tocto1n1

plus, les résultats visés doivent être de qualité raisonnable voire bonne sans l'implication humaine de manière significative.

Le projet *Eurotra* s'est fait remarquer en termes de nombre des personnels impliqués et de coûts ainsi que par rapport à la large distribution géographique des groupes de recherche faisant partie du projet. Pendant la période 1980-1990, approximativement deux cents chercheurs, principalement des universitaires, ont participé à *Eurotra* dans les différents pays de la communauté européenne. Les industriels se sont impliqués tardivement dans le but de développer « un prototype grandeur nature » (Loffler-Laurian, 1996 : 40).

Le groupe de recherche ayant comme tâche l'administration du projet était établi au Luxembourg. D'autres groupes des pays faisant partie de la communauté européenne sont chargés chacun de l'étude de la langue du pays respectif pendant que deux autres groupes situés à Dublin et au Luxembourg ont des tâches spécifiques telles la terminologie du domaine et la documentation du projet. Les centres de recherche impliqués dans le projet, leurs tâches spécifiques et la ville où se situe chaque centre figurent dans le Tableau 1 suivant (Hutchins et Somers, 1992). De temps en temps, d'autres groupes de recherche, comme l'*ISSCO*[26] de Genève et le *GETA* de Grenoble, se sont aussi impliqués dans le projet *Eurotra*.

[26] *Instituto per gli Studi Semantici e Cognitivi*

Tableau 1. Les participants du projet *Eurotra* (Hutchins et Somers, 1992)

Tâche	Groupe de recherche	Ville
Administration	*DG XIII-B, CEC*	Luxembourg
danois	*Københavns Universitet*	Copenhague
néerlandais	*Rijksuniversiteit Utrecht*	Utrecht
	Katholieke Universiteit Leuven	Louvain
anglais	*UMIST (University of Manchester Institute of Science and Technology)*	Manchester
	University of Essex	Colchester
français	*Université de Nancy II*	Nancy
	Université Paris VII	Paris
	Université de Liège	Liège
allemand	*IAI (Institut für Angewandte Informationswissenschaft)*	Sarrebruck
	IKP (Institut für Kommunikations-forschung und Phonetik)	Bonn
grec	*Eurotra Greece*	Athènes
	Panepistemio tou Rethymnou	Crète
italien	*Gruppo Dima*	Turin
	Università di Pisa	Pise
portugais	*Universidade de Lisboa*	Lisbonne
espagnol	*Universidad de Barcelona*	Barcelone
	Universidad Autónoma de Madrid	Madrid
Terminologie	*Dublin City University*	Dublin
Documentation	*CRETA (Centre de Recherches et d'Études en Traduction Automatique)*	Luxembourg

Comme nous l'avons précisé auparavant, les systèmes par transfert sont caractérisés par l'étape de transformation (appelée aussi *mappage*) des représentations des langues sous la forme d'arbres étiquetés. Ainsi, dans le système *Eurotra*, ont été proposées des séries de transformations telles les suivantes (Hutchins, 1994) : un arbre morphologique est transformé dans un arbre syntaxique, un arbre syntaxique dans un arbre sémantique, un arbre d'interface du texte source dans un arbre correspondant du texte cible, etc. Un arbre est soumis à des conditions précises, possède une structure et des traits syntaxiques ou sémantiques particuliers. Ce sont les règles de formation qui testent les arbres, c'est-à-dire leur structure et les relations représentées sont vérifiées par une grammaire. Ensuite, un arbre est éloigné s'il n'est pas vérifié par les règles grammaticales du niveau d'analyse considéré (morphologique, syntaxique, sémantique, etc.). Les grammaires et les règles de transformation posent les conditions qui restreignent les solutions de transfert d'un niveau à un autre et notamment d'un texte source à un texte cible (Hutchins, 1994). La Figure 4 ci-dessous comprend un exemple de règles de formation et de transformation qui interviennent dans les trois étapes du

processus de traduction (analyse, transfert, synthèse ou génération) dans *Eurotra* (Hutchins, 1994 : 6).

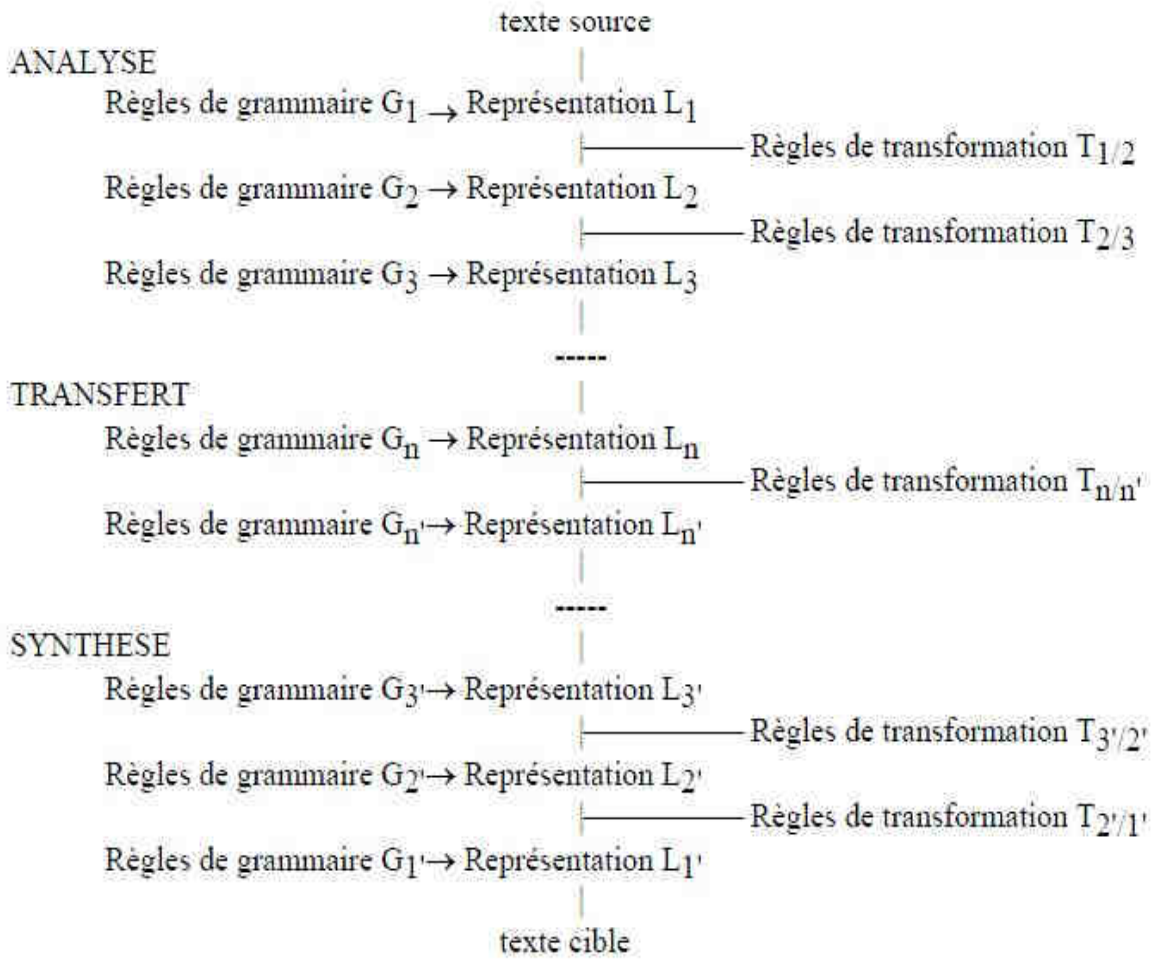


Figure 4. Règles de formation et de transformation dans *Eurotra* (Hutchins, 1994 : 6)

L'avantage d'un tel système est sa modularité qui permet facilement l'ajout d'autres paires de langues, mais il reste toutefois très coûteux en ressources humaines et matérielles (cf. sous-section 2.1.1.2.1.1.). Les performances du système *Eurotra* ont été évaluées par un comité formé par la Communauté Économique Européenne (*CEE*) en 1989 à cet effet. Le rapport issu de cette démarche de la *CEE*, dénommé le *Rapport Danzin* (1990), formulait des perspectives pour *Eurotra* qui dépassait tout autre système de traduction automatique au niveau syntaxique. Cependant, le niveau sémantique n'était pas assez développé et le système n'était pas capable de fournir une traduction de qualité élevée. Les traductions nécessitaient donc une pré- ou post-édition.

Ainsi, le projet *Eurotra* n'a finalement pas abouti au stade industriel. Il a cependant initié et promu des recherches linguistiques et computationnelles de base importantes dans le domaine de la traduction automatique. Après cet échec partiel d'*Eurotra*, les regards se sont tournés vers les systèmes d'aide à la traduction qui seront présentés plus loin, dans la sous-section 2.1.2.

Dans la sous-section suivante, sera décrite la deuxième méthode de l'approche experte indirecte et notamment la méthode par langue pivot ou l'interlangue.

2.1.1.2.2. L'interlangue

Une tendance de la recherche en traduction automatique dans les années 1980 et 1990 a été le développement des systèmes utilisant une langue pivot (dénommée aussi *interlangue*) pour représenter le sens indépendamment d'une langue donnée (Koehn, 2010). Cette langue pivot est une langue neutre qui représente le sens des textes utilisés. Il s'agit notamment d'une représentation au niveau conceptuel des textes. L'intérêt pour développer ce genre de systèmes se base sur le fait que « la traduction suppose l'expression du sens dans différentes langues et une théorie correcte du sens semble donc aborder ce problème à un niveau plus essentiel que le niveau faible de la mise en correspondance des unités lexicales ou syntaxiques » (Koehn, 2010 : 16).

Les traits distinctifs de cette méthode sont la langue pivot et des bases de connaissances dans le domaine des textes qui doivent être traduits (Hutchins, 1994). Ainsi, le texte source d'entrée est analysé initialement par un module utilisant des dictionnaires et une grammaire de la langue source. La sortie de cette étape consiste dans la représentation abstraite du texte source en langage pivot. Un dictionnaire bilingue (langue source - langue cible) sert pour trouver les correspondances dans la langue cible. Ensuite, à partir de cette représentation, un module de génération fournit la traduction du texte source d'entrée à l'aide de dictionnaires et d'une grammaire de la langue cible. La Figure 5 suivante schématise l'architecture de l'approche par langue pivot.

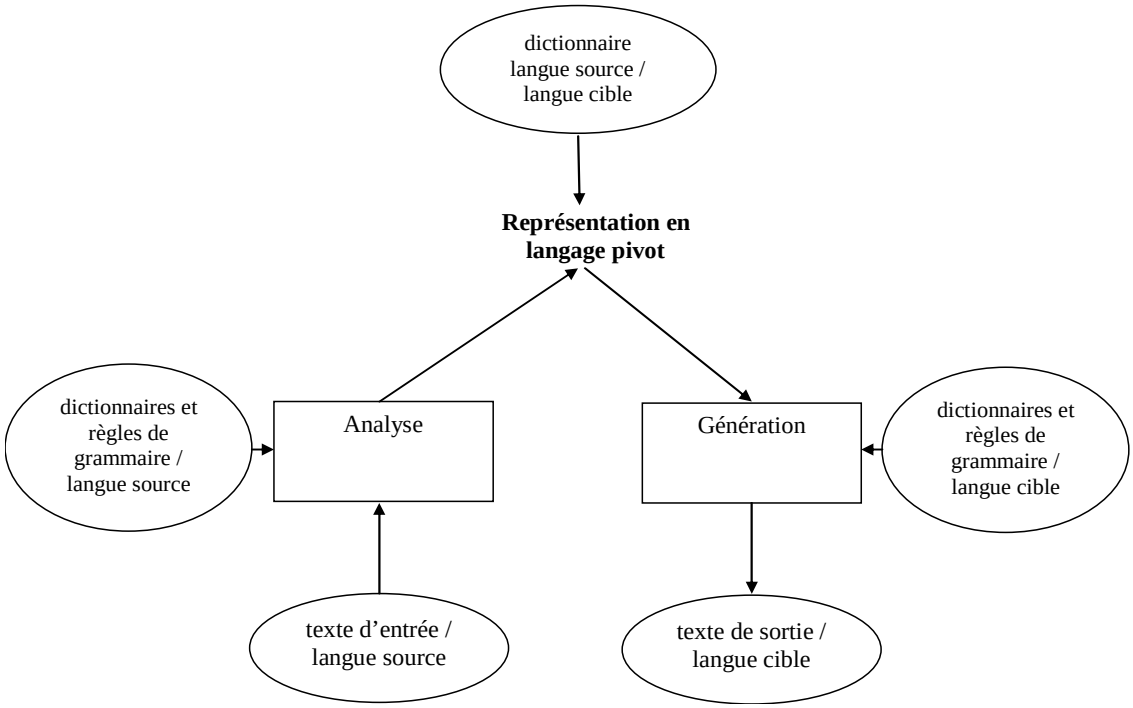


Figure 5. L'architecture du système de traduction automatique par langue pivot

Si l'interlangue a été bien construite pour un domaine donné, cette approche a l'avantage d'être plus aisément implémentable pour de nouvelles paires de langues qu'une approche par transfert. En effet, cette méthode requiert seulement les grammaires de la langue source et de la langue cible par comparaison à l'approche par transfert qui demande en plus une base de règles de transfert bilingues.

La limite de cette approche consiste toutefois dans la difficulté de définir une interlangue correcte pour différents domaines. Comme cette méthode nécessite la définition de tous les concepts possibles pour les langues considérées, la difficulté réside dans l'identification de ces concepts pouvant présenter des différences importantes d'une langue à l'autre. Par exemple, le concept du roumain *nepot* - une « personne considérée en rapport avec ses grands-parents ou ses oncles et ses tantes »[27] - correspond en français à deux concepts : *neveu* « personne considérée en rapport avec ses oncles et tantes[28] » et *petit-fils* « personne considérée en rapport avec ses grands-parents[29] ». De plus, il existe des cas où certains

[27] Définition extraite du *Dictionnaire explicatif de la langue roumaine*, 1998.
[28] La définition de *neveu* dans *Le Nouveau Petit Robert*, 2002, est la suivante : « fils du frère, de la sœur, du beau-frère, ou de la belle-sœur ; fils du cousin germain ou d'une cousine germaine ».
[29] La définition de *petit-fils* dans *Le Nouveau Petit Robert*, 2002, est la suivante : « fils d'un fils ou d'une fille par rapport à un grand-père ou à une grand-mère ».

concepts n'apparaissent pas d'une langue à l'autre. Un exemple intéressant dans ce sens est le concept *natsukashii* du japonais. Celui-ci est traduit le plus souvent par *nostalgie* dans les langues européennes même si, en réalité, il ne renvoie pas au sentiment de tristesse lié à un endroit, un événement du passé, etc. mais plutôt au bonheur ressenti grâce à ce retour en arrière.

Un exemple de système par langue pivot est *CATALYST* inauguré en 1992 et développé à l'Université de Carnegie Mellon en collaboration avec l'entreprise Caterpillar Tractor. Le but de ce projet est la traduction sans post-édition des manuels techniques de la compagnie. Un autre exemple est le système *Pangloss* (Hovy, 1993) réalisé en collaboration par les Universités de Southern California, de New Mexico State et de Carnegie Mellon. Le but de ce système est la traduction de rapports des journaux concernant les fusions des affaires. En Europe, le projet *VERBMOBIL* (Schulze-Furhoff et Abbou, 1992) a été aussi partiellement basé sur l'approche interlangue. Ce projet vise la traduction du dialogue parlé dans les affaires commerciales internationales de l'allemand vers l'anglais.

Nous venons de voir les trois méthodes de traduction automatique de l'approche experte (la méthode directe, le transfert et l'interlangue). Dans la sous-section suivante, seront présentés les systèmes commerciaux d'aide à la traduction développés dans un contexte où, d'un côté, les Japonais montaient en puissance et misaient sur une société de l'information et du commerce international qui montrait un besoin imminent des traductions à l'échelle mondiale et où, d'un autre côté, le projet *Eurotra* prenait fin en Europe, montrant les limites de la traduction automatique.

2.1.2. Les systèmes d'aide à la traduction

Pendant l'intervalle 1980-1990 ont lieu le tournant japonais et l'automatisation de la communication (Léon, 2002)[30]. Il s'agit d'une nouvelle étape pour la traduction automatique déclenchée par le développement des micro-ordinateurs, des outils de traitement de texte et aussi par le contexte de mondialisation de la consommation. C'est l'étape de la commercialisation. La traduction des modes d'emplois et des descriptifs de produits devient une condition indispensable afin de commercialiser mondialement ces produits (Léon, 2002).

[30] http://histoire-cnrs.revues.org/3461#tocto1n1

En 1982, les Japonais annoncent le projet « 5e génération » à l'*ICOT* (*Institute for New Generation Computer Technology*), ayant comme perspective la société basée sur l'information. Ce projet est soutenu par le ministère du commerce et de l'industrie *MITI* (*Japan's Ministry of International Trade and lndustry*) du Japon. Ainsi, voient le jour les systèmes interactifs de traduction automatique assistée par des humains et les stations de travail de traduction assistée par ordinateur (*TAO*). Ces systèmes mettent à la disposition des traducteurs humains des instruments utiles pour effectuer des traductions dans un environnement de travail automatisé.

Les stations de travail de traduction assistée par ordinateur peuvent comprendre plusieurs outils et fonctionnalités (Hutchins, 1993) :

- l'accès automatique à des dictionnaires et bases de données terminologiques ;
- des outils de gestion de ressources terminologiques et de traitement de textes multilingues ;
- des dispositifs de reconnaissance optique de caractères ;
- la réception et la transmission électroniques des textes ;
- des concordanciers (outils qui fournissent les contextes gauche et droit des mots ou des séquences de mots) ;
- le stockage et l'accès des traductions déjà effectuées pour la réutilisation partielle ou la révision ;
- des outils d'alignement des traductions bilingues déjà réalisées ; Ces outils mettent en correspondance automatiquement les phrases des textes traduis pour leur utilisation ultérieure. L'ensemble de ces phrases alignées forme une mémoire de traduction.

Léon (2002)[31] remarque qu'à partir des années quatre-vingts, « la traduction automatique n'est plus confinée à une utilisation 'maison' mais fait partie de l'ingénierie linguistique, au même titre que les dictionnaires électroniques, les bases de données terminologiques et la génération de texte ».

[31] http://histoire-cnrs.revues.org/3461#tocto1n1

Pendant les années quatre-vingt-dix, des systèmes d'aide à la traduction sont développés aussi par des compagnies comme *Trados*[32]. D'autres systèmes disponibles actuellement sont les suivants : *Déjà Vu*[33], *OmegaT*[34], *Wordfast*[35], *Transit Satellite PE*[36], *IBM Translation Manager*[37], etc.

Soulignons que les systèmes d'aide à la traduction exploitent des mémoires de traduction. Ainsi, les efforts considérables pour construire de telles ressources importantes pour la traduction ont mené, avec le temps, au développement des approches empiriques basées sur ce genre de données textuelles. En effet, l'idée d'utiliser des traductions déjà effectuées par des humains, donc considérées comme fiables, semblait intéressante pour développer les recherches dans cette voix. L'approche empirique de la traduction automatique fera l'objet de la sous-section suivante.

2.1.3. L'approche empirique

Au sein de l'approche empirique se sont distinguées deux grandes méthodes basées sur les corpus bilingues parallèles : les systèmes basés sur l'exemple (cf. sous-section suivante) et les systèmes statistiques (cf. sous-section 2.1.3.2.).

2.1.3.1. Les systèmes à base d'exemples

Les systèmes basés sur l'exemple (appelés aussi systèmes par analogie) ont été développés particulièrement au Japon, dans les années 1980. C'est Makoto Nagao qui a proposé cette méthode en 1984 (Nagao, 1984) en se basant sur le constat que l'être humain fait appel à des exemples de traduction afin de traduire d'une langue source vers une langue cible. Donc, un système modélisant cet aspect serait approprié pour la traduction automatique. D'autres travaux illustrant l'intérêt japonais pour les systèmes à base d'exemples sont les recherches menées à Hitachi (Kaji *et al.*, 1992) et à Kyoto (Utsuro *et al.*, 1992) (Hutchins, 1993).

Pour développer un système par analogie, il est nécessaire, tout d' abord, de construire une base d'exemples de traduction à partir de grandes collections de textes déjà traduits

[32] http://www.translationzone.com/trados.html
[33] http://www.atril.com/fr
[34] http://www.omegat.org/en/omegat.html
[35] http://www.champollion.net/
[36] http://www.star-spain.com/en/tecnologia/transit.php
[37] http://www-03.ibm.com/software/products/en/application

auparavant par des humains, dans un domaine donné, appelés aussi des corpus parallèles. Ces exemples de traduction peuvent être des couples bilingues (source - cible) de phrases, de séquences ou de mots. Exploitant cette base préalablement constituée, le système fonctionne en trois étapes :

1) Premièrement, le système extrait les séquences sources trouvées dans la base d'exemples qui sont les plus similaires à la phrase source d'entrée ; Cette étape est appelée de *correspondance*.

2) Ensuite, le système extrait les séquences cibles correspondantes dans la base d'exemples aux séquences sources retenues antérieurement ; Cette étape peut être considérée comme un *alignement* effectué entre les séquences sources et cibles retenues.

3) Finalement, les séquences cibles obtenues pendant la deuxième étape sont ordonnées afin de générer une traduction potentielle pour la phrase source d'entrée ; Il s'agit d'une étape d'*adaptation*.

Quant à la première étape du processus de traduction et notamment le calcul de similarité entre la phrase source d'entrée et les séquences sources contenues dans la base d'exemples, plusieurs méthodes sont proposées. Ainsi, Brown (1996) découpe la phrase source d'entrée en séquences et recherche ensuite, pour chacune d'entre elles, les séquences sources les plus similaires dans la base d'exemples. En revanche, Veale et Way (1997) cherchent les similarités au niveau de la phrase source toute entière. D'autres systèmes (Nagao, 1984 ; Alp et Turhan, 2008) proposent de chercher les similarités au niveau des arbres syntaxiques représentant la phrase source d'entrée et les phrases sources contenues dans la base d'exemples. En effet, la phrase source d'entrée et les phrases sources de la base d'exemples sont traitées par un analyseur syntaxique et, ensuite, le calcul de similarité se fait au niveau des arbres obtenus. Nakazawa *et al.* (2006) utilisent un analyseur morphologique et un lexique bilingue afin d'améliorer la performance d'un système anglais - japonais. D'autres méthodes exploitent aussi des règles linguistiques pour améliorer leurs résultats (Langlais et Gotti, 2006). Un système entièrement basé sur l'analogie est *ALEPH* (Lepage et Denoual, 2005). Irimia (2008) développe une approche à base d'exemples qui intègre aussi le roumain en combinaison avec l'anglais. Un autre système incorporant le roumain mais en combinaison avec l'allemand, cette fois-ci, est proposé par Gavrilă (2012).

À la différence d'un système expert qui nécessite des ressources humaines et matérielles considérables, l'avantage d'un système à base d'exemples est que celui-ci s'améliore facilement à mesure que d'autres corpus parallèles deviennent disponibles pour différentes paires de langues. Ce fait peut constituer toutefois un inconvénient pour des langues moins dotées, c'est-à-dire des langues pour lesquelles des ressources langagières de ce genre ne sont pas largement disponibles. Un autre avantage des systèmes à base d'exemples est qu'ils exploitent des traductions effectuées par des experts humains, donc les traductions cibles proposées sont supposées avoir une qualité au moins comparable aux traductions fournies par les systèmes experts par l'utilisation de moyens moins coûteux.

Une deuxième catégorie de systèmes faisant partie de l'approche empirique est constituée par les systèmes statistiques qui seront brièvement décrits dans la sous-section suivante. Comme notre projet de thèse se base sur l'approche statistique de la traduction automatique, celle-ci sera décrite en détail dans la section 2.2. de ce chapitre, qui lui est spécialement dédiée.

2.1.3.2. Les systèmes statistiques

La traduction automatique statistique utilise également de grands corpus bilingues parallèles, donc des traductions déjà effectuées par des humains en quantité considérable. La différence principale entre cette approche et celle à base d'exemples consiste dans le fait que la traduction statistique utilise les résultats fournis par un corpus aligné (un ensemble d'exemples compilés), alors que la traduction par analogie utilise les exemples directement pendant le processus de traduction (Boitet, 2008).

Les systèmes statistiques se sont imposés entre autres facteurs grâce, d'une part, au développement d'ordinateurs puissants en termes de vitesse et de capacités de stockage ainsi que, d'autre part, grâce à la généralisation du *World Wide Web* à partir des années 2000. Ce dernier a facilité l'accès à des ressources textuelles dans différentes paires de langues. Sur ce terrain propice, l'approche statistique, exploitant de grandes quantités de données textuelles monolingues et bilingues parallèles, s'est développée dans la dernière décennie.

L'idée de la traduction automatique statistique a vu le jour au *IBM*[38] lors du projet *Candide* (Brown *et al.*, 1988), grâce au développement prometteur des techniques statistiques dans le domaine de la reconnaissance de la parole. Ainsi, par la modélisation de la traduction comme un problème d'optimisation statistique (cf. section 2.2.), le projet *Candide* a donné à la traduction automatique un fondement mathématique solide (Koehn, 2010).

Les travaux de référence (Brown *et al.*, 1990 ; Brown *et al.*, 1993) proposant des systèmes lexicaux, ont posé les fondements des développements ultérieurs dans le domaine. Ces systèmes exploitent des corpus bilingues parallèles alignés lexicalement, c'est-à-dire que les mots sources et cibles sont mis en correspondance à l'intérieur des couples de phrases bilingues parallèles. À partir de ces corpus alignés, le système extrait sa propre table de traduction représentant un ensemble d'équivalents lexicaux bilingues, par le biais de calculs statistiques. Ainsi, chaque paire d'équivalents de traduction extraite se voit assignée une probabilité de traduction. Ensuite, le système génère une possible traduction pour une phrase source donnée en entrée, en combinant des données bilingues fournies par sa table de traduction et des données monolingues de la langue cible, obtenues à partir d'un corpus monolingue. Ces données représentent des séquences de mots cibles munies de leurs probabilités qui vérifient la syntaxe de la traduction en langue cible. L'ensemble d'équivalents de traduction, extraits à partir d'un corpus bilingue parallèle aligné lexicalement et munis de leurs probabilités, forme un *modèle de traduction* (cf. sous-section 2.2.1.3.2.). L'ensemble des séquences cibles, calculées statistiquement à partir d'un corpus monolingue de la langue cible, forme un *modèle de langue* (cf. sous-section 2.2.1.3.1.). La traduction proprement dite est réalisée par un décodeur algorithmique qui combine le modèle de traduction et le modèle de langue construits au préalable (cf. sous-section 2.2.1.3.3.).

Afin de diminuer les erreurs spécifiques à une traduction mot-à-mot proposée par ces premières méthodes statistiques (par exemple des séquences plus ou moins compositionnelles non traduites, l'ordre incorrecte des mots cibles, etc.), se sont développés les systèmes à base de séquences (Och *et al.*, 1999 ; Zens *et al.*, 2002 ; Zhang *et al.*, 2003 ; Koehn *et al.*, 2003). Ces systèmes n'utilisent plus le mot dans le processus de traduction mais plutôt la séquence de mots, étant censée, par exemple, mieux gérer le ré-ordonnancement local des mots au niveau de la phrase cible.

[38] http://www.ibm.com/fr/fr/

Des travaux plus récents et notamment les systèmes factorisés (Koehn et Hoang, 2007 ; Birch *et al.*, 2007 ; Avramidis et Koehn, 2008 ; Tufiș *et al.*, 2008b ; Ceaușu et Tufiș, 2011) intègrent aussi, dans le processus de traduction, des informations linguistiques (morphosyntaxiques, syntaxiques, etc.) associées aux mots, afin d'améliorer encore les résultats des méthodes purement statistiques. Par l'utilisation de ces techniques mixtes, les systèmes factorisés donnent des résultats comparables aux méthodes expertes, pour différentes paires de langues, et s'avèrent moins coûteux en temps et en ressources humaines. *EuroMatrix*[39] (2006-2009) constitue un exemple de projet ayant comme objectif le développement des systèmes factorisés pour les langues de l'Union Européenne.

Comme pour les méthodes à base d'exemples, les systèmes statistiques ont l'avantage de s'améliorer facilement dès que de nouveaux corpus parallèles sont disponibles, en désavantageant cependant les langues moins dotées en ressources langagières. Ainsi, les performances de tels systèmes varient en fonction du volume des données utilisées.

Des systèmes commerciaux de traduction automatique statistiques ont été développés par des entreprises importantes comme *IBM*, *Language Weaver*[40], *Microsoft*[41] et *Google*[42]. Comme nous l'avons déjà mentionné auparavant, l'entreprise *Systran*[43], spécialiste de traduction automatique, a mis sur le marché un moteur hybride de traduction (Schwenk *et al.*, 2009) combinant des règles linguistiques et des techniques statistiques. Notons que pendant la campagne d'évaluation *Workshop on Statistical Machine Translation* (2009), ce système a obtenu la première place pour la traduction de l'anglais vers le français, en dépassant donc le système statistique pur *Google Translate*[44] qui, lui aussi, a fait ses preuves dans le domaine de la traduction automatique, en gagnant, par exemple, le titre de meilleur système de traduction lors de la campagne d'évaluation *NIST 2006*[45]. Cela montre la puissance des méthodes hybrides qui tirent profit des avantages des deux approches, empirique et experte. Parmi les exemples de projets plus récents exploitant des méthodes hybrides de traduction automatique, relevons *EuroMatrixPlus* (2009-2012)[46] pour les langues de l'Union Européenne et

[39] http://www.euromatrix.net/index.html
[40] http://www.languageweaver.com
[41] http://www.microsoft.com/fr-fr/default.aspx
[42] http://www.google.com/intl/fr_fr/about/company/
[43] http://www.systran.fr/systran
[44] http://translate.google.com/#en/fr/
[45] http://www.itl.nist.gov/iad/mig//tests/mt/2006/doc/mt06eval_official_results.html
[46] http://www.euromatrixplus.net/

PRESEMT (2010-2012)[47] pour le tchèque, l'anglais, l'allemand, le grec, le norvégien (langues sources) et l'anglais, l'allemand, l'italien (langues cibles).

Les systèmes statistiques purs ou factorisés intégrant le roumain, en combinaison avec l'anglais, sont les systèmes proposés par Marcu et Munteanu (2005), Ceaușu (2009), Ceaușu et Tufiș (2011), Dumitrescu *et al.* (2012), Tufiș et Dumitrescu (2012). Le roumain est intégré aussi par le système purement statistique *Google Translate*. Notons que la plupart des systèmes existants pour le roumain proposent des traductions de et vers l'anglais.

Nous venons de voir différentes approches et systèmes implémentés faisant partie, d'une part, du paradigme expert de la traduction automatique et, d'autre part, du paradigme empirique. Ces approches ont été présentées dans une perspective historique de la traduction automatique. Les méthodes expertes, de par l'utilisation des règles linguistiques complexes, donnent de bons résultats pour différentes paires de langues (p. ex. *Systran* utilisant le transfert jusqu'en 2009), mais elles s'avèrent très coûteuses en temps et en ressources humaines. Les systèmes hybrides combinant des règles linguistiques et des techniques statistiques (p. ex. *Systran* depuis 2009) donnent des résultats encore meilleurs qu'une approche purement linguistique, basée sur la méthode par transfert, ou qu'une approche exclusivement statistique. En outre, les systèmes statistiques factorisés, comme le projet *EuroMatrix*, peuvent donner des résultats comparables aux systèmes experts en utilisant moins de temps et de personnels (cf. section 2.2.).

Comme notre étude porte sur la paire de langues français - roumain, nous tournerons maintenant notre regard du côté de la France et de la Roumanie pour explorer le contexte historique de la recherche en traduction automatique dans ces pays (Léon, 2001, 2002 ; Iancu, 2007) (cf. sous-section suivante).

2.1.4. La traduction automatique en France et en Roumanie

En France, la recherche en traduction automatique a connu un certain retard dû en grande partie au retard en informatique. Afin de développer des machines de calculs françaises, le *Centre national de la recherche scientifique* (*CNRS*)[48] crée l'Institut Blaise-Pascal (*IBP*) en 1946. Une année plus tard, le *CNRS* initie le développement d'une machine numérique mais

[47] *Pattern REcognition-based Statistically Enhanced MT*: http://www.presemt.eu/
[48] http://www.cnrs.fr/

le projet s'arrête en 1952 à cause d'une erreur de conception. Ainsi, en 1955, le *CNRS* achète une première machine de provenance anglaise pour l'*IBP*. Ensuite, en 1959, la politique gaullienne soutenant la recherche impulse la dotation des universités en calculateurs électroniques et trois grands centres informatiques voient ainsi le jour à Paris, Grenoble et Toulouse. Cet état des choses va favoriser la création de centres de recherche en traduction automatique à Paris et à Grenoble. Ainsi, en décembre 1959, le *CNRS* crée au sein de l'*IBP* le *Centre d'étude pour la traduction automatique* (*CETA*) avec deux pôles : le *CETAP* à Paris, sous la direction d'Aimé Sestier, et le *CETAG* à Grenoble, sous la direction de Bernard Vauquois. Également, en avril 1959, est fondée l'*Association pour la traduction automatique et la linguistique appliquée* (*ATALA*)[49] présidée par Emile Delavenay ayant un « rôle moteur » (Léon, 2001 : 92) dans le développement des recherches en traduction automatique en France. Le personnel fondateur de l'*ATALA* comprend des mathématiciens, des ingénieurs, des traducteurs, des linguistes, des documentalistes. Cette association se voit plutôt comme un forum de discussion sur la traduction automatique, la documentation automatique et la linguistique appliquée, qu'un centre de recherche (Léon, 2002). Un deuxième centre pour la traduction automatique, le *Groupe de Traduction Automatique*, est créé par le *CNRS* en mai 1960 à l'Université de Nancy. Il existe de ce fait un décalage d'environ une dizaine d'années de la France par rapport aux États-Unis et à la Grande-Bretagne en ce qui concerne les recherches en traduction automatique.

Quant au *CETA*, celui-ci est fondé sur une convention entre le *CNRS*, la *Direction des études et fabrications d'armement* (*DEFA*) et le *Comité d'action scientifique de défense du centre d'exploitation scientifique et technique* (*CASDEN*) du ministère de la Défense (Léon, 2002). Le *CETA* a comme mission « l'étude et la conception d'une méthode pour la traduction automatique notamment du russe en français et l'étude de l'organisation générale d'une machine pour cette fin » (Léon, 2002)[50]. Ainsi, les objectifs concernent l'armée et le contre-espionnage, comme aux États-Unis, et l'enjeu de la traduction automatique, dans le contexte de la guerre froide, réside principalement dans la traduction des articles scientifiques et techniques russes.

[49] Actuellement, cette association est devenue *L'Association pour le traitement automatique des langues* : http://www.atala.org/.
[50] http://histoire-cnrs.revues.org/3461#ftn19

Les orientations de la traduction automatique en France ont été définies par le rapport intitulé « Comment doit être organisé à l'échelle française l'effort pour la traduction automatique ? », rédigé par A. Sestier en novembre 1959. Ce rapport propose comme méthode l'analyse de la langue source qui est le russe. Dans cette optique, tout développement nécessite une technologie linguistique en tant que fondement scientifique d'une traduction automatique de qualité raisonnable. Cette technologie linguistique doit inventorier et classer des faits linguistiques, syntaxiques, morphologiques et lexicaux sur la base d'un corpus de textes sources (Léon, 2002).

Toutefois, cette idée que le développement d'une technologie linguistique est suffisant pour mettre en place des dispositifs fiables de traduction automatique est abandonnée autant que le travail en linguistique fondamentale. Cette situation pousse progressivement Sestier à démissionner et le *CETAP* est dissout en octobre 1962. La démission de Sestier a été aussi influencée, entre autres, par le rapport de Bar-Hillel qui présentait négativement la plupart des résultats et des activités des centres de recherche en traduction automatique à travers le monde et particulièrement aux États-Unis (Léon, 2002).

Le deuxième groupe de recherche sur la traduction automatique en France, celui de Nancy, est créé par Schneider et Culioli en 1960. C'est le seul groupe qui est dirigé par des linguistes. Les recherches de ce groupe porteront sur l'anglais en tant que langue source et sur le français et l'espagnol en tant que langues cibles afin de traduire des ouvrages scientifiques.

Les centres de recherche sur la traduction automatique créés à cette époque comprennent très peu de linguistes. Maurice Gross ayant reçu une formation d'ingénieur et Yves Gentilhomme ayant une formation mathématique faisaient partie du *CETAP*. Après la dissolution de ce dernier en 1962, Gross et Gentilhomme vont se consacrer aux langages formels. Les linguistes de formation littéraire qui se sont investis dans des recherches concrètes de traduction automatique sont Bernard Pottier et Guy Bourquin. En revanche, les linguistes « institutionnellement établis » (Léon, 2001 : 93) s'impliquent vraiment dans le domaine. Certains d'entre eux sont membres fondateurs de l'*ATALA*, comme Marcel Cohen, David Cohen ou Antoine Culioli, tandis que d'autres appartiennent aux comités d'évaluation du *CETAP* et du *CETAG*, comme Émile Benveniste, Georges Gougenheim, Michel Lejeune, André Martinet, Bernard Quemada, Jean Train, etc. (Léon, 2001).

Rappelons que le rapport *ALPAC* (cf. sous-section 2.1.1.), paru en 1966 aux États-Unis, met fin aux financements des recherches en traduction automatique pratiquement partout dans le monde. Celui-ci révèle, entre autres, le manque de performance de la traduction automatique et son inutilité dans le contexte de l'époque. Toutefois, le *CNRS*, étant moins concerné par les enjeux économiques de la traduction automatique et intéressé par la linguistique appliquée, soutient encore le *CETA*. Celui-ci se voit recommander par le conseil scientifique de décembre 1966 de diversifier les domaines d'applications des méthodes proposées. Ainsi, la traduction automatique n'a plus de priorité, devant « apparaître comme un produit dérivé d'études linguistiques et logiques plus générales » (Léon, 2002)[51]. Soutenu par le *CNRS*, le *CETA* se développe. Pendant la seconde conférence internationale sur le traitement automatique des langues (août 1967), celui-ci présente sa première expérience de traduction automatique.

Cependant, en 1971, le *CETA* est moins prospère, perdant des moyens, ses chercheurs et son statut de laboratoire propre du *CNRS* mais sa reconversion se fait de manière progressive, ce qui lui donne une certaine stabilité. De ce fait, le *CETA*, devenu le *GETA*[52] dans la même année et sous la même direction de B. Vauquois, développe le système *Ariane-78* basé sur l'approche de transfert. Ce système est considéré comme un des plus performants de cette période. Néanmoins, il n'aboutira pas à une commercialisation. Le *GETA* a contribué aussi par des études au projet *Eurotra* (1977-1994) (cf. sous-section 2.1.1.2.1.2.). Dans ce projet, ont participé aussi l'Université de Nancy II et l'Université Paris VII (Hutchins et Somers, 1992).

Le *GETA* a aussi été dirigé par Christian Boitet et il a fusionné depuis 2007 avec le *GEOD*[53] en *GETALP*[54], se trouvant désormais sous la direction d'Hervé Blanchon. Les recherches sont menées encore, entre autres, sur la traduction automatique (*TA*) et sur la traduction assistée par ordinateur (*TAO*) (Boitet, 1993ab ; Boitet et Blanchon, 1994).

Quant à la traduction automatique en Roumanie, les premières recherches de grammaire effectuées dans ce but sont dues au mathématicien Grigore Moisil, considéré le père de l'informatique en Roumanie. Moisil a étudié spécialement le verbe du roumain (Ștefan et

[51] http://histoire-cnrs.revues.org/3461#tocto1n1
[52] *Groupe d'étude pour la traduction automatique*
[53] *Groupe d'étude sur l'oral et le dialogue*
[54] *Groupe d'étude pour la traduction automatique et le traitement automatisé des langues et de la parole*

Nicolau, 1981, *cit. in* Iancu, 2007). Il a initié les premières expériences de traduction automatique sur l'ordinateur *MECIPT* (*Maşină Electronică de Calcul a Institutului Politehnic Timişoara - Machine Électronique de Calcul de l'Institut Polytechnique de Timişoara*) ayant 50 itérations par seconde[55]. C'est Erika Nistor qui a élaboré des algorithmes de traduction automatique de l'anglais vers le roumain et en 1959 ont vu le jour les premières traductions à l'*Institut Polytechnique de Timişoara* (Ştefan et Nicolau, 1981, *cit. in* Iancu, 2007). Ainsi, la première phrase anglaise traduite en roumain *« You explain the development of science and we help describe the examples* » est devenue célèbre en étant publiée dans tous les journaux de l'époque[56]. D'autres études ayant comme finalité la traduction automatique des textes ont été réalisées par Minerva Bocşa qui a conçu des programmes pour identifier les caractéristiques des textes en roumain, russe et allemand. Ces études sont basées sur des indices comme la fréquence des lettres, la longueur moyenne des mots et des phrases, la fréquence des mots, etc. (Ştefan et Nicolau, 1981, *cit. in* Iancu, 2007).

Des recherches effectives plus récentes en traduction automatique - statistique (factorisée) et à base d'exemples - concernant le roumain en combinaison avec l'anglais (Tufiş *et al.*, 2008b ; Irimia, 2008 ; Ceauşu, 2009 ; Ceauşu et Tufiş, 2011 ; Dumitrescu *et al.*, 2012 ; Tufiş et Dumitrescu, 2012 ; Dumitrescu *et al.*, 2013 ; Tufiş *et al.*, 2013ab ; Boroş *et al*., 2013) sont menées sous la direction de l'académicien Dan Tufiş à l'*Institut de Recherches en Intelligence Artificielle « Mihai Drăgănescu »* (*ICIA*)[57]. L'*ICIA* a été créé en 1994 dans le cadre de l'Académie Roumaine de Bucarest comme un centre de compétence dans le domaine de l'Intelligence Artificielle.

Dans cette section, nous avons présenté les différentes approches de la traduction automatique qu'elle soit experte ou empirique, dans leurs contextes historiques, afin de faire un inventaire des approches existantes dans le domaine. Comme notre projet de thèse porte sur la paire de langues français - roumain, des aspects historiques de la traduction automatique en France et en Roumanie ont été aussi abordés. Dans la section suivante, nous nous pencherons de façon détaillée sur l'approche statistique de la traduction automatique car c'est celle-ci qui fait l'objet de notre étude.

[55] http://www.unibuc.ro/prof/vlada_m/Did_you_know.php
[56] http://www.unibuc.ro/prof/vlada_m/Did_you_know.php
[57] http://www.racai.ro/en/about-us/about-racai/

2.2. La traduction automatique statistique

La traduction automatique statistique utilise la théorie mathématique de distribution et d'estimation probabiliste mise en œuvre par Frederick Jelinek au *IBM T. J. Watson Research Center*. Un rôle majeur dans le développement des systèmes de traduction automatique statistique ont aussi les travaux de Brown *et al.* (1990) et Brown *et al.* (1993), qui décrivent un modèle lexical.

Les systèmes de traduction automatique statistique fonctionnent par l'apprentissage d'un modèle probabiliste de traduction à partir de corpus parallèles bilingues alignés et d'un modèle probabiliste de la langue cible, à partir d'un corpus monolingue. Ces notions seront expliquées plus loin dans la section. La maximisation du modèle de traduction et du modèle de langue donne la meilleure traduction de la phrase source. Les probabilités des modèles de langue et de traduction sont générées indépendamment et sont représentées par des ensembles de tables. Par rapport au modèle traditionnel (le système *expert*), le modèle probabiliste de traduction s'améliore de manière automatique en fonction de nouvelles données d'entraînement.

Plusieurs approches de traduction automatique statistique ont utilisé la méthode d'alignement des corpus parallèles au niveau lexical (Brown *et al.*, 1990 ; Brown *et al.*, 1993 ; Vogel *et al.*, 1996 ; Tillmann *et al.*, 1997 ; Niessen *et al.*, 1998). L'utilisation d'un tel alignement pose des problèmes importants à la qualité de la traduction, parce qu'il n'est pas capable, par exemple, de prendre en compte les dépendances entre les mots ou les groupes de mots. Pour palier ces problèmes, les travaux ultérieurs ont développé des modèles à base de séquences de mots (Och *et al.*, 1999 ; Koehn *et al.*, 2003), qui obtiennent des performances significativement meilleures par rapport aux modèles lexicaux. Par l'utilisation des séquences de mots, les systèmes de traduction automatique statistique prennent en compte certaines contraintes locales sur l'ordre des mots.

Toutefois, les modèles courants de traduction automatique statistique à base de séquences sont limités à la mise en correspondance des groupes de mots sans présentation d'informations linguistiques explicites (morphologiques, syntaxiques ou sémantiques). De ce fait, les modèles hybrides de traduction automatique statistique (dénommés des modèles factorisés) utilisent de plus en plus d'informations linguistiques pendant le processus de traduction, en tentant par cela d'améliorer les résultats des approches statistiques à base de séquences pures.

En effet, les modèles factorisés (Koehn et Hoang, 2007 ; Birch *et al.*, 2007 ; Avramidis et Koehn, 2008 ; Tufiș *et al.*, 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011) permettent d'exploiter les différents niveaux de prétraitement linguistique des corpus : lemmatisation, étiquetage, *chunking* (analyse syntaxique partielle du corpus), désambiguïsation des sens, etc. Dans l'optique de ces modèles, le mot représente un vecteur de facteurs linguistiques (le lemme, la partie du discours, les propriétés morphosyntaxiques ou syntaxiques, etc.) et ces facteurs sont pris en compte pendant le processus de traduction. Ces modèles se concentrent sur le développement des corpus d'apprentissage de taille et de qualité significatives afin d'améliorer les résultats des approches purement statistiques. La qualité des corpus d'apprentissage peut être mesurée en évaluant la qualité des résultats du prétraitement du corpus (p. ex. étiquetage, lemmatisation) et de l'alignement propositionnel ou lexical.

Dans les sous-sections suivantes, seront présentés quelques fondements mathématiques (Brown *et al.*, 1990 ; Brown *et al.*, 1993 ; Knight, 1999[58]) exploités par les systèmes de traduction automatique statistique, afin de comprendre les notions de base de tels systèmes. Ensuite, différentes approches utilisées dans le domaine de la traduction automatique statistique y seront également décrites.

2.2.1. Probabilités fondamentales de la traduction automatique statistique

En termes mathématiques, la probabilité qu'un événement souhaité d'un ensemble d'événements possibles se réalise, est donnée par le calcul du rapport entre le nombre d'événements souhaitables et d'événements possibles. Par exemple, pour calculer la probabilité que l'événement **n** de l'ensemble d'événements possibles **N** se réalise, la formule suivante est utilisée :

$$P(n) = \frac{nombre\_cas\_favorables}{nombre\_cas\_possibles} = \frac{n}{N}$$

Le nombre d'événements favorables est toujours inférieur ou égal au nombre d'événements possibles. Par conséquent, la probabilité P(**n**) a des valeurs comprises entre 0 et 1 inclusivement.

[58] http://ccl.pku.edu.cn/doubtfire/NLP/Parsing/Unification_based_parsing/Kevin%20Knight.htm (*A Statistical MT Tutorial Workbook* - Knight, 30 avril 1999)

2.2.1.1. Principe de base de la traduction automatique statistique

En traduction automatique statistique, le principe de base est de considérer qu'une phrase **e** de la langue cible est une possible traduction d'une phrase **f** de la langue source. Évidemment, certaines traductions ont plus de chances que d'autres. La formule mathématique qui formule ce principe figure ci-dessous (Knight, 1999)[59] :

$$P(f,e) = P(f) \times P(e \mid f)$$

où :

- P(**f**) : *probabilité a priori* qui constitue la chance que l'événement **f** ait lieu. En traduction automatique statistique, cela veut dire que pour une phrase **f** de la langue source, P(**f**) constitue la chance que cette phrase **f** s'est produite à un moment donné.

- P(**e**|**f**) : *probabilité conditionnelle* qui constitue la chance que l'événement **e** ait lieu à condition que l'événement **f** ait eu lieu. Dans le cas de la traduction automatique statistique, cela signifie que si **f** est une phrase produite de la langue source, alors P(**e**|**f**) est la chance que la phrase **e** de la langue cible soit la traduction de la phrase **f**.

- P(**f,e**) : *probabilité conjointe* qui est le produit des probabilités P(**f**) et P(**e**|**f**) et constitue la chance que les événements **f** et **e** aient lieu en même temps. En traduction automatique statistique, P(**f,e**) est la probabilité que la phrase **e** soit la traduction de la phrase **f**.

Si **f** et **e** ne s'influencent pas réciproquement, la probabilité P(**f,e**) est le produit des probabilités de chaque événement et est donnée par la formule :

$$P(f,e) = P(f) \times P(e)$$

Toutes ces probabilités ont des valeurs comprises entre 0 et 1 inclusivement.

Afin de comprendre les notations des formules mathématiques utilisées en traduction automatique statistique, les notions de *somme* et *produit* des nombres entiers seront décrites dans la sous-section suivante.

[59] http://ccl.pku.edu.cn/doubtfire/NLP/Parsing/Unification_based_parsing/Kevin%20Knight.htm

2.2.1.2. Somme et produit

Pour représenter symboliquement la somme (a) et le produit (b) des nombres entiers de 1 à n, nous pouvons écrire :

a) $$\sum_{i=1}^{n} i = 1 + 2 + 3 + ... + n$$

b) $$\prod_{i=1}^{n} i = 1 \times 2 \times 3 \times ... \times n$$

Dans le cas où il existe un facteur commun *k* des termes de la somme des nombres entiers de 1 à n, nous pouvons diviser chaque terme de la somme par le facteur commun *k* et multiplier la nouvelle somme par *k*. Alors, nous pouvons écrire :

$$\sum_{i=1}^{n} i \times k = k + 2k + 3k + ...nk = k \sum_{i=1}^{n} i$$

Nous donnerons ci-dessous deux formules utiles en ce qui concerne les probabilités utilisées en traduction automatique statistique. Ainsi, la somme des probabilités *a priori* P(**f**) et des probabilités *conditionnelles* P(**e**|**f**) est égale à 1 :

$$\sum_{f} P(f) = 1$$

$$\sum_{e} P(e \mid f) = 1$$

À partir de ces considérations, nous allons voir dans les sous-sections suivantes leur application pour un système lexical de traduction automatique statistique.

2.2.1.3. Systèmes lexicaux de traduction automatique statistique

Le problème de la traduction automatique statistique se pose ainsi (Brown *et al.*, 1990 ; Brown *et al.*, 1993) : pour une phrase donnée ***T*** en langue cible, on cherche la phrase ***S*** de la langue source pour laquelle le système de traduction a produit la phrase ***T***. La chance d'obtenir des erreurs devient minimum en choisissant la phrase ***S*** la plus probable pour la

phrase ***T***. De ce fait, on cherche la phrase ***S*** maximisant la probabilité ***P(S|T)*** qui peut être obtenue à partir du théorème de Bayes donné ci-dessous :

$$P(S \mid T) \times P(T) = P(S) \times P(T \mid S)$$

À partir de cette égalité, nous pouvons obtenir la probabilité ***P(S|T)*** :

$$P(S \mid T) = \frac{P(S) \times P(T \mid S)}{P(T)}$$

Comme le dénominateur de cette équation (la probabilité ***P(T)***) ne dépend pas de la phrase source ***S,*** nous pouvons chercher la phrase source ***S*** qui maximise le produit au numérateur $P(S) \times P(T \mid S)$ et nous obtenons l'équation suivante :

$$\arg\max P(S \mid T) = \arg\max P(S) \times P(T \mid S)$$

où la distribution de probabilité ***P(S)*** représente le modèle de langue de la phrase source ***S*** et la distribution de probabilité ***P(T|S)*** représente le modèle de traduction de la phrase cible ***T*** étant donné la phrase source ***S***.

Ainsi, la meilleure traduction d'une phrase source se calcule selon la logique suivante : pour une phrase source ***S*** donnée, la tâche du système est de trouver la phrase cible ***T*** qui est la plus probable parmi toutes les traductions possibles. Cette phrase cible ***T*** maximise la probabilité ***P(T|S)***, par la maximisation du produit $P(T) \times P(S \mid T)$ et nous obtenons l'équation fondamentale de la traduction automatique statistique :

$$\arg\max P(T \mid S) = \arg\max P(T) \times P(S \mid T)$$

En effet, un système de traduction automatique statistique demande trois méthodes :

- une méthode pour calculer les probabilités du modèle de langue de la langue cible ***P(T)*** ; Ce modèle fournit un ensemble de phrases cibles munies de leurs probabilités et nécessaires pour tester une traduction potentielle.

- une méthode pour calculer les probabilités du modèle de traduction ***P(S|T)*** ; Ce modèle fournit une table de traduction contenant des paires bilingues de mots / séquences source - cible, munies aussi de leurs probabilités, qui sont extraites à

partir d'un corpus parallèle aligné au niveau lexical. Celui-ci teste si une phrase cible donnée est une traduction possible.

- une méthode de décodage algorithmique utilisant les modèles construits auparavant et fournissant la traduction la plus probable parmi un ensemble de traductions possibles : *argmax.*

Dans les sous-sections suivantes, seront expliquées les notions de modèle de langue, modèle de traduction et décodage, telles que décrites par Brown *et al.* (1990), dans le cadre d'un système lexical de traduction automatique statistique.

2.2.1.3.1. Modèle de langue

Un modèle de langue est construit par l'utilisation d'un modèle appelé *n-grammes*.

À partir d'une chaîne de mots $s_1 s_2 ... s_n$ nous pouvons obtenir la probabilité d'un mot donné s_n de la chaîne par le calcul du produit des probabilités des mots qui le précèdent dans la chaîne, selon la formule :

$$P(s_1 s_2 ... s_n) = P(s_1) \times P(s_2 \mid s_1) ... P(s_n \mid s_1 s_2 ... s_{n-1})$$

Ainsi, à tout point dans la phrase, nous devons connaître la probabilité d'un mot s_j étant donné son historique $s_1 s_2 ... s_{j-1}$ .Un problème qui s'y pose est la longueur de l'historique d'un mot donné. Pour résoudre cet inconvénient, il est nécessaire de considérer un historique de taille réduite et fixe, ce qu'on appelle modèles *n-grammes.* Par ailleurs, comme il existe une multitude d'historiques pour un mot donné, il n'est pas possible de traiter séparément chacun de ces paramètres de probabilité. Pour en réduire leur nombre, la solution est de mettre chaque historique dans une classe d'équivalence. Dans un modèle de langue *n-grammes*, deux historiques sont équivalents si les n-1 mots sont identiques. En effet, deux historiques sont équivalents s'ils finissent par le même mot dans le cas d'un modèle bi-grammes, par les mêmes deux mots dans le cas d'un modèle tri-grammes, etc. L'opération suivante est de permettre à la probabilité d'un mot de dépendre de l'historique seulement par la classe d'équivalence.

Les modèles *n-grammes* ont été très utiles dans la reconnaissance de la parole et ont l'avantage d'être facile à construire et à utiliser. Pour voir la puissance d'un modèle tri-grammes (construit initialement pour un système de reconnaissance de la parole) Brown *et al.* (1990) utilisent une technique appelée *sac de traduction* (*bag translation*) de l'anglais vers l'anglais. Cette technique suppose plusieurs opérations : couper une phrase en mots, placer les mots dans le *sac de traduction* et essayer de récupérer la phrase initiale compte tenu du *sac de traduction.* Pour classer les différents arrangements des mots du *sac de traduction*, Brown *et al.* (1990) utilisent le modèle *n-grammes*. Ainsi, un arrangement S est mieux qu'un arrangement S', si la probabilité P(S) est plus importante que la probabilité P(S'). Parmi un échantillon de 100 phrases, les auteurs prennent en compte seulement les 38 phrases comprenant chacune moins de 11 mots. La restriction de la longueur de la phrase est due au fait que le nombre des arrangements possibles augmente exponentiellement avec la longueur de la phrase, mais aussi la probabilité des erreurs. Les résultats obtenus par Brown *et al.* (1990) sont divisés ainsi : des phrases correctes (24 (63%)), des phrases qui ne sont pas des reproductions fidèles des phrases initiales, mais gardent le même sens (8 (21%)) et des phrases incorrectes (6 (16%)). Le pourcentage des phrases considérées correctes (84%) montre qu'un modèle *n-grammes* donne de bons résultats pour construire des modèles de langue.

La deuxième méthode demandée par un système de traduction automatique statistique constitue le modèle de traduction qui sera présenté dans la sous-section suivante.

2.2.1.3.2. Modèle de traduction

Pour construire un modèle de traduction, Brown *et al.* (1990) partent de l'idée que la traduction d'une phrase de la langue source (anglais) vers une phrase de la langue cible (français) est générée à partir de la phrase en anglais mot par mot. Par exemple, pour la paire de phrases (*Jean aime Marie* | *John loves Mary*), le mot *John* produit le mot *Jean*, *loves* produit *aime* et *Mary* produit *Marie*. Ainsi, un mot de la langue source est aligné avec un mot qu'il produit de la langue cible. Par conséquent, ce modèle de traduction doit utiliser un corpus parallèle aligné au niveau lexical, c'est-à-dire aligné mot-à-mot. Un exemple d'alignement mot-à-mot pour un couple de phrases anglais - français (*he often eats an apple…* vs. *il mange souvent une pomme…*) se trouve dans la Figure 6 ci-dessous :

Figure 6. Alignement mot-à-mot pour la paire de langues anglais - français

Ce modèle de traduction représente un modèle lexical de traduction. Dans un tel modèle, le processus de traduction est décomposé en petites étapes de calcul de probabilités de *fertilité*, de *traduction* et de *distorsion*, qui seront expliquées plus loin.

Dans un alignement lexical, il existe des cas où un mot de la langue source est aligné avec plusieurs mots de la langue cible. Ce problème est résolu par le calcul de la mesure appelée la *fertilité* (*fertility*) des mots de la langue source. La *fertilité* représente le nombre de mots de la phrase cible produits par un mot de la langue source. Cette mesure résout aussi certains cas où les mots n'ont pas d'équivalents d'une langue à l'autre.

Dans le processus d'alignement, les mots près du début ou de la fin de la phrase cible ont tendance à s'aligner avec les mots près du début ou de la fin de la phrase source. Mais ce cas de figure n'est pas toujours possible. En effet, un mot de la phrase cible peut apparaître très loin du mot de la phrase source qui l'a produit. Cet effet est appelé *distorsion*. Par exemple, les distorsions permettent aux adjectifs de précéder les noms en anglais, mais de suivre les noms en français.

Pour expliquer le modèle de traduction Brown *et al.* (1990) prennent comme exemple la paire de phrases alignées *Le chien est battu par Jean / John does beat the dog* dont la notation est la suivante :

(*Le chien est battu par Jean*|*John* (6) *does beat* (3,4) *the* (1) *dog* (2)),

où pour chaque mot de la phrase source sont données entre parenthèses les listes des positions des mots de la phrase cible avec lesquels les mots de la phrase source sont alignés.

Ainsi, dans cet exemple, le mot *John* produit le mot *Jean*, *does* ne produit aucun mot, *beat* produit *est battu*, *the* produit *Le*, *dog* produit *chien* et *par* n'est produit par aucun mot de l'anglais (dans le modèle de traduction, c'est un mot considéré <*null*> qui va produire *par*).

La probabilité de cet alignement est calculée par le produit des probabilités de *fertilité* des mots de la langue source et des probabilités de *traduction* :

P(fertilité=1|*John*) × P(*Jean*|*John*) ×

P(fertilité=0|*does*) ×

P(fertilité=2|*beat*) × P(*est*|*beat*)P(*battu*|*beat*) ×

P(fertilité=1|*the*) × P(*Le*|*the*) ×

P(fertilité=1|*dog*) × P(*chien*|*dog*) ×

P(fertilité=1<*null*>) × P(*par*<*null*>).

À partir du calcul ci-dessus, on peut observer, par exemple, que si la fertilité d'un mot source est égale à 0 (P(fertilité=0|*does*)) le mot n'est pas traduit, ce qui résout les cas où les mots n'ont pas d'équivalents d'une langue à l'autre. Nous voyons donc la nécessité de calculer les probabilités de *fertilité* des mots de la langue source.

Pour calculer le facteur de *distorsion* des probabilités, Brown *et al.* (1990) considèrent que la position d'un mot cible dépend seulement de la longueur de la phrase cible et de la position du mot source. Ainsi, la probabilité de *distorsion* a la forme suivante :

P (*i*|*j*, *l*),

où : *i* est la position du mot cible dans la phrase cible ;

j est la position du mot source dans la phrase source ;

l est la longueur de la phrase cible.

En conclusion, les paramètres du modèle de traduction de Brown *et al.* (1990) sont les suivants : un set de probabilités de fertilité *P(n|e)* pour chaque mot source *e* et pour chaque fertilité *n* de 0 jusqu'à une limite considérée (25 dans ce cas), un set de probabilités de traduction P (f|e) et un set de probabilités de distorsion P (i|j, l), avec *i*, *j* et *l* compris entre 1 et 25.

Ces notions fondamentales ont mené au développement des modèles de traduction *IBM* (Brown *et al.*, 1993) qui sont devenus la référence des travaux ultérieurs en traduction automatique statistique. Ces modèles seront présentés plus loin, dans le chapitre 3 de ce travail, car ils ont été implémentés dans l'aligneur lexical *GIZA++* (Och et Ney, 2000, 2003), outil que nous avons utilisé pour notre système de traduction automatique.

À partir du modèle de langue et de traduction préalablement construits, le système de traduction automatique statistique procède à l'étape de décodage (Brown *et al.*, 1990) présentée dans la sous-section suivante.

2.2.1.3.3. Décodage

La recherche automatique de la phrase source *S* qui maximise le produit *P(S)P(T|S)*, est une tâche difficile à effectuer à cause du nombre significatif des phrases à essayer. Pour simplifier cette tâche, Brown *et al.* (1990) utilisent une variante de la méthode du modèle de reconnaissance de la parole (Bahl *et al.*, 1983). Ainsi, Brown *et al.* (1990) maintiennent une liste partielle d'hypothèses d'alignement. Cette liste contient, au début, seulement une entrée correspondant à l'hypothèse que la phrase cible est générée à partir d'une séquence de mots sources inconnue. Une telle entrée est notée comme dans cet exemple (*Jean aime Marie*|*), où l'astérisque tient la place d'une séquence de mots source inconnue. La recherche automatique procède par itérations. Chaque itération étend les entrées les plus prometteuses de la liste. Une entrée est étendue par l'addition d'un ou de plusieurs mots. Par exemple, une entrée initiale (*Jean aime Marie*|*) peut être étendue avec une ou plusieurs des entrées suivantes (Brown *et al.*, 1990) :

(*Jean aime Marie*|*John* (1)*),

(*Jean aime Marie*|**loves* (2)*),

(*Jean aime Marie*|**Mary* (3)),

(*Jean aime Marie*|*Jeans* (1)*).

La recherche prend fin quand il existe un alignement complet dans la liste, qui est nettement plus prometteur que tout autre alignement incomplet (Brown *et al.*, 1990).

Les paramètres du modèle de langue et de traduction sont estimés automatiquement à partir d'une base de données significative des paires de phrases source - cible, par l'utilisation d'un algorithme statistique qui optimise l'adéquation entre les modèles et les données.

Toutefois, il existe parfois, parmi les résultats du décodage, des phrases sources obtenues qui ne sont pas des traductions acceptables. Ainsi, quand il existe des erreurs de recherche de la traduction la plus probable, les problèmes du système sont liés aux modèles de langue et de traduction construits. Pour améliorer les résultats, de nouveaux paramètres doivent être implémentés aux modèles de langue et de traduction construits.

En conclusion, les modèles lexicaux utilisent des corpus alignés au niveau lexical et constituent la base des systèmes développés ultérieurement dans le domaine de la traduction automatique statistique. Toutefois, les modèles lexicaux n'intègrent pas d'informations linguistiques, le mot étant considéré comme une simple occurrence (token[60]). Par conséquent, dans les premiers systèmes d'alignement tels que IBM 1 (cf. chapitre 3, sous-section 3.1.2.), les mots sont complètement indépendants les uns des autres et la traduction se fait au niveau des formes des mots sans aucune motivation linguistique. De ce fait, l'utilisation des alignements lexicaux pose des problèmes majeurs à la qualité de la traduction, car ces systèmes s'avèrent incapables de détecter les dépendances entre les mots et les groupes de mots. Ainsi, le système lexical gère difficilement l'ordre correcte des mots dans la phrase et la traduction des séquences plus ou moins compositionnelles, telles que : noms composés, collocations, etc. Le modèle lexical de traduction automatique statistique rencontre aussi des problèmes importants dans le cas du choix lexical concernant les ambiguïtés liées à la langue source.

Vu les inconvénients des modèles lexicaux, nous saisissons le besoin des travaux ultérieurs de développer des systèmes qui orientent le processus de traduction au niveau des groupes de mots (des *chunks*), afin de prendre en compte certaines contraintes locales sur l'ordre des mots. Pour ces systèmes, l'unité de traduction n'est donc plus le mot mais plutôt la séquence de mots. Les expériences ont montré que les modèles à base de séquences obtiennent des résultats nettement plus performants par rapport aux modèles lexicaux.

[60] Un token est une unité lexicale ou un signe de ponctuation.

Ainsi, les approches ultérieures ont mis en place des modèles de traduction à base de séquences de mots (Och *et al.*, 1999; Koehn *et al.*, 2003), afin de réduire le nombre de problèmes rencontrés par les modèles lexicaux. Les modèles à base de séquences utilisent le contexte local des mots. De ce fait, ces modèles sont capables de détecter des dépendances entre les mots et les groupes de mots et de traduire des séquences plus ou moins compositionnelles en bloc. Ces systèmes résolvent aussi des problèmes du choix lexical face aux ambiguïtés provenant de la langue source (p. ex. le mot *ring* anglais traduit par *plaque*, *anneau* ou *sonner* en français, mais la séquence *electric ring*, traduite par *plaque électrique* et non pas par *anneau électrique* (Lavecchia, 2010)). Toutefois, l'inconvénient des systèmes à base de séquences est le fait que ces systèmes traitent seulement des séquences constituées de mots contigus dans la phrase. Par conséquent, ces modèles ne prennent pas en compte les collocations discontinues dans la phrase. De plus, comme l'alignement des corpus parallèles se fait par une simple mise en correspondances des groupes, sans aucune motivation linguistique, les traductions fournies sont parfois incohérentes.

Vu ces considérations, les modèles de traduction automatique statistique à base de séquences de mots seront présentés dans la section suivante.

2.2.2. Systèmes de traduction automatique statistique à base de séquences

Plusieurs travaux ont contribué à l'amélioration des résultats des systèmes de traduction automatique statistique par l'utilisation des modèles à base de séquences de mots (Och *et al.*, 1999 ; Yamada et Knight, 2001 ; Marcu et Wong, 2002 ; Koehn *et al.*, 2003). Ces systèmes utilisent des corpus alignés au niveau des mots et des séquences de mots. Un exemple d'alignement à base de séquences est donné dans la Figure 7 ci-dessous pour un couple de phrases anglais - français (*he often eats a yellow apple in the kitchen…* vs. *il mange souvent une pomme jaune dans la cuisine…*).

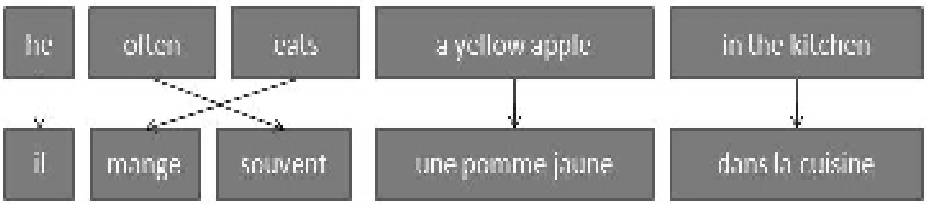


Figure 7. Alignement au niveau des séquences pour la paire de langues anglais - français

L'idée fondamentale de ces systèmes est de mettre en œuvre deux types d'alignement : un alignement lexical au niveau des séquences et un autre à l'intérieur de ces séquences (Och *et al.*, 1999).

Le système de traduction automatique statistique à base de séquences (Koehn *et al.*, 2003) fonctionne selon le même principe que celui décrit dans la sous-section 2.2.1.3., c'est-à-dire par la maximisation du modèle de langue et de traduction préalablement construits.

Le processus de traduction automatique à base de séquences se divise en trois étapes :

- Tout d'abord, une phrase donnée est divisée en séquences de mots ;
- Ensuite, chaque séquence est traduite à partir des tables d'équivalents de traduction ;
- Finalement, les séquences sont réordonnées en fonction de l'ordre des mots de la langue cible.

Le schéma du processus de traduction dans un système de traduction automatique statistique à base de séquences est présenté dans la Figure 8 ci-dessous. L'exemple donné dans cette figure concerne le couple de phrases anglais - français (*he often eats a yellow apple in the kitchen…* vs. *il mange souvent une pomme jaune dans la cuisine…*).

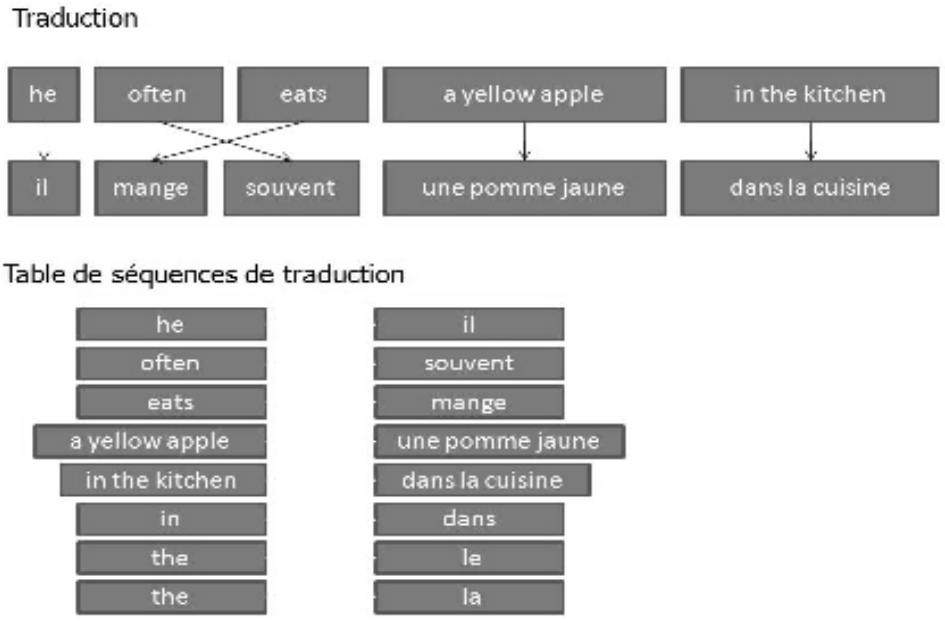


Figure 8. Schéma du processus de traduction dans un système de traduction automatique statistique à base de séquences (exemple pour les langues anglais - français)

Pendant l'étape de décodage, une phrase source **f** est divisée en une séquence de *I* segments $\overline{f}_1^{\mathrm{I}}$.

Ensuite, chaque segment source $\overline{f}_{\mathrm{i}}$ de l'ensemble $\overline{f}_1^{\mathrm{I}}$ est traduit dans un segment cible $\overline{e}_{\mathrm{i}}$. Les segments cibles peuvent être réordonnés. La traduction des segments est modelée par une distribution des probabilités conditionnelles, notée $\phi\left(\overline{f}_i \mid \overline{e}_i\right)$.

Le ré-ordonnancement des segments de la phrase cible est modelé par une distribution relative des probabilités de distorsion, notée $d\left(a_i - b_{i-1}\right)$,

où :

- a_i est la position de départ du segment source traduit dans le $i^{ème}$ segment cible ;
- b_{i-1} est la position de fin du segment source traduit dans le $(i\text{-}1)^{ème}$ segment cible.

La distribution de la probabilité de distorsion *d(.)* est calculée à l'aide d'un modèle de probabilité jointe qui sera défini plus loin. Un autre modèle de distorsion plus simple peut être utilisé alternativement :

$d\left(a_i - b_{i-1}\right) = \alpha^{|a_i - b_{i-1} - 1|}$, où le paramètre α a une valeur appropriée.

Pour optimiser la performance du système, Koehn *et al.* (2003) introduisent un facteur ω pour chaque mot cible généré, en plus du modèle de langue trigramme noté P_{LM}, destiné à calibrer la longueur des sorties. La valeur de ce facteur est habituellement supérieure à 1. Celui-ci pénalise les sorties plus longues.

Ainsi, dans le modèle de Koehn *et al.* (2003), la meilleure traduction e_{best} pour une phrase source **f** donnée est obtenue par la maximisation du produit des probabilités de traduction et de la probabilité que la phrase fasse partie du modèle de la langue cible, incluant le facteur additionnel ω à une puissance égale à la longueur des segments de la phrase cible :

$$e_{best} = \arg\max_e P(e \mid f)$$

$$= \arg\max_e P(f \mid e) P_{LM}(e) \omega^{length(e)}$$

Le modèle de traduction *p(f|e)* est le produit des distributions de probabilités conditionnelles $\phi(\overline{f_i} \mid \overline{e_i})$ et de probabilités de distorsion $d(a_i - b_{i-1})$ calculées pour chaque segment de *i=1* jusqu'à *I* :

$$p(\overline{f}_1^I \mid \overline{e}_1^I) = \prod_{i=1}^{I} \phi(\overline{f_i} \mid \overline{e_i}) d(a_i - b_{i-1})$$

La fonction ϕ est estimée par le calcul de la fréquence relative :

$$\phi(\overline{f} \mid \overline{e}) = \frac{count(\overline{f}, \overline{e})}{\sum_{\overline{f}} count(\overline{f}, \overline{e})}$$

Une manière de valider la qualité d'une paire de séquences de traduction est de voir dans quelle mesure les mots composants sont traduits l'un par l'autre (Koehn *et al.*, 2003).

Ainsi, il est nécessaire de calculer la distribution des probabilités de traduction lexicale $w(f \mid e)$, qui est donnée par la fréquence relative :

$$w(f \mid e) = \frac{count(f, e)}{\sum_{f'} count(f', e)}$$

Un token spécial de la langue cible *NULL* est rajouté à chaque phrase cible et aligné avec chaque mot non-aligné de la langue source.

À partir d'une paire de segments $\overline{f}$, $\overline{e}$ et d'un alignement au niveau des mots *a* entre les positions des mots de la langue source *i* de 1 jusqu'à *n* et les positions des mots de la langue cible *j* de 0 jusqu'à *m*, il est possible d'obtenir le poids lexical p_w par le biais de la formule :

$$p_w(\overline{f} \mid \overline{e}, a) = \prod_{i=1}^{n} \frac{1}{|\{j \mid (i, j) \in a\}|} \sum_{\forall (i,j) \in a} w(f_i \mid e_j)$$

Dans le cas où il existe plusieurs alignements *a* pour une paire de segments $(\overline{f}, \overline{e})$, l'alignement ayant le poids lexical le plus élevé est pris en compte :

$$p_w\left(\overline{f} \mid \overline{\mathrm{e}}\right) = \max_a \; p_w\left(\overline{f} \mid \overline{\mathrm{e}}, a\right)$$

Le poids lexical est utilisé comme un facteur additionnel. Ainsi, le modèle $p(f|e)$ devient :

$$p\left(\overline{f}_1^I \mid \overline{\mathrm{e}}_1^I\right) = \prod_{i=1}^{I} \phi\left(\overline{f_i} \mid \overline{e_i}\right) d\left(a_i - b_{i-1}\right) p_w\left(\overline{f_i} \mid \overline{e_i}, a\right)^{\lambda},$$

où le paramètre λ définit la force du poids lexical, ayant une valeur d'environ 0.25.

Dans toutes les expériences effectuées, Koehn *et al.* (2003) utilisent les mêmes données d'apprentissage, le modèle de langue trigrammes (Seymore et Rosenfeld, 1997) et un décodeur.

Koehn *et al.* (2003) développent un décodeur pour comparer différents modèles de traduction à base de séquences. Ce décodeur utilise un algorithme de recherche en faisceau (*beam search*). Ainsi, la phrase cible est obtenue par des traductions partielles appelées des *hypothèses*, générées de gauche à droite.

Nous allons expliquer cet algorithme tel que utilisé par Koehn *et al.* (2003). Initialement, l'algorithme démarre par une hypothèse nulle. Les nouvelles hypothèses sont étendues à partir des hypothèses existantes par la traduction d'un segment source. Ainsi, un segment de mots sources non-traduits et une possible traduction en langue cible sont sélectés. Le segment cible est attaché aux séquences de sortie déjà existantes. Les mots sources sont considérés traduits et la probabilité de l'hypothèse de traduction est actualisée. La probabilité la plus élevée des hypothèses de traduction finales représente la sortie de la recherche.

Les hypothèses de traduction sont stockées dans des piles. Une pile s_m contient toutes les hypothèses de traduction de *m* mots sources. Pour chaque pile, seulement un faisceau des meilleures hypothèses est gardé.

À l'aide du décodeur développé, Koehn *et al.* (2003) comparent la performance de trois méthodes pour construire des tables de traduction à base de séquences. Koehn *et al.* (2003)

montrent que l'extraction heuristique basée sur l'alignement lexical donne de meilleures performances par rapport à d'autres méthodes à base de séquences telles que l'utilisation de la syntaxe (Yamada et Knight, 2001) ou de la probabilité jointe (Marcu et Wong, 2002). Ces trois méthodes, ainsi qu'une étude comparative de leurs résultats (Koehn *et al.*, 2003), seront présentées dans les sous-sections suivantes.

2.2.2.1. Modèle utilisant des alignements lexicaux

La première méthode prise en compte par Koehn *et al.* (2003) apprend des alignements au niveau des séquences à partir des corpus alignés au niveau lexical (Och et Ney, 2000) par le biais de l'outil *GIZA++* (Och et Ney, 2000, 2003)[61] implémentant les modèles *IBM* (Brown *et al.*, 1993). Toutes les paires de séquences alignées qui sont conformes avec l'alignement lexical sont recueillies. En effet, les mots d'une paire de séquences collectée doivent être alignés seulement à l'intérieur de la paire et non pas avec d'autres mots extérieurs à la paire considérée (Och *et al.*, 1999). À partir des paires de séquences collectées, la distribution des probabilités de traduction des séquences est estimée par une fréquence relative.

Koehn *et al.* (2003) améliorent l'alignement lexical réalisé par *GIZA++* avec un certain nombre d'heuristiques. Ainsi, un corpus parallèle est aligné bi-directionnellement au niveau lexical : à partir de la langue source vers la langue cible et inversement. Afin de symétriser ces deux alignements, il est possible d'utiliser des heuristiques telles que l'intersection ou l'union. Si les deux alignements sont intersectés, une précision élevée de l'alignement et une confidence élevée des points d'alignement peuvent être obtenues. En revanche, si les deux alignements sont réunis, un rappel significatif de l'alignement est obtenu par l'addition des points d'alignement supplémentaires.

La collecte des points d'alignement commence par l'intersection des alignements bidirectionnels et s'étend par l'addition des points supplémentaires à partir de l'union de ces alignements. Plusieurs questions en fonction desquelles les points d'alignement additionnels sont sélectés doivent être prises en compte (Koehn *et al.*, 2003) :

- Dans quel alignement le point d'alignement potentiel se trouve-t-il ? (Langue source - cible ou inversement ?)

[61] L'outil d'alignement lexical *GIZA++* (Och et Ney, 2000, 2003) sera décrit plus en détail dans le chapitre 3.

- Le point voisin du point potentiel a-t-il déjà établi des points ?

- La notion de points « voisins » signifie-t-elle des points directement adjacents ou des points adjacents en diagonale ?

- Est-ce le mot cible ou le mot de la langue source que le point potentiel connecte non-aligné jusqu'ici ? Sont-ils non alignés tous les deux ?

- Quelle est la probabilité lexicale du point potentiel ?

Un exemple qui illustre l'algorithme d'extraction de séquences proposé par Koehn *et al.* (2003) sur un couple de phrases parallèles français - anglais est donné dans la Figure 9 suivante (Lavecchia, 2010 : 46). Ces phrases (*je vous invite à vous lever pour cette minute de silence* vs. *please rise then for this minute's silence*) sont extraites du corpus parallèle *Europarl*[62]. Les alignements bidirectionnels (français - anglais et vice-versa) des deux phrases considérées sont obtenus avec *GIZA++*. Ceux-ci sont représentés visuellement par l'intermédiaire des matrices d'alignement.

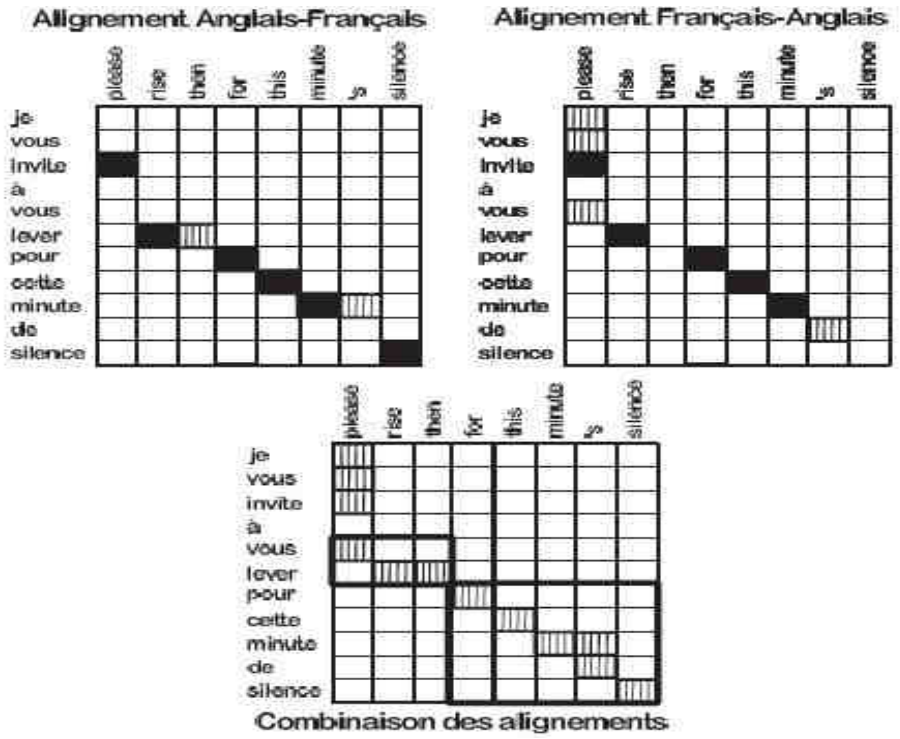


Figure 9. Exemple d'extraction de séquences français - anglais selon la méthode de Koehn *et al.* (2003) (Lavecchia, 2010 : 46)

[62] Ce corpus sera décrit plus loin dans la sous-section 2.2.2.4.

Le fonctionnement de la base heuristique (Och *et al.*, 1999) se déroule comme suit : Premièrement, les deux alignements lexicaux sont intersectés. Ensuite, les nouveaux points d'alignement qui sont additionnés sont seulement les points qui existent par l'union des deux alignements lexicaux. Il est nécessaire également qu'un nouveau point d'alignement connecte au moins un mot non-aligné précédemment (Koehn *et al.*, 2003).

Dans leur approche heuristique, Koehn *et al.* (2003) s'intéressent, en premier, seulement aux points d'alignement directement adjacents. Ils vérifient les points d'alignement pour le premier mot cible, pour le deuxième, etc. Cette opération itérative s'arrête quand aucun point d'alignement ne peut plus être rajouté. Finalement, les points d'alignement non-adjacents sont aussi rajoutés selon les mêmes critères.

À partir des alignements ainsi combinés, toutes les paires de séquences consistantes avec le nouvel alignement, c'est-à-dire les paires dont les mots sources et cibles sont alignés à l'intérieur de ces mêmes paires, sont extraites. Dans la Figure 9, ce type de paires de séquences est marqué par un trait épais.

La deuxième méthode testée dans le cadre de travail de Koehn *et al.* (2003), notamment le modèle syntaxique de traduction automatique statistique à base de séquences, sera présentée dans la sous-section suivante.

2.2.2.2. Modèle syntaxique

Afin de résoudre les problèmes rencontrés par les systèmes qui n'utilisent pas d'informations linguistiques (comme, par exemple, le problème de dépendances entre les mots ou les groupes de mots), les modèles syntaxiques (Wu, 1997 ; Yamada et Knight, 2001 ; Imamura, 2002) tiennent compte du fait que le ré-ordonnancement des mots au niveau de la phrase est géré par la syntaxe.

Ainsi, dans l'approche syntaxique de traduction automatique statistique (Wu, 1997 ; Yamada et Knight, 2001 ; Imamura, 2002), les groupes de mots sont considérés des séquences de traduction s'ils représentent des constituants, c'est-à-dire des sous-arbres dans un arbre syntaxique (p. ex. le groupe nominal). De ce fait, le ré-ordonnancement des mots est limité au ré-ordonnancement des constituants dans les arbres d'analyse syntaxique bien formés. En effet, seulement la traduction des séquences qui couvrent tous les sous-arbres syntaxiques est possible. Aussi est-il important de voir si cette limitation a un effet positif ou négatif sur le

système. Les paires de séquences syntaxiques sont identifiées à partir des corpus alignés au niveau lexical et annotées avec des arbres syntaxiques générés par des analyseurs syntaxiques statistiques (Collins, 1997 ; Schmidt et Schulte im Walde, 2000).

Le processus de traduction dans le modèle syntaxique de traduction automatique statistique se déroule en quatre étapes :

1) le ré-ordonnancement des constituants syntaxiques de la phrase source selon la syntaxe de la phrase cible ;

2) l'insertion des mots des constituants de la phrase cible, n'ayant pas d'équivalent dans la phrase source, aux constituants de la phrase source ;

3) la traduction des constituants syntaxiques ;

4) la lecture des feuilles de l'arbre syntaxique obtenu à l'étape antérieure, qui fournit la phrase cible.

Le schéma du processus de traduction du modèle syntaxique de traduction automatique statistique est donné dans la Figure 10 suivante. Cet exemple concerne le couple de phrases anglais - japonais (*he adores listening to music* vs. *Kare ha ongaku wo kiku no ga daisukidesu* 'il adore écouter de la musique') (Yamada et Knight, 2001).

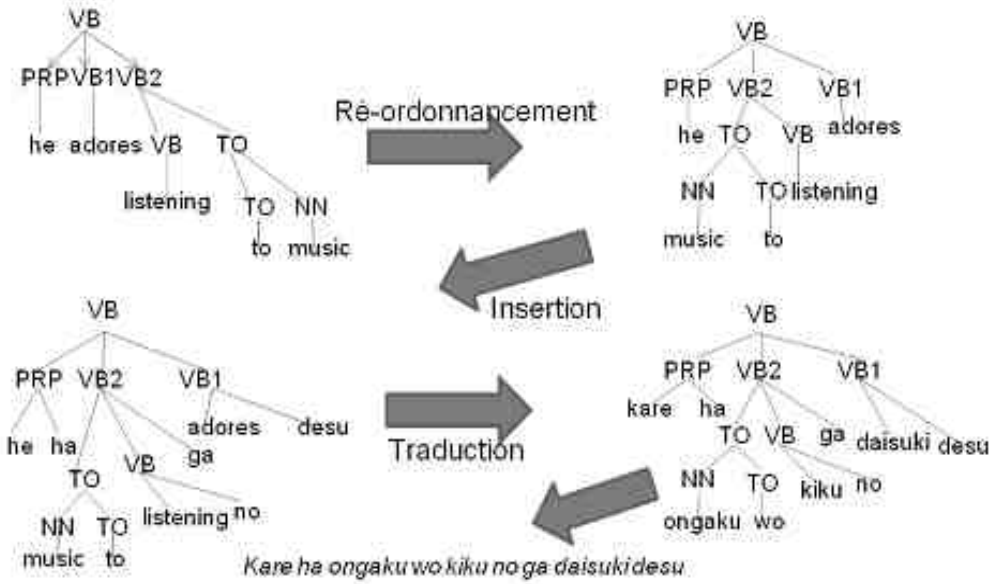


Figure 10. Schéma du processus de traduction dans un modèle syntaxique de traduction automatique statistique à base de séquences (Yamada et Knight, 2001)

Koehn *et al.* (2003) incluent aussi dans leur modèle syntaxique seulement les séquences qui constituent des sous-arbres dans un arbre d'analyse syntaxique. Ainsi, leur modèle apprend les paires de séquences basées sur la syntaxe qui représentent un sous-ensemble des paires de séquences apprises sans information syntaxique (cf. sous-section antérieure).

La troisième méthode testée par Koehn *et al.* (2003) est le modèle des séquences jointes qui sera décrit dans la sous-section suivante.

2.2.2.3. Modèle des séquences jointes

Le modèle des séquences jointes (Marcu et Wong, 2002) fonctionne par l'apprentissage des alignements au niveau des séquences directement à partir des corpus parallèles. Ainsi, le modèle de traduction de Marcu et Wong (2002) suppose que les correspondances lexicales peuvent être établies non seulement au niveau lexical, mais aussi au niveau de la séquence. Pour que le système apprenne des correspondances lexicales, Marcu et Wong (2002) développent un modèle de probabilité jointe basé sur les séquences, qui génère simultanément les phrases source et cible à partir d'un corpus parallèle. Le modèle d'apprentissage de Marcu et Wong (2002) utilise deux paramètres statistiques :

- une distribution de probabilité jointe $\varphi(\bar{e}, \bar{f})$ constituant la probabilité que les segments $\bar{e}$ et $\bar{f}$ sont des équivalents de traduction ;

- une distribution jointe *d(i, j)* constituant la probabilité qu'un segment sur une position *i* est traduit en un segment sur une position *j*.

Les trois méthodes d'apprentissage automatique de séquences qui ont été décrites antérieurement (cf. sous-sections 2.2.2.1., 2.2.2.2. et 2.2.2.3.), ont été évaluées et comparées par Koehn *et al.* (2003) afin de rendre compte de leur performance. Cette évaluation sera présentée dans la sous-section suivante.

2.2.2.4. Évaluation des méthodes

Pour comparer les résultats des trois méthodes prises en compte, Koehn *et al.* (2003) exploitent la partie allemand -> anglais du corpus parallèle *Europarl*[63] (1996-2001) contenant

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

environ 20 millions de mots par langue. Ce corpus est composé des travaux du Parlement Européen des années 1996-2011. Il est disponible pour 21 langues officielles de l'Union Européenne et comprend environ 30 millions de mots par langue. Un échantillon de 1 755 phrases est constitué afin de tester les méthodes considérées. La mesure d'évaluation utilisée est le score *BLEU*[64] (Papineni *et al.*, 2002) qui sera présenté et discuté ci-dessous.

2.2.2.4.1. Le score *BLEU*

Le score *BLEU* (Papineni *et al.*, 2002) s'avère être la mesure d'évaluation la plus usitée en traduction automatique. Celui-ci est exprimé souvent en pourcentage étant compris entre 0 et 1. Il compare une traduction candidate *c* avec une traduction de référence *r* (réalisée par les humains) au niveau des mots, des bigrammes, des trigrammes, etc. Le score *BLEU* se calcule selon la formule suivante (Patry et Langlais, 2005) :

$$BLEU\left(c,r\right)=BP\times e^{\sum_{n=1}^{N}\frac{n-grammes_{c}\cap n-grammes_{r}}{N\times\left|n-grammes_{c}\right|}}$$

où :

- N : la longueur maximale des n-grammes considérés (p. ex. 1, 2, 3, 4) ;
- $n-grammes_{c}$: l'échantillon de n-grammes des phrases candidates (*c*) ;
- $n-grammes_{r}$: l'échantillon de n-grammes des phrases de référence (*r*) ;
- BP : la *brevity penalty* qui est décrite par la formule suivante :

$$BP=\min\left(1,e^{\frac{|c|}{|r|}}\right)$$

Le rôle du coefficient BP dans la formule de calcul du score *BLEU* est d'éviter que le score ne favorise pas les traductions candidates courtes pour lesquelles $\left|n-grammes_{c}\right|$ est petit, ce qui augmente le quotient dans l'exponentielle de la formule du *BLEU* de façon artificielle (Sadat *et al.*, 2006).

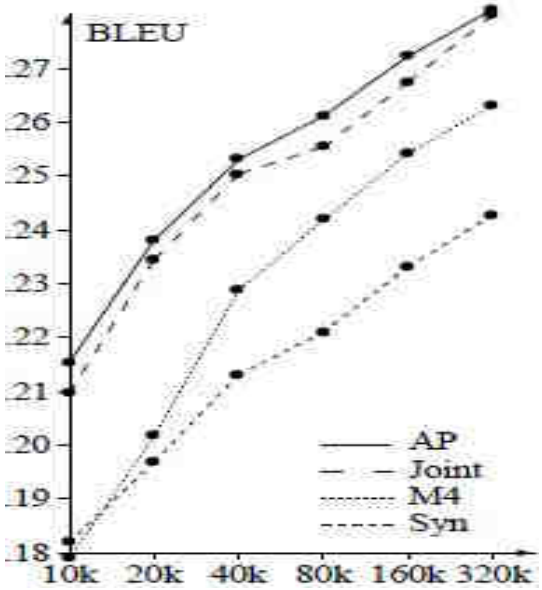
[64] *Bilingual Evaluation Understudy*

Le score *BLEU* a comme inconvénient le fait qu'il peut éliminer du calcul des n-grammes traduites correctement par le système de traduction mais qui ne font pas partie de l'échantillon de référence. En effet, on sait que la traduction humaine est souvent subjective et qu'une phrase source peut présenter plusieurs variantes de traduction. Dans ces cas, si une variante de traduction proposée correctement par le système ne se trouve pas parmi les phrases de référence, le score *BLEU* ne la prendra pas en compte alors qu'elle s'avère correcte lors d'une évaluation humaine. De ce fait, une évaluation humaine reste toujours nécessaire pour rendre réellement compte de la qualité d'une traduction proposée par un système de traduction automatique, même si cela est une tâche très coûteuse en termes de temps et de ressources humaines.

Cependant, l'avantage du score *BLEU* réside dans le fait qu'il peut prendre en compte plusieurs traductions comme référence, étant mieux corrélé avec l'évaluation humaine si plusieurs traductions de référence sont utilisées (Ceaușu, 2009).

2.2.2.4.2. Comparaison des méthodes

Koehn *et al.* (2003) comparent les trois méthodes à base de séquences prises en considération. Les résultats de ces méthodes sont représentés dans le graphique de la Figure 11 suivante (Koehn *et al.*, 2003 : 51). La performance du modèle lexical IBM 4 (M4) (cf. chapitre 3, sous-section 3.1.2.) y est également représentée.



(*Taille du corpus d'entraînement*)

AP = Modèle à base de séquences appris à partir des alignements lexicaux ; *Joint* = Modèle des séquences jointes ; *Syn* = Modèle syntaxique à base de séquences

Figure 11. Comparaison des résultats des méthodes à base de séquences (Koehn *et al.*, 2003 : 51)

Les expériences effectuées par Koehn *et al.* (2003) montrent que le modèle d'apprentissage de tous les segments à partir des alignements lexicaux (Och *et al.*, 1999) obtient des résultats meilleurs que le modèle des séquences jointes (Marcu et Wong, 2002). De plus, la seule utilisation des segments basés sur la syntaxe (Wu, 1997 ; Yamada et Knight, 2001 ; Imamura, 2002) s'avère nuisible au système parce que cette méthode conduit à l'élimination d'une quantité importante de paires de séquences. En effet, la seule utilisation des sous-arbres syntaxiques ne conduit pas à de meilleures paires de séquences, mais à la réduction considérable des paires de séquences et à la perte de connaissances importantes. Par exemple, la séquence en allemand *es gibt* qui signifie *there is* en anglais, est traduite par *it gives* en anglais. Comme les séquences *es gibt* et *there is* ne sont pas des constituants syntaxiques, leur traduction d'une langue à l'autre s'avère erronée.

Vu ces considérations, nous voyons que le choix de la méthode d'apprentissage de séquences de traduction devient important, afin d'atteindre un niveau de performance élevé dans le processus de traduction.

En conclusion, les expériences de Koehn *et al.* (2003) montrent qu'aucune des trois méthodes comparées ne contribuent de façon significative à la qualité de la traduction. Par contre, les résultats obtenus par les modèles à base de séquences sont meilleurs par rapport aux résultats fournis par les modèles lexicaux (Brown *et al.*, 1993). Les meilleurs résultats sont obtenus par l'utilisation de petites séquences de jusqu'à trois mots. De plus, l'ajout du poids lexical pendant le processus de traduction améliore la performance des résultats.

Le modèle syntaxique n'est pas capable de prendre en compte des alignements importants au niveau des séquences. Par conséquent, ce modèle ne donne pas de meilleures performances par rapport aux modèles à base de séquences sans utilisation de la syntaxe.

À partir des résultats obtenus par Koehn *et al.* (2003), nous pouvons conclure que les meilleures performances sont obtenues par l'utilisation de l'approche heuristique. En effet, il est très important de savoir comment les séquences sont extraites. Ainsi, le choix du bon alignement heuristique est plus important que le modèle utilisé pour créer les alignements lexicaux initiaux (Koehn *et al.*, 2003).

Pour améliorer les résultats fournis par les modèles standards de traduction automatique statistique à base de séquences, une autre démarche possible serait de faire appel à des ressources linguistiques déjà existantes (dictionnaires, bases de données terminologiques)

pour différentes langues. Sadat *et al.* (2006) proposent une telle méthode dans le projet *Portage* (Groupe de Technologies Langagières Interactives (*GTLI*), Conseil national de recherches, Canada, 2006) qui sera décrit dans la sous-section suivante.

2.2.2.5. Modèle utilisant des ressources terminologiques

Dans le projet *Portage*, Sadat *et al.* (2006) adoptent une approche statistique à base de séquences qui utilise deux corpus parallèles alignés et, de plus, un dictionnaire de terminologie bilingue français - anglais, afin d'améliorer les résultats de la traduction automatique statistique. Les corpus utilisés sont le *Hansard*, un corpus français - anglais aligné et composé des débats de la *Chambre des Communes* du *Parlement canadien*, et le corpus *Europarl* (cf. sous-section 2.2.2.4.) extrait des registres du *Parlement européen* et disponible aussi pour la paire de langues français - anglais.

Le système *Portage* fonctionne selon les cinq phases principales ci-dessous (Sadat *et al.*, 2006) :

1. *prétraitement* par tokénisation des données bruitées avec traduction de quelques mots ou séquences générés à partir de règles ; La tokénisation est appliquée aux deux corpus (source et cible) et prend en compte les espaces pour identifier les mots, les nombres et la ponctuation. Ensuite, les textes sont mis en minuscule afin de simplifier le travail pendant la phase d'entraînement des données.

2. *entraînement* pour construire les modèles de langue (de la langue cible) et de traduction à base de séquences ;

3. *décodage* pour traduire le texte source utilisant des hypothèses de traduction ;

4. *ré-ordonnancement* afin d'obtenir une ou plusieurs hypothèses de traduction, suivi du *ré-ordonnancement* des hypothèses obtenues pour maximiser la performance du système ;

5. *post-traitement* des résultats de sortie en redonnant un format adéquat aux traductions fournies par le système, en fonction de la langue cible.

Le système *Portage* comprend aussi un module dénommé *nombre et date* à base de règles, développé afin de repérer des nombres et des dates dans le texte source et de détecter leurs équivalents de traduction dans le texte cible.

Pendant l'étape d'alignement des corpus parallèles utilisés, Sadat *et al.* (2006) emploient l'algorithme de Moore (2002) afin de mettre en correspondance les lignes des corpus de sorte que la $i^{ème}$ ligne dans le corpus cible représente la traduction de la $i^{ème}$ ligne dans le corpus source. Les corpus alignés comprendront le même nombre de lignes.

Lors de l'étape de décodage, le système *Portage* combine quatre composants principaux :

1. un modèle trigramme en langue cible ;
2. un modèle de traduction à base de séquences ;
3. un modèle de distorsion qui mesure les différences dans l'ordre des mots en langues source et cible ;
4. un modèle de longueur qui mesure les différences de longueur des phrases entre les deux langues.

Pour améliorer les résultats du système *Portage*, Sadat *et al.* (2006) ont fait appel à un dictionnaire terminologique bilingue anglais - français, le *Grand Dictionnaire Terminologique* (*GDT*, L'Office québécois de la langue française). Ce lexique comprend des termes, des synonymes, des acronymes, des définitions, des unités phraséologiques, des exemples d'utilisation et des observations dans différents domaines. Il fournit environ 3 millions de termes français - anglais appartenant au vocabulaire industriel, scientifique et commercial, dans 200 domaines d'activité (Sadat *et al.*, 2006).

En plus du modèle de traduction obtenu par l'utilisation des corpus parallèles, Sadat *et al.* (2006) construisent un modèle de traduction probabiliste à partir des paires de termes en langue source et des alternatives de traduction dans le *GDT*. Concernant chaque terme ou phrase source, leurs probabilités sont considérées équiprobables : par exemple, pour le mot français *port*, si le lexique propose trois traductions en anglais *haven*, *harbor* et *port*, le modèle P(t|s) aura les probabilités $P(haven|port)$=0.33, $P(harbor|port)$=0.33, $P(port|port)$ = 0.33.

L'évaluation de *Portage* a montré que l'incorporation du *GDT* dans le système a amélioré les résultats de traduction automatique statistique, conformément au score *BLEU*. En fait, par l'utilisation d'une liste d'équivalents de traduction extraits du *GDT* dans l'entraînement d'un modèle de traduction automatique statistique, en plus des modèles de traduction et de langue obtenus à partir des corpus utilisés, la performance du système peut être augmentée. De plus, pour améliorer les résultats de la traduction automatique statistique, Sadat *et al.* (2006) envisagent d'intégrer de l'information morphologique à la phase du décodage.

Toutefois, d'une part, il n'existe pas actuellement des ressources linguistiques disponibles pour certaines paires de langues, moins dotées, comme c'est le cas de la paire de langues étudiées dans le cadre de notre travail. De ce fait, l'implémentation des systèmes de traduction automatique statistique faisant appel à des ressources linguistiques telles que dictionnaires, bases de données terminologiques, lexiques, etc., s'avère difficile à réaliser pour ces langues. D'autre part, les modèles probabilistes standard à base de séquences faisant appel à des ressources terminologiques (comme *Portage* - Sadat *et al.*, 2006) ou non, n'incorporent pas d'informations linguistiques dans les modèles de traduction et de langue eux-mêmes. Ce fait rend ces systèmes moins performants par rapport aux systèmes plus récents qui utilisent des techniques de factorisation permettant d'exploiter la contribution de divers facteurs linguistiques (morphologiques, syntaxiques) à la qualité de la traduction.

Ainsi, les systèmes qui permettent l'exploitation d'informations linguistiques et qui donnent des résultats performants en termes de scores statistiques d'évaluation et de cohérence grammaticale, sont les systèmes factorisés (Koehn et Hoang, 2007 ; Birch *et al.*, 2007 ; Tufiș *et al.*, 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011 ; Tufiș et Dumitrescu, 2012). Ces systèmes ne considèrent plus le mot comme un simple token, mais comme un vecteur de facteurs linguistiques, tels que : formes de mots, lemmes, parties du discours, traits morphosyntaxiques, etc. Les systèmes factorisés s'avèrent performants parce qu'ils offrent la possibilité de moduler les ressources linguistiques utilisées. Ainsi, comme les contraintes locales sur les mots apparaissent au niveau morphologique, l'intégration du niveau morphologique riche au processus de traduction améliore nettement les résultats. De plus, l'information syntaxique intégrée dans la phrase source permet l'extraction de l'information linguistique demandée par la phrase cible (p. ex. la catégorie grammaticale du cas des noms) et le système est capable de générer la forme fléchie correcte du mot. L'inconvénient des

systèmes factorisés est le fait que leurs performances dépendent des paires de langues considérées et aussi du sens de la traduction.

Dans les sous-sections suivantes seront présentés divers modèles factorisés de traduction automatique statistique (Koehn et Hoang, 2007 ; Ceaușu et Tufiș, 2011 ; Tufiș et Dumitrescu, 2012 ; Birch *et al.*, 2007 ; Avramidis et Koehn, 2008) pour différentes paires de langues.

2.2.2.6. Modèles factorisés

Comme nous l'avons vu auparavant, les modèles courants de traduction automatique à base de séquences sont limités à la mise en correspondance des *chunks* sans présentation d'informations linguistiques explicites (morphologiques, syntaxiques ou sémantiques). Selon Koehn et Hoang (2007), les raisons de la nécessité d'intégrer des informations linguistiques dans le modèle de traduction pendant les étapes de prétraitement et post-traitement du corpus sont les suivantes :

- Les modèles de traduction qui prennent en compte des représentations linguistiques comme, par exemple, les lemmes à la place des formes de mots, peuvent donner de meilleurs résultats en termes de scores statistiques d'évaluation ; En effet, ces modèles peuvent réduire les problèmes liés au phénomène bien connu de la dispersion des données dans les corpus d'entraînement. La dispersion des données fait référence à la répartition des formes fléchies dans les corpus d'apprentissage pour différentes paires de langues. En fait, les formes fléchies des mots ont des distributions différentes dans un corpus d'entraînement, surtout en ce qui concerne les langues riches morphologiquement. Ainsi, les formes qui sont moins fréquentes et leurs équivalents auront des probabilités faibles et ne seront pas traduites. En revanche, si toutes les formes fléchies d'un mot sont regroupées sous un même lemme, leur probabilité de traduction est ainsi augmentée.

- De nombreux aspects de la traduction peuvent être mieux expliqués au niveau morphologique, syntaxique ou sémantique. La disponibilité des informations linguistiques dans le modèle de traduction permet la modélisation directe des aspects de la traduction, comme par exemple : le ré-ordonnancement au niveau de la phrase est principalement basé sur la syntaxe, les contraintes locales sur les mots apparaissent au niveau morphologique, etc.

Les modèles factorisés de traduction automatique (Koehn et Hoang, 2007) représentent une extension des modèles de traduction automatique statistique à base de séquences, par l'intégration des annotations supplémentaires au niveau du mot, linguistiques, ou par la génération automatique de classes de mots. Nous comprenons par classe de mots le regroupement des mots en fonction de leurs traits morphosyntaxiques communs (p. ex. la classe du nom commun, genre masculin, nombre singulier). La méthode des n-classes obtenues automatiquement (Kneser et Ney, 1993) prédit non plus un mot en fonction de n-1 mots qui le précèdent dans la chaîne, mais une classe de mots par rapport à n-1 classes la précédant. Par conséquent, l'avantage de cette méthode est le fait qu'elle résout le problème de traduction des mots inconnus dans le corpus d'apprentissage. En effet, si un mot d'une classe donnée n'existe pas dans le corpus, on lui attribue empiriquement une probabilité d'être traduit par un mot de la même classe.

La Figure 12 suivante illustre la représentation des mots dans un système factorisé de traduction automatique statistique (Koehn et Hoang, 2007).

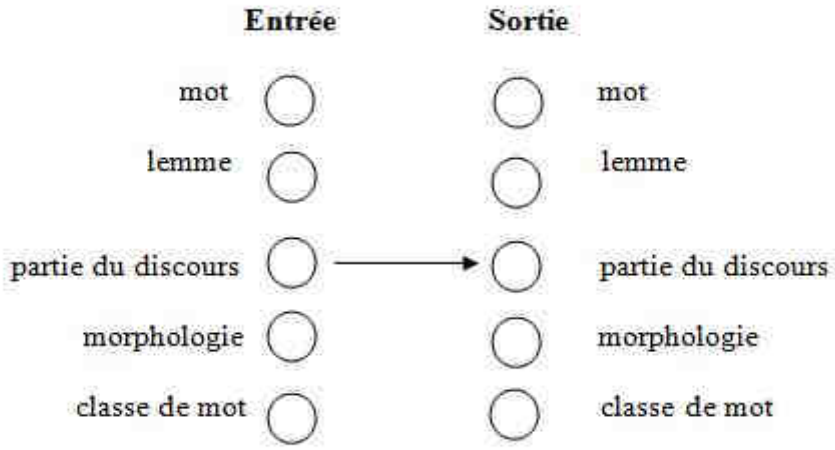


Figure 12. Représentation factorisée des mots annotés en entrée et en sortie dans un système statistique de traduction automatique factorisée (Koehn et Hoang, 2007)

Dans l'optique des modèles factorisés, la traduction des lemmes et des facteurs linguistiques se fait séparément et ces informations sont combinées, afin de générer les formes fléchies des mots en sortie. En effet, le processus de traduction dans le modèle factorisé de traduction automatique se divise en une séquence d'étapes de transformation dites de « mappage ». Pendant ces étapes ont lieu la traduction des facteurs linguistiques donnés en entrée en facteurs linguistiques de sortie et la génération des formes fléchies des mots cibles, en fonction des facteurs linguistiques de sortie.

Dans le système de Koehn et Hoang (2007), qui incorpore l'analyse morphologique et la génération, le processus de traduction se déroule en trois étapes de mappage (voir aussi Figure 13 ci-dessous) :

1) Traduction des lemmes d'entrée en lemmes de sortie ;
2) Traduction des facteurs morphologiques et parties de discours ;
3) Génération des formes fléchies des mots en sortie à partir des lemmes et des facteurs linguistiques traduits antérieurement.

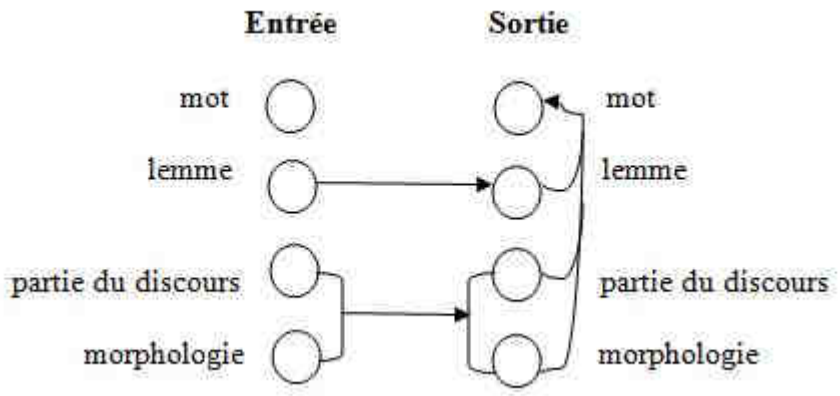


Figure 13. Exemple de modèle de traduction factorisé : analyse morphologique et génération des formes fléchies des mots cibles (Koehn et Hoang, 2007)

Ces étapes de mappage différencient le modèle factorisé de traduction automatique des modèles standard à base de séquences. Dans le modèle de Koehn et Hoang (2007), toutes les étapes de traduction fonctionnent au niveau de la séquence, alors que toutes les étapes de génération fonctionnent au niveau du mot.

Pour illustrer les étapes de mappage du modèle factorisé, Koehn et Hoang (2007) donnent comme exemple la traduction du mot *häuser* de l'allemand en anglais. Ainsi, la représentation du mot *häuser* en allemand est la suivante : forme du mot *häuser* | lemme *haus* | partie du discours *NN* | nombre *pluriel* | cas *nominatif* | genre *neutre*.
Les trois étapes de mappage sont appliquées comme suit :

1) Traduction : Mappage des lemmes
    - *haus → house, home, building, shell*
2) Traduction : Mappage de la morphologie
    - *NN|pluriel-nominatif-neutre → NN|pluriel, NN|singulier*
3) Génération : Génération des formes des mots
    - *house|NN|pluriel → houses*

- *house|NN|singulier → house*
- *home|NN|pluriel → homes*
- …

L'application de ces trois étapes à une séquence d'entrée est appelée l'*expansion* de la séquence. Dans le cas où il existe plusieurs choix lexicaux pour chaque étape, chaque séquence d'entrée peut être étendue dans une liste d'options de traduction. Ainsi, la représentation du *häuser* en allemand *häuser|haus|NN|pluriel-nominatif-neutre* peut être étendue comme suit (Koehn et Hoang, 2007) :

1) Traduction : Mappage des lemmes

{ *?|house|?|?, ?|home|?|?, ?|building|?|?, ?|shell|?|?* }

2) Traduction : Mappage de la morphologie

{ *?|house|NN|pluriel, ?|home|NN|pluriel, ?|building|NN|pluriel, ?|shell|NN|pluriel, ?|house|NN|singulier, ...*}

3) Génération : Génération des formes fléchies des mots cibles

{ *houses|house|NN|pluriel, homes|home|NN|pluriel, buildings|building|NN|pluriel, shells|shell|NN|pluriel, house|house|NN|singulier, ...*}

Vu ces considérations, les méthodes d'entraînement de corpus et de décodage dans un système factorisé (Koehn et Hoang, 2007) seront présentées ci-dessous.

Pour entraîner un modèle factorisé, il faut tout d'abord annoter un corpus parallèle avec des facteurs additionnels (étiquettes morphosyntaxiques et syntaxiques, lemmes). Ensuite, le corpus parallèle doit être aligné au niveau lexical. Les méthodes d'alignement peuvent opérer sur la forme des mots ou sur les facteurs additionnels (par exemple, l'utilisation des lemmes pendant l'étape d'alignement améliore la qualité de l'alignement - voir chapitre 4).

Chaque étape de mappage constitue une composante du modèle global. En effet, du point de vue du modèle d'entraînement, cela signifie qu'il faut apprendre des tables de traduction et de génération à partir d'un corpus parallèle aligné au niveau lexical et de définir des méthodes statistiques appropriées pour faire les bons choix parmi les résultats ambigus des étapes de mappage.

Koehn et Hoang (2007) obtiennent les modèles des étapes de traduction à partir des corpus parallèles alignés au niveau lexical selon la méthodologie des systèmes à base de séquences. Ainsi, pour les facteurs spécifiés en entrée et en sortie, des séquences mappées sont extraites.

Les distributions de génération sont estimées seulement en sortie. Le modèle de génération est appris sur une base mot-à-mot. Par exemple, pour l'étape de génération qui mappe la forme des mots à partir des parties du discours, une table contenant des entrées de type (fish,NN) est construite. Différents scores statistiques peuvent être définis à partir d'une telle table. Koehn et Hoang (2007) utilisent à la fois les distributions de probabilités conditionnelles p(fish|NN) et p(NN|fish).

Dans le cadre de travail des modèles factorisés, les modèles de langue n-grammes fonctionnent sur les facteurs ou les sets de facteurs spécifiés.

Ainsi, les modèles factorisés sont une combinaison de plusieurs composantes : modèle de langue, modèle de ré-ordonnancement, étapes de traduction, étapes de génération. Ces composantes définissent une ou plusieurs fonctions de traits qui sont combinées dans le modèle log-linéaire :

$$p(e \mid f) = \frac{1}{z} \exp \sum_{i=1}^{n} \lambda_i h_i(e, f)$$

Dans la formule ci-dessus, Z constitue une constante qui n'est pas prise en compte en pratique. Pour calculer la probabilité d'une traduction *e* d'une phrase donnée en entrée *f*, il faut évaluer chaque fonction de traits h_i. Par exemple, la fonction de traits pour un modèle de langue bi-grammes est la suivante :

$$\begin{aligned} h_{LM}(e, f) &= p_{LM}(e) \\ &= p(e_1)p(e_2|e_1)...p(e_m|e_{m-1}), \end{aligned}$$

où : *m* constitue le nombre de mots e_i dans la phrase *e*.

Les fonctions de traits introduites par les étapes de traduction et de génération des modèles factorisés seront décrites ci-dessous.

La traduction d'une phrase d'entrée *f* dans une phrase de sortie *e* se décompose dans un set de traductions de segments $\{(\overline{f_j}, \overline{e_j})\}$.

Pour l'étape de traduction, chaque fonction de traits h_τ est définie sur une paire de segments $(\overline{f_j}, \overline{e_j})$, pour une fonction de score τ donnée :

$$h_\tau(e, f) = \sum_j \tau\left(\overline{f_j}, \overline{e_j}\right)$$

Concernant l'étape de génération, chaque fonction de traits h_G, pour une fonction de score γ donnée, est définie seulement sur les mots de sortie e_k :

$$h_G(e, f) = \sum_k \gamma(e_k)$$

Les fonctions de traits suivent les fonctions de score (τ, γ) acquises pendant l'entraînement des tables de traduction et de génération : par exemple, une fonction de score pour un modèle de génération qui est une distribution de probabilités conditionnelles entre les facteurs d'entrée et de sortie $\gamma(fish, NN, \text{singulier}) = p(NN|fish)$.

La fonction de poids λ_i dans le modèle log-linéaire est obtenue par l'utilisation de la méthode d'entraînement de taux d'erreurs minimum (Och, 2003). Cette méthode optimise les paramètres du système en maximisant un score d'évaluation automatique à partir d'un corpus de développement.

Pendant l'étape de décodage, Koehn et Hoang (2007) adaptent l'algorithme heuristique de recherche en faisceau des modèles à base de séquences (voir sous-section 2.2.2.) pour les modèles factorisés où de multiples tables de traduction et de génération doivent être consultées et combinées, afin de trouver la meilleure hypothèse de traduction. Étant donné que toutes les étapes de mappage opèrent sur un même segment, les expansions de ces étapes peuvent être pré-calculées avant le décodage et stockées comme des options de traduction. De ce fait, l'algorithme heuristique de recherche en faisceau fondamental ne change pas. Comme les expansions des étapes de mappage créent un nombre significatif d'options de traduction, leur manipulation s'avère difficile en pratique. De ce fait, Koehn et Hoang (2007) réduisent les expansions et le nombre d'options de traduction pour un segment d'entrée à un nombre maximum (50 par défaut).

La méthode de Koehn et Hoang (2007) est implémentée par le décodeur open - source *MOSES*[65] (Koehn *et al.*, 2007) que nous allons décrire plus en détail dans le chapitre 3, car c'est ce décodeur qui sera également utilisé dans le cadre de notre travail. *MOSES* représente un système de traduction automatique à base de séquences qui, de plus, est capable de prendre

[65] http://www.statmt.org/moses/

en compte des modèles de traduction factorisés. Celui-ci a une grande popularité dans le domaine de la traduction automatique statistique mais il existe aussi d'autres décodeurs développés : *Pharaoh* (Koehn, 2004), *Ramses*[66] (Patry *et al.*, 2006), *Phramer*[67] (Olteanu *et al.*, 2006), *Marie*[68] (Crego et Marino, 2007a). Tous ces décodeurs ont en commun le fait qu'ils utilisent une fonction de score incrémentale pour trouver la meilleure traduction d'une phrase source donnée en entrée, par le biais de la programmation dynamique.

Dans leur approche, Koehn et Hoang (2007) s'intéressent davantage à l'intégration d'informations morphologiques dans le modèle de traduction au niveau du mot et moins aux informations syntaxiques exploitées par d'autres travaux du domaine (Wu, 1997 ; Yamada et Knight, 2001). Leur cadre de travail s'applique également à des modèles intégrant des classes de mots définies statistiquement ou qui présentent d'autres annotations au niveau du mot. Les expériences effectuées dans ce sens par Koehn et Hoang (2007) seront présentées dans la sous-section suivante.

2.2.2.6.1. Niveau morphologique dans les modèles factorisés

Koehn et Hoang (2007) expliquent le choix de leur approche par le fait que la représentation factorisée du mot annoté morphologiquement évite les problèmes posés par les systèmes de traduction automatique statistique classiques. En effet, les systèmes classiques prennent en compte seulement les formes des mots dans le modèle de traduction et chaque forme de mot est considérée comme un token (par exemple, les mots *house* et *houses* sont complètement indépendants). Par conséquent, chaque instance du mot *house* dans le corpus d'apprentissage n'ajoute aucune information pour la traduction du mot *houses*. De plus, dans les cas où la traduction du mot *house* est connue dans le système et le mot *houses* est inconnu, le système est incapable de traduire le mot *houses*. Si ce problème ne s'impose pas fortement dans le cas de l'anglais qui est une langue avec une flexion morphologique limitée, il est très significatif pour les langues avec une flexion morphologique riche. De ce fait, il est préférable pour ces langues d'utiliser des corpus lemmatisés dans le modèle de traduction et de mettre en commun les différentes formes de mots ayant le même lemme. De cette façon, de nouveaux mots inconnus dans le corpus d'apprentissage peuvent être traduits. Dans leur cadre de travail,

[66] http://smtmood.sourceforge.net
[67] http://www.phramer.org
[68] http://gps-tsc.upc.es/soft/marie

Koehn et Hoang (2007) ont effectué plusieurs expériences de l'anglais vers l'allemand, espagnol et tchèque, afin d'évaluer l'efficacité des modèles factorisés construits :

1) L'utilisation des sorties enrichies syntaxiquement ;

2) L'analyse morphologique et la génération ;

3) L'utilisation de classes de mots générées automatiquement ;

4) L'intégration de l'étape de *recasing* (Lita *et al.*, 2003; Wang *et al.*, 2006).

1) Le Tableau 2 suivant comprend les résultats obtenus par Koehn et Hoang (2007), évalués par le biais du score *BLEU* (Papineni *et al.*, 2002), quant à l'utilisation des sorties enrichies syntaxiquement. Les systèmes anglais - allemand et anglais - espagnol sont entraînés sur le corpus *Europarl* (respectivement 751 088 phrases et 40 000 phrases) et les systèmes anglais - tchèque sur le corpus *Wall Street Journal* (20 000 phrases).

Tableau 2. Résultats expérimentaux avec des sorties enrichies syntaxiquement (parties du discours, morphologie) (Koehn et Hoang, 2007)

anglais - allemand

Modèle	***BLEU***
Formes de mots	18,04%
Formes de mots + parties du discours	18,15%
Formes de mots + parties du discours + morphologie	18,22%

anglais - espagnol

Formes de mots	23,41%
Formes de mots + morphologie	24,66%
Formes de mots + parties du discours + morphologie	24,25%

anglais - tchèque

Formes de mots	25,82%
Formes de mots + morphologie	27,04%
Formes de mots + cas/nombre/genre	27,45%
Formes de mots + verbe (personne, temps, aspect) / prépositions (lemme, cas)	27,62%

Premièrement, le système de Koehn et Hoang (2007) traduit la forme de mots et génère des facteurs additionnels en sortie (parties du discours) à partir des traductions des formes de mots (cf. Figure 14 suivante).

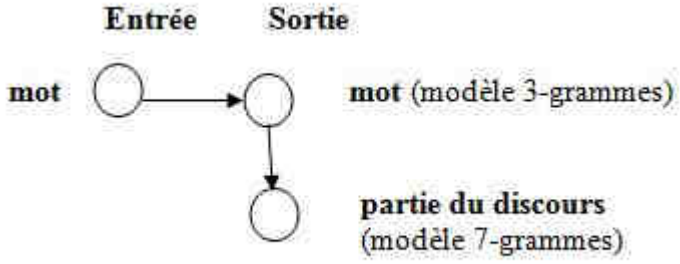


Figure 14. Sorties enrichies syntaxiquement (Koehn et Hoang, 2007)

Par l'ajout des informations morphologiques et syntaxiques superficielles dans les modèles factorisés, Koehn et Hoang (2007) utilisent des modèles de langue qui prennent en compte des séquences de mots plus longues (modèles de langue 7-grammes), afin de soutenir la cohérence syntaxique des sorties.

Pour les systèmes anglais - allemand, l'ajout des facteurs morphologiques et parties du discours aux sorties exploitées avec des modèles 7-grammes n'améliore pas beaucoup le score *BLEU*. Le modèle intégrant à la fois des parties du discours et la morphologie (*BLEU* 18,22% vs. modèle - formes 18,04%) assure une meilleure cohérence grammaticale. Le système basé sur les formes fournit souvent des phrases comme *zur (to) zwischenstaatlichen (inter-governmental) methoden (methods)*, avec un décalage entre le déterminant (singulier) et le nom (pluriel), alors que l'adjectif est ambigu. Les évaluations de l'accord à l'intérieur des groupes nominaux ont montré que le modèle factorisé réduit les erreurs de désaccord pour les segments ayant une longueur égale ou supérieure à 3 de 15% à 4% (Koehn *et al.*, 2007).

Les systèmes anglais - espagnol montrent des améliorations du score *BLEU* au niveau morphologique de 1,25% et au niveau morphologique et parties du discours à la fois, de 0,84%. Toutefois, les améliorations sur l'ensemble du corpus *Europarl* sont plus petites.

Les expériences effectuées sur les systèmes anglais - tchèque montrent qu'il faut examiner soigneusement quels facteurs morphologiques doivent être utilisés. En effet, l'ajout de tous les facteurs morphologiques fournit des résultats moins performants (27,04% *BLEU*), par rapport à l'addition seulement du cas, nombre et genre (27,45% *BLEU*) ou l'utilisation des facteurs morphologiques des verbes (personne, temps, aspect) et des prépositions (lemme et cas) (27,62% *BLEU*). Tous ces modèles ont des scores bien au-dessus par rapport au modèle qui ne prend en compte que la forme des mots (25,82%).

2) Les expériences suivantes de Koehn et Hoang (2007) portent sur le modèle motivé par la morphologie pour les langues allemand - anglais. Ainsi, au lieu de traduire la forme des mots, le système traduit les lemmes et les facteurs morphologiques séparément et ensuite génère la forme fléchie des mots en sortie.

Le Tableau 3 ci-dessous donne les résultats obtenus par les modèles construits dans cette optique (Koehn et Hoang, 2007).

Tableau 3. Résultats expérimentaux avec analyse morphologique et génération (Koehn et Hoang, 2007)

allemand - anglais

Modèle	***BLEU***
Formes des mots	18,19%
+ parties du discours *LM* (modèle de langue)	19,05%
Lemme/morphologie	14,46%
Modèle alternatif lemme/morphologie	19,47%

Les systèmes allemand - anglais sont entraînés sur 52 185 phrases du corpus *News Commentary*[69]. L'analyse morphologique et l'étiquetage avec des parties du discours du corpus allemand sont réalisés par *LoPar* (Schmidt et Schulte im Walde, 2000). Le corpus anglais est étiqueté par l'étiqueteur de Brill (Brill, 1995) et lemmatisé par un lemmatiseur simple basé sur les résultats d'étiquetage.

Par l'utilisation d'un modèle de langue basé sur les parties du discours le score *BLEU* augmente de 18,19% à 19,05%. Toutefois, passer d'une étape de mappage des traductions de formes de mots à une étape de mappage des lemmes et facteurs morphologiques montre une détérioration du score *BLEU* de 18,19% à 14,46%. De ce fait, il semble que la possibilité d'obtenir des traductions à partir des traductions des lemmes et des facteurs morphologiques en ignorant la forme des mots en entrée pose des problèmes à la qualité des résultats. Pour palier ce problème, Koehn et Hoang (2007) mettent en place un chemin alternatif de la méthode : les options de traduction peuvent être obtenues à partir du modèle basé sur les formes des mots ou à partir du modèle basé sur les lemmes et les facteurs morphologiques.

Ainsi, pour les formes de mots représentatives dans le corpus d'entraînement, Koehn et Hoang (2007) mettent en œuvre des étapes de mappage des formes de mots. Par contre, pour les formes de mots moins représentatives ou inconnues dans le corpus, ils décomposent les

[69] http://www.statmt.org/wmt07/

formes de mots en lemmes et informations morphologiques qui sont ensuite mappés séparément. Cette méthode alternative donne des résultats plus performants que le modèle utilisant la forme des mots et le modèle de langue basé sur les parties du discours (*BLEU* 19,47% vs. 19,05%). Le modèle qui mappe les lemmes et la morphologie a traduit 687 mots additionnels (le corpus de test contient 3 276 formes de mots inconnues vs. 2 589 lemmes inconnus, par conséquent leur différence constitue 687 mots inconnus ayant des lemmes connus).

3) Koehn et Hoang (2007) ont effectué des expériences en utilisant aussi des classes de mots générées automatiquement. Par le regroupement des mots par le biais de leurs similarités contextuelles, il est possible de trouver statistiquement des similarités qui peuvent donner des modèles plus généralisés et robustes.

Les modèles sont entraînés sur la tâche IWSLT 2006 (39 953 phrases). Le Tableau 4 suivant donne les résultats obtenus par un modèle basé sur la forme des mots par rapport à un modèle utilisant les classes de mots générées automatiquement, pour la paire de langues anglais - chinois.

Tableau 4. Résultats expérimentaux avec des classes de mots générées automatiquement obtenues par le regroupement de mots en fonction de leurs similarités contextuelles (Koehn et Hoang, 2007)

anglais - chinois

Modèle	***BLEU***
Formes de mots	19,54%
Formes de mots + classes de mots	21,10%

Par rapport à un modèle utilisant les formes de mots, l'addition des classes de mots aux sorties peut être exploitée par un modèle 7-grammes. Dans un tel modèle, le score *BLEU* s'améliore de 1,5%.

4) Pour démontrer la versatilité des modèles factorisés de traduction automatique, Koehn et Hoang (2007) prennent en compte la tâche de *recasing* (Lita *et al.*, 2003; Wang *et al.*, 2006).

Dans un système de traduction automatique statistique, le corpus d'entraînement est mis en minuscule pour généraliser les différentes variantes d'écriture d'un mot à un seul token (par

exemple the, The, THE). De ce fait, le système nécessite une étape de post-traitement des sorties pour restaurer les mots en sortie.

Par l'utilisation des modèles factorisés, il est possible d'intégrer cette étape dans le modèle en ajoutant une étape de génération. Le Tableau 5 suivant montre les résultats d'une telle méthode pour la paire de langues chinois - anglais.

Tableau 5. Résultats expérimentaux avec l'intégration de l'étape de *recasing* (*IWSLT 2006*) (Koehn et Hoang, 2007)

chinois - anglais

Modèle de *recasing*	***BLEU***
Étapes standard : système de traduction automatique statistique + *recasing*	20,65%
Modèle factorisé intégrant l'étape de *recasing* (optimisé)	21,08%

À partir du Tableau 5 ci-dessus, on peut observer que le modèle factorisé qui intègre l'étape de *recasing* donne des résultats plus performants en termes du score *BLEU*, par rapport à l'approche standard (21,08% vs. 20,65%).

En conclusion, Koehn et Hoang (2007) décrivent plusieurs expériences (de l'anglais vers l'allemand, l'espagnol et le tchèque) pour montrer comment différents facteurs linguistiques (morphologiques, classes de mots) peuvent être manipulés dans un système factorisé, afin d'augmenter la performance des résultats de la traduction automatique statistique.

Les résultats obtenus par les différents modèles de langue et de traduction construits montrent que l'intégration des facteurs morphologiques et des parties du discours au système factorisé donne de meilleurs résultats, par rapport à un modèle statistique standard basé sur la forme des mots. De plus, l'utilisation des lemmes résout le problème de traduction des mots inconnus dans le corpus d'entraînement. La méthode de traduction des lemmes et des étiquettes morphologiques et de génération des formes de mots en sortie nuit à la performance des résultats. De ce fait, il faut prendre en compte également les formes des mots. Ainsi, il faut utiliser une méthode de traduction basée sur les formes pour les mots représentatifs dans le corpus d'entraînement et une méthode de mappage des lemmes et des facteurs morphologiques pour les mots moins représentatifs et inconnus dans le corpus d'entraînement. Cette méthode alternative obtient de meilleures performances par rapport au modèle basé sur les parties du discours et la forme des mots et résout le problème de

traduction des mots inconnus dans le corpus d'entraînement. Des modèles factorisés plus généralisés et robustes peuvent être obtenus par l'intégration des classes de mots générées automatiquement qui prennent en compte les similarités contextuelles des mots.

Concernant le roumain, des systèmes de traduction automatique factorisés, avec l'anglais comme langue source ou cible, ont été proposés par Tufiș *et al*. (2008b), Ceaușu (2009)[70], Ceaușu et Tufiș (2011), Tufiș et Dumitrescu (2012), Dumitrescu *et al*. (2012).

Ceaușu et Tufiș (2011) décrivent des expériences où les descriptions morphosyntaxiques sont utilisées afin de traduire et de générer l'information morphologique dans un système factorisé, construit de l'anglais vers une langue riche morphologiquement comme le roumain. Ils montrent qu'à partir d'un corpus dont la taille est relativement réduite, la configuration proposée obtient de meilleurs résultats qu'un système de base uniquement construit sur les séquences. Ils montrent aussi que la configuration retenue améliore les résultats d'un système de base à partir d'un corpus plus large, dans le cas où un corpus supplémentaire de la langue cible est disponible.

Un premier corpus parallèle utilisé est *SEE-ERA.net* qui a été constitué dans le cadre du projet *SEE-ERA.net* (Tufiș *et al*., 2008b) visant la construction des modèles factorisés pour des langues slaves et balkaniques (bulgare, grec, slovène, roumain) de et vers l'anglais. D'autres langues comme le tchèque, le français et l'allemand sont incorporées aussi. *SEE-ERA.net* est basé sur le corpus parallèle multilingue juridique et administratif *JRC-Acquis*[71] (Steinberger *et al*., 2006), étant lemmatisé et étiqueté par des parties du discours et des propriétés morphosyntaxiques (pour les langues bulgare, tchèque, grec, anglais, slovène et roumain). Le corpus comprend environ 1,4 millions de tokens par langue.

Un deuxième corpus parallèle exploité est *STAR* disponible en roumain et en anglais. Ce corpus a été construit dans le cadre du projet *STAR*[72]. Il s'agit d'un corpus parallèle contenant principalement des textes du domaine juridique et journalistique. Celui-ci comprend aussi un corpus monolingue roumain « équilibré » à base de textes littéraires, de documents journalistiques et scientifiques. *STAR* est composé des textes provenant des corpus suivants :

[70] Le système développé par Ceaușu (2009) sera décrit en détail dans le chapitre 3, car il s'agit du système que nous avons adapté pour la paire de langues étudiées.
[71] Ce corpus sera décrit dans le chapitre 3 de ce travail, puisqu'il fait partie des corpus disponibles également pour le français et le roumain.
[72] http://www.racai.ro/star/

- le corpus juridique et administratif *DGT-TM* (*Directorate-General for Translation - Translation Memory*)[73] ;

- le corpus médical *EMEA* (*European Medicines Agency*) extrait à partir du corpus *OPUS* (Tiedemann, 2009) ;

- le corpus journalistique *SE Times* (*Southeast European Times corpus*) obtenu à partir du corpus *OPUS* (Tiedemann, 2009) ;

- le corpus journalistique anglais - roumain *NAACL* (Martin *et al*., 2005) utilisé lors de la compétition d'alignement de *NAACL* (*Association of Computational Linguistics, North American Chapter*) *2005* ;

- le corpus monolingue roumain « équilibré » (20 millions de tokens).

STAR contient environ 27 millions de tokens par langue. Il est aussi segmenté lexicalement, lemmatisé, étiqueté par des parties du discours et des propriétés morphosyntaxiques.

Pour améliorer la traduction vers les langues riches morphologiquement, les hypothèses suivantes peuvent être prises en considération (Ceaușu et Tufiș, 2011 ; Tufiș et Dumitrescu, 2012) :

- aligner et traduire les lemmes (à la place des formes de mots) peut réduire de manière significative le nombre des classes d'équivalence, surtout en ce qui concerne les langues riches morphologiquement qui présentent un nombre important de formes fléchies ;

- les mots traduits ont la tendance de garder la partie du discours de la langue source ou montrent des préférences pour certaines catégories lexicales - p. ex. un nom peut être traduit par un nom ou un verbe ;

- le ré-ordonnancement des mots de la phrase cible peut être amélioré si des modèles de langue construits sur les parties du discours ou les étiquettes morphosyntaxiques sont utilisés.

À partir du corpus *SEE-ERA.net*, plusieurs configurations de modèles factorisés ont été effectuées par le biais du décodeur *MOSES* (Koehn *et al*., 2007) pour les langues anglais -

[73] http://ipsc.jrc.ec.europa.eu/index.php/Traineeships/197/0/. Ce corpus sera décrit plus en détail dans le chapitre 3, car il est exploité aussi dans le cadre de notre travail.

grec, anglais - bulgare, anglais - slovène et anglais - roumain. Le corpus d'entraînement comprend 57 000 paires de phrases, le corpus de développement contient 500 paires de phrases et le corpus de test est composé de 1 000 phrases alignées. La taille du corpus d'entraînement est très limitée mais, comme les expériences le montreront, l'utilisation des informations linguistiques compense le déficit des données brutes (Ceaușu et Tufiș, 2011).

Les modèles de langue ont été construits en utilisant l'ensemble du corpus d'entraînement. Ces modèles sont basés sur la forme des mots (modèles 4-grammes) et sur les étiquettes de parties du discours et morphosyntaxiques (modèles 5-grammes).

Les systèmes de base ont été entraînés en utilisant les lemmes dans l'alignement et un modèle supplémentaire de ré-ordonnancement lexicalisé (Tillman, 2004) qui est censé fonctionner mieux que le modèle par défaut basé sur la distance. En effet, le modèle utilisant la distance s'applique dans une fenêtre de tokens et fournit un coût de ré-ordonnancement en fonction de la différence entre les positions des séquences sources et cibles (cf. sous-section 2.2.2.1.). À la différence de ce modèle, le ré-ordonnancement lexicalisé calcule, pour chaque séquence source, sa probabilité d'être traduite de façon monotone, échangé ou discontinue (à distance) par rapport à la traduction de la séquence précédente. Ce modèle est largement utilisé (il est intégré dans *MOSES* (Koehn *et al*., 2007)) car, par rapport à d'autres modèles qui exploitent seulement l'information de la langue source (Collins *et al*., 2005 ; Popovic et Ney, 2006 ; Crego et Marino, 2007b), il prend en compte les formes des mots au niveau de la langue source et cible (Crego et Yvon, 2010).

Le Tableau 6 suivant comprend quelques configurations factorisées pour la paire de langues anglais - roumain (Ceaușu et Tufiș, 2011) et leur évaluation en termes du score *BLEU* (Papineni *et al*., 2002). Les meilleurs résultats ont été obtenus par le système factorisé qui traduit les lemmes et les propriétés morphosyntaxiques avant de générer les formes des mots à partir des sorties obtenues antérieurement (voir configuration 4 du Tableau 6). Les résultats dépassent ceux fournis par le système de base de 1% (de 51,76 à 52,76).

Tableau 6. Les différentes configurations factorisées et leur évaluation pour la paire de langues anglais - roumain (Ceaușu et Tufiș, 2011)

Configu-rations	Modèles de traduction	Modèles de génération	Modèles de langue	Modèles de ré-ordonnancement	BLEU
1	formes de mots	-	formes de mots	formes de mots	51,76
2	lemmes	lemme -> formes de mots	formes de mots	distance	51,79
3	lemmes POS	lemmes -> POS lemmes, POS -> formes de mots	POS formes de mots	distance	52,31
4	lemmes MSD	lemmes -> MSD lemmes, MSD -> formes de mots	MSD formes de mots	distance	**52,76**
5	lemmes MSD	lemmes -> MSD lemmes, MSD -> formes de mots	MSD formes de mots	formes de mots	46,39
6	lemmes MSD	lemmes -> MSD lemmes, MSD -> formes de mots	MSD formes de mots	MSD	45,77

POS : *Part-of-Speech* (parties du discours) ;

MSD : *Morpho-Syntactic Descriptors* (descriptions morphosyntaxiques).

Le système factorisé retenu utilise des modèles de langue construits sur la forme des mots et sur les propriétés morphosyntaxiques. Le modèle de ré-ordonnancement est basé sur la distance. Il semble que les modèles lexicalisés soient redondants quand des modèles de langue basés sur les propriétés morphosyntaxiques (ou les parties du discours) sont utilisés (voir les configurations 5 et 6 du Tableau 6 où les résultats sont nettement déprécies) (Ceaușu et Tufiș, 2011).

Les scores *BLEU* performants obtenus dans le cadre de ces expériences peuvent être expliqués par les caractéristiques du corpus juridique utilisé (vocabulaire limité, des séquences longues qui se répètent le long du texte) (Ceaușu et Tufiș, 2011).

Le Tableau 7 suivant présente les scores *BLEU* calculés pour les systèmes de base et factorisés (présentant la même configuration 4 du Tableau 6) anglais - bulgare, anglais - grec, anglais - slovène, en comparaison avec les systèmes anglais - roumain.

Tableau 7. Évaluation des systèmes de base et factorisés anglais - bulgare, anglais - grec, anglais - slovène, anglais - roumain (Ceaușu et Tufiș, 2011)

Direction de traduction	BLEU Systèmes de base	BLEU Systèmes factorisés
anglais - bulgare	38,94	39,60
anglais - grec	42,22	43,07
anglais - slovène	40,73	42,68
anglais - roumain	51,76	52,76

Les systèmes factorisés obtiennent des scores améliorés par rapport aux systèmes de base pour toutes les paires de langues traitées. Les scores *BLEU* significativement élevés des systèmes anglais - roumain par rapport aux autres systèmes, s'expliquent par la qualité des ressources lexicales utilisées pour la segmentation lexicale et l'étiquetage des textes pour cette paire de langues. Pendant la segmentation lexicale, les expressions idiomatiques et la terminologie ont été corrélés dans le corpus anglais - roumain (Ceaușu et Tufiș, 2011).

Le système anglais - roumain a été construit aussi en exploitant le corpus *STAR* (1,5 millions paires de phrases) afin de vérifier si la configuration factorisée retenue améliore les résultats d'un système de base quand un corpus plus volumineux est utilisé.

Le système de base a été développé de la même manière que celui présenté antérieurement (utilisation des lemmes pendant l'alignement, intégration d'un modèle lexicalisé de ré-ordonnancement) et testé sur un ensemble de 1 000 phrases. Les scores *BLEU* obtenus lors de ces expériences figurent dans le Tableau 8 ci-dessous :

Tableau 8. Le système factorisé anglais - roumain exploitant le corpus *STAR* (Ceaușu et Tufiș, 2011)

Systèmes	BLEU
Système de base	53,82
Système factorisé (corpus parallèle)	53,41
Système factorisé (corpus monolingue supplémentaire)	54,52

Par l'utilisation d'un corpus plus volumineux, les systèmes anglais - roumain obtiennent de meilleurs scores par rapport aux premières expériences mais, cette fois-ci, le modèle factorisé n'améliore pas les résultats du système de base.

Pour augmenter la valeur du score *BLEU* du système factorisé, Ceaușu et Tufiș (2011) partent de l'idée que l'étape de génération s'appuie seulement sur la langue cible et ils intègrent alors le corpus monolingue *STAR* dans la partie roumaine du corpus parallèle, afin de construire les modèles de génération et d'entraîner le modèle de langue basé sur les propriétés morphosyntaxiques. Cette opération améliore le score *BLEU* de 0,7 points par rapport au système de base (voir le Tableau 8 ci-dessus). Ce résultat montre qu'il est possible d'améliorer les performances d'un système statistique factorisé utilisant de corpus plus volumineux, quand un corpus supplémentaire de la langue cible est disponible.

Afin d'améliorer encore les résultats de la traduction automatique statistique anglais - roumain (dans les deux sens du processus de traduction), Tufiș et Dumitrescu (2012) décrivent une méthode de traduction automatique appelée « en cascade ». Ils partent de l'intuition que la même méthodologie utilisée pour traduire d'une langue A vers une langue B peut être appliquée pour corriger partiellement les erreurs initiales de traduction. Cette démarche suit plusieurs étapes :

- Dans un premier temps, le meilleur système *S1* doit être construit ; Pour ce faire, un modèle factorisé est entraîné à partir d'un corpus bilingue parallèle (Ceaușu et Tufiș, 2011).

- Ensuite, la partie source du corpus parallèle est entièrement traduite par le biais du système *S1* ; Les résultats de sortie de ce système forment avec la partie cible du corpus parallèle initial un nouveau corpus.

- Puis, le nouveau corpus parallèle est utilisé pour entraîner un deuxième système *S2* ;

- Finalement, les deux systèmes *S1* et *S2* sont enchaînés (*S1+S2*). Le processus de traduction se déroule donc en cascade, c'es-à-dire le texte source d'entrée du *S1* est traduit dans le texte cible de sortie et, ensuite, ce texte cible devient le texte d'entrée du système *S2* qui va produire la traduction finale. Cette technique est censée corriger certaines erreurs de la traduction initiale.

Le corpus bilingue parallèle anglais - roumain utilisé lors des expériences menées a été construit dans le cadre du projet *ACCURAT FP-7*[74]. Ce corpus est obtenu à partir des sources suivantes :

1) le corpus juridique et administratif *DGT-TM*[75] (contenant environ 650 000 phrases) ;

2) le corpus médical *EMEA* (Tiedemann, 2009) (environ 994 000 phrases) ;

3) la partie anglais - roumain du thesaurus multilingue *Eurovoc*[76] (environ 6 500 termes[77] bilingues traités comme des phrases parallèles courtes) ;

4) *PHP*[78] (traduction du manuel de *PHP*, environ 30 000 phrases) ;

5) *KDE*[79] (traduction de l'interface *Linux KDE*, 114 000 phrases) ;

6) le corpus journalistique *SETIMES*[80] (environ 170 000 phrases).

Le corpus entier comprend environ 1 950 000 phrases. Il est nettoyé, segmenté lexicalement, lemmatisé et annoté par des parties du discours et des propriétés morphosyntaxiques.

Les annotations sont exploitées par *MOSES* (Koehn *et al*., 2007) pour construire les systèmes factorisés dans les deux sens du processus de traduction. Après l'étape de nettoyage, la taille du corpus est réduite à environ 1 250 000 phrases parallèles. Le corpus de test extrait contient 1 200 paires de phrases.

Plusieurs configurations de modèles ont été testées afin de construire le système *S1*. Dans le Tableau 9 suivant figurent les configurations ayant obtenu des scores *BLEU* performants (Tufiș et Dumitrescu, 2012).

[74] *Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation* (http://www.accurat-project.eu/)
[75] http://ipsc.jrc.ec.europa.eu/index.php/Traineeships/197/0/
[76] http://eurovoc.europa.eu/
[77] La notion de *terme* est définie dans le chapitre 4, sous-section 4.1.3.2.
[78] http://www.php.net/
[79] http://docs.kde.org/
[80] http://www.setimes.com/

Tableau 9. Configurations de modèles pour *S1* (Tufiș et Dumitrescu, 2012)

Modèle #	Détails
#1	t0-0 m0
#2	t1-1 g1-0 m0
#3	t1-1 g1-3 t3-3 g1,3-0 , m0m3
#4	t1-1 g1-3 t3-3 g1,3-0 , m0m3 r0
#5	t1-1 g1-3 t3-3 g1,3-0 , m0m3 r3

t : l'étape de traduction ; g : l'étape de génération ; m : modèle de langue ; r : modèle de ré-ordonnancement ; 0 : facteur formes des mots ; 1 : facteur lemme ; 3 : facteur propriétés morphosyntaxiques.

Pour la direction roumain -> anglais, le meilleur score *BLEU* (57.01 points) a été obtenu par le système #3 basé sur les lemmes et les propriétés morphosyntaxiques (les formes des mots sont générées à partir de ces facteurs traduits au préalable). Ce système utilise aussi deux modèles de la langue cible : un modèle qui est construit sur les formes des mots et un autre étant basé sur les étiquettes morphosyntaxiques.

Concernant la direction anglais -> roumain, le système retenu (#2) a obtenu un score *BLEU* (53.94 points) moins élevé que le système construit en sens inverse. Celui-ci génère les formes des mots à partir des lemmes traduits en langue cible et exploite un modèle de langue construit sur les formes des mots.

Les parties monolingues du corpus parallèle d'entraînement ont été ensuite traduites par le biais des systèmes construits. Puis, ces corpus ont été également prétraités (lemmatisés et annotés par des parties du discours et propriétés morphosyntaxiques) et nettoyés. Après l'étape de nettoyage, le corpus résulté comprend environ 1 110 000 phrases pour la direction roumain -> anglais et 1 010 000 phrases en sens inverse.

À partir de ces corpus parallèles intermédiaires, plusieurs configurations de modèles ont été testées pour construire le système *S2* (voir Tableau 10 suivant).

Tableau 10. Configurations de modèles pour *S2* (Tufiș et Dumitrescu, 2012)

Modèle #	Détails
#1	t0-0 m0
#2	t1-1 g1-0 m0
#3	t1-1 g1-2 t2-2 g1,2-0 m0,m2
#4	t1-1 g1-3 t3-3 g1,3-0 m0,m3
#5	t1-1 g1-3 t3-3 g1,3-0 m0,m3 r3
#6	t1-1 g1-2 t2-2 g2-3 t3-3 g1,3-0 m0,m2,m3
#7	t0,1-0,1 m0
#8	t0,1,2-0,1,2 m0,m2
#9	t1,2-t1,2 m0,m2

t : l'étape de traduction ; g : l'étape de génération ; m : modèle de langue ; r : modèle de ré-ordonnancement ; 0 : facteur formes des mots ; 1 : facteur lemme ; 2 : facteur partie du discours ; 3 : facteur propriétés morphosyntaxiques.

Le corpus de test du système *S2* est le même qu'initialement (1 200 phrases), à la différence qu'il est traduit avec les modèles retenus dans la première phase.

Pour la direction roumain -> anglais, le meilleur score *BLEU* (60.90 points) a été obtenu par le système #7 qui traduit les formes des mots et les lemmes sources en formes des mots et lemmes cibles et utilise un modèle de langue basé sur la forme des mots (cf. Tableau 10 ci-dessus). Les résultats obtenus montrent une amélioration significative (de 3,89 points) en termes du score *BLEU* du système en cascade *S1+S2* par rapport au système initial *S1* (de 57,01 à 60.90).

Quant à la direction de traduction anglais -> roumain, le meilleur score *BLEU* (54.44 points) a été fourni par les systèmes #7 et #8 (cf. Tableau 10). Les deux configurations fournissent le même score mais des temps de traduction différents (167 s pour #7 et 287 s pour #8). Les résultats du système en cascade *S1+S2* sont meilleurs de 0,50 points que ceux fournis par le premier système *S1* (de 53,94 à 54,44).

Les expériences présentées ci-dessus montrent de très bons scores *BLEU* pour la paire de langues anglais - roumain. De plus, ces résultats montrent bien qu'un système factorisé en cascade améliore les scores *BLEU* d'un système de base dans les deux directions du processus de traduction (Tufiș et Dumitrescu, 2012).

Dans cette sous-section, nous venons de voir comment les modèles factorisés qui utilisent des facteurs morphologiques dans le processus de traduction améliorent de manière significative les résultats de la traduction automatique statistique à base de séquences (Koehn et Hoang, 2007 ; Ceaușu et Tufiș, 2011). En outre, les résultats d'une telle approche peuvent être encore améliorés en utilisant des systèmes en cascade (Tufiș et Dumitrescu, 2012). Nous allons nous intéresser également aux modèles factorisés exploitant des facteurs syntaxiques (Birch *et al.*, 2007 ; Avramidis et Koehn, 2008), dans la sous-section suivante.

2.2.2.6.2. Niveau syntaxique dans les modèles factorisés

Birch *et al.* (2007) exploitent des modèles factorisés qui utilisent des étiquettes syntaxiques *CCG (Combinatorial Categorial Grammar)* comme facteurs au niveau des mots de la langue source ou cible. Dans les modèles factorisés, l'utilisation des méta-catégories *CCG* fournit des informations syntaxiques riches au niveau des mots, puisque ces étiquettes décrivent la structure de grammaire dans le lexique.

Les *CCG* ont des lexiques riches syntaxiquement et un ensemble d'opérateurs combinatoires qui rassemblent des arbres syntaxiques. Une catégorie du lexique est attribuée à chaque mot dans la phrase. Cette catégorie peut être atomique (ex. S, NP) ou complexe (ex. S\S, (S\NP)/NP). Les catégories complexes ont des formes générales de type α/β ou $\alpha\backslash\beta$ (c'est-à-dire α se combine avec β à gauche ou à droite, α et β étant aussi des catégories). Un exemple de *CCG* pour une phrase de l'anglais (Birch *et al.*, 2007) est donné dans la Figure 15 suivante :

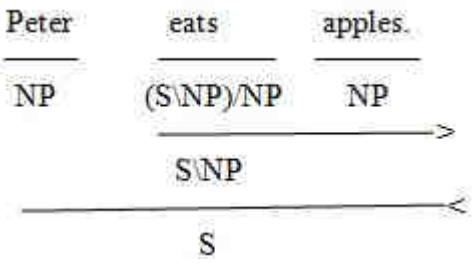


Figure 15. Exemple de *CCG* pour une phrase de l'anglais (Birch *et al.*, 2007)

Dans cet exemple, le procédé de dérivation est le suivant : *eats* est combiné avec *apples* dans le cadre du fonctionnement de l'application en avant. Mais *eats* peut être conçu comme une fonction qui prend un syntagme nominal NP (*apples*) à droite et retourne un S\NP. De la même manière, la phrase *eats apples* peut être conçue comme une fonction qui prend un

syntagme nominal NP (*Peter*) à gauche et retourne une phrase S. Cette opération est appelée *application en arrière*.

Une phrase avec ses catégories *CCG* comprend la plupart des informations existantes dans l'analyse complète. Ces catégories sont lexicalisées et, par conséquent, elles peuvent être intégrées facilement dans un modèle factorisé. Le corpus de travail peut être enrichi avec des méta-catégories *CCG* par le biais d'un méta-catégoriseur (Clark, 2002). Les méta-catégories ont été introduites par Bangalore et Joshi (1999) dans le but d'augmenter l'efficience de l'analyse du corpus par la réduction du nombre des structures attribuées à chaque mot du corpus (Birch *et al.*, 2007).

Le modèle factorisé de Birch *et al.* (2007) est donné dans la Figure 16 ci-dessous, où les mots cible et les méta-catégories *CCG* sont mappés en sortie à partir du mot source.

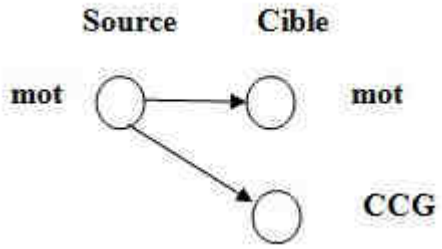


Figure 16. Modèle factorisé avec les mots source qui déterminent les mots cible et les méta-catégories *CCG* (Birch *et al.*, 2007)

Les expériences de Birch *et al.* (2007) montrent que l'intégration des étiquettes *CCG* comme facteurs linguistiques pour construire des modèles de langue n-grammes pour la langue cible et des modèles de traduction, améliore la qualité des résultats de traduction.

Le corpus d'entraînement pour les langues néerlandais - anglais est extrait d'*Europarl* (855 677 phrases). Le corpus de développement (500 phrases) et le corpus de test (2 000 phrases) proviennent du *ACL Workshop on Building and Using Parallel Texts*[81]. Les expériences sont réalisées avec *MOSES* (Koehn *et al.*, 2007).

Le Tableau 11 suivant montre les résultats de traduction des modèles factorisés (Birch *et al.*, 2007) pour le couple de langues néerlandais - anglais.

[81] http://www.statmt.org/wpt05/

Tableau 11. Résultats des modèles factorisés utilisant des facteurs comme mots (w), étiquettes des parties du discours (p), méta-catégories (c), pour les mots sources (s) ou cibles (t) (Birch *et al.*, 2007)

néerlandais - anglais

Modèle	***BLEU***
s_w, t_w	23,97%
s_w, t_{wp}	24,11%
s_w, t_{wc}	24,42%
s_w, t_{wpc}	**24,43%**

À partir du Tableau 11 ci-dessus, les résultats en termes du score *BLEU* montrent que le modèle utilisant des *CCG* en langue cible (modèle s_w, t_{wc}) obtient de meilleurs résultats par rapport au modèle standard basé sur la forme de mots (s_w, t_w) : 24,42% vs. 23,97%. Les modèles utilisant des *CCG* (modèles s_w, t_{wc} et s_w, t_{wpc}) donnent de meilleures performances par rapport aux modèles intégrant des étiquettes de parties du discours (modèle s_w, t_{wp}) : 24,42% et 24,43% vs. 24,11%. Les meilleurs résultats sont obtenus par le modèle (s_w, t_{wpc}) qui utilisent à la fois des *CCG* et des étiquettes des parties du discours (24,43%), mais la différence avec un modèle utilisant seulement les *CCG* n'est pas très significative. L'utilisation des *CCG* donne de meilleurs résultats puisque ces étiquettes sont plus informatives que les étiquettes des parties du discours, en contenant le contexte syntaxique du mot.

Les méta-catégories *CCG* sont utilisées également pour diriger le processus de traduction. Ainsi, les méta-catégories *CCG* intégrées dans la phrase source permettent au décodeur de prendre des décisions basées sur la structure des entrées. Par exemple, la sous-catégorisation des verbes aide à sélectionner la traduction correcte. Pour utiliser de multiples dépendances entre les facteurs de la phrase source, Birch *et al.* (2007) proposent l'utilisation de deux modèles de traduction log2Pr(t_w|s_{wc}) et log2Pr(t_w|s_w) combinés selon la méthode *logarithmic opinion pool (LOP)* (Smith *et al.*, 2005). Par l'utilisation de cette méthode, l'ensemble du modèle log-linéaire est combiné en utilisant une constante multiplicative qui conduit à de meilleures valeurs de paramètres.

Le Tableau 12 ci-dessous comprend les résultats des modèles construits pour la paire de langues allemand - anglais (Birch *et al.*, 2007). Les modèles utilisent les données proposées

dans le cadre du *NAACL 2006 Workshop on Statistical Machine Translation*[82], comme suit : 751 088 phrases pour l'entraînement, 500 phrases pour le développement et 3064 pour le test.

Tableau 12. Méta-catégories *CCG* utilisées comme facteurs dans la phrase source ; les modèles simples sont combinés selon deux démarches : le modèle log-linéaire et *LOP* des modèles log-linéaires (Birch *et al.*, 2007)

allemand - anglais

Modèle	***BLEU***
s_w, t_w	23,30%
s_{wc}, t_w	19,73%
Log-linéaire	23,29%
LOP	**23,46%**

À partir du Tableau 12 ci-dessus, on peut observer que le modèle général s_w, t_w obtient de meilleurs résultats par rapport au modèle s_{wc}, t_w intégrant des *CCG* au niveau des mots sources : 23,30% vs. 19,73%. Cela est dû au fait qu'il existe des mots dans la phrase de test qui ont été vus avant, mais sans des *CCG*. Par contre, la combinaison des deux modèles en utilisant la méthode *LOP* donne des résultats plus performants : 23,46%. Toutefois, la pertinence de l'intégration des informations syntaxiques au niveau de la phrase source reste à démontrer par plus d'expériences (Birch *et al.*, 2007) et également sur d'autres paires de langues.

Ainsi, les expériences de Birch *et al.* (2007) sur des modèles factorisés intégrant des méta-catégories *CCG* dans la phrase source ou cible montrent des résultats plus performants par rapport aux modèles factorisés qui n'incluent pas ces informations syntaxiques. Nous voyons donc l'importance des informations syntaxiques dans les modèles factorisés, qui améliorent la qualité de la traduction en termes de scores statistiques et de cohérence grammaticale.

Nous avons observé que les performances des systèmes factorisés en termes de scores statistiques (le score *BLEU* plus ou moins élevé par paires de langues) diffèrent en fonction des paires de langues considérées et aussi du sens de la traduction. Ainsi, nous avons vu dans les expériences de Birch *et al.* (2007) que l'intégration des étiquettes syntaxiques et de parties du discours améliore la qualité de la traduction automatique à partir des langues morphologiquement riches (néerlandais, allemand) vers l'anglais qui est une langue moins riche du point de vue morphologique.

[82] http://www.statmt.org/wpt06/

Les approches ultérieures (Avramidis et Koehn, 2008) mènent des expériences afin de tester l'efficacité de la syntaxe dans les modèles factorisés concernant la traduction à partir des langues pauvres du point de vue morphologique (p. ex. anglais) vers des langues riches morphologiquement (p. ex. grec, tchèque). Dans une étude pour tester la qualité de la traduction automatique pour les langues du corpus *Europarl*, Koehn (2005) montre que traduire vers les langues riches morphologiquement est plus difficile que de traduire à partir de ces langues. En effet, générer des informations morphologiques riches à partir des langues pauvres morphologiquement est plus difficile. Par exemple, le groupe nominal en anglais reste le même en position sujet ou objet dans la phrase. Par contre, les mots d'un groupe nominal en grec sont fléchis en fonction de leur position sujet ou objet dans la phrase. De ce fait, le simple mappage lexical des groupes nominaux en anglais vers les groupes nominaux en grec rencontre des problèmes à cause du manque d'information sur leur rôle dans la phrase et le choix de la forme fléchie correcte est difficile.

Dans leur approche, Avramidis et Koehn (2008) ajoutent des informations linguistiques au niveau des mots de la langue source. Ils utilisent la syntaxe de la phrase source afin d'extraire l'information du cas des noms et de la personne des verbes. Ensuite, ils annotent les mots correspondants de la langue cible en utilisant ces informations. Avramidis et Koehn (2008) mettent l'accent sur l'accord du cas des noms et la personne des verbes puisque ces éléments représentent les plus fréquentes erreurs du système statistique de base.

L'accord du cas des noms, adjectifs et articles est principalement défini par le rôle syntaxique de chaque groupe nominal dans la phrase. Ainsi, le cas nominatif définit les noms en position sujet dans la phrase, l'accusatif montre habituellement l'objet direct des verbes et le datif définit l'objet indirect des verbes bi-transitifs.

Dans leur approche, Avramidis et Koehn (2008) utilisent la syntaxe selon la méthode similaire *Semantic Role Labelling* (Carreras et Marquez, 2005 ; Surdeanu et Turmo, 2005).

Pour l'anglais, qui décrit habituellement un ordre fixe de mots, sujet - verbe - objet, l'utilisation de l'analyse syntaxique permet d'annoter facilement le sujet et l'objet dans les phrases. À partir de telles annotations, un modèle de traduction factorisé est entraîné pour mapper la paire mot - cas avec l'inflexion correcte du nom cible. Étant donné la restriction de l'accord, tous les mots qui déterminent le nom (adjectifs, articles, déterminants) doivent suivre le cas du nom, de sorte que leur cas susceptible doit être identifié à partir du cas du

nom. Pour manipuler ces informations, Avramidis et Koehn (2008) utilisent un analyseur syntaxique qui réalise les arbres syntaxiques pour chaque phrase en anglais. Les arbres sont analysés premièrement en profondeur et les cas sont identifiés à l'aide des *patrons sous-arbres* spécifiés manuellement. Avramidis et Koehn (2008) utilisent la séquence des nodes dans l'arbre afin de repérer le rôle syntaxique de chaque groupe nominal. Un exemple d'analyse syntaxique d'une phrase source anglaise (*We resolved the issue of Kossovo Bosnia Herzegovina or relations with Serbia*) est donné dans la Figure 17 ci-dessous (Avramidis et Koehn, 2008).

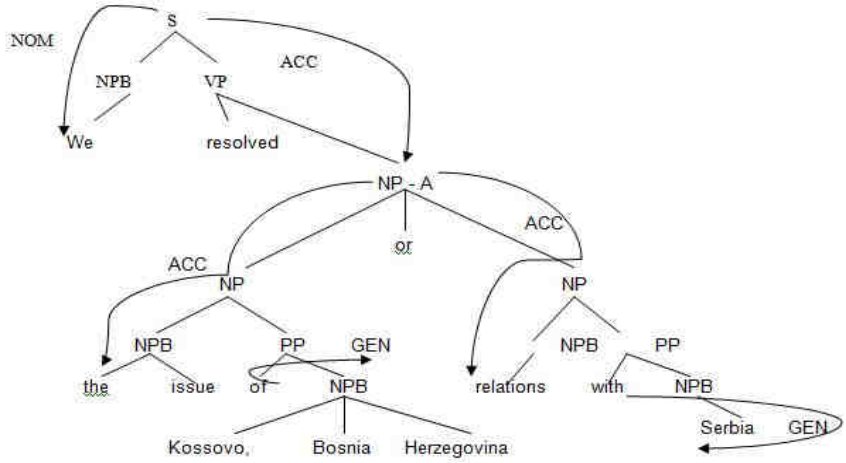


Figure 17. Les étiquettes du cas sont attribuées pendant l'analyse en profondeur de l'arbre syntaxique (anglais) basée sur des patrons sous-arbres (Avramidis et Koehn, 2008)

Premièrement, l'algorithme identifie le sous-arbre *S-(NPB-VP)* et l'étiquette du cas nominatif est appliquée au node *NPB* (qui correspond au pronom *we*). L'exemple de l'accusatif montre comment les cas sont transférés aux sous-arbres imbriqués. Des règles similaires sont appliquées pour couvrir l'ensemble des patrons de séquences de nodes.

Pour les verbes, l'étiquette de la personne est extraite à partir du sujet de la phrase.

Pendant l'étape de décodage avec *MOSES* (Koehn *et al.*, 2007), Avramidis et Koehn (2008) combinent le modèle log-linéaire avec un chemin alternatif de traduction des tables entraînées avec moins ou aucun facteur, selon la méthode de Birch *et al.* (2007). Ainsi, il est possible de couvrir les cas où un mot apparaît avec un facteur avec lequel le mot n'a pas été entraîné.

Pour entraîner les modèles factorisés anglais - grec, Avramidis et Koehn (2008) utilisent le corpus *Europarl* (440 082 phrases) et pour les modèles anglais - tchèque ils utilisent le corpus

News Commentary (57 464 phrases). Deux sets de phrases de test (environ 2000 phrases par set) sont utilisés pour chaque paire de langues. Les corpus sont analysés syntaxiquement par l'analyseur de Collins (Collins, 1997) et étiquetés avec les parties du discours par l'étiqueteur de Brill (Brill, 1992).

Les résultats de traduction anglais - grec et anglais - tchèque évalués par le biais du score *BLEU* (Papineni *et al.*, 2002) sont donnés dans les Tableaux 13 et 14 suivants :

Tableau 13. Résultats de la traduction de l'anglais vers le grec (Avramidis et Koehn, 2008)

	BLEU	***BLEU***
Set	**devtest**	**Test07**
Modèle de base	18,13%	18,05%
Personne	18,16%	18,17%
Parties du discours + personne	18,14%	18,16%
Personne + cas	18,08%	**18,24%**
Chemin alternatif : parties du discours	**18,21%**	**18,20%**

Avramidis et Koehn (2008) utilisent plusieurs combinaisons d'étiquettes pour les modèles factorisés. Certaines combinaisons sont affectées par les problèmes posés par la faible densité des données. Les meilleurs résultats sont obtenus par l'utilisation à la fois de la personne des verbes et le cas des noms : 18,24%. La méthode de traduction est plus efficiente pendant le deuxième test. Mais le meilleur score est obtenu par l'utilisation du chemin alternatif des modèles factorisés avec parties du discours : 18,21%, 18,20%.

Tableau 14. Résultats de traduction anglais - tchèque (Avramidis et Koehn, 2008)

	BLEU	***BLEU***
Set	**Devtest**	**Test**
Modèle de base	12,08%	12,34%
Personne + cas chemin alternatif : parties du discours	11,98	11,99%
Personne chemin alternatif : mot	12,23%	12,11%
Cas chemin alternatif : mot	**12,54%**	**12,51%**

Afin d'améliorer les résultats de traduction pour la paire de langues anglais - tchèque, Avramidis et Koehn (2008) utilisent une table de traduction mot-à-mot. Cette démarche est

due à la petite taille du corpus utilisé. Par la combinaison des étiquettes du cas et de la personne, les résultats de traduction sont aggravés, ce qui montre que l'application de la méthode à d'autres langues demande des utilisations différentes des attributs pour chacune d'entre elles.

Ainsi, enrichir les données source peut être utile pour la traduction de l'anglais vers le tchèque qui est une langue riche morphologiquement. Les expériences montrent des résultats améliorés par l'utilisation des facteurs du cas des noms avec un chemin alternatif.

Dans leur approche, Avramidis et Koehn (2008) montrent comment la performance des modèles factorisés peut être augmentée dans le cas des traductions à partir de l'anglais vers des langues riches de point de vue morphologique, par l'intégration des informations linguistiques dans les phrases de la langue source. Même si une langue source comme l'anglais manque d'attributs morphologiques demandés par la langue cible, ce manque peut être compensé par la syntaxe de la phrase source. En effet, à partir de l'analyse syntaxique de la phrase anglaise, il est possible d'extraire l'information du cas des noms et de la personne des verbes de la langue cible riche morphologiquement. Avramidis et Koehn (2008) montrent que ce type d'information peut être facilement intégré dans un système factorisé de traduction automatique statistique par le prétraitement du corpus de la langue source.

Dans cette sous-section (2.2.2.6.), nous avons vu comment les informations linguistiques de divers niveaux (morphologique, syntaxique) peuvent être exploitées afin d'améliorer les résultats des systèmes statistiques standard à base de séquences. Les problèmes résolus, en partie, par les systèmes factorisés sont principalement ceux liés à la dispersion des données dans les corpus d'entraînement, aux contraintes locales sur les mots ou au ré-ordonnancement. Mais un autre problème important rencontré par les systèmes à base de séquences est le traitement des expressions poly-lexicales (collocations) comme les noms composés, les expressions idiomatiques, les verbes à particule qui ne sont pas toujours traduites en langue cible. Nous présenterons dans la sous-section suivante l'approche de Ramisch *et al*. (2013) qui tentent à évaluer le degré de difficulté des systèmes statistiques standard à traduire des expressions comme les verbes à particule de l'anglais vers le français.

2.2.2.7. Modèles intégrant des collocations

La traduction des expressions poly-lexicales est l'une des difficultés rencontrées par les systèmes de traduction automatique statistique. Ces structures peuvent être des noms composés (*dry run, vacuum cleaner*), des expressions idiomatiques (*set the bar high*) et des verbes à particule (*think through, sit down*) (Ramisch *et al*., 2013). Les techniques simples employées afin d'intégrer ces structures dans les systèmes statistiques sont l'ajout des équivalents collocationnels au corpus d'entraînement ou directement dans les tables de traduction, munis de probabilités attribuées de manière artificielle (Ren *et al*., 2009 ; Bouamor *et al*., 2012).

Nous présenterons ici l'approche de Ramisch *et al*. (2013) qui évalue la difficulté des systèmes de traduction automatique statistique à base de séquences (Koehn *et al*., 2003) et hiérarchiques (Chiang, 2007) à traduire les verbes à particule anglais en français. Dans un système hiérarchique, les unités de traduction sont représentées par des arbres à la place des simples mots. Même si ces systèmes peuvent traduire des séquences comme les noms composés, par exemple, ils ne sont pas capables de traduire toutes les expressions dans leurs différents contextes. Ramisch *et al*. (2013) définissent l'expression poly-lexicale comme une combinaison d'au moins deux unités lexicales qui présentent un comportement idiosyncrasique à un certain niveau d'analyse linguistique (Baldwin et Kim, 2010).

Cette approche est représentative pour la paire de langues anglais - français car l'évaluation effectuée utilise un protocole soigneusement mis au point, pour étudier une classe de constructions plus flexible (les verbes à particule). Ce protocole se base sur une évaluation manuelle qui prend en compte tant l'adéquation que la fluence de la traduction. L'évaluation manuelle, étant basée sur l'avis expert des annotateurs, présente donc l'avantage d'être plus fiable qu'une évaluation uniquement automatique où les scores habituellement utilisés (comme le score *BLEU* (Papineni *et al*., 2002), par exemple) comparent simplement les n-grammes d'une traduction candidate et d'une autre considérée de référence. Ces scores ne fournissent donc pas de détails sur la nature des erreurs de traduction (voir la sous-section 2.2.2.4.1. pour une discussion plus détaillée concernant les limites du score *BLEU*).

En outre, l'objectif de cette étude n'est pas d'améliorer un système statistique par l'incorporation des verbes à particule mais plutôt d'évaluer la difficulté de traduire ces constructions. Cela peut aider à concevoir des modèles plus fondés linguistiquement afin de

traiter les expressions poly-lexicales par les systèmes de traduction automatique, par rapport aux méthodes existantes basées sur des heuristiques (Ramisch *et al*., 2013).

Les verbes à particule sont composés par un verbe principal (*take*) qui est combiné avec une préposition (*take on* en *take on a challenge*) ou un adverbe (*take away* en *I take away your books*). Ces constructions présentent une variabilité syntaxique et sémantique élevée.

Du point de vue syntaxique, ces verbes sont intransitifs (*the aircraft takes off*) ou transitifs (*he took off his shoes*). D'autres verbes peuvent apparaître dans les deux constructions à la fois ayant un sens lié (*the band broke up, the government broke up monopolies*) ou différent (*the aircraft takes off, he took off his shoes*). Dans leur étude, Ramisch *et al*. (2013) traitent seulement les constructions transitives divisées en deux catégories :

- verbe - particule (*put off*, *give up*) ; Dans ce cas, la particule est dépendante syntaxiquement et sémantiquement du verbe.

- verbe - préposition (*talk about*, *rely on*). La préposition dépend de l'objet et constitue un groupe prépositionnel complément d'un verbe régulier.

De plus, dans les cas d'homographie entre les particules, les prépositions et les adverbes (*up, out, in, off*), certaines constructions deviennent ambigües de point de vue syntaxique, ce qui rend difficile leur identification et traduction automatiques (*eat up [ten apples]*, *eat [up in her room]*, *eat [up to ten apples]*).

Du point de vue sémantique, les verbes à particule peuvent être décrits ainsi :

a) littéraux ou compositionnels (*take away*) ;

b) aspectuels ou semi-idiomatiques (*fix up*) ;

c) combinaisons idiomatiques (*pull off*) (Bolinger, 1971).

Dans leurs expériences, Ramisch *et al*. (2013) mettent en place des systèmes statistiques à base de séquences et hiérarchiques exploitant un même corpus parallèle. La principale différence entre les deux types de systèmes est la manière de représenter les équivalences dans le modèle de traduction. En effet, si le premier type de système exploite la séquence de mots, le deuxième utilise des grammaires indépendantes du contexte qui permettent l'intégration des variables dans la table de traduction. Par exemple, le système à base de

séquences énumère toutes les possibilités d'apparition des séquences (*make up*, *make it up*, *make the story up*, etc.), alors que le système hiérarchique peut les substituer par une seule variable (*make X up*).

Le corpus utilisé est la partie anglais - français du *TED Talks*[83] (Cettolo *et al*., 2012) qui représente une collection de transcriptions de conférences publiques sur différents sujets.

La taille du corpus est de 141 390 phrases alignées comprenant environ 2,5 millions de tokens par langue. Le corpus est segmenté lexicalement. De plus, la partie monolingue anglaise est analysée syntaxiquement.

Ramisch *et al*. (2013) repèrent également les phrases contenant des verbes à particule en anglais. Les 2 071 phrases obtenues sont utilisées dans l'entraînement des systèmes mis en place (systèmes 1 et 2) et comme ensemble de test pour l'analyse manuelle des erreurs identifiées. En effet, le système 1 utilise la première moitié de cet ensemble dans l'entraînement et la deuxième moitié pour le test, tandis que dans le système 2 ces ensembles de données sont inversées. La description des corpus d'entraînement, de développement et de test pour les deux systèmes figure dans le Tableau 15 suivant.

Tableau 15. Les corpus utilisés pour les systèmes construits (Ramisch *et al*., 2013)

Corpus	Phrases / Système 1	Phrases / Système 2
Corpus d'entraînement	137 319	137 319
Corpus d'entraînement contenant les verbes à particule	1 034	1 037
Corpus de développement	2 000	2 000
Corpus de test contenant les verbes à particules	1 037	1 034
Total	141 390	141 390

Les systèmes construits ont été entraînés par le biais de *MOSES* (Koehn *et al*., 2007), en exploitant ses paramètres par défaut. Les modèles de langue français sont des modèles 5-grammes estimés à partir de la partie monolingue française du corpus d'entraînement.

[83] https://wit3.fbk.eu/

Afin d'identifier les verbes à particule trois étapes ont été suivies. Dans un premier temps, l'outil *MWEToolkit* (Ramisch *et al*., 2010) a été appliqué afin d'extraire les phrases contenant les constructions du type *Verbe + Objet + Particule* (où le verbe représente un verbe plein, l'objet est une séquence ayant une longueur comprise entre 1 et 5 mots qui ne sont pas des verbes, la particule est une préposition ou un adverbe dépendants syntaxiquement du verbe). À l'issue de cette étape un ensemble de phrases contenant des verbes à particule a été extrait. Cet ensemble a été ensuite filtré par des heuristiques et validé manuellement.

Les phrases anglaises de test ont été traduites en français par le biais des deux systèmes de traduction automatique construits. Les traductions ont été évaluées selon un protocole soigneusement élaboré que nous présenterons brièvement ci-dessous[84] et annotées à l'aide du système *BLAST* (Stymne, 2011). Le protocole mis au point permet aux annotateurs humains de rendre compte de la qualité des traductions des verbes à particule selon deux critères comme (Ramisch *et al.*, 2013) :

- l'adéquation ; l'annotateur évalue le degré de la préservation du sens d'origine dans la traduction.
- la fluence ; l'annotateur rend compte de la qualité de la traduction au niveau grammatical sans tenir compte de son sens.

Pour évaluer l'adéquation de la traduction, la notation se situe sur une échelle de 3 à 0, comme suit :

- 3- *FULL* - les équivalents français présentent le sens d'origine ;
- 2- *PARTIAL* - le sens peut être déduit sans faire référence à la phrase anglaise ;
- 1- *NONE* - le sens n'est pas conservé dans la traduction ;
- 0 - *UNABLE TO JUDGE* - il existe un problème lié à la phrase anglaise empêchant l'annotateur de la comprendre.

[84] Le guide détaillé ainsi que les données présentées dans Ramisch *et al.* (2013) figurent à l'adresse suivante : http://cameleon.imag.fr/xwiki/bin/view/Main/Phrasal_verbs_annotation.

Afin de rendre compte de la fluence de la traduction, la notation se situe sur une échelle de 4 à 1, comme suit :

- 4 - *FLUENT* - les équivalents français ne montrent ni fautes d'orthographe ni de syntaxe ;
- 3 - *NON-NATIVE* - la forme du verbe et/ou son accord avec le sujet ou l'objet sont erronés ;
- 2 - *DISFLUENT* - les équivalents français rendent la phrase incohérente du point de vue syntaxique ;
- 1 - *INCOMPREHENSIBLE* - le verbe à particule n'est pas traduit.

La validation du protocole est effectuée en calculant un score d'accord entre les annotateurs (Artstein et Poesio, 2008) à partir de la distribution totale des catégories de traductions définies.

Ramisch *et al.* (2013) analysent les résultats d'annotation manuelle fournis par quatre annotateurs humains sur un échantillon de 750 phrases correspondant à 500 phrases source. Ils trouvent que dans la moitié des phrases les verbes à particule ont été traduits de manière similaire par les deux systèmes construits, tandis que dans l'autre moitié ces constructions ont été traduites différemment.

Tout d'abord, Ramisch *et al.* (2013) ont étudié la qualité globale de la traduction indépendamment du système qui l'a fournie. Ils ont observé ainsi que la qualité de la traduction s'avère médiocre. À titre d'illustration, Ramisch *et al.* (2013) donnent comme exemple la construction *boil down* qui signifie *reduce* ou *come down* (*résumer*) qui a été traduite littéralement comme *bouillir descendu* par le système hiérarchique et comme *furoncle jusqu'* par le système à base de séquences. Un autre exemple de traduction erronée est la séquence française *penser à travers* proposée pour la construction *think through* (*repenser*, *réfléchir*).

Les systèmes ont été aussi évalués en termes du score *BLEU*. Ainsi, le système à base de séquences a montré un score de 29,5 points, tandis que le système hiérarchique a obtenu un score moins élevé de 25,1. Ces résultats ont été comparés à ceux fournis par *Google*

Translate[85] ayant un score *BLEU* plus performant de 32,3. Les résultats obtenus sont considères toutefois acceptables pour les expériences effectuées compte tenu de la taille réduite du corpus d'entraînement (Ramisch *et al.*, 2013).

En termes d'adéquation et de fluence, les traductions ont obtenu une moyenne de 1,73 pour l'adéquation (sur une échelle de 1 à 3) et une moyenne de 2,57 pour la fluence (sur une échelle de 1 à 4). Par conséquent, environ la moitie des traductions montre des problèmes liés au sens ou à la grammaire. Les problèmes liés à l'adéquation sont légèrement plus fréquentes que ceux liés à la fluence.

Dans la Figure 18 suivante, apparaît le graphique représentant la proportion des traductions jugées comme *FULL*, *PARTIAL* et *NONE* pour l'adéquation (Ramisch *et al*., 2013 : 59).

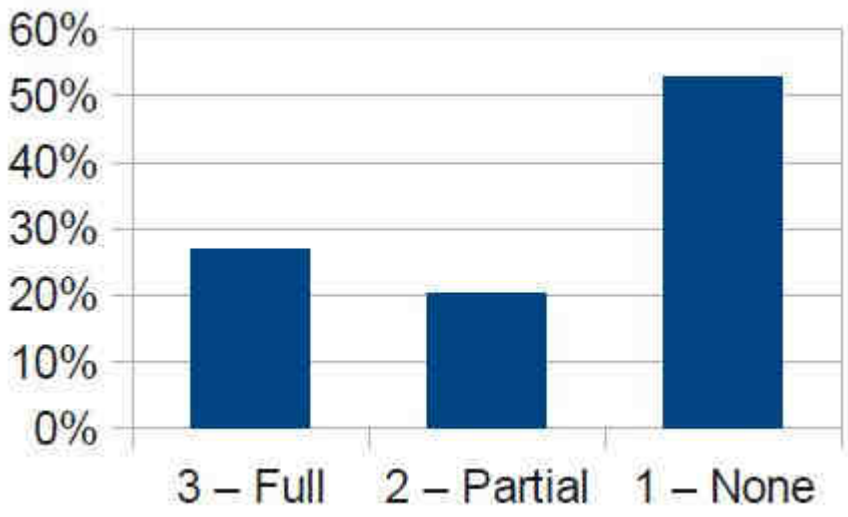


Figure 18. La proportion des traductions jugées comme *FULL*, *PARTIAL* et *NONE* pour l'adéquation (Ramisch *et al*., 2013 : 59)

Le graphique de la Figure 18 ci-dessus montre que seulement 27% des traductions des verbes à particule gardent le sens d'origine, environ 20% présentent un sens qui est partiellement conservé par rapport à la construction anglaise et 57% des traductions sont inutiles. Par conséquent, force est de constater que ces constructions ne sont pas correctement traitées par les systèmes de traduction automatique mis en place (Ramisch *et al*., 2013).

Comparativement, le système à base de séquences obtient de meilleurs résultats (2,67 pour la fluence et 1,75 pour l'adéquation) par rapport au système hiérarchique (2,46 pour la fluence et

[85] http://translate.google.com/?hl=fr#en/fr/

1,72 pour l'adéquation). Dans la Figure 19 suivante, apparaît le graphique montrant la proportion des traductions jugées comme *FULL*, *PARTIAL* et *NONE* pour l'adéquation fournie par les deux systèmes construits (Ramisch *et al*., 2013 : 60).

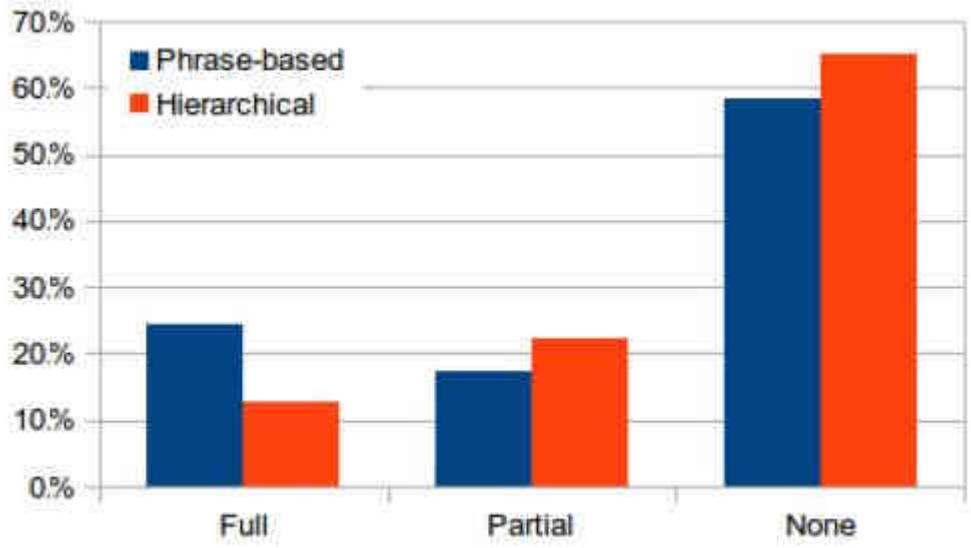


Figure 19. La proportion de différentes traductions jugées comme *FULL, PARTIAL* ou *NONE* pour l'adéquation (Ramisch *et al*., 2013 : 60)

Ramisch *et al*. (2013) ont comparé aussi la moyenne des scores obtenus par les phrases traduites de manière similaire par les deux systèmes à celle des phrases traduites différemment. Ils ont trouvé ainsi que les traductions similaires prouvent une meilleure qualité (2,82 pour la fluence et 2,07 pour l'adéquation) que les traductions différentes (2,44 pour la fluence et 1,56 pour l'adéquation). Ils considèrent que ce résultat est un trait potentiellement utile dans les modèles afin d'estimer automatiquement la qualité de la traduction.

Dans leurs expériences, Ramisch *et al*. (2013) prouvent que les systèmes statistiques standard traduisent correctement seulement 27% des verbes à particule. Ils montrent également qu'un système statistique à base de séquences obtient de meilleurs résultats par comparaison à un système hiérarchique. En outre, quand les systèmes traduisent les verbes à particule de manière similaire, la qualité de la traduction est améliorée.

Ainsi, Ramisch *et al*. (2013) montrent en premier lieu que les constructions flexibles comme les verbes à particule de l'anglais ne sont pas traitées de manière appropriée par les systèmes statistiques, même si actuellement l'approche statistique a prouvé ses performances. Par conséquent, plus de la moitié des traductions fournies révèle des problèmes de sens et/ou de grammaire. Le traitement approprié des constructions ainsi que leur traduction restent donc un

défi pour les modèles actuels de traduction automatique statistique. Un projet européen récent qui se concentre sur le traitement des constructions par des outils de *TAL* est l'*Action COST*[86] *IC 1207 - PARSEME*[87] (2013-2017) visant l'utilisation de connaissances collocationnelles dans l'analyse syntaxique (le *parsing*).

Nous venons de voir dans la sous-section 2.2.2. antérieure différents systèmes de l'approche de traduction automatique statistique à base de séquences qui donnent des résultats significativement meilleurs que les simples systèmes lexicaux. L'un des avantages des systèmes à base de séquences est l'utilisation du contexte local des mots qui résout, en partie, le problème du ré-ordonnancement des mots de la phrase cible. Cependant, cette approche rencontre encore des difficultés de ré-ordonnancement surtout quand il s'agit des réarrangements plus longs. Cet aspect est abordé seulement par l'utilisation des séquences longues qui ne sont pas toujours présentes dans le corpus d'entraînement à cause du phénomène de la dispersion des données (Crego et Marino, 2007b). Ainsi, des méthodes n-grammes bilingues ont été proposées, entre autres, afin d'améliorer le ré-ordonnancement dans les systèmes de traduction automatique statistique (Casacuberta et Vidal, 2004 ; Marino *et al*., 2006 ; Crego et Marino, 2007b ; Crego et Yvon, 2010 ; Crego *et al*., 2011 ; Le *et al*., 2012). L'approche des n-grammes bilingues sera présentée dans la sous-section suivante.

2.2.3. Systèmes de traduction automatique statistique à base de n-grammes bilingues

Le *et al*. (2012) décrivent des systèmes à base de n-grammes bilingues construits pour le français et l'allemand avec l'anglais comme langue source ou cible[88]. Le décodeur utilisé est *n*-code[89] (Crego *et al*., 2011), un système de traduction automatique statistique *open-source* qui implémente les méthodes n-grammes bilingues (Casacuberta et Vidal, 2004 ; Marino *et al*., 2006 ; Crego et Marino, 2007b). Dans cette approche, la traduction est réalisée en deux étapes : le ré-ordonnancement des mots sources suivi d'une étape de traduction monotone. La première étape exploite un ensemble de règles de réécriture apprises automatiquement qui

[86] http://www.cost.eu/about_cost
[87] *Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing* (http://www.cost.eu/domains_actions/ict/Actions/IC1207)
[88] Ces systèmes ont été présentés dans le cadre du *7th Workshop on Statistical Machine Translation (shared translation task, NAACL 2012, Montréal : http://www.statmt.org/wmt12/)* (*LIMSI - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur- WMT'2012).*
[89] http://ncode.limsi.fr/

effectue le ré-ordonnancement des mots sources. L'application de ces règles mène à la formation d'un graphe d'état fini de tous les réarrangements possibles qui est ensuite recherché afin de trouver la meilleure traduction.

Selon cette approche, à partir d'une phrase source *s* de *I* mots, la meilleure hypothèse de traduction $\hat{t}$ est définie comme la séquence de *J* mots qui maximise une combinaison linéaire de fonctions de traits (cf. formule suivante) :

$$\hat{t} = \underset{t,a}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(a,s,t) \right\} \quad \text{(a)}$$

où λ_m constitue le poids associé à une fonction de traits (h_m) et *a* représente un alignement entre les séquences sources et cibles.

La forme particulière du modèle de traduction représente l'une des principales différences entre l'approche n-grammes (Casacuberta et Vidal, 2004 ; Marino *et al*., 2006 ; Crego et Marino, 2007b) et les systèmes standard à base de séquences (Och *et al.*, 1999 ; Zens *et al.*, 2002 ; Koehn *et al*., 2003). Les modèles de traduction n-grammes sont basés sur une décomposition spécifique de la probabilité jointe P(s,t) d'une paire de phrases. Ainsi, une paire de phrases notée (s,t) est formée d'une séquence de *L* unités de traduction bilingues appelées *tuples* qui définissent une segmentation jointe de la paire $(s, t) = u_1 \ldots u_L$. Dans l'approche de Marino *et al*. (2006), cette segmentation est un « sous-produit » du ré-ordonnancement de la phrase source, qui est obtenu à partir des alignements lexicaux effectués au préalable. Ainsi, à partir de tels alignements, les *tuples* définissent une segmentation unique et monotone de chaque paire de phrases, ce qui permet d'obtenir un ensemble de séquences moins volumineux par rapport aux méthodes standard (Marino *et al*., 2006). La Figure 20 suivante (Le *et al*., 2012 : 332) montre un exemple d'une paire de phrases français - anglais (*à recevoir le prix nobel de la paix* vs. *to receive the nobel peace prize*) segmentée en *tuples*. La phrase française d'origine (*org*) a été réordonnée en fonction de l'ordre des mots cibles. Les *tuples* représentent une correspondance $u = (\bar{s}, \bar{t})$ entre une séquence source $\bar{s}$ et cible $\bar{t}$.

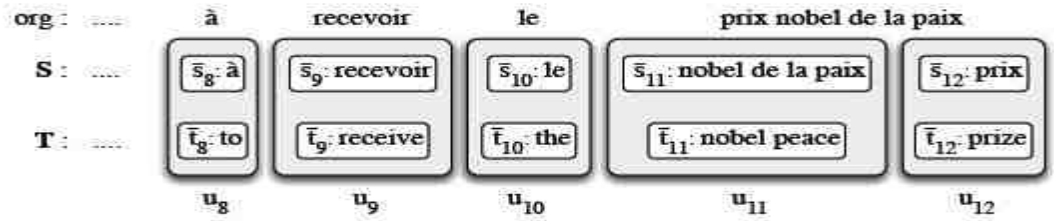


Figure 20. Exemple d'une paire de phrases français - anglais segmentée en *tuples* (Le *et al.*, 2012 : 332)

Selon la méthode n-grammes, la probabilité jointe d'une paire de phrases segmentée se décompose ainsi :

$$P(s,t) = \prod_{i=1}^{L} P\left(u_i \middle| u_{i-1},...,u_{i-n+1}\right) \quad \text{(b)}$$

Pendant l'étape d'entraînement, les *tuples* sont donc extraits à partir d'un corpus parallèle aligné lexicalement de telle sorte qu'une segmentation unique du corpus bilingue soit effectuée. Un modèle de traduction n-grammes de base est estimé ainsi à partir d'un corpus d'entraînement composés de séquences de *tuples*.

Pour optimiser les systèmes, l'algorithme *MERT* (*Minimum Error Rate Training*) (Och, 2003) qui maximise un score d'évaluation (le score *BLEU* (Papineni *et al.*, 2002)) sur un corpus de développement est utilisé.

Pendant l'étape de décodage, les phrases sources sont représentées comme des treillis de mots (graphes de mots) qui comprennent les hypothèses de ré-ordonnancement les plus prometteuses, de manière à reproduire les modifications dans l'ordre des mots qui ont été introduites lors de l'extraction des *tuples*. Le ré-ordonnancement est effectué au moyen de règles apprises automatiquement à partir des alignements lexicaux. Par exemple, dans la Figure 20, la règle [*prix nobel de la paix - nobel de la paix prix*] reproduit l'inversion des mots français observée lors de la traduction du français vers l'anglais (Le *et al.*, 2012).

Afin de généraliser de telles règles, l'information sur les parties du discours est utilisée à la place des formes de mots (Crego et Marino, 2007b). En effet, il est largement reconnu que les informations structurelles (parties du discours, *chunks*, dépendances syntaxiques) sont nécessaires pour modéliser les différences concernant l'ordre des mots d'une langue à l'autre, puisqu'elles offrent la possibilité d'apprendre des généralisations entre les langues que les

modèles simplement basés sur la forme des mots ne sont pas capables de collecter à partir de données d'apprentissage (Crego et Yvon, 2010). D'ailleurs, afin d'améliorer le ré-ordonnancement dans un système de traduction automatique statistique, Crego et Yvon (2010) ont proposé une approche estimée comme un modèle de langue n-grammes standard exploitant l'information morphosyntaxique (les parties du discours) des langues source et cible. Ce modèle dépasse les performances d'un modèle lexicalisé classique (Tillman, 2004) (cf. sous-section 2.2.2.6.1.) car, à part la robustesse qu'offre une approche n-grammes pour modéliser les langues, l'information linguistique morphosyntaxique inclue dans le modèle de ré-ordonnancement joue un rôle important dans la prédiction des différences systématiques dans l'ordre des mots entre les paires de langues. (Crego et Yvon, 2010)[90].

Toutefois, un premier problème rencontré par un modèle n-grammes réside dans le fait qu'il doit traiter un vocabulaire de base assez grand compte tenu que les unités exploitées sont des paires bilingues. Ainsi, à cause de la dispersion des données d'apprentissage, ce modèle rencontre des difficultés importantes d'estimation de paramètres. En outre, les niveaux source et cible jouent un rôle symétrique : lors du décodage le niveau source est connu et seulement le niveau cible doit être prédit (Le *et al*., 2012).

Afin de surmonter ces problèmes, la probabilité jointe de la paire de phrases P(s,t) peut être factorisée par la décomposition des *tuples* en unités sources et cibles et la prise en compte des mots comme unités de base dans le modèle de traduction n-grammes. La décomposition des séquences en mots est introduite dans ce cas seulement comme un moyen d'atténuer les problèmes d'estimation de paramètres. Soulignons que les unités de traduction restent encore les paires de séquences dérivées à partir de la segmentation bilingue des paires de phrases en *tuples* (Le *et al*., 2012).

Notons s_i^k dénotant le mot k^{th} du *tuple* source $\overline{s_i}$. À partir de l'exemple donné dans la Figure 20, s_{11}^1 représente le mot source *nobel*, s_{11}^4 dénotes le mot source *paix* et $h^{n-1}\left(t_i^k\right)$ fait référence à la séquence de n-1 mots précédant t_i^k dans la phrase cible.

[90] Une description détaillée des expériences effectuées se trouve dans les travaux de Crego et Yvon (2010).

Ainsi, dans l'exemple de la Figure 20, $h^3\left(t_{11}^2\right)$dénotes le contexte des trois mots *receive the nobel* associés au mot *peace* (t_{11}^2).

À partir des notations ci-dessus, l'équation (b) devient :

$$P\left(a, s, t\right) = \prod_{i=1}^{L}\left[\prod_{k=1}^{|\overline{t_i}|} P\left(t_i^k \middle| h^{n-1}\left(t_i^k\right), h^{n-1}\left(s_{i+1}^1\right)\right)\right.$$

$$\left.\times \prod_{k=1}^{|\overline{s_i}|} P\left(s_i^k \middle| h^{n-1}\left(t_i^1\right), h^{n-1}\left(s_i^k\right)\right)\right] \quad \text{(c)}$$

Cette nouvelle formulation présente l'avantage que les vocabulaires contiennent seulement des mots et sont, par conséquent, moins volumineux que les vocabulaires composés des *tuples*. Ces modèles sont ainsi censés mieux gérer les problèmes de la dispersion des données (Le *et al*., 2012). En outre, la formule (c) implique maintenant deux modèles : le premier terme représente le modèle de traduction, tandis que le deuxième terme est vu comme un modèle de ré-ordonnancement. Ainsi, le modèle de traduction prédit seulement la phrase cible, étant donné ses contextes sources et cibles (Le *et al*., 2012) (cf. formule ci-dessous) :

$$P\left(s, t\right) = \prod_{i=1}^{L}\left[\prod_{k=1}^{|\overline{s_i}|} P\left(s_i^k \middle| h^{n-1}\left(s_i^k\right), h^{n-1}\left(t_{i+1}^1\right)\right)\right.$$

$$\left.\times \prod_{k=1}^{|\overline{t_i}|} P\left(t_i^k \middle| h^{n-1}\left(s_i^1\right), h^{n-1}\left(t_i^k\right)\right)\right] \quad \text{(d)}$$

La difficulté majeure des modèles n-grammes basés sur les équations (c) et (d) ci-dessus réside dans l'estimation fiable de leurs paramètres dont le nombre croit de manière exponentielle avec l'ordre du modèle. Ce problème est accentué dans le traitement automatique des langues par la question de la dispersion des données (Le *et al*., 2012).

Dans leur approche, Le *et al*. (2012) utilisent l'architecture spécifique d'un réseau neuronal (*the Structured OUtput Layer* ou modèle *SOUL*) (Bengio *et al*., 2003 ; Schwenk *et al*., 2007 ; Le *et al*., 2011) qui permet de prendre en compte des vocabulaires plus larges par

rapport au modèles n-grammes standard. Cette approche peut ainsi être utilisée pour estimer des modèles n-grammes exploitant des vocabulaires volumineux tant en ce qui concerne les modèles de la langue cible que les modèles de traduction.

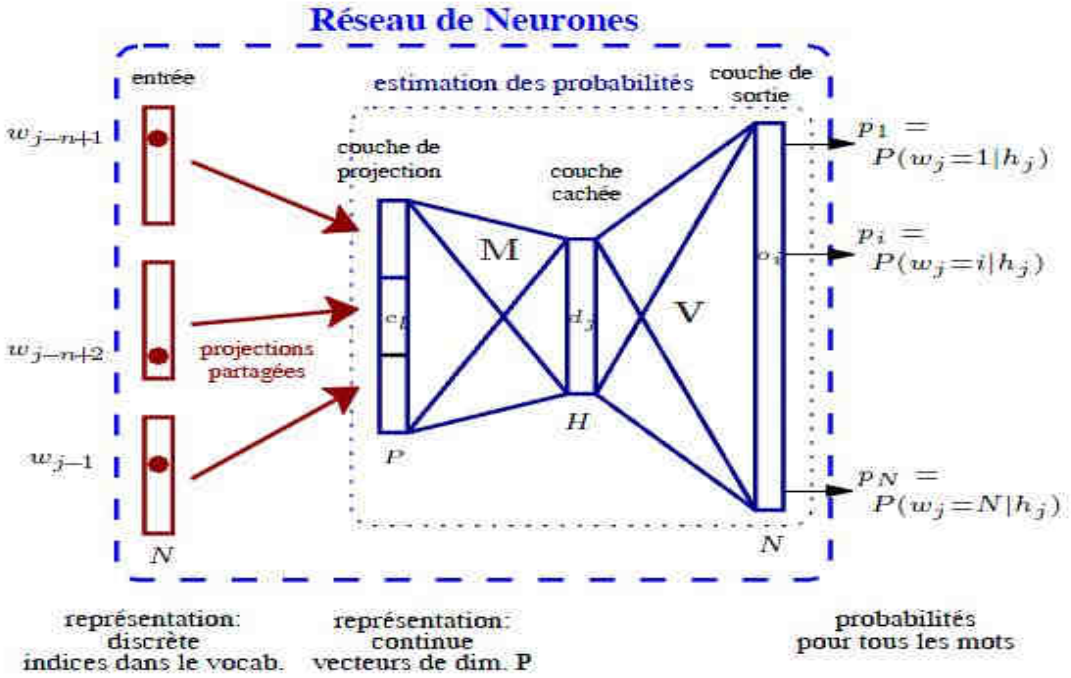
Nous présenterons ici l'architecture du réseau de neurones multi-couches (complètement connecté) telle que décrite par Schwenk *et al.* (2007). Ce réseau est utilisé afin d'apprendre conjointement la projection des mots dans un espace continu et d'estimer les probabilités n-grammes. Le réseau prend en entrée les *n-1* mots précédents du vocabulaire et fournit en sortie les probabilités a-posteriori pour tous les mots du vocabulaire (cf. formule ci-dessous) :

$$P\left(w_j = i \middle| w_{j-n+1}, \ldots, w_{j-2}, w_{j-1}\right) = P\left(w_j = i \middle| h_j\right) \qquad \forall i \in [1, N]$$

où :

- N : la taille du vocabulaire ;
- h_j : le contexte $w_{j-n+1}, \ldots, w_{j-1}$.

La Figure 21 suivante illustre l'architecture du modèle de langage neuronal (Schwenk *et al.*, 2007 : 258).



h_j dénomme le contexte $w_{j-n+1},\ldots, w_{j-1}$; P est la taille d'une projection, et H et N correspondent à la dimension de la couche cachée et de sortie, respectivement.

Figure 21. Architecture du modèle de langage neuronal (Schwenk *et al.*, 2007 : 258)

Les entrées du réseau sont projetées sur un espace continu (voir la couche *P* dans la Figure 21 antérieure). Les autres couches (de projection, cachée, de sortie) sont utiles pour calculer les probabilités de manière non-linéaire. Le réseau estime directement les probabilités de tous les mots du vocabulaire pour le même contexte. En effet, la valeur de la ième sortie correspond à la probabilité du n-gramme $P\left(w_j = i \middle| h_j\right)$ (Schwenk *et al.*, 2007).

Afin d'intégrer des modèles *SOUL* dans un système de traduction automatique statistique, Le *et al*. (2012) suivent deux étapes :

1) l'utilisation des modèles de langue et de traduction n-grammes à repli (*back-off*) classiques afin de produire la liste des *k* traductions les plus probables ;

2) le calcul des probabilités d'un modèle m-grammes *SOUL* pour chaque hypothèse et le ré-ordonnancement de la liste des *k* meilleures hypothèses. (Dans les expériences effectuées, la taille du contexte utilisée pour le modèle *SOUL* est m=10 et k=300).

Les corpus exploités disponibles pour les paires de langues traitées (*Shared Task: Machine Translation - 7th Workshop on Statistical Machine Translation, NAACL 2012)*[91] sont extraits principalement du corpus parallèle multilingue *Europarl*[92] (cf. sous-section 2.2.2.4.). Les corpus parallèles utilisés sont prétraités (segmentés lexicalement, étiquetés par des parties du discours)[93].

Dans le Tableau 16 suivant, figurent les résultats obtenus pour les systèmes de base et ceux intégrant des modèles *SOUL* développés pour les paires de langues traitées dans les deux directions du processus de traduction (Le *et al*., 2012). Les performances des systèmes sont évaluées en termes du score *BLEU* sur les corpus de test (3 003 phrases) de 2011 et 2012 distribués dans le cadre du *Shared Task : Machine Translation*.[94]

Tableau 16. Les résultats en termes du score *BLEU* pour les systèmes de base et ceux intégrant des modèles *SOUL* (Le *et al*., 2012)

Direction	**Systèmes**	**BLEU**	
		test 2011	*test 2012*
anglais - français	Système de base	32,0	28,9
	+ *SOUL*	33,4	29,9
français - anglais	Système de base	30,2	30,4
	+ *SOUL*	31,1	31,5
anglais - allemand	Système de base	15,4	16,0
	+ *SOUL*	16,6	17,0
allemand - anglais	Système de base	21,8	22,9
	+ *SOUL*	22,8	23,9

Les systèmes de base sont construits par l'utilisation des modèles n-grammes standard (*back-off*) tant pour les modèles monolingues de la langue cible que pour ceux bilingues. Dans les méthodes n-grammes standard l'ordre est limité à n=4. Le *et al*. (2012) notent que le système

[91] Les corpus disponibles se trouvent à l'adresse suivante : http://www.statmt.org/wmt12/translation-task.html.
[92] http://www.statmt.org/europarl/
[93] Plus de détails sur ces corpus et leur prétraitement figurent dans les travaux de Le *et al*. (2012).
[94] http://www.statmt.org/wmt12/translation-task.html

de base français - anglais construit avec *n*-code (Crego *et al*., 2011) a obtenu, par exemple, un score *BLEU* amélioré de 0,5 points par rapport à un système entraîné à l'aide de *MOSES* (Koehn *et al*., 2007) avec la même configuration de données, dans les deux directions de traduction.

L'intégration des modèles de la langue cible et de traduction *SOUL* a amélioré les résultats d'un point *BLEU* par rapport aux systèmes de base pour les deux paires de langues traitées, dans les deux sens du processus de traduction (cf. Tableau 16).

2.3. Bilan du chapitre

Dans ce chapitre, nous avons présenté les deux approches importantes de la traduction automatique - l'experte et l'empirique -, dans leurs contextes historiques, afin de pouvoir révéler les diverses méthodes qu'elles mettent en œuvre ainsi que leurs avantages et leurs inconvénients pour différentes paires de langues.

L'approche experte est basée sur des connaissances linguistiques apportées par des experts humains, tandis que l'empirique tire les connaissances exploitées à partir de grosses collections de textes bilingues ou multilingues appelés des corpus parallèles. D'un côté, l'approche experte concerne trois méthodes principales : la traduction automatique directe, le transfert et l'interlangue. D'un autre côté, l'approche empirique comprend les méthodes à base d'exemples et les méthodes statistiques.

Du point de vue historique, les systèmes développés jusqu'au rapport *ALPAC* (qui montrait les limites de la traduction automatique et mettait ainsi fin en 1966 aux recherches dans le domaine presque partout dans le monde) ont été généralement construits sur la méthode directe de la traduction (Hutchins, 1994). Il s'agit d'une méthode de traduction mot-à-mot qui exploite des règles basiques et un dictionnaire bilingue.

L'avantage de cette méthode est qu'elle s'avère aisément implémentable en requérant peu de ressources et de règles de traduction basiques. Celle-ci peut être efficace pour des traductions littérales ou dans des domaines limités comme la météorologie (le système *TAUM-METEO* - Université de Montréal). En revanche, cette méthode rencontre des problèmes importants dans le cas de la traduction oblique ou le transfert lexical est réalisé plutôt par groupes de mots. Par conséquent, la traduction directe n'est pas capable de fournir des équivalents pour les

séquences plus ou moins compositionnelles et le ré-ordonnancement s'avère généralement incorrect dans l'absence des connaissances linguistiques explicites liées à la phrase cible. Ces inconvénients majeurs font qu'aujourd'hui cette approche n'est plus guère implémentée.

Tandis que la traduction directe était la plus utilisée jusqu'au rapport *ALPAC*, l'approche qui a dominé ensuite jusqu'à la fin des années quatre-vingts a été celle à base de règles linguistiques (d'analyse syntaxique, lexicales, de désambiguïsation, de transformation des arbres syntaxiques, etc.) (Hutchins, 1994). Cette approche (appelée aussi *indirecte*) comprend les systèmes par transfert et les systèmes par langue pivot (ou l'interlangue).

Le transfert utilise des dictionnaires et des règles linguistiques monolingues et bilingues pour assurer la traduction d'un texte source vers la langue cible. Comme il s'agit d'une méthode modulaire, celle-ci présente l'avantage d'être assez facilement adaptable à d'autres paires de langues par la construction des grammaires et des règles de transfert appropriées. Les résultats fournis s'avèrent performants, car ces systèmes utilisent des connaissances linguistiques avancées. Toutefois, les systèmes par transfert montrent l'inconvénient d'être très coûteux en temps et en personnels car, pour construire les ressources linguistiques requises par ces systèmes, il faut faire appel à des spécialistes pour chaque couple de langues considérées. En outre, la description d'une grammaire exhaustive nécessaire pour chaque langue représente une tâche fastidieuse demandant un temps considérable.

Des systèmes industrialisés qui utilisent la méthode du transfert et qui ont survécu à la crise provoquée par les conclusions du rapport *ALPAC* sont les suivants : *Systran* (société *Systran*), *Logos* (société *Logos*), *METAL* (Université de Texas), etc. D'autres systèmes faisant partie des rescapés de l'histoire de la traduction automatique mais n'ayant pas parvenu au stade commercial sont *SUSY* (Maas, 1977) ou *Ariane* (Boitet *et al.*, 1982). Ceux-ci ont été appuyés par de grands centres de recherche, n'étant principalement pas préoccupés par les enjeux économiques de la traduction automatique, comme le *CNRS* (Léon, 2002)[95]. Un autre survivant à la crise générale, soutenu par des raisons politiques et développé pour la communauté européenne, est le projet *Eurotra* (King, 1981) (1977-1994). Notons que celui-ci a connu une grande envergure en termes du nombre de personnels impliqués mais aussi des coûts et de l'étendu géographique des différents centres de recherches y faisant partie. Même si le projet *Eurotra* n'a pas abouti finalement à une industrialisation, celui-ci a initié et promu

[95] http://histoire-cnrs.revues.org/3461#tocto1n1

des développements linguistiques et computationnels de base qui se sont avérés importants pour le développement d'un système de traduction automatique multilingue (Hutchins et Somers, 1992).

La deuxième méthode de l'approche experte - l'interlangue ou la langue pivot - a constitué la tendance de la recherche en traduction automatique dans les années 1980 et 1990 (Koehn, 2010). La méthode utilise une langue intermédiaire (neutre) pour traduire un texte source dans la langue cible. Cette langue intermédiaire représente le sens des textes utilisés. Il s'agit donc d'une représentation des textes au niveau conceptuel.

Même si l'interlangue se montre utile pour la traduction automatique car elle vise le transfert du sens à partir de la langue source vers la langue cible, ce qui constitue en fait le vrai but de la traduction automatique, plutôt que de mettre simplement en correspondance des unités de traduction, cette approche s'avère compliquée quant à la définition de tous les concepts d'une langue. En effet, il est difficile d'identifier tous les concepts possibles des langues, car ceux-ci peuvent présenter des différences importantes d'une langue à l'autre. Cependant, si l'interlangue a été correctement construite pour un domaine donné, cette approche montre l'avantage d'être plus facilement implémentable pour d'autres paires de langues que la méthode par transfert. Cela est dû au fait que l'interlangue demande seulement les grammaires de la langue source et cible, par rapport au transfert qui requiert aussi un ensemble de règles de transfert bilingues.

Des exemples de systèmes exploitant l'approche interlangue sont les suivants : *CATALYST* (Université de Carnegie Mellon), *Pangloss* (Hovy, 1993), *VERBMOBIL* (Schulze-Furhoff et Abbou, 1992), etc.

Dans un contexte où, d'une part, les Japonais montaient en puissance et comptaient sur une société de l'information montrant un besoin accru des traductions à l'échelle mondiale et où, d'autre part, le projet *Eurotra* prenait fin en Europe, révélant ainsi les limites de la traduction automatique, les regards se sont tournés vers les systèmes commerciaux d'aide à la traduction. Ceux-ci exploitent, entre autres, des mémoires de traduction, autrement dit des ensembles de traductions déjà effectuées par des humains, dont les phrases sont mises en correspondance pour leur utilisation ultérieure.

Les efforts considérables pour construire des mémoires de traduction ont mené ainsi, le long du temps, à la mise en place des approches empiriques - basées sur l'exemple (ou par analogie) et statistiques - utilisant ce genre de données textuelles. En effet, cette idée de réutiliser les traductions fiables (effectuées au préalable par des traducteurs humains) paraît encourageante pour développer les recherches dans cette direction.

Ainsi, ces méthodes exploitent des corpus parallèles de taille importante afin d'en extraire, à l'aide des calculs statistiques, des exemples de traduction ou des séquences bilingues utilisées ensuite dans le processus de traduction. Les deux approches se différencient principalement par le fait que la traduction statistique exploite les résultats fournis par un corpus aligné (un ensemble d'exemples compilés), pendant que la traduction par analogie utilise les exemples directement pendant le processus de traduction (Boitet, 2008).

Par rapport aux systèmes experts coûteux en temps et en ressources humaines, les systèmes par analogie (Nagao, 1984 ; Kaji *et al.*, 1992 ; Utsuro *et al.*, 1992 ; Brown, 1996 ; Veale et Way, 1997 ; Alp et Turhan, 2008 ; Nakazawa *et al.*, 2006 ; Langlais et Gotti, 2006 ; Lepage et Denoual, 2005 ; Irimia, 2008 ; Gavrilă, 2012) présentent l'avantage de s'améliorer facilement dès que de nouveaux corpus deviennent disponibles. Ce fait est néanmoins un inconvénient pour les langues moins dotées en ressources langagières. En outre, comme les traductions exploitées sont réalisées par des humains, les systèmes à base d'exemples montrent aussi l'avantage que la qualité des traductions fournies s'avère au moins comparable à celle des traductions effectuées par les systèmes expertes, en utilisant des moyens moins coûteux.

Les systèmes statistiques ont pris de l'ampleur entre autres facteurs grâce au développement d'ordinateurs puissants en termes de vitesse et de capacités de stockage ainsi que grâce à la généralisation du *World Wide Web* à partir des années 2000, qui a permis la mise à disposition de bases de données textuelles pour différentes paires de langues. Ainsi, sur ce terrain propice, cette approche nécessitant de grandes quantités de données textuelles s'est développée dans la dernière décennie.

Les premiers systèmes de traduction automatique statistique ont vu le jour au *IBM* (Brown *et al.*, 1988 ; Brown *et al.*, 1990 ; Brown *et al.*, 1993) et exploitaient des alignements lexicaux. Ces systèmes ont encouragé les recherches ayant posé les fondements des développements ultérieurs dans le domaine. Par l'utilisation des alignements au niveau du mot, ces systèmes s'avèrent néanmoins incapables de prendre en compte les dépendances entre les mots ou les

séquences. Comme il s'agit plutôt d'une traduction du type mot-à-mot, ces méthodes rencontrent également des difficultés pour établir l'ordre correct des mots de la phrase cible. En outre, ces systèmes n'arrivent pas à traduire correctement des structures plus ou moins compositionnelles (noms composés, expressions).

Toutefois, ces approches ont ouvert la voix aux systèmes de traduction automatique statistique à base de séquences (Och *et al.*, 1999 ; Zens *et al.*, 2002 ; Zhang *et al.*, 2003 ; Koehn *et al.*, 2003) qui obtiennent des performances significativement meilleures que les systèmes lexicaux. Ces méthodes n'utilisent plus le mot comme unité de traduction mais plutôt la séquence de mots qui rend ces systèmes capables de mieux gérer le ré-ordonnancement local au niveau de la phrase cible. Cela est dû au fait que l'utilisation des séquences permet l'exploitation du contexte local des mots. De plus, la traduction de structures plus ou moins compositionnelles devient possible même si elle n'est pas, bien évidemment, entièrement résolue.

Plusieurs approches tentent d'extraire des séquences à partir des corpus parallèles. Certaines méthodes (Och *et al.*, 1999 ; Koehn *et al.*, 2003) exploitent des alignements lexicaux effectués dans les deux sens du processus d'alignement et, ensuite, ils mettent en œuvre des heuristiques, comme l'intersection ou l'union, pour obtenir les séquences de traduction. D'une part, Koehn *et al*. (2003) comparent la méthode utilisant des alignements lexicaux avec le modèle de probabilité jointe qui génère simultanément les segments source et cible à partir d'un corpus parallèle (Marcu et Wong, 2002). D'autre part, ils font aussi une comparaison avec la méthode qui extrait des séquences fondées syntaxiquement (Wu, 1997 ; Yamada et Knight, 2001 ; Imamura, 2002).

Koehn *et al.* (2003) montrent que le modèle utilisant des alignements lexicaux (Och *et al.*, 1999) obtient de meilleurs résultats par rapport au modèle des séquences jointes (Marcu et Wong, 2002). En outre, la seule utilisation des segments basés sur la syntaxe (Wu, 1997 ; Yamada et Knight, 2001 ; Imamura, 2002) s'avère nuisible au système car cette méthode mène à l'élimination d'une quantité considérable de paires de séquences. Ainsi, ils en concluent qu'aucune des trois méthodes comparées ne contribuent de manière significative à la qualité de la traduction. En revanche, les résultats obtenus par les modèles à base de séquences sont meilleurs par rapport à ceux montrés par les modèles lexicaux (Brown *et al.*, 1993).

Comme pour les méthodes à base d'exemples, les systèmes statistiques ont l'avantage de s'améliorer facilement au moment où de nouveaux corpus parallèles sont disponibles. Ces systèmes désavantagent toutefois les langues moins dotées en ressources langagières. Ainsi, les performances de ces systèmes sont dépendantes du volume des données utilisées.

Afin d'améliorer les résultats des systèmes statistiques à base de séquences, il est possible d'utiliser des ressources linguistiques externes comme les dictionnaires, les bases de données terminologiques, les lexiques, etc. Un système qui intègre un dictionnaire de terminologie bilingue pour améliorer les résultats d'un système anglais - français est *Portage* (Sadat *et al.*, 2006). Le dictionnaire exploité est le *Grand Dictionnaire Terminologique* (*GDT*, L'Office québécois de la langue française).

Cependant, d'un côté, ce genre de ressources linguistiques n'est pas disponible pour certaines paires de langues moins dotées, comme c'est le cas également de la paire français - roumain. De ce fait, l'entraînement des systèmes de traduction automatique statistique intégrant des ressources linguistiques telles que dictionnaires, bases de données terminologiques, etc., s'avère difficile pour ces langues. D'un autre côté, les modèles statistiques standard à base de séquences incorporant ou non des ressources externes n'intègrent pas d'informations linguistiques dans les modèles de traduction et de langue eux-mêmes. Ce fait rend ces systèmes moins performants par comparaison avec les approches plus récentes qui utilisent des techniques de factorisation. Ces techniques permettent d'exploiter de facteurs linguistiques de divers niveaux (morphologique, syntaxique) afin d'améliorer la qualité de la traduction.

Dans les systèmes de traduction automatique factorisés (Koehn et Hoang, 2007 ; Birch *et al.*, 2007 ; Tufiș *et al.*, 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011 ; Tufiș et Dumitrescu, 2012), le mot n'est plus seulement un token (comme il est considéré dans les approches de traduction automatique statistique standard), mais un vecteur de facteurs linguistiques, tels que : formes de mots, lemmes, parties du discours, étiquettes morphosyntaxiques ou syntaxiques, classes de mots, etc. Les expériences menées dans ce sens ont montré que l'intégration du niveau morphologique riche au processus de traduction (Koehn et Hoang, 2007) améliore nettement les résultats de la traduction, par rapport à un système classique. En effet, en connaissant le lemme et les propriétés morphosyntaxiques d'un mot, par exemple, le système factorisé est capable de générer en sortie la forme fléchie correcte d'un mot. Cette technique permet de traduire des mots inconnus ou moins représentatifs dans le corpus d'apprentissage et réduit

ainsi les problèmes liés au phénomène bien connu de la dispersion des données. La dispersion fait référence aux distributions différentes des formes fléchies dans un corpus. Ce phénomène est beaucoup plus important pour les langues riches du point de vue morphologique (tchèque, allemand, français, roumain, etc.), car celles-ci présentent un nombre important de formes fléchies pour un seul mot. Les occurrences moins fréquentes ont ainsi des probabilités de traduction plus faibles. De ce fait, si l'on regroupe toutes les formes d'un mot sous un même lemme, on augmente sa probabilité de traduction.

Ceauşu et Tufiş (2011) améliorent les résultats d'un système de base anglais - roumain par l'utilisation, entre autres, des facteurs comme les lemmes et les propriétés morphosyntaxiques des unités lexicales, à partir initialement d'un corpus de taille relativement limitée. En outre, par l'utilisation d'un corpus plus large, ils montrent que les résultats peuvent être encore améliorés quand un corpus monolingue supplémentaire de la langue cible est aussi disponible.

Pour améliorer encore plus les résultats de la traduction factorisée anglais - roumain, Tufiş et Dumitrscu (2012) décrivent une méthode de traduction appelée *en cascade*. Cette technique réutilise les sorties d'un premier système factorisé comme données d'entrée d'un deuxième système qui est capable de corriger ainsi certaines erreurs des traductions initiales. Ils obtiennent ainsi de très bons scores statistiques pour la paire de langues traitées.

Birch *et al.* (2007) ont combiné le niveau morphologique avec le niveau syntaxique de la phrase cible, en obtenant des résultats plus performants par rapport aux modèles standard et aux modèles factorisés utilisant seulement la morphologie. Les approches de Birch *et al.* (2007), Avramidis et Koehn (2008) ont exploité aussi l'intégration de l'information syntaxique dans la phrase source qui permet l'extraction de l'information linguistique demandée par la phrase cible et, par conséquent, la génération des formes fléchies correctes des mots cibles. Les résultats ont été améliorés en utilisant des systèmes factorisés combinés. Ainsi, l'enjeu des modèles factorisés est de voir, au niveau des expériences, quels facteurs linguistiques et comment ces facteurs peuvent être combinés et intégrés dans le processus de traduction, afin d'améliorer la qualité des résultats pour différentes paires de langues, riches ou moins riches de point de vue morphologique.

Les systèmes factorisés se montrent ainsi performants puisqu'ils présentent l'avantage de moduler les ressources linguistiques utilisées. Comme les contraintes locales sur les mots apparaissent au niveau morphologique, l'incorporation de ce niveau au processus de

traduction améliore les résultats d'un système de base de manière significative. En outre, l'information syntaxique incorporée dans la phrase source rend possible l'extraction de l'information linguistique demandée au niveau de la phrase cible (p. ex. la catégorie grammaticale du cas des noms). Ainsi, le système peut générer la forme fléchie correcte du mot. De plus, par l'utilisation de ces techniques mixtes, les systèmes factorisés fournissent des résultats comparables aux méthodes expertes, en s'avérant moins coûteux en temps et en ressources humaines, pour différentes paires de langues. Cependant, l'inconvénient des systèmes factorisés réside dans le fait que leurs résultats dépendent des paires de langues considérées et aussi du sens de la traduction.

Les problèmes qui sont partiellement résolus par les systèmes factorisés sont principalement ceux concernant la dispersion des données dans les corpus d'entraînement, les contraintes locales sur les mots ou le ré-ordonnancement au niveau de la phrase cible. Mais un autre problème important rencontré par les systèmes à base de séquences est le traitement des expressions poly-lexicales qui n'apparaissent pas toujours traduites en langue cible (Ramisch *et al*., 2013).

Ainsi, Ramisch *et al*. (2013) tentent d'évaluer la difficulté des systèmes statistiques standard à traduire des expressions comme les verbes à particule anglais en français, par la mise en place d'un protocole d'évaluation soigneusement élaboré. Ce protocole s'avère fiable car il est basé sur une évaluation manuelle des résultats. Ils montrent que ces constructions ne sont pas traitées de manière appropriée par les systèmes statistiques, malgré le fait qu'aujourd'hui cette approche a prouvé ses performances. En conséquence, plus de la moitié des traductions révèle des problèmes de sens et/ou de grammaire. Ainsi, le traitement approprié des constructions (collocations) et leur traduction constituent encore un défi pour les modèles actuels de traduction automatique statistique. Un projet européen récent qui vise le traitement des collocations par les applications de *TAL* est l'*Action COST IC 1207 - PARSEME*[96] (2013-2017). Ce projet exploite des connaissances collocationnelles dans l'analyse syntaxique.

Soulignons que parmi les avantages des systèmes à base de séquences se trouve l'utilisation du contexte local des mots censée résoudre le problème du ré-ordonnancement au niveau de la phrase cible. Toutefois, ce problème n'est pas entièrement résolu, surtout quand il s'agit des réarrangements plus longs. Cet aspect est abordé seulement par l'exploitation des séquences

[96] http://www.cost.eu/domains_actions/ict/Actions/IC1207

longues qui ne sont pas toujours présentes dans le corpus d'entraînement à cause du phénomène de la dispersion des données (Crego et Marino, 2007b).

Afin d'améliorer le ré-ordonnancement dans les systèmes statistiques, des méthodes n-grammes bilingues ont été proposées entre autres (Casacuberta et Vidal, 2004 ; Marino *et al*., 2006 ; Crego et Marino, 2007b ; Crego et Yvon, 2010 ; Crego *et al*., 2011 ; Le *et al*., 2012). Dans cette approche, le ré-ordonnancement est réalisé au niveau de la langue source par l'utilisation des règles apprises automatiquement, observées dans la traduction à partir de la langue source vers la langue cible. Ces règles peuvent exploitées des informations linguistiques telles que les parties du discours qui permettent d'apprendre des généralisations entre les langues (Crego et Yvon, 2010), par rapport aux simples formes de mots qui ne le font pas.

Dans leurs expériences de traduction français - anglais, Le *et al*. (2012) notent que le système de base utilisant de n-grammes bilingues (entraîné avec *n*-code (Crego *et al*., 2011)) a amélioré les résultats d'un système à base de séquences (entraîné avec *MOSES* (Koehn *et al*., 2007)) d'un demi point *BLEU*, à partir des mêmes données. Ils présentent également des expériences anglais - français et anglais - allemand où les méthodes n-grammes bilingues sont améliorées par l'utilisation des modèles de traduction et de langue appelés *SOUL* (Schwenk *et al*., 2007). Ces modèles utilisent l'architecture spécifique du réseau de neurones multi-couches qui permet le traitement des vocabulaires plus larges, par rapport au modèles n-grammes standard.

De notre côté, nous avons adopté l'approche factorisée de la traduction automatique (Koehn et Hoang, 2007 ; Birch *et al*., 2007 ; Tufiș *et al*., 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011) car, comme nous l'avons vu dans ce chapitre, il s'agit de l'approche appropriée pour des langues moins dotées en ressources langagières et riches morphologiquement, comme c'est le cas du français et du roumain.

En effet, il n'existe pas actuellement, à notre connaissance, de ressources linguistiques considérables comme les dictionnaires électroniques, les bases de données terminologiques, les lexiques, etc., qui puissent être intégrées dans un système de traduction automatique statistique français - roumain (comme dans le système *Portage* (Sadat *et al*., 2006) qui exploite un dictionnaire de terminologie bilingue afin d'améliorer les résultats d'un système purement statistique français - anglais). Le seul dictionnaire dont nous avons disposé est le

dictionnaire d'expressions poly-lexicales (collocations) *Verbe + Nom* (Todirașcu *et al*., 2008) que nous avons tenté d'exploiter pendant nos expériences d'alignement lexical (cf. chapitre 4, section 4.4.), car, comme ils l'ont montré aussi Ramisch *et al*. (2013), les constructions représentent encore un défi pour les systèmes statistiques de traduction automatique. Néanmoins, ce dictionnaire est limité en taille et il incorpore une seule classe de collocations (cf. chapitre 4, section 4.4.). En outre, les corpus parallèles disponibles pour cette paire de langues sont aussi peu nombreux (voir chapitre suivant, section 3.2.). Notons aussi que des projets européens de traduction automatique plus récents comme *LetsMT!*[97], *ACCURAT*[98] ou *TTC*[99] ne proposent pas non plus de ressources pour la paire de langues étudiées.

Comme nous l'avons vu dans ce chapitre, par l'utilisation des techniques de factorisation, il est possible d'atténuer le problème du déficit des données dans les corpus d'entraînement. En effet, en prenant en compte les lemmes et les propriétés morphosyntaxiques dans le processus de traduction, par exemple, le système est capable de générer en sortie les formes fléchies correctes des mots cibles. Donc, ce type de système aide au choix de la variante morphologique correcte en langue cible. Celui-ci peut traduire ainsi des mots inconnus dans le système ou moins représentatifs dans le corpus d'apprentissage (Koehn et Hoang, 2007). Ces techniques résolvent donc, en partie, le problème de la dispersion des données qui est beaucoup plus intense dans le cas des langues riches du point de vue morphologique.

Dans le cadre de cette étude, nous avons utilisé seulement le niveau morphologique riche (Koehn et Hoang, 2007 ; Tufiș *et al.*, 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011) dans les systèmes factorisés mis en place. Nous n'avons pas intégré, pour le moment, d'informations syntaxiques (Birch *et al.*, 2007 ; Avramidis et Koehn, 2008), car nous n'avons pas disposé initialement d'analyseurs syntaxiques pour la paire de langues étudiées. Mais cela représente l'une des perspectives de ce travail. De plus, l'utilisation des systèmes factorisés *en cascade* (Tufiș et Dumitrescu, 2012) constitue aussi une ouverture envisageable pour l'amélioration ultérieure des résultats.

Soulignons aussi que nous ne nous sommes pas concentrés, pour le moment, sur l'amélioration du ré-ordonnancement (Crego et Marino, 2007b ; Crego et Yvon, 2010 ; Le *et al*., 2012), car l'ordre des mots est généralement libre en roumain et plus fixe en français.

[97] http://project.letsmt.eu/
[98] http://www.accurat-project.eu/
[99] *Terminology Extraction, Translation Tools and Comparable Corpora* (http://www.ttc-project.eu/)

Néanmoins, les deux langues présentent un ordre de mots plus flexible que l'anglais, par exemple, qui montre un ordre standard (Sujet - Verbe - Complément). Nous avons donc donné la priorité au prétraitement spécifique des corpus pour les langues riches morphologiquement (lemmatisation, désambigüisation morphologique des lemmes (Ceauşu, 2009), annotation morphosyntaxique *stratifiée* Tufiş (1999, 2000), etc.) (cf. chapitre suivant) et à l'alignement lexical qui, comme nous l'avons vu dans ce chapitre, joue aussi un rôle important dans la qualité de la traduction.

Ainsi, nous avons tout d'abord prétraité les corpus utilisés (cf. chapitre suivant). Ensuite, nous nous sommes concentrés sur le développement d'un système d'alignement lexical car, à notre connaissance, il n'existe pas de systèmes spécialement conçus pour la paire de langues étudiées. De plus, la qualité de l'alignement influence aussi les résultats de traduction. De ce fait, nous avons développé un système hybride combinant des techniques statistiques et des informations linguistiques (lemmes, propriétés morphosyntaxiques) afin d'améliorer les résultats d'un système de base purement statistique. Ainsi, notre démarche est d'étudier aussi l'influence des informations linguistiques tant dans le processus d'alignement lexical (cf. chapitre 4) que dans celui de traduction (cf. chapitre 5).

Dans le chapitre suivant, seront présentés l'architecture du système de traduction automatique factorisé mis en place (cf. section 3.1.), ainsi que les corpus disponibles pour la paire de langues traitées et leur prétraitement spécifique (cf. section 3.2.).

3. Le système de traduction automatique statistique factorisée français - roumain

Notre étude se situe dans la lignée des travaux du projet *SEE-ERA.net* (Tufiș *et al.*, 2008b) ayant pour objectif la construction de systèmes de traduction automatique statistique factorisée pour des langues slaves et balkaniques (bulgare, grec, serbe, slovène), parmi lesquelles nous retrouvons également le roumain. Les systèmes développés utilisent l'anglais comme langue source ou cible. Ces systèmes exploitent des corpus parallèles lemmatisés, étiquetés (par parties de discours et propriétés morphosyntaxiques) et alignés aux niveaux propositionnel et lexical, mais aussi une combinaison de facteurs linguistiques tels que les formes des mots, les lemmes, les parties du discours ou les propriétés morphosyntaxiques. En outre, nous avons élaboré une méthodologie linguistique contrastive pour l'alignement lexical qui nous a permis d'adapter le système à la paire de langues étudiées.

Nous avons ainsi développé un système de traduction automatique statistique factorisée pour la paire de langues français - roumain. Ce système a été implémenté initialement pour l'anglais et le roumain[100] (Ceaușu, 2009) à l'*Institut de Recherche en Intelligence Artificielle*[101] de l'Académie Roumaine de Bucarest.

Le décodeur utilisé est *MOSES* (Koehn *et al.*, 2007) (cf. sous-section 3.1.1.), avec différentes configurations de paramètres linguistiques optimisés (lemmes et descriptions morphosyntaxiques) établies en fonction du sens du processus de traduction. Ce décodeur a une popularité élevée dans le domaine de la traduction automatique statistique, car il présente l'avantage d'être un système open-source. Les paramètres du décodeur sont appris en utilisant le corpus parallèle anglais - roumain *SEEERA.NET* (Tufiș *et al.*, 2008b) basé sur l'*Acquis Communautaire*[102] et comprenant environ 60 000 paires de phrases alignées. Ce corpus, de

[100] http://www.racai.ro/webservices/FactoredTranslation.aspx
[101] http://www.racai.ro/Home/TabId/36/Default.aspx
[102] Le corpus parallèle *Acquis Communautaire* sera décrit en détail plus loin, dans ce chapitre.

taille beaucoup plus réduite que le corpus entier disponible pour la paire de langues anglais - roumain (800 000 paires de phrases), est utilisé pour optimiser le temps de calcul[103].

Les paramètres du décodeur sont optimisés grâce à l'utilisation de l'application *MERT* (Bertoldi *et al.*, 2009) sur 200 paires de phrases, n'étant pas présentes dans le corpus d'entraînement. Cette application fait partie de la distribution de *MOSES* (cf. sous-section 3.1.1.). Pour calculer le score d'évaluation automatique *BLEU* (Papineni *et al.*, 2002), Ceaușu (2009) utilise 400 paires de phrases. Ces phrases ne figurent pas non plus dans le corpus d'entraînement.

À partir des paramètres ainsi établis, le système final anglais - roumain (Ceaușu, 2009) est entraîné sur l'ensemble du corpus d'apprentissage (environ 800 000 paires de phrases). Celui-ci a obtenu des scores *BLEU* performants : 43,29 pour le système anglais -> roumain et 51,02 pour le système roumain -> anglais. Le corpus total d'entraînement (environ 30 millions de mots) pour ce système anglais - roumain est constitué des corpus suivants (Ceaușu, 2009) :

- *JRC-Acquis-Ro* (Ceaușu, 2008) basé sur *JRC-Acquis* (Steinberger *et al.*, 2006) et notamment sur 6 085 documents communs en roumain et en anglais dont les différences de longueur sont minimales ; Ce corpus comprend environ 13 millions de mots par langue.

- *SEEERA.NET* anglais - roumain (Tufiș *et al*., 2008b) basé aussi sur *JRC-Acquis* mais dont l'annotation (segmentation lexicale, lemmatisation, étiquetage morphosyntaxique) a été corrigée par des experts humains ; Ce corpus contient environ 1 million et demi de mots par langue.

- *NAACL* (Martin *et al.*, 2005), corpus journalistique anglais - roumain utilisé pendant la compétition d'alignement de *NAACL* (*Association of Computational Linguistics, North American Chapter*) 2005. La taille du corpus *NAACL* est d'approximativement 800 000 mots par langue.

De plus, le corpus d'entraînement intègre 52 774 paires de définitions fournies par l'ontologie lexicale *Ro-Wordnet* et les listes d'équivalents de traduction sont enrichies par 121 176

[103] Si le corpus entier avait été utilisé pour tester différentes configurations du système, la manipulation aurait duré plusieurs dizaines de jours avec un ordinateur *Intel Quad-Core Xeon* 5420 2.5 GHz, 4 Gb de mémoire (Ceauşu, 2009).

équivalents extraits de la même ontologie. L'évaluation est faite sur 1 000 phrases et l'optimisation du système sur 400 phrases. Ces échantillons ne font pas partie du corpus d'entraînement.

Les résultats obtenus ont été comparés à ceux fournis par *Google Translate*[104] : le score *BLEU* de 39,01 pour la direction de traduction anglais -> roumain et de 40,27 en sens inverse. Il est évident que le système factorisé anglais - roumain (Ceaușu, 2009) a obtenu des résultats significativement meilleurs que le système purement statistique à base de séquences *Google Translate*. Cela montre que l'intégration des informations linguistiques dans le processus de traduction a une importance cruciale afin d'améliorer la performance des systèmes statistiques pures. Les bons résultats du système factorisé anglais - roumain (Ceaușu, 2009) ont également été influencés par l'exploitation de l'ontologie lexicale *Ro-Wordnet*.

Quant à notre système factorisé français - roumain, celui-ci exploite les corpus parallèles disponibles pour la paire de langues étudiées (cf, section 3.2.) et le décodeur *MOSES* (Koehn *et al.*, 2007) avec les autres instruments qui font partie de sa distribution (cf. sous-section 3.1.1.). L'architecture du système construit ainsi que les corpus intégrés et leur prétraitement seront présentés dans les sections suivantes 3.1. et 3.2.

3.1. L'architecture du système de traduction automatique

Notre système utilise un corpus bilingue parallèle lemmatisé, étiqueté, annoté et aligné aux niveaux propositionnel et lexical. L'étiqueteur appliqué est *TTL* (Ion, 2007 ; Todirașcu *et al.*, 2011). Cet outil ainsi que les corpus exploités seront décrits plus loin, dans la section 3.2.

La méthode d'alignement lexical élaborée (Navlea et Todirașcu, 2013, 2014) sera détaillée, dans le chapitre 4. Cette méthode a comme point de départ l'application de l'outil statistique *GIZA++* (Och et Ney, 2000, 2003) qui sera présenté dans la sous-section 3.1.2.

Pour construire des modèles de langue de la langue cible, l'application *SRILM* (Stolcke, 2002) a été utilisée (cf. sous-section 3.1.1.). Afin d'optimiser les paramètres du décodeur *MOSES* (Koehn *et al.*, 2007) (cf. sous-section 3.1.1.), la procédure *MERT* (Bertoldi *et al.*, 2009) a été appliquée (cf. sous-section 3.1.1.). La Figure 22 suivante comprend l'architecture du système de traduction automatique statistique factorisée français - roumain.

[104] http://translate.google.com/

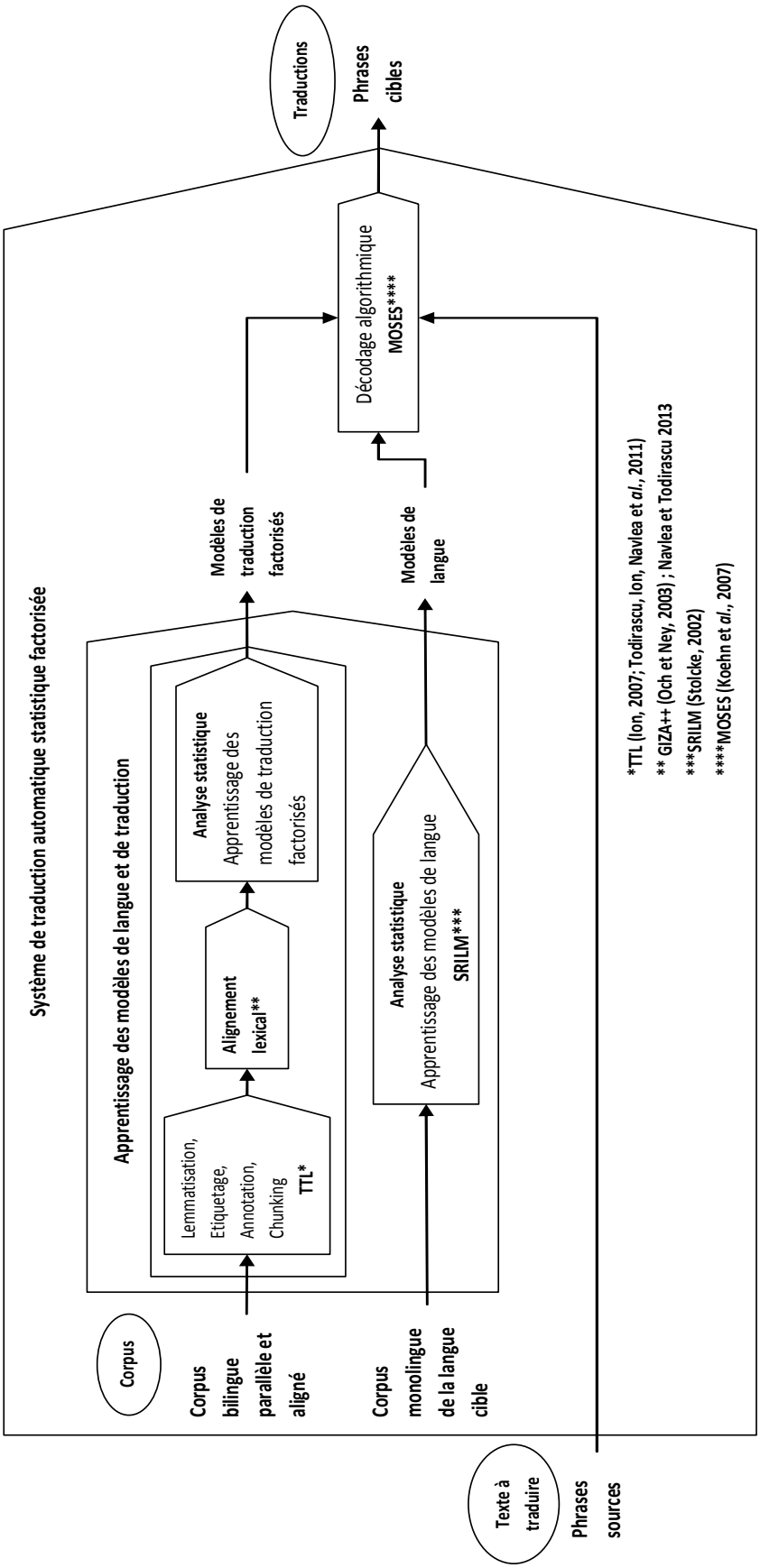


Figure 22. L'architecture du système de traduction automatique statistique factorisée

Pour développer le système de traduction automatique, nous avons appliqué les huit étapes principales suivantes :

1. constitution de corpus monolingues en langue cible et bilingues parallèles alignés au niveau propositionnel ;

2. prétraitement des corpus (segmentation lexicale, lemmatisation, étiquetage et annotation) ;

3. alignement lexical des corpus bilingues parallèles ;

4. construction de modèles de traduction purement statistiques et factorisés dans les deux sens du processus de traduction ;

5. construction de modèles de langue en langue cible ;

6. évaluation des systèmes de traduction construits ;

7. optimisation des paramètres du décodeur pour chaque système construit ;

8. évaluation des systèmes optimisés.

Ainsi, l'une de nos priorités a été de construire les ressources linguistiques requises par *MOSES* (Koehn *et al.*, 2007) afin d'adapter le décodeur à la paire de langues étudiées. Ce décodeur sera décrit dans la sous-section suivante.

3.1.1. Le décodeur *MOSES*

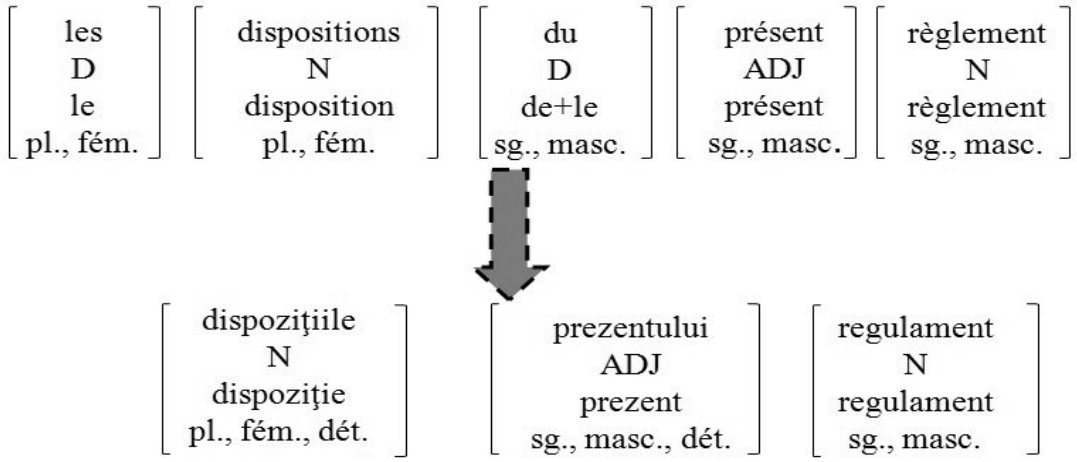
Pour adapter *MOSES* (Koehn *et al.*, 2007) à une nouvelle paire de langues, dans notre cas pour le français et le roumain, il est nécessaire de :

- construire des modèles de traduction factorisés, en utilisant un corpus bilingue parallèle lemmatisé, étiqueté et aligné aux niveaux propositionnel et lexical ;

- construire des modèles de langue de la langue cible (basés sur la forme d'occurrence des mots mais aussi sur d'autres facteurs linguistiques présents dans le corpus d'entraînement), par l'utilisation d'un corpus monolingue.

MOSES est un décodeur à base de séquences mais il est capable de prendre en compte également des modèles de traduction factorisés. Celui-ci étend donc la traduction de séquences par l'utilisation des informations linguistiques (lemmes, étiquettes morphosyntaxiques, etc.) associées aux mots.

Un exemple de traduction factorisée du français vers le roumain est donné dans la Figure 23. Cet exemple concerne la paire bilingue de séquences *les dispositions du présent règlement* vs. *dispoziţiile prezentului regulament*. Dans cette figure, chaque mot est représenté comme un vecteur de facteurs linguistiques (formes de mots, étiquettes de parties de discours, lemmes, descriptions morphosyntaxiques).



D : déterminant ; N : Nom ; ADJ : Adjectif ; sg. : singulier ; pl. : pluriel ; fém. : féminin ; masc. : masculin ; dét. :déterminé.

Figure 23. Traduction factorisée du français vers le roumain

Un système factorisé met en correspondance les facteurs linguistiques associés aux mots avant d'exploiter ces facteurs pour générer en sortie d'autres facteurs, comme les formes fléchies correctes des mots, par exemple. Ci-dessous figure une configuration possible de traduction du nom *règlement*, du français vers le roumain. Les étiquettes utilisées sont celles du projet *Multext*[105] (les étiquettes *MSD*[106]) pour le français (Ide et Véronis, 1994) et le roumain (Tufiș et Barbu, 1997). Ces étiquettes apparaissent aussi dans notre corpus (cf. sous-section 3.2.). De plus, chaque lemme est accompagné des deux premiers caractères de l'étiquette morphosyntaxique afin de désambigüiser morphologiquement les lemmes (Tufiș *et*

[105] http://aune.lpl.univ-aix.fr/projects/multext/
[106] Morpho-Syntactic Descriptors

al., 2005b, 2006). Par exemple, un même lemme peut être un nom commun (*employé_Nc*) ou un adjectif qualificatif participial (*employé_Af*). Ces lemmes seront différenciés du point de vue morphologique par l'étiquette correspondante.

Tout d'abord, le système traduit les lemmes :

règlement_Nc → regulament_Nc

Dans un deuxième temps, il traduit les étiquettes des parties du discours :

NSN → NSN, NSRY, NSOY, NPN, NPRY, NPOY

L'étiquette de partie du discours NSN (nom, singulier, sans déterminant) du corpus français est mise en correspondance avec toutes les étiquettes possibles concernant le nom dans le corpus roumain : nom singulier ou pluriel, non déterminé (NSN, NPN) ou nom singulier ou pluriel, déterminé dans les cas[107] nominatif-accusatif (NSRY, NPRY) ou génitif-datif (NSOY, NPOY). Ces étiquettes réduites appelées *C-tag*[108] (Tufiș, 1999, 2000) sont dérivées à partir des étiquettes *MSD* proposées par le projet *Multext-EAST* (Dimitrova *et al.*, 1998) (voir Annexes 2 et 3).

Dans un troisième temps, le système traduit les étiquettes morphosyntaxiques des unités lexicales considérées :

Ncms-- → Ncms-n, Ncmsry, Ncmsoy, Ncfp-n, Ncfpry, Ncfpoy

L'étiquette morphosyntaxique *Ncms--* (nom commun, masculin, singulier) du corpus français est mise en correspondance avec toutes les étiquettes possibles du nom commun dans le corpus roumain : nom commun, masculin, singulier, non déterminé (Ncms-n) ou déterminé dans les cas nominatif-accusatif (Ncmsry) ou génitif-datif (Ncmsoy) ; nom commun, féminin, pluriel, non déterminé (Ncfp-n) ou déterminé dans les cas nominatif-accusatif (Ncfpry) ou génitif - datif (Ncfpoy). Il s'agit des étiquettes *MSD* proposées par le projet *Multext* (voir Annexes 2 et 3).

Enfin, le système génère toutes les formes fléchies du mot cible à partir des résultats des étapes antérieures :

---

[107] Le nom en roumain présente plusieurs cas (cf. chapitre 4, sous-section 4.1.3.).
[108] Nous reviendrons sur l'utilité de ces étiquettes dans le chapitre 5, section 5.1.

regulament_Nc | *NSN* | *Ncms-n* | → *regulament*

regulament_Nc | *NSRY* | *Ncmsry* | → *regulamentul*

regulament_Nc | *NSOY* | *Ncmsoy* | → *regulamentului*

regulament_Nc | *NPN* | *Ncfp-n* | → *regulamente*

regulament_Nc | *NPRY* | *Ncfpry* | → *regulamentele*

regulament_Nc | *NPOY* | *Ncfpoy* | → *regulamentelor*

Ainsi, le système génère toutes les formes fléchies du mot cible (*regulament*, *regulamentul*, *regulamentului*, etc.) à partir du lemme, des étiquettes des parties du discours et des étiquettes morphosyntaxiques. Par conséquent, cette méthode permet, par exemple, la traduction des mots supplémentaires inconnus initialement dans le système. À titre d'illustration, si le mot d'entrée *règlement* est connu et la forme *règlements* est inconnue dans le système, celui-ci peut générer la forme cible *regulamente* à partir du lemme connu et des descriptions linguistiques de sortie *regulament_Nc|NPN|Ncfp-n*. Cette technique est utile surtout pour les langues riches morphologiquement, comme le sont le français et le roumain.

Dans sa distribution standard, *MOSES* comprend les applications nécessaires pour construire des modèles de traduction (statistiques purs et factorisés).

Afin de construire des modèles de langue de la langue cible (cf. chapitre 2, sous-section 2.2.1.3.1.) pour notre système de traduction, nous avons utilisé l'application appelée *SRILM* (Stolcke, 2002). Cette application apprend automatiquement un modèle de langue, en exploitant un corpus monolingue. En effet, *SRILM* (Stolcke, 2002) permet l'obtention des modèles de langue basés sur les formes des mots, sur les lemmes ou sur tout autre facteur linguistique présent dans le corpus d'entraînement (étiquettes de partie du discours, morphosyntaxiques, etc.).

À partir des modèles de langue et de traduction construits au préalable, le décodeur *MOSES* cherche la traduction la plus probable, par le biais d'un algorithme de recherche en faisceau (*beam search*) (Koehn *et al.*, 2003 ; Koehn et Hoang, 2007) (cf. chapitre 2, sous-sections 2.2.2.et 2.2.2.6.).

Pour évaluer les systèmes de traduction construits, le score *BLEU* (Papineni *et al.*, 2002) (cf. sous section 2.2.2.4.1) est également disponible dans la distribution de *MOSES*.

Après évaluation, pour chaque système de traduction construit, les paramètres du décodeur - c'est-à-dire les poids des fonctions de traits utilisées par *MOSES* (cf. sous-section 2.2.2.6. où la formule mathématique appliquée par les systèmes factorisés est décrite) - peuvent être optimisés itérativement au moyen de l'application appelée *MERT* (Bertoldi *et al.*, 2009). Cette application est aussi intégrée par *MOSES*. Celle-ci implémente l'algorithme *Minimal Error Rate Training* (Och, 2003) qui maximise un score d'évaluation (le score *BLEU* par défaut) sur un corpus de développement non utilisé pendant l'étape d'entraînement.

La première étape pour construire des modèles de traduction dans *MOSES* est l'alignement lexical du corpus bilingue parallèle utilisé, dans les deux sens du processus de traduction. Pour la réaliser, *MOSES* fait appel à l'outil statistique d'alignement *GIZA++* (Och et Ney, 2000, 2003) qui sera décrit dans la sous-section ci-dessous.

3.1.2. L'aligneur lexical *GIZA++*

L'aligneur *GIZA++* (Och et Ney, 2000, 2003) implémente les cinq modèles génératifs *IBM* bien connus (Brown *et al.*, 1993), qui proposent des alignements mot-à-mot (cf. chapitre 2, sous-section 2.2.1.3.2.). Les modèles *IBM* sont des modèles de traduction très complexes que *GIZA++* entraîne successivement, du plus simple (*IBM 1*) au plus sophistiqué (*IBM 5*), à partir d'un corpus bilingue parallèle aligné au niveau propositionnel. En effet, les paramètres d'un modèle sont utilisés comme point initial de départ par le modèle suivant. À côté des modèles *IBM*, *GIZA++* implémente aussi le modèle *HMM* (Vogel *et al.*, 1996). Ces modèles seront présentés tout au long de cette section.

GIZA++ apprend itérativement les paramètres des modèles de traduction utilisés par le biais de l'algorithme dénommé *Expectation Maximization* (*EM*) (Dempster *et al.*, 1977). Cet algorithme suit les quatre étapes suivantes (Koehn, 2010) :

1) l'initialisation du modèle par des distributions de probabilités uniformes, c'est-à-dire que chaque mot d'entrée peut être traduit avec une probabilité égale dans n'importe quel mot de sortie ;

2) l'application du modèle au corpus parallèle (*Expectation - E*) ; Si, initialement, tous les alignements sont possibles, par la suite ce sont les alignements les plus probables qui sont recherchés.

3) l'apprentissage du modèle à partir du corpus parallèle (*Maximization - M*) ; Le modèle calcule les probabilités de traduction des mots.

4) l'itération des étapes 2 et 3 jusqu'à ce que les paramètres aient une valeur optimale.

Dans l'optique des modèles *IBM*, un mot source peut avoir comme équivalent de traduction zéro ou un seul mot cible dans un corpus bilingue parallèle. Dans le premier cas, un token spécial appelé *NULL* est généré dans la phrase cible et sera le correspondant du mot source concerné. En revanche, un seul mot cible peut avoir comme équivalent de traduction zéro, un ou plusieurs mots source. Par conséquent, les modèles *IBM* sont asymétriques, d'où la nécessité des approches à base de séquences ultérieures pour les symétriser, par des heuristiques comme l'union ou l'intersection des alignements bidirectionnels (Koehn *et al.*, 2003) (cf. sous-section 2.2.2.1.). Le meilleur alignement unidirectionnel est dénommé *alignement Viterbi*.

Dans leur travail, Brown *et al.* (1993) ont mis en œuvre les modèles *IBM*, du plus simple au plus complexe, afin d'estimer les probabilités de traduction des mots.

Ainsi, *IBM* 1 calcule les probabilités lexicales de traduction entre deux mots avec la mention que tous les alignements sont vus comme étant équiprobables. En effet, les positions des mots sources et cibles dans deux phrases parallèles considérées ne sont pas prises en compte et donc l'alignement effectué ne dépend pas des positions des mots. Cela a un grand désavantage car les alignements à des distances trop longues peuvent nuire considérablement à la performance du modèle, si l'on considère que les phrases bilingues parallèles gardent une certaine similitude syntaxique au niveau des groupes de mots d'une langue à l'autre.

Le modèle *IBM* 2 considère alors que les alignements ne sont pas équiprobables et introduit un modèle d'alignement afin de prendre en compte les positions des mots sources et cibles dans deux phrases parallèles considérées. Ainsi, chaque alignement à l'intérieur de ces phrases a une probabilité qui se calcule en fonction des mots appartenant à l'alignement considéré et de la taille des phrases. Il s'agit du calcul de la distorsion (cf. sous-section

2.2.1.3.2). Ce modèle a comme paramètres les probabilités de traduction ainsi que les probabilités du modèle d'alignement.

Les modèles *IBM* suivants - 3, 4 et 5 - introduisent la notion de fertilité (cf. sous-section 2.2.1.3.2.), car un mot peut avoir comme équivalent de traduction non seulement un seul mot mais une séquence de mots. Rappelons que la fertilité représente le nombre de mots générés par un mot source dans la langue cible. Dans ces modèles, le processus de traduction fonctionne en trois étapes. Un exemple de traduction selon le modèle *IBM* 3 est donné dans la Figure 24 suivante. L'exemple concerne le couple de phrases français - roumain *Comisia examinează următorul text* (RO) vs. *La Commission procède à l'examen du texte suivant* (FR).

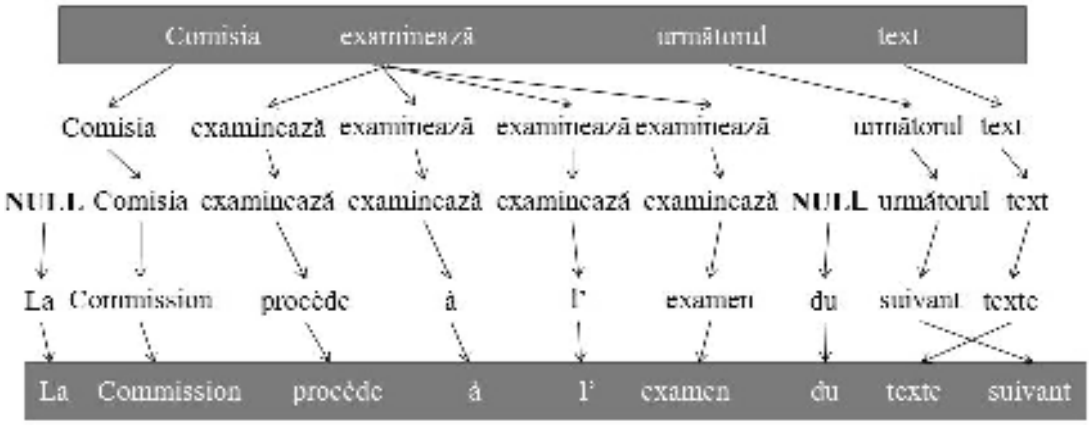


Figure 24. Processus de traduction dans le modèle *IBM* 3

Dans la Figure 24 ci-dessus, le modèle *IBM* 3 calcule premièrement les probabilités de fertilité. Par exemple, le mot roumain *examinează* 'examine' produit 4 mots *procède à l'examen*, tandis que les autres mots produisent un seul mot en français.

Ensuite, les probabilités de traduction sont calculées. Le mot spécial *NULL* est aussi traduit.

Finalement, les mots traduits sont réordonnés en calculant les probabilités de distorsion (cf. sous-section 2.2.1.3.2.). La distorsion est estimée comme dans le modèle 2 et, notamment, en fonction de la position des mots et de la taille des phrases.

À la différence du modèle 3, les modèles 4 et 5 suivants introduisent la notion de distorsion relative, c'est-à-dire que l'alignement des mots est dépendant de l'alignement des mots voisins. En effet, les mots ne s'alignent pas de façon arbitraire mais plutôt en fonction de leurs relations de dépendance locale qui sont généralement gardées d'une langue à l'autre.

Après démonstration de la relative déficience des modèles *IBM* 3 et 4, Brown *et al.* (1993) ont développé le modèle 5, très proche du modèle 4. Néanmoins, le modèle *IBM* 5 ne s'est pas révélé plus performant que son prédécesseur (Och et Ney, 2000)[109].

Le modèle *HMM* (Voget *et al.*, 1996) a été conçu comme une variante du modèle *IBM* 2 qui effectue un alignement lexical indépendamment des alignements voisins. En ce qui concerne les langues indo-européennes, l'idée de départ se base sur le constat que l'alignement des mots ne se fait pas aléatoirement mais en fonction des mots existants à proximité. Cette idée est en concordance avec les modèles *IBM* 4 et 5 calculant les probabilités de distorsion relative. En effet, si un mot source sur une position *i* apparaît aligné avec un mot cible sur une position *j*, le mot source suivant prouve la tendance de s'aligner avec un mot cible qui est à proximité de *j*. C'est un modèle Markovien (*Hidden Markov Model - HMM*) (cf. sous-section 2.2.1.3.1) qui peut modéliser un tel aspect de l'alignement. Ainsi, selon le modèle *HMM* (Voget *et al.*, 1996) qui en est un, l'alignement d'un mot source sur la position *i* dépendra de l'alignement du mot source précédent sur la position *i-1* et de la taille de la phrase cible.

Malgré tous ces efforts des modèles *IBM* et HMM pour modéliser l'alignement entre les mots à l'intérieur des phrases parallèles, *GIZA++* fournit un alignement lexical de base qui nécessite toujours des améliorations afin d'être performant pour la traduction automatique statistique.

Les expériences ont montré que les performances de *GIZA++* dépendent du domaine et de la taille du corpus d'entraînement. La qualité des résultats de cet aligneur est directement proportionnelle avec le volume des données d'entraînement, car cet outil calcule un score d'association entre les mots. De ce fait, les formes hapax (les formes de fréquence 1) ont moins de chances de s'aligner correctement. De même, la longueur des phrases alignées peut poser des problèmes importants à l'aligneur lexical dans le cas où les phrases sont trop longues (plus de 100 tokens) ou s'il y a une grande différence entre les longueurs des phrases formant un couple bilingue. De ce fait, les phrases bilingues parallèles d'entrée sont sélectionnées au préalable en fonction de leur longueur, afin d'améliorer les performances d'alignement lexical de *GIZA++* et, par conséquent, de la traduction automatique statistique.

[109] Une description plus détaillée des modèles *IBM* se trouve dans les travaux de référence de Brown *et al.* (1993).

Les résultats de *GIZA++* peuvent aussi dépendre des langues n'ayant pas une syntaxe similaire.

En outre, les modèles *IBM* ne sont pas symétriques. Par exemple, un seul mot cible peut correspondre à plusieurs mots source mais pas inversement. De plus, l'aligneur n'est pas capable de collecter des alignements multiples, dans les cas où plusieurs mots sources correspondent à plusieurs mots cibles. L'aligneur ne collecte donc pas les alignements des séquences plus ou moins compositionnelles. Comme l'unité d'alignement est le mot, *GIZA++* ne détecte pas non plus les dépendances entre les groupes de mots. Une difficulté importante rencontrée par ces modèles de traduction statistiques reste également le ré-ordonnancement efficace des mots cibles. En effet, ces modèles envisagent une certaine monotonie de l'alignement lexical, c'est-à-dire que si un mot source sur une position *i* est aligné à un mot cible sur une position *j* alors un mot source sur une position *i' > i* sera aligné à un mot cible sur une position *j' > j*. Or, en fonction des langues traitées, l'ordre des mots de la langue cible est plus ou moins différente de l'ordre des correspondants dans la langue source.

Toutefois, l'aligneur fournit un alignement lexical de base (mot-à-mot) à partir duquel les approches ultérieures à base de séquences ont construit des modèles de traduction plus performants, par l'utilisation des heuristiques de symétrisation (intersection / union) des alignements bidirectionnels (Koehn *et al.*, 2003) (cf. sous-section 2.2.2.1.). Cet alignement de base produit un certain nombre d'erreurs et nécessite donc la mise en place de stratégies et de méthodes d'amélioration. Les modules supplémentaires que nous avons développés afin d'améliorer les sorties de *GIZA++* seront brièvement présentés dans la sous-section suivante et détaillés dans le chapitre 4, spécialement dédié au système d'alignement lexical français - roumain.

3.1.3. Modules supplémentaires d'alignement lexical

Dans notre approche, *GIZA++* a fourni un alignement lexical de base (cf. chapitre 4, sous-section 4.1.) que nous avons amélioré par l'heuristique d'intersection (Koehn *et al.*, 2003) et par l'implémentation de modules d'alignement lexical supplémentaires. L'intersection des alignements bidirectionnels (langue source - langue cible et vice-versa) sert à l'amélioration de la précision du système car cette heuristique collecte seulement les alignements communs, considérés donc sûrs, vu qu'ils sont repérés dans les deux sens du processus d'alignement. De plus, afin d'améliorer le rappel du système, des modules supplémentaires sont nécessaires

pour collecter des alignements corrects qui n'ont pas été détectés automatiquement par *GIZA++*.

Ces modules ont été développés à l'issue d'une analyse linguistique détaillée des erreurs générées par le système d'alignement lexical de base. Ces erreurs concernent principalement l'alignement des cognats et l'alignement des morphèmes ou des lexèmes au niveau morphosyntaxique et stylistique (cf. chapitre 4). Ainsi, les modules implémentés sont les suivants :

1. un module d'identification et d'alignement automatique de cognats (Navlea et Todirașcu, 2011a,b,e, 2012) ; Les cognats représentent des paires bilingues de mots similaires au niveau orthographique et/ou phonétique et susceptibles d'avoir un sens commun tels que *produit* (FR) / *produs* (RO), par exemple. Ce module (cf. chapitre 4, section 4.2.) est nécessaire car *GIZA++*, bien que détectant les cognats les plus fréquents dans le corpus, il ne collecte pas les alignements des cognats moins fréquents ou apparaissant une seule fois dans le corpus (les formes hapax).

2. un module utilisant un ensemble de règles heuristiques contextuelles morphosyntaxiques et stylistiques (Navlea et Todirașcu, 2010ab, 2011cd ; Navlea et Havași, 2012) ; Ce module collecte les alignements manquants au niveau des structures morphosyntaxiques qui sont différentes d'une langue à l'autre (cf. sous-section 4.1.3.) et au niveau des contraintes stylistiques en traduction juridique (cf. sous-section 4.1.4.).

Précisons que des erreurs fréquentes d'alignement lexical ont également été repérées au niveau des collocations. Celles-ci représentent des expressions poly-lexicales où les mots entretiennent une relation lexico-syntaxique (Todirașcu *et al.*, 2008). Ainsi, le système d'alignement lexical initial intégrait également un dictionnaire de collocations *Verbe + Nom* disponible pour le français et le roumain (Todirașcu *et al.*, 2008). Mais, comme à l'issue de l'évaluation du système, nous avons remarqué une faible amélioration due au fait qu'il existe très peu de collocations communes entre le dictionnaire et le corpus de test utilisé (cf. sous-section 4.4.4.), le module d'alignement de collocations n'est plus inclus dans le système d'alignement lexical intégré par le système de traduction automatique développé. Le système d'alignement lexical ainsi que son évaluation seront décrits en détail dans le chapitre 4.

Dans la section suivante, seront présentés les corpus bilingues parallèles et monolingues disponibles pour le français et le roumain, ainsi que les corpus sélectionnés et prétraités afin de construire notre système de traduction automatique.

3.2. Les corpus

La première étape suivie pour construire le système de traduction automatique statistique factorisée réside dans la constitution de corpus bilingues parallèles et monolingues. Nous avons tout d'abord procédé à un inventaire des corpus disponibles pour la paire de langues étudiées. Nous présenterons ces corpus dans la sous-section suivante, en nous efforçant de motiver le choix des corpus sélectionnés.

3.2.1. Les corpus bilingues parallèles et monolingues disponibles pour le français et le roumain

Il existe très peu de corpus parallèles dans la paire de langues français - roumain. Les corpus bilingues parallèles disponibles appartiennent principalement aux domaines juridique - administratif et politique (cf. Tableau 17 ci-dessous).

Tableau 17. Les corpus bilingues parallèles disponibles dans la paire de langues français - roumain

Source du corpus	Domaine	Nombre de mots / FR	Nombre de mots / RO
JRC-Acquis (Steinberger *et al*., 2006)	juridique et administratif	5 828 169	5 357 017
DGT-TM (Steinberger *et al.*, 2012)	juridique et administratif	9 953 360	9 142 291
Site Web de la Commission Européenne	politique	200 590	185 476
Site Web du Parlement Européen	politique	137 422	126 366
Sites Web des compagnies aériennes roumaines (*TAROM*, *Blue Air*)	aéronautique	33 757	29 596
TOTAL	**-**	**16 153 298**	**14 840 746**

Le Tableau 17 comprend les corpus bilingues parallèles disponibles pour la paire de langues français - roumain avec la source de chaque corpus, le domaine auquel ils appartiennent et

leur taille (nombre de mots). La taille du corpus entier s'élève à environ 30 millions de mots : 16 153 298 pour le français et 14 840 746 pour le roumain.

Le premier corpus (si l'on considère la chronologie) bilingue parallèle représentatif disponible, aussi bien pour le français que le roumain, est le corpus parallèle juridique *JRC-Acquis* (Steinberger *et al.*, 2006). Il est basé sur le corpus parallèle multilingue *Acquis Communautaire*, composé de la législation de l'Union Européenne des années 1950 jusqu'à présent. L'*Acquis Communautaire* est disponible dans 231 paires de langues unidirectionnelles (462 au total) obtenues à partir de 22 sur les 23 langues officielles de l'Union Européenne.

JRC-Acquis est un ensemble de textes multilingues parallèles alignés au niveau des paragraphes, disponible gratuitement au format *XML* (*eXtensible Markup Language*). *XML* est un langage informatique qui permet de structurer l'information dans les documents par le biais des balises. Une balise est un marqueur encadré par les chevrons (<marqueur>) et peut comporter une balise de fin (</marqueur>). Celle-ci marque les zones de texte afin de faciliter la recherche d'information et le partage des connaissances. L'un des avantages du *XML* réside dans le fait que les balises peuvent être définies manuellement pour des applications propres comme, par exemple, l'annotation morphosyntaxique d'un texte. Concernant un corpus parallèle au format *XML* et aligné au niveau des paragraphes, une balise de type <paragraphe> va marquer les paragraphes du corpus. Chaque paragraphe est mis en correspondance avec le paragraphe traduit en langue cible par le biais d'un attribut identifiant unique.

À partir de *JRC-Acquis*, nous avons extrait un sous-ensemble de 228 174 paires de phrases alignées 1:1 (c'est-à-dire qu'à une seule phrase source correspond une seule phrase cible), choisies parmi l'ensemble de documents communs en français et en roumain. Ont été préférées les phrases en correspondance de 1 à 1, aux phrases alignées 1 à plusieurs, par exemple, car les systèmes actuels d'alignement lexical et de traduction automatique statistique utilisent seulement ce type d'alignement propositionnel (1:1). En effet, ces systèmes ne peuvent prendre en compte le contexte linguistique et situationnel d'une paire de phrases alignées, ce qui représente une des limites de la traduction automatique statistique.

L'autre grand corpus parallèle représentatif disponible pour le français et le roumain est *DGT-TM* (*Directorate-General for Translation - Translation Memory*)[110] (Steinberger *et al.*, 2012) basé également sur l'*Acquis Communautaire*. Comme son nom l'indique et à la différence de *JRC-Acquis*, *DGT-TM* est une mémoire de traduction, c'est-à-dire une base constituée d'un ensemble de phrases alignées démunies de leur contexte. Dans ce corpus, l'avantage est que la plupart des alignements ont été réalisés manuellement. *DGT-TM* est disponible gratuitement au format *TMX* (*Translation Memory eXchange*). TMX est un standard du *XML*, quasi classique en matière d'échange de mémoires de traduction.

À partir de *DGT-TM*, nous avons extrait les 490 962 paires de phrases alignées 1:1, constituant des équivalences de traduction en français et en roumain.

Hormis ces deux corpus à forte spécialisation juridique et administrative, nous avons constitué d'autres corpus parallèles à partir du Web, dans l'idée d'obtenir une plus grande richesse lexicale et d'explorer d'autres domaines. Ainsi, nous avons constitué ces corpus dans le cadre du projet *CAP*[111] (Todirașcu et Navlea, 2010 ; Todirașcu *et al*., 2014), réalisé en collaboration avec le projet européen *CLARIN*[112] visant la construction des ressources linguistiques et des outils pour les humanités. Il s'agit de corpus parallèles multilingues alignés au niveau propositionnel, disponibles dans les langues français, anglais, allemand et roumain. Ces corpus parallèles appartiennent au domaine politique et aéronautique et comprennent approximativement 400 000 mots par langue. Pour notre projet de thèse, nous en avons extrait la partie français - roumain.

Nous avons constitué ces corpus manuellement, en prenant en compte différents critères comme la disponibilité des textes bilingues et, notamment, des domaines, la fiabilité des sources, la qualité des traductions humaines de ces textes.

Il existe relativement peu de domaines présents sur le Web pour la paire de langues étudiées. Nous avons repéré essentiellement deux domaines bien représentés sur le Web : les compagnies aériennes roumaines (*TAROM*[113], *Blue Air*[114]) et, de nouveau, les institutions européennes. Nous avons donc plutôt privilégié le domaine de la politique, en tenant compte

[110] http://ipsc.jrc.ec.europa.eu/index.php/Traineeships/197/0/
[111] L'objectif de ce projet est l'étude de la relation de hiérarchie « chef » en contexte multilingue.
[112] http://www.clarin.eu/
[113] https://www.tarom.ro/fr/
[114] http://www.blueairweb.com/Page-D-Accueil/

de la fiabilité élevée des sites Web multilingues des institutions européennes et de la qualité élevée des traductions humaines fournies par ces sites. Les textes parallèles concernant ces institutions ont été collectés sur les sites Web du Parlement Européen[115] et de la Commission Européenne[116]. Nous avions eu au départ l'idée de constituer un corpus incluant davantage de langue générale, mais cela n'a pas été possible en raison du manque de ressources disponibles. Il nous a donc semblé préférable, pour des raisons d'homogénéité, de rester dans les domaines juridique - administratif et politique des institutions européennes. C'est la raison pour laquelle nous avons choisi de ne plus inclure, dans le système final, les corpus parallèles des compagnies aériennes constitués par nos soins.

Lors de l'étape de constitution manuelle des corpus à partir du Web, nous avons spécifié, pour chaque texte, l'auteur, la date et sa source indiquée par son *URL*. Les données collectées ont été nettoyées en éliminant les éléments non-textuels : images, notes de bas de page, tableaux, etc. Pour résoudre le problème de l'absence des diacritiques pour la plupart des textes roumains collectés à partir du Web, nous avons utilisé le système qui récupère les diacritiques *Diac+* (Tufiș et Ceaușu, 2008).

Les corpus parallèles collectés nécessitent un alignement propositionnel et lexical afin de pouvoir être exploités par les systèmes de traduction automatique statistiques. Ainsi, nous avons aligné ces corpus au niveau propositionnel en utilisant l'aligneur *Alinea*[117] (Kraif, 2001) qui sera présenté dans la sous-section 3.2.2. Concernant l'alignement lexical, nous avons mis en place notre propre méthode qui sera décrite en détail dans le chapitre 4.

Par la suite, nous avons dû encore limiter la quantité et le domaine du corpus car l'utilisation du corpus entier pour réaliser différentes configurations du système de traduction automatique aurait pris un temps considérable - plusieurs dizaines de jours de temps de calcul sur un ordinateur de type PC (Intel(R) Core(M) 15 CPU 2,67 GHz, 3,9 Gio de mémoire). Nous avons donc opté finalement pour *DGT-TM*, car la plupart des alignements au niveau propositionnel avaient été réalisés manuellement, au contraire des autres corpus dont l'alignement propositionnel avait été effectué automatiquement, ce qui impliquait un certain nombre d'erreurs.

[115] http://www.europarl.europa.eu/parliament/public/staticDisplay.do?id=146&language=fr
[116] http://ec.europa.eu/index_fr.htm
[117] http://w3.u-grenoble3.fr/kraif/index.php?option=com_content&task=view&id=27&Itemid=43

Pour identifier les paramètres des systèmes de traduction automatique français - roumain, nous avons utilisé 64 923 paires de phrases alignées 1:1 extraites à partir du corpus juridique et administratif *DGT-TM*. Ces phrases sont complètes, c'est-à-dire qu'elles commencent par une majuscule et se terminent par un signe de ponctuation. De plus, chaque phrase comprend 100 tokens (mots et signes de ponctuation) au maximum, comme requis par le décodeur utilisé *MOSES* (Koehn *et al*., 2007).

Afin de construire des modèles de langue appropriés pour un système factorisé (des modèles basés sur les formes des mots ou sur les facteurs linguistiques associés aux mots), un corpus monolingue français lemmatisé et étiqueté, comprenant 480 764 phrases extraites de *JRC-Acquis* (Steinberger *et al*., 2006), est disponible.

Quant au roumain, nous avons exploité des modèles de langue déjà construits (Tufiș *et al*., 2013a). Il s'agit des modèles développés à partir du corpus juridique *JRC-Acquis* (Steinberger *et al*., 2006) (cf. chapitre 5).

Dans les sous-sections suivantes, seront présentés l'alignement propositionnel des corpus constitués (sous-section 3.2.2.), ainsi que le prétraitement des corpus utilisés (sous-section 3.2.3.).

3.2.2. L'alignement propositionnel des corpus parallèles constitués

Les corpus parallèles que nous avons constitués ont été également alignés au niveau propositionnel à l'aide d'*Alinea* (Kraif, 2001), un aligneur propositionnel dépendant de la paire de langues à traiter.

Comme *Alinea* (Kraif, 2001) n'intègre pas la paire de langues français - roumain, nous avons utilisé ses paramètres par défaut. *Alinea* effectue un alignement propositionnel fondé sur des paramètres tels que :

- les transfuges ; Ceux-ci sont définis comme des « chaînes de caractères invariantes dans le passage à la traduction : les noms propres, les données numériques, certains sigles, les numéros de chapitre, etc. » (Kraif, 2001 : 252).

- les cognats ; Concernant les cognats, Kraif (2001) donne une définition opératoire de ceux-ci : « Deux unités *U* et *U*' sont des cognats si et seulement si :

1) *U* et *U*’ ont un lien étymologique (emprunt, origine commune) perceptible dans leur signifiant.

2) on peut trouver deux phrases (*P*,*P*’) dont l’une est la traduction de l’autre, et dans lesquelles *U* et *U*’ sont en relation d’équivalence. » (Kraif, 2001 : 255). Les transfuges peuvent être vus comme « des cognats particuliers » (Kraif, 2001 : 255).

- le rapport des longueurs des phrases (en tokens ou en nombre de caractères) ;
- les lexiques bilingues.

Des algorithmes d’alignement récursifs combinent ces paramètres pour fournir un alignement propositionnel optimal. Dans *Alinea*, une phrase source peut avoir comme équivalent de traduction de zéro à trois phrases cibles et inversement. De plus, deux phrases sources peuvent être traduites par deux phrases cibles. Une fois le corpus parallèle aligné, seules les paires de phrases en correspondance de 1 à 1 sont sélectionnées pour la traduction automatique.

Pour aligner un corpus parallèle au niveau propositionnel, *Alinea* procède à une étape de segmentation en phrases du corpus parallèle utilisé et de segmentation des phrases en mots.

La segmentation en phrases d’un corpus s’avère une tâche difficile pour un logiciel d’alignement. En effet, quand bien même on pourrait considérer qu’une phrase est délimitée par une majuscule au début et un signe de ponctuation (*« . », « ? », « ! », « ... »*) à la fin, un signe de ponctuation comme le point (« . ») ne suffirait pas pour faire cette délimitation. En effet, un point peut apparaître aussi dans l’un des cas suivants : une abréviation (p. ex. *op. cit.*, *N.B.*), une liste numérotée (p. ex. *1. 2. 3.*), un nombre (p. ex. 2.5), une adresse Web (p. ex. *www.domaine.fr*). Qui plus est, le point ne suffit pas non plus comme séparateur de mots dans le cas des abréviations ou des nombres avec décimales. Se pose alors la question de savoir ce que sont la phrase et le mot pour un logiciel d’alignement.

Faisant référence à des textes du domaine politique, auquel notre corpus appartient aussi, « fortement structurés par les énumérations » (Kraif, 2001 : 234), et notamment à des extraits d’un rapport du Parlement Européen, Kraif (2001) observe, d’une part, que « si l’on admet une définition extensive de la phrase basée sur le noyau verbal régissant, on peut aboutir à des

phrases atteintes de gigantisme, formant un texte entier s'étalant sur plusieurs pages. Mais si l'on se base seulement sur certains indices typographiques, tels que les sauts à la ligne, n'importe quel syntagme, dûment énuméré, pourra revendiquer son statut de phrase. » (Kraif, 2001 : 235).

Kraif (2001) observe, d'autre part, que le niveau sémantique ne suffit pas non plus à rendre compte des limites de la phrase, comme le disent Rastier *et al.* (1994 : 115-116, *cit. in* Kraif 2001 : 235) : « à l'autonomie syntaxique qui refléterait la complétude et l'autosuffisance de la prédication, on doit opposer les relations sémantiques qui rattachent toute phrase à son contexte linguistique et situationnel. Si bien que le découpage d'un texte en phrases n'est pas si simple et la recherche d'un point n'y suffit pas. »

Au vu de ces considérations, Kraif (2001 : 235-236) propose une définition opératoire de la phrase comme « unité textuelle minimale dont les frontières sont marquées par des indices typographiques de début et de fin ». Les indices à prendre en compte sont les suivants : un point, un point d'exclamation, un point d'interrogation, un point virgule, deux points, un tiret, une position en début de ligne, une marque de paragraphe, une marque d'alinéa, etc. Mais pour faire la différence entre un point marquant la fin d'une phrase et le point d'une abréviation, des règles basées sur des expressions régulières s'avèrent nécessaires.

Dans ses expérimentations, Kraif (2001) définit donc un ensemble de règles syntaxiques simples pour effectuer la segmentation dans les cas de figure les plus fréquents. Les marqueurs de fin de phrase retenus sont les suivants : le point, le point virgule, les deux points et le saut de ligne. Parmi ceux-ci, seul le point marquant une abréviation est ambigu. Les points ont donc été ignorés dans les cas ci-dessous (Kraif, 2001) :

- point non suivi d'un espace et d'une majuscule ;
- point qui suit une lettre majuscule isolée (quant aux sigles), une lettre minuscule isolée (p. ex. *i.e.*, *c.à.d.*, *n.*) ou une abréviation standard (p. ex. cf., *pp.*, *Mr*.).

Le mot a lui aussi une définition opératoire (Kraif, 2001 : 287) : « toute chaîne de caractères (lettre ou chiffre) comprise entre deux séparateurs (caractère non-alphanumérique) ». Les séparateurs peuvent être : un espace, un retour de chariot, une tabulation, des signes de ponctuation - point, virgule, point virgule, point d'exclamation, point d'interrogation, guillemet -, une apostrophe, un tiret, des signes marquant les parenthèses - parenthèse,

crochets, accolades -, d'autres caractères non alphanumériques - pourcentage, symboles monétaires, etc. Les exceptions qui s'imposent par rapport à cette règle sont les suivantes (Kraif, 2001) :

- pour reconnaitre les sigles et les acronymes, le point ne constitue pas un séparateur s'il est immédiatement précédé et suivi d'une lettre majuscule ;

- pour reconnaitre les nombres, le point ou la virgule ne constituent pas des séparateurs s'ils sont immédiatement précédés et suivis d'un chiffre.

Les règles ainsi définies sont dépendantes des corpus et des langues à traiter. Des dictionnaires d'abréviations *ad hoc* peuvent aussi être nécessaires.

Dans sa version standard, quant au français, *Alinea* prend en compte les séparateurs suivants :

- de paragraphes (nouvelle ligne : « */n* » ou « */n/n* ») ;

- de phrases (« . », « ; », « : », « ? » , « ! ») ;

- de tokens (tabulation, espace, nouvelle ligne : « \t », « \s », « \n ») ;

- et les séparateurs pouvant apparaître dans des chaînes numériques telles que *3,2*, *01/01/2001*, etc. (« . », « , », « - », « / »).

Alinea inclut aussi une liste d'abréviations susceptibles de comporter un point, en fonction des langues à traiter.

Ainsi, nous avons adapté *Alinea* à la paire de langues étudiées, en écrivant le fichier de segmentation que l'aligneur demande comme ressource pour le roumain et nous avons fourni une liste d'abréviations *ad hoc* à partir de nos corpus parallèles constitués pour les deux langues.

Pour illustrer les critères de la segmentation en phrases dans *Alinea*, des exemples d'alignement propositionnel français - roumain effectué par l'aligneur apparaissent dans la Figure 25 suivante ; ils sont affichés dans le navigateur bi-textuel. Ci-dessous figurent également deux échantillons de texte extraits du corpus *Parlement Européen* constitué par nos soins, qui comprennent les trois phrases alignées dans la Figure 25 (il s'agit des phrases mises en gras des échantillons A et B).

A.

Résolution du Parlement européen sur les perspectives du plan d'action de Doha pour le développement au lendemain de la 7e conférence ministérielle de l'OMC

Le Parlement européen,

– vu l'accord du GATT, chapitre IV, articles 36 "Principes et objectifs" et 37 "Engagements",

– vu la déclaration ministérielle de Doha de l'OMC du 14 novembre 2001,

– vu la déclaration ministérielle de Hong Kong de l'OMC du 18 décembre 2005,

[…]

A. considérant que le cycle de Doha a été lancé dans le but de corriger les déséquilibres existant dans le système du commerce international, partant de la conviction qu'un système multilatéral fondé sur des règles plus justes et plus équitables peut contribuer à un commerce équitable dans l'intérêt du développement de tous les continents,

B. considérant que cet objectif correspond aux principes fondamentaux de l'OMC, le système de commerce multilatéral reposant sur les règles du GATT élaborées en 1947, le préambule précisant que les rapports dans le domaine commercial et économique doivent être orientés vers le relèvement des niveaux de vie, la réalisation du plein emploi, l'accord précisant que la réalisation de ces objectifs est particulièrement urgente pour les parties contractantes moins développées, cependant que le préambule de l'accord de 1994 instituant l'OMC ajoutait que ces objectifs doivent être poursuivis conformément à la finalité du développement durable, tout en s'efforçant de protéger et de préserver l'environnement,

[…]

1. se déclare fermement convaincu qu'un cadre efficace et révisé pour le commerce multilatéral est nécessaire pour mettre en place un système économique plus équilibré et plus équitable s'inscrivant dans une nouvelle gouvernance mondiale, dans l'intérêt du développement et de l'éradication de la pauvreté; réaffirme dès lors le rôle indispensable de l'OMC et la nécessité d'améliorer les règles du commerce multilatéral ;

2. rappelle l'engagement pris en 2001 à Doha par l'ensemble des membres de l'OMC de mener à bien le cycle de négociations concernant le développement à l'effet de corriger les déséquilibres existants du système de commerce et à mettre ce dernier au service du développement ;

[…]

25. charge son Président de transmettre la présente résolution au Conseil et à la Commission, ainsi qu'aux gouvernements et aux parlements des États membres et au directeur général de l'OMC.[118]

B.

Rezoluția Parlamentului European referitoare la perspectivele Agendei de Dezvoltare de la Doha (ADD) în urma celei de-a 7-a Conferințe Ministeriale a OMC

Parlamentul European,

– având în vedere Acordul GATT, capitolul IV, articolele 36 (Principii și obiective) și 37 (Angajamente),

– având în vedere Declarația ministerială de la Doha a Organizației Mondiale a Comerțului (OMC) din 14 noiembrie 2001,

– având în vedere Declarația ministerială de la Hong Kong a Organizației Mondiale a Comerțului (OMC) din 18 decembrie 2005,

[…]

A. întrucât Runda de la Doha a fost lansată cu obiectivul de a rectifica dezechilibrele existente în sistemul comercial internațional, care se bazează pe convingerea că un sistem multilateral, bazat pe norme mai corecte și mai echitabile, poate contribui la un comerț echitabil care să favorizeze dezvoltarea tuturor continentelor ;

B. întrucât obiectivul acestei runde corespunde principiilor de bază ale OMC, sistemul comercial multilateral fiind bazat pe normele GATT elaborate în 1947, în ale căror preambul se menționează că „relațiile din domeniul comerțului și activitățile economice ar trebui dezvoltate în vederea ridicării standardelor de viață, a asigurării unui loc de muncă pentru toți...", acordul continuând cu declarația conform căreia „realizarea acestor obiective are un caracter deosebit de urgent pentru părțile contractante mai puțin dezvoltate", iar în preambulul la acordul din 1994, prin care se înființa OMC, se adaugă precizarea că aceste obiective trebuie realizate „în conformitate cu obiectivul dezvoltării durabile, vizând atât protecția, cât și conservarea mediului" ;

[…]

1. este ferm convins că este nevoie de un cadru comercial multilateral eficient și reformat pentru a dezvolta un sistem economic mai echilibrat și mai echitabil ca parte a unei guvernanțe globale noi care să se afle în serviciul dezvoltării și al eradicării sărăciei; reafirmă, prin urmare, rolul indispensabil al OMC și necesitatea îmbunătățirii normelor comerciale multilaterale ;

---

[118] http://www.europarl.europa.eu/sides/getDoc.do?type=MOTION&reference=B7-2009-0189&format=XML&language=FR

2. reamintește angajamentul asumat la Doha în 2001 de toți membrii OMC privind încheierea unei runde de negocieri axate pe dezvoltare, care să vizeze rectificarea dezechilibrelor existente în sistemul comercial și utilizarea comerțului în serviciul dezvoltării ;
[…]
25. încredințează Președintelui sarcina de a transmite prezenta rezoluție Consiliului, Comisiei, guvernelor și parlamentelor statelor membre, precum și Directorului General al OMC.[119]

Dans les échantillons A et B ci-dessus, trois types d'énumérations sont imbriquées au sein d'une unité complexe de longueur assez importante, de sorte que la recherche du point s'avère insuffisante pour délimiter une phrase. *Alinea* prend en compte les unités minimales « dûment énumérés », marqués par des signes typographiques de début (tiret, lettre majuscule, numéro) et de fin (virgule, point virgule, point) afin d'effectuer la segmentation phrastique du texte d'entrée. Par conséquent, un syntagme tel que *Le Parlement européen* suivi d'une virgule, est considéré comme une phrase (voir Figure 25 suivante).

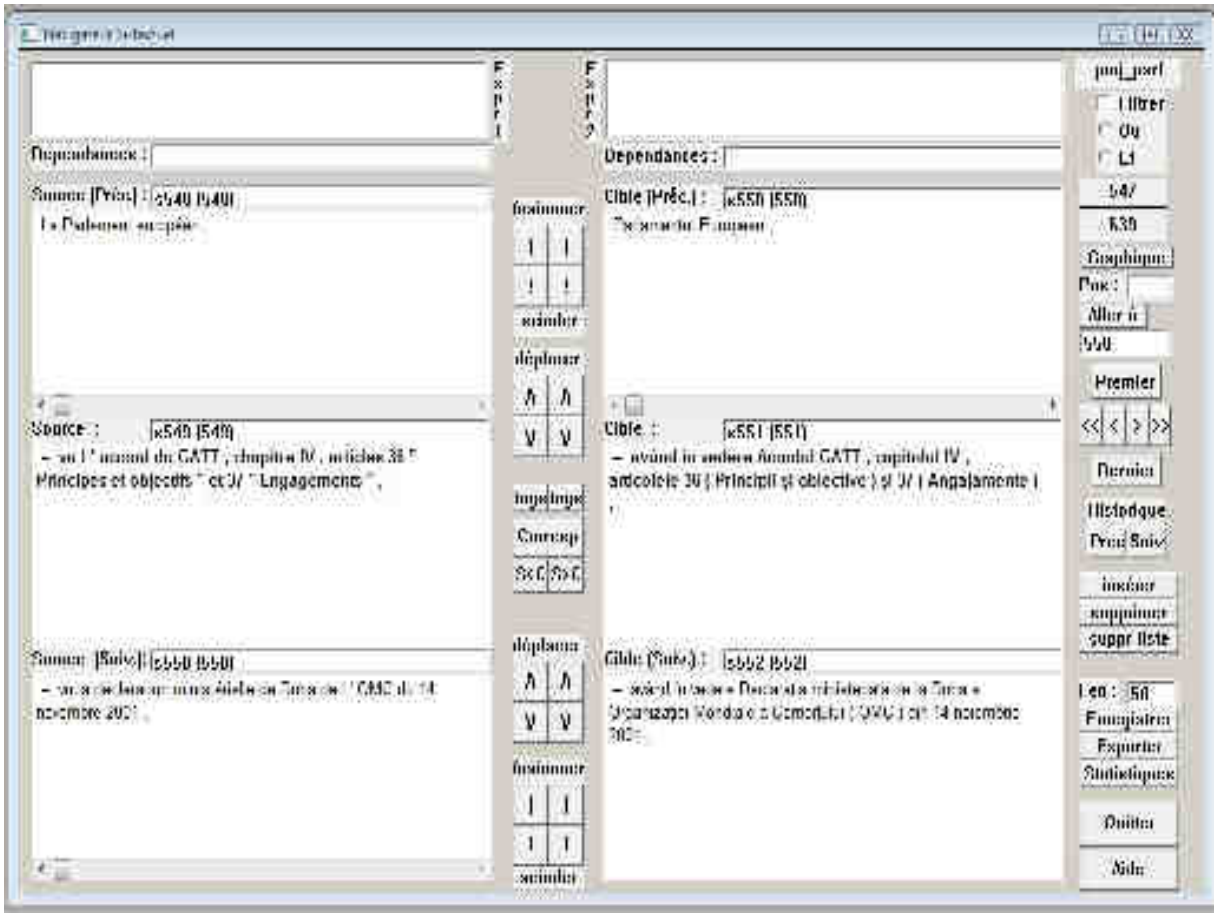


Figure 25. Exemple de phrases alignées français - roumain avec *Alinea* (Kraif, 2001)

[119]http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+MOTION+B7-2009-0189+0+DOC+XML+V0//RO

La segmentation dépend aussi du corpus et des langues à traiter car, dans l'échantillon du texte français, la deuxième énumération est marquée par une virgule, tandis qu'en roumain celle-ci est marquée par un point virgule.

Une fois alignés au niveau propositionnel, les corpus de travail passent par une étape de prétraitement. Dans la sous-section suivante, sera présentée cette étape de prétraitement concernant les corpus utilisés.

3.2.3. Le prétraitement des corpus utilisés

Le prétraitement des corpus consiste dans la segmentation lexicale (la tokénisation), la lemmatisation, l'étiquetage morphosyntaxique et l'annotation au niveau des *chunks* (découpage en groupes nominaux, prépositionnels, adjectivaux et adverbiaux simples, non-récursifs, c'est-à-dire ne pouvant pas être répétés un nombre indéfini de fois en application de la même règle). Ce type de prétraitement est requis par les systèmes factorisés du fait que ceux-ci utilisent des « facteurs » (informations linguistiques) associés aux unités lexicales dans le processus de traduction. Nous avons prétraité nos corpus en appliquant l'étiqueteur *TTL* disponible en français (Todirașcu *et al.*, 2011) et en roumain (Ion, 2007) comme service Web[120]. Nous avons aussi travaillé au développement des ressources linguistiques pour l'annotation en français[121], vu que *TTL* était initialement disponible seulement pour l'anglais et le roumain.

TTL est un étiqueteur probabiliste fonctionnant par apprentissage automatique à partir d'un corpus d'entraînement étiqueté et lemmatisé au préalable. Ainsi, nous avons étiqueté, lemmatisé et corrigé les erreurs d'étiquetage et de lemmatisation du corpus d'entraînement de *TTL*, comprenant environ 1 million de tokens. Ce corpus est composé de textes juridiques et administratifs extrais de *JRC-Acquis* (Steinberger *et al.*, 2006) (498 788 mots) et de textes journalistiques extraits du corpus *Le Monde* (488 543 mots) (Todirașcu *et al.*, 2011). L'étiquetage et la lemmatisation du corpus sont réalisés par *Tree Tagger*[122] (Schmid, 1994). De plus, le corpus est traité par *Flemm*[123] (Namer, 2000) pour réaliser une analyse

[120] https://weblicht.sfs.uni-tuebingen.de/
[121] La préparation des ressources linguistiques pour *TTL* en français a fait l'objet d'un stage de 4 mois, effectué par nos soins, au sein de l'*U.R. LiLPa* de l'Université de Strasbourg (en 2008).
[122] Les jeux d'étiquettes utilisées par *Tree Tagger* (Schmid, 1994) figurent dans l'Annexe 1.
[123] Les étiquettes utilisées par *Flemm* (Namer, 2000) sont celles du projet *Multext* (http://aune.lpl.univ-aix.FR/projects/multext/) pour le français (Ide et Véronis, 1994). Un exemple d'étiquettes concernant les adjectifs figure dans l'Annexe 2.

morphologique flexionnelle du corpus (l'enrichissement des étiquettes fournies par *Tree Tagger* par des propriétés morphologiques). Ensuite, afin de corriger les erreurs produites par la lemmatisation et l'étiquetage, nous avons proposé une méthode automatique (développée en Perl) basée sur des règles morphosyntaxiques de correction (19 règles) des erreurs systématiques identifiées, telles que : des adjectifs participiaux traités comme des verbes au participe, des noms analysés comme des adjectifs, l'absence du genre et/ou du nombre pour les noms, etc. Enfin, les erreurs ne pouvant pas être résolues par des règles de correction automatique ont été corrigées manuellement.

TTL tokénise, lemmatise et annote les corpus par des descriptions morphosyntaxiques et au niveau des *chunks*. Les résultats fournis par *TTL* sont en format *XCES* (*Corpus Encoding Standard for XML*). Ce format est un standard basé sur *XML* et utilisé dans les technologies du langage pour annoter des corpus par des informations linguistiques. Celui-ci utilise un set d'annotation considéré comme standard, qui marque le texte à différents niveaux d'analyse linguistique (syntaxique, morphosyntaxique, sémantique) et la ponctuation, par exemple. Les sorties de *TTL* comprennent l'ensemble des étiquettes *MSD* du projet *Multext*[124] pour le français (Ide et Véronis, 1994) et le roumain (Tufiș et Barbu, 1997). Un exemple de phrases alignées 1:1 et prétraitées avec *TTL* apparaît dans la Figure 26 suivante. Cet exemple concerne le couple de phrases français - roumain ci-dessous :

(FR) *Les États membres communiquent à la Commission les méthodes qu'ils utilisent.*
(RO) *Statele membre comunică Comisiei metodele pe care le utilizează.*

Ces phrases sont extraites du corpus *DGT-TM* (Steinberger *et al*., 2012).

[124] http://aune.lpl.univ-aix.FR/projects/multext/

français

```
<seg lang="fr"><s id="ttlfr.41">
<w lemma="le" ana="Da-mp"
chunk="Np#1">Les</w>
<w lemma="État" ana="Ncmp"
chunk="Np#1">États</w>
<w lemma="membre" ana="Ncmp"
chunk="Np#1">membres</w>
<w lemma="communiquer"
ana="Vmsp3p"
chunk="Vp#1">communiquent</w>
<w lemma="à" ana="Spa"
chunk="Pp#1">à</w>
<w lemma="le" ana="Da-fs"
chunk="Pp#1,Np#2">la</w>
<w lemma="commission" ana="Ncfs"
chunk="Pp#1,Np#2">Commission</w>
<w lemma="le" ana="Da-fp"
chunk="Np#3">les</w>
<w lemma="méthode" ana="Ncfp"
chunk="Np#3">méthodes</w>
<w lemma="que" ana="Pr">qu’</w>
<w lemma="il" ana="Pp3mp"
chunk="Vp#2">ils</w>
<w lemma="utiliser" ana="Vmip3p"
chunk="Vp#2">utilisent</w>
<c>.</c></s></seg>
```

roumain

```
<seg lang="ro"><s
id="ttlro.40">
<w lemma="stat" ana="Ncfpry"
chunk="Np#1">Statele</w>
<w lemma="membru" ana="Ncfp-
n" chunk="Np#1">membre</w>
<w lemma="comunica"
ana="Vmis3s"
chunk="Vp#1">comunică</w>
<w lemma="comisie"
ana="Ncfsoy"
chunk="Np#2">Comisiei</w>
<w lemma="metodă"
ana="Ncfpry"
chunk="Np#3">metodele</w>
<w lemma="pe"
ana="Spsa">pe</w>
<w lemma="care" ana="Pw3--
r">care</w>
<w lemma="el" ana="Pp3fpa----
----w" chunk="Vp#2">le</w>
<w lemma="utiliza"
ana="Vmip3"
chunk="Vp#2">utilizează</w>
<c>.</c></s></seg>
```

Figure 26. Phrases alignées français - roumain au format *XCES* prétraitées avec *TTL*

Dans l’exemple présenté dans la Figure 26 ci-dessus, l’attribut *lemma* comprend les informations sur les lemmes, l’attribut *ana* a trait aux propriétés morphosyntaxiques, tandis que l’attribut *chunk* marque les groupes reconnus (nominaux, prépositionnels, adjectivaux et

verbaux non-récursifs). L'interprétation des étiquettes morphosyntaxiques apparaissant dans ls Figure 26 se trouve dans l'Annexe 2 (pour le français) et dans l'Annexe 3 (pour le roumain).

Ces informations sont exploitées par le système de traduction automatique factorisé mais aussi par le système d'alignement lexical développé. Ce système d'alignement lexical sera présenté en détail dans le chapitre suivant, alors que le système de traduction fera l'objet du chapitre 5.

3.3. Bilan du chapitre

Dans ce chapitre, nous avons présenté l'architecture du système de traduction automatique statistique factorisée français - roumain. Ce système est construit par l'intermédiaire du décodeur *MOSES* (Koehn *et al*., 1997) qui propose aussi les instruments nécessaires à la construction de modèles de langue et de traduction basés sur la simple formes des mots ou factorisés.

Le système de traduction automatique utilise des corpus parallèles lemmatisés, étiquetés et alignés au niveau propositionnel et lexical. Pour la paire de langues étudiées, deux corpus parallèles importants du domaine juridique et administratif sont disponibles dans le domaine public : *JRC-Acquis* (Steinberger *et al*., 2006) et *DGT-TM* (Steinberger *et al*., 2012) respectivement. Afin d'explorer d'autres domaines et d'obtenir une plus grande richesse lexicale, nous avons initialement constitué trois autres corpus parallèles pour notre système. Ces corpus appartiennent au domaine politique et aéronautique. Ils ont été nettoyés et alignés au niveau propositionnel par *Alinea* (Kraif, 2001) que nous avons adapté pour la paire de langues étudiées. Au début, nous voulions avoir un corpus incluant davantage de langue générale mais cela n'a pas été possible à cause du manque de ressources sur le Web pour le roumain et le français. Ensuite, pour des raisons d'homogénéité, nous avons choisi le domaine juridique - administratif et politique pour construire le système de traduction. Finalement, nous avons dû encore limiter la quantité et le domaine du corpus, car l'exploitation du corpus entier pour réaliser différentes configurations du système de traduction automatique aurait pris un temps considérable. Nous avons donc opté pour *DGT-TM*, car la plupart des alignements au niveau propositionnel avaient été réalisés manuellement, à la différence des autres corpus dont l'alignement propositionnel automatique fournissait un certain nombre d'erreurs. Néanmoins, nous envisageons d'exploiter dans l'avenir tous les corpus disponibles et donc

nos trois corpus parallèles constitués restent des ressources importantes pour notre système de traduction automatique.

Pour construire des systèmes de traduction factorisés, les corpus utilisés nécessitent un prétraitement (segmentation lexicale, lemmatisation, étiquetage morphosyntaxique). Concernant le roumain, le seul étiqueteur disponible était *TTL* (Ion, 2007). Afin d'avoir le même étiqueteur pour les deux langues étudiées, nous avons aussi travaillé sur le développement des ressources linguistiques pour *TTL* en français (Todirașcu *et al*., 2011). Ainsi, un corpus monolingue français d'environ 1 million de tokens des domaines juridique et journalistique a été constitué pour l'entraînement. Ce corpus est lemmatisé, étiqueté et corrigé.

Quant à l'alignement lexical des corpus parallèles utilisés, l'outil statistique *GIZA++* (Och et Ney, 2000, 2003) a été initialement appliqué. Cet outil implémente les modèles *IBM* (Brown *et al*., 1993) qui sont asymétriques. Par exemple, un seul mot cible peut avoir comme équivalent de traduction plusieurs mots sources mais l'inverse n'est pas vrai, d'où la nécessité ressentie par les approches ultérieures (Koehn *et al*., 2003) de les symétriser (intersection / union des alignements bidirectionnels) afin d'obtenir des alignements à base de séquences. Par l'utilisation des séquences, le système exploite le contexte local des mots et gère mieux l'ordre des mots de la langue cible, qui reste l'un des problèmes majeurs rencontrés par les modèles IBM caractérisés plutôt par la monotonie des alignements effectués. Or, l'ordre des mots de la langue cible est souvent différente de l'ordre des mots de la langue source pour différentes paires de langues.

Nous avons ainsi obtenu un alignement lexical de base (mot-à-mot) par le biais de *GIZA++*, que nous avons amélioré par le développement de modules supplémentaires implémentant la reconnaissance et l'alignement automatique des cognats (Navlea et Todirașcu, 2011a,b,e, 2012) et des règles heuristiques linguistiques (Navlea et Todirașcu, 2010ab, 2011cd ; Navlea et Havași, 2012). Comme cet aligneur n'effectue pas d'alignements multiples (plusieurs mots sources pouvant correspondre à plusieurs mots cibles), nous avons également effectué des expériences intégrant un dictionnaire de collocations (Todirașcu *et al*., 2008). Ce module ne fait pas partie de l'alignement lexical intégré par le système de traduction, car il a amélioré de façon non significative l'alignement lexical global lors de l'évaluation du module. Le système d'alignement lexical développé ainsi que son évaluation seront présentés dans le chapitre suivant.

4. Le système d'alignement lexical français - roumain

Comme nous l'avons vu dans l'état de l'art, l'alignement lexical des corpus bilingues parallèles représente une étape fondamentale dans la construction des modèles de traduction requis par les systèmes de traduction automatique statistique. De ce fait, de nombreux travaux se concentrent sur le développement de méthodes d'alignement lexical purement statistiques (Brown *et al.*, 1990 ; Brown *et al.*, 1993 ; Och et Ney, 2000, 2003 ; Lardilleux et Lepage, 2008) ou intégrant des informations linguistiques pour accroître la performance de l'alignement et améliorer ainsi les résultats de la traduction automatique statistique (Tiedemann, 2003 ; Cherry et Lin, 2003 ; Tufiș *et al.*, 2006 ; Schrader, 2006 ; Hermjakob, 2009 ; Cendejas *et al*., 2009 ; Pal *et al*., 2013). Toutefois, certaines approches (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007) montrent qu'une amélioration substantielle de l'alignement ne mène qu'à une faible amélioration des résultats de traduction automatique statistique finaux. Ainsi, nous avons cherché à améliorer l'alignement lexical afin de pouvoir étudier l'impact de cette amélioration sur les résultats finaux du système de traduction automatique développé.

Nous avons construit un système hybride d'alignement lexical. Ce système utilise des corpus bilingues parallèles juridiques et administratifs lemmatisés, étiquetés et alignés au niveau propositionnel. Celui-ci combine des techniques statistiques et des règles heuristiques ayant un fondement linguistique. Le système intègre aussi un module d'identification et d'alignement automatique des cognats (Navlea et Todirașcu, 2011abe, 2012). De manière générale, les cognats sont définis comme les paires bilingues de mots qui se ressemblent au niveau orthographique et possèdent un sens commun. Dans notre approche, les cognats sont des équivalents de traduction présentant des similarités orthographiques et / ou phonétiques au niveau des lemmes et susceptibles de partager un sens commun (*document vs. document* ; *phase vs. fază*). De plus, le système initial intègre également un dictionnaire de collocations (Todirașcu *et al.*, 2008), c'est-à-dire des expressions poly-lexicales dont les mots entretiennent une relation lexico-syntaxique. Mais, comme nous l'avons déjà précisé antérieurement, le dictionnaire de collocations n'apparaît plus dans le système d'alignement lexical final, à cause de l'amélioration négligeable de ses résultats, révélée par l'évaluation du système d'alignement présentée plus loin (cf. sous-section 4.4.4).

L'architecture du système d'alignement lexical français - roumain développé est présentée dans la Figure 27 suivante :

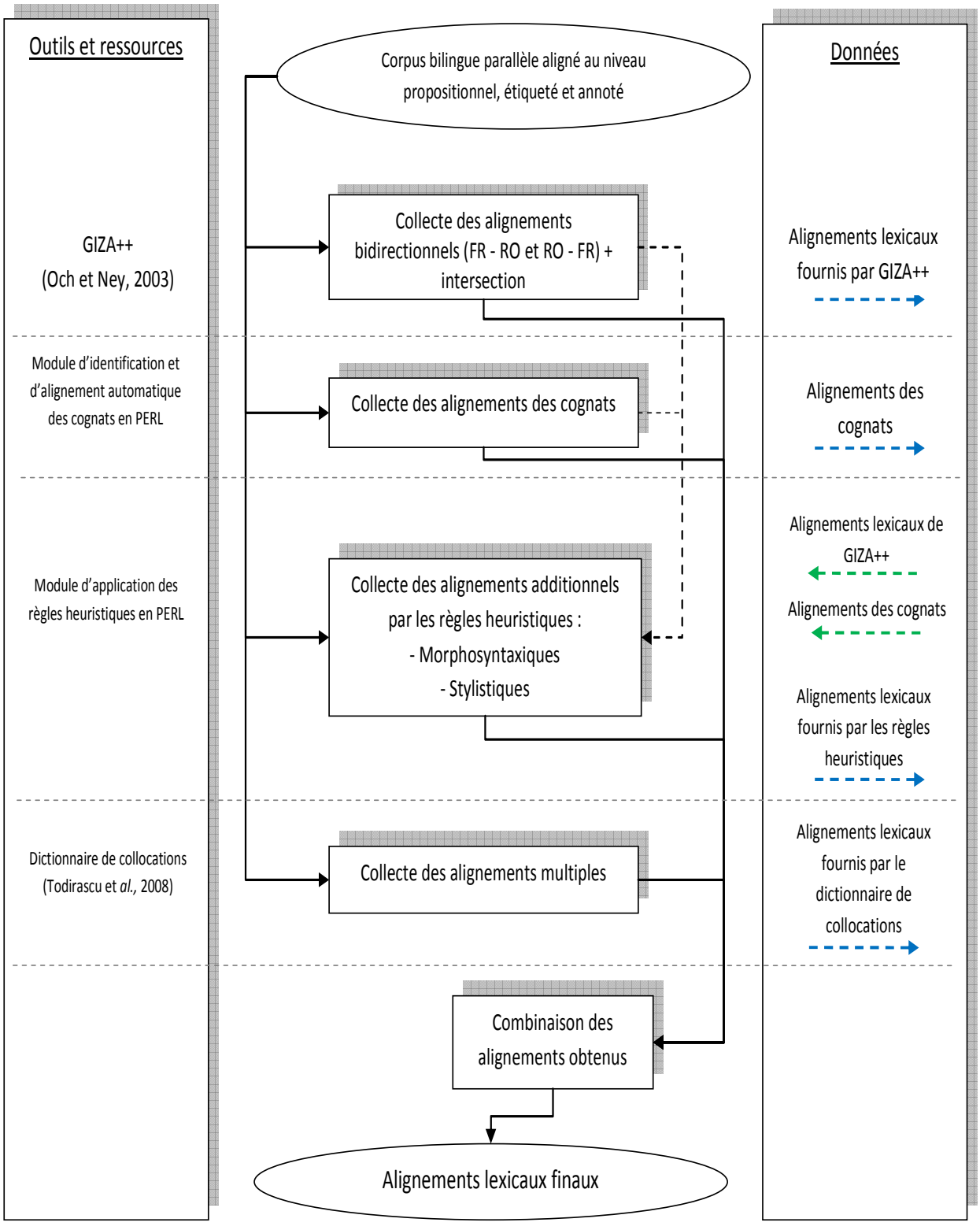


Figure 27. L'architecture du système d'alignement lexical français - roumain

Pour construire notre système d'alignement lexical, nous avons suivi les cinq étapes ci-dessous :

1. D'abord, l'application de l'outil statistique *GIZA++* (Och et Ney, 2000, 2003), présenté dans la sous-section 3.1.2., afin de construire le système d'alignement lexical de base ; Cet outil est appliqué dans les deux directions du processus d'alignement (FR - RO et RO - FR). Ensuite, l'intersection des alignements bidirectionnels (Koehn *et al.*, 2003) est obtenue afin d'augmenter la précision du système d'alignement lexical de base.

2. Puis, l'analyse linguistique des erreurs de l'alignement lexical de base ; Cette étape permet de définir des règles heuristiques et des stratégies pour corriger une partie de ces erreurs et augmenter ainsi la performance du système. Plusieurs catégories d'erreurs ont été identifiées et concernent principalement les cognats, les erreurs produites à cause des différences morphosyntaxiques entre les deux langues étudiées, les contraintes stylistiques en traduction et les collocations.

3. Ensuite, l'identification et l'alignement automatique des cognats (Navlea et Todirașcu, 2011abe, 2012) ; Cette étape est nécessaire car *GIZA++* ne détecte pas tous les cognats existants dans le corpus et ne dispose pas non plus de règles spécifiques dans ce but.

4. Après, l'application d'un ensemble de règles heuristiques morphosyntaxiques (Navlea et Todirașcu, 2010ab, 2011cd) et stylistiques (Navlea et Havași, 2012) pour compléter les alignements à l'intérieur des *chunks* (Tufiș *et al.*, 2006) ;

5. Enfin, l'application d'un dictionnaire de collocations (Todirașcu *et al.*, 2008) afin de corriger les alignements au niveau des collocations et d'étudier son impact sur les résultats du système d'alignement lexical final.

Le système d'alignement lexical de base ainsi que son évaluation feront l'objet de la section suivante. Ensuite, l'analyse linguistique des erreurs du système, suivie d'une étude du corpus parallèle au niveau de la traduction humaine, y seront également présentées. Nous avons effectué ces travaux afin de révéler les erreurs du système d'alignement lexical de base dues à des phénomènes linguistiques ou traductionnelles divers et de proposer des solutions pour résoudre une partie de ces erreurs et améliorer ainsi la performance du système.

4.1. Le système d'alignement lexical de base

Par alignement lexical de base, nous entendons un alignement mot-à-mot fourni par une méthode purement statistique. Pour construire un tel système, nous avons utilisé l'aligneur statistique *GIZA++* (Och et Ney, 2000, 2003) (décrit dans la sous- section 3.1.2.). Comme il a été déjà vu auparavant, *GIZA++* implémente les modèles génératifs *IBM* (Brown *et al.*, 1993) proposant des alignements mot-à-mot. Dans l'optique de ces modèles, un seul mot de la langue source peut avoir zéro ou un seul équivalent dans la langue cible. En revanche, un seul mot de la langue cible peut correspondre à zéro, un seul ou plusieurs mots de la langue source. Par conséquent, ces modèles ne réalisent pas d'alignements multiples (plusieurs mots de la langue source qui peuvent correspondre à plusieurs mots de la langue cible). De ce fait, nous avons utilisé des règles heuristiques et un dictionnaire de collocations pour collecter de tels alignements dans le système d'alignement lexical développé.

Nous allons présenter dans la sous-section suivante comment le système d'alignement lexical a été construit.

4.1.1. La construction du système d'alignement lexical de base

Afin d'entraîner le système d'alignement lexical de base, nous avons exploité un corpus bilingue parallèle extrait à partir de *DGT-TM* (Steinberger *et al*., 2012) (cf. sous-section 3.2.1.). Le corpus utilisé comprend 64 923 paires de phrases alignées 1:1, soient 1 818 768 tokens pour le français et 1 554 233 tokens pour le roumain. Comme le système d'alignement lexical final est construit pour être intégré dans un système de traduction automatique statistique factorisée - *MOSES* (Koehn *et al.*, 2007) (cf. sous-section 3.1.1.) -, chaque phrase sélectionnée est complète, c'est-à-dire elle commence par une majuscule et finit par un signe de ponctuation. De plus, chaque phrase présente une longueur comprise entre 5 et 100 tokens comme requis par *MOSES.* Ce corpus d'entraînement est également tokenisé, lemmatisé et étiqueté[125] (cf. sous-section 3.2.1.).

Pour évaluer le système d'alignement lexical, nous avons utilisé un corpus de test de 1 000 phrases alignées 1:1 extraites aussi à partir de *DGT-TM* (Steinberger *et al*., 2012). Ce corpus

[125] Le prétraitement du corpus d'entraînement ainsi que l'application des étapes ultérieures de notre méthodologie d'alignement lexical sur ce corpus ont été réalisés avec la collaboration de GONCHAROVA Yuliya, étudiante en master de *Linguistique, Informatique, Traduction*, pendant son stage de 6 mois au sein de l'U.R. *LiLPa* de l'Université de Strasbourg. Nous remercions l'*U.R. LiLPa* pour le financement de ce stage.

de test comprend 33 036 tokens en français et 28 645 en roumain. À part sa taille réduite, il possède toutes les caractéristiques du corpus d'entraînement. Ainsi, celui-ci est tokenisé, lemmatisé et étiqueté, les phrases composantes sont complètes et limitées en longueur. Ce corpus ne fait pas partie du corpus d'entraînement.

Nous avons aligné manuellement le corpus de test au niveau lexical[126] au moyen de l'application *MtKit* (Tufiș *et al.*, 2005a) permettant la visualisation des alignements. Comme référence d'alignement, le guide proposé par Melamed (1998) dans le cadre du projet Blinker, pour la paire de langues français - anglais[127] a été utilisé. Ce guide présente comme règle générale d'alignement manuel - à côté des règles spécifiques proposées - la recherche des correspondances minimales (mot-à-mot) à l'intérieur des phrases parallèles de l'échantillon de référence construit. Cette règle est appliquée même dans le cas de la traduction oblique. Toutefois, en fonction des particularités morphosyntaxiques de la paire de langues étudiées ici, le français et le roumain, et aussi du domaine juridique et administratif du corpus utilisé, nous avons défini également des règles supplémentaires d'alignement lexical. Ces règles seront décrites en détail plus loin, dans les sous-sections 4.1.3. et 4.1.4.

Comme nous avons effectué aussi des expériences intégrant un dictionnaire de collocations *Verbe + Nom* (Todirașcu *et al.*, 2008) dans le processus d'alignement lexical, deux corpus de référence alignés manuellement ont été construits : l'un qui ne contient pas les alignements des collocations (dénommé *RefA*) et un autre qui contient ces alignements (dénommé *RefB*). Pour ce faire, nous avons validé manuellement un ensemble de 470 paires d'équivalents collocationnels existants dans le corpus dont au moins un était une collocation de type *Verbe + Nom* dans la paire (voir sous-section 4.4.4. pour plus de détails concernant les statistiques liées aux collocations dans le corpus de référence).

Pour construire le système d'alignement lexical de base, nous avons préparé, tout d'abord, le corpus utilisé dans le format d'entrée requis par *GIZA++*. De plus, à la place des formes des mots, ont été utilisés les lemmes afin de réduire la dispersion des données dans le corpus d'entraînement, car celle-ci pose problèmes aux méthodes statistiques dans le cas des langues riches morphologiquement (cf. chapitre 2, sous-section 2.2.2.6.), comme le sont le français et

[126] Afin d'aligner manuellement le corpus de test, nous avons collaboré avec HAVAŞI Sebastian-Flaviu, étudiant en master de Traductologie à l'Université Babeș-Bolyai de Cluj-Napoca, Roumanie, pendant son stage de 4 mois au sein de l'*U.R. LiLPa* de l'Université de Strasbourg.
[127] http://archive.org/stream/arxiv-cmp-lg9805004/cmp-lg9805004#page/n3/mode/2up

le roumain. En effet, les formes fléchies des mots ont des distributions variées dans les langues source et cible. Par conséquent, les formes fléchies moins fréquentes et leurs traductions auront des probabilités faibles et ne seront pas alignées. Si l'on regroupe toutes les formes fléchies d'un mot sous un même lemme, on augmente ainsi leur probabilité de traduction. Rappelons que, de plus, à chaque lemme sont assignés les deux premiers caractères de l'étiquette morphosyntaxique afin de désambigüiser morphologiquement les lemmes (Tufiș *et al.*, 2005b, 2006) et améliorer ainsi la performance de *GIZA++*. Nous donnons un exemple à titre d'illustration : le même lemme (*écrit*) peut être un nom commun (*Nc*) ou un adjectif qualificatif participial (*Af*). Les deux premiers caractères de l'étiquette morphosyntaxique assignés distinguent entre les deux lemmes : *écrit_Nc* vs. *écrit_Af.*

Dans le but d'augmenter la précision du système d'alignement lexical de base, ont été réalisés des alignements bidirectionnels (FR - RO et RO - FR) avec *GIZA++* et, ensuite, leur intersection a été effectuée (Koehn *et al.*, 2003). Cette heuristique (l'intersection des alignements bidirectionnels) permet de conserver seulement les alignements considérés comme sûrs du fait qu'ils sont repérés dans les deux sens du processus de mise en correspondance automatique. C'est l'alignement obtenu à l'issue de cette étape qui a été donc retenu en tant qu'alignement lexical de base dans notre approche.

L'évaluation des résultats du système d'alignement lexical de base sera présentée dans la sous-section suivante.

4.1.2. L'évaluation du système de base

Le système d'alignement lexical développé a été évalué en termes de précision, rappel, F-mesure et score *AER - Alignment Error Rate* (Och et Ney, 2003), par comparaison avec nos deux alignements de référence constitués manuellement : respectivement *RefA* et *RefB*. Ces mesures d'évaluation se calculent généralement en fonction des alignements sûrs et probables qui peuvent se révéler pendant l'étape de constitution d'un corpus de référence. En effet, la tâche d'alignement manuel s'avère être assez difficile à cause des cas ambigus tels que l'alignement au niveau des paraphrases, explicitations, omissions de sens ou ajouts, etc. De ce fait, une distinction entre les alignements sûrs et probables s'avère nécessaire dans le processus d'évaluation automatique.

Ces mesures d'évaluation seront expliquées ci-dessous telles qu'elles sont utilisées afin d'évaluer des systèmes d'alignement lexical (Och et Ney, 2003 ; Fraser et Marcu, 2007) :

Soit *A* un alignement hypothétique, c'est-à-dire un ensemble de liens entre les mots des phrases sources et les mots des phrases cibles d'un corpus parallèle. Soit aussi un alignement de référence comprenant les alignements sûrs *S* et les alignements probables *P*. En fonction de ces paramètres, les formules de calcul des mesures mentionnées ci-dessus sont les suivantes :

1) Précision : $P_r = \frac{|A \cap P|}{|A|}$

La précision mesure la proportion entre les alignements probables fournis par le système, c'est-à-dire communs avec les alignements de référence probables, par rapport au nombre total d'alignements fournis par le système.

2) Rappel : $R = \frac{|A \cap S|}{|S|}$

Le rappel mesure la proportion entre les alignements sûrs fournis par le système, c'est-à-dire communs avec les alignements de référence sûrs, par rapport au nombre total d'alignements surs de référence existants.

Notons qu'afin d'assurer la cohérence des mesures, l'ensemble des alignements probables est défini comme l'union des alignements sûrs et probables (Allauzen et Wisniewski, 2009).

3) F-mesure : $F = \frac{2.P_r.R}{P_r + R}$

La F-mesure est obtenue par la combinaison de la précision et du rappel. Cette mesure se rapproche de la moyenne si la précision et le rappel sont rapprochés, tandis qu'elle diminue si la précision et le rappel sont éloignés.

4) *Alignment Error Rate* (*AER*) : $AER(A, S, P) = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$

Le score *AER* (Och et Ney, 2003) est dérivé à partir de la F-mesure. Ce score mesure la proportion entre le nombre total des alignements sûrs et probables fournis par le système y

compris les alignements fournis par le système et ceux de référence qui sont sûrs, par rapport au nombre total des alignements fournis par le système et ceux de référence qui sont sûrs. Mesurant un taux d'erreurs d'alignement, ce score doit être minimal pour que le système soit considéré performant.

Notons que s'il n'y a pas de différence entre les alignements sûrs et probables dans l'alignement de référence, alors l'*AER* devient le complémentaire de la F-mesure, étant obtenu par l'application de la formule suivante :

$$AER = 1\text{-F-mesure}$$

Comme l'alignement lexical manuel est une tâche assez coûteuse en temps et, de plus, comme pour bien distinguer entre les alignements sûrs et probables nous considérons qu'elle demande le concours de plusieurs spécialistes en traduction humaine français - roumain, nous n'avons pas fait de distinction entre les alignements sûrs et probables dans nos alignements de référence. Par conséquent, nous avons considéré tous les alignements comme sûrs afin de calculer le score *AER*. Ainsi, la formule du score *AER* retenue pour notre évaluation est la suivante :

$$AER = 1\text{-F-mesure}$$

Le Tableau 18 ci-dessous comprend les résultats d'évaluation du système d'alignement lexical de base en utilisant les deux corpus de référence, respectivement *RefA* (contenant les alignements des collocations) et *RefB* (sans alignement des collocations).

Tableau 18. L'évaluation du système d'alignement lexical de base

Système	**Précision (%)**		**Rappel (%)**		**F-mesure (%)**		***AER* (%)**	
	RefA	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***
Système de base (*GIZA++* et intersection)	95,56	95,51	52,91	49,71	68,11	65,39	31,89	34,61

Le système de base a obtenu une précision élevée (95,56% / 95,51%) mais un rappel faible (52,91% / 49,71%) qui ont conduit à un score *AER* élevé de 31,89% pour *RefA* et de 2,72% plus élevé pour *RefB* (34,60%).

Afin d'améliorer les résultats du système d'alignement lexical de base, nous avons procédé à l'analyse linguistique des erreurs d'alignement, en proposant également des solutions pour résoudre une partie de ces erreurs. De plus, nous avons réalisé une étude du corpus bilingue parallèle juridique au niveau de la traduction humaine afin de repérer des erreurs d'alignement dues aux procédés de traduction utilisés dans le corpus et nous avons proposé aussi des solutions pour diminuer le taux d'erreurs à ce niveau.

Dans la sous-section suivante, seront présentées l'analyse linguistique et la classification des erreurs de l'alignement lexical de base effectuées, tandis que l'étude du corpus bilingue parallèle juridique au niveau de la traduction humaine figurera dans la sous-section 4.1.4.

4.1.3. L'analyse linguistique et la classification des erreurs de l'alignement lexical de base

À l'issue de l'analyse linguistique des erreurs de l'alignement lexical de base, nous avons relevé plusieurs catégories d'erreurs systématiques, tout en proposant des solutions pour diminuer leur taux. Ces erreurs concernent principalement les cognats, les erreurs dues aux différences morphosyntaxiques entre les deux langues étudiées et aux contraintes stylistiques en traduction, les collocations et les termes poly-lexicaux du domaine.

Tout d'abord, pour résoudre une partie des erreurs d'alignement liées aux cognats, nous avons développé un module d'identification et d'alignement automatique des cognats (Navlea et Todirașcu, 2011abe, 2012) qui sera présenté dans la section 4.2.

Ensuite, afin de diminuer le taux d'erreurs apparues à cause des différences morphosyntaxiques entre les deux langues étudiées et à cause des contraintes stylistiques en traduction, un ensemble de règles heuristiques morphosyntaxiques (Navlea et Todirașcu, 2010ab, 2011cd) et stylistiques (Navlea et Havași, 2012) a été défini. Ces règles exploitent les étiquettes morphosyntaxiques et les lemmes du corpus et seront exemplifiées plus loin, dans la section 4.3.

Finalement, pour résoudre une partie des erreurs d'alignement liées aux collocations, nous avons appliqué un dictionnaire de collocations de type *Verbe + Nom* disponible pour le français et le roumain (Todirașcu *et al.*, 2008), en étudiant aussi son impact dans le processus d'alignement. Ce dictionnaire sera décrit dans la section 4.4. Nous avons aussi complété le

dictionnaire avec une classe spécifique de collocations *Nom 1 déverbal + (préposition) + Nom 2* (cf. sous-section 4.4.2.4.). Par contre, pour les autres classes de collocations (*Nom + Adjectif*, *Adverbe + Adjectif*, etc.), des ressources externes ne sont pas disponibles.

En outre, concernant les problèmes d'alignement apparaissant au niveau des termes poly-lexicaux du domaine, nous avons exploité les cognats identifiés (cf. sous-section 4.2.1.) afin d'aligner les termes contenant des cognats (cf. sous-section 4.2.3.).

La terminologie grammaticale utilisée dans ce travail repose en grande partie sur les travaux concernant la *Grammaire méthodique du français* (Riegel *et al*., 2009) et la *Gramatica limbii române* 'Grammaire de la langue roumaine' (coord. Guţu-Romalo, 2005).

Comme nous avons pu le constater lors de l'analyse linguistique des erreurs du système d'alignement lexical de base, des erreurs fréquentes apparaissent à cause des différences morphosyntaxiques entre le français et le roumain. Malgré le fait que les deux langues étudiées sont des langues latines, elles présentent des différences morphosyntaxiques importantes qui seront discutées ci-dessous.

Par rapport au français, en plus du genre et du nombre, le nom du roumain possède aussi la catégorie grammaticale du cas : nominatif, accusatif, génitif, datif, vocatif. Le cas représente « la catégorie grammaticale qui exprime les rapports et les fonctions syntaxiques du nom dans le cadre de l'énoncé » (*GLR*[128], 2005 : 70). Ainsi, le nominatif exprime la fonction syntaxique prototypique de sujet, l'accusatif de complément direct, le génitif d'attribut génitival (complément du nom en français) et le datif de complément indirect. Un nom en vocatif ne présente pas de fonction syntaxique. Le cas se réalise par des désinences spécifiques (dans le paradigme flexionnel) et par des morphèmes proclitiques (p. ex. *lui* marquant le génitif-datif : *zilele* ***lui*** *august* 'les jours de l'août' ; *al, a, ai, ale* marquant le génitif : *un caiet* ***al*** *elevului* 'un cahier de l'élève') ou affixes proclitiques (p. ex. *pe* provenu de la préposition grammaticalisée *pe* 'sur' qui marque l'accusatif exprimant le complément direct : *îl văd* ***pe*** *tata* 'je vois mon père') (*GLR*, 2005 : 83-84). Le cas est aussi spécifique aux parties du discours appartenant à la classe du nom : le pronom, l'adjectif et le numéral. Concernant le genre, le roumain présente aussi le genre neutre.

[128] *Gramatica limbii române*, volume I, *Cuvântul* 'Le mot'.

Le déterminant défini (de type article) du français est toujours proclitique et forme un mot séparé (***la*** *directive*) alors qu'en roumain il est toujours enclitique et fusionne avec le nom (*directivă vs. directiv**a**)* ou l'adjectif antéposé (p. ex. *prezent**a** directivă* où le déterminant défini *a* est accolé à l'adjectif *prezentă* qui fusionne en *prezent**a***).

D'autres différences concernent les morphèmes supplémentaires ou différents d'une langue à l'autre (prépositions, particules d'infinitif, etc.) ou les structures syntaxiques propres à chaque langue. Par exemple, en français, l'utilisation du pronom clitique empêche l'occurrence simultanée du complément direct ou indirect dans l'énoncé (*Le Conseil **l**'a nommé comme président.* vs. *Le Conseil nomme comme président **l'ancien chef.***). En revanche, les clitiques à l'accusatif et au datif du roumain s'expriment simultanément avec le complément direct ou indirect (*Comisia **l**-a numit președinte **pe fostul șef** al misiunii de pace*). En français (langue non pro-drop), le pronom sujet accompagne généralement le verbe, tandis qu'en roumain (langue pro-drop), le plus souvent, il ne se réalise pas lexicalement, étant inclus dans la terminaison du verbe. Concernant l'ordre des mots dans la phrase, le français présente plutôt un ordre standard (Sujet - Verbe - Complément), alors qu'en roumain l'ordre des mots est généralement libre.

Le Tableau 19 suivant présente les classes d'erreurs les plus fréquentes dans le corpus de test étudié, repérées au niveau morphosyntaxique, ainsi que leur distribution par des classes *ad hoc*. Ces résultats permettent de connaitre les classes d'erreurs prédominantes auxquelles il appartiendra d'accorder la priorité afin de diminuer efficacement le taux d'erreurs d'alignement lexical.

Tableau 19. Distribution par classes des erreurs fréquentes d'alignement lexical au niveau morphosyntaxique

Classes d'erreurs d'alignement lexical français - roumain		Distribution (%)
1	Déterminants	42,77
2	Collocations, termes poly-lexicaux	26,22
3	Compléments du nom introduit par *de* (FR) vs. Noms au génitif (RO)	11,07
4	Pronoms relatifs	6,25
5	Modes et temps verbaux	5,13
6	Négation	1,57
7	Compléments d'objet indirect / objet second introduits par *à* (FR) vs. Noms au datif (RO)	1,17
8	Pronoms adverbiaux *en* et *y*	0,49
9	Déterminants numéraux ordinaux	0,34
10	Autres	5,00

Les catégories d'erreurs apparaissant dans le Tableau 19 ci-dessus seront discutées en ordre décroissant selon leur fréquence, dans les sous-sections qui suivent.

4.1.3.1. Déterminants

Tout d'abord, nous avons remarqué que les erreurs les plus fréquentes appartiennent à la classe des déterminants - définis (de type article) et possessifs - qui représentent environ 43% du nombre total d'erreurs dans le corpus étudié. De ce fait, nous nous sommes d'abord concentrés sur la résolution de cette classe d'erreurs, en appliquant les règles heuristiques morphosyntaxiques définies dans ce sens. Ci-dessous figurent des exemples pour illustrer cette classe d'erreurs d'alignement lexical.

Comme il a été déjà précisé auparavant, le déterminant défini (de type article) du français (*le, la, les*) est toujours proclitique, tandis qu'en roumain celui-ci (*-l,-a, -i, -le, -lui, -lor*) est toujours enclitique et fusionne avec le nom ou l'adjectif antéposé. De ce fait, le déterminant défini du français n'est pas aligné avec son correspondant roumain. Dans la Figure 28 suivante, est donné un exemple d'erreur concernant l'alignement du déterminant défini *le* du français avec *-l* du roumain, fusionné avec le nom (*text* vs. *textu**l*** 'texte') au nominatif-accusatif. Cet exemple d'alignement est marqué dans la Figure 28 par une ligne pointillée,

comme d'ailleurs tous les autres exemples d'alignement erronés discutés ici. La règle morphosyntaxique définie pour résoudre ce type d'erreur y figure aussi.

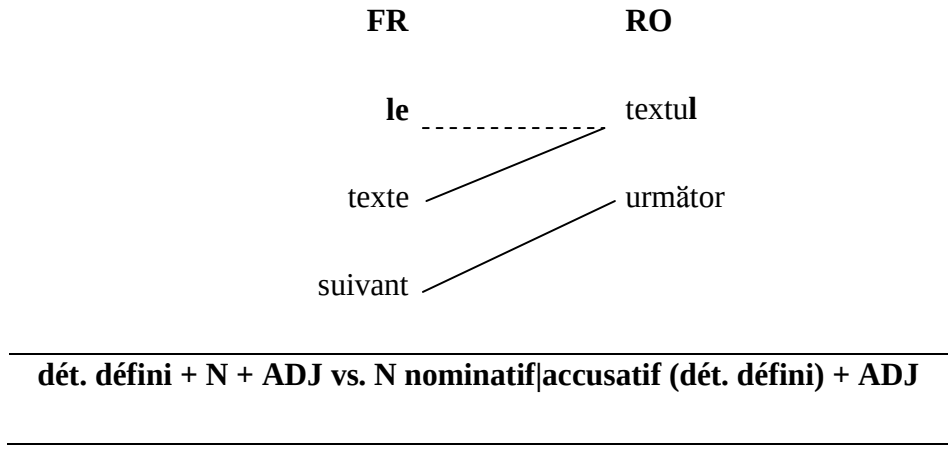


Figure 28. Alignement des déterminants définis du français avec le nom du roumain

Toutes les règles présentées ici décrivent les conditions morphosyntaxiques à remplir par les contextes bilingues français - roumain afin de pouvoir corriger un alignement erroné. Par exemple, la règle donnée dans la Figure 28 est interprétée ainsi : si dans la langue source (français) un déterminant défini est suivi d'un nom qui est suivi d'un adjectif et, en même temps, dans la langue cible (roumain) un nom dans le cas nominatif-accusatif déterminé défini est suivi d'un adjectif, alors le déterminant défini du français s'aligne avec le nom du roumain. Les abréviations et les signes utilisés dans l'écriture de ces règles heuristiques sont explicités dans l'Annexe 4. Précisons que ces règles ont été transposées aussi en pseudo-code (cf. sous-section 4.3.1.), ensuite implémentées en Perl.

La Figure 29 suivante illustre un exemple d'erreur où le déterminant défini *le* du français n'est pas aligné avec le déterminant défini *-l* du roumain accolé à l'adjectif antéposé (*următor* vs. *următorul* 'suivant') au nominatif - accusatif. La règle morphosyntaxique définie pour résoudre ce type d'erreur est aussi indiquée dans la figure.

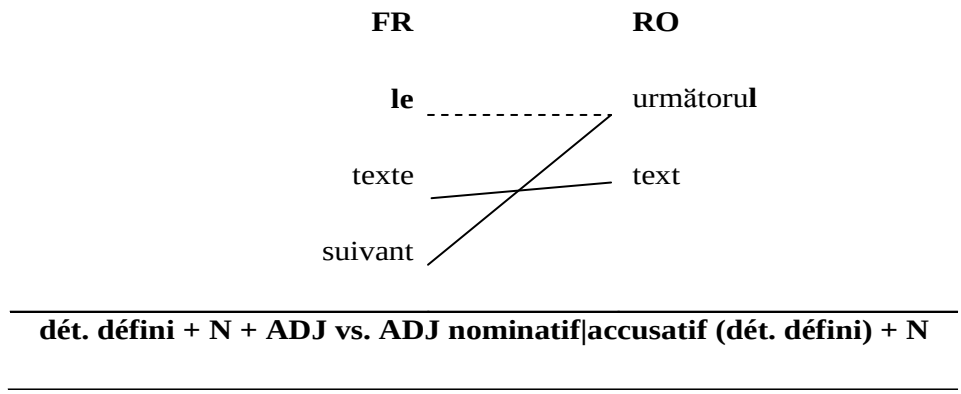


Figure 29. Alignement des déterminants définis du français avec l'adjectif du roumain

Des règles similaires ont été définies pour aligner le déterminant défini du français avec également le nom au génitif-datif ainsi qu'avec l'adjectif antéposé au génitif-datif.

Concernant le déterminant possessif, celui-ci précède le nom en français, alors qu'en roumain, dans les corpus du domaine juridique et administratif, il suit toujours le nom. Dans notre corpus spécialisé, seulement les formes de la troisième personne du singulier (*sa*, *son* vs. *sa*, *său*, *sale*) ou du pluriel (*ses*, *leur*, *leurs* vs. *lor*) sont présentes. Ces déterminants ne sont pas systématiquement alignés. La Figure 30 suivante illustre un exemple de ce type d'erreur et la règle morphosyntaxique correspondante définie pour les aligner.

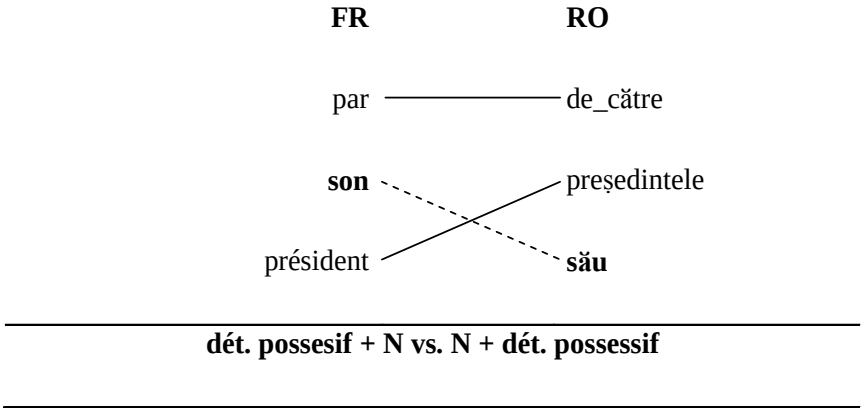


Figure 30. Alignement des déterminants possessifs présents dans les deux langues

Souvent, le déterminant possessif du français n'apparaît pas traduit en roumain (*à la date de* ***sa*** *publication au journal… vs. la data publicări****i*** *în Jurnalul…*) et, par conséquent, le déterminant possessif du français n'est pas aligné. Nous illustrons ce type d'erreur dans la Figure 31 ci-dessous ainsi que la règle morphosyntaxique définie afin d'aligner le déterminant

possessif *sa* avec le nom correspondant (*publicării*) du nom français qu'il détermine (*publication*).

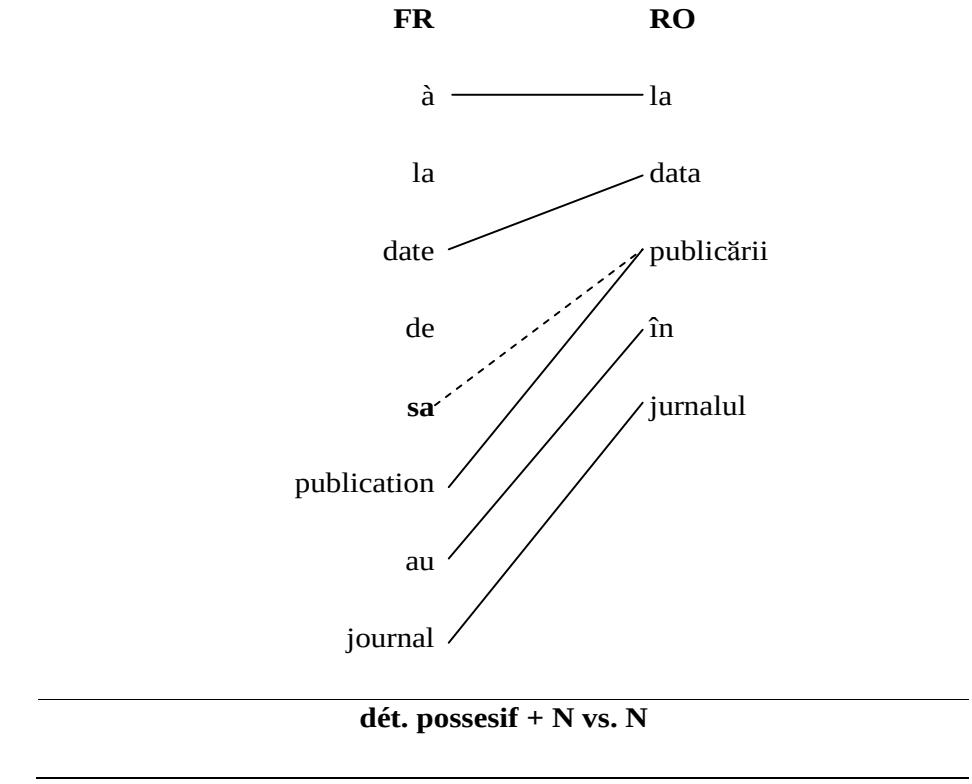


Figure 31. Alignement du déterminant possessif présent en français avec le nom du roumain

Nous avons identifié des cas où le déterminant possessif du français est traduit en roumain par un pronom réfléchi (forme non accentuée) au datif possessif (…*les groupements ont* ***leu****r siège…* vs. …*grupările* ***își*** *au sediul…*). La séquence roumaine …*grupările își au sediul…* peut être paraphrasée par …*grupările au sediul* ***lor****…* 'les groupements ont **leur** siège' où *grupările* 'les groupements' représente le possesseur et *sediul* 'siège' est l'objet possédé. À cause de cette différence, le déterminant possessif du français *leur* et le pronom au datif possessif du roumain *își* restent non-alignés. La Figure 32 ci-dessous montre ce cas d'erreur d'alignement et donne la règle morphosyntaxique de correction proposée. Cette règle aligne le déterminant possessif *leur* avec le pronom *își*.

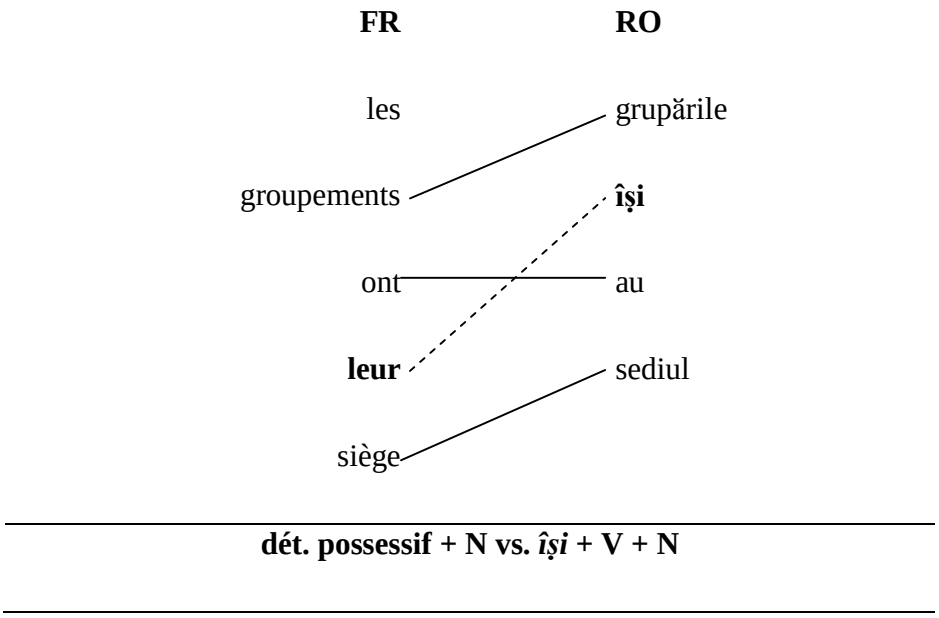


Figure 32. Alignement du déterminant possessif du français avec le datif possessif du roumain

4.1.3.2. Collocations et termes poly-lexicaux

Une autre classe d'erreurs significative (approximativement 26%) est représentée par l'ensemble des erreurs apparaissant au niveau des collocations et des termes poly-lexicaux présents dans notre corpus juridique et administratif. Il s'agit de structures poly-lexicales plus ou moins compositionnelles qui ne sont pas alignées en bloc et qui posent problème pour l'alignement lexical à cause du manque de ressources externes (dictionnaires de collocations, bases de données terminologiques) pour le français et le roumain.

Rappelons que, dans notre approche, nous entendons par collocations les expressions poly-lexicales dont les mots entretiennent une relation lexico-syntaxique (Todirașcu *et al.*, 2008). Les collocations peuvent avoir comme équivalents de traduction aussi bien des collocations qu'une seule unité lexicale. Par exemple, une collocation *Verbe + Nom* comme *avoir le droit* a comme équivalent roumain la collocation *a avea dreptul*, mais *procéder à l'examen* se traduit par le verbe *a examina* 'examiner'. De plus, une collocation verbo-nominale comme *mettre en application* peut être traduite par sa variante nominalisée *punerea în aplicare* 'la mise en application' (on parle alors de transposition - voir sous-section 4.1.4. qui présente les procédés de traduction rencontrés dans le corpus).

Dans notre alignement lexical de base, les collocations ne sont pas alignées en bloc. Par conséquent, des unités lexicales appartenant aux collocations restent non-alignées. Dans

l'exemple de la Figure 33 (… *la Commission n'a pas pris de mesures raisonnables*… vs. … *Comisia nu a luat măsurile necesare*…), le déterminant indéfini *de* appartenant à la collocation verbo-nominale *ne pas prendre de mesures* (forme négative) n'est pas aligné.

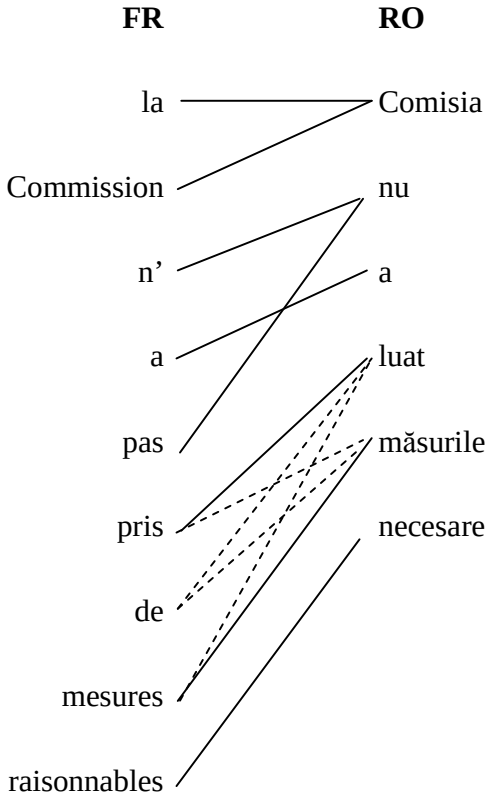


Figure 33. Alignement en bloc des collocations

Ainsi, pour résoudre les erreurs d'alignement au niveau des collocations, des ressources externes comme les dictionnaires sont nécessaires. Certes, il existe des méthodes purement statistiques (cf. section 4.4.) pouvant s'appliquer *ad hoc* à partir d'un corpus afin d'identifier les collocations vues comme des cooccurrences fréquentes de mots (Cowie, 1981 ; Quasthoff, 1998). Ces méthodes calculent des cooccurrences fréquentes de mots susceptibles de former des collocations, par l'utilisation des scores d'association entre les mots d'un corpus. Mais comme nous partons de l'idée que la distinction entre les collocations et les simples cooccurrences se fait uniquement manuellement selon des critères sémantiques (Todirașcu *et al.*, 2008), nous avons choisi d'utiliser un dictionnaire afin de corriger les erreurs survenues au niveau des collocations.

Le seul dictionnaire disponible pour la paire de langues étudiées est celui de Todirașcu *et al.* (2008) contenant des collocations de type *Verbe + Nom* et leurs divers équivalents de traduction. Nous avons appliqué ce dictionnaire en étudiant aussi son impact dans le processus d'alignement lexical. La description détaillée du dictionnaire ainsi que l'algorithme d'alignement des collocations figurent plus loin, dans la section 4.4 dédiée en entier à ce type de structures poly-lexicales. Pour les autres classes de collocations (*Adverbe + Adjectif, Nom + Adjectif*, etc.), des ressources externes ne sont pas disponibles pour la paire de langues étudiées. Néanmoins, nous avons proposé une méthode pour compléter le dictionnaire avec les collocations *Nom 1 déverbal + (préposition) + Nom 2* (cf. sous-section 4.4.2.4.).

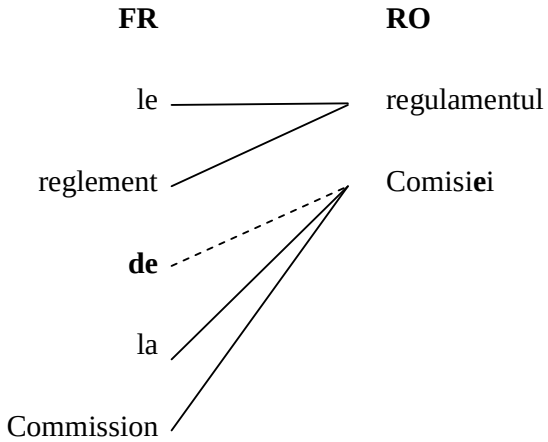
Concernant les termes poly-lexicaux, ils représentent, dans notre travail, des « combinaisons lexicales spécialisées », c'est-à-dire « des groupes associés à un domaine de connaissances » (L'Homme et Meynard, 1998 : 201). Comme le tri entre les termes poly-lexicaux et les collocations de la langue générale suppose la connaissance de la structure conceptuelle du domaine de spécialité et nécessite, en conséquence, la validation des résultats par un spécialiste du domaine, nous ne nous proposons pas de faire ici une distinction nette entre ceux-ci. Le seul moyen que nous avons trouvé pour collecter des alignements au niveau des termes poly-lexicaux, et diminuer ainsi les erreurs d'alignement à leur niveau, est l'exploitation des termes poly-lexicaux que l'étiqueteur *TTL* (Ion, 2007) reconnaît et marque dans le corpus roumain (p. ex. *integrare_economică* 'intégration économique', *organizarea_comună_a_pieţelor* 'l'organisation commune des marchés') et l'ensemble des cognats identifiés dans le corpus (Navlea et Todirașcu, 2012). En effet, à partir de ces termes poly-lexicaux du roumain, nous avons recueilli des alignements au niveau des termes poly-lexicaux correspondants du français, par l'intermédiaire des cognats existants dans une paire bilingue de termes, comme décrit dans la sous-section 4.2.3.

4.1.3.3. Compléments du nom introduit par *de* (FR) *vs.* Noms au génitif (RO)

Des erreurs d'alignement lexical fréquentes (approximativement 11%) apparaissent également au niveau du complément du nom introduit par la préposition *de* en français (*Nom 1* ***de Nom 2***), ayant comme correspondant en roumain un nom au génitif. Précisons que la préposition *de* du français peut apparaître fusionnée avec le déterminant défini *le* (singulier) ou *les* (pluriel) en donnant les formes *du* ou *des*. Le nom au génitif du roumain (*Nom 2*) peut

être marqué de plusieurs façons : par désinence spécifique et par morphèmes supplémentaires de génitif. Dans le corpus étudié, nous avons identifié les cas suivants :

1) Le *Nom 1* est au nominatif-accusatif et toujours déterminé défini tandis que le *Nom 2* est au génitif marqué par désinence et déterminé défini (p. ex. *le règlement* ***de*** *la Commission* vs. *regulamentu**l** Comisi**ei***) ; À cause de cette différence, la préposition *de* et, le cas échéant, les formes *du* ou *des* du français ne sont pas alignées avec le nom au génitif du roumain. La Figure 34 ci-dessous illustre ce type d'erreur d'alignement lexical et donne aussi la règle morphosyntaxique définie pour pallier à ce problème.



de + dét. déf. + N vs. N (nominatif - accusatif, dét. déf.) + N (génitif, dét. déf.)

Figure 34. Alignement de la préposition *de* du français avec le nom au génitif du roumain

2) Le *Nom 2* au génitif du roumain est, de plus, précédé par un morphème de génitif : *al* (singulier, masculin), *a* (singulier, féminin), *ai* (pluriel, masculin), *ale* (pluriel, féminin). Ce morphème apparaît seulement dans certaines conditions syntaxiques, comme suit :

    a) le *Nom 1* présente un déterminant indéfini (p. ex. ***une*** *diminution* ***de*** *la franchise* vs. ***o*** *descreștere* ***a*** *scutirii*) (voir Figure 35 suivante) ;

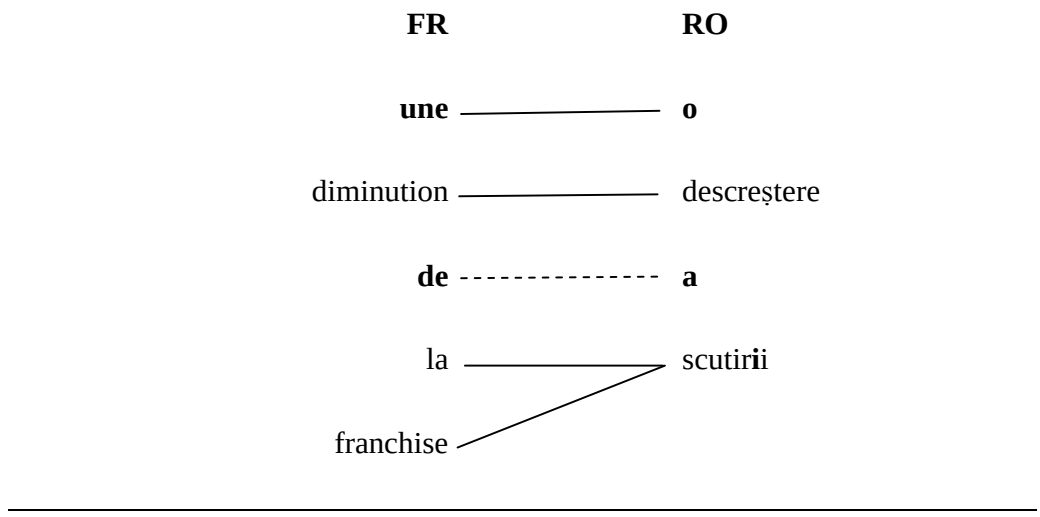


Figure 35. Alignement de la préposition *de* du français avec le morphème de génitif du roumain (a)

b) le *Nom 1* est suivi d'un ou plusieurs adjectifs (*l'activité* ***professionnelle de*** *la personne* vs. *activitatea* ***profesională a*** *persoanei*) ;

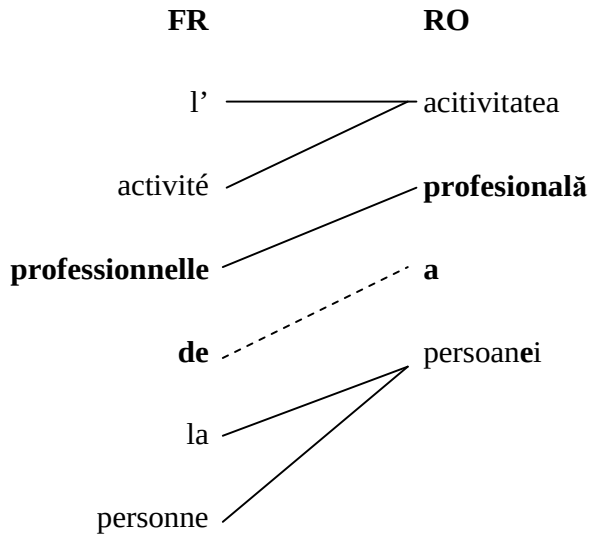


Figure 36. Alignement de la préposition *de* du français avec le morphème de génitif du roumain (b)

c) le *Nom 1* présente le déterminant zéro (p. ex. *le comité de* ***gestion du*** *lait* vs. *Comitetul de* ***gestionare a*** *laptelui*).

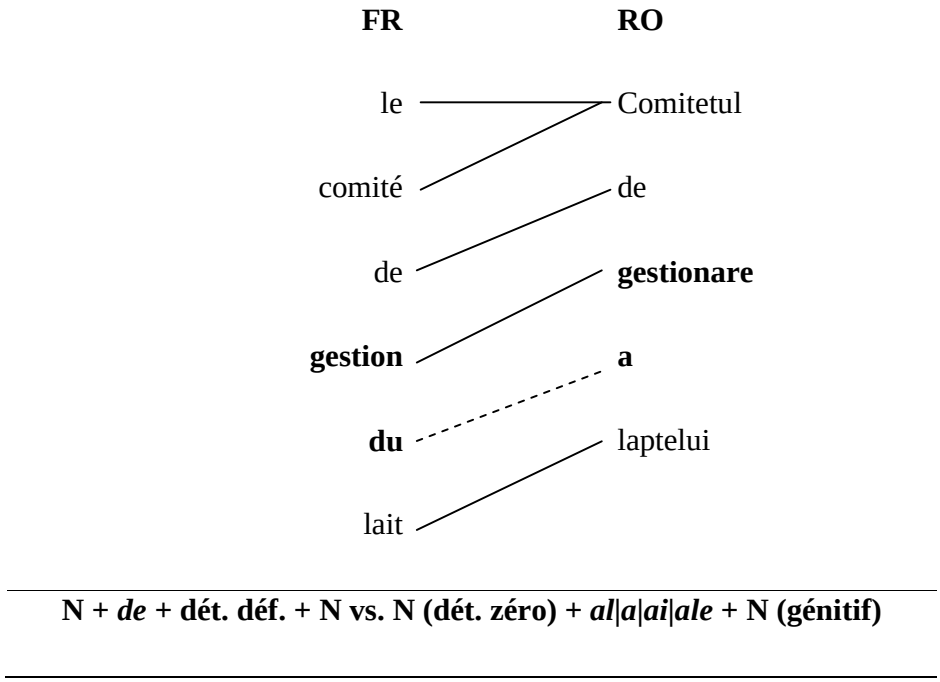
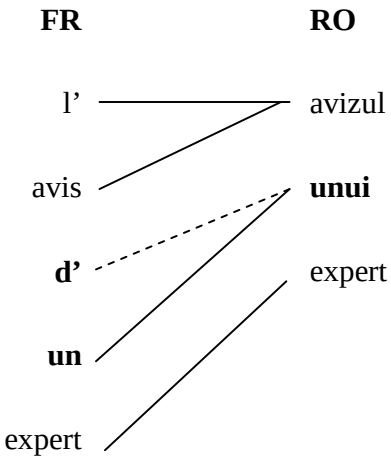


Figure 37. Alignement de la préposition *de* du français avec le morphème de génitif du roumain (c)

Concernant ces trois cas discutés ci-dessus, la préposition *de* et, éventuellement, les formes *du* ou *des* du français et le morphème de génitif du roumain ne sont pas alignés dans le corpus étudié. Les Figures 35, 36, 37 ci-dessous présentent des exemples de ce type d'erreurs et aussi les règles morphosyntaxiques permettant de solutionner ces problèmes.

3) Lorsque le *Nom 2* du roumain est précédé par un déterminant indéfini, de type article, ayant la forme de génitif (*unui* 'd'un', *unei* 'd'une', *unor* 'de/des'), la préposition *de* du français n'apparaît pas alignée avec le déterminant indéfini du roumain. Un exemple de ce type d'erreur ainsi que la règle morphosyntaxique correspondante sont illustrés dans la Figure 38 suivante.



N + *de* + dét. indéf. + N vs. N + dét. indéfini, forme de génitif + N (génitif)

Figure 38. Alignement de la préposition *de* du français avec le déterminant indéfini, forme de génitif, du roumain

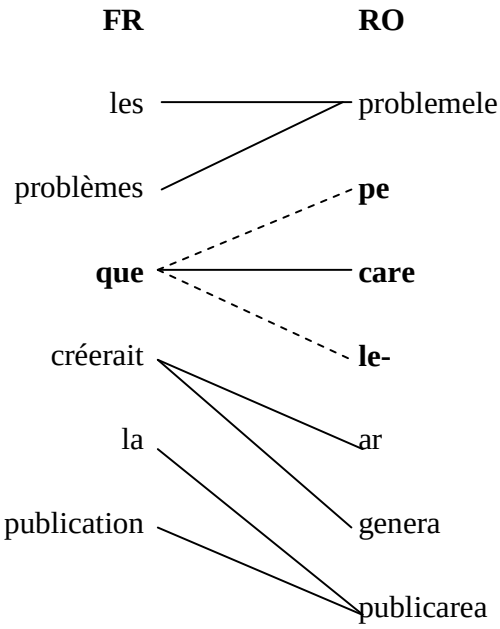
De même, quand le *Nom 2* du roumain est précédé par un adjectif au génitif (p. ex. *les dispositions **de** la **présente** directive* vs. *dispozițiile **prezentei** directiv**e***), un déterminant démonstratif (au génitif) (p. ex. *les dispositions **de cette** directive* vs. *dispozițiile **acestei** directiv**e***) ou un déterminant indéfini (au génitif) (p. ex. *les dispositions **de chaque** directive* vs. *dispozițiile **fiecărei** directiv**e***), la préposition *de* du français n’apparaît pas alignée avec l’adjectif ou le déterminant du roumain. Pour ces cas, des règles morphosyntaxiques de correction ont été aussi définies.

4.1.3.4. Pronoms relatifs

La classe d’erreurs d’alignement lexical suivante (approximativement 6%) est représentée par les erreurs apparues au niveau de l’alignement des pronoms relatifs compléments d’objet direct ou indirect.

Dans les subordonnées relatives du français, le complément d’objet direct est exprimé par le pronom relatif *que* alors qu’en roumain celui-ci est doublement exprimé par le pronom relatif *care*, à l’accusatif, toujours précédé par l’affixe *pe*, et le pronom personnel *îl, -l, o, îi, i, le* (formes non-accentuées). À cause de cette différence, dans l’exemple de la Figure 39, le pronom *que* du français n’est pas aligné avec l’affixe *pe* et le pronom personnel *le* du roumain. Il s’agit de l’analyse des séquences bilingues parallèles …*les problèmes **que** créerait*

la publication… vs. … *problemele* ***pe care le****-ar genera publicarea…*). La Figure 39 donne aussi la règle morphosyntaxique de correction définie.



N + *que* + V vs. N + *pe* + *care* (accusative) + *îl* | *-l* | *o* | *îi* | *-i* | *le* + V

Figure 39. Alignement du pronom relatif *que* du français avec le morphème *pe* et le pronom personnel *le-* du roumain

Concernant le complément d'objet indirect, dans les subordonnées relatives du français, il s'exprime à l'aide de la préposition *à* ou *auprès de* et le pronom relatif *lequel* pendant qu'en roumain, celui-ci est doublement exprimé par le pronom relatif *care*, au datif, et les formes non-accentuées du pronom personnel *îi, -i, le, li.* À cause de cette différence, dans l'exemple donné dans la Figure 40 suivante (… *les activités* ***auxquelle****s s'applique la présente directive…* vs. … *activităţile* ***cărora li*** *se aplică prezenta directivă…*), le pronom relatif *auxquelles* (forme contractée de *à + lesquelles*) du français n'est pas aligné avec le pronom personnel *li* du roumain. La Figure 40 comprend aussi la règle morphosyntaxique de correction proposée.

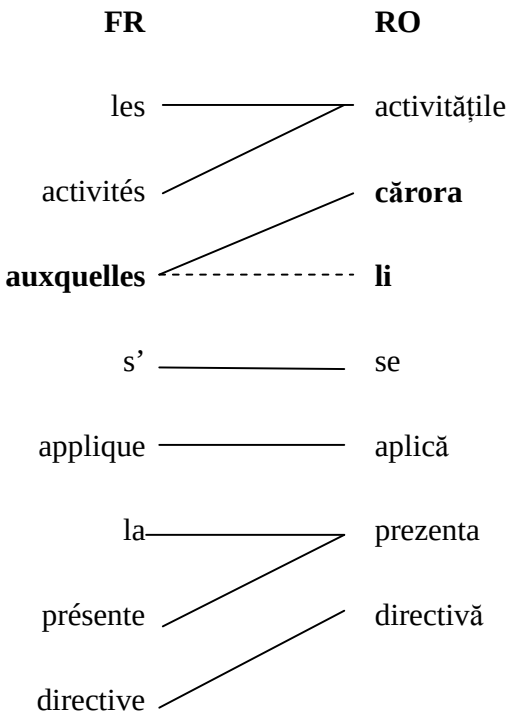


Figure 40. Alignement du pronom relatif *auxquelles* du français avec le pronom personnel *li* du roumain

Dans les cas ou le pronom relatif n’apparaît pas d’une langue à l’autre, il n’est pas aligné. De ce fait, nous avons défini aussi une règle morphosyntaxique qui aligne le pronom relatif présent dans une langue avec le nom correspondant à l’antécédent nominal qu’il reprend, dans l’autre langue. La Figure 41 suivante présente un cas d’alignement du pronom relatif *care* du roumain, en position de sujet, avec le nom *personnes* et son déterminant défini *les* du français.

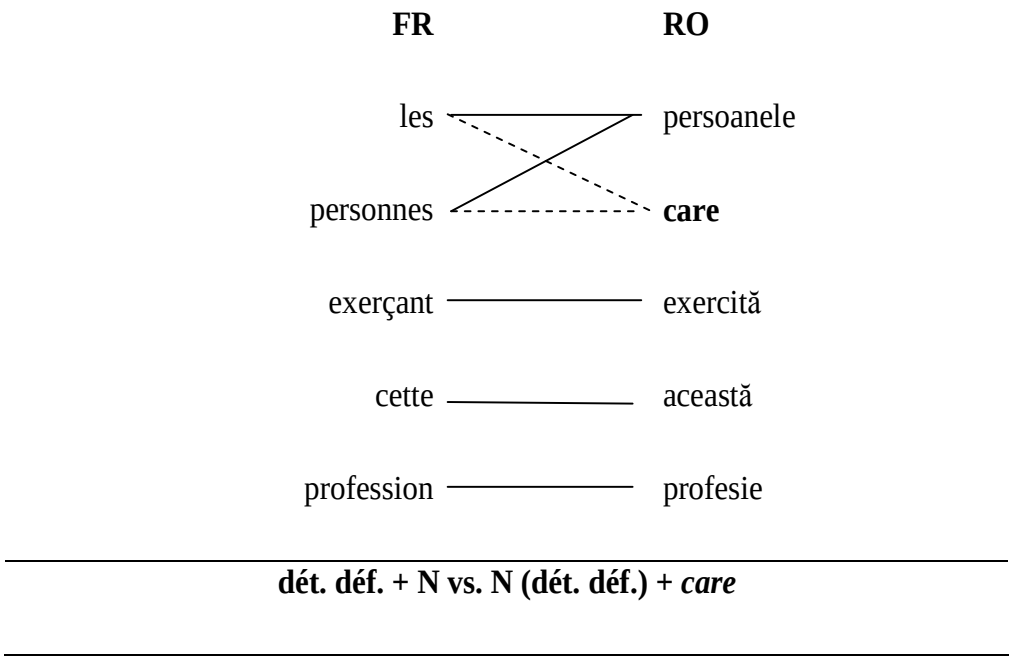


Figure 41. Alignement du pronom relatif *care* du roumain avec le nom correspondant à l'antécédent nominal qu'il reprend

4.1.3.5. Modes et temps verbaux

La classe d'erreurs d'alignement lexical suivante (environ 5%) concerne les erreurs survenues au niveau du verbe, et notamment dans l'expression de différents modes et temps verbaux, d'une langue à l'autre.

Ainsi, par rapport au français, l'infinitif roumain possède une particule supplémentaire *a* (p. ex. ***a*** *rămâne*) qui, évidemment, n'est pas alignée avec l'infinitif français (*demeurer*) dans le corpus étudié. Ci-dessous, la Figure 42 illustre ce type d'erreur et la règle morphosyntaxique de correction correspondante.

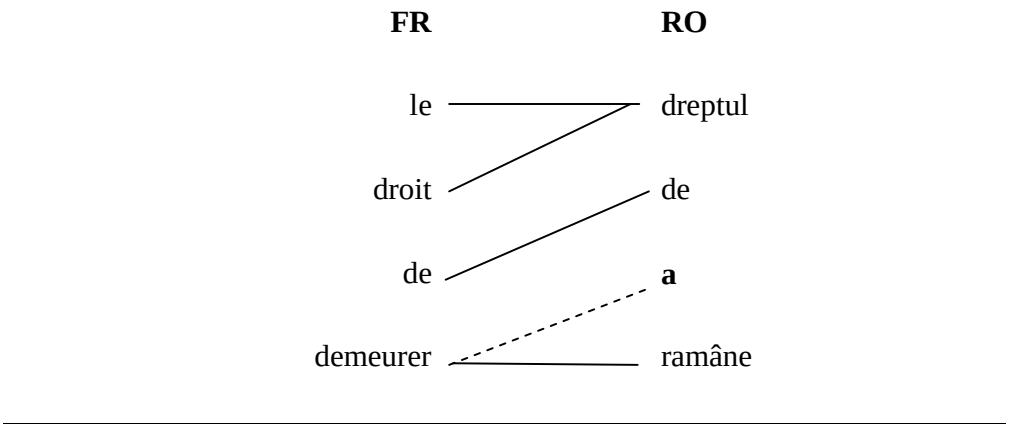


Figure 42. Alignement du l'infinitif français avec la particule supplémentaire d'infinitif *a* du roumain

Souvent, l'infinitif français (p. ex. *solliciter*) se traduit en roumain par le subjonctif (*să solicite*). Le subjonctif roumain possède, en plus, la particule *să*. Par conséquent, cette particule n'est pas alignée avec l'infinitif français. Un exemple d'erreur relevée au niveau de ces structures apparaît dans la Figure 43 ci-dessous, accompagné aussi de sa règle morphosyntaxique de correction.

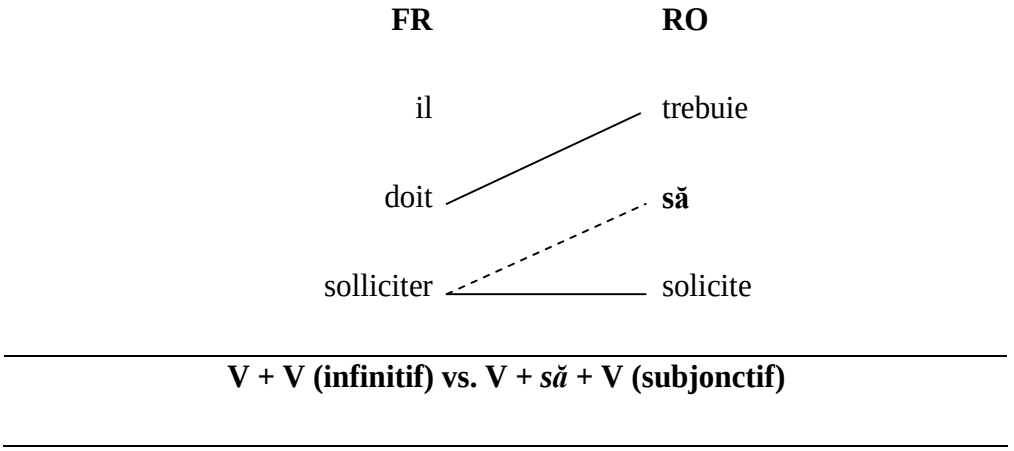


Figure 43. Alignement du l'infinitif français avec la particule du subjonctif *să* du roumain

Si l'infinitif français est précédé par la préposition *de*, cette préposition reste non-alignée dans le corpus. La règle morphosyntaxique proposée aligne la préposition *de* avec la particule du subjonctif *să* du roumain (voir Figure 44 suivante).

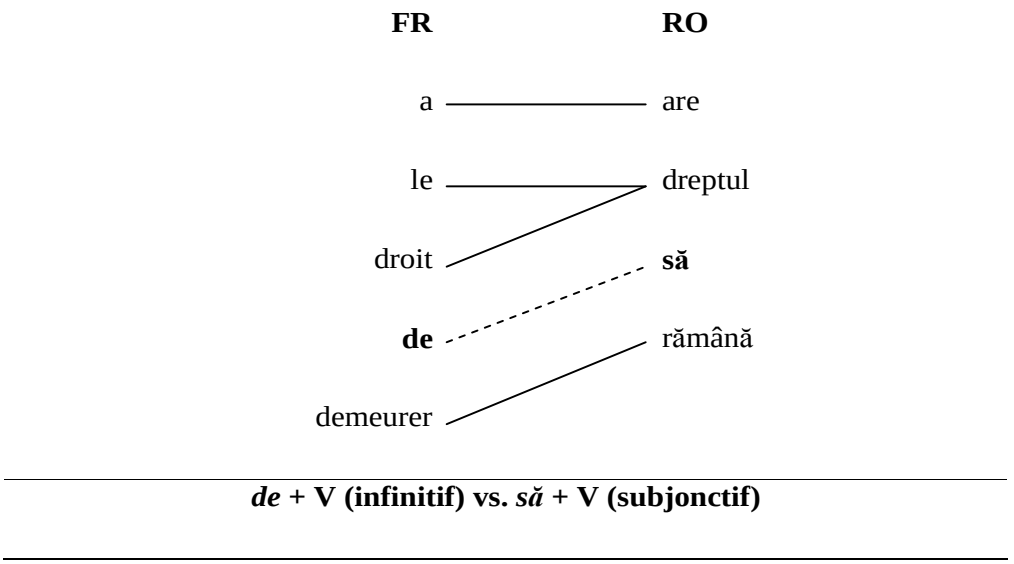


Figure 44. Alignement de la préposition *de* du français avec la particule du subjonctif să du roumain

Des erreurs d'alignement lexical apparaissent aussi au niveau des verbes réfléchis et notamment quand seulement un des verbes (source ou cible) est réfléchi. Dans ce cas, le pronom réfléchi *se* (toujours à la troisième personne du singulier ou du pluriel) n'est pas aligné avec le verbe correspondant. Dans la Figure 45 suivante en est illustré un exemple et la règle morphosyntaxique proposée.

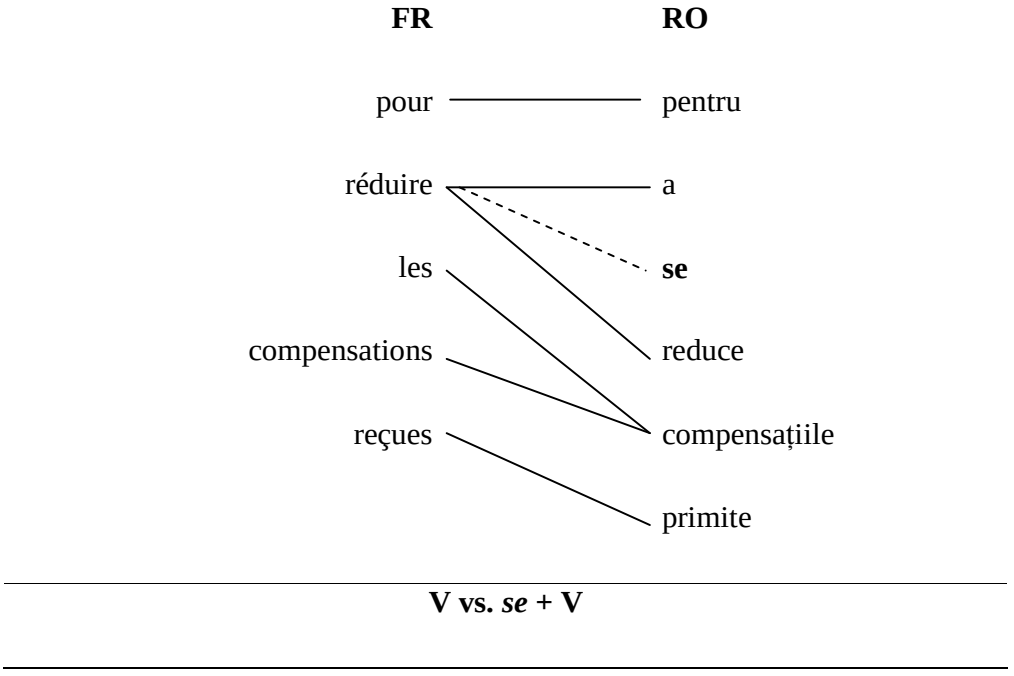


Figure 45. Alignement du verbe français avec le pronom réfléchi *se* du roumain

Au niveau de la correspondance des temps verbaux utilisés dans le corpus étudié, nous avons repéré des cas où le présent de l'indicatif français se traduit par un présent du conditionnel optatif en roumain. À cause de cette différence, le verbe auxiliaire *a avea* 'avoir' reste non-aligné. La Figure 46 suivante en donne un exemple et sa règle morphosyntaxique de correction.

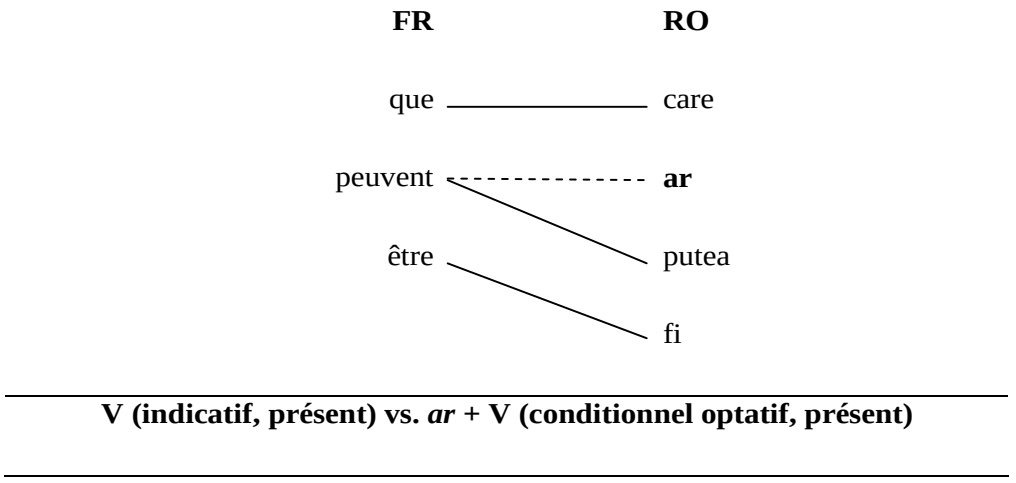


Figure 46. Alignement du verbe français à l'indicatif présent avec le verbe auxiliaire *ar* du présent du conditionnel optatif roumain

De plus, le futur de l'indicatif (comme *il pourra*) et le présent du conditionnel optatif (comme *il pourrait*) du français s'expriment à l'aide des terminaisons spécifiques à ces modes et temps tandis qu'en roumain le verbe auxiliaire *a avea* 'avoir' est nécessaire pour exprimer ces modes et temps (*el **va** putea* 'il pourra' ; *el **ar** putea* 'il pourrait'). Par conséquent, pour ces cas, le verbe auxiliaire roumain reste non aligné. Ce verbe est toujours à la troisième personne du singulier ou du pluriel et présente les formes *ar* (conditionnel optatif, présent), *va* et *vor* (indicatif, futur). Ainsi, en prenant en compte ces caractéristiques morphosyntaxiques, nous avons également défini des règles heuristiques pour aligner ces structures.

Les classes d'erreurs d'alignement lexical suivantes (la négation (1,57%), les compléments d'objet indirect / objet second introduits par la préposition *à* du français ayant comme équivalent un nom au datif en roumain (1,17%) et les numéraux ordinaux ou les déterminants numéraux (0,34%)) sont moins fréquentes dans le corpus étudié, mais leur nombre peut augmenter significativement si le volume du corpus de travail devient important, car elles représentent des erreurs systématiques. De ce fait, pour ces classes d'erreurs nous avons aussi défini des heuristiques morphosyntaxiques.

4.1.3.6. Négation

En ce qui concerne la négation, celle-ci peut être totale ou partielle. La négation totale porte sur la proposition entière et s'exprime, en français, par le biais de *pas* ou *point*, associés à *ne* (*GMF*, 2009 : 698) (p. ex. … *qu'elle* ***ne*** *se considère* ***pas*** *liée par*…) alors qu'en roumain elle s'exprime par l'adverbe de négation *nu* (… *că* ***nu*** *se consideră obligată prin*…). Cette différence fait que les deux particules de négation du français ne sont pas alignées avec l'équivalent du roumain dans notre corpus de travail. Quant à la négation totale, nous avons identifié seulement la négation réalisée par *ne* … *pas* (1,57%), La Figure 47 ci-dessous illustre un exemple d'alignement concernant la négation totale en français et en roumain. En utilisant les formes de ces particules de la négation, une règle de correction a été proposée.

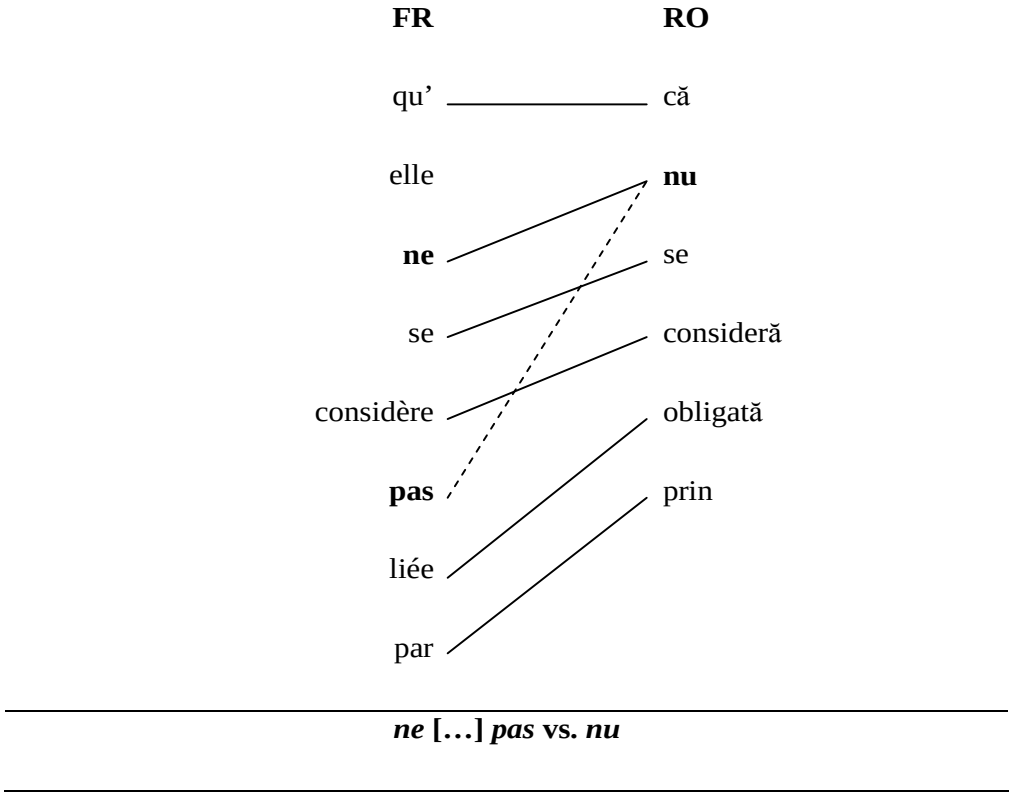


Figure 47. Alignement des particules de la négation totale *ne* … *pas* du français avec la particule *nu* du roumain

La négation partielle qui « porte sur une partie seulement de la proposition » et « s'exprime au moyen de mots négatifs associés à *ne* » (*GMF*, 2009 : 698) comme les pronoms négatifs (*Si* ***aucune*** *des parties contractantes* ***ne*** *le dénonce*… vs. *în cazul în care* ***niciuna*** *din părțile contractante* ***nu*** *îl denunță*…) ou les déterminants négatifs (… ***aucun*** *risque* ***n****'a été établi*… vs. … ***nu*** *a stabilit* ***niciun*** *risc*…), ne pose pas de problèmes particuliers à l'alignement lexical

car les structures à négation partielle sont similaires dans les deux langues étudiées. De ce fait, les pronoms ou les déterminants négatifs sont généralement alignés.

4.1.3.7. Compléments d'objet indirect / objet second introduits par *à* (FR) *vs.* Noms au datif (RO)

Des erreurs d'alignement lexical (1,17%) sont rencontrées aussi au niveau de l'objet second introduit par la préposition *à* du français (*Les États membres communiquent* ***à la Commission*** *les textes*…) ou du complément d'objet indirect introduit par *à* (… pour *se conformer* ***à la*** *présente* ***directive***…), ayant comme équivalent, en roumain, un nom au datif (*Statele membre comunică* ***Comisiei*** *textele*… ; … *pentru a se conforma prezent****ei directive***…). Précisons que la préposition *à* du français peut fusionner avec le déterminant défini *le* (singulier) ou *les* (pluriel) en donnant les formes *au* ou *aux.* En roumain, le nom au datif est marqué par la désinence spécifique du datif et fusionne avec le déterminant défini enclitique, forme de datif, (p. ex. *Comisi****ei*** 'à la Commission', *Comisi****ilor*** 'aux Commissions'). Les erreurs d'alignement repérées concernent la préposition *à* du français qui n'est pas alignée avec le nom du roumain. De même, quand le nom du roumain est précédé par un adjectif (au datif) (*prezent****ei*** *directiv****e*** 'à la présente directive'), un déterminant démonstratif (au datif) (*acest****ei*** *directiv****e*** **'**à cette directive'), indéfini (au datif) (*fiecăr****ei*** *directiv****e*** 'à chaque directive') ou indéfini de type article (forme de datif) (*unei directive* 'à une directive'), la préposition *à* du français reste non alignée avec ceux-ci. Un exemple d'erreur d'alignement lexical (et sa règle morphosyntaxique de correction) survenue au niveau de l'objet second introduit par *à* en français, qui a comme correspondant en roumain un nom au datif, est donné dans la Figure 48 suivante.

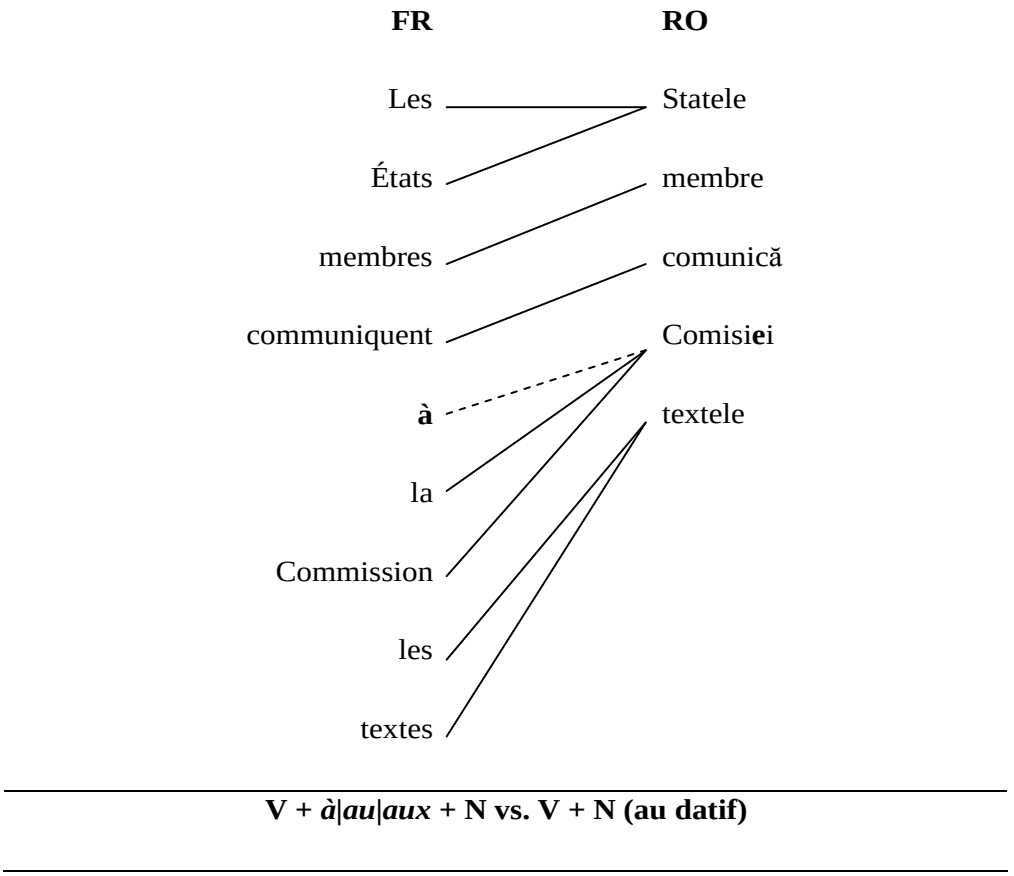


Figure 48. Alignement de la préposition *à* introduisant un complément d'objet second en français avec le nom du roumain au datif

La Figure 49 suivante montre un exemple d'erreur où la préposition *à* introduisant un complément d'objet indirect en français n'est pas alignée avec l'adjectif antéposé *prezentei* 'présente' du roumain, au datif. La règle morphosyntaxique de correction définie y figure aussi.

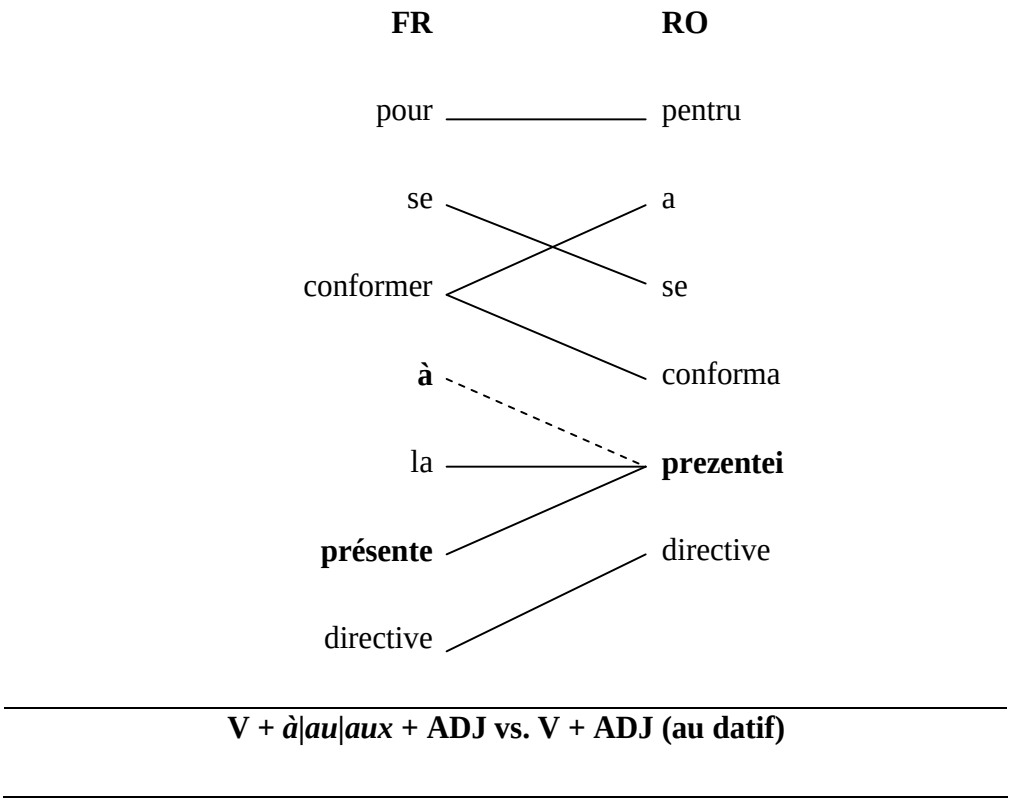


Figure 49. Alignement de la préposition *à* introduisant un complément d'objet indirect du français avec l'adjectif antéposé du roumain au datif

4.1.3.8. Pronoms adverbiaux *en* et *y*

Les pronoms adverbiaux *en* et *y* (anaphoriques ou cataphoriques) du français n'ont pas d'équivalents directs en roumain. De ce fait, ils restent non alignés par le système d'alignement lexical de base. Comme dans la phrase française ils remplacent un groupe, il faut les mettre en correspondance avec l'équivalent de ce groupe dans la phrase roumaine. Dans les exemples 1 et 2 ci-dessous, le pronom *en* a comme équivalent, en roumain, le groupe prépositionnel « despre aceasta » 'sur cela' pendant que dans les exemples 3 et 4, le pronom *y* est traduit par le groupe nominal « la procedură » '(participent) *à la procédure'*.

1) *Les États membres qui font usage de la faculté visée au présent paragraphe **en** informent la Commission et…* (FR)

2) *Statele membre care uzează de posibilitatea menționată în prezentul alineat informează Comisia **despre aceasta** și…* (RO)

3) *Les personnes effectuant des procédures ou **y** prenant part, ainsi que…* (FR)

4) *Persoanele care efectuează proceduri sau care participă **la procedură**, precum și…* (RO)

Le taux de ces erreurs est très faible dans le corpus étudié (0,49%) mais il peut croître en fonction du domaine et du volume des corpus de travail. Pour ce type d'erreurs, des études sur des corpus juridiques et administratifs plus volumineux restent encore nécessaires, afin de rendre compte de leur impact sur l'alignement lexical français - roumain. De plus, à partir des seules informations linguistiques présentes dans notre corpus de travail, nous ne pouvons pas définir d'heuristiques morphosyntaxiques afin de résoudre les erreurs d'alignement au niveau de ces pronoms. En effet, des techniques de modélisation du phénomène anaphorique, à l'intérieur des phrases parallèles, deviennent nécessaires. Même si nous ne nous sommes pas proposé de s'en occuper dans le cadre de cette étude, l'utilisation de telles techniques représente l'une des perspectives de ce travail.

4.1.3.9. Déterminants numéraux ordinaux

La dernière classe d'erreurs (0,34%) repérée est représentée par les erreurs apparues au niveau de l'emploi des déterminants numéraux ordinaux (*le cinquième alinéa vs. al cincilea alineat*). En français, ce déterminant se construit à partir du numéral cardinal auquel s'ajoute la terminaison spécifique *-ième* (*cinquième*) étant précédé du déterminant défini *le* ou *la*. À la différence du français, le déterminant numéral ordinal du roumain se forme à partir du numéral cardinal mais en ajoutant les éléments homonymes avec l'article défini enclitique qui varie en genre ***-le*** (masculin - neutre : *patru+**le**+a* 'quatrième') ou ***-a*** (féminin : *patru+**a*** > *patr**a*** 'quatrième') et la particule déictique ***a*** (pour masculin-neutre : *patru+le+**a***). En plus, ce déterminant est précédé par les morphèmes *al / a* qui varient en genre (***a**l patrulea* 'le quatrième', ***a** patra* 'la quatrième') (*GLR*, 2005 : 303). À cause de ces différences morphosyntaxiques, le déterminant défini du français *le* ou *la* et le morphème *al* ou *a* du roumain ne sont pas toujours alignés, comme dans l'exemple de la Figure 50 suivante, où la règle de correction définie est donnée aussi.

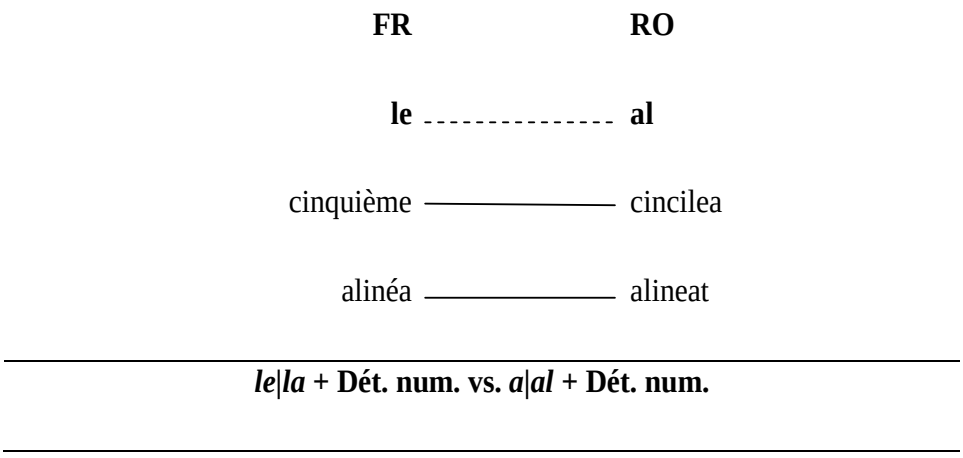


Figure 50. Alignement du déterminant défini *le/la* du français avec le morphème *a*/*al* du roumain dans le cas des déterminants numéraux ordinaux

Dans d'autres cas, le déterminant numéral ordinal du roumain est précédé par un morphème supplémentaire *cel / cea* (un pronom semi-indépendant - *GLR*, 2005 : 306) et la préposition *de (cel de-al cincilea ; cea de-a cincea)* tandis que la construction française reste la même (*le cinquième ; la cinquième*). De ce fait, le déterminant défini *le* du français n'est pas aligné avec les morphèmes supplémentaires et la préposition *de* du roumain, comme dans l'exemple illustré dans le Figure 51 ci-dessous, accompagné par sa règle de correction proposée.

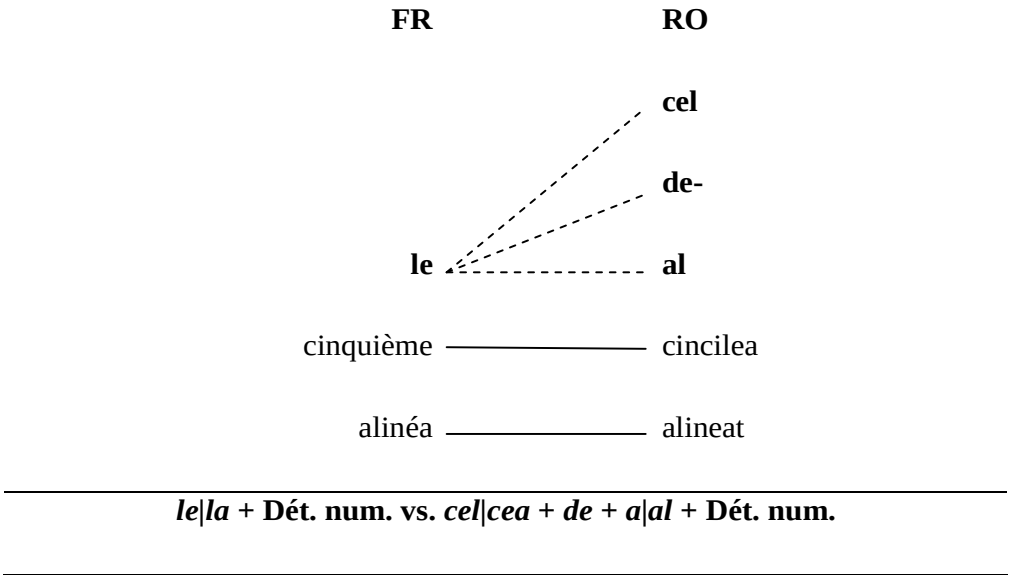
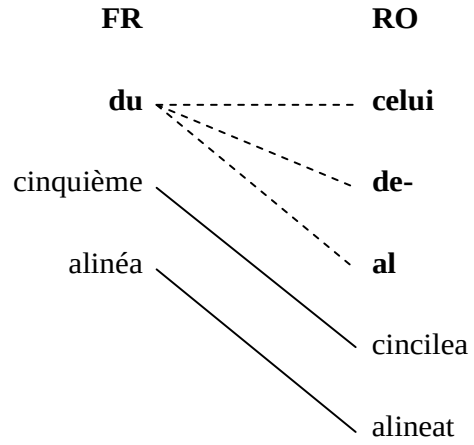


Figure 51. Alignement du déterminant défini le/la du français avec les morphèmes supplémentaires et la préposition *de* du déterminant numéral ordinal du roumain

Quand le déterminant numéral ordinal du français est précédé par l'une des prépositions *de* ou *à* ou de leurs formes fusionnées avec le déterminant défini *du*, respectivement *au*, en roumain,

le nom précédé par le déterminant numéral ordinal est au génitif ou au datif et, par conséquent, le pronom semi-indépendant *cel / cea* est au génitif-datif *celui / celei*. Dans ces cas, les prépositions françaises *de* ou *à* (ou leurs formes contractées) ne sont pas alignées avec les morphèmes supplémentaires et la préposition *de* du roumain. La Figure 52 ci-dessous donne un exemple de ce type d'erreur d'alignement lexical ainsi que sa règle morphosyntaxique de correction.



***du*|*de la* + Dét. num. vs. *celui*|*celei* + *de* + *a*|*al* + Dét. num.**

Figure 52. Alignement du déterminant défini *le* fusionné avec la préposition *de* du français avec les morphèmes supplémentaires et la préposition *de* du déterminant numéral ordinal du roumain

De plus, lorsqu'un nom français précédé par un déterminant numéral ordinal remplit la fonction syntaxique de complément circonstanciel de temps (*entre en vigueur le troisième jour*), en roumain, ce complément est toujours introduit par la préposition *în* 'dans' (*intră în vigoare* ***în*** *a treia zi*). À cause de cette différence, la préposition *în* du roumain n'est pas alignée. Une règle morphosyntaxique de correction a été également définie pour résoudre ce type d'erreur (voir Figure suivante).

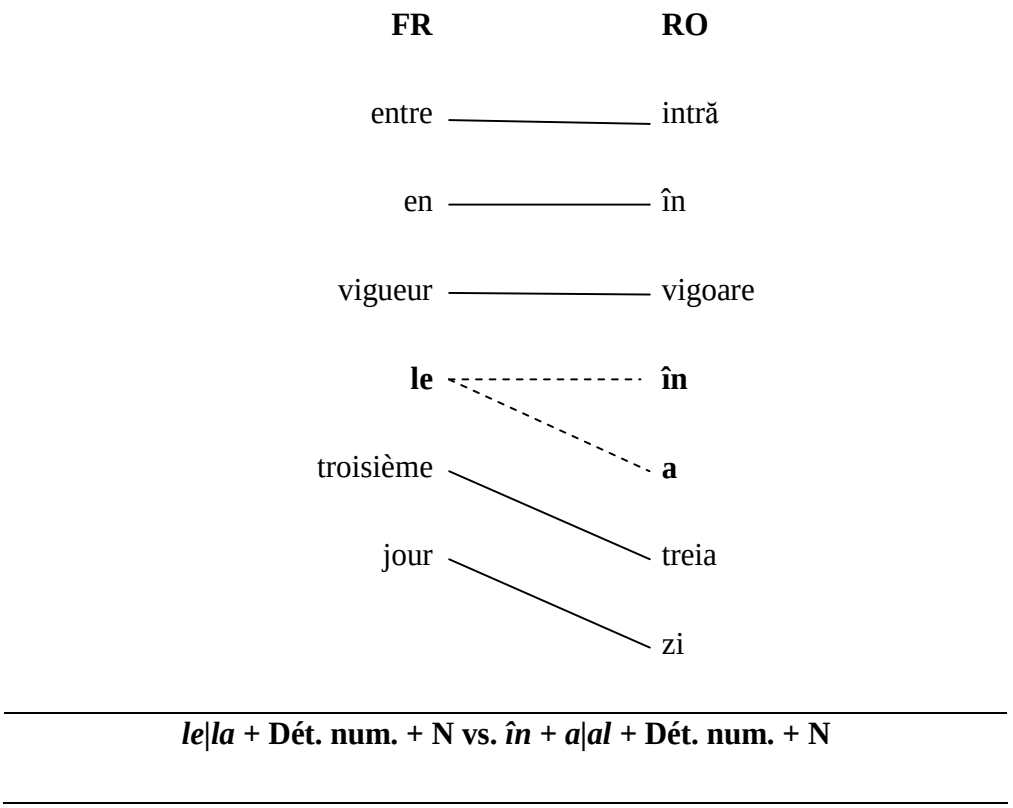


Figure 53. Alignement du déterminant défini *le*|*la* du français avec la préposition *în* et le morphème *a*|*al* du roumain

De même, dans l'expression de la date (*entre en vigueur* ***le 1er avril 1974*** vs. *intră în vigoare* ***la 1 aprilie 1974***), le français utilise le numéral ordinal alors qu'en roumain le complément circonstanciel de temps est introduit par l'une des prépositions *la* 'à' / *în* 'dans'/ *din* 'de'/ *până la* 'jusque là'. De ce fait, les prépositions du roumain ne sont pas alignées. La Figure 54 suivante illustre un exemple de ce type d'erreur ainsi qua sa règle morphosyntaxique de correction.

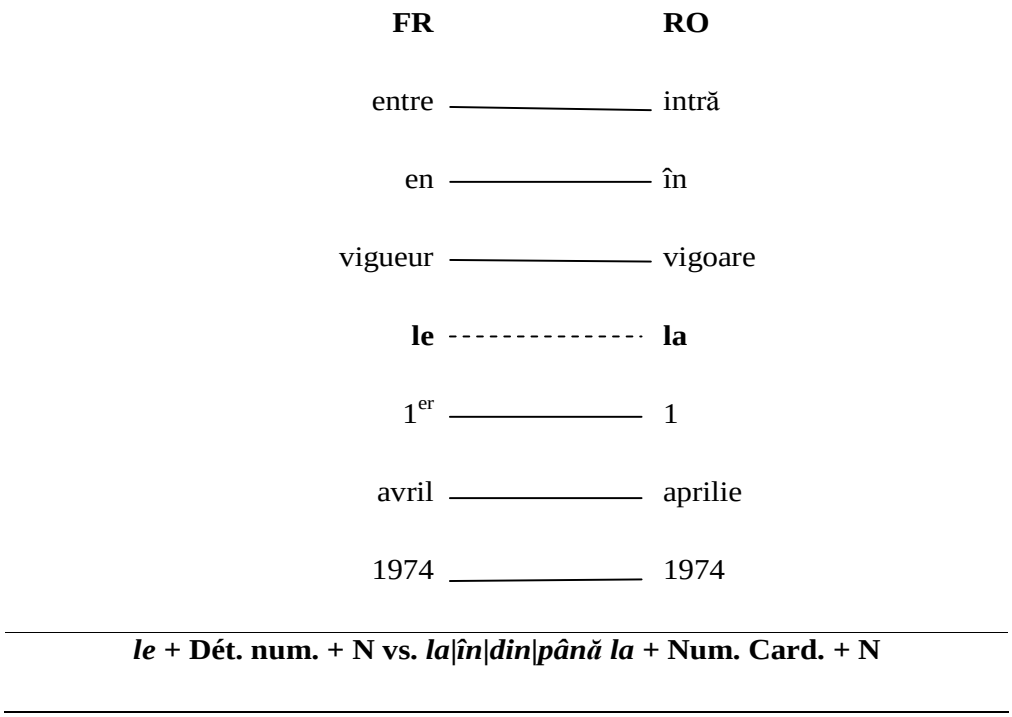


Figure 54. Alignement du déterminant défini *le* du français avec la préposition introduisant un complément circonstanciel de temps du roumain

4.1.3.10. Discussion

Les erreurs d'alignement lexical relevées et leur distribution par classes dépendent des langues étudiées, du domaine du corpus utilisé et aussi du volume des données et ne peuvent pas être généralisées. Mais elles restent une indication importante pour effectuer des alignements lexicaux des corpus bilingues parallèles juridiques et administratifs français - roumain (l'impact des règles heuristiques définies pour résoudre une partie de ces erreurs d'alignement lexical est montré dans la sous-section 4.3.2., où le module implémentant ces règles est évalué).

Toutefois, des études similaires restent encore nécessaires sur d'autres corpus de test du même domaine ou des domaines différents afin d'enrichir la base de règles morphosyntaxiques définies, car nous sommes conscients que les classes d'erreurs repérées et discutées ici sont limitées au corpus de test utilisé.

Nous avons inclus dans la catégorie *Autres* (5%) les erreurs pour lesquelles nous n'avons pas pu définir des règles heuristiques morphosyntaxiques, comme les erreurs apparues au niveau

des pronoms anaphoriques, des contraintes stylistiques en traduction humaine ou des procédés de traduction utilisés dans le domaine juridique et administratif de notre corpus.

Ainsi, nous avons effectué également une étude du corpus bilingue parallèle juridique et administratif au niveau de la traduction humaine, pour pouvoir expliquer les erreurs dues aux contraintes stylistiques ou aux procédés de traduction utilisés dans le corpus. Nous présenterons cette étude dans la sous-section suivante.

4.1.4. L'étude du corpus parallèle juridique au niveau de la traduction humaine

Il convient tout d'abord de signaler que l'étude du corpus au niveau de la traduction humaine a révélé une série d'erreurs d'alignement lexical dues aux contraintes stylistiques en traduction ou à d'autres procédés de la traduction oblique qui seront exemplifiés plus loin. Parmi ces erreurs, les seules que nous avons pu traiter par des règles de correction sont celles produites à cause des contraintes stylistiques en traduction. Pour corriger ces erreurs, nous avons défini un ensemble de règles exploitant les marques stylistiques spécifiques du corpus juridique étudié. Pour les autres catégories d'erreurs, nous n'avons pas disposé de ressources appropriées (listes de paraphrases, par exemple). Ainsi, les erreurs dues aux contraintes stylistiques en traduction surviennent principalement dans le cas des extensions monolexicales avec renforcement de sens à valeur stylistique n'apparaissant pas d'une langue à l'autre (déterminants indéfinis : … *peut procéder à des consultations* vs. … *poate efectua* ***orice*** *consultare* 'peut effectuer **toute** consultation' ; certains adjectifs qualificatifs : … *la notification* vs. … ***respectiva*** *notificare* 'la **respective** notification', etc.). D'autres erreurs de ce type concernent l'expression de la voix d'une langue à l'autre (par exemple, la préférence de traduction de la voix active ayant pour sujet le pronom personnel *on* du français par la voix pronominale en roumain : … ***on*** *ajuste le prélèvement* vs. … *taxa* ***se*** *corectează* 'le prélèvement s'ajuste') ou les structures impersonnelles (la préférence de traduction des structures impersonnelles du français par des structures pronominales en roumain : … ***il est*** *effectué* vs. … ***se*** *efectuează* 's'effectue').

Toutefois, les catégories d'erreurs repérées dépendent des langues étudiées, du domaine et du volume du corpus juridique et administratif utilisé et peuvent varier en fonction de ces paramètres.

Le discours juridique est caractérisé par l'utilisation des constructions impersonnelles, de la voix passive, l'emploi de la troisième personne du singulier, etc. (Cornu, 1995). Au niveau de la cohésion, le discours juridique présente comme marques stylistiques spécifiques des lexèmes, tels que :

- pronoms et déterminants démonstratifs (*ce* - *acest*) ;
- pronoms et déterminants indéfinis (*tout* - *orice*) ;
- pronoms relatifs (*qui* - *care*) ;
- pronoms personnels (*il* - *el*) et réfléchis (*se* - *îşi*), exclusivement à la troisième personne ;
- adverbes (*ainsi* - *astfel*) ;
- adjectifs (*tel* - *asemenea*) exprimant la similarité (Stoichițoiu-Ichim, 2000).

Au vu de ces considérations, nous ne pouvons pas généraliser les catégories d'erreurs relevées. Pour des corpus relatifs à des domaines différents, des études similaires restent encore nécessaires afin d'identifier les classes d'erreurs spécifiques. Cependant, ces résultats permettent d'améliorer l'alignement lexical des corpus parallèles juridiques français - roumain.

Dans l'Union Européenne, parmi les 23 langues officielles, il existe trois langues dites procédurales, l'anglais, le français et l'allemand, dans lesquelles sont rédigés les textes sources concernant la législation européenne. Si, statistiquement, l'anglais domine largement (plus de 70% des textes sources), il n'est jamais absolument sûr de pouvoir déterminer avec certitude la langue source parmi les trois langues procédurales ayant servi de fondement à la traduction vers le roumain. Une autre remarque s'impose aussi au niveau européen, c'est que chaque version linguistique du même texte fait foi dans sa langue, autrement dit chaque version linguistique constitue en soi un texte source. Nous arrivons pour les textes européens à une sorte de paradoxe où dans un corpus parallèle, il n'est plus possible de définir le sens de la traduction, bien que nous comprenions par corpus bilingue parallèle un ensemble de textes écrits dans une langue source et traduits dans une langue cible.

Nous avons étudié le parallélisme de notre corpus juridique et administratif français - roumain par le biais de la dichotomie *perte* et *gain* dans le transfert traductionnel (Hervey et Higgins, 2002), afin de rendre compte des problèmes rencontrés par notre système d'alignement lexical au niveau de la traduction juridique humaine. Même si nous ne connaissons pas avec exactitude quelle est la langue source de notre corpus, nous sommes partis de l'hypothèse que le français est la langue source et le roumain est la langue cible. Cette hypothèse est basée sur le fait que nous savons que le roumain est l'une des langues cibles et le français l'une des langues sources en traduction, dans le cadre des institutions européennes.

L'étude du parallélisme du corpus est limitée au niveau phrastique puisque les unités maximales de correspondance dans le corpus sont les phrases. Ainsi, nous avons effectué une évaluation traductionnelle au niveau de la phrase du français vers le roumain.

Nous avons étudié la dichotomie de *perte* et *gain* dans la traduction juridique français - roumain par l'intermédiaire des procédés de traduction (Vinay et Darbelnet, 1977 ; Delisle *et al.*, 1999 ; Molina et Albir, 2002 ; Cristea, 2007) rencontrés dans le corpus étudié. Ainsi, ont été identifiés des procédés de la traduction directe comme :

- le calque (*décision-cadre* vs. *decizie-cadru*) ;
- l'emprunt (… *au développement de* ***l'acquis*** *de Schengen* vs. … *la dezvoltarea* ***acquis-ului*** *Schengen*) ;
- la traduction littérale (***l'État membre veille à communiquer à la Commission*** vs. ***Statul membru trebuie să comunice Comisiei***).

Les procédés ci-dessus ne posent pas de problèmes significatifs à l'alignement lexical automatique du fait qu'ils fournissent des équivalents directs dans la langue cible. En revanche, d'autres procédés rencontrés au niveau de la traduction oblique posent problème à l'aligneur automatique à cause des phénomènes linguistiques complexes qu'ils traitent, comme le manque de correspondants directs pour certaines unités de traduction dans la langue cible, les contraintes stylistiques en traduction, etc. Ces procédés seront énumérés et illustrés par des exemples ci-dessous.

Certaines contraintes stylistiques au niveau de la traduction oblique sont résolues par l'étoffement (Delisle *et al.*, 1999). En effet, dans le corpus analysé, l'étoffement prend la

forme des extensions monolexicales avec renforcement de sens à valeur stylistique (certains adjectifs qualificatifs, déterminants indéfinis, etc.). Nous entendons par extension monolexicale un type spécifique d'étoffement qui suppose la supplémentation d'un noyau d'extension (nom ou verbe) par un lexème à valeur stylistique, afin de respecter les spécificités stylistiques du discours juridique en renforçant le sens mais sans en ajouter (Molina et Albir, 2002). Ainsi, ces extensions monolexicales peuvent être rencontrées dans la langue cible tandis qu'elles n'apparaissent pas dans la langue source (p. ex. … *la directive* vs. … ***prezenta** directivă* 'la **présente** directive'). Ces marques peuvent aussi manquer de la langue cible, alors qu'elles sont présentes dans la langue source, phénomène appelé dépouillement (Vinay et Darbelnet, 1977) (p. ex. ***Chaque** État membre…* vs. *Statele membre*…).

D'autres procédés de la traduction oblique souvent repérés dans ce corpus parallèle juridique sont les suivants :

- la transposition (le changement de catégorie grammaticale) (p. ex. ***la visualisation** du micronoyau est facilitée* vs. *micronucleul **se vizualizează** mai ușor* 'le micronoyau **est visualisé** plus facilement') ;

- l'équivalence (la traduction des structures figées sources par des structures figées cibles) (p. ex. ***au sens du** paragraphe 1* vs. ***în înțelesul** alineatului (1)*).

Enfin, les procédés de la traduction oblique également identifiés dans le corpus étudié sont l'explicitation et la paraphrase interlinguale.

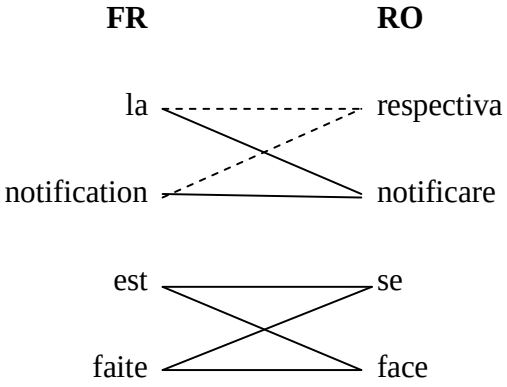
L'explicitation suppose la transformation de l'implicite de la langue source dans des précisions dans la langue cible (Molina et Albir, 2002) (p. ex. *la charge est déterminée* vs. *sarcina **financiară pe care aceasta o antrenează** se stabilește* 'la charge **financière que celle-ci** entraîne est déterminée').

La paraphrase interlinguale et son correspondant sont ancrés dans une relation d'équivalence qui repose sur une invariabilité du sens global des deux et qui est issue d'une variabilité lexico-syntaxique. Celle-ci relève du système de la langue et non pas du discours (Cristea, 2007). Un exemple de paraphrase interlinguale extrait de notre corpus est le suivant : *sur justification de* vs. *la prezentarea unei dovezi privitoare la…* 'à la présentation d'une preuve liée à…'.

À partir de cette étude du corpus bilingue parallèle au niveau de la traduction humaine, nous avons repéré ainsi des erreurs d'alignement lexical au niveau stylistique, dans le cas des extensions monolexicales avec renforcement de sens à valeur stylistique, des voix et des structures impersonnelles. Ces erreurs seront exemplifiées dans les sous-sections suivantes.

4.1.4.1. Extensions monolexicales

Nous avons compté dans la catégorie des extensions monolexicales avec renforcement de sens à valeur stylistique certains adjectifs qualificatifs, tels que : *respectiv* 'respectif', *prezent* 'présent', *asemenea* 'tel', *anumit* 'certain', etc. Ces adjectifs n'apparaissent pas alignés quand ils ne sont pas présents d'une langue à l'autre au niveau de la traduction. Dans l'exemple de la Figure 55 suivante, l'adjectif qualificatif *respectiv* 'respectif' du roumain n'est pas aligné en bloc avec le nom *notification* et le déterminant défini *la* du français.



Dét. déf. + N vs. *respectiv|prezent|asemenea|anumit* … + N

Figure 55. Alignement des extensions monolexicales avec renforcement de sens à valeur stylistique

Nous avons collecté tous ces adjectifs rencontrés dans le corpus étudié et nous avons défini des règles stylistiques exploitant leurs lemmes et les étiquettes morphosyntaxiques des noms qu'ils déterminent, afin de solutionner les erreurs liées à leur alignement (cf. Figure 55 ci-dessus).

Les autres lexèmes avec renforcement de sens à valeur stylistique repérés sont les suivants : déterminants indéfinis (*chaque - fiecare ; tout - oricare / orice*), déterminants démonstratifs (*ce - acest*) et verbes (*devoir - trebui*). Dans le cas du dépouillement en roumain, nous avons relevé l'adverbe *notamment* du français.

4.1.4.2. Voix

Concernant les voix, nous avons remarqué des erreurs d'alignement au niveau de la voix passive du français qui se traduit souvent par la voix pronominale en roumain (… *le texte* ***est remplacé*** *par*… vs. … *textul* ***se înlocuiește*** *cu* … 'le texte se remplace par'). Dans ce cas, les marques linguistiques de chaque structure (la présence du verbe auxiliaire *être* suivi du participe passé du verbe (FR) vs. le pronom réfléchi *se* 'se' (ayant toujours la forme de la troisième personne du singulier ou du pluriel) suivi du verbe (RO)) ne sont pas alignées en bloc. Pour ce type d'erreurs, une règle stylistique de correction à été définie (cf. Figure 56 ci-dessous).

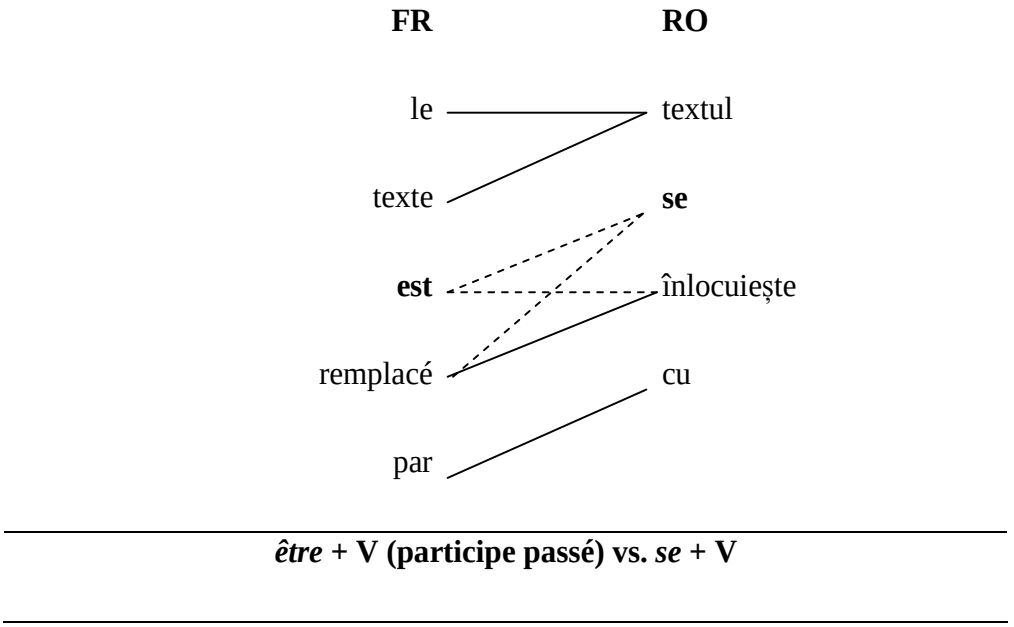


Figure 56. Alignement de la voix passive du français avec la voix pronominale du roumain

De même, la voix active ayant pour sujet le pronom personnel *on* du français se traduit par la voix pronominale en roumain (***On entend*** *par*… vs. ***Se înțelege*** *prin*… 'S'entend par'). La Figure 57 suivante donne un exemple d'erreur où le pronom personnel *on* du français et le verbe qui le suit ne sont pas alignés en bloc avec le pronom réfléchi *se* suivi du verbe en roumain. La règle stylistique proposée y figure aussi.

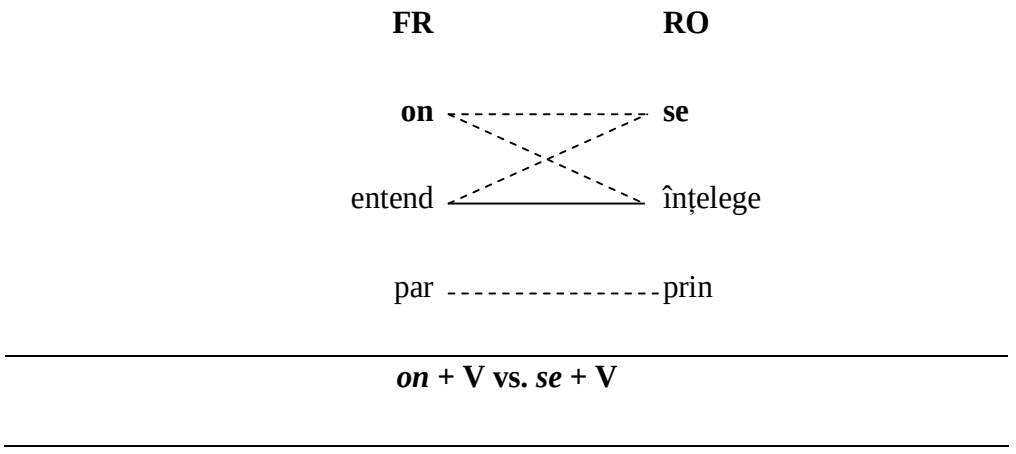


Figure 57. Alignement de la voix active du français ayant pour sujet le pronom personnel *on* avec la voix pronominale du roumain

4.1.4.3. Constructions impersonnelles

Des erreurs d'alignement lexical ont été relevées aussi au niveau des constructions impersonnelles qui s'expriment différemment d'une langue à l'autre. Les erreurs repérées dans ce sens concernent des équivalents français - roumain, tels que : *il est nécessaire* (FR) vs. *este / e necesar* 'est nécessaire' (RO) ; *il est effectué* (FR) vs. *se efectuează* 's'effectue' (RO) ; *il existe* (FR) vs *există* 'existe' (RO).

Ces constructions impersonnelles du français (langue non pro-drop) sont caractérisées par la présence du pronom impersonnel *il* « référentiellement vide » (*GMF*, 2009 : 750), « occupant la place canonique du sujet non pourvue » (*GMF*, 2009 : 750), alors qu'en roumain (langue pro-drop) le pronom s'absente. À cause de cette différence, le pronom *il* n'est pas aligné au niveau du groupe verbal. Pour ce type d'erreurs, des règles d'alignement lexical de correction ont été également proposées. Gojun (2010) définit aussi des règles afin d'aligner le pronom sujet avec les groupes verbaux pour une autre langue pro-drop comme l'italien en combinaison avec l'anglais qui est une langue non pro-drop.

Dans les couples comme *il est nécessaire* (FR) vs. *este / e necesar* 'est nécessaire' (RO), le pronom impersonnel *il* du français est suivi du verbe *être* et, plus précisément, de la forme *est*, toujours à la troisième personne du singulier, et d'un adjectif (*nécessaire*), pendant qu'en roumain la construction impersonnelle est composée seulement du verbe *a fi* 'être' (la forme *este* ou *e* à la troisième personne du singulier) et d'un adjectif (*necesar*). La Figure 58

suivante présente un exemple d'erreur ou les éléments composants de ces constructions ne sont pas alignés en bloc. Une règle stylistique de correction a aussi été définie (cf. Figure 58).

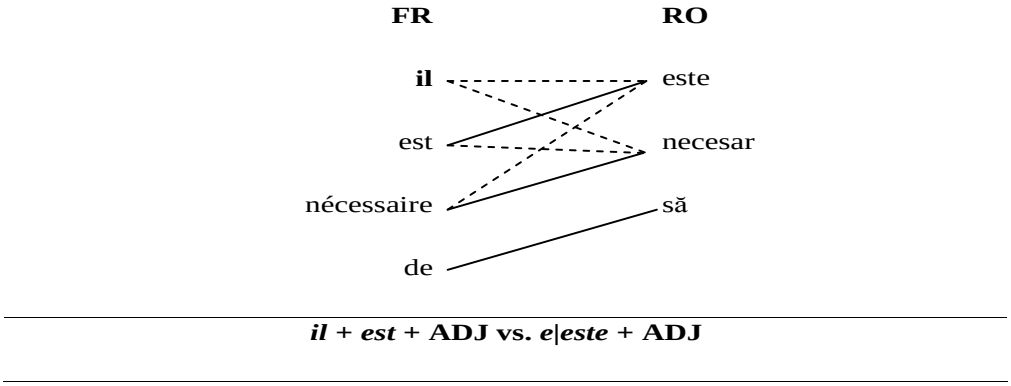


Figure 58. Alignement des constructions impersonnelles en français et en roumain

D'autres erreurs d'alignement lexical existent aussi au niveau des formes impersonnelles du passif en français (*il est effectué*) vs. les formes pronominales en roumain (*se efectuează* 's'effectue'). Dans ce cas, le pronom impersonnel *il* du français est suivi d'une forme passive (*être* et un verbe au participe passé), alors que le roumain préfère la construction pronominale composée du pronom réfléchi *se* 'se' et d'un verbe (toujours a la troisième personne). À cause de cette différence, les éléments des deux constructions ne sont pas alignés en bloc. Pour ce type d'erreurs une règle stylistique a été aussi proposée (cf. Figure 59 ci-dessous).

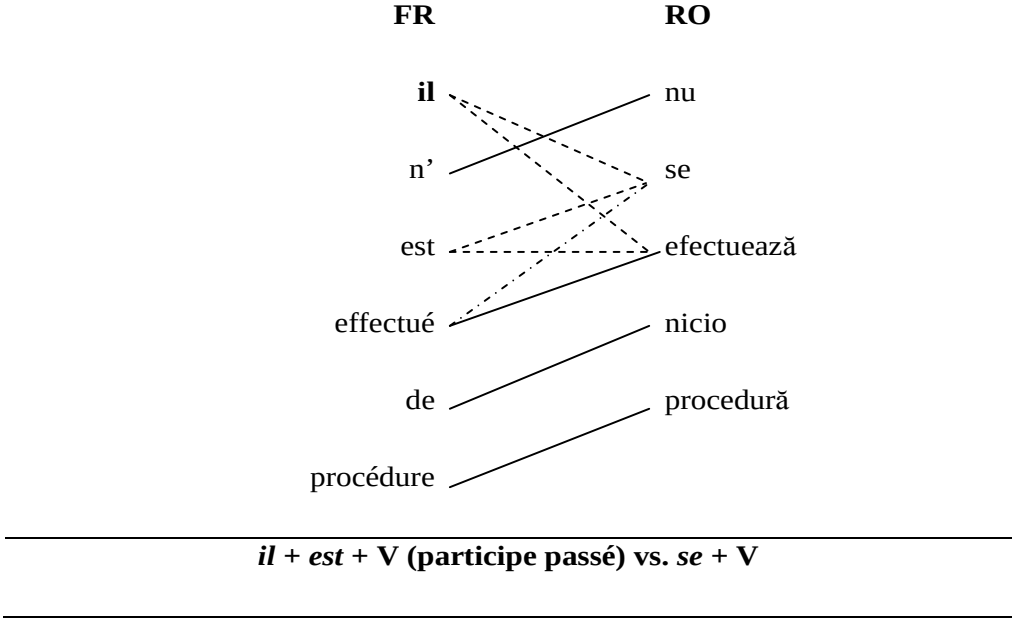


Figure 59. Alignement des formes impersonnelles du passif en français avec les formes pronominales en roumain

Des problèmes d'alignement concernent aussi les verbes impersonnels du français, car ils sont toujours précédés par le pronom impersonnel *il*, tandis qu'en roumain le verbe apparaît sans pronom (*il existe* vs. *există* ; *il résulte* vs. *rezultă* ; *il convient* vs. *trebuie* ; etc.). Dans ce cas, le pronom *il* reste toujours non aligné (cf. Figure 60 ci-dessous).

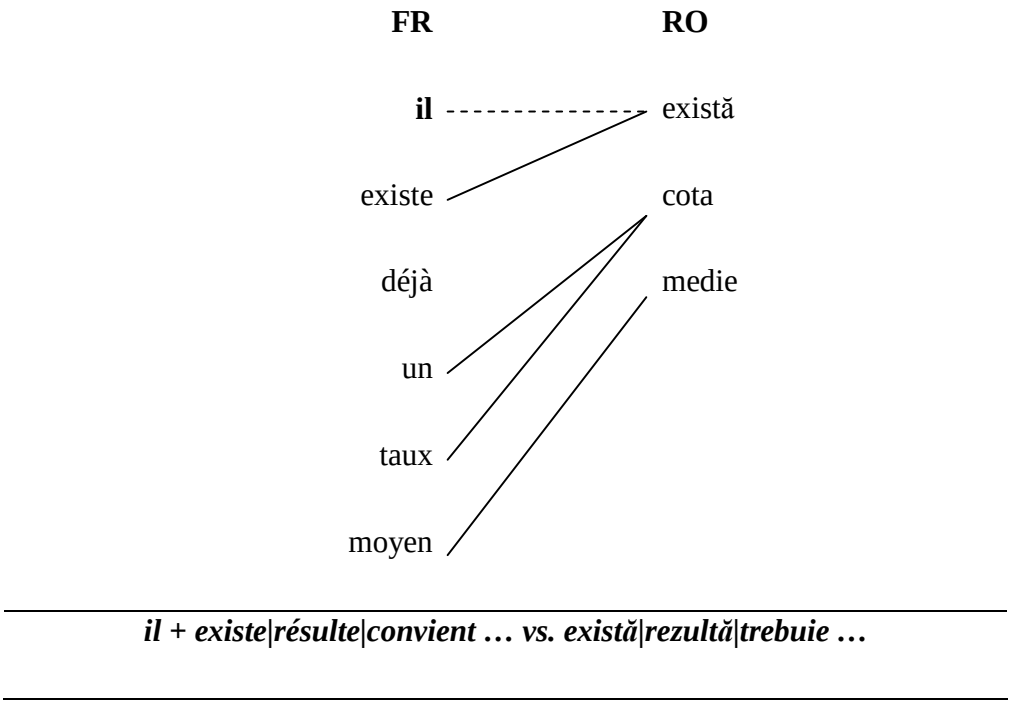


Figure 60. Alignement des verbes impersonnels en français et en roumain

Nous avons collecté tous ces verbes impersonnels à partir du corpus parallèle de test et leurs équivalents de traduction, en définissant des règles stylistiques de correction basées sur leurs formes (cf. Figure 60 ci-dessus). Cette liste n'est pas exhaustive car elle dépend, évidemment, du corpus de test utilisé. Par conséquent, des études stylistiques sur d'autres corpus de test restent encore nécessaires afin d'enrichir la liste des verbes impersonnels français - roumain collectée, qui sera utile pour des applications ultérieures sur d'autres corpus.

Pour les autres catégories d'erreurs repérées au niveau des paraphrases ou des équivalences, nous n'avons pas pu définir de règles en utilisant les informations linguistiques disponibles dans le corpus. Nous n'avons pas disposé non plus de ressources externes pour résoudre ces erreurs d'alignement (listes de paraphrases ou de locutions, par exemple). Dans le cas des transpositions, comme elles supposent un changement de catégorie grammaticale, nous avons exploité les cognats pour résoudre les erreurs d'alignement apparaissant à leur niveau (cf. section suivante). En ce qui concerne les explicitations, nous n'avons pas remarqué d'erreurs

d'alignement lexical à leur niveau, car ces structures représentent des précisions dans la langue cible qui n'apparaissent pas dans la langue source.

4.1.5. Discussion

Les règles heuristiques morphosyntaxiques et stylistiques, définies à l'issue de l'analyse linguistique et de l'étude du corpus bilingue parallèle au niveau de la traduction humaine, sont automatisables dans un module supplémentaire d'alignement lexical dépendant de la paire de langues étudiées (cf. section 4.3.). Ces règles dépendent aussi du corpus de test utilisé (de son domaine et de sa taille) et la base de ces règles peut toujours être améliorée.

Le module d'application des règles heuristiques mis en place et son impact sur le système d'alignement lexical seront présentés, plus loin, dans la section 4.3. Néanmoins, nous sommes conscients que, pour rendre compte de la fiabilité des règles heuristiques définies, avec plus de précision, celles-ci doivent être appliquées et évaluées sur d'autres corpus de test. Mais ces expériences représentent l'une des perspectives de cette thèse.

Toutefois, comme nous avons effectué des expériences d'alignement lexical seulement sur un seul corpus de test, une possibilité de rendre compte de la fiabilité de notre base de règles heuristiques, avec plus de précision, a été d'étudier son impact dans le système de traduction automatique statistique factorisé lui-même. Nous évaluerons son influence sur ce système dans le chapitre 5.

À partir de l'analyse linguistique effectuée, la première étape suivie pour améliorer la performance de notre système d'alignement lexical consiste dans l'identification et l'alignement automatique des cognats. Le module développé ainsi que son évaluation seront présentés dans la section suivante.

4.2. L'identification et l'alignement automatique des cognats

Nous avons déjà défini comme cognats les équivalents de traduction ayant une forme identique d'une langue à l'autre ou présentant des similarités aux niveaux orthographique (*information* (FR) - *informare* (RO) et/ou phonétique (*phase (*FR) - *fază* (RO)). Les cognats représentent des indices lexicaux importants dans un système d'alignement lexical français - roumain, utilisant des corpus bilingues parallèles juridiques, pour trois raisons principales :

i. Le français et le roumain sont deux langues apparentées de la famille des langues romanes. Le roumain appartient, quant à lui, au groupe des langues romanes orientales. Celui-ci est parlé par environ 29 millions de locuteurs dans le monde entier (Trandabăț *et al.*, 2012). Mais le roumain est principalement parlé en Roumanie et en République de Moldavie. Une grande partie du vocabulaire roumain est issu directement du latin. Toutefois, il existe également un substrat thrace ainsi qu'un superstrat slave ainsi qu'un apport de mots allemands, français, grecs, hongrois et turcs auquel s'ajoute un apport récent de mots anglais.

ii. Du fait que nous nous appuyons sur un corpus à dominante juridique, *DGT-TM* (Steinberger *et al.*, 2012), il nous appartient de souligner que le roumain a repris la terminologie juridique française par des emprunts (*acquis* (FR) - *acquis* (RO)) et des calques (*décision-cadre* (FR) - *decizie-cadru* (RO)). Les cognats s'avèrent ainsi très utiles pour détecter également des termes bilingues poly-lexicaux à partir des corpus parallèles juridiques et administratifs (*coopération européenne* vs. *cooperare europeană*), tant qu'aucune ressource terminologique externe n'est utilisée. Rappelons que nous entendons par termes polylexicaux des « combinaisons lexicales spécialisées », c'est-à-dire « des groupes associés à un domaine de connaissances » (L'Homme et Meynard, 1998 : 201).

iii. Les cognats sont utilisés généralement pour repérer des équivalents de traduction sûrs. Lorsqu'ils sont intégrés dans un système d'alignement lexical, ils peuvent améliorer la qualité des modèles de traduction statistiques (Kondrak *et al.*, 2003). Comme notre objectif final est la mise en place d'un système factorisé de traduction automatique, nous nous sommes concentrés, tout d'abord, sur le développement d'une méthode d'identification automatique des cognats pour la paire de langues étudiées.

Dans notre approche, nous avons utilisé une combinaison originale de techniques statistiques, informations linguistiques et ajustements orthographiques. Dans un premier temps, nous avons défini et appliqué au corpus un ensemble d'ajustements orthographiques afin de diminuer les différences formelles entre les mots d'une paire bilingue. Ces ajustements sont basés sur les similarités orthographiques et/ou phonétiques entre les mots des paires bilingues. Ensuite, nous avons combiné des techniques statistiques (des méthodes n-grammes (Simard *et al.*, 1992), fréquences, extractions itératives, etc.) à des informations linguistiques (lemmes, parties du discours) pour détecter automatiquement les cognats à partir du corpus bilingue

parallèle utilisé. Nous avons évalué le module développé, en faisant aussi une comparaison des résultats obtenus avec ceux fournis par des méthodes statistiques pures afin d'étudier l'impact des informations linguistiques utilisées sur la qualité des résultats. Nous montrerons plus loin que l'utilisation des informations linguistiques a augmenté significativement la performance de la méthode.

Toutefois, le repérage automatique des cognats à partir des textes multilingues parallèles est une tâche difficile. Cela est dû aux similarités orthographiques et/ou phonétiques importantes entre des mots ayant un sens différent (les faux amis). Pour distinguer les cognats anglais - français des faux amis et des mots non-apparentés, Inkpen *et al.* (2005) développent des classifieurs basés sur plusieurs dictionnaires et des listes de cognats construites manuellement. Ainsi, Inkpen *et al.* (2005) classifient les paires bilingues de mots en plusieurs catégories :

1. cognats (*texte* (FR) - *text* (EN) ; *sens* (FR) - *sense* (EN)) ;
2. faux amis (*blesser* (FR) - *bless* (EN)) ;
3. cognats partiels (dépendant du contexte) (*facteur* (FR) - *factor* ou *mailman* (EN)) ;
4. cognats génétiques (ayant une origine commune) (*chef* (FR) - *head* (EN)) ;
5. paires de mots non-apparentées (*glace* (FR) - *ice* (EN) et *glace* (FR) - *chair* (EN)).

Par rapport à l'approche proposée par Inkpen *et al.*, (2005), notre méthode identifie des cognats sans toutefois faire la différence entre cognats, cognats partiels et cognats génétiques. Notre but était d'améliorer le système d'alignement lexical de base par l'incorporation d'informations linguistiques au processus d'alignement. Ainsi, nous avons cherché à obtenir une précision élevée de la méthode développée, en combinant des méthodes n-grammes (Simard *et al.*, 1992) à des informations linguistiques. De plus, notre méthode élimine une partie des faux amis et des paires de mots non-apparentées en utilisant la fréquence des candidats cognats dans le corpus utilisé. Cette approche ne demande pas d'exploiter des ressources extérieures (dictionnaires, listes de cognats) Ainsi, notre méthode peut être facilement adaptable à d'autres langues apparentées.

Plusieurs approches exploitent les similarités orthographiques entre les mots d'une paire bilingue afin d'identifier automatiquement les cognats à partir du corpus. Une méthode efficace est celle appelée 4-grammes (Simard *et al.*, 1992) : deux mots sont considérés comme

cognats s'ils possèdent au moins 4 caractères et si leurs 4 premiers caractères sont identiques (p. ex. ***auto****risation* (FR) vs. ***auto****rizare* (RO)).

Certaines méthodes exploitent le coefficient de *Dice* (Adamson et Boreham, 1974 ; Brew et McKelvie, 1996). Ce score d'association calcule le rapport entre le nombre de caractères des bigrammes communs aux deux mots considérés et le nombre total des bigrammes des deux mots. Le coefficient de *Dice* est calculé selon la formule suivante :

$$Dice(mot_1, mot_2) = \frac{2 \times nombre\_bigrammes\_communs}{nombre\_total\_bigrammes(mot_1) + nombre\_total\_bigrammes(mot_2)}$$

Cette méthode attribue un score à chaque paire bilingue de mots prise en compte à l'intérieur d'une paire de phrases parallèles. Ce score doit se situer au-dessus d'un seuil établi empiriquement pour pouvoir sélectionner les cognats considérés comme pertinents parmi les autres paires candidates. De ce fait, il peut avoir des valeurs différentes en fonction des paires de langues considérées mais aussi du corpus utilisé.

D'autres approches calculent le rapport entre le nombre de caractères (ordonnés et pas nécessairement contigus) de la sous-chaîne maximale (*SCM*) commune aux deux mots et la longueur du mot le plus long (Melamed, 1999 ; Kraif, 1999). La formule qui calcule la mesure *SCM* figure ci-dessous :

$$mesure_{SCM}(mot_1, mot_2) = \frac{longueur(sous\_chaîne\_commune(mot_1, mot_2))}{\max(longueur(mot_1), longueur(mot_2))}$$

Comme dans le cas du coefficient de *Dice*, la valeur de la mesure *SCM* doit aussi être supérieure (ou égale) à un seuil établi empiriquement afin de pouvoir garder les paires de mots considérés comme cognats. Ce seuil dépend aussi des langues traitées et du corpus utilisé.

Nous avons implémenté les trois méthodes mentionnées ci-dessus (4-grammes, le coefficient de *Dice* et la mesure *SCM*) afin de faire une comparaison des résultats fournis avec ceux donnés par notre propre méthode. Cette évaluation comparative sera présentée dans la sous-section 4.2.2.

De manière similaire, d'autres approches calculent, d'un côté, la distance entre deux mots, qui représente le nombre minimum de substitutions, insertions et suppressions utilisées pour

transformer un mot dans un autre (Wagner et Fischer, 1974 ; Tufiș *et al.*, 2006 ; Ceaușu, 2009). Cette mesure exploite également les ressemblances au niveau orthographique entre les mots d'une paire bilingue, étant appelée la distance de Levenshtein (Levenshtein, 1966). D'autre côté, certaines approches estiment la distance phonétique entre deux mots appartenant à une paire bilingue (Oakes, 2000). Les similarités entre ces mots sont évaluées dans ce cas au niveau des phonèmes. Les valeurs de ces mesures doivent aussi se situer au-dessus d'un seuil calculé empiriquement.

Une méthode qui intègre également le roumain mais en combinaison avec l'anglais calcule la similarité orthographique, au sein des paires bilingues de mots, comme une mesure Levenshtein (Tufiș *et al.*, 2006 ; Ceaușu, 2009). Le seuil de cette mesure, établi empiriquement, a la valeur de 0.42 pour la paire de langues anglais - roumain. De plus, cette mesure est calculée en fonction d'une étape de normalisation dépendante de la paire de langues traitées, qui élimine les diacritiques, les consonnes doubles, certains suffixes, etc. pour diminuer les différences orthographiques entre les mots d'une paire bilingue. Ces opérations ont été définies manuellement. Mais elles peuvent être définies de manière automatique si des listes de cognats roumain - anglais sont disponibles (Tufiș *et al.*, 2006). Nous avons aussi appliqué cette étape de normalisation et, plus précisément, l'élimination des diacritiques et des consonnes doubles, mais, à la différence de cette approche, nous avons défini des règles d'ajustements orthographiques dépendantes mais aussi non dépendantes du contexte phonétique. Ces règles seront décrites dans la sous-section suivante.

À notre connaissance, aucune méthode spécifique à la paire de langues étudiée n'est disponible, à l'exception du système multilingue développé par Bergsma et Kondrak (2007) pour les cinq langues romanes (français, italien, espagnol, portugais et roumain).

Cette approche exploite l'idée de la transitivité entre les équivalents de traduction à travers les langues pour tenter d'améliorer l'extraction automatique des cognats. En effet, si un mot *x* (d'une langue notée *X*) est l'équivalent d'un mot *y* (d'une langue notée *Y*) et le mot *y* est également l'équivalent d'un mot *z* (d'une langue notée *Z*), alors, par transitivité, les mots *x* et *z* représentent aussi des équivalents de traduction. Concernant les cognats, un exemple illustrant la transitivité est le suivant : si *cuore* (italien) est cognat avec *cœur* (français) et, à part, *cœur* (français) est cognat avec *corazon* (espagnol), alors *cuore* (italien) et *corazon* (espagnol) sont aussi des cognats (Bergsma et Kondrak, 2007).

La méthode proposée par Bergsma et Kondrak (2007) fournit en sortie des séries de cognats (compris comme des mots similaires au niveau orthographique ayant une origine commune) à travers les cinq langues romanes, à la différence des méthodes traditionnelles qui s'appliquent et donnent des résultats par paires de langues. Cette approche prend en compte la contrainte de transitivité de la relation de *cognation* à travers les langues traitées et maximise une fonction de scores calculés pour les cognats, par l'exploitation d'une technique de programmation linéaire (*Integer Linear Programming - ILP*). Un système *open-source* basé sur cette technique est *lp_solve*[129]. Celui-ci est utilisé par Bergsma et Kondrak (2007) dans leurs expériences d'identification automatique des cognats.

Étant donné une mesure de similarité et la condition de transitivité de la relation de *cognation*, le système classifie les sets des cognats et des mots non apparentés selon un certain seuil de la mesure de similarité utilisée. Dans cette approche, la mesure implémentée est la sous-chaîne maximale (*SCM*) (Melamed, 1999) qui calcule, rappelons-le, le rapport entre le nombre de caractères de la sous-chaîne maximale commune aux deux mots appartenant à une paire bilingue et la longueur du mot le plus long.

Bergsma et Kondrak (2007) effectuent des expériences à partir des séries de mots partageant un même sens (par exemple : *juste* (français), *giusto* (italien), *derecho* (espagnol), *direito* (portugais), *drept* (roumain)). La tâche consiste à vérifier quels mots de chaque série constituent des cognats et à partager cette série dans des groupes de cognats.

Les données de référence utilisées (*gold-standard*) sont extraites de *Comparative Indoeuropean Data Corpus* (Dyen *et al.*, 1992). Ce corpus est disponible en 95 langues de la famille indoeuropéenne. Celui-ci est composé de listes de mots représentant 200 sens de base, munis de l'information sur les cognats (validés par des linguistes).

Ainsi, les 200 séries de mots ont été extraites dans les cinq langues romanes considérées. À partir de ces données, cette méthode exploitant la transitivité de la relation de *cognation*, obtient une bonne performance (*F-mesure* de 61,6%).

De notre côté, la démarche était d'évaluer l'influence des informations linguistiques dans l'identification automatique des cognats. Ainsi, nous avons développé (en Perl) une méthode à base de n-grammes (Simard *et al.*, 1992) et d'informations linguistiques (lemmes, parties du

[129] http://lpsolve.sourceforge.net/

discours) et nous avons comparé les résultats obtenus avec ceux fournis par des méthodes purement statistiques, comme le coefficient de *Dice* et la mesure *SCM*. La méthode proposée ainsi que son évaluation individuelle et comparative feront l'objet des sous-sections suivantes.

4.2.1. Le module d'identification automatique des cognats

D'après la théorie, la détection automatique des cognats permet d'augmenter la performance d'un système d'alignement lexical. La nécessité de développer un module d'identification automatique des cognats a donc été l'une de nos priorités, surtout que *GIZA++* (Och et Ney, 2000, 2003) ne repère pas tous les cognats du corpus et ne prévoit pas non plus de règles spécifiques dans ce but. À partir du corpus parallèle de travail (lemmatisé, étiqueté et aligné au niveau propositionnel - cf. section 3.2), notre méthode exploite les informations linguistiques associées aux unités lexicales, telles que les lemmes et les étiquettes morphosyntaxiques. Ainsi, nous considérons comme cognats les paires bilingues de mots qui remplissent les conditions linguistiques suivantes :

1. leurs lemmes sont des équivalents de traduction dans deux phrases parallèles ;

2. leurs lemmes sont identiques ou présentent des similarités orthographiques et/ou phonétiques ;

3. leurs lemmes ont la même partie du discours ou appartiennent à la même classe d'équivalence de catégorie lexicale (par exemple, un nom peut être traduit par un nom, un verbe ou un adjectif) ; De plus, nous avons filtré les mots courts comme les prépositions et les conjonctions afin de diminuer le bruit.

Notre méthode vise prioritairement l'obtention d'une précision élevée afin d'être intégrée dans le système d'alignement lexical français - roumain. Ainsi, pour augmenter la précision de la méthode, nous avons combiné les conditions linguistiques mentionnées ci-dessus avec d'autres stratégies de désambiguïsation des données d'entrée (extractions itératives des cognats considérés les plus fiables, extraction des candidats cognats les plus fréquents pour les cas ambigus). La configuration choisie dans ce sens sera présentée plus loin.

Au niveau orthographique, nous avons classifié les cognats identifiés dans plusieurs catégories :

1. transfuges (nombres, certains sigles et acronymes, ainsi que les signes de ponctuation) ;

2. cognats identiques (***document*** vs. ***document***) ;

3. cognats similaires remplissant une des conditions suivantes :

   a) 4-grammes (Simard *et al.*, 1992) ; Les cognats ont au moins les 4 premiers caractères du lemme identiques. La longueur des lemmes est égale ou supérieure à 4 (***prod****uit* vs. ***prod****us*) ;

   b) 3-grammes ; Les cognats ont les 3 premiers caractères identiques et la longueur de leurs lemmes est égale ou supérieure à 3 (***act****e* vs ***act***) ;

   c) 8-bigrammes ; Les cognats possèdent une sous-chaîne commune de caractères ordonnés parmi les 8 premiers bigrammes au niveau du lemme. Nous procédons par comparaison des bigrammes avec la condition qu'au moins un caractère de chaque bigramme est commun aux deux lemmes. Cette condition permet les sauts d'un caractère différent (***s****o****u****s****c****r****i****r****e*** vs. ***su****b****s****c****r****i****e***). Dans ce cas, les lemmes ont une longueur supérieure à 7 ;

   d) 4-bigrammes ; Les cognats possèdent une sous-chaîne commune de caractères ordonnés parmi les 4 premiers bigrammes au niveau du lemme. Au moins un caractère de chaque bigramme est commun aux deux lemmes. Dans ce cas, nous considérons aussi bien les lemmes courts (longueur égale ou inférieure à 7) (***gr****o****up****e* vs ***grup***) que les lemmes longs (longueur supérieure à 7) (*h****omol****ogué* vs. ***omol****ogat*).

Dans un premier temps, nous avons appliqué un ensemble d'ajustements orthographiques constitué empiriquement, au niveau des lemmes. Ces ajustements comprennent des simples opérations comme l'élimination des diacritiques et des consonnes doubles (Tufiș *et al.*, 2006 ; Ceaușu, 2009) mais aussi un ensemble de règles d'ajustements basées sur le repérage des correspondances phonétiques. En effet, comme le français a une écriture étymologique et le roumain possède une écriture généralement phonétique, nous avons identifié des correspondances phonétiques au niveau des lemmes et, ensuite, des ajustements orthographiques ont été effectués du français vers le roumain. Par exemple, les cognats *phase* (FR) vs. *fază* (RO) deviennent ***f****a****z****e*** (FR) vs. ***f****a****z****a*** (RO). Ainsi, dans cet exemple, deux ajustements sont réalisés : le groupe consonantique *ph* [f] du français devient [f] comme en

roumain et le *s* [z] intervocalique du français devient [z] comme en roumain. Nous avons fait aussi des ajustements dans les cas ambigus (*ch* ([ʃ] ou [k])) en considérant les deux variantes en étapes successives : *ma**ch**ine* (FR) vs. *maşină* (RO) ; ***ch**lorure* (FR) vs. ***c**lorură* (RO). Les règles ainsi définies afin d'effectuer ces ajustements orthographiques sont sensitives au contexte phonétique comme dans l'exemple *é**q**uilibre* vs. *e**c**hilibru* où ***q*** (qu(+i) (médial)) devient [k], comme en roumain, ou non sensitives au contexte (*sto**ck**age* vs. *sto**c**are*) où le groupe consonantique ***ck*** du français devient aussi [k], comme en roumain. L'ensemble d'ajustements orthographiques appliqué au corpus utilisé figure dans le Tableau 20 ci-dessous :

Tableau 20. Ajustements orthographiques appliqués au corpus parallèle français - roumain

Niveaux d'ajustements orthographiques	FR	RO	Exemples
signes diacritiques	é-e ; à-a…	î-i ; â-a, ă-a…	d**é**p**ô**t - d**e**p**o**zit
lettres identiques contigües	cc-c ; dd-d…	cc-c ; nn-n…	ra**pp**ort - ra**p**ort
groupes consonantiques	ph th dh ck cq ch ch	f [f] t [t] d [d] c [k] c [k] ș [ʃ] c [k]	**ph**ase - **f**ază mé**th**ode - me**t**odă a**dh**érent - a**d**erent sto**ck**age - sto**c**are gre**cq**ue - gre**c** ma**ch**ine - maș**i**nă **ch**lorure - **c**lorură
q	q (final) qu(+i) (médial) qu(+e) (médial) qu(+a) que (final)	c [k] c [k] c [k] c(+a) [k] c [k]	cin**q** - cin**c**i é**q**uilibre - e**c**hilibru mar**q**uer - mar**c**a **q**ualité - **c**alitate prati**q**ue - practi**c**ă
s intervocalique	*v* + s + *v*	*v* + z + *v*	pré**s**ent - pre**z**ent
w	w	v	**w**agon - **v**agon
y	y	i	**y**aourt - **i**aurt

Dans un deuxième temps, nous avons appliqué sept étapes d'extractions itératives des cognats par catégorie identifiée, dans l'ordre qui permet l'obtention d'une précision élevée de chaque étape (cf. Tableau 21 suivant).

Tableau 21. Étapes d'extraction automatique de cognats français - roumain

Étapes d'extraction par catégorie de cognats	Même catégorie lexicale	Fréquence des candidats ambigus	Suppression des données d'entrée	Précision (%)
1 : transfuges			x	100
2 : cognats identiques	x		x	100
3 :4-grammes (longueur des lemmes >= 4)	x	x	x	99,05
4 : 3-grammes (longueur des lemmes >= 3)	x	x	x	93,13
5 : 8-bigrammes (lemmes longs, longueur >7)	x		x	95,24
6 : 4-bigrammes (lemmes longs, longueur > 7)	x			75
7 : 4-bigrammes (lemmes courts, longueur =< 7)	x	x		65,63

Cette procédure a été appliquée en premier aux candidats ayant la même catégorie lexicale (N-N - un nom équivalent à un nom, V-V - un verbe équivalent à un verbe, etc.). Ensuite, la même procédure a été utilisée pour les candidats présentant des équivalences de catégorie lexicale (un nom (N) peut être traduit par un verbe (V) ou un adjectif (ADJ)).

De plus, pour limiter le bruit des résultats, deux stratégies supplémentaires de désambiguïsation des données d'entrée ont été appliquées.

Tout d'abord, les candidats ambigus (un même lemme source apparaît avec deux ou plusieurs candidats cibles : *autorité* (FR) vs. *autoritate* | *autorizare* (RO) 'autorité | autorisation') ont été filtrés en calculant la fréquence des paires candidates dans le corpus étudié. En effet, nous avons sélectionné les candidats ambigus munis de leurs fréquences dans le corpus et la paire candidate la plus fréquente a été retenue. Cette opération est très efficace pour augmenter la précision des résultats mais, dans certains cas, elle décroît le rappel par la perte des candidats cognats pertinents. Ainsi, concernant les déverbaux, certains lemmes français présentant une seule forme ont comme équivalents de traduction en roumain deux formes différentes : *information* (FR) vs. *informaţie* ou *informare* (RO) *; manifestation* (FR) *vs. manifestaţie* ou *manifestare* (RO). Ces paires ont été récupérées en utilisant des expressions régulières basées sur les terminaisons spécifiques des lemmes (*ion* (FR) vs. *ţie*|*re* (RO)).

Une autre stratégie de désambiguïsation des données d'entrée et notamment la suppression du corpus des cognats considérés fiables (précision élevée), à la fin de chaque étape d'extraction, a été également utilisée. Par exemple, les cognats identiques *transport* vs. *transport* obtenus pendant l'étape d'extraction correspondante et supprimés des données d'entrée, éliminent

l'occurrence du candidat *transport* vs. *tranzit* comme 4-grammes cognats dans l'étape suivante.

Après avoir extrait les candidats ayant la même catégorie lexicale (N-N : *transport* - *transport*), nous avons appliqué la même méthode d'extraction pour les cognats présentant des équivalences de catégorie lexicale (V-N : *transporter* - *transport*, ADJ-N : *transporté* - *transport*). Dans ce cas, seulement les cognats 4-grammes ont été gardés, puisque nous avons observé une diminution importante de la précision pour les autres catégories considérées (les étapes 4-7 présentées dans la Tableau 21).

Nous avons évalué notre méthode de détection automatique des cognats en faisant aussi une comparaison des résultats avec ceux fournis par des méthodes purement statistiques, pour étudier l'impact des informations linguistiques utilisées sur la qualité des résultats obtenus. L'étape d'évaluation a montré que l'utilisation des informations linguistiques augmente significativement la performance de la méthode. Ces expériences seront décrites dans la sous-section suivante.

4.2.2. L'évaluation du module et la comparaison de méthodes

L'évaluation de notre module d'identification automatique des cognats a été faite en termes de précision, rappel et F-mesure, par rapport à une liste de cognats de référence constituée manuellement, à partir du corpus parallèle de test (décrit dans la sous-section 3.2.1.). Cette liste contient 2 034 cognats français - roumain. Ensuite, nous avons comparé les résultats obtenus avec ceux fournis par les méthodes statistiques suivantes :

1) le calcul de la mesure *SCM* (Melamed, 1999) qui prend en compte la longueur de la sous-chaîne maximale (*SCM*) de caractères communs aux deux mots d'une paire bilingue et la longueur du mot le plus long (cf. formule ci-dessous) :

$$mesure_{SCM}(mot_1, mot_2) = \frac{longueur(sous\_chaîne\_commune(mot_1, mot_2))}{\max(longueur(mot_1), longueur(mot_2))}$$

   Les mots sont considérés comme cognats si la valeur de la mesure *SCM* est supérieure ou égale à 0.68 (seuil établi empiriquement à partir de notre corpus).

2) le calcul du coefficient de *Dice* (Adamson et Boreham, 1974 ; Brew et McKelvie, 1996) qui prend en compte le nombre de caractères des bigrammes communs aux deux

mots considérés et le nombre total des bigrammes des deux mots (cf. formule suivante) :

$$Dice(mot_1, mot_2) = \frac{2 \times nombre\_bigrammes\_communs}{nombre\_total\_bigrammes(mot_1) + nombre\_total\_bigrammes(mot_2)}$$

Dans ce cas, les mots sont retenus comme cognats si la valeur de ce coefficient est supérieure ou égale à 0,62 (valeur établie empiriquement à partir de notre corpus).

3) 4-grammes (Simard *et al.*, 1992) ; Les mots sont considérés comme cognats s'ils comprennent au moins 4 caractères et si leurs premiers 4 caractères sont identiques.

Ces méthodes s'appliquent généralement pour des mots ayant une longueur égale ou supérieure à 4 pour réduire le bruit. En outre, les paires de cognats sont recherchées dans des phrases parallèles alignées.

Nous avons implémenté les trois méthodes mentionnées ci-dessus en utilisant le corpus contenant l'ensemble d'ajustements orthographiques effectués au préalable au niveau des lemmes (cf. Tableau 20). De plus, la méthode 4-grammes (Simard *et al.*, 1992) a été appliquée aussi sur le corpus sans ajustements orthographiques afin de tester l'influence de ces ajustements sur les résultats d'une méthode à base de n-grammes.

Dans le Tableau 22 ci-dessous figurent les valeurs des scores d'évaluation obtenues pour toutes les méthodes implémentées.

Tableau 22. Évaluation du module développé et comparaison avec d'autres méthodes

Méthodes	Précision	Rappel	F-mesure
SCM + Ajustements	**44,13%**	58,95%	**50,47%**
DICE + Ajustements	56.47%	60.91%	58.61%
4-grammes - Sans ajustements	90,85%	**47,84%**	62,68%
4-grammes + Ajustements	91,55%	72,42%	80,87%
Notre méthode	**94,78%**	**89,18%**	**91,89%**

Notre méthode a extrait 1 814 cognats corrects sur 1 914 candidats cognats fournis, obtenant ainsi les meilleurs scores (précision=94,78% ; rappel=89,18% ; f-mesure=91,89%), par rapport aux autres méthodes implémentées. La méthode 4-grammes appliquée sur le corpus initial (sans ajustements orthographiques) a obtenu une bonne précision (90,85%) mais un rappel faible de 47,84%. L'étape d'ajustements orthographiques au niveau des lemmes a amélioré nettement le rappel de la méthode 4-grammes de presque 25% (de 47,84% à 72,42%). Cela s'explique par les spécificités du corpus juridique utilisé où les termes provenant du français par des emprunts ou des calques sont nombreux.

Les scores les plus faibles ont été obtenus par la mesure *SCM* (f-mesure=50,47%), suivie par la méthode basée sur le coefficient de *Dice* (f-mesure=58,61%). Ces approches générales produisent beaucoup de bruit dû aux ressemblances formelles importantes entre des mots qui ne présentent aucun lien au niveau sémantique. Leurs résultats pourraient être améliorés en combinant des techniques statistiques avec d'autres informations (équivalences de catégorie lexicale, par exemple) ou en combinant plusieurs scores d'association.

Toutefois, notre méthode d'identification automatique des cognats a extrait un ensemble de candidats ambigus, tels que : *disposition* (FR) - *dispoziţie* 'disposition' | *dispozitiv* 'dispositif' (RO), *directive* | *direction* (FR) - *directivă* 'directive' (RO). Comme nous l'avons déjà précisé auparavant, nous avons éliminé une partie de ces erreurs en gardant la paire candidate la plus fréquente (*disposition* - *dispoziţie*, *directive* - *directivă).* Néanmoins, dans le cas où ces candidats ont des fréquences égales, les paires candidates erronées ne sont pas éliminées. D'autres erreurs restantes concernent les candidats hapax incorrects qui ne peuvent pas être éloignés par le critère de la fréquence car ils apparaissent une seule fois (p. ex. *numéro* - *nume* 'nom', *compléter* - *compune* 'composer'). De plus, certains cognats n'ont pas été identifiés : *heure* - *oră*, *semaine* - *săptămână*, *lieu* - *loc*. Ces erreurs concernent les cognats présentant des similarités orthographiques et/ou phonétiques très faibles.

En conclusion, notre méthode d'identification automatique des cognats fournit des résultats performants par rapport à des méthodes statistiques pures. Cependant, les résultats dépendent de la paire de langues étudiées, du domaine du corpus utilisé et aussi du volume des données. Des études similaires restent encore nécessaires sur d'autres types de corpus parallèles français - roumain, afin de pouvoir généraliser. Toutefois, notre méthode s'avère efficace pour l'identification des cognats à partir des corpus juridiques.

Le module développé aligne également les candidats retenus après leur détection automatique. Nous avons évalué l'impact de l'identification et de l'alignement des cognats sur le système d'alignement lexical de base. Nous avons montré que l'intégration des cognats dans le processus d'alignement lexical améliore significativement les résultats (Navlea et Todirașcu, 2012). Les expériences effectuées dans ce sens seront présentées dans la sous-section suivante.

4.2.3. L'évaluation de l'alignement des cognats

Après avoir aligné les cognats, nous avons également mis en correspondance les termes poly-lexicaux contenant des cognats identifiés au préalable. Ensuite, ces alignements ont été rajoutés à ceux fournis par le système de base. L'alignement résultant a été évalué en termes de précision, rappel, f-mesure et score *AER*, par comparaison avec nos deux alignements de référence: *RefA* - sans alignement de collocations, respectivement *RefB* - avec alignement de collocations (cf. sous-section 4.1.1.). Les conditions de l'évaluation sont les mêmes que pour le système de base (alignements nuls ignorés, aucune distinction entre les alignements sûrs et possibles dans les alignements de référence) (cf. sous-section 4.1.1.).

Afin de trouver la meilleure solution pour exploiter les cognats pendant le processus d'alignement lexical, nous avons effectué des expériences dans deux directions :

1) filtrer les résultats du système de base par les alignements des cognats ;

2) enrichir les résultats du système de base par les alignements des cognats.

Pour ce faire, nous avons utilisé seulement le corpus de référence *RefA*, c'est-à-dire le corpus ne contenant pas les alignements des collocations. Les résultats d'évaluation obtenus lors de ces expériences d'alignement lexical figurent dans le Tableau 23 suivant :

Tableau 23. L'évaluation de l'alignement des cognats

Systèmes	**Précision (%)**	**Rappel (%)**	**F-mesure (%)**	***AER* (%)**
	RefA	***RefA***	***RefA***	***RefA***
1. système de base	95,56	52,91	68,11	31,89
2. système de base + filtrage par cognats	96,51	42,31	58,83	41,17
3. système de base + alignements des cognats	95,13	54,53	69,32	30,68
4. système de base + alignements des cognats + alignements des termes poly-lexicaux contenant des cognats	95,23	55,72	70,30	29,70

Dans un premier temps, nous avons filtré les résultats du système de base par les alignements des cognats afin d'augmenter la précision du système. En effet, nous avons gardé seulement les alignements des cognats corrects fournis par le système de base (*sens - sens, améliorer - ameliora*). Dans ce cas, la précision du système a augmenté de 95,56% à 96,51%, mais nous avons remarqué une diminution importante du rappel de 52,91% à 42,31%, ce qui a augmenté le score *AER* de 9,28% (de 31,89% à 41,17%). Cela est dû principalement à une synonymie importante dans le corpus (*sens - sens* vs. *sens - înţeles* ; *améliorer - ameliora* vs. *améliorer - îmbunătăţi*). Mais notre système n'intègre pas encore des techniques pour gérer la synonymie.

Dans un deuxième temps, nous avons enrichi les résultats du système de base par les alignements des cognats afin d'augmenter le rappel du système. Dans ce cas, le rappel a augmenté de 1,62% (de 52,91% à 54,53%) et le score *AER* a diminué de manière intéressante de 1,21% (de 31,89% à 30,68%). De plus, nous avons exploité les cognats retenus pour collecter les alignements des termes ploylexicaux contenant des cognats identifiés. Cette étape concerne les termes fournis en roumain par l'étiqueteur *TTL* (Ion, 2007). En effet, l'étiqueteur reconnaît un ensemble de termes poly-lexicaux en roumain (*cooperare_europeană* 'coopération_européenne', *autorizaţie_de_transport* 'autorisation_de_transport'), tandis qu'en français ces termes sont tokenisés. Un exemple d'alignement de ces termes apparaît dans la Figure 61 ci-dessous :

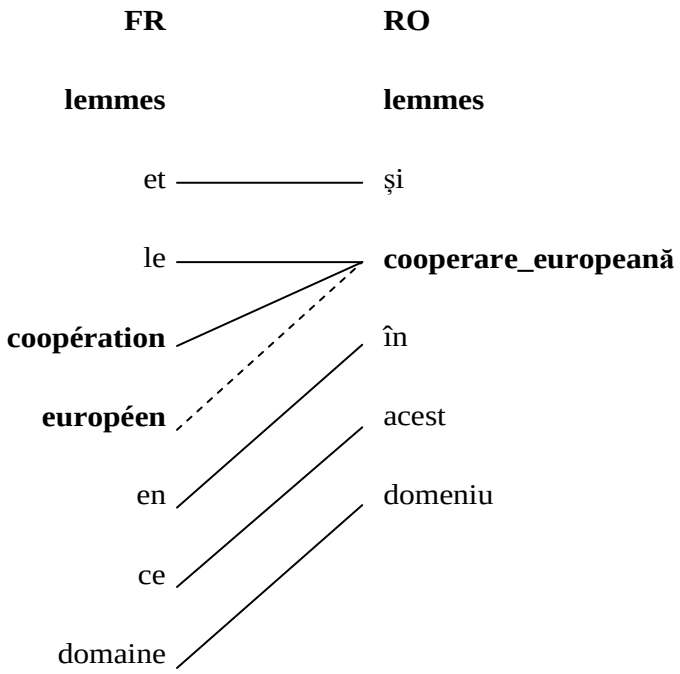


Figure 61. Exemple d'alignement des termes poly-lexicaux reconnus par *TTL* en roumain

Ont été obtenus 477 alignements supplémentaires corrects qui ont augmenté le rappel de 1,19% (de 54,53% à 55,72%) et ont diminué le score *AER* de 0,98% (de 30,68% à 29,70%), par rapport au système additionnant seulement les alignements des cognats 1:1.

Ainsi, l'intégration des cognats dans le processus d'alignement lexical a amélioré la performance du système : le rappel a augmenté de presque 3% (de 52,91% à 55,72%) et le score *AER* a diminué de manière intéressante d'environ 2% (de 31,89% à 29,70%), par rapport au système de base.

Nous avons évalué l'influence des cognats identifiés sur le système de base en utilisant aussi le corpus de référence *RefB*, c'est-à-dire le corpus contenant les alignements des collocations.

Le Tableau 24 suivant comprend les scores obtenus pour *RefB*, par rapport à *RefA*.

Tableau 24. Évaluation de l'influence des cognats sur le système de base

Systèmes	Précision (%)		Rappel (%)		F-mesure (%)		*AER* (%)	
	RefA	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***
1 : **système de base**	95,56	95,51	52,91	49,71	68,11	65,39	31,89	34,61
2 : système de base + **cognats**	95,23	95,18	55,72	52,35	70,30	67,55	29,70	32,45

Concernant le corpus de référence *RefB*, le rappel a augmenté de 2,64% (de 49,71% à 52,35%) et le score *AER* a diminué de 2,16% (de 34,60% à 32,45%), par rapport au système de base. Ainsi, les résultats obtenus pour *RefB* sont comparables avec ceux de *RefA* qui diminue le score *AER* de 2,19% (de 31,89% à 29,70%).

4.2.4. Bilan de la section

Nous venons de présenter dans cette section un module d'identification et d'alignement automatique des cognats développé pour le français et le roumain. Ce module utilise une combinaison originale de techniques statistiques, informations linguistiques et ajustements orthographiques entre les mots des paires bilingues de lemmes, ce qui augmente significativement la performance de ses résultats. Celui-ci obtient aussi de meilleurs scores par rapport à des méthodes statistiques traditionnelles comme le coefficient de *DICE* ou la mesure *SCM*.

Toutefois, les résultats dépendent des langues étudiées, du domaine et de la qualité de l'étiquetage du corpus, ainsi que du volume des données. D'autres expériences sur des corpus appartenant à d'autres domaines restent encore nécessaires afin de pouvoir généraliser, mais la méthode s'avère efficace pour le domaine juridique étudié.

Dans la section suivante, nous allons nous intéresser au module implémentant les règles heuristiques morphosyntaxiques et stylistiques définies à l'issue de l'analyse linguistique des erreurs du système d'alignement lexical de base (cf. sous-section 4.1.3.) et de l'étude du corpus bilingue parallèle au niveau de la traduction humaine (cf. sous-section 4.1.4.). Ainsi, nous y allons étudier l'impact des règles heuristiques définies sur la qualité de l'alignement lexical construit auparavant.

4.3. Les règles heuristiques morphosyntaxiques et stylistiques

Le module développé est intégré dans le système d'alignement lexical final et implémente un ensemble de 37 règles heuristiques morphosyntaxiques et stylistiques. Nous allons expliquer l'algorithme proposé et allons donner des exemples de règles dans la sous-section suivante.

4.3.1. L'implémentation des règles heuristiques

Les règles définies sont destinées à corriger les erreurs d'alignement lexical dues aux différences morphosyntaxiques entre les deux langues étudiées ainsi qu'aux contraintes stylistiques imposées par la traduction juridique. Ces règles utilisent des informations linguistiques telles que les étiquettes morphosyntaxiques et les lemmes associés aux unités lexicales, ainsi que leurs formes fléchies. Les règles définies s'appliquent donc à partir des corpus bilingues parallèles alignés au niveau propositionnel, étiquetés et lemmatisés.

Il s'agit des règles contextuelles qui vérifient des conditions simultanées aux niveaux morphosyntaxique ou stylistique dans deux phrases bilingues parallèles afin de collecter des alignements manquants. Les nouveaux alignements sont rajoutés en fonction des alignements obtenus lors des étapes antérieures du processus d'alignement lexical. En effet, l'algorithme fait appel à un fichier intermédiaire contenant les alignements des étapes précédentes. Si un alignement existe dans ce fichier intermédiaire et les conditions morphosyntaxiques ou stylistiques sont satisfaites autour de cet alignement, un nouvel alignement est rajouté à l'alignement final. Dans le cas contraire, une nouvelle recherche est effectuée. De plus, la recherche s'effectue dans une fenêtre de 10. Cette distance a été établie empiriquement. La limitation de l'espace de recherche est nécessaire afin d'éviter les alignements entre des unités lexicales très éloignées, à l'intérieur des phrases parallèles, qui, généralement, ne sont pas corrects. Cette restriction est possible si l'on considère qu'en traduction les phrases tendent à avoir une longueur et un ordre des mots comparables, même si ce n'est pas toujours le cas. Toutefois, pour notre corpus de travail, cette restriction est adéquate car, dans la traduction juridique, la plupart des phrases parallèles remplissent les critères mentionnés.

L'avantage de cet algorithme, par rapport à ceux purement statistiques utilisant seulement les distances locales entre les mots source et cible, consiste dans la collecte des alignements considérés précis, car il exploite des schémas morphosyntaxiques définies manuellement. Par contre, le rappel est plus faible car la recherche est limitée à une fenêtre de 10 et les

alignements existants en dehors de cette limite ne sont évidemment pas collectés. Or, il existe aussi des cas ou les traductions commencent par la fin de la phrase source et, si les phrases en question ont une longueur supérieure à 10, aucun alignement ne sera collecté, même si les conditions morphosyntaxiques ou stylistiques sont remplies. D'autres situations où la distance entre les unités lexicales correspondantes devient plus longue sont illustrées par les traductions qui présentent beaucoup d'explicitations ou renforcements de sens. Donc, l'algorithme proposé s'avère efficace pour les phrases qui gardent une longueur et un ordre de mots semblables, critères qui sont généralement remplis par les phrases du corpus juridique étudié.

La première catégorie de règles heuristiques définies est constituée par les règles morphosyntaxiques qui s'avèrent très productives en représentant la plupart (60,52%) des alignements obtenus. Rappelons que les règles heuristiques morphosyntaxiques concernent l'alignement des déterminants et des unités lexicales qui entrent dans des structures morphosyntaxiques spécifiques à chaque langue, tels que les pronoms relatifs, personnels ou réfléchis, certaines prépositions, les particules d'infinitif ou subjonctif, etc. (cf. sous-section 4.1.3.).

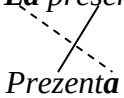
Soient les deux exemples de séquences parallèles français - roumain (1 et 2) ci-dessous accompagnés de leurs règles heuristiques respectives transposées en pseudo-code. Ces règles morphosyntaxiques concernent l'alignement des déterminants définis. En effet, celles-ci mettent en correspondance le déterminant défini *le, la, les* du français qui précède le nom (1) ou l'adjectif antéposé (2) avec le nom (1) ou l'adjectif antéposé (2) au singulier ou au pluriel du roumain présentant le déterminant défini en tant que suffixe.

1) ***La*** *directive est…* (FR)

*Directiv**a** este…* (RO)

Règle : if nr_phrase_fr = nr_phrase_ro && source=fr && |i-j| < 11 && lema[i]="le" && ana[i+1]="N*" && target=ro && ana[j]="N*ry" && aligned(i+1,j) then align(i,j)

2) ***La*** *présente directive est…* (FR)



*Prezent**a** directivă este…* (RO)

Règle : if nr_phrase_fr = nr_phrase_ro && source=fr && |i-j| < 11 && lema[i]="le" && ana[i+1]="A*" && target=ro && ana[j]="A*ry" && aligned(i+1,j) then align(i,j)

Où :

- nr_phrase_fr : la position de la phrase source dans le corpus parallèle ;
- nr_phrase_ro : la position de la phrase cible dans le corpus parallèle ;
- && : signifie « et » ;
- source=fr : la langue source est le français ;
- i : la position du mot source dans le phrase source ;
- j : la position du mot cible dans la phrase cible ;
- |i-j| < 11 : la recherche s'effectue dans une fenêtre de 10 ;
- lema[i] : le lemme du mot source sur la position i ;
- ana[i+1] : l'étiquette morphosyntaxique du mot source sur la position i+1 ;
- target=ro : la langue cible est le roumain ;
- ana[j] : l'étiquette morphosyntaxique du mot cible sur la position j ;
- aligned(i+1,j) : condition qui vérifie si le mot source sur la position i+1 est déjà aligné au mot cible sur la position j pendant les étapes d'alignement précédentes ;
- align(i,j) : fonction qui aligne le mot source sur la position i au mot cible sur la position j.
- "N*": l'étiquette identifie tous les noms ;
- "N*ry": l'étiquette identifie tous les noms (N) au nominatif-accusatif (r), présentant le déterminant défini (y) ;
- "A*": l'étiquette identifie tous les adjectifs (A) ;
- "A*ry": l'étiquette identifie tous les adjectifs au nominatif-accusatif (r), présentant le déterminant défini (y).

Nous avons aussi tenu compte des éléments supplémentaires qui peuvent apparaître à l'intérieur des séquences bilingues parallèles à aligner comme dans l'exemple 3 suivant. Ces éléments sont représentés par des variables (*k* - pour le français et *p* - pour le roumain) dans la syntaxe des règles définies. Cet exemple concerne l'alignement de la préposition *de* du français avec le morphème de génitif *al* du roumain.

3) … *une proposition d'élaboration **d**'un nouveau règlement technique mondial*… (FR)

… *o propunere de elaborare **a** unei noi norme tehnice mondiale*…(RO)

Règle : if nr_phrase_fr = nr_phrase_ro && source=fr && |i-j| < 11 && lema[i]="de|de+le" && 1<=k && k<=5 && ana[i+k]="^N" && target=ro && lema[j]="al" && 1<=p && p<=5 && ana[j+p]="N*oy|N*-n|N*on" && ana[j+1] ne "^Mo" && aligned(i+k,j+p) then align(i,j)

Où :

- nr_phrase_fr : la position de la phrase source dans le corpus parallèle ;
- nr_phrase_ro : la position de la phrase cible dans le corpus parallèle ;
- && : signifie « et » ;
- source=fr : la langue source est le français ;
- i : la position du mot source dans le phrase source ;
- j : la position du mot cible dans la phrase cible ;
- |i-j| < 11 : la recherche s'effectue dans une fenêtre de 10 ;
- lema[i] : le lemme du mot source sur la position i ;
- k : variable ayant des valeurs entre 1 et 5 ;
- ana[i+k] : l'étiquette morphosyntaxique du mot source sur la position i+k ;

- target=ro : la langue cible est le roumain ;
- lema[j] : le lemme du mot cible sur la position j ;
- p : variable ayant des valeurs entre 1 et 5 ;
- ana[j+p] : l'étiquette morphosyntaxique du mot cible sur la position j+p ;
- ana[j+1] : l'étiquette morphosyntaxique du mot cible sur la position j+1 ;
- aligned(i+k,j+p) : condition qui vérifie si le mot source sur la position i+k est déjà aligné au mot cible sur la position j+p pendant les étapes d'alignement précédentes ;
- align(i,j) : fonction qui aligne le mot source sur la position i au mot cible sur la position j.
- "^N" : l'étiquette identifie tous les noms ;
- "N*oy|N*-n|N*on" : l'étiquette identifie les noms au génitif-datif (o) déterminés défini (y) et non déterminés (n) ;
- "^Mo" : l'étiquette identifie les numéraux ordinaux.

Dans l'exemple ci-dessus, *k* et *p* peuvent prendre des valeurs comprises entre 1 et 5. Ces variables peuvent manquer en ce qui concerne certaines règles (comme dans les exemples 1 et 2), mais elles peuvent aussi être présentes simultanément ou individuellement selon le cas. Pour chaque règle définie, nous avons établi empiriquement les limites inférieures et supérieures des variables supplémentaires introduites.

La deuxième catégorie de règles heuristiques définies est représentée par les règles stylistiques. Rappelons que (cf. sous-section 4.1.4.) ces règles concernent l'alignement des lexèmes spécifiques au style juridique et administratif, qui ne sont pas présents d'une langue à l'autre comme certains adjectifs qualificatifs (*présent - prezent, certain - anumit, respectif - respectiv*) ou déterminants indéfinis (*tout - tot, chaque - fiecare*). Quand ils apparaissent dans la langue cible et ne sont pas présents dans la langue source, ces lexèmes renforcent le sens du nom qu'ils précèdent et présentent une valeur stylistique. D'autres règles stylistiques concernent les mises en correspondance des structures passives ou impersonnelles à l'intérieur de deux phrases bilingues parallèles.

Soit l'exemple suivant :

4) *La notification est…* (FR)

***Respectiva** notificare este…* (RO)

Règle : if nr_phrase_fr = nr_phrase_ro && source=fr && |i-j| < 11 && ana[i]="Da*" && ana[i+1]="^N" && target=ro && lema[j]="respectiv|prezent|asemenea|anumit" && ana[j+1]="^N" && aligned(i+1,j+1) then align(i,j) && align(i+1,j)

Où :

- nr_phrase_fr : la position de la phrase source dans le corpus parallèle ;
- nr_phrase_ro : la position de la phrase cible dans le corpus parallèle ;
- && : signifie « et » ;

- source=fr : la langue source est le français ;
- i : la position du mot source dans le phrase source ;
- j : la position du mot cible dans la phrase cible ;
- |i-j| < 11 : la recherche s'effectue dans une fenêtre de 10 ;
- ana[i] : l'étiquette morphosyntaxique du mot source sur la position i ;
- ana[i+1] : l'étiquette morphosyntaxique du mot source sur la position i+1 ;
- target=ro : la langue cible est le roumain ;
- lema[j]="respectiv|prezent|asemenea|anumit" : le lemme du mot cible sur la position j ;
- ana[j+1] : l'étiquette morphosyntaxique du mot cible sur la position j+1 ;
- aligned(i+1,j+1) : condition qui vérifie si le mot source sur la position i+1 est déjà aligné au mot cible sur la position j+1 pendant les étapes d'alignement précédentes ;
- align(i,j) : fonction qui aligne le mot source sur la position i au mot cible sur la position j ;
- align(i+1,j) : fonction qui aligne le mot source sur la position i+1 au mot cible sur la position j.

La règle ci-dessus vérifie si dans deux phrases bilingues parallèles un nom de la langue cible est précédé par l'un des adjectifs qualificatifs *respectiv* 'respectif', *prezent* 'présent', *asemenea* 'tel', *anumit* 'certain', pendant que le nom correspondant de la langue source n'est pas déterminé par un adjectif. Dans ce cas, l'adjectif qualificatif cible *respectiva* sera aligné avec le nom source *notification* et son déterminant défini *la.*

Après l'application des règles heuristiques, nous avons évalué leur influence sur la qualité des résultats de l'alignement lexical. L'évaluation effectuée sera présentée dans la sous-section suivante.

4.3.2. L'évaluation du module à base de règles heuristiques

L'alignement lexical obtenu à l'issue de l'étape d'application des règles heuristiques définies a été évalué en termes de précision, rappel, f-mesure et score *AER* par comparaison avec nos deux alignements de référence (*RefA*, sans alignement des collocations ; *RefB*, avec alignement des collocations - cf. sous-section 4.1.1.). Précisons que les conditions de l'évaluation sont les mêmes que pour l'évaluation des étapes précédentes d'alignement lexical (alignements nuls ignorés, aucune distinction entre les alignements sûrs et possibles dans les alignements de référence) (cf. sous-section 4.1.1.).

Les valeurs des scores d'évaluation obtenues figurent dans le Tableau 25 suivant :

Tableau 25. Évaluation du module d'application des règles heuristiques

Systèmes	Précision (%)		Rappel (%)		F-mesure (%)		*AER* (%)	
	RefA	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***
1 : **système de base**	95,56	95,51	52,91	49,71	68,11	65,39	31,89	34,61
2 : système de base + **cognats**	95,23	95,18	55,72	52,35	70,30	67,55	29,70	32,45
3 : système de base + cognats + **règles heuristiques**	92,55	92,21	63,15	60,91	75,07	73,36	24,93	26,64

Les règles heuristiques ont une contribution significative à l'amélioration du score *AER* par rapport à l'étape précédente d'alignement lexical, c'est-à-dire d'identification et d'alignement automatique des cognats. Ces règles ont déprécié la précision du système de 2,68% pour *RefA* et de 2,97% pour *RefB* mais ont augmenté son rappel de 7,43% pour *RefA* et de 8.56% pour *RefB*, ce qui a mené à l'amélioration significative du score *AER* de 4,77% pour *RefA* et de 5,81% pour *RefB*.

Dans cette section, a été présenté et évalué le module implémentant les règles heuristiques morphosyntaxiques et stylistiques définies manuellement afin d'améliorer les résultats du système d'alignement lexical de base.

Dans la section suivante, nous allons décrire la dernière étape du système d'alignement lexical construit et notamment l'intégration d'un dictionnaire de collocations au processus d'alignement afin de corriger les erreurs fournies au niveau des collocations par le système statistique de base. Ainsi, nous allons présenter les outils et les ressources disponibles pour différentes paires de langues afin d'exploiter les collocations dans les systèmes d'alignement lexical. La sous-section 4.4.1. présente la notion de collocation prise en compte dans notre approche. Ensuite, dans la sous-section 4.4.2., le dictionnaire exploité dans notre système d'alignement lexical sera décrit. Puis, la sous-section 4.4.3. concerne l'algorithme d'alignement des collocations. Enfin, l'évaluation de cette dernière étape d'alignement lexical figure dans la sous-section 4.4.4.

4.4. L’application d’un dictionnaire de collocations français - roumain

Afin d’améliorer les résultats des méthodes d’alignement lexical requises par les systèmes de traduction automatique statistique, des ressources externes comme les dictionnaires de collocations peuvent y être intégrées, sachant que les collocations sont très fréquentes dans un corpus de la langue générale volumineux ou même dans un corpus juridique et administratif comme le nôtre. Des méthodes d’extraction automatique de collocations à partir de corpus monolingues ou multilingues sont aussi appliquées *ad hoc* afin d’améliorer les résultats des systèmes d’alignement lexical (Ren *et al.*, 2009).

Certaines méthodes d’alignement lexical appliquent les collocations à partir de l’étape de segmentation lexicale (Lamber et Banchs, 2005), avant de commencer le processus d’alignement et montrent ainsi une amélioration de la traduction automatique statistique. D’autres méthodes utilisent des listes de collocations ou dictionnaires pour collecter des alignements multiples (Wehrli *et al.*, 2009 ; Tiedemann, 1999).

Les approches utilisant des dictionnaires de collocations multilingues ou monolingues concernent différentes langues comme l’anglais et le danois (Braasch et Olsen, 2000), le français et l’allemand (Blumenthal, 2007), l’espagnol (*DICE* (Ramos *et al.*, 2010)), le français (*LAF*, *Dico* (Polguère, 2000)). Pour la paire de langues étudiées, il existe seulement le dictionnaire de collocations de Todirașcu *et al.* (2008). Ce dictionnaire comprend aussi la paire de langues français - allemand. C’est ce dictionnaire qui sera exploité dans notre approche afin de collecter des alignements multiples. La limite de ce dictionnaire réside dans le fait qu’il comprend seulement la classe des collocations verbo-nominales (*Verbe + Nom*) (Gledhill, 2007). Celui-ci sera décrit plus loin.

Certains de ces dictionnaires sont construits manuellement et comprennent seulement des listes partielles de collocations. D’autres dictionnaires sont disponibles pour des catégories spécifiques comme les noms de sentiment (*DICE* (Ramos *et al.*, 2010)) ou les collocations français - allemand *Nom + Nom* (Blumenthal, 2007). Toutefois, ces dictionnaires sont incomplets et leur disponibilité est limitée à certaines langues ou paires de langues. C’est pourquoi, d’autres approches se concentrent sur les méthodes d’extraction automatique de collocations à partir de corpus monolingues (Hausmann, 1979 ; Evert, 2004 ; Tutin, 2004 ; Heid et Ritz, 2005 ; Ritz et Heid, 2006 ; Seretan, 2008 ; Todirașcu *et al.*, 2008 ; Wehrli *et al.*,

2009 ; Bouamor *et al.*, 2012). Il existe ainsi plusieurs approches de détection automatique de collocations divisées en trois catégories : statistiques, linguistiques et hybrides.

D'une part, les méthodes statistiques (Cowie, 1981 ; Quasthoff, 1998 ; Evert, 2004) considèrent les collocations comme des cooccurrences fréquentes de mots. Ces méthodes utilisent de gros corpus et obtiennent une faible précision due au nombre élevé des candidats existants. D'autre part, les approches linguistiques (Hausmann, 1979 ; Grossmann et Tutin, 2003) considèrent que les collocations sont liées par des relations syntaxiques, sémantiques ou pragmatiques. Au niveau syntaxique, les collocations sont extraites à partir de corpus annotés (Tutin, 2004 ; Seretan, 2008). Ces méthodes obtiennent une précision élevée par rapport aux approches purement statistiques, mais elles sont dépendantes de la qualité de l'annotation syntaxique. Enfin, les approches hybrides tentent de tirer profit des deux méthodes - statistique et linguistique respectivement - en les combinant (Ritz et Heid, 2006 ; Todirașcu *et al.*, 2008 ; Wehrli *et al.*, 2009). Ces méthodes ne sont pas directement utilisables dans les systèmes d'alignement lexical requis par les systèmes de traduction automatique statistique, vu les résultats imprécis nécessitant une intervention manuelle pour la sélection des candidats.

Nous venons de mentionner quelques dictionnaires et méthodes d'extraction automatique de collocations pouvant être exploités dans des systèmes d'alignement lexical afin d'améliorer leurs résultats. Dans la sous-section suivante, nous allons donner la définition des collocations prise en compte dans notre étude et nous allons présenter aussi quelques problèmes liés à l'alignement des collocations en contexte.

4.4.1. Collocations verbo-nominales en contexte

Le dictionnaire de collocations (Todirașcu *et al.*, 2008) exploité dans notre approche a été construit dans le cadre d'une approche contextualiste (Halliday, 1985 ; Banks, 2005 ; Gledhill, 2007). Comme nous l'avons déjà mentionné, ce dictionnaire comprend seulement la classe de collocations verbo-nominales (Gledhill, 2007). De ce fait, notre étude vise seulement cette classe de collocations.

Nous considérons les collocations « des expressions poly-lexicales, semi-figées, parfois discontinues, ayant un comportement morphosyntaxique et sémantique très particulier mais imprévisible » (Gledhill et Todirașcu, 2008 : 137). Ainsi, les collocations satisfont trois critères (Gledhill, 2007 ; Gledhill et Todirașcu, 2008 ; Todirașcu *et al.*, 2008) :

a) le critère de la fréquence (selon lequel les éléments composant une collocation sont des cooccurrences fréquentes dans la même phrase et dans des contextes similaires) ;

b) le critère syntaxique (en fonction duquel les éléments des collocations ne sont pas indépendants mais ils sont liés par certaines propriétés syntaxiques) ; Ainsi, plusieurs classes de collocations ont été identifiées (Hausmann, 2004) : *Nom + Nom* (le deuxième nom est un modifieur du premier), *Verbe + Nom* (le nom est le complément d'objet direct du verbe), etc.

c) le critère sémantique (selon lequel le sens d'une collocation n'est pas compositionnel, c'est-à-dire qu'il n'est pas possible d'établir le sens global de la collocation à partir du sens des éléments qui la composent).

Afin de pouvoir élaborer un algorithme d'alignement des collocations, nous avons procédé à une étude empirique du corpus parallèle concernant les collocations verbo-nominales et leurs équivalents de traduction suivant la méthodologie présentée dans Todirașcu *et al.* (2008). Nous avons constaté qu'il est difficile de repérer automatiquement les équivalents de traduction des collocations verbo-nominales à cause de deux raisons principales : d'une part, leurs équivalents de traduction varient souvent d'une langue à l'autre et, d'autre part, leur comportement morphosyntaxique est aussi varié (des modifieurs peuvent apparaître entre le verbe et le nom, la distance entre les éléments collocationnels est parfois longue) (Todirașcu *et al.*, 2008). Ces cas de figure seront illustrés ci-dessous par des exemples.

Nous avons identifié plusieurs cas de traduction d'une collocation verbo-nominale dans notre corpus :

a) collocation verbo-nominale directement équivalente (*prendre des mesures* vs. *a lua măsuri*) ;

b) collocation verbo-nominale synonyme (*prendre des mesures* vs. *a adopta măsuri* 'adopter des mesures') ;

c) une seule unité lexicale (*avoir l'intention* vs. *a intenționa* 'intentionner') ;

d) collocation nominalisée (*remplir sa mission* vs. *îndeplinirea misiunii*) ;

e) paraphrase (*prendre effet* vs. *a fi valabil* 'être valable').

De plus, des modifieurs (adverbes, adjectifs) ou des groupes prépositionnels peuvent être présents entre le verbe et le nom. Dans ces cas, la distance entre les éléments collocationnels est parfois longue comme dans les exemples (1) et (2) ci-dessous concernant la collocation *prendre les dispositions* :

1) *Il **prend** également, après consultation de cet État, **les dispositions** prévues auxdits paragraphes*. (FR)

2) *El **ia**, de asemenea, **măsurile** prevăzute la alineatele menționate, după consultarea acestui stat*. (RO)

Dans (1) un adverbe et un groupe prépositionnel apparaît entre le verbe et le nom, tandis qu'en (2) le groupe prépositionnel se trouve à une distance plus longue, après le nom.

D'autres contextes montrent des collocations verbo-nominales étendues comme dans les exemples (3) et (4) suivants :

3) *[…] les Etats membres […] ne **maintiennent** aucune **mesure** […]* (FR)

4) *[…] statele membre nu **menţin în vigoare măsuri** […]* (RO)

Dans ces cas, le verbe *maintenir* du français est traduit en roumain par une collocation verbo-nominale *a menţine în vigoare* 'maintenir en vigueur' suivi par son co-occurrent *măsuri* 'mesures'.

Il existe aussi des cas où les verbes des collocations se trouvent en rapport de coordination étant suivis d'un seul nom (*ne maintenir ou instituer aucune mesure vs. a nu menţine sau a nu institui nicio măsură*).

Les différents problèmes discutés ci-dessus doivent être pris en considération pour l'élaboration de l'algorithme d'alignement des collocations verbo-nominales dans des corpus bilingues parallèles français - roumain. Ce genre de difficultés où la traduction des collocations ou des termes poly-lexicaux se réalise par des constructions structurellement non-isomorphiques (par exemple, une collocation verbo-nominale qui est l'équivalente d'une construction nominalisée), peut être traité au moyen des règles de correspondance (voir le

projet *TTC*[130]). De notre côté, nous avons intégré dans le processus d'alignement un dictionnaire de collocations du type *Verbe + Nom*, munies aussi de leurs propriétés morphosyntaxiques contextuelles (Todirașcu *et al.*, 2008). En outre, afin de couvrir partiellement les cas de traduction non-isomorphique, nous avons enrichi le dictionnaire initial par des collocations nominales du type *Nom 1 déverbal + (préposition) + Nom 2*. Pour ce faire, nous avons développé une méthode supplémentaire qui consiste dans la génération automatique de cette classe de collocations à partir des entrées du dictionnaire.

Dans la sous-section suivante, nous allons décrire le dictionnaire de collocations (Todirașcu *et al.*, 2008) exploité dans notre approche. Tout d'abord, les propriétés des collocations verbo-nominales (Todirașcu *et al.*, 2008) seront spécifiées dans la sous-section 4.4.2.1. Puis, la structure du dictionnaire sera décrite dans la sous-section 4.4.2.2. Ensuite, la méthode d'extraction automatique de collocations utilisée par Todirașcu *et al.* (2008) pour alimenter le dictionnaire sera présentée dans la sous-section 4.4.2.3. Enfin, la classe de collocations nominales du type *Nom 1 déverbal + (préposition) + Nom 2*, ainsi que la méthode de génération automatique mise en place feront l'objet de la sous-section 4.4.2.4.

4.4.2. Le dictionnaire de collocations français - roumain

Le dictionnaire de collocations utilisé (Todirașcu *et al.*, 2008) contient des entrées trilingues français - roumain - allemand, leurs équivalents de traduction et leurs propriétés. Ce dictionnaire est construit à partir du corpus multilingue parallèle juridique et administratif *JRC-Acquis* (Steinberger *et al.*, 2006) qui est aussi lemmatisé et étiqueté. Comme nous l'avons déjà mentionné, il est limité à la classe des collocations verbo-nominales. Pour notre projet de thèse, nous avons exploité seulement la partie français - roumain de ce dictionnaire. Comme celui-ci contient aussi les propriétés des collocations *Verbe + Nom*, nous allons présenter ces propriétés telles que décrites par Todirașcu *et al.* (2008), dans la sous-section suivante.

4.4.2.1. Les propriétés des collocations verbo-nominales

À partir de l'analyse linguistique du corpus exploité, Todirașcu *et al.* (2008) concluent que les collocations ont un comportement syntaxique fortement dépendant du contexte. Cette étude

[130] http://www.ttc-project.eu/

fournit un ensemble de propriétés morphosyntaxiques caractérisant le comportement des collocations verbo-nominales pour chaque langue étudiée. Ces propriétés ainsi que le critère de la fréquence sont utilisés afin d'extraire automatiquement ces collocations à partir du corpus pour alimenter le dictionnaire.

Du point de vue sémantique, les collocations verbo-nominales sont analysées dans le cadre d'une approche contextualiste et notamment la grammaire systémique et fonctionnelle (Halliday, 1985 ; Banks, 2005 ; Gledhill, 2007). Selon cette approche, un verbe dans ses différents contextes exprime un des trois procès majeurs suivants (Banks, 2005) :

1) *Matériel ;* « Les procès matériels sont les procès qui expriment des changements dans le monde physique » (Banks, 2005 : 38), comme par exemple « *marcher* », « *donner* ».

2) *Mental ;* « Les procès mentaux concernent les événements qui ont lieu au niveau cérébral » (Banks, 2005 : 38), comme les procès cognitifs (« *penser* », « *croire* »), de perception (« *voir* », « *entendre* ») et affectifs (« *aimer* », « *détester* »).

3) *Relationnel ;* « Les procès relationnels expriment les relations entre deux entités, ou entre une entité et une propriété, sans événement ni changement physique » (Banks, 2005 : 38). Ce type de procès exprime des états (« *être* », « *avoir* »). Les procès exprimant la création d'un état (les procès de devenir) font partie également de cette catégorie.

Selon Gledhill (2007), le seul critère qui fait la différence entre une cooccurrence et une construction verbo-nominale est le rôle sémantique joué par le nom. Ainsi, dans une cooccurrence *Verbe + Nom* (par exemple, *faire une maison*) le complément exprime le rôle sémantique de « *cible* » ou « *effectué* » / « *affecté* », alors que dans une construction *Verbe + Nom* (par exemple, *faire une observation*), le nom exprime « *la portée* », notion reprise de Halliday (1985). « La portée est le rôle sémantique exprimé par tout élément qui désigne ou délimite le procès mais qui n'est pas le prédicateur » (Gledhill, 2007, *cit. in* Gledhill et Todirașcu, 2008 : 142). Le Complément de portée n'est pas modifié ou qualifié par le procès mais il contribue à l'expression du procès (Gledhill, 2007).

Todirașcu *et al.* (2008) identifient deux classes de collocations verbo-nominales :

1) *prédicateurs complexes* qui présentent des propriétés morphosyntaxiques fixes (par exemple, *tenir compte*, *faire l'objet*) ; Dans ces structures, le nom figure systématiquement sans déterminant ou montre une préférence accentuée pour le déterminant défini. De plus, aucun modifieur (adverbe, adjectif, proposition relative) ne peut apparaître. Également, le verbe montre une forte préférence pour certaines propriétés (par exemple, la voix active). Parfois, une préposition spécifique peut faire partie d'une collocation (*mettre **en** application*). Du point de vue sémantique, le nom exprime *la portée* (Gledhill, 2007). Celui-ci peut être un *nom de portée* (*tenir compte*) ou un *Adverbial de portée* (*mettre en application*).

2) *prédicats complexes* qui présentent un comportement morphosyntaxique variable : le nombre du nom peut varier entre le singulier et le pluriel, les déterminants (définis, indéfinis, démonstratifs, possessifs) ainsi que les modifieurs peuvent apparaître. Dans ce cas, le nom est un *Complément de portée* (par exemple, *prendre des mesures*).

Conformément à cette étude linguistique, Todirașcu *et al.* (2008) proposent une méthode hybride d'extraction automatique de collocations et de leurs propriétés morphosyntaxiques à partir du corpus. Cette méthode combine des techniques statistiques et des filtres linguistiques. Ainsi, des filtres qui exploitent les étiquettes morphosyntaxiques et les lemmes des unités lexicales sont définis afin de modéliser le comportement morphosyntaxique des collocations en contexte. Ces filtres sont appliqués ensuite aux contextes des candidats collocationnels obtenus par le biais des techniques statistiques (cf. sous-section 4.4.2.3.), afin de repérer automatiquement les candidats considérés pertinents. Après l'extraction automatique, les candidats obtenus sont toutefois classés manuellement, en fonction du critère sémantique illustré auparavant, afin d'alimenter le dictionnaire.

Dans la sous-section suivante, sera présentée la structure du dictionnaire utilisé, tandis que la méthode d'extraction automatique de collocations appliquée pour alimenter le dictionnaire (Todirașcu *et al.*, 2008) sera détaillée dans la sous-section d'après.

4.4.2.2. La structure du dictionnaire

Le dictionnaire de collocations verbo-nominales (Todirașcu *et al.*, 2008) contient une liste de 250 entrées trilingues (français - roumain - allemand).

La Figure 62 suivante comprend un exemple d'entrée pour le français.

```
<entry id= "1">
   <te  lang= "fr">
          <complexitem>
<construction>prendre+en+compte</construction>
<v_spec><lemma>prendre</lemma>
<voice freq=100>active</voice>
</v_spec>
<prep>en</prep>
<n_spec><lemma>compte</lemma>
  <det freq="100">null</det>
  <nb freq="100">sg</nb>
</n_spec>
<c_spec>
<colloc_spec>
<required_args case="acc"> object </required_args>
<colloc_type> compl_predicator </colloc_type>
</colloc_spec>

<colloc_documentation>
<colloc_LL value="2999.854" corpus="ACQ"/>
<examples>
<example> ... </example>
</examples>
</colloc_documentation>
</c_spec>
          </complexitem>
</te>
</entry>
```

Figure 62. Un exemple d'entrée du dictionnaire pour le français (Todirașcu *et al.*, 2008)

Chaque entrée contient trois sections pour chaque langue représentées par l'élément <te> avec l'attribut « lang ». Chaque équivalent de traduction comprend un élément <complexitem> ou un autre dénommé <simpleitem> pour représenter les cas ou les collocations sont traduites par une collocation ou une seule unité lexicale.

Le verbe et ses propriétés spécifiques (temps, mode, voix, nombre, personne) appartenant à une collocation sont représentés par l'élément <v_spec>, tandis que le nom et ses propriétés spécifiques (nombre, genre, cas) sont représentés par <n_spec>. Ensuite, les propriétés spécifiques de la collocation (les arguments requis, le type de la collocation - prédicat complexe ou prédicateur complexe) sont marquées par l'élément <colloc_spec>. La fréquence des propriétés du verbe ou du nom dans un corpus donné est également représentée par l'attribut « freq ».

Pour l'exemple donné dans la Figure 62 (*prendre en compte*), le nom présente toujours le déterminant zéro (« null ») représenté par l'élément <det> et l'attribut « freq » qui a la valeur 100, ce qui veut dire que tous les contextes partagent cette propriété. Le nombre singulier <nb> a aussi la fréquence 100, donc il est systématiquement utilisé. L'information stockée

dans l'élément <required_args> montre que le complément d'objet direct est en accusative (« case »), alors que l'élément <colloc_type> marque le type de la collocation de prédicateur complexe. La collocation apparaît toujours à la voix active, donc la fréquence, dans ce cas, est égale à 100.

Comme nous venons de le voir, les entrées du dictionnaire ne sont pas une simple liste de collocations et de leurs équivalents de traduction, mais une liste exhaustive contenant également leurs propriétés morphosyntaxiques contextuelles et décrivant le comportement syntaxique de ces collocations. Ces informations linguistiques associées à chaque entrée sont aussi exploitées dans le processus d'alignement lexical pour étudier l'influence du dictionnaire sur les résultats d'alignement.

À partir des entrées du dictionnaire, nous avons saisi la possibilité de générer automatiquement les collocations nominales équivalentes sémantiquement de type *Nom 1 déverbal + (préposition) + Nom 2* (p. ex. *mettre en application* vs. *mise en application* (FR) ; *pune în aplicare* vs. *punere în aplicare* (RO)). En effet, ces collocations nominales se forment par la conversion des verbo-nominales initiales. Le verbe (***prendre*** *des mesures*) devient un nom déverbal (***prise*** *de mesures*) par transposition (le changement de catégorie grammaticale). Les deux structures sont donc équivalentes du point de vue sémantique. Comme la transposition est un procédé très usité en traduction humaine, il est aussi fréquemment employé dans le corpus parallèle juridique étudié. De ce fait, nous avons considéré utile cette démarche de tenter d'obtenir automatiquement un ensemble supplémentaire de collocations nominales à partir du dictionnaire et l'enrichir aussi par ces nouvelles données.

Pour ce faire, nous avons constaté au départ que la transformation du verbe dans le nom déverbal peut se réaliser de manière automatique, en exploitant des indices morphologiques des déverbaux roumains (les terminaisons en *-RE*) (cf. sous-section 4.4.2.4.) et des ressources externes comme le lexique de déverbaux français *VerbAction*[131] (Hathout *et al.*, 2002 ; Tanguy et Hathout, 2002) (cf. sous-section 4.4.2.4.). Notons que pour le roumain il n'existe pas de telles ressources.

[131] Ce lexique sera décrit plus loin dans la sous-section 4.4.2.4.

En même temps, pour la génération automatique des collocations nominales complètes, nous avons eu besoin de connaître les propriétés morphosyntaxiques de ces constructions afin de voir si elles héritent les propriétés de leurs correspondants verbo-nominaux ou différentes variations peuvent apparaître.

Ainsi, pour étudier les propriétés contextuelles de ces collocations nominales et leur comportement morphosyntaxique par rapport à leurs correspondants d'origine, nous avons appliqué, dans un premier temps, l'approche contextualiste d'extraction automatique de collocations proposée par Todirașcu *et al.* (2008). Cette méthode nous a permis tout d'abord l'obtention d'un ensemble de collocations de type *Nom 1 déverbal + (préposition) + Nom 2*, à partir du corpus, comme nous le décrirons dans la sous-section 4.4.2.4. Ensuite, nous avons étudié en contexte le comportement morphosyntaxique de ces collocations, nous avons relevé leurs propriétés et défini les filtres morphosyntaxiques correspondants, comme il est proposé dans Todirașcu *et al*. (2008). Puis, nous avons aussi effectué une comparaison des collocations *Nom 1 déverbal + (préposition) + Nom 2* avec les verbo-nominales d'origine pour vérifier si leurs propriétés sont héritées lors de la conversion ou des différences morphosyntaxiques peuvent se révéler. Enfin, nous avons généré automatiquement les constructions nominales en nous basant sur l'étude linguistique contextuelle effectuée (cf. sous-section 4.4.2.4.).

Dans la sous-section suivante, sera ainsi présentée la méthode d'extraction automatique de collocations de Todirașcu *et al.* (2008). Ensuite, dans la sous-section 4.4.2.4. sera décrite la méthode que nous avons mise en place pour générer automatiquement les collocations de type *Nom 1 déverbal + (préposition) + Nom 2* à partir des entrées du dictionnaire (Todirașcu *et al.*, 2008).

4.4.2.3. La méthode d'extraction automatique des collocations

Afin de repérer automatiquement les candidats collocationnels à partir du corpus, Todirașcu *et al.* (2008) développent une méthode hybride combinant des techniques statistiques et des filtres linguistiques. Cette méthode est valable pour toutes les paires collocationelles (*Verbe + Nom ; Nom + Nom, Nom + Adjectif*, etc.). Celle-ci a été mise en œuvre et ensuite appliquée

par Todirașcu *et al*. (2008) pour alimenter le dictionnaire des collocations verbo-nominales. Cette approche nécessite des corpus monolingues étiquetés et lemmatisés[132].

La méthode développée suit trois étapes principales :

1) l'extraction statistique des candidats collocationnels à partir de chaque corpus monolingue utilisé ;

2) la sélection des candidats considérés comme pertinents par l'application des filtres morphosyntaxiques ;

3) la classification manuelle des candidats au niveau sémantique afin d'alimenter le dictionnaire par les collocations retenues.

La première étape d'extraction automatique de collocations consiste dans l'application du module statistique développé (Todirașcu *et al.*, 2008) sur chaque corpus monolingue exploité. Ce module cherche toutes les paires collocationnelles (p. ex. *Verbe + Nom ; Nom + Nom, Nom + Adjectif*, etc.) à l'intérieur d'une fenêtre de 11. Les scores statistiques utilisés afin de détecter des collocations selon le critère statistique sont la déviation des distances (Smadja et McKeown, 1990) et le *Log-Likelihood* (LL) (Dunning, 1993). Ces deux scores statistiques seront expliqués ci-dessous :

1) ***la moyenne et la dispersion*** (Smadja et McKeown, 1990). La dispersion mesure la déviation des distances par rapport à la moyenne, pour deux mots du corpus *w1* et *w2*.

La dispersion représente le carré de la déviation standard :

$$\sigma^2 = \frac{\sum_{i=1}^{n} (d_i - \mu)^2}{n-1}$$

Où :

n : le nombre d'apparitions de la paire de mots *w1 - w2* ;
di : la distance calculée entre les mots *w1* et *w2*, pour l'occurrence *i* de la paire *w1 - w2* ;
μ : la moyenne arithmétique des distances *di.*

[132] Le corpus utilisé par Todirașcu *et al.* (2008) est *JRC-Acquis* (Steinberger *et al*., 2006), étiqueté et lemmatisé pour l'allemand et le français par *TreeTagger* (Schmid,1994) et pour le roumain par *TTL* (Ion, 2007).

L'interprétation de ces mesures est la suivante :

a) si la dispersion est limitée, la moyenne montre la distance usuelle entre deux mots du corpus ;

b) dans le cas où la distance entre deux mots du corpus est constante, la dispersion est 0 ;

c) si les distances calculées entre deux mots du corpus ont une répartition aléatoire, la dispersion a des valeurs importantes.

Dans l'approche de Todirașcu *et al.* (2008), les candidats retenus sont toutes les paires collocationnelles (p. ex. *Verbe + Nom ; Nom + Nom,* etc.*)* ayant une déviation standard inférieure à 1,5 (Manning et Schütze, 1999).

2) ***le log-likelihood*** (Dunning, 1993). Ce test utilise un rapport de probabilité entre deux hypothèses H1 et H2 décrites ci-dessous :

H1 : il n'existe pas de rapport entre les mots, ils sont donc indépendants :

$$\boxed{P\langle w_2 \mid w_1 \rangle = p = P\langle w_2 \mid -w_1 \rangle}.$$

Dans ce cas, la probabilité que le mot 2 (w2) et que le mot 1 (w1) se retrouvent ensemble est égale à la probabilité que le mot 2 apparaisse dans un contexte qui ne contient pas le mot 1.

H2 : il existe un rapport de dépendance entre les 2 mots :

$$\boxed{P\langle w_2 \mid w_1 \rangle = p_1 \neq p_2 = P\langle w_2 \mid -w_1 \rangle}$$

Dans ce cas, la probabilité que le mot 2 (w2) et que le mot 1 (w1) apparaissent ensemble est différente de la probabilité que le mot 2 apparaisse indépendamment du mot 1.

L(H) représente la vraisemblance d'une hypothèse, soit L(H1) la vraisemblance que l'hypothèse 1 soit vérifiée et L(H2) que l'hypothèse 2 le soit. La méthode calcule ensuite le logarithme du rapport des vraisemblances L (H1) et L (H2) :

$$\boxed{\log \lambda = \log \frac{L(H1)}{L(H2)}}$$

Soit le tableau de contingence suivant :

	Mot 2 présent	**Mot 2 absent**
Mot 1 présent	**a**	**b**
Mot 1 absent	**c**	**d**

Où :

a : le nombre de fois où le mot 2 est présent lorsque le mot 1 l'est aussi ;
b : le nombre de fois où le mot 2 est absent dans un contexte où le mot 1 est présent ;
c : le nombre de fois où le mot 2 est présent dans un contexte où le mot 1 est absent ;
d : le nombre de fois où le mot 2 est absent lorsque le mot 1 l'est aussi.

Dans ce cas-là, la vraisemblance des deux hypothèses suit la loi binomiale B(k,n,p) (la probabilité d'obtenir *k* succès dans un ensemble de *n* mots et avec une probabilité de réussite de chaque essais égale à *p*) et on a :

L(H1)=B(a,a+b,p).B(c,c+d,p)

L(H2)=B(a,a+b,p1).B(c,c+d,p2)

$p = \frac{a+c}{N}$, $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, avec N le nombre total de mots du document.

La loi binomiale traduit bien la vraisemblance des hypothèses car, par définition, elle traduit, pour B(a,a+b,p), par exemple, la probabilité d'obtenir *a* fois le mot 2 parmi *a+b* mots 1 présents et avec une probabilité de succès de chaque essai égale à *p* (le même raisonnement doit être appliqué aux autres lois binomiales).

En posant LL=-2.LOG(λ), où LL représente le score obtenu, il est clair de constater que plus LL sera élevé et plus la vraisemblance de l'hypothèse H2 sera vérifiée. La méthode pourra donc conclure que les deux mots forment une collocation si le score dépasse un certain seuil fixé.

Dans l'approche de Todirașcu *et al.* (2008), sont sélectionnées toutes les paires collocationnelles (p. ex. *Verbe + Nom ; Nom + Nom*, etc.) dont le score LL est supérieur à 9.

La deuxième étape du système d'extraction automatique de collocations de Todirașcu *et al.* (2008) consiste dans la sélection des candidats en fonction des filtres morphosyntaxiques définis, selon une étude contrastive des propriétés des collocations.

Enfin, les candidats collocationnels obtenus auparavant sont classifiés manuellement au niveau sémantique, le critère de la fréquence et le critère morphosyntaxique n'étant pas suffisants pour distinguer entre les collocations (p. ex. *prendre des mesures*) et les simples cooccurrences (p. ex. *prendre des échantillons*).

À l'issue de cette dernière étape, sont gardées manuellement seulement les collocations verbo-nominales et sont éloignées les simples cooccurrences en fonction du critère sémantique. Par conséquent, le système d'extraction automatique de collocations n'est pas utilisé pour alimenter le dictionnaire directement, car la classification manuelle reste impérieusement une étape intermédiaire.

C'est pour la même raison que, dans notre projet de thèse, nous n'avons pas appliqué la méthode d'extraction automatique de collocations de Todiraşcu *et al.* (2008) directement dans le processus d'alignement. Mais nous avons intégré le dictionnaire dans ce processus, en tant que ressource externe, afin d'étudier son influence sur les résultats d'alignement lexical finaux.

De plus, nous avons enrichi le dictionnaire de collocations verbo-nominales français - roumain initial (Todiraşcu *et al.*, 2008) avec les collocations nominales correspondantes, obtenues par conversion : *prendre des décisions* > *prise de décisions* ; *mettre en application* > *mise en application*. Pour identifier les propriétés de ces constructions nominales, nous avons suivi la méthodologie proposée dans Todiraşcu *et al.* (2008). Ensuite, nous avons obtenu ces collocations par génération automatique à partir des entrées du dictionnaire. Dans la sous-section suivante, sera décrite cette méthode mise en œuvre par nos soins.

4.4.2.4. L'enrichissement du dictionnaire par des collocations nominales

Dans notre corpus, les collocations verbo-nominales sont souvent traduites par les collocations nominales correspondantes (*Nom 1 déverbal + (préposition) + Nom 2*) ou vice-versa, comme dans les exemples (1) et (2) ci-dessous. Il s'agit alors du procédé de traduction de *transposition*.

1) *Les États membres communiquent […] le nom des services qu'ils ont désignés pour* ***effectuer ces contrôles****.* (FR)

Statele membre comunică […] numele serviciilor pe care le-au desemnat pentru ***efectuarea acestor controale****.* (RO)

2) *L'employeur ne peut être contraint* ***au paiement de cotisations*** *majorées.* (FR)

Un angajator nu poate fi constrâns ***să plătească cotizații*** *majorate.* (RO)

Comme nous l'avons déjà précisé auparavant, afin d'étudier les propriétés contextuelles de ces collocations nominales et voir également si celles-ci gardent les propriétés des collocations d'origine ou d'éventuelles variations peuvent intervenir, nous avons appliqué la méthode d'extraction automatique de collocations contextualiste proposée par Todirașcu *et al.* (2008), pour les collocations nominales de type *Nom 1 déverbal + (préposition) + Nom 2*.

Ainsi, les cooccurrences *Nom + Nom* ont été extraites statistiquement (Todirașcu *et al*., 2008) (cf. sous-section antérieure) à partir du corpus *JRC-Acquis* (Steinberger *et al.*, 2006) en français et en roumain. Ensuite, nous avons étudié en contexte les 300 premiers candidats de type *Nom 1 déverbal + (préposition) + Nom 2* afin de relever leurs propriétés morphosyntaxiques et définir aussi les filtres linguistiques correspondants. Un exemple de ces candidats et leurs contextes respectifs en roumain figure dans l'Annexe 5. Comme nous avons remarqué que leurs propriétés ne sont pas discriminantes, comme il est aussi valable dans le cas des collocations verbo-nominales d'origine (Todirașcu *et al*., 2008), nous avons classifié manuellement les candidats obtenus en fonction du rôle sémantique de *portée* joué par le deuxième nom sur le plan de la collocation verbo-nominale (Todirașcu *et al.*, 2008). En effet, nous avons vérifié en contexte si sur le plan de la nominalisation du verbe, le deuxième nom apparaît en tant que projection de Complément de portée (*prendre des mesures > prise de mesures*) ou d'Adverbial de portée (*mettre en application > mise en application*) (cf. sous-section 4.4.2.1.).

Toutes les autres catégories présentant des projections d'autres compléments ou du sujet ont été éloignées manuellement :

a) Complément d'objet affecté (*prendre des échantillons > prise d'échantillons*) ou effectué (*instituer le lieu > institution du lieu*) ;

b) Complément indirect (*publier au journal > la publication au journal*) ;

c) Circonstant (*réagir au feu > la réaction au feu*) ;

d) Sujet (*le conseil décide > la décision du conseil*).

À l'issue de cette étude, nous avons vu que les noms déverbaux sont compatibles avec les marques flexionnelles du nom et le déterminant, tandis que les *Noms 2* montrent des préférences morphosyntaxiques (un nombre, un cas, une préposition, un certain déterminant). Des modifieurs (adjectifs qualificatifs ou groupes prépositionnels) peuvent aussi apparaître au niveau du *Nom 2*.

Dans le Tableau 26 suivant figurent les propriétés morphosyntaxiques repérées du *Nom 2* pour le français et le roumain.

Tableau 26. Propriétés morphosyntaxiques du *Nom 2* pour le français et le roumain

Propriétés morphosyntaxiques du *Nom 2*	**Valeurs pour le français**	**Valeurs pour le roumain**
Nombre	singulier, pluriel	singulier, pluriel
Cas	-	accusatif, génitif
Préposition	*de, en, à, sous*	*Ø, de, în, la, sub*
Déterminants (articles)	Ø, défini, indéfini, démonstratif, possessif, fusion préposition / article défini	Ø, défini, indéfini, génitival
Qualification	adjectif, groupe prépositionnel	adjectif, génitif, groupe prépositionnel

Ø : l'élément zéro

En fonction des propriétés morphosyntaxiques contextuelles des collocations nominales relevées, nous avons aussi défini des filtres linguistiques pour les deux langues étudiées, suivant la méthodologie proposée dans Todirașcu *et al.* (2008) pour les collocations verbo-nominales. Ces filtres identifiant les collocations nominales à partir du corpus figurent dans le Tableau 27 suivant. L'explication des symboles qui apparaissent dans l'écriture des filtres définis figure dans l'Annexe 6.

Tableau 27. Filtres morphosyntaxiques identifiant des collocations nominales de type *Nom 1 déverbal + (préposition) + Nom 2* pour le français et le roumain

Filtres morphosyntaxiques FR	**Filtres morphosyntaxiques RO**
Nom1 déverbal - préposition *de* - **Nom2** [déterminant défini | zéro, singulier | pluriel] - (préposition *de* / Projection du Sujet | Projection du Complément indirect)	**Nom1 déverbal** [forme en *-RE*, féminin] - préposition *de* - **Nom2** [déterminant zéro, singulier | pluriel, accusatif] - (préposition *de către* / Projection du Sujet)
-	**Nom1 déverbal** [forme en *-RE*, féminin] - (modifieur) - **Nom2** [déterminant défini | indéfini, singulier | pluriel, génitif]
-	**Nom1 déverbal** [forme en *-RE*, féminin, déterminant zéro] - *a* [déterminant génital féminin, singulier] - **Nom2** [déterminant défini | indéfini, singulier | pluriel, génitif]
Nom1 déverbal - (modifieur) - préposition *en, à, sous* - (modifieur) - **Nom2** [déterminant zéro, singulier] - (préposition *de* / Projection du Complément direct | Sujet)	**Nom1 déverbal** [forme en *-RE*, féminin] - préposition *în, la, sub* - **Nom2** [déterminant zéro, singulier, accusatif] (déterminant génital *a* / Projection du Complément direct)

Par la comparaison des collocations nominales avec les verbo-nominales d'origine, nous avons remarqué que les collocations nominales présentant la projection d'un Adverbial de portée (*mettre en application > mise en application ; a pune în aplicare > punere în aplicare*) gardent les propriétés des collocations verbo-nominales d'origine. En revanche, pour l'autre catégorie identifiée et notamment les collocations nominales présentant la projection d'un Complément de portée (*prendre des mesures > prise de mesures ; a lua măsuri > luare de măsuri, luarea măsurilor, luare a măsurilor*) des variations morphosyntaxiques apparaissent. Nous allons décrire ces variations morphosyntaxiques, plus loin, à l'aide des exemples concrets. Ces constats nous ont permis de générer les collocations de type *Nom 1 déverbal + (préposition) + Nom 2* de manière automatique à partir du dictionnaire (Todirașcu *et al.*, 2008), comme il sera décrit ci-dessous.

Dans les deux langues étudiées, la conversion des collocations verbo-nominales se réalise par la nominalisation du verbe (***mettre** en valeur > **mise** en valeur* vs. ***a pune** în valoare > **punere** în valoare*). En roumain, le déverbal est généralement une nominalisation à base d'infinitif ayant la forme en *-RE* (*GLR*[133], 2005) et le genre féminin : *lua + re > luare* 'prise' ; *îndeplini*

[133] *Gramatica limbii române* 'Grammaire de la langue roumaine'

+ re > îndeplinire ‘accomplissement’. Cet indice morphologique peut être exploité pour identifier automatiquement les déverbaux du roumain. Par contre, en français, les déverbaux présentent plusieurs terminaisons (*prise, accomplissement, repérage, information*) et l’indice morphologique n’est plus suffisant pour les repérer automatiquement. De ce fait, des ressources externes comme les lexiques de déverbaux sont nécessaires. Nous avons utilisé ainsi le lexique *VerbAction* (Hathout *et al*., 2002 ; Tanguy et Hathout, 2002) qui sera décrit plus loin.

En ce qui concerne les collocations nominales présentant la projection d’un Adverbial de portée, celles-ci gardent la préposition des collocations verbo-nominales initiales *en / à / sous* (FR) vs. *în / la / sub* (RO), pendant que le *Nom 2* reste invariable (déterminant zéro, nombre singulier), comme dans les exemples suivants :

1) *L’acceptation et **la prise en charge** des produits livrés est subordonnée à la vérification.…* (FR)

2) *… y compris l’objectif de **la mise à disposition** d’informations.…* (FR)

3) *À partir de **la mise sous contrôle** et au plus tôt le 16 octobre de la campagne de commercialisation en cause,…* (FR)

4) *… pentru a facilita **punerea în aplicare** a dispoziţiilor preconizate,…* (RO)

   ‘... pour faciliter la mise en application des dispositions préconisées …’

5) *Pentru **aducerea la îndeplinire** a prevederilor articolelor 2, 3 şi 4.…* (RO)

   ‘Pour l’accomplissement des prévoyances des articles 2, 3 et 4…’

6) *… **punerea** efectivului **sub control** oficial,…* (RO)

   ‘… la mise de l’effectif sous contrôle officiel…’

Quant aux collocations nominales présentant la projection d’un Complément de portée, le *Nom 2* est précédé par la préposition *de* et le déterminant zéro (*prendre des mesures > la prise de mesures* (FR) vs. *a lua măsuri > luare de măsuri* (RO)). Toutefois, la préposition *de* peut manquer en roumain : *a lua măsuri* vs. *luarea măsurilor*. Dans ce cas, le deuxième nom est

toujours en génitif (*măsurilor*). Dans d'autres cas, un déterminant génitival *a* (forme de féminin et au singulier) peut apparaître en roumain : *luare a măsurilor*.

Au vu de ces considérations, nous avons obtenu les collocations nominales présentant la projection d'un Complément de portée du roumain par génération automatique à partir des collocations verbo-nominales fournies par le dictionnaire, comme dans l'exemple : *a lua decizii* 'prendre de décisions' > *(lua* + ***re)*** + ***de*** + *decizii ; (lua* + *re**a)** decizii**lor** ; (lua* + ***re)*** + ***a*** + *decizii**lor*** 'prise de décisions'. Quant aux collocations présentant la projection d'un Adverbial de portée, le déverbal est suivi par la préposition correspondante (*pune **în** aplicare* 'mettre en application' > ***pune+re în** aplicare* 'mise en application').

En ce qui concerne le français, nous avons généré automatiquement les déverbaux à partir des verbes des collocations verbo-nominales fournies par le dictionnaire à l'aide d'un lexique de déverbaux français *VerbAction* (Hathout *et al*., 2002 ; Tanguy et Hathout, 2002). Ce lexique est composé d'un ensemble de 9 393 couples de verbes et leurs noms déverbaux équivalents qui remplissent deux conditions : le nom est morphologiquement lié au verbe et le nom dénomme l'action ou l'activité exprimée par le verbe. *VerbAction* est disponible en format *XML/TEI*. Un exemple de couple formé par un verbe et son nom correspondant extrait du lexique *VerbAction* est donné dans la Figure 63 suivante :

```
<couple>
    <verb>
      <lemma>informer</lemma>
      <tag>Vmn----</tag>
    </verb>
    <noun gender="feminine" number="singular">
      <lemma>information</lemma>
      <tag>Ncfs</tag>
    </noun>
</couple>
```

Vmn : Verbe principal, infinitif ; *Ncfs* : Nom commun, féminin, singulier

Figure 63. Un exemple de couple verbe - nom déverbal extrait du lexique *VerbAction* (Hathout *et al*., 2002 ; Tanguy et Hathout, 2002)

Nous avons comparé automatiquement le lexique *VerbAction* aux entrées du dictionnaire afin de générer le déverbal français à partir du verbe des collocations *Verbe + Nom* trouvé dans le dictionnaire et, simultanément, dans le lexique. Le déverbal est suivi par la préposition *de* pour les collocations présentant la projection d'un Complément de portée (*prendre des mesures* > ***prise de** mesures*) et par la préposition correspondante dans le cas des collocations présentant la projection d'un Adverbial de portée (*mettre en application* > ***mise en** application*).

Certaines collocations verbo-nominales, précisons-le, ne peuvent pas être nominalisées (*donner suite* > *donation de suite (FR) / *a da curs* >*dare de curs (RO)). De ce fait, dans un dernier temps, nous avons validé les collocations nominales obtenues par comparaison automatique avec les corpus monolingues utilisés pour les deux langues de travail. Cette opération permet en même temps de collecter la totalité des formes attestées dans le corpus.

Les collocations ainsi obtenues ont été ajoutées au dictionnaire munies de leurs propriétés identifiées lors de notre étude. Une fois enrichi, celui-ci a été appliqué au corpus parallèle utilisé afin de corriger les erreurs d'alignement lexical au niveau des collocations et d'étudier également son impact sur les résultats du système d'alignement global. L'algorithme élaboré dans ce sens sera décrit dans la sous-section suivante.

4.4.3. L'alignement des collocations

Les erreurs d'alignement lexical repérées au niveau des collocations verbo-nominales sont dues, en grande partie, à leur variabilité syntaxique élevée. Ainsi, une simple liste de collocations *Verbe + Nom* et leurs équivalents de traduction n'est pas suffisante pour collecter les alignements multiples concernant ces collocations. Des unités lexicales additionnelles appartenant à une collocation telles que les déterminants, certains modifieurs ou les auxiliaires doivent être alignées en bloc avec les autres composants de celle-ci. De ce fait, les propriétés morphosyntaxiques représentées dans le dictionnaire sont aussi prises en compte afin de compléter l'alignement lexical au niveau de ces expressions poly-lexicales.

Tout d'abord, l'algorithme cherche les collocations (*Cfr*) et leurs équivalents de traduction (*Cro*) accompagnés de leurs propriétés morphosyntaxiques dans le dictionnaire (*DICOCOLLOC*). Puis, sont sélectionnées les phrases sources contenant les occurrences des collocations à partir du corpus parallèle d'entrée (*PARALLELCORPUS*). Ensuite, les

équivalents de traduction sont cherchés dans les phrases cibles correspondantes trouvées dans le corpus parallèle d'entrée. Une fois les positions du verbe et du nom identifiées, l'algorithme cherche les autres unités lexicales susceptibles de faire partie des collocations (auxiliaires, déterminants, modifieurs) dans les contextes du verbe et du nom. Si les catégories lexicales de ces unités correspondent à l'une des propriétés des collocations (par exemple, à un déterminant ou un modifieur) dans la phrase source et une catégorie compatible existe également dans la phrase cible, alors les positions de ces unités sont rajoutées aux autres alignements. La recherche s'effectue dans une fenêtre de 11 mots, cette distance étant considérée suffisante pour collecter les alignements au niveau des collocations étendues.

Dans la Figure 64 suivante figure l'algorithme élaboré afin d'aligner en bloc les unités lexicales appartenant aux collocations visées.

Input : collocation pairs (*Cfr*, *Cro*), in *DICOCOLLOC* and *PARALLELCORPUS*
Output : The list ALIGNMWE of aligned positions.

```
for each (Cfr,Cro) in DICOCOLLOC
        Cfr=(Vfr, Nfr);
        Cro=(Vro, Nro);
        Propfr = properties(Cfr, DICOCOLLOC);
        Propro=properties(Cro, DICOCOLLOC);
        for each (sentfr,sentro) in PARALLELCORPUS
                if (Cfr in words(sentfr)) and  (Cro in words(sentro))
                then
                        positionv=position(Vfr,sentfr);
                        positionn=position(Nfr, sentfr) ;

                        contextefr=contexte(positionv) and contexte(positionn)
                        for each w in contextfr
                        if(cat(w) in proprfr) then
                          positionw=position(w, Sentfr)
                        add(positionw, ALIGNFR)

positionv_ro=position(Vro, Sentro) ;
                        positionn_ro=position(Nro, Sentro)
                        contextero=contexte(positionv_ro) and contexte(positionn_ro)
                        for each wro in contextero
                        if(category(wro) in proprro) then
                          positionro=position(wro, sentro)
                        add(positionro, ALIGNRO)
                end
        end

for each p in ALIGNFR
        for each r in ALIGNRO
                add(p,r, ALIGNMWE);
end
end
end;
```

Word() extrait la liste de mots d'une phrase donnée ; *category* retourne la catégorie lexicale d'un mot donné ; *position* retourne la position d'un mot dans la phrase.

Figure 64. L'algorithme d'alignement des collocations *Verbe + Nom*

Nous avons appliqué l'algorithme décrit ci-dessus au corpus parallèle utilisé et avons évalué l'impact du dictionnaire sur les résultats finaux du système d'alignement lexical développé. Dans la sous-section suivante, sera présentée l'évaluation de l'alignement des collocations effectuée sur le corpus de test utilisé.

4.4.4. L'évaluation de l'alignement des collocations

L'alignement lexical obtenu à l'issue de l'étape d'alignement des collocations *Verbe + Nom* fournies par le dictionnaire utilisé (Todirașcu *et al*., 2008) a été évalué en termes de précision, rappel, f-mesure et *AER* par comparaison avec nos deux alignements de référence (*RefA*, sans alignement des collocations et *RefB*, avec alignement des collocations - cf. sous-section 4.1.1.). Rappelons qu'afin d'aligner les collocations *Verbe + Nom* dans le corpus de référence *RefB*, nous avons validé manuellement 470 paires d'équivalents de traduction français - roumain dont au moins un est une collocation *Verbe + Nom* dans la paire. Précisons que les conditions d'évaluation sont les mêmes que pour l'évaluation des étapes précédentes d'alignement lexical (alignements nuls ignorés, aucune distinction entre les alignements sûrs et possibles dans les alignements de référence) (cf. sous-section 4.1.1.). Le Tableau 28 ci-dessous comprend les valeurs des scores d'évaluation obtenues :

Tableau 28. Évaluation de l'alignement des collocations fournies par le dictionnaire

Systèmes	**Précision (%)**		**Rappel (%)**		**F-mesure (%)**		***AER* (%)**	
	RefA	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***	***RefA***	***RefB***
1 : **système de base**	95,56	95,51	52,91	49,71	68,11	65,39	31,89	34,61
2 : système de base + **cognats**	95,23	95,18	55,72	52,35	70,30	67,55	29,70	32,45
3 : système de base + cognats + **règles heuristiques**	92,55	92,21	63,15	60,91	75,07	73,36	24,93	26,64
4 : système de base + cognats + règles heuristiques + **collocations**	92,38	92,13	63,99	61,39	75,61	73,68	24,39	26,32

L'application du dictionnaire comme dernière étape du processus d'alignement lexical a amélioré le score *AER*, mais pas de manière significative. D'une part, le dictionnaire a diminué la précision du système de 0,17% pour *RefA* et de 0,08% pour *RefB* et, d'autre part, celui-ci a amélioré le rappel de 0,84% pour *RefA* et de 0,48% pour *RefB.* Cela a mené à l'amélioration du score *AER* de seulement 0,54% pour *RefA* et de 0,32% pour *RefB.*

Afin d'expliquer le résultat obtenu, nous avons cherché à vérifier combien de collocations existantes dans le corpus de test étaient alignées à l'aide du dictionnaire. Pour ce faire, nous avons comparé automatiquement la liste des collocations fournies par le dictionnaire avec la liste extraite manuellement à partir du corpus de test lors de la construction du corpus de référence contenant l'alignement des collocations (*RefB*). Ensuite, nous avons analysé manuellement les collocations verbo-nominales ayant comme équivalent de traduction une collocation nominalisée dans le corpus de test, pour voir également combien de ces collocations nominales existaient dans le corpus et, parmi celles-ci, combien étaient alignées par l'intermédiaire du dictionnaire. Le Tableau 29 suivant comprend les statistiques concernant ces questions.

Tableau 29. Statistiques concernant les collocations étudiées à partir du corpus de test et du dictionnaire

Ressources		**Collocations FR**	**Collocations RO**
Corpus de test	Équivalents collocationnels totaux	470	470
	VN par langue	339	347
	VN uniques par langue (< = > VN, N, Autre)	254	245
	VN< = > N	2	21
Dictionnaire		250	250
VN communes Corpus de test - Dictionnaire		18	15
N communes Corpus de test - Dictionnaire		-	2
VN supplémentaires Corpus de test		236	230

VN : collocation verbo-nominale ; N : collocation nominalisée ; **< = > : «** équivalent à ».

Parmi les 470 équivalents collocationnels bilingues existants dans le corpus de test, 339 représentent des collocations verbo-nominales françaises tandis que 347 sont des collocations roumaines. Apres l'élimination des occurrences répétées, le nombre total des collocations

verbo-nominales s'élève à 254 pour le français, respectivement 245 pour le roumain. Parmi ces ensembles, 21 collocations verbo-nominales françaises ont comme équivalent de traduction en roumain une collocation nominale, pendant que seulement 2 collocations nominales françaises sont traduites par une verbo-nominale roumaine.

La raison pour laquelle l'amélioration du score *AER* est assez faible réside dans le fait qu'il existe très peu de collocations communes entre le corpus de test utilisé et le dictionnaire de collocations : 18 pour le français et 15 pour le roumain. De plus, seulement 2 collocations nominales du roumain sont alignées avec leurs correspondants verbo-nominaux français par l'intermédiaire du dictionnaire. Concernant le français, aucune collocation nominale n'est alignée avec son homologue verbo-nominal roumain. De ces faits, des expériences d'alignement des collocations *Verbe + Nom* restent encore nécessaires sur d'autres corpus de test afin de rendre compte de l'impact du dictionnaire utilisé sur la qualité des résultats finaux de manière plus précise.

Comme pour notre corpus de test l'amélioration du score *AER* est assez faible, nous avons choisi de ne plus intégrer le dictionnaire dans le système d'alignement lexical final. Mais nous projetons dans l'avenir d'entreprendre les expériences d'alignement des collocations étudiées sur d'autres corpus de test, appartenant au même domaine juridique et administratif, dans le but de voir dans quelles conditions (corpus de test plus volumineux, taille du dictionnaire augmentée) l'amélioration plus significative du score *AER* devient possible. Pour ce faire, nous augmenterons le volume des corpus de test. De plus, nous enrichirons le dictionnaire avec l'ensemble des collocations verbo-nominales supplémentaires validées par nos soins dans le cadre de notre projet de thèse : 236 pour le français et 230 pour le roumain (cf. Tableau 29). Ces collocations sont accompagnées de leurs équivalents de traduction repérés dans le corpus parallèle et aussi de leurs propriétés morphosyntaxiques contextuelles. Celles-ci possèdent également les équivalents nominaux générés automatiquement selon la méthode décrite antérieurement, dans la sous-section 4.4.2.4.

Nous venons de décrire l'étape d'application du dictionnaire de collocations utilisé dans le processus d'alignement lexical. Notre conclusion est qu'une ressource externe telle que le dictionnaire peut améliorer l'alignement lexical final mais, pour notre corpus de test, cela ne s'est pas réalisé de manière considérable. De ce fait, d'autres études (l'utilisation de corpus de test de taille plus grande, l'amélioration du dictionnaire) restent encore à effectuer dans ce

sens. C'est pour cette raison que cette étape d'alignement lexical ne fait plus partie du système d'alignement lexical final. Ce système sera évalué dans la sous-section suivante.

4.5. L'évaluation globale du système d'alignement lexical

Cette évaluation est basée sur les quatre mesures suivantes : la précision, le rappel, la f-mesure et le score *AER*. Comme le système final ne comprend plus l'alignement des collocations, l'évaluation est réalisée en utilisant le corpus de référence *RefA*, c'est-à-dire le corpus ne contenant pas l'alignement des collocations étudiées.

Rappelons que les conditions d'évaluation prises en compte sont les suivantes :

1) tous les alignements nuls de l'alignement final et de celui de référence sont ignorés ;

2) il n'existe aucune distinction entre les alignements sûrs et possibles dans l'alignement de référence.

Le Tableau 30 ci-dessous illustre les valeurs des scores d'évaluation obtenues en ce qui concerne le système d'alignement lexical final :

Tableau 30. Évaluation globale du système d'alignement lexical final

Systèmes	**Précision (%)**	**Rappel (%)**	**F-mesure (%)**	***AER* (%)**
	RefA	***RefA***	***RefA***	***RefA***
1 : **système de base**	95,56	52,91	68,11	31,89
2 : système de base + **cognats**	95,23	55,72	70,30	29,70
3 : système de base + cognats + **règles heuristiques**	92,55	63,15	75,07	24,93

Le système de base construit par l'application de *GIZA++* (Och et Ney, 2003) dans les deux sens du processus d'alignement lexical et leur intersection (Koehn *et al.*, 2003) a obtenu une précision élevée de 95,56% mais un faible rappel de 52,91, ce qui a mené à un score *AER* élevé de 31,89%.

Ensuite, l'intégration du module d'identification et d'alignement automatique de cognats a déprécié la précision de seulement 0,33% mais a augmenté le rappel de 2,81%. Cela a conduit à une amélioration intéressante du score *AER* de 2,19%.

Finalement, l'application des règles heuristiques morphosyntaxiques et stylistiques a diminué la précision du système de 2,68% mais a augmenté, en même temps, de manière significative son rappel de 7,43%. Cette dernière étape d'alignement lexical a mené à une importante amélioration du score *AER* de 4,77% par rapport à l'étape précédente.

L'utilisation des informations linguistiques (lemmes, propriétés morphosyntaxiques) dans les modules d'alignement lexical développés basés sur les cognats et les heuristiques linguistiques a eu ainsi un impact positif sur les résultats du système d'alignement lexical final, en baissant le score *AER* de manière significative d'environ 7%, par rapport au système de base purement statistique. Une autre configuration qui pourrait être testée consiste dans l'application des règles heuristiques définies directement sur le système de base afin d'évaluer leur efficacité par rapport à un système statistique pur.

Cependant, les résultats obtenus dépendent de la taille du corpus de test utilisé, des annotations d'alignement effectuées dans le corpus de référence mais aussi des langues traitées. Des études similaires restent encore à effectuer sur d'autres corpus, de taille plus importante, pour pouvoir généraliser.

4.6. Discussion

Le système d'alignement lexical mis en place est un système hybride dépendant de la paire de langues incorporées, qui combine des techniques statistiques et des heuristiques linguistiques pour améliorer les résultats d'un système statistique pur.

Pour pouvoir rendre compte avec plus de précision des améliorations obtenues par l'incorporation des informations linguistiques dans le processus d'alignement, une comparaison avec les résultats d'autres systèmes français - roumain ou avec ceux fournis par d'autres méthodes dépendantes de la paire de langues et incluant le roumain, serait intéressante.

Cependant, une comparaison effective pourrait être faite si les systèmes utilisent les mêmes corpus de test et de référence. Ainsi, nous avons rencontré plusieurs difficultés par rapport à

cette démarche comparative. En effet, il n'existe pas, à notre connaissance, d'autres systèmes autonomes spécialement développés pour le français et le roumain. De plus, des corpus de référence construits pour cette paire de langues ne sont pas disponibles. De ce fait, nous avons constitué notre propre corpus de référence pour pouvoir évaluer les résultats du système d'alignement lexical développé. Ce corpus est relativement réduit (environ 30 000 tokens par langue) car le processus d'alignement manuel est une tâche coûteuse en temps et en ressources humaines.

Ainsi, nous avons pu faire une comparaison approximative des résultats obtenus avec ceux fournis par trois aligneurs disponibles incluant le roumain mais en combinaison avec l'anglais : *YAWA, MEBA* et *COWAL* (Tufiș *et al.*, 2006). Ces systèmes ont été évalués par leurs auteurs contre le corpus de référence roumain - anglais (*Gold Standard*) proposé dans le cadre de *Word Alignment Shared Tasks* (*HLT-NAACL 2003*)[134] (Mihalcea et Pederson, 2003).

YAWA fournit des alignements lexicaux par l'exploitation des techniques statistiques et des informations linguistiques telles que : les mots pleins, les *chunks* et les heuristiques. Quant à *MEBA*, celui-ci effectue des alignements itérativement en utilisant plusieurs traits comme : les cognats, les mots pleins, l'équivalence de traduction, l'affinité de catégorie grammaticale (p. ex. un nom peut être traduit par un nom ou un verbe), les collocations, la localité (l'alignement des mots se trouvant dans le voisinage des mots déjà alignées), etc. Un lien devient ainsi, par le calcul de cet ensemble de traits, un objet structuré, indépendant du corpus et même des unités qui le composent, pouvant être manipulé dans des traitements ultérieurs. Cette procédure est appelée par Tufiș *et al.* (2006) *la réification de l'alignement*. Les traits calculés pour chaque lien sont utilisés pour combiner les alignements fournis par les deux aligneurs mentionnés ci-dessous. L'aligneur résultant est ainsi *COWAL*[135] (Tufiș *et al.*, 2006).

À la différence de notre méthode, les cognats sont calculés à l'aide de la mesure *Levenshtein* (Levenshtein, 1966). De plus, avant le calcul de ce score, le corpus est passé par une étape de normalisation afin de diminuer les différences orthographiques entre les mots d'une paire bilingue. Les opérations ainsi effectuées sur le corpus sont l'élimination des diacritiques, des

[134] http://www2.sims.berkeley.edu/research/conferences/hlt-naacl03/

[135] *COWAL* a été classé en première position, parmi les 37 systèmes participants, dans le cadre du *Shared Task on Word Alignment (Romanian - English track) (ACL 2005 Workshop on Building and Using Parallel Corpora Data-driven Machine Translation and Beyond* (Martin *et al.*, 2005)) (Tufiș *et al.*, 2006).

.

consonnes doubles, de certains suffixes, etc. Dans notre méthode, nous avons aussi appliqué deux de ces opérations simples (la suppression des diacritiques et des consonnes doubles) mais nous avons défini, en outre, un ensemble de règles sensitives et non sensitives au contexte phonétique (cf. section 4.2.), pour diminuer encore plus les différences orthographiques entre les mots d'une paire bilingue.

Concernant les heuristiques exploitées, *YAWA* utilise un ensemble de règles simples établies empiriquement afin d'aligner les mots appartenant aux *chunks*, qui sont restés non alignés pendant l'étape précédente (un exemple de ces règles figure dans l'Annexe 7). Dans notre approche, les règles définies sont plus complexes et en même temps plus précises (étant définies manuellement). Celles-ci sont basées certaines sur la structure morphosyntaxique de chaque langue et d'autres sur le style juridique du corpus étudié (cf. section 4.3.).

De plus, *MEBA* collecte des alignements au niveau des collocations. Le système construit des listes de bi-grammes par le biais de la mesure d'association entre les mots appelée *log-likelihood*[136] (Dunning, 1993) et le calcul de la fréquence d'occurrence minimale pour filtrer les candidats, à partir de chaque partie monolingue du corpus parallèle d'entraînement. Dans notre approche, nous avons initialement utilisé un dictionnaire de collocations (Todirașcu *et al.*, 2008) comme ressource externe qui n'est plus intégré dans le système final à cause de l'impact non significative sur les résultats globaux (cf. sous-section 4.4.4.).

Les trois aligneurs présentés ci-dessus ont été évalués en termes de précision, rappel et F-mesure. Ceux-ci fournissent des valeurs performantes pour la F-mesure, comme suit : *YAWA* - 81.22%, *MEBA* - 80,17% et *COWAL* - 83,30% (Tufiș *et al.*, 2006). Au vu de ces résultats, ces systèmes obtiennent de meilleurs résultats que ceux fournis par notre système d'environ 5% (*MEBA*), 6% (*YAWA*) et 8% (*COWAL*).

Cette comparaison reste toutefois approximative car, comme nous l'avons déjà précisé auparavant, les systèmes n'utilisent pas les mêmes corpus de test et les mêmes annotations dans les corpus de référence. De plus, ces systèmes sont dépendants des langues qu'ils traitent. Ainsi, le français et le roumain sont deux langues riches morphologiquement qui demandent des traitements automatiques complexes pour pouvoir manipuler leur morphologie riche tandis que l'anglais a une morphologie plus simple.

[136] Ce score d'association a été décrit dans la sous-section 4.4.2.3.

Au vu de ces considérations, nous ne pouvons pas, pour le moment, tirer une conclusion nette concernant les résultats de notre méthode par rapport aux autres systèmes incluant le roumain qui sont disponibles. Des études ultérieures où les systèmes utilisent les mêmes corpus sont donc encore nécessaires dans ce sens, pour pouvoir effectuer une comparaison efficiente des résultats.

Pour rendre compte des améliorations obtenues dans le cadre de notre système d'alignement lexical, nous testerons également les modules d'alignement mis en place, séparément, dans le système de traduction automatique factorisé lui-même (cf. chapitre suivant).

4.7. Bilan du chapitre

Ce chapitre a été entièrement consacré au système d'alignement lexical qui est intégré dans le système de traduction automatique factorisé (cf. chapitre suivant). Comme l'alignement lexical est une étape de premier ordre dans la construction des systèmes de traduction automatique statistique, nous nous sommes concentrés sur le développement d'une approche hybride (Tufiș *et al.*, 2006 ; Schrader, 2006 ; Hermjakob, 2009 ; Cendejas *et al*., 2009 ; Pal *et al*., 2013), afin d'améliorer les résultats d'un système purement statistique (Brown *et al.*, 1990 ; Brown *et al.*, 1993 ; Och et Ney, 2000, 2003) pour la paire de langues traitées. Néanmoins, certaines approches ont révélé que l'amélioration significative de l'alignement mène seulement à une faible amélioration des résultats de traduction automatique (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007). Ainsi, notre démarche était d'évaluer l'influence des informations linguistiques incorporées (lemmes, heuristiques morphosyntaxiques, etc.) sur les résultats du système d'alignement lexical mais aussi du système de traduction automatique français - roumain (cf. chapitre suivant).

Pour construire le système d'alignement lexical, nous avons choisi le corpus juridique et administratif disponible gratuitement *DGT-TM* (Steinberger *et al*., 2012), déjà aligné au niveau propositionnel. L'avantage de celui-ci, par rapport aux autres corpus disponibles (cf. chapitre 3), est que l'alignement propositionnel est validé manuellement. Nous en avons extrait un ensemble de 64 923 paires de phrases alignées 1:1, constituant environ un million et demi de tokens pas langue. La taille du corpus a été limitée car les étapes d'étiquetage et d'alignement lexical se sont avérées coûteuses en temps et en ressources humaines. En outre, l'entraînement des systèmes de traduction factorisés sur des quantités plus importantes de données requiert un temps considérable (Ceaușu, 2009). Ainsi, nous avons estimé que cette

quantité de données est suffisante pour effectuer nos premières expériences de traduction automatique français - roumain et étudier ainsi l'influence des informations linguistiques incorporées sur les résultats du système de traduction, en fonction de la direction d'entraînement (cf. chapitre suivant). Mais l'utilisation du corpus entier est l'une des perspectives de ce travail, car nous sommes conscients qu'une quantité plus importante de données d'entraînement peut améliorer les résultats de la traduction automatique statistique, surtout pour des langues riches morphologiquement comme sont le français et le roumain.

Le corpus d'entraînement a été tout d'abord segmenté lexicalement, lemmatisé et étiqueté au moyen de l'étiqueteur *TTL* (Ion, 2007 ; Todirașcu *et al*., 2011) (cf. chapitre 3). Ensuite, il a été aligné lexicalement selon la méthodologie proposée (Navlea et Todirașcu, 2013, 2014).

Si des corpus parallèles d'entraînement sont disponibles pour la paire de langues étudiées, ce n'est pas aussi le cas des corpus de référence. À notre connaissance, il n'existe pas de corpus de référence pour le français et le roumain. De ce fait, afin d'évaluer le système d'alignement lexical, deux variantes d'un corpus de référence ont été construites manuellement : une variante qui ne contient pas l'alignement des collocations (appelée *RefA*) et une autre qui comprend ces alignements (dénommée *RefB*). Cette démarche a été motivée par le fait que le système d'alignement lexical initial intègre un dictionnaire de collocations (Todirașcu *et al*., 2008) (cf. section 4.4.).

Ce corpus de test est aussi extrait de *DGT-TM*. Il comprend 1 000 phrases alignées constituant environ 30 000 tokens par langue et ne fait pas partie du corpus d'entraînement. Sa taille est relativement réduite car l'alignement lexical manuel constitue une tâche coûteuse en termes de temps et de personnels. Nous avons également segmenté lexicalement, lemmatisé et étiqueté ce corpus.

Pour aligner le corpus de test au niveau lexical, nous avons utilisé le guide proposé par Melamed (1998) pour la paire de langue français - anglais. De plus, en fonction des spécificités morphosyntaxiques des langues étudiées, ainsi que du style juridique du corpus, nous avons également défini des règles d'annotations supplémentaires (cf. sous-sections 4.1.3. et 4.1.4.). L'ensemble de ces règles a été constitué à l'issue de l'étude des erreurs d'alignement lexical fournies par le système statistique de base, construit à l'aide de *GIZA++* (Och et Ney, 2000, 2003).

La première étape du système d'alignement lexical consiste ainsi dans l'application de l'outil *GIZA++* (cf. sous-section 4.1.1.) dans les deux sens du processus d'alignement et l'obtention de l'alignement de base par l'heuristique d'intersection (Koehn *et al.*, 2003). Cette heuristique fournit des alignements considérés comme sûrs du moment qu'ils sont repérés dans les deux directions d'alignement. Cependant, ce système de base génère en sortie un certain nombre d'erreurs pour lesquelles nous avons tenté de proposer des stratégies de correction automatique afin de diminuer leur taux. Pour ce faire, nous avons procédé à l'analyse linguistique des erreurs fournies par ce système de base (cf. sous-section 4.1.3.) et à l'étude du corpus parallèle au niveau de la traduction humaine (cf. 4.1.4.) afin de repérer éventuellement des erreurs dues aux procédés de traduction usités dans le corpus.

Les erreurs repérées concernent principalement l'alignement des cognats, les problèmes dus aux différences morphosyntaxiques entre les deux langues étudiées et aux contraintes stylistiques en traduction, les collocations et les termes poly-lexicaux du domaine. À partir de ces catégories d'erreurs, nous avons mis en place trois stratégies d'amélioration des sorties fournies par *GIZA++* :

- le développement d'un module d'identification et d'alignement automatique des cognats (Navlea et Todirașcu, 2011abe, 2012) ;
- l'implémentation d'un module à base de règles heuristiques morphosyntaxiques et stylistiques (Navlea et Todirașcu, 2010ab, 2011cd ; Navlea et Havași, 2012) ;
- l'intégration d'un dictionnaire de collocations (Todirașcu *et al.*, 2008) comme ressource externe (Navlea et Todirașcu, 2014).

La méthode de détection et d'alignement automatique de cognats utilise une combinaison originale de techniques statistiques et d'informations linguistiques comme les lemmes, les parties du discours, les équivalences de catégorie lexicale. Cette approche exploite des méthodes n-grammes (Simard et *al.*, 1992) et des techniques de désambigüisation des données d'entrée comme les extractions itératives des candidats, la suppression des données d'entrée des candidats considérés fiables à l'issue de chaque étape d'extraction, le calcul des fréquences pour filtrer les cas ambigus.

Cette méthode applique initialement une étape de normalisation au corpus d'entrée (Tufiș et *al.*, 2006 ; Ceaușu, 2009), afin de diminuer les différences orthographiques entre les paires

bilingues de lemmes des unités lexicales. Cette étape utilise un ensemble d'ajustements orthographies simples comme la suppression des diacritiques et l'élimination des consonnes doubles (Tufiș *et al.*, 2006 ; Ceaușu, 2009) ainsi que des règles d'ajustements définies empiriquement. Ces règles sont basées sur le repérage des correspondances phonétiques à l'intérieur des paires bilingues de lemmes et peuvent être sensitives ou non sensitives au contexte phonétique (cf. sous-section 4.2.1.). L'originalité de cette étape réside dans la définition et l'application de ces règles au corpus parallèle d'entrée. En effet, celles-ci exploitent l'idée que le français a une écriture étymologique tandis que le roumain possède une écriture généralement phonétique. Ainsi, il est possible de repérer les correspondances phonétiques et remplacer les graphèmes du français par les phonèmes du roumain (p. ex. ***ph****a****s****e* - ***f****a****z****ă** vs. ***f****a****z****e* - ***f****a****z****a*). Cette étape d'ajustements orthographiques a augmenté le rappel d'une simple méthode 4-grammes (Simard *et al*., 1992) de manière significative de presque 25% (cf. sous-section 4.2.2.).

La méthode a obtenu de bonnes performances en termes de F-mesure et de meilleurs scores par rapport aux méthodes traditionnels comme le coefficient de *Dice* (Adamson et Boreham, 1974 ; Brew et McKelvie, 1996) et le calcul de la sous-chaîne maximale (Melamed, 1999).

Cependant, les résultats obtenus sont dépendants des langues traitées, du domaine et de la taille du corpus utilisé, ainsi que de la qualité de l'étiquetage du corpus. De ce fait, les résultats obtenus ne peuvent pas être généralisés et des expériences sur d'autres corpus de test restent encore nécessaires. De plus, la méthode fournit encore des erreurs dans le cas des candidats ambigus, des formes hapax ou des candidats présentant des similarités orthographiques et/ou phonétiques très faibles.

Comme cette méthode ne demande pas de ressources externes telles que des lexiques, dictionnaires ou des listes de cognats validées par des experts, elle peut être facilement adaptée à d'autres langues apparentées.

Concernant l'intégration des cognats dans le processus d'alignement lexical (cf. sous-section 4.2.3.), ceux-ci ont amélioré le score *AER* de manière intéressante d'environ 2% par rapport au système de base. Les alignements des cognats seront aussi évalués dans le système de traduction automatique factorisé lui-même (cf. chapitre suivant).

Après l'étape d'identification et d'alignement automatique des cognats, nous avons appliqué un ensemble de 37 règles heuristiques morphosyntaxiques et stylistiques. Ces règles ont été définies afin de corriger les erreurs du système d'alignement lexical de base apparues à cause des différences morphosyntaxiques entre les deux langues étudiées et des contraintes stylistiques au niveau de la traduction humaine juridique.

La combinaison des techniques statistiques et des heuristiques linguistiques (Schrader, 2006 ; Tufiș *et al.*, 2006 ; Hermjakob, 2009 ; Cendejas *et al*., 2009 ; Pal *et al*., 2013) n'est pas une idée originale mais la base de règles définies est spécifique à la paire de langues français - roumain et, à notre connaissance, il n'existe pas d'autres ressources de ce type pour cette paire de langues. Cette ressource peut donc être utilisée pour d'autres expériences d'alignement lexical français - roumain.

Cette base de règles peut aussi servir de guide d'alignement lexical manuel pour la constitution d'autres corpus de référence français - roumain. Celle-ci peut avoir la même utilité pour d'autres paires de langues incluant le roumain ou le français, car ces règles décrivent des structures morphosyntaxiques spécifiques à chaque langue, étant basées donc sur des connaissances linguistiques de type contrastif. Ainsi, à partir des structures morphosyntaxiques qui posent problème à l'alignement automatique connues, en français ou en roumain, il est facile de repérer la structure équivalente dans une autre langue à partir des corpus alignés manuellement, par exemple, et, par conséquent, de définir la règle correspondante pour la nouvelle paire de langues (dans le cas, bien évidemment, où cette structure est différente et nécessite donc la définition d'une règle). De plus, cette base contient des règles fondées sur la connaissance du style juridique du corpus étudié et des procédés de traduction usités dans le corpus (cf. sous-section 4.1.4.). L'originalité de cette démarche consiste dans la définition de certaines règles stylistiques par la collecte des lexèmes avec renforcement de sens à valeur stylistique, qui n'apparaissent pas d'une langue à l'autre, dans le cas des procédés de traduction comme l'étoffement ou le dépouillement. L'ensemble de ces lexèmes comprend certains adjectifs qualificatifs, certains déterminants indéfinis, démonstratifs, certains verbes et adverbes (cf. sous-section 4.1.4.).

Le module implémentant les règles heuristiques a amélioré de manière significative les résultats du système d'approximativement 5%. Néanmoins, ces résultats sont dépendants de la paire de langues étudiées, du volume et du domaine du corpus utilisé mais aussi des annotations d'alignement effectuées dans le corpus de référence. De ce fait, des études

ultérieures restent encore nécessaires sur d'autres corpus de test afin de pouvoir généraliser et d'enrichir également la base de règles proposées.

Une autre méthode que nous envisageons de mettre en place pour enrichir la base d'heuristiques définies consiste dans l'apprentissage automatique des règles à partir du corpus annoté et aligné manuellement au niveau lexical. Comme nous n'avons pas disposé initialement de corpus de référence annotés, nous n'avons pas pu utiliser la méthode d'apprentissage automatique supervisée. De ce fait, notre démarche première a été la définition manuelle des règles. Celles-ci ont toutefois l'avantage d'être plus précises du point de vue grammatical par rapport à des règles apprises automatiquement.

La méthode d'apprentissage automatique offre la possibilité, d'un côté, d'acquérir un ensemble de règles plus volumineux que celui obtenu manuellement. D'un autre côté, cette méthode propose des règles qui sont plus conformes aux corpus utilisés que celles définies manuellement. En effet, si le corpus présente des erreurs systématiques d'étiquetage, par exemple, l'apprentissage automatique exploite ces erreurs dans la définition des règles. Cette approche a ainsi deux avantages principaux : l'adaptabilité et la portabilité entre les paires de langues ou entre les corpus utilisés (Ozdowska, 2006).

Comme les résultats d'évaluation obtenus dépendent du corpus de test utilisé, afin de rendre compte de la fiabilité de la base de règles définies, avec plus de précision, nous avons évalué leur influence sur le système de traduction factorisé lui-même (cf. chapitre suivant).

La dernière étape du système consiste dans l'intégration d'un dictionnaire de collocations *Verbe + Nom* (Todirașcu *et al*., 2008) dans le processus d'alignement lexical. Ce dictionnaire contient 250 entrées pour le roumain et le français. Chaque collocation verbo-nominale est munie de ses propriétés morphosyntaxiques contextuelles. De plus, nous avons enrichi le dictionnaire par l'ensemble des collocations nominales correspondantes de type *Nom 1 déverbal + (préposition) + Nom 2*. Pour ce faire, nous avons mis en place une méthode de génération automatique des collocations nominales à partir des entrées du dictionnaire. Cette méthode comprend deux étapes principales :

1) l'étude des collocations *Nom 1 déverbal + (préposition) + Nom 2* en contexte afin de repérer leurs propriétés morphosyntaxiques, selon la méthodologie proposée par Todirașcu *et al*. (2008) ; Pendant cette étape, nous avons relevé les propriétés

morphosyntaxiques des collocations nominales et défini les filtres linguistiques correspondants (cf. sous-section 4.4.2.4.).

2) la génération automatique des collocations nominales françaises à partir des entrées du dictionnaire au moyen du lexique *VerbAction* (Hathout *et al*., 2002 ; Tanguy et Hathout, 2002) et de la morphologie des déverbaux roumains qui se forment généralement à partir de l'infinitif des verbes d'origine (verbe à l'infinitif suivi de la terminaison *-RE*) ; L'originalité de cette démarche consiste dans la combinaison des ressources et des indices morphologiques afin d'automatiser la collecte des collocations nominales bilingues à partir du dictionnaire, comme une solution moins coûteuse en temps et en ressources humaines par rapport à la définition manuelle des équivalences.

Cependant, le dictionnaire n'est plus intégré par le système d'alignement lexical final car l'évaluation des résultats globaux a révélé une amélioration négligeable du score *AER*. (cf. sous-section 4.4.4.). Cela est dû au fait qu'il existe très peu de collocations communes entre le dictionnaire et le corpus de test utilisé. De ce fait, nous envisageons dans l'avenir d'effectuer des expériences sur d'autres corpus de test (de taille plus volumineuse) afin de pouvoir étudier, avec plus de précision, l'influence du dictionnaire dans le processus d'alignement. De plus, nous enrichirons le dictionnaire avec l'ensemble des collocations verbo-nominales supplémentaires que nous avons validées manuellement lors de la construction du corpus de référence contenant les alignements des collocations (*RefB*) (cf. sous-section 4.4.4.). Cette ressource comprend environ 230 collocations verbo-nominales par langue.

L'évaluation globale du système développé a montré que l'utilisation des cognats et des règles heuristiques linguistiques dans le processus d'alignement lexical a amélioré nettement les résultats finaux d'environ 7%, par rapport au système de base purement statistique. L'inconvénient de ce système est néanmoins le fait qu'il est dépendant de la paire de langues étudiées et du corpus utilisé. De plus, la définition manuelle des règles basées, d'une part, sur des connaissances linguistiques de type contrastif et, d'autre part, sur l'étude du style juridique du corpus et des procédés de traduction est une tâche coûteuse en temps. De ce fait, il est utile de combiner la définition manuelle des règles avec des techniques d'apprentissage automatique. Mais, pour ce faire, des corpus de référence annotés et alignés lexicalement doivent être construits. Soulignons que, pour la paire de langues étudiées, ce genre de ressources n'est pas encore disponible.

Toutefois, comme le système développé est modulaire, il présente l'avantage d'être facilement adaptable à d'autres paires de langues apparentées.

Nous avons testé les améliorations obtenues dans le cadre de notre approche d'alignement lexical dans le système de traduction automatique factorisé lui-même. Les expériences effectuées, dans les deux directions du processus de traduction, seront présentées dans le chapitre suivant.

5. Expériences de traduction automatique statistique (factorisée) français - roumain

Commençons par rappeler ci-dessus les étapes suivies afin de développer notre système de traduction automatique statistique factorisée, dans les deux sens du processus de traduction :

1. la constitution de corpus monolingues en langue cible et des corpus bilingues parallèles alignés au niveau propositionnel ;

2. le prétraitement des corpus (segmentation lexicale, lemmatisation, étiquetage morphosyntaxique et annotation au niveau des *chunks*), par l'application de *TTL* (Ion, 2007 ; Todirașcu *et al.*, 2011) (cf. chapitre 3, sous-section 3.2.3.) ;

3. l'alignement lexical de base (purement statistique) des corpus bilingues parallèles, par l'utilisation de *GIZA++* (Och et Ney, 2000, 2003) (cf. chapitre 4, sous-section 4.1.1.) ;

4. l'étude des erreurs de l'alignement lexical de base, leur classification et les propositions d'amélioration du système de base (cf. chapitre 4, sous-sections 4.1.3. et 4.1.4.) ;

5. la mise en place du système hybride d'alignement lexical :

    5.1. le développement d'un module d'identification et d'alignement automatique des cognats n'ayant pas été, dans un premier temps, alignés par *GIZA++* (cf. chapitre 4, section 4.2.) ;

    5.2. la définition et l'implémentation de règles de correction morphosyntaxiques et stylistiques, pour d'autres types d'erreurs récurrentes de l'alignement lexical de base (cf. chapitre 4, section 4.3.) ;

    5.3. l'application d'un dictionnaire de collocations *Verbe + Nom* (Todirașcu *et al.*, 2008), afin de corriger les erreurs au niveau de ces constructions (cf. chapitre 4, section 4.4.) ;

6. l'évaluation automatique du système d'alignement ;

7. la construction de modèles de langue en langue cible pour le système de traduction, par l'utilisation de l'application dénommée *SRILM* (Stolcke, 2002) (cf. chapitre 3, sous-section 3.1.1.) ;

8. la construction de modèles de traduction automatique statistiques et factorisés, par le biais des scripts appropriés du décodeur utilisé *MOSES*[137] (Koehn *et al.*, 2007) (cf. chapitre 3, sous-section 3.1.1.) ;

9. l'évaluation automatique des systèmes de traduction développés ;

10. l'optimisation des paramètres du décodeur par l'utilisation de l'application appelée *MERT* (Bertoldi *et al.*, 2009) (cf. chapitre 3, sous-section 3.1.1.) ;

11. l'évaluation automatique des systèmes optimisés.

Ce sont les étapes de 7 à 11 présentées ci-dessus qui feront l'objet de ce chapitre.

Comme nous l'avons vu dans le chapitre précédent, nous sommes partis, d'un côté, de l'idée qu'un alignement lexical combinant des techniques statistiques et des informations linguistiques (Tiedemann, 2003 ; Cherry et Lin, 2003 ; Tufiș *et al.*, 2006 ; Schrader, 2006 ; Hermjakob, 2009 ; Cendejas *et al.*, 2009 ; Pal *et al.*, 2013) montre de meilleures performances par rapport à un alignement purement statistique (Brown *et al.*, 1990 ; Brown *et al.*, 1993 ; Och et Ney, 2000, 2003) et améliore ainsi les résultats des systèmes statistiques de traduction. D'un autre côté, nous avons aussi pris en considération l'idée qu'une amélioration considérable de l'alignement ne mène pas nécessairement à une amélioration importante des résultats de traduction finaux (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007).

Ainsi, afin de vérifier ces deux constats pour la paire de langues étudiées, nous avons amélioré les résultats de notre système d'alignement lexical de base (purement statistique) par l'exploitation des cognats et des heuristiques linguistiques. Cette amélioration est illustrée par un score *AER* diminué de façon significative d'environ 7% (cf. chapitre 4, section 4.5.).

Dans nos expériences de traduction automatique, nous vérifierons, tout d'abord, si l'utilisation de facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction (Koehn et Hoang, 2007 ; Avramidis et Koehn, 2008 ; Tufiș *et al.*, 2008b ; Ceaușu,

[137] Version installée pour *Ubuntu 10.10 - 11.04 32bit / 64bit* (4169).

2009 ; Ceaușu et Tufiș, 2011) améliore les résultats de la traduction automatique, par comparaison avec un système de base, purement statistique. Ensuite, nous évaluerons, d'une part, l'impact de l'alignement lexical développé sur les résultats du système de traduction factorisé mis en place, dans les deux sens du processus de traduction. D'autre part, nous étudierons l'influence de chaque module supplémentaire d'alignement lexical implémenté sur le système de traduction factorisé et notamment le module d'identification et d'alignement automatique des cognats (cf. chapitre 4, section 4.2.) et celui intégrant la base des règles heuristiques linguistiques constituée (cf. chapitre 4, section 4.3.).

Dans la section suivante, figureront les corpus utilisés afin de construire et d'évaluer les systèmes de traduction automatique français - roumain, dans les deux sens du processus de traduction.

5.1. Les corpus utilisés

Les corpus exploités dans nos expériences de traduction automatique et notamment les corpus d'entraînement, de test et de développement, sont décrits dans le Tableau 31 ci-dessous. Leur taille est donnée en nombre de couples de phrases alignées, de mots et de tokens (mots et signes de ponctuation).

Tableau 31. La description des corpus utilisés en nombre de couples de phrases alignées, tokens et mots

Corpus extraits de *DGT-TM*	**Taille du corpus**	**français**	**roumain**
Corpus d'entraînement	phrases alignées	64 923	64 923
	tokens	1 818 768	1 554 233
	mots	1 678 011	1 402 069
Corpus de test	phrases alignées	300	300
	tokens	7 798	6 551
	mots	6 963	6 040
Corpus de développement	phrases alignées	300	300
	tokens	10 563	8 732
	mots	9 275	7 815

Comme nous l'avons déjà précisé dans le chapitre 3 (section 3.2.), où nous avons décrit en détail les corpus monolingues et bilingues parallèles disponibles pour la paire de langues étudiées, nous avons utilisé un ensemble de 64 923 couples de phrases alignées (représentant environ un million et demi de tokens par langue) extraits du corpus juridique et administratif *DGT-TM* (Steinberger *et al.*, 2012). La taille et le domaine du corpus ont été limités car, d'une part, le prétraitement et l'alignement lexical des corpus bilingues parallèles, selon notre approche hybride, se sont avérées des tâches coûteuses en temps et en ressources humaines et, d'autre part, l'entraînement et l'optimisation des systèmes de traduction factorisés requièrent également un temps considérable (cf. chapitre 3, sous-section 3.2.1.). Rappelons aussi que nous avons opté pour le corpus *DGT-TM* car la plupart des alignements propositionnels sont effectués manuellement et donc, par rapport aux autres corpus alignés automatiquement, *DGT-TM* est supposé à contenir moins d'erreurs au niveau de l'alignement propositionnel.

Les corpus de test et de développement retenus comprennent chacun un ensemble de 300 phrases alignées représentant environ 7 000 tokens par langue en ce qui concerne le corpus de test et 9 000 tokens par langue quant au corpus de développement. Ceux-ci ne font pas partie du corpus d'entraînement.

Précisons aussi qu'afin de construire les systèmes de traduction purement statistiques, les corpus utilisés sont en état brut ne comprenant aucun prétraitement linguistique, tandis que les corpus exploités par les systèmes factorisés sont lemmatisés, étiquetés par des parties du discours et des propriétés morphosyntaxiques (cf. chapitre 3, sous-section 3.2.3.) et préparés dans le format approprié requis par *MOSES*. La Figure 65 suivante comprend un exemple de phrases alignées extrait du corpus d'entraînement des systèmes factorisés : *L'arbitre ainsi nommé préside l'arbitrage.* (FR) vs. *Arbitrul astfel ales prezidează arbitrajul.* (RO)

français	**roumain**
L'|le^Da|DAS|Da-ms arbitre|arbitre^Nc|NSN|Ncms ainsi|ainsi^R|R|R nommé|nommer^Vm|VP|Vmps-s préside|présider^Vm|VS3|Vmsp3s l'|le^Da|DAS|Da- ms arbitrage|arbitrage^Nc|NSN|Ncms .|.|.|.	Arbitrul|arbitru^Nc|NSRY|Ncmsry astfel|astfel^Rg|R|Rgp ales|ales^Rg|R|Rgp prezidează|prezida^Vm|V3|Vmip3 arbitrajul|arbitraj^Nc|NSRY|Ncmsry .|.|.|.

Figure 65. Exemple de phrases alignées françaises - roumaines extraites du corpus d'entraînement des systèmes factorisés

Dans l'exemple de la Figure 65, chaque token ne représente plus une simple forme du mot (comme dans le cas des systèmes statistiques purs), mais un vecteur de facteurs linguistiques tels que : la forme des mots, le lemme, les parties du discours et les propriétés morphosyntaxiques.

Analysons les tokens *arbitrage* du français et son correspondant roumain *arbitraj* qui sont représentés de la manière suivante : *arbitrage|arbitrage^Nc|NSN|Ncms* vs. *arbitrajul|arbitraj^Nc|NSRY|Ncmsry*. Dans cet exemple, le token (la forme du mot *arbitrage / arbitraj*) est accompagné de son lemme *arbitrage / arbitraj* suivi par les deux premiers caractères de son étiquette morphosyntaxique (*^Nc* - nom commun) : *arbitrage^Nc* vs. *arbitraj^Nc*. Cette opération s'avère utile car elle désambigüise morphologiquement le lemme (Tufiș *et al.*, 2005b, 2006) (cf. chapitre 3, sous-section 3.1.1.). Dans *MOSES*, la forme du mot est considéré le facteur linguistique 0, tandis que le lemme qui suit est considéré le facteur linguistique 1. Ensuite, le facteur linguistique 2 (l'étiquette *NSN* française de notre exemple, qui signifie nom au singulier, non déterminé, ou l'étiquette *NSRY* - nom au singulier, cas nominatif-accusatif, déterminé - de son correspondant roumain *arbitrul*) est représenté par un ensemble d'étiquettes morphosyntaxiques réduites dénommées *C-tag* (Tufiș, 1999, 2000) et dérivées à partir des étiquettes morphosyntaxiques (*MSD*) du projet *Multext-EAST* (Dimitrova *et al*., 1998), présentes dans notre corpus d'entraînement (cf. chapitre 3, sous-section 3.2.3.). Enfin, le facteur linguistique 3 est représenté par les étiquettes *MSD* qui décrivent les propriétés morphosyntaxiques des mots (l'étiquette *Ncms* du nom français - nom commun, masculin, singulier ou l'étiquette correspondante roumaine *Ncmsry* - nom commun, masculin, singulier, cas nominatif-accusatif, déterminé).

Précisons que l'annotation du corpus par des étiquettes *C-tag* et *MSD* est dénommée *annotation stratifiée* (Tufiș, 1999, 2000) et résout le problème du nombre élevé des classes d'équivalence en ce qui concerne l'annotation par les étiquettes *MSD*, pour les langues riches morphologiquement, présentant des paradigmes flexionnelles importantes (Ceaușu, 2009), comme c'est aussi le cas du français et du roumain. Cette technique est d'autant plus utile que nos ressources d'entraînement ont des dimensions limitées.

Remarquons que notre corpus d'entraînement présente quatre facteurs linguistiques associés aux mots : la forme des mots (0), le lemme (1), la partie du discours (2) et les propriétés morphosyntaxiques (3). À partir de ces facteurs, *MOSES* peut entraîner plusieurs configurations de modèles de traduction fournies par l'utilisateur, pendant différentes étapes

de traduction et de génération, afin de proposer une configuration optimale du modèle, en fonction de la direction de traduction.

Quant aux modèles de langue construits en langue cible, que *MOSES* exploite également dans le processus de traduction, nous avons utilisé des modèles déjà disponibles pour le roumain (Tufiș *et al*., 2013a)[138]. Ceux-ci sont des modèles de langue 5-grammes développés sur la forme des mots ou sur les étiquettes morphosyntaxiques (*MSD*) et sont créés à partir des corpus juridiques, journalistiques et médicaux. Comme en français nous avons disposé pour la construction de modèles de langue d'un seul corpus monolingue lemmatisé et étiqueté du domaine juridique (cf. chapitre 3, sous-section 3.2.1.), nous avons utilisé le modèle de langue roumain appartenant à ce domaine. Ce modèle a été obtenu à partir du corpus *JRC-Acquis* (Steinberger *et al*., 2006) en roumain.

Afin d'effectuer des expériences similaires dans les deux directions du processus de traduction, nous avons donc construit des modèles de langue 5-grammes français à base de formes de mots et de propriétés morphosyntaxiques (*MSD*) (comme en roumain), à partir du corpus juridique français lemmatisé et étiqueté, contenant 480 764 phrases extraites de *JRC-Acquis* (Steinberger *et al*., 2006) (cf. chapitre 3, sous-section 3.2.1.).

Les systèmes de traduction automatique statistiques et factorisés développés seront présentés dans la section suivante. L'évaluation des systèmes ayant comme langue source le roumain sera présentée dans la sous-section 5.2.1., tandis que l'évaluation des systèmes intégrant comme langue source le français figurera dans la sous-section 5.2.2.

5.2. Les systèmes de traduction automatique développés

Les expériences menées dans les deux sens du processus de traduction ont été réalisées par l'utilisation de la distribution standard du décodeur utilisé *MOSES* (Koehn *et al.*, 2007) (cf. chapitre 3, sous-section 3.1.1.). Deux sets d'expériences ont été effectués dans les deux directions de traduction, comme il est décrit de manière générale, ci-dessous.

Dans un premier temps, l'aligneur lexical *GIZA++* (Ocn et Ney, 2000, 2003) (cf. chapitre 3, sous-section 3.1.2.) a été appliqué dans les deux sens du processus de traduction.

[138] Nous remercions M. Ștefan Dumitrescu de l'*Institut de Recherche en Intelligence Artificielle* (ICIA) de l'Académie Roumaine de Bucarest, pour le partage des modèles de langue pour le roumain.

Ensuite, afin d'entraîner les modèles de traduction à base de séquences (Koehn *et al.*, 2003) (statistiques purs ou factorisés), l'heuristique appelée *grow-diag-final* (par défaut dans *MOSES* pour les deux types de modèles de traduction) a été initialement appliquée. Cette heuristique exploite l'intersection et l'union des alignements bidirectionnels (cf. chapitre 2, sous-section 2.2.2.1.). En effet, l'intersection de ces alignements est tout d'abord effectuée et les alignements communs sont alors sélectionnés. Ceux-ci sont considérés théoriquement fiables du moment qu'ils ont été repérés dans les deux sens du processus d'alignement. Ensuite, des liens supplémentaires provenant de l'union des alignements bidirectionnels sont rajoutés s'ils se trouvent dans le voisinage des alignements communs. Pour un deuxième set d'expériences, l'heuristique dénommée *union* a été aussi appliquée. Celle-ci prend en compte seulement l'union des alignements lexicaux bidirectionnels, autrement dit les liens repérés dans les deux sens du processus d'alignement sont mis ensemble et les liens répétés sont ajoutés une seule fois. En effet, cette heuristique considère les alignements communs et tous les alignements supplémentaires trouvés dans les deux directions, à la différence de *grow-diag-final* qui, à côté des alignements communs, prend en compte seulement les voisins de ces alignements. On peut considérer que *grow-diag-final* est plus fondée linguistiquement que *union*, car elle cherche, en théorie, à collecter plutôt des constituants que des groupes aléatoires. En outre, *union* ajoute bien évidemment des alignements plus nombreux que *grow-diag-final* et certains d'entre eux sont considérés à priori moins fiables, étant donné qu'ils ont été repérés dans une seule direction d'alignement et se trouvent en même temps à distance par rapport aux alignements communs. Par conséquent *grow-diag-final* est censée théoriquement mieux fonctionner que *union*. Toutefois, nous avons testé les deux heuristiques dans le cadre de nos expériences de traduction français - roumain afin d'étudier réellement l'impact de chacune sur les résultats finaux.

Suite à cette étape d'entraînement, des tables d'équivalents de traduction sont construites. *MOSES* combine ces tables de traduction et celles des modèles de langue en langue cible pour fournir la meilleure traduction d'une phrase source donnée en entrée.

Enfin, les systèmes sont évalués par le biais du score *BLEU* (par défaut dans *MOSES*) (cf. chapitre 2, sous-section 2.2.2.4.1.), optimisés par *MERT* (cf. chapitre 3, sous-section 3.1.1.) et réévalués. Dans nos expériences, nous avons appelé *BLEU 1* le score *BLEU* avant l'optimisation du système de traduction et *BLEU 2* le même score après son optimisation.

Dans cette section, seront décrits les systèmes de traduction automatique construits dans les deux sens du processus de traduction. Pour chacun des deux sets d'expériences mentionnés auparavant, nous avons entraîné un système de base, purement statistique, et trois variantes d'un système factorisé utilisant des alignements lexicaux différents, comme il est décrit plus loin.

Tout d'abord, nous avons construit le système de traduction automatique purement statistique (dans les deux sens du processus de traduction), considéré comme base pour nos comparaisons des résultats de différents systèmes entraînés. Celui-ci a été obtenu à partir du corpus parallèle d'entraînement brut, sans aucune information linguistique incorporée. Le système ayant comme langue cible le français exploite donc un modèle de langue que nous avons entraîné sur la forme des mots, à partir du corpus monolingue utilisé (cf. chapitre 3, sous-section 3.2.1.). Il s'agit d'un modèle 5-grammes réalisé par *SRILM* (Stolcke, 2002) (cf. chapitre 3, sous-section 3.1.1.), comme d'ailleurs tous les autres modèles de langue utilisés. En sens inverse, le système incorporant le roumain comme langue cible utilise un modèle de langue 5-grammes déjà disponible (cf. section antérieure), qui exploite également la forme des mots.

Ensuite, nous avons mis en place le système de traduction automatique factorisé dans les deux directions de la traduction. Celui-ci a été construit à partir du corpus parallèle d'entraînement lemmatisé et étiqueté (cf. section 5.1. de ce chapitre). Les modèles de langue appropriés pour un système factorisé (c'est-à-dire les modèles construits sur la forme des mots et sur différents facteurs linguistiques présents dans le corpus) que nous avons utilisés dans les deux sens du processus de traduction, sont les modèles développés sur la forme des mots et sur les étiquettes morphosyntaxiques (*MSD*) (cf. section antérieure).

Le choix initial des paramètres du système factorisé a été déterminé par les résultats du système anglais - roumain développé par Ceaușu (2009), qui a obtenu les meilleurs scores *BLEU* pour la configuration exploitant les lemmes et les étiquettes morphosyntaxiques (*MSD*) (cf. chapitre 3). De plus, le système roumain -> anglais utilise un modèle de distorsion (ré-ordonnancement) construit sur les étiquettes *MSD* de la langue source et de la langue cible, étant censé fonctionner mieux que le modèle par défaut utilisant la seule distance (Ceaușu et Tufiș, 2011) (cf. chapitre 2, sous-section 2.2.2.6.1.). Rappelons que le modèle basé sur la distance réordonne les mots cibles en fonction de la différence entre les positions des séquences sources et cibles. À la différence de ce modèle, le ré-ordonnancement exploitant les

formes des mots ou les étiquettes morphosyntaxiques calcule pour chaque séquence source la probabilité qu'elle soit traduite de manière monotone, échangé ou discontinue par rapport à la traduction de la séquence qui la précède.

Dans ces premières expériences français - roumain, nous avons utilisé seulement le modèle de distorsion par défaut (exploitant la distance) car, par rapport à l'anglais qui a un ordre de mots plutôt fixe, nécessitant donc des modèles de ré-ordonnancement plus performants en tant que langue cible, l'ordre des mots en français est plus flexible, voire libre en roumain. Une autre raison pour laquelle nous avons initialement utilisé seulement le modèle par défaut réside dans le fait que nous avons déjà exploité des modèles de langue développés sur les étiquettes morphosyntaxiques, qui ont le rôle d'améliorer le ré-ordonnancement des mots de la phrase cible, par rapport à un modèle construit sur la simple forme des mots (Ceauşu et Tufiş, 2011 ; Tufiş et Dumitrescu, 2012) (cf. chapitre 2, sous-section 2.2.2.6.1.). La combinaison de ces modèles de langue et de distorsion exploitant les formes des mots ou les étiquettes *MSD*, peut s'avérer redondante, ayant une influence négative sur les résultats de la traduction (Ceauşu et Tufiş, 2011) (cf. chapitre 2, sous-section 2.2.2.6.1.). Toutefois, l'utilisation des modèles de ré-ordonnancement construits sur les étiquettes *MSD* ou sur les formes des mots fait partie des perspectives de ce travail afin de pouvoir étudier leur impact sur la qualité de la traduction français - roumain.

Le système factorisé mis en place est basé ainsi sur les lemmes et les propriétés morphosyntaxiques (les étiquettes *MSD* du corpus) et utilise le modèle de langue construit sur la forme des mots et celui développé sur les étiquettes *MSD*, en langue cible. Les étapes du processus de traduction pour ce système sont les suivantes :

1) La traduction du lemme ;

2) La traduction de l'étiquette *MSD* ;

3) La génération de la forme des mots à partir du lemme et de l'étiquette *MSD* traduits auparavant.

Une autre configuration possible du système factorisé, qui pourrait aussi être entraînée à partir des données d'entraînement disponibles est représentée par le système développé sur les lemmes. Ce système utilise un modèle de langue exploitant la forme des mots et procède en deux étapes :

1) La traduction du lemme ;

2) La génération de la forme du mot à partir du lemme.

Cette configuration reste encore à vérifier dans nos expériences de traduction automatique ultérieures. En outre, nous envisageons dans l'avenir d'entraîner des modèles de langue utilisant l'information sur les parties du discours, pour les deux langues étudiées, afin de pouvoir développer des systèmes factorisés qui puissent modéliser également ce facteur. Nous n'avons pas encore effectué ces expériences car, pour le moment, nous ne disposons pas de corpus monolingues roumains incorporant ce facteur.

Le système factorisé entraîné dans le cadre de ces expériences est basé sur les lemmes et les étiquettes *MSD*, comme il a été vu antérieurement. De plus, afin de tester l'influence de notre système d'alignement lexical sur les résultats de la traduction de ce système, ainsi que l'influence de chaque module développé (exploitation de cognats et de règles linguistiques dans l'alignement lexical), nous avons entraîné encore deux systèmes factorisés utilisant des alignements lexicaux différents et notamment :

1) le système factorisé incorporant aussi les alignements supplémentaires fournis par le module d'identification et d'alignement automatique des cognats ;

2) le système factorisé intégrant également les alignements fournis par les cognats et par le module implémentant la base de règles heuristiques linguistiques constituée.

Dans la sous-section suivante, figurera l'évaluation des systèmes présentés ci-dessus, intégrant le roumain en tant que langue source.

5.2.1. L'évaluation des systèmes de traduction automatique du roumain vers le français

Le Tableau 32 suivant comprend l'évaluation des systèmes de traduction automatique développés, du roumain vers le français. Ceux-ci ont été obtenus par l'utilisation de l'heuristique *grow-diag-final*. L'évaluation est faite, rappelons-le, en termes de score *BLEU 1* (le score d'évaluation avant l'optimisation du système) et *BLEU 2* (le score après son optimisation).

Tableau 32. Systèmes de traduction automatique du roumain vers le français / Heuristique *grow-diag-final*

Systèmes roumain -> français	**Facteurs linguistiques / modèles de traduction**	**Facteurs linguistiques / modèles de la langue cible**	**BLEU1**	**BLEU2**
Statistique	formes des mots	formes des mots	**29,38**	**30,11**
Factorisé	lemmes et *MSD*	formes des mots et *MSD*	47,05	48,34
Factorisé + **cognats**	lemmes et *MSD*	formes des mots et *MSD*	46,94	48,24
Factorisé + **cognats** + **règles linguistiques**	lemmes et *MSD*	formes des mots et *MSD*	47,29	**48,48**

Le système de base, purement statistique, obtient un score *BLEU 1* de 29,38 et un score *BLEU 2* de 30,11. Le système factorisé montre des scores performants et dépasse de manière significative le système de base : de 17,67 points pour *BLEU 1* et de 18,23 pour *BLEU 2*. En revanche, les alignements additionnels des cognats diminuent légèrement les scores *BLEU 1* et *2* du système factorisé : de 0,11 pour le *BLEU 1* et de 0,10 pour le *BLEU 2*. Ensuite, les alignements lexicaux fournis par les cognats et les règles linguistiques améliorent de 0,35 points le *BLEU 1* et de 0,24 points le *BLEU 2* du système antérieur. Finalement, ce système apprécie légèrement de 0,24 points le *BLEU 1* et de 0,14 le *BLEU 2* du système factorisé. L'amélioration finale par rapport au système de base devient de 17,91 points pour le *BLEU 1* et de 18,37 pour le *BLEU 2* et s'avère toujours très importante.

Le graphique donné dans la Figure 66 ci-dessous montre l'évolution des scores *BLEU 1* et *BLEU 2* pour ce premier set d'expériences.

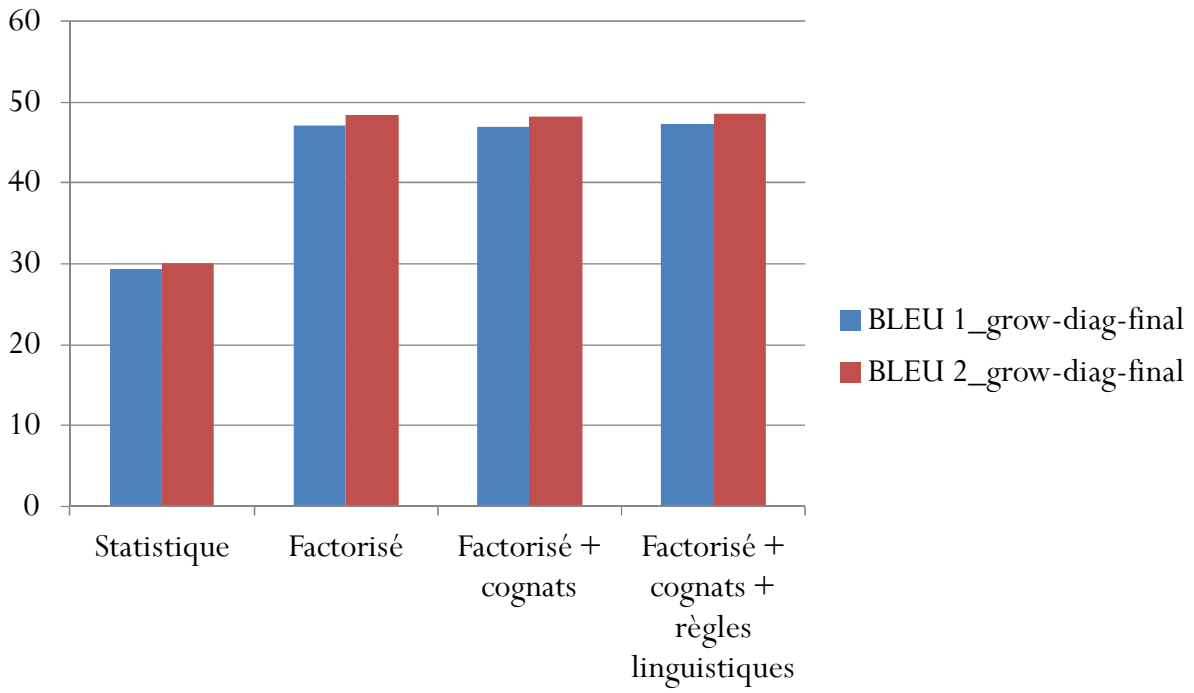


Figure 66. L'évolution des scores *BLEU 1* et *BLEU 2* des systèmes de traduction automatique du roumain vers le français / Heuristique *grow-diag-final*

Cet ensemble d'expériences montre bien que l'intégration de facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction automatique améliore de façon significative les résultats de la traduction roumain -> français, par rapport à un système de base statistique pur. En outre, une amélioration significative du système d'alignement lexical, par l'incorporation des informations linguistiques dans le processus d'alignement, peut améliorer les résultats d'un système factorisé (de 0,24 points le score *BLEU 1* et de 0,14 le *BLEU 2*, dans notre cas), même si cette amélioration s'avère légère.

Regardons maintenant le deuxième set d'expériences effectué par le biais de l'heuristique *union*, cette fois-ci. Celle-ci a été utilisée pour tenter d'obtenir une éventuelle amélioration des résultats des systèmes développés du roumain vers le français. Le Tableau 33 suivant comprend les scores *BLEU 1* et *2* ainsi obtenus.

Tableau 33. Systèmes de traduction automatique du roumain vers le français / Heuristique *union*

Systèmes roumain -> français	**Facteurs linguistiques / modèles de traduction**	**Facteurs linguistiques / modèles de la langue cible**	**BLEU1**	**BLEU2**
Statistique	formes des mots	formes des mots	**29,31**	**30,12**
Factorisé	lemmes et *MSD*	formes des mots et *MSD*	46,84	47,88
Factorisé + **cognats**	lemmes et *MSD*	formes des mots et *MSD*	**46,95**	**48,24**
Factorisé + **cognats** + **règles linguistiques**	lemmes et *MSD*	formes des mots et *MSD*	46,74	48,13

Le système de base obtient, dans ce cas, un score *BLEU 1* de 29,31 et un *BLEU 2* de 30,12, qui s'avèrent comparables aux scores fournis par *grow-diag-final*. Le système factorisé améliore aussi de façon significative les scores *BLEU 1* et *2*, par rapport au système de base (de 17,53 points le *BLEU 1* et de 17,76 le *BLEU 2*), mais présente des scores inferieurs à ceux obtenus par le biais de l'heuristique par défaut (la diminution du *BLEU 1* est de 0,21 et celle du *BLEU 2* de 0,46). Remarquons que les alignements additionnels des cognats améliorent cette fois-ci (même si légèrement) le score *BLEU 1* du système factorisé de 0,11 et le *BLEU 2* de 0,36. En revanche, les alignements supplémentaires des règles linguistiques semblent avoir un impact négatif sur les résultats antérieurs et, contrairement aux premières expériences, ceux-ci déprécient faiblement ces résultats (de 0,21 pour le score *BLEU 1* et de 0,11 pour le *BLEU 2*). La dépréciation finale du score *BLEU 1* par rapport au système factorisé est ainsi de 0,10 tandis que le score *BLEU 2* s'avère légèrement amélioré de 0,25.

Le graphique donné dans la Figure 67 suivante illustre l'évolution des scores *BLEU 1* et *BLEU 2* pour ce deuxième set d'expériences.

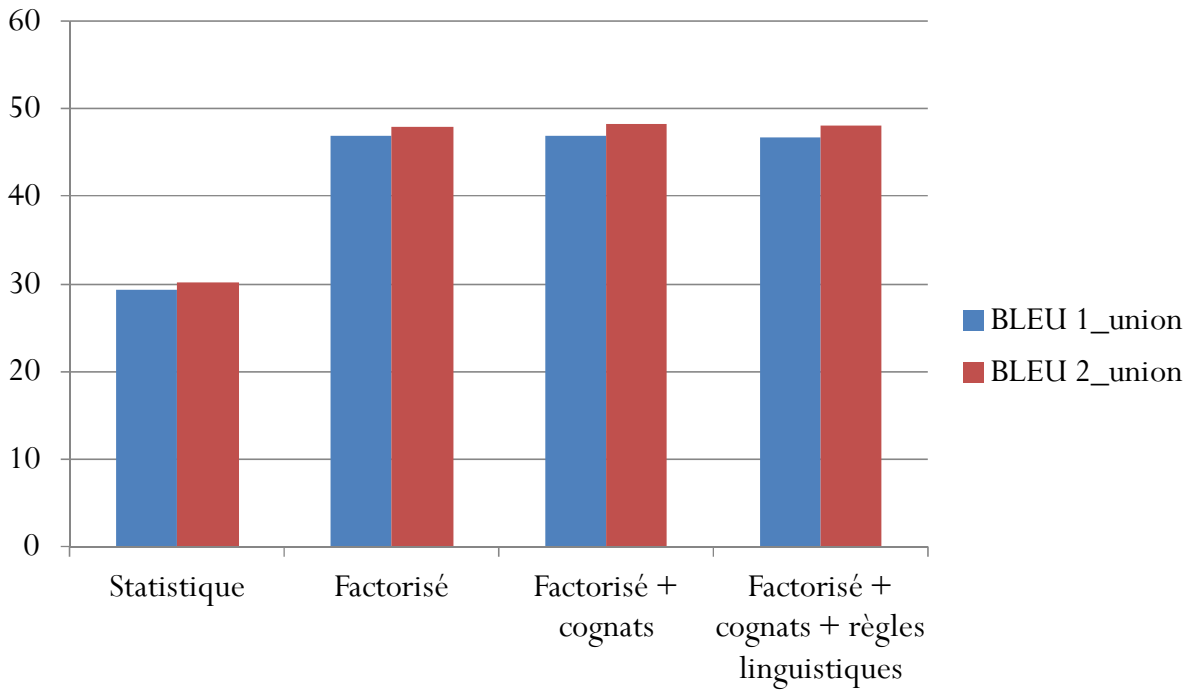


Figure 67. L'évolution des scores *BLEU 1* et *BLEU 2* des systèmes de traduction automatique du roumain vers le français / Heuristique *union*

Dans ce cas, l'intégration des facteurs linguistiques dans le processus de traduction automatique améliore aussi de manière significative les scores *BLEU 1* et *2* du système factorisé, par rapport au système de base, mais ces scores restent légèrement inférieurs à ceux obtenus lors des premières expériences. En outre, l'intégration des cognats pendant le processus d'alignement lexical produit aussi une légère amélioration des résultats du système factorisé, alors que l'incorporation des règles linguistiques déprécie faiblement les résultats antérieurs. Les deux modules appliqués en même temps améliorent toujours le système factorisé, même si cette amélioration reste néanmoins légère.

En conclusion, le système qui obtient les meilleurs scores lors de ces expériences est le système factorisé intégrant aussi les alignements additionnels des cognats. Pourtant, ces résultats finaux sont légèrement inférieurs à ceux obtenus initialement.

Les résultats obtenus lors de ces deux ensembles d'expériences présentés dans cette sous-section sont toutefois comparables, comme le témoigne aussi le graphique donné dans la Figure 68 suivante. Ce graphique illustre l'évolution de chaque score *BLEU* fourni par les systèmes entraînés à l'aide de *grow-diag-final* et d'*union*.

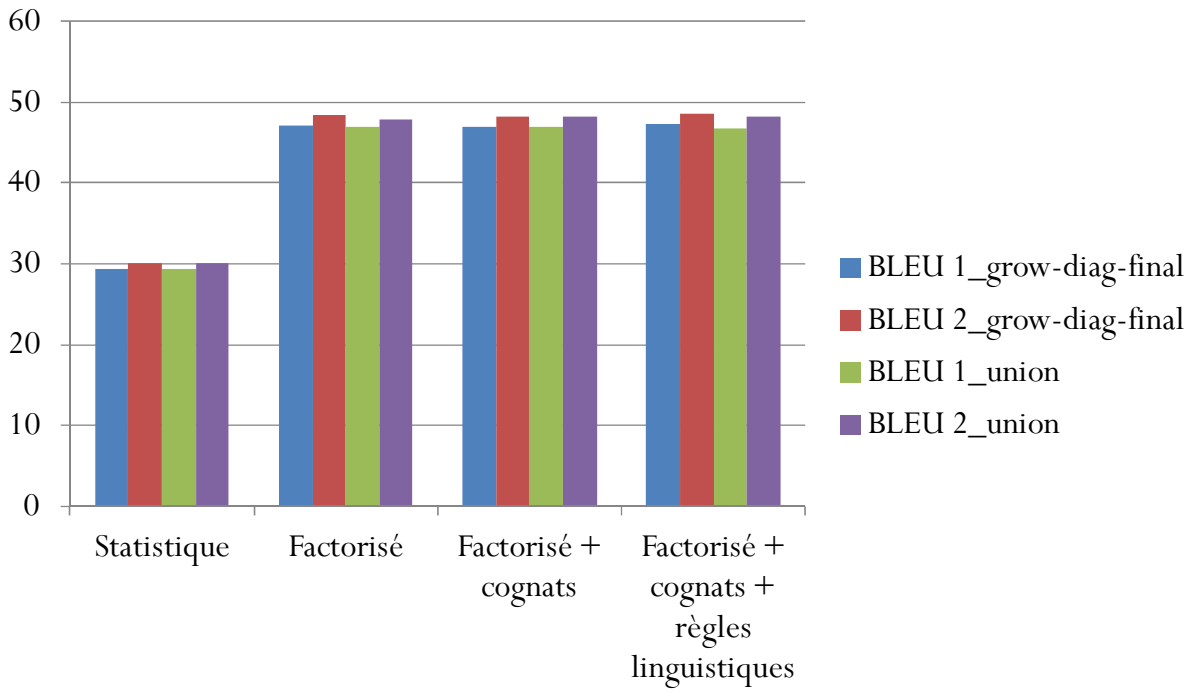


Figure 68. L'évolution des scores *BLEU 1* et *BLEU 2* des systèmes de traduction automatique du roumain vers le français / Heuristiques *grow-diag-final* et *union*

Comme le montre également le graphique ci-dessus, le score *BLEU* optimal est obtenu par le système factorisé utilisant l'heuristique par défaut *grow-diag-final*. Ce système intègre à la fois les alignements des cognats et des règles linguistiques. Dans cette direction de traduction, il semblerait que les alignements erronés ajoutés par l'heuristique *union* aient déprécié les résultats finaux par comparaison avec *grow-diag-final*, confirmant ainsi les constats de départ concernant les deux heuristiques (cf. section 5.2.). Les différences ne se sont pourtant pas avérées importantes entre les deux et les résultats restent comparables.

Récapitulons ci-dessous les trois cas de figure remarqués quant aux améliorations / dépréciations apportées par les modules supplémentaires d'alignement lexical pour la direction de traduction roumain -> français (cf. le graphique antérieur) :

1) Les alignements des cognats ont légèrement déprécié les résultats du système factorisé obtenu par *grow-diag-final*, mais ont apprécié de manière similaire ceux fournis par le système développé par *union*, devançant même le système suivant (+ cognats + règles + *union*) ; Il semblerait que les liens erronés ajoutés par les cognats aient eu un impact négatif sur le premier système, tandis que les liens corrects ajoutés aient influencé positivement les résultats du deuxième système.

2) Contrairement aux cognats, les alignements fournis par les règles linguistiques ont légèrement apprécié les résultats du système factorisé précédent (+ cognats + *grow-diag-final*), mais ont déprécié faiblement ceux donnés par le système (+ cognats + *union*) ; Dans ce cas, il paraîtrait que les liens corrects fournis par ce module ont influencé positivement les résultats du premier système, alors que les liens incorrects ont eu un impact négatif sur le deuxième système.

3) Les alignements des cognats et des règles appliqués à la fois ont amélioré les systèmes factorisés obtenus dans les deux sets d'expériences, après l'optimisation de ces systèmes. Ces améliorations se sont avérées néanmoins légères pour cette direction de traduction roumain -> français.

Nous pouvons en conclure que les scores finaux dépendent de l'heuristique utilisée. De plus, il faut tester toutes les configurations avec les deux heuristiques pour choisir celle qui se comporte le mieux.

À partir des résultats obtenus suite à ces deux sets d'expériences de traduction automatique, le système final retenu roumain -> français a la configuration donnée dans le Tableau 34 suivant. Ce système est basé sur les lemmes et les étiquettes *MSD*. Celui-ci exploite également un modèle de langue construit sur la forme des mots et un autre modèle développé sur les étiquettes *MSD* et montre des scores d'évaluation performants (le *BLEU 1* de 47,29 et le *BLEU 2* de 48,48 points).

Tableau 34. Configuration du système de traduction automatique du roumain vers le français

Système roumain -> français	**Facteurs linguistiques / modèles de traduction**	**Facteurs linguistiques / modèles de la langue cible**	**BLEU1**	**BLEU2**
Factorisé + **cognats** + **règles linguistiques**	lemmes et *MSD*	formes des mots et *MSD*	47,29	48,48

Par comparaison avec les résultats rapportés par Ceaușu (2009) pour la même configuration de facteurs linguistiques (lemmes et étiquettes *MSD*), entraînée du roumain vers l'anglais, le score *BLEU* effectif obtenu par notre système est inférieur au *BLEU* du système roumain ->

anglais qui est de 51,74 points. Ce dernier a été entraîné (pendant l'étape du choix de ses paramètres) à partir d'environ 60 000 couples de phrases alignées, qui comprennent environ un million et demi de mots, comme le corpus utilisé dans nos expériences (cf. chapitre 3). Rappelons qu'à la différence de notre approche, le système roumain -> anglais utilise de plus un modèle de distorsion basé sur les étiquettes *MSD* car, comme nous l'avons déjà mentionné auparavant, l'anglais nécessite en tant que langue cible des modèles de ré-ordonnancement plus performants qu'une langue présentant un ordre de mots plus flexible, comme le français et le roumain. Enfin, le système roumain -> anglais final a obtenu un score *BLEU* de 51,02 pour la même configuration mais entraînée, cette fois-ci, à partir du corpus entier de 800 000 couples de phrases alignées (contenant environ 30 millions de mots) (cf. chapitre 3). Dans sa version finale, ce système exploite aussi une ontologie et montre un très bon score *BLEU* (cf. chapitre 3 où celui-ci est présenté en détail). Dans notre approche, nous n'utilisons pas encore d'ontologies. De plus, rappelons que, dans le cadre de cette étude, nos expériences de traduction automatique sont basées seulement sur le corpus d'entraînement limité en taille, de 64 923 couples de phrases alignées. La taille du corpus a dû être réduite, d'une part, pour optimiser le temps d'entraînement des différentes configurations (cf. chapitre 3). D'autre part, cela est dû au fait que l'étiquetage et l'alignement lexical, nécessaires pour notre méthode hybride, se sont avérées des tâches coûteuses en temps et en ressources humaines (cf. chapitre 4). Mais nous projetons dans l'avenir d'étendre, bien évidemment, notre système par l'utilisation du corpus d'entraînement entier disponible pour le français et le roumain et représentant aussi environ 30 millions de mots (cf. chapitre 3, sous-section 3.2.1.).

Le Tableau 35 suivant comprend deux exemples de phrases traduites du roumain vers le français, tant par notre système de traduction automatique factorisée retenu que par le système commercial purement statistique *Google Translate*[139]. Ces exemples sont aussi accompagnés par leur traduction de référence. Ceux-ci proviennent du corpus de test utilisé, ayant extrait du *DGT-TM* (Steinberger *et al.*, 2012).

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

Tableau 35. Exemples de traductions roumain -> français fournies par notre système factorisé et par *Google Translate*

Phrases sources / roumain	**Traductions de référence**	**Phrases cibles / Notre système factorisé**	**Phrases cibles / Système *Google Translate***
Statele membre comunică Secretariatului General al Consiliului și Comisiei textul dispozițiilor care transpun în dreptul intern obligațiile pe care le impune prezenta decizie-cadru.	Les Etats membres **communiquent** au Secrétariat général du Conseil et à la Commission **le texte** des dispositions transposant dans leur droit national les obligations que leur impose la présente décision-cadre.	Les états membres **communiquent** au secrétariat général du Conseil et à la Commission **le texte** des dispositions transposant en droit national les obligations que leur impose la présente décision-cadre.	Les Etats membres **informent** le Secrétariat général du Conseil et la Commission **le texte** des dispositions transposant dans leur droit national les obligations que leur impose la présente décision-cadre.
In sensul prezentei directive , un serviciu paneuropean de comunicații mobile terestre digitale celulare **înseamnă un serviciu** public de radiotelefonie celulară asigurat în fiecare dintre statele membre în conformitate cu o specificație comună , care prevede , în special , ca toate **semnalele vocale *să fie*** codificate sub formă **de** cifre binare înainte de **a fi transmise** prin radio și care permite utilizatorilor ce **beneficiază de un serviciu** într-un stat membru **să poată** avea acces **și** la **serviciul existent** în oricare alt stat membru.	**Aux fins de** la présente directive , un service européen de communications mobiles terrestres **publiques** cellulaires numériques **signifie un service public de radiotéléphonie cellulaire *qui est assuré*** dans chacun des États membres **selon une spécification** commune ***prévoyant notamment*** que tous les signaux **vocaux sont** encodés **sous forme de** chiffres binaires **avant la transmission radio** et qui permet aux **usagers *bénéficiant d' un service*** dans un État membre **d' avoir également accès *au service*** existant dans **un autre** État membre.	Au sens de la présente directive, un service paneuropéen de communications mobiles terrestres cellulaires numériques signifie un service public de **radiotelefonie** cellulaire assuré dans chacun des états membres conformément à une spécification commune, qui prévoit, notamment, que tous_**les signaux vocale *être*** codés sous forme de chiffres binaires avant d' être **transmises** par radio et qui permet aux utilisateurs bénéficiant d'un service dans un état membre **peuvent** avoir accès **et** au service existant dans un autre état membre.	Aux fins de la présente directive, un service paneuropéen de communications cellulaires numériques terrestres mobiles**, un service** de radio cellulaire public fourni dans chacun des États membres, conformément à une spécification commune, qui stipule en particulier que tous les signaux vocaux sont encodés sous la forme **d'**chiffres binaires avant d'être transmis par la radio et permet aux utilisateurs qui **reçoivent un service** dans un État membre d'avoir accès aux **services existants** dans tout État membre.

Le premier exemple du Tableau 35 illustre une phrase source roumaine correctement traduite par notre système factorisé, tandis que la traduction fournie par *Google Translate* contient une erreur due à la synonymie. En effet, ce système propose pour le verbe *communiquer* appartenant à la collocation discontinue dans la phrase *communiquer […] le texte*, le verbe *informer* (*informer […] le texte**) qui n'est pas approprié dans le contexte donné. Le sens d'origine est pourtant conservé dans les deux traductions.

Dans le deuxième exemple, qui concerne une phrase source plus complexe que la première, les deux systèmes rencontrent plusieurs difficultés. Dans cet exemple, les deux systèmes ne commettent pas les mêmes erreurs au niveau grammatical. De plus, le sens de la phrase cible proposée par *Google Translate* s'avère, cette fois-ci, altéré tandis que la traduction fournie par le système factorisé garde le sens de la phrase source.

Le système factorisé montre ainsi des difficultés en ce qui concerne l'accord du participe passé (*les signaux […] avant d'être* ***transmises*** vs. *les signaux […] avant d'être* ***transmis***). Une autre erreur de ce système est la traduction du subjonctif par l'infinitif (*să fie* vs. *être*) ou par l'indicatif présent (*să poată* vs. *peuvent*). Même si dans d'autres contextes cette traduction peut être possible, dans ce cas elle n'est pas appropriée. En outre, le système propose comme équivalent de traduction pour l'adverbe *și* du roumain (qui signifie *également, aussi* en français) la conjonction *et*, car *și* 'et' du roumain peut être aussi une conjonction de coordination et le système n'est donc pas capable de résoudre cette ambigüité liée à la langue source. Enfin, les mots *radiotelefonie* 'radiotéléphonie' et *vocale* 'vocaux' n'ont pas été traduits. Il s'agit des mots inconnus ou moins représentatifs dans le corpus d'apprentissage.

Quant au *Google Translate*, celui-ci montre encore des problèmes dus à la synonymie. En effet, il traduit le verbe *a beneficia* 'bénéficier' appartenant à la collocation *a beneficia de un serviciu* 'beneficier d'un service' par son synonyme *recevoir*, mais ce choix s'avère inapproprié dans le contexte donné (… *utilizatorilor ce beneficiază de un serviciu* 'aux usagers bénéficiant d'un service' vs. … *aux utilisateurs qui reçoivent un service*). De plus, il omet parfois des mots de la traduction et cette omission peut altérer le sens global de la phrase, comme dans l'exemple donné à l'égard du verbe *a însemna* 'signifier' du roumain qui n'est pas traduit en langue cible (*înseamnă un serviciu* 'signifie un service' vs. , *un service*). Un signe de ponctuation qui n'apparaît pas dans la phrase source (la virgule) apparaît dans la phrase cible avant la séquence *un service* donnant lieu à une énumération qui dénature le sens

de la phrase par la suite. D'autres erreurs de nature morphologique apparaissent dans la phrase cible comme le singulier qui est traduit par le pluriel : *serviciul existent* vs. *services existants*.

Ces deux exemples de traduction automatique discutés ci-dessus montrent que le système factorisé roumain -> français a traduit mieux les phrases sources données en entrée que le système statistique pur *Google Translate*. En outre, dans le deuxième exemple, la phrase cible fournie par le système factorisé garde le sens de la phrase source, tandis que celle donnée par *Google Translate* montre un sens légèrement altéré.

Par ailleurs, à partir du deuxième exemple, nous pouvons voir les limites de la mesure d'évaluation automatique utilisée : le score *BLEU* (cf. chapitre 2, sous-section 2.2.2.4.1.). En effet, ce score est estimé à partir des n-grammes communs entre la traduction fournie par un système automatique et sa traduction de référence (effectuée par les humains). Par conséquent, des n-grammes correctement traduits par le système mais n'apparaissant pas dans la traduction de référence ne sont pas pris en compte. Or la traduction manuelle est souvent une tâche subjective et pour une phrase source donnée en entrée il peut y avoir donc plusieurs variantes de traduction, comme dans notre exemple. Ainsi, à partir de la traduction proposée par notre système et sa traduction de référence, nous avons identifié des n-grammes correctement traduits par le système mais ne faisant pas partie de la traduction de référence. Ces cas sont les suivants :

- l'utilisation des synonymes dans la traduction de référence : ***Aux fins de*** *la présente directive* vs. ***Au sens de*** *la présente directive* ; ***selon*** *une spécification* vs. ***conformément*** *à une spécification* ; *encodés* vs. *codés* ; *usagers* vs. *utilisateurs* ;

- des adjectifs participiaux traduits par des relatives dans la référence : *service […]* ***qui est assuré*** vs. *service […]* ***assuré*** ;

- des relatives traduites par des participes présents dans la référence : *spécification commune* ***prévoyant*** vs. *spécification commune,* ***qui prévoit*** ;

- des verbes traduits par les déverbaux correspondants dans la référence : *avant* ***la transmission*** *radio* vs. *avant d'****être transmises**** *par radio.* Il s'agit du procédé de traduction appelé transposition, autrement dit le changement de catégorie grammaticale dans la traduction.

Quand ces variantes de traduction proposées par le système n'apparaissent pas dans l'échantillon de référence, le score *BLEU* ne les prend pas en considération et les résultats de l'évaluation s'avèrent ainsi approximatifs, d'où l'intérêt de combiner l'évaluation automatique d'un système avec son évaluation manuelle. De ce fait, nous envisageons dans l'avenir de réaliser également une évaluation manuelle des résultats de traduction obtenus afin de proposer d'éventuelles solutions aux problèmes identifiés.

Dans la sous-section suivante, sera présentée l'évaluation automatique des systèmes de traduction développés du français vers le roumain.

5.2.2. L'évaluation des systèmes de traduction automatique du français vers le roumain

Comme il a déjà été précisé auparavant, nous avons construit les mêmes systèmes dans les deux directions de traduction automatique. Ainsi, les systèmes présentés antérieurement ont été également entraînés du français vers le roumain. Le Tableau 36 ci-dessous comprend l'évaluation de ces systèmes en termes de scores *BLEU 1* et *2*. Ceux-ci ont été construits tout d'abord par le biais de l'heuristique par défaut *grow-diag-final*.

Tableau 36. Systèmes de traduction automatique du français vers le roumain / Heuristique *grow-diag-final*

Systèmes français -> roumain	**Facteurs linguistiques / modèles de traduction**	**Facteurs linguistiques / modèles de la langue cible**	**BLEU1**	**BLEU2**
Statistique	formes des mots	formes des mots	**23,88**	**24,15**
Factorisé	lemmes et *MSD*	formes des mots et *MSD*	23,15	25,33
Factorisé + **cognats**	lemmes et *MSD*	formes des mots et *MSD*	22,83	25,83
Factorisé + **cognats** + **règles linguistiques**	lemmes et *MSD*	formes des mots et *MSD*	23,74	**25,98**

Il est clair de constater dès le début que, de manière générale, les valeurs des scores *BLEU 1* et *2* obtenues par ces systèmes français -> roumain sont significativement inférieures à celles fournies par les systèmes inverses roumain -> français présentés dans la sous-section antérieure.

Le système de base statistique pur a obtenu un score *BLEU 1* de 23,88 et un *BLEU 2* de 24,15. Ceux-ci sont inférieurs d'environ 5 points à ceux fournis par le système de base roumain -> français (*BLEU 1* de 29,38 et *BLEU 2* de 30,11 - cf. sous-section antérieure). Contrairement au système factorisé roumain -> français, le système construit vers le roumain n'améliore pas de manière significative les résultats, par rapport au système de base. Celui-ci montre même un impact négatif sur l'évolution du score *BLEU 1*. En effet, ce système factorisé déprécie le *BLEU 1* de presque un point (0,73), mais améliore le *BLEU 2* de 1,18. L'intégration des alignements des cognats diminue encore légèrement le score *BLEU 1* de 0,32, mais améliore le *BLEU 2* d'un demi-point (de 25,33 à 25,83), par rapport au système factorisé. Enfin, l'incorporation des alignements lexicaux fournis par les règles linguistiques dans le système factorisé produit une amélioration intéressante de presque un point (0,91) du *BLEU 1* et une légère amélioration de 0,15 du *BLEU 2*, par rapport au système précédent. Ainsi, l'amélioration totale du système factorisé par l'intégration des alignements fournis par les cognats et les règles linguistiques dépasse un demi-point pour les deux scores (0,59 pour le *BLEU 1* et 0,65 pour le *BLEU 2*). Par comparaison avec le système de base, le *BLEU 1* final est toutefois légèrement déprécié de 0,14, mais le *BLEU 2* final est convenablement amélioré de presque 2 points (1,83).

Le graphique apparaissant dans la Figure 69 suivante illustre l'évolution des scores *BLEU 1* et *BLEU 2* pour ces premières expériences de traduction automatique français -> roumain.

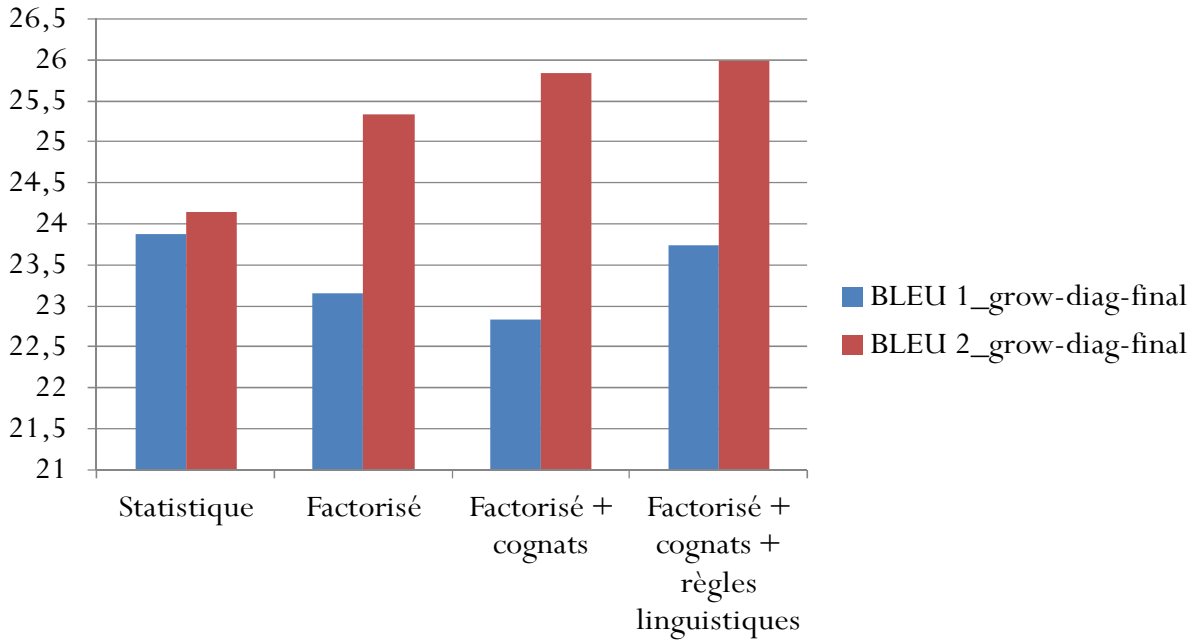


Figure 69. L'évolution des scores *BLEU 1* et *BLEU 2* des systèmes de traduction automatique du français vers le roumain / Heuristique *grow-diag-final*

Ces expériences montrent, d'un côté, que l'utilisation de facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction automatique français -> roumain améliore les résultats de la traduction automatique, par rapport à un système de base statistique pur. Cependant, cette fois-ci, cette amélioration (d'environ un point) ne s'est pas avérée très importante, comme dans le cas du système roumain -> français (cf. sous-section antérieure). De plus, cette amélioration est obtenue seulement après l'optimisation du système (voir le graphique ci-dessus). D'un autre côté, l'exploitation des cognats et des règles linguistiques peut améliorer les résultats d'un système factorisé (de plus d'un demi-point, dans ce cas), même si cette amélioration ne se montre pourtant pas très importante. Remarquons que lors de ces expériences, les deux modules supplémentaires d'alignement lexical ont amélioré successivement le *BLEU* 2 du système factorisé, à la différence du système roumain -> français où les cognats dépréciaient légèrement les résultats et seulement les règles les amélioraient (cf. sous-section antérieure). En outre, les alignements des cognats ont apprécié le *BLEU 2* convenablement d'un demi-point, tandis que les règles de seulement 0,15 points.

Nous avons aussi entraîné les systèmes de traduction automatique français -> roumain par le biais de l'heuristique *union*. Les valeurs des scores d'évaluation obtenues pendant ces expériences figurent dans le Tableau 37 suivant :

Tableau 37. Systèmes de traduction automatique du français vers le roumain / Heuristique *union*

Systèmes français -> roumain	**Facteurs linguistiques / modèles de traduction**	**Facteurs linguistiques / modèles de la langue cible**	**BLEU1**	**BLEU2**
Statistique	formes des mots	formes des mots	**23,53**	**23,68**
Factorisé	lemmes et *MSD*	formes des mots et *MSD*	22,44	25,90
Factorisé + **cognats**	lemmes et *MSD*	formes des mots et *MSD*	23,37	25,87
Factorisé + **cognats** + **règles linguistiques**	lemmes et *MSD*	formes des mots et *MSD*	**22,91**	**26,87**

Dans ces expériences, les valeurs des scores *BLEU 1* et *2* restent encore, de manière générale, significativement inférieures à celles fournies par les systèmes roumain -> français (cf. sous-section antérieure). Le système de base a un score *BLEU 1* de 23,53 et un *BLEU 2* de 23,68. Remarquons que ces scores sont légèrement inferieures à ceux du système de base français -> roumain obtenu par *grow-diag-final*. Comme dans les expériences précédentes, le système factorisé influence négativement l'évolution du score *BLEU 1* qui baisse d'environ un point (1,09), par rapport au *BLEU 1* du système de base, mais améliore le *BLEU 2* d'environ 2 points (2,22). De nouveau, l'amélioration obtenue ne s'avère pas très importante comme elle l'était dans le cas du système factorisé roumain -> français. En revanche, le système factorisé intégrant aussi les alignements des cognats montre une amélioration intéressante du score *BLEU 1* de presque un point (0,93) par rapport au système factorisé initial et une très faible dépréciation (de 0,03) en ce qui concerne le *BLEU 2.* Ensuite, le système intégrant également les alignements fournis par les règles linguistiques déprécie le *BLEU 1* de presque un demi-point (0,46) par rapport au système antérieur, mais produit une amélioration intéressante du *BLEU 2* d'exactement un point. Enfin, l'amélioration totale du système factorisé obtenue grâce aux alignements fournis par les cognats et les règles linguistiques s'avère cette fois-ci intéressante, étant de presque un demi-point (0,47) concernant le *BLEU 1* et de presque un point (0,97) quant au *BLEU 2*.

Notons que ces résultats s'avèrent aussi plus intéressants que ceux du système roumain -> français retenu, où l'amélioration du système factorisé par les cognats et les règles linguistiques était de 0,24 points pour le score *BLEU 1* et de seulement 0,14 pour le *BLEU 2* (cf. sous-section antérieure). Cependant, par comparaison avec le système de base, le score *BLEU 1* est déprécié de plus d'un demi-point (0,62), mais le *BLEU 2* est amélioré convenablement d'environ 3 points (3,19). Remarquons aussi que le système final construit par *union* montre cette fois-ci un score *BLEU 2* apprécié de manière intéressante de presque un point (0,89), par rapport au système obtenu par *grow-diag-final*.

Dans la Figure 70 suivante est donné le graphique illustrant l'évolution des scores *BLEU 1* et *BLEU 2* des systèmes français -> roumain obtenus par *union*.

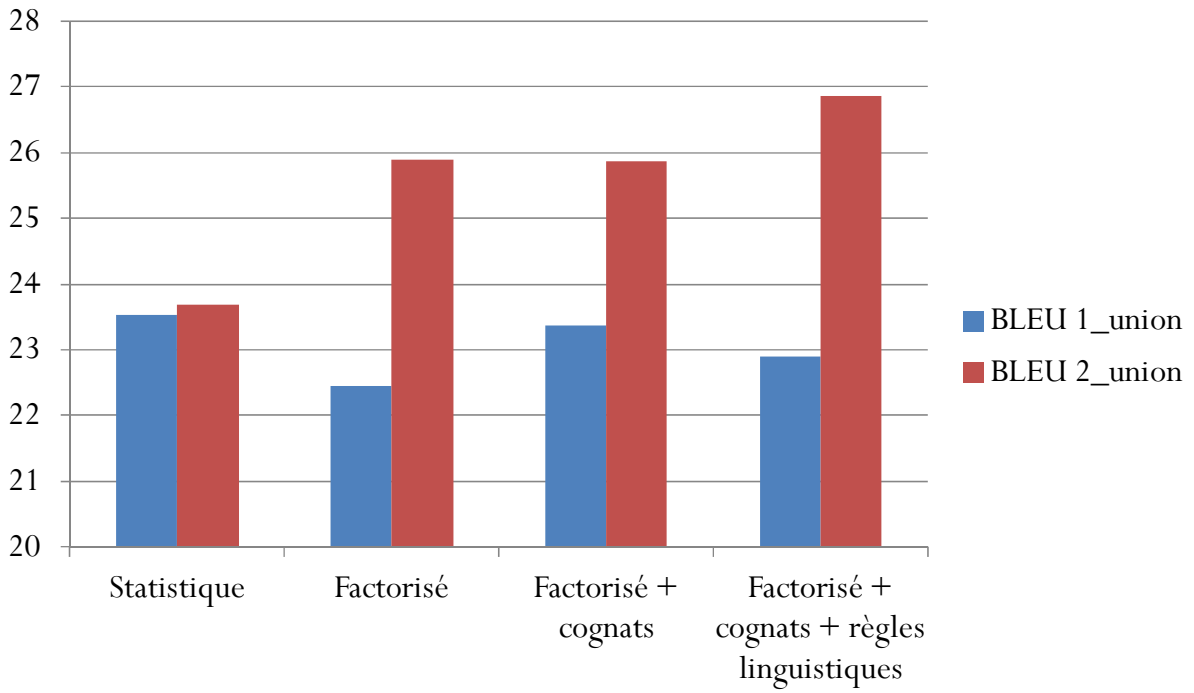


Figure 70. L'évolution des scores *BLEU 1* et *BLEU 2* des systèmes de traduction automatique du français vers le roumain / Heuristique *union*

Ce deuxième ensemble d'expériences de traduction automatique français -> roumain montre, comme lors des premières d'expériences, que l'intégration de facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction améliore les résultats par rapport à un système de base purement statistique, après l'optimisation du système. Néanmoins, cette amélioration, obtenue seulement après l'optimisation du système, ne s'avère pas très importante (2,22), surtout en comparaison avec le système factorisé roumain -> français où l'amélioration du score *BLEU 2* était d'environ 18 points (cf. sous-section

antérieure). En outre, l'intégration des alignements fournis par les cognats et les règles linguistiques peut apprécier les résultats d'un système factorisé de manière intéressante (de presque un point (0,97), après l'optimisation du système, dans nos expériences).

Notons que les valeurs des scores d'évaluation fournies dans le cadre de ces deux sets d'expériences restent cependant, de manière générale, significativement inférieures aux valeurs obtenues dans le sens inverse de la traduction (cf. sous-section antérieure). Ces résultats peuvent être dus au fait que le roumain est plus riche morphologiquement que le français. Par conséquent, le phénomène bien connu de la dispersion des données d'entraînement devient plus important en roumain qu'en français. Rappelons que ce phénomène fait référence aux répartitions variées des formes fléchies dans un corpus d'entraînement. Par exemple, le roumain présente la catégorie du cas (nominatif, accusatif, génitif, datif, vocatif) (cf. chapitre 4, sous-section 4.1.3.) qui n'existe pas en français. Celui-ci est spécifique aux noms et aux parties du discours appartenant à la classe du nom : le pronom, l'adjectif et le numéral. Il se réalise, entre autres, au moyen des désinences spécifiques qui forment des paradigmes flexionnelles riches par rapport au français. À titre d'illustration, pour les formes du nom *règlement / règlements* du français, le roumain présente comme équivalents les formes suivantes : *regulament, regulamentul, regulamentului / regulamente, regulamentele, regulamentelor.* Ces formes peuvent avoir des distributions variées dans les corpus d'apprentissage. Ainsi, celles qui sont moins fréquentes et leurs correspondants auront des probabilités de traduction faibles et ne seront pas fournies en sortie. Une solution pour résoudre en partie ce problème de dispersion des données est la lemmatisation des corpus d'entraînement. En effet, si toutes les formes fléchies d'un mot sont regroupées sous un même lemme, leur chance d'être traduites est augmentée. De plus, si les informations morphosyntaxiques sont présentes dans les corpus, le système factorisé peut fournir en sortie la variante morphologique correcte du mot cible en fonction du lemme et des propriétés morphosyntaxiques traduits au préalable. Même si nos corpus d'entraînement sont passés par la lemmatisation et l'étiquetage morphosyntaxique qui devraient résoudre en quelque sorte le problème de la dispersion des données, il semble que la taille du corpus d'entraînement pour la direction de traduction français -> roumain devrait être plus volumineuse. De ce fait, nous envisageons dans des travaux ultérieurs d'augmenter la taille du corpus d'entraînement afin de tenter une éventuelle amélioration des résultats pour cette direction de traduction.

Le graphique donné dans la Figure 71 suivante montre l'évolution de chaque score *BLEU* obtenu dans le cadre de ces deux sets d'expériences de traduction automatique français -> roumain : *BLEU 1* et *2* calculés lors des expériences basées sur l'heuristique *grow-diag-final* et *BLEU 1* et *2* fournis pendant les expériences exploitant l'heuristique *union*.

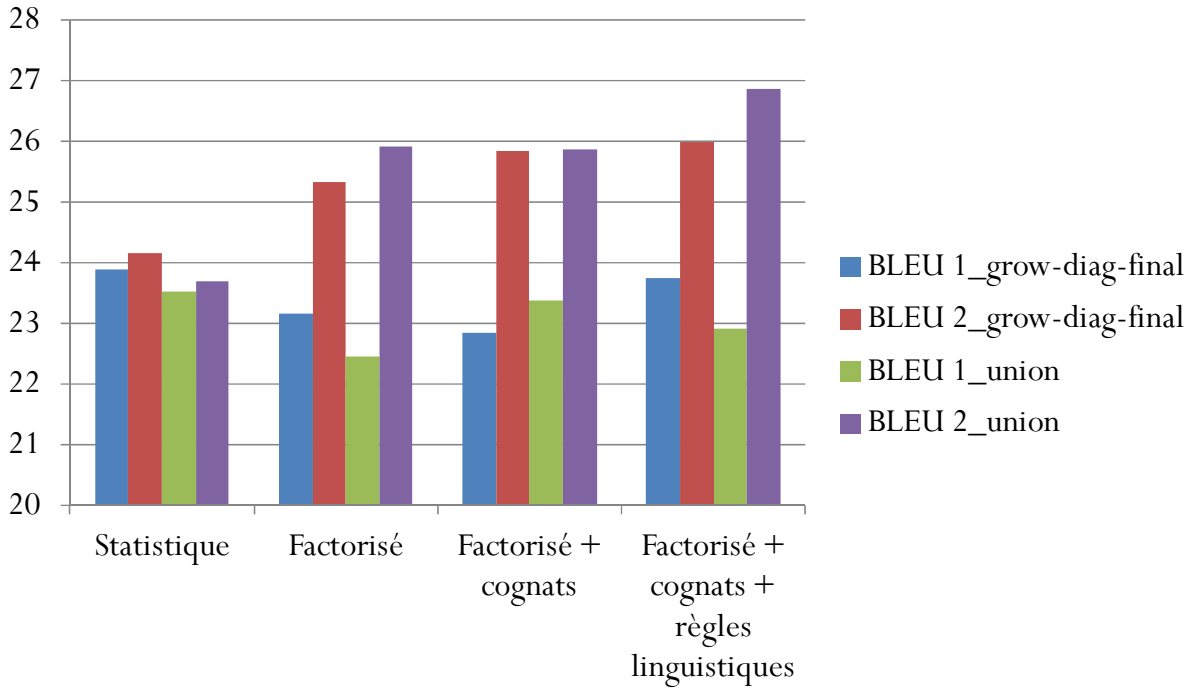


Figure 71. L'évolution des scores *BLEU 1* et *BLEU 2* des systèmes de traduction automatique du français vers le roumain / Heuristiques *grow-diag-final* et *union*

Comme il peut aussi être vu dans le graphique ci-dessus, le score *BLEU 2* optimal est fourni par le système factorisé exploitant l'heuristique *union*. Celui-ci incorpore aussi à la fois les alignements lexicaux supplémentaires fournis par les cognats et par l'application des règles linguistiques. Concernant cette direction de traduction, il semblerait que les alignements erronés ajoutés par l'heuristique *grow-diag-final* aient déprécié les résultats finaux par comparaison avec *union*, après l'optimisation du système. La différence entre les scores *BLEU 2* obtenus est cette fois-ci plus importante (0,89) que pour le système roumain -> français (0,24).

Récapitulons ci-dessous les trois cas de figure observés concernant les améliorations / dépréciations apportées par les modules supplémentaires d'alignement lexical pour la direction de traduction français -> roumain (cf. le graphique antérieur) :

1) Les alignements des cognats ont apprécié d'un demi-point les résultats du système factorisé utilisant *grow-diag-final* (après l'optimisation du système) et de presque un point (0,93) les résultats fournis par le système obtenu par *union* (avant l'optimisation du système) ; notons qu'une dépréciation non-significative (de 0,03 après l'optimisation du système) s'est observée pour le deuxième système. On pourrait donc considérer que les alignements corrects ajoutés par les cognats ont influencé positivement les deux systèmes.

2) Les alignements fournis par les règles linguistiques ont légèrement amélioré les résultats du système factorisé (+ cognats + *grow-diag-final*), après l'optimisation du système ; remarquons qu'ils ont apprécié le *BLEU 2* du système (+ cognats + *union*) de manière intéressante d'exactement un point. Les liens corrects fournis par ce module ont eu donc un impact positif sur les deux systèmes.

3) Les cognats et les règles appliqués à la fois ont amélioré les résultats des deux systèmes factorisés. Ces améliorations se sont avérées plutôt intéressantes pour cette direction de traduction français -> roumain (de 0,65 pour *grow-diag-final* et de 0,97 pour *union*).

Comme dans le cas des systèmes roumain -> français, nous pouvons en conclure, que les résultats finaux dépendent de l'heuristique utilisée mais aussi du sens de la traduction. De plus, il faut vérifier toutes les combinaisons à l'aide des deux heuristiques pour trouver la configuration optimale. Comme les scores obtenus pour les systèmes français –> roumain se sont avérés largement inférieurs à ceux des systèmes roumain -> français à cause des différences morphologiques importantes entre les deux langues, force est de constater que les résultats dépendent aussi de la taille du corpus en fonction de la direction de traduction. En effet, il semblerait que la traduction vers le roumain nécessite des données d'entraînement plus volumineuses afin d'améliorer ses résultats. En outre, étant donné que les règles linguistiques exploitées dans l'alignement lexical sont dépendantes de la paire de langues étudiées ainsi que du domaine du corpus utilisé, les résultats de traduction dépendent aussi de ces paramètres.

Le système final retenu pour la direction de traduction français -> roumain a la configuration donnée dans le Tableau 38 suivant. Celui-ci utilise les lemmes et les étiquettes *MSD* et un modèle de langue développé sur la forme des mots et un autre modèle construit sur les

étiquettes *MSD*. Étant donné que les valeurs des scores *BLEU 1* et *2* obtenues par ce système ne sont pas importantes comme dans le système construit en sens inverse (cf. sous-section antérieure), des expériences basées sur une quantité plus volumineuse du corpus d'entraînement restent encore nécessaires afin de tenter une éventuelle amélioration des résultats, comme nous l'avons déjà observé auparavant.

Tableau 38. Configuration du système de traduction automatique du français vers le roumain

Systèmes français -> roumain	**Facteurs linguistiques / modèles de traduction**	**Facteurs linguistiques / modèles de la langue cible**	**BLEU1**	**BLEU2**
Factorisé + **cognats** + **règles linguistiques** (*union*)	lemmes et *MSD*	formes des mots et *MSD*	22,91	26,87

Les résultats français -> roumain obtenus sont fortement inferieurs à ceux rapportés par Ceaușu (2009) pour le système anglais -> roumain utilisant la même configuration de facteurs linguistiques (lemmes et *MSD*), entraînée à partir d'environ 60 000 phrases alignées, comme dans notre cas. Le système anglais -> roumain n'utilise pas de modèle de distorsion construit sur les étiquettes *MSD* comme dans la direction inverse (cf. sous-section antérieure). Le score *BLEU* effectif rapporté pour ce système (Ceaușu, 2009) est de 52,31 points. En revanche, le système final exploitant le corpus entier d'entraînement (800 000 phrases alignées) obtient un score *BLEU* moins élevé de 43,29 points. De notre côté, une fois de plus, nous voyons l'intérêt d'augmenter la taille du corpus d'entraînement pour la traduction vers le roumain.

Le Tableau 39 suivant comprend les deux exemples de traduction déjà discutés dans la sous-section antérieure mais, cette fois-ci, la traduction est faite du français vers le roumain, tant par le système factorisé retenu que par *Google Translate*[140].

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

Tableau 39. Exemples de traductions français -> roumain fournies par notre système factorisé et par *Google Translate*

Phrases sources / français	**Traductions de référence**	**Phrases cibles / Notre système factorisé**	**Phrases cibles / Système *Google Translate***
Les Etats membres communiquent au Secrétariat général du Conseil et à la Commission le texte des dispositions transposant dans leur droit national les obligations que leur impose la présente décision-cadre.	Statele membre comunică Secretariatului General al Consiliului și Comisiei textul dispozițiilor care transpun în dreptul intern obligațiile pe care le impune prezenta decizie-cadru.	statele membre comunică Secretariatului General al Consiliului şi Comisiei textul dispoziţiilor care transpun în legislaţia lor internă obligaţiile pe care le impune prezenta decizie-cadru.	Comunică Secretariatului General al Consiliului şi Comisiei textul dispoziţiilor care transpun în legislaţia lor națională obligațiile care le revin în temeiul prezentei decizii-cadru, **statele membre**.
Aux fins de la présente directive , un service européen de communications mobiles terrestres publiques cellulaires numériques signifie un service public de radiotéléphonie cellulaire qui est assuré dans chacun des États membres selon une spécification commune prévoyant notamment que tous les signaux vocaux sont encodés sous forme de chiffres binaires avant la transmission radio et qui permet aux usagers bénéficiant d' un service dans un État membre d' avoir également accès au service existant dans un autre État membre.	In sensul prezentei directive , un serviciu paneuropean de comunicații mobile terestre digitale celulare înseamnă un serviciu public de radiotelefonie celulară asigurat în fiecare dintre statele membre în conformitate cu o specificație comună, care prevede , în special , ca toate semnalele vocale să fie codificate sub formă de cifre binare înainte de a fi transmise prin radio și care permite utilizatorilor ce beneficiază de un serviciu într-un stat membru să poată avea acces și la serviciul existent în oricare alt stat membru.	în sensul prezentei directive, serviciul european de comunicaţii mobile digitale celulare publice terestre înseamnă un serviciu_public de **radiotelefonică celulare** care este **asigurată** în fiecare **statelor_membre** conform unei specificaţii comune, **care prevăd** în_special că toate **semnalizările *vocaux*** să fie ***encodés*** sub formă de cifre binare înaintea **transmisia** radio şi care permite **utilizatorului *să beneficieze* într-un serviciu stat_membru** are, de_asemenea, accesul la serviciile existente într- un alt stat membru.	In sensul prezentei directive, un serviciu european de **servicii publice terestre de comunicații mobile** celulare digitale înseamnă **o rețelelor publice de celulare** este oferit în fiecare dintre statele membre **a unei specificații comune, inclusiv cerințe care toate semnalele vocale sunt codificate** în cifre binare înainte de transmisia radio și permite utilizatorilor care **primesc un serviciu** într-un stat membru de a avea, de asemenea, acces la serviciul într-un alt stat membru.

Dans le premier exemple du Tableau 39 ci-dessus, la phrase source française est correctement traduite par notre système factorisé, alors que *Google Translate*, même s'il fournit une traduction qui garde le sens de la phrase source, rencontre des difficultés pour établir un ordre

adéquat des mots cibles (le sujet *statele membre* ‘les États membres’ apparaît en fin de phrase, ce qui n’est pas spécifique au domaine juridique et administratif du corpus utilisé).

Quant au deuxième exemple pris en compte, les deux systèmes fournissent des traductions plus erronées, au niveau grammatical, que les traductions obtenues en sens inverse (cf. sous-section antérieure). Sémantiquement, la phrase cible fournie par le système factorisé a un sens légèrement altéré par rapport au sens de la phrase source, tandis que la phrase cible donnée par *Google Translate* s’avère incohérente. Cela peut être dû à la complexité de cette phrase source et, en même temps, au fait que la traduction automatique fonctionne moins bien du français vers le roumain qu’inversement, comme nous l’avons déjà observé auparavant.

Dans le cas de cet exemple, le système factorisé montre aussi des difficultés concernant l’accord du participe passé (*qui est assuré* vs. *care este* ***asigurată*** ‘qui est assurée’**)** ou du sujet avec le prédicat (*selon une* ***spécification*** *commune* ***prévoyant*** vs. *conform unei* ***specificaţii*** *comune, care* ***prevăd*** **‘**selon une **spécification** commune qui prévoient*’**).**

Une autre erreur de traduction apparaît au niveau du cas du nom en roumain, dans les situations suivantes :

- l’utilisation du nominatif / accusatif à la place du génitif (*avant la transmission radio* vs. *înaintea* ***transmisia*** *radio* à la place de *înaintea transmisiei radio*) ;

- l’emploi du génitif / datif à la place de l’accusatif (*dans chacun des États membres* vs. *în fiecare statelor_membre* au lieu de *în fiecare dintre statele membre***).**

Le système produit aussi des erreurs au niveau du participe présent qui se traduit en roumain par une relative. Le participe présent est traduit ainsi par un verbe au subjonctif à la place d’une relative, ce qui altère évidemment le sens de la phrase (*qui permet aux usagers* ***bénéficiant d****’un service* vs. *care permite utilizatorului* ***să beneficieze într****-un serviciu* à la place de *care permite utilizatorilor ce beneficiază de un serviciu*). De plus, la préposition *într-* ‘dans’ du roumain qui aurait dû précéder le nom suivant dans la phrase (***într****-un stat membru* ‘dans un État membre’) apparaît devant le nom *serviciu* ‘service’ et l’ordre des mots s’avère donc erroné.

D'autres erreurs apparaissent à cause de la synonymie dans le corpus d’apprentissage. Dans cet exemple, le système propose pour le nom *signaux* un équivalent erroné pour le contexte

donné *semnalizările* ‘signalisations’ à la place de *semnale* ‘signaux’. En effet, dans le corpus d’apprentissage il existe des constructions comme *signal gestuel* traduite par *gest de semnalizare* ‘geste de signalisation’ et le système apprend donc les équivalents *signal - semnalizare* ‘signalisation’ qui semblent avoir une probabilité de traduction plus élevée que les correspondants *signal - semnal* ‘signal’ présents aussi dans le corpus d’entraînement.

Nous avons aussi relevé une erreur due au procédé de traduction transposition (le changement de catégorie grammaticale) et au déficit des données dans le corpus d’entraînement. En effet, les constructions comme *installation de radiotéléphonie* existantes dans le corpus sont traduites en roumain comme *instalație radiotelefonică* ‘installation radiotéléphonique’ et le système apprend donc les équivalents *radiotéléphonie - radiotelefonică*, qui ne sont pas erronés en soi, car il est connu qu’un nom peut être traduit par un verbe ou un adjectif, par exemple. De plus, il n’y a pas de contextes dans le corpus d’apprentissage où le nom *radiotéléphonie* est traduit en roumain par *radiotelefonie*. C’est aussi pour cette raison que, dans l’autre sens de la traduction (cf. sous-section antérieure - Tableau 35), le nom *radiotelefonie* du roumain est traité comme un mot inconnu et n’apparaît pas traduit en sortie. Le système propose donc pour la séquence *un service public de radiotéléphonie cellulaire* la traduction *un serviciu public de radiotelefonică celulare* ‘un service public de radiotéléphonique cellulaires’ qui s’avère inappropriée dans ce contexte puisqu’elle modifie en quelque sorte le sens d’origine.

Des erreurs de nature morphologique comme le pluriel qui est traduit par le singulier ou inversement apparaissent également : *aux usagers* vs. *utilizatorului* ; *service existant* vs. *serviciile existente*.

Enfin, notons que les mots *vocaux* et *encodés* n’ont pas été traduits, car ils sont des mots inconnus ou moins représentatifs dans le corpus d’entraînement.

Quant au système *Google Translate*, celui-ci montre aussi des problèmes dus à la synonymie, tout comme dans le sens inverse de la traduction (cf. sous-section antérieure). En effet, celui-ci traduit, par exemple, le verbe *bénéficier* appartenant à la collocation *bénéficier d’un service* par son synonyme *a primi* ‘recevoir’ qui est inapproprié dans le contexte donné (… *aux usagers bénéficiant d’un service* vs. … *utilizatorilor care* ***primesc*** *un serviciu*).

En outre, *Google Translate* produit dans la langue cible des séquences plus ou moins cohérentes qui dénaturent le sens global de la phrase. Ces séquences sont mises en gras dans les exemples suivants :

- *un service européen de communications mobiles terrestres publiques cellulaires numériques* vs. *un serviciu european de* ***servicii publice terestre de comunicații mobile celulare digitale*** ‘un service européen de **services publics terrestres de communications mobiles cellulaires numériques**’ ;

- *signifie un service public de radiotéléphonie cellulaire* vs. *înseamnă* ***o rețelelor publice de celulare*** ‘signifie **une des réseaux publiques de cellulaires’** ;

- *selon une spécification commune prévoyant notamment que tous les signaux vocaux sont encodés* vs. ***a unei specificații comune, inclusiv cerințe care*** *toate semnalele vocale sunt codificate* ‘**d’une spécification commune, inclusivement des demandes qui** tous les signaux vocaux sont codifiés’.

Les deux exemples de traduction automatique discutés ci-dessous montrent que le système factorisé construit du français vers le roumain a traduit mieux les phrases sources données en entrée, par rapport au système statistique pur *Google Translate*. Même si pour le premier exemple les deux systèmes ont donné des phrases cibles gardant le sens de la phrase source, dans le deuxième exemple, concernant une phrase plus complexe, la phrase cible proposée par le système factorisé est plus cohérente que celle fournie par *Google Translate*. Par ailleurs, les résultats de traduction du français vers le roumain sont moins bons que ceux obtenus en sens inverse (cf. sous-section antérieure).

5.3. Bilan du chapitre

Ce chapitre a été dédié aux expériences de traduction automatique que nous avons effectuées pour la paire de langues français - roumain, dans les deux sens du processus de traduction.

Dans le cadre de nos expériences, nous avons cherché tout d’abord à vérifier si l’intégration de facteurs linguistiques au processus de traduction (Koehn et Hoang, 2007 ; Avramidis et Koehn, 2008 ; Tufiș *et al.*, 2008b ; Ceaușu, 2009 ; Ceaușu et Tufiș, 2011) améliore les résultats de la traduction automatique statistique, en ce qui concerne la paire de langues étudiées.

En outre, nous avons tenté de vérifier encore deux constats liés au rapport entre un alignement lexical performant et les résultats de la traduction automatique statistique, pour le roumain et le français. D'une part, nous sommes partis de l'idée qu'un alignement lexical exploitant aussi bien des techniques statistiques que des informations linguistiques (Tiedemann, 2003 ; Cherry et Lin, 2003 ; Tufiș *et al.*, 2006 ; Schrader, 2006 ; Hermjakob, 2009 ; Cendejas *et al.*, 2009 ; Pal *et al.*, 2013) donne de meilleurs résultats par rapport à un alignement lexical basé uniquement sur les statistiques (Brown *et al.*, 1990 ; Brown *et al.*, 1993) et, par conséquent, améliore aussi les résultats de la traduction automatique statistique. D'autre part, nous avons également pris en considération que le lien entre les résultats de l'alignement et ceux de traduction n'est pas clairement prouvé. Parfois, on constate une faible amélioration des résultats de traduction (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007).

Ainsi, nous avons développé un système d'alignement lexical hybride, autrement dit un système qui combine des techniques statistiques et des informations linguistiques, dont l'amélioration par rapport à un système purement statistique a été illustrée par un score *AER* diminué de manière significative d'environ 7% (cf. chapitre 4, section 4.5.). Ensuite, nous avons donc évalué l'impact de cet alignement lexical sur les résultats du système statistique factorisé développé. De plus, comme ce système d'alignement lexical intègre deux modules supplémentaires exploitant des cognats et des règles heuristiques linguistiques (cf. chapitre 4), nous avons aussi tenté d'étudier l'influence de chaque module supplémentaire sur les résultats de la traduction automatique statistique factorisée, en ce qui concerne la paire de langues étudiées.

Afin d'effectuer nos expériences de traduction automatique, nous avons utilisé la distribution standard du décodeur *MOSES* (Koehn *et al.*, 2007) (cf. chapitre 3, sous-section 3.1.1.) qui permet l'entraînement de systèmes de traduction automatique statistiques purs et factorisés, leur évaluation et leur optimisation.

Nos données d'entraînement et de test disponibles dans le cadre de ces premières expériences de traduction automatique français - roumain comprennent :

- un corpus parallèle juridique aligné au niveau propositionnel et lexical (cf. chapitre 4), étant aussi lemmatisé, annoté par des étiquettes des parties du discours (*C-tag*) (Tufiș, 1999, 2000) et des propriétés morphosyntaxiques (*MSD*) ; Ce corpus contient 64 923 couples de phrases alignées (environ un million et demi de tokens par langue). Sa

taille est limitée afin d'optimiser le temps d'entraînement. Une autre raison pour laquelle nous avons limité la taille du corpus réside dans le fait que l'étiquetage et l'alignement lexical selon notre approche hybride se sont avérés des étapes coûteuses en temps et en ressources humaines. Nous avons donc considéré que cette quantité de données est suffisante dans le cadre de ces premières expériences de traduction automatique français - roumain, pour déterminer les paramètres des systèmes développés.

- un corpus de test et un corpus de développement comprenant chacun 300 couples de phrases alignées, lemmatisées et annotées ;

- un corpus monolingue français du domaine juridique (480 764 phrases tirées de *JRC-Acquis* (Steinberger *et al*., 2006)), lemmatisé et annoté par des étiquettes *MSD* (cf. chapitre 3, sous-section 3.2.1.) ; À partir de ce corpus, nous avons entraîné seulement deux modèles de langue 5-grammes pour le français en tant que langue cible : un modèle basé sur la forme des mots et un modèle construit sur les étiquettes *MSD.*

- deux modèles de langue 5-grammes déjà disponibles pour le roumain en tant que langue cible (Tufiș *et al.*, 2013a) (cf. section 5.1. de ce chapitre) : un modèle développé sur la forme des mots et un autre modèle exploitant les étiquettes *MSD* ; Ceux-ci sont entraînes à partir des corpus juridiques.

Ainsi, en fonction des ressources mentionnées ci-dessus, 8 systèmes développés dans les deux directions de la traduction (16 en total) sont devenus disponibles, suite à l'étape d'entraînement. Pendant cette étape, deux heuristiques ont été utilisées afin de construire les modèles de traduction statistiques et factorisés : l'heuristique par défaut dans *MOSES (*appelée *grow-diag-final*) et celle dénommée *union*. La première exploite tant l'intersection que l'union des alignements lexicaux bidirectionnels, alors que la deuxième utilise seulement l'union de ces alignements (Koehn *et al.*, 2003). Les systèmes développés sont les suivants :

- un système statistique ; Celui-ci est basé sur la forme des mots et utilise un modèle de la langue cible développé également sur la forme des mots. Celui-ci est construit au moyen des deux heuristiques utilisées (*grow-diag-final* et *union*). Il est considéré comme base pour nos résultats comparatifs.

- un premier système factorisé utilisant les lemmes et les étiquettes *MSD* ; Celui-ci exploite un modèle de langue construit sur la forme des mots et un autre modèle entraîné sur les étiquettes *MSD*. Cette configuration est obtenue par le biais des deux heuristiques mentionnées et donne ainsi naissance a deux systèmes différents.

- un deuxième système factorisé ayant la même con*fig*uration que le précédent mais incorporant aussi les alignements lexicaux additionnels fournis par le module d'identification et d'alignement automatique des cognats ; Cette configuration est aussi entraînée par le biais des deux heuristiques prises en compte.

- un troisième système factorisé présentant la même configuration que le premier et le deuxième mais intégrant aussi les alignements lexicaux supplémentaires fournis par les cognats et par le module implémentant la base de règles linguistiques constituée. Dans ce cas, les deux heuristiques sont également utilisées.

Après les étapes d'évaluation des systèmes, ont été identifiés ceux ayant obtenu les meilleurs scores *BLEU 1* (le score *BLEU* avant l'optimisation du système) et *BLEU 2* (le score *BLEU* après son optimisation), par rapport au système de base considéré. Ainsi, les systèmes suivants ont été retenus :

- Pour la direction de traduction du roumain vers le français, le système obtenu est factorisé et incorpore dans le processus de traduction les lemmes et les étiquettes *MSD*. Celui-ci utilise un modèle de langue basé sur la forme des mots et un autre modèle construit sur les étiquettes *MSD*. Il utilise l'heuristique par défaut *grow-diag-final*. De plus, il intègre aussi les alignements lexicaux fournis par les cognats et l'application des règles linguistiques. Ce système obtient de scores *BLEU 1* et *2* performants : 47,29 points pour le *BLEU 1* et 48,48 pour le *BLEU 2* respectivement.

- Quant à la direction de traduction du français vers le roumain, le système retenu est aussi factorisé et intègre dans le processus de traduction les lemmes et les étiquettes *MSD*. Celui-ci exploite un modèle de langue entraîné sur la forme des mots et un autre modèle développé sur les étiquettes *MSD*. Contrairement au système roumain -> français retenu, celui-ci utilise l'heuristique *union*. En outre, il intègre également les alignements lexicaux fournis par les cognats et l'application des règles linguistiques. Ce système obtient des scores *BLEU 1* et *2* significativement inférieurs à ceux fournis

par le système roumain -> français. Son score *BLEU 1* est de 22,91 points, tandis que son score *BLEU 2* s'élève à 26,87 points.

Les valeurs des scores *BLEU 1* et *BLEU 2* des systèmes retenus montrent, d'une part, que les résultats de traduction dépendent de l'heuristique utilisée mais aussi de la direction de traduction. D'autre part, ces résultats prouvent que, pour la même quantité de données d'entraînement, la traduction automatique roumain -> français fonctionne beaucoup mieux que la traduction automatique français -> roumain (il y a environ 24 points de décalage entre les scores *BLEU 1* et 21 points entre les scores *BLEU 2* obtenus). Ce phénomène pourrait être expliqué par le fait que le roumain est une langue plus riche morphologiquement que le français (par exemple, le roumain présente la déclinaison selon la catégorie du cas, ce qui ne se passe pas en français car le cas n'existe pas dans cette langue). Par conséquent, en roumain, la dispersion des données est plus importante qu'en français et même si nos corpus d'entraînement sont lemmatisés et étiquetés par des propriétés morphosyntaxiques (la lemmatisation et l'étiquetage étant supposés de résoudre le problème de la dispersion), cela ne peut substituer le besoin d'une quantité de données plus importante en ce qui concerne la direction de traduction français -> roumain. Nous pouvons donc en conclure que les résultats de la traduction automatique statistique factorisée français - roumain dépendent aussi du volume des données d'entraînement, en fonction de la direction de traduction. De ce fait, nous envisageons dans nos expériences ultérieures d'augmenter la taille du corpus parallèle d'entraînement pour les systèmes français -> roumain afin de tenter d'améliorer éventuellement les résultats en termes de score *BLEU*. Ces résultats dépendent également de la paire de langues étudiées et du domaine du corpus (les règles linguistiques bilingues exploitées dans l'alignement lexical ont été définies à partir des corpus juridiques et administratifs).

Les expériences effectuées vers le français montrent aussi que l'intégration de facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction a amélioré de manière significative les résultats de la traduction automatique factorisée, par rapport à un système de base purement statistique (de 17,67 points le score *BLEU 1* et de 18,23 le *BLEU 2*, en ce qui concerne les résultats du système retenu). En revanche, dans le sens inverse de la traduction, c'est-à-dire du français vers le roumain, cette amélioration ne s'est pas avérée très importante, étant de 2,22 après l'optimisation du système. De plus, une amélioration significative du système d'alignement lexical, par l'incorporation des

informations linguistiques dans le processus d'alignement, peut améliorer légèrement les résultats d'un système factorisé roumain -> français (de 0,24 points le score *BLEU 1* et de 0,14 le *BLEU 2*, en ce qui concerne le système retenu). Toutefois, dans le sens inverse de la traduction, du français vers le roumain, le système factorisé retenu, intégrant à la fois les alignements lexicaux additionnels fournis par les cognats et l'application des règles linguistiques, a amélioré de manière intéressante les résultats (de presque un point (0,97) le score *BLEU*, après l'optimisation du système).

Quant aux modules supplémentaires des cognats et des règles linguistiques appliqués séparément, ceux-ci peuvent améliorer ou déprécier les scores des systèmes factorisés, en fonction de l'heuristique utilisée mais aussi du sens de la traduction. De ce fait, il convient de les tester toujours séparément pour trouver la configuration qui se comporte le mieux pour chaque ensemble d'expériences.

La configuration du système factorisé basé sur les lemmes qui utilise un modèle de langue exploitant la forme des mots reste encore à implémenter et à évaluer. En outre, nous envisageons dans l'avenir d'entraîner des modèles de langue sur les parties du discours et pouvoir ainsi développer des systèmes factorisés construits sur ce facteur.

Nous sommes également conscients des limites de la mesure d'évaluation utilisé : le score *BLEU* (Papineni *et al*., 2002) (cf. chapitre 2, sous-section 2.2.2.4.1.). En effet, comme celui-ci fonctionne par la comparaison d'une traduction fournie par le système avec l'une de référence, ce score peut éliminer du calcul des n-grammes correctement traduites automatiquement mais ne faisant pas partie de l'échantillon de référence (dans le cas où une phrase source a plusieurs variantes de traduction humaine). De ce fait, une évaluation manuelle reste toujours nécessaire afin de rendre réellement compte de la qualité d'une traduction automatique. Ainsi, nous projetons de réaliser une évaluation manuelle des résultats de traduction proposés par les systèmes finaux, afin de repérer les erreurs de traduction, les classifier et tenter de proposer d'éventuelles solutions de correction.

Nous compléterons aussi l'évaluation automatique par l'utilisation d'autres mesures comme *NIST*[141] (Doddington, 2002) et *METEOR*[142] (Banerjee et Lavie, 2005). *NIST* représente une variante du score *BLEU* (Papineni *et al*., 2002). Si le *BLEU* est censé mesurer la fluence de la

[141] *National Institute of Standards and Technology*
[142] *Metric for Evaluation of Translation with Explicit ORdering*

traduction par rapport à une traduction de référence, le score *NIST* mesure plutôt son adéquation. À la différence de ces deux scores, *METEOR* est capable de prendre en compte aussi les synonymes des n-grammes à comparer, en augmentant ainsi l'efficacité de l'évaluation automatique.

Notre finalité ultime est d'étendre le système développé par l'utilisation du corpus parallèle d'entraînement entier disponible pour la paire de langues étudiées (comprenant environ 30 millions de mots - cf. chapitre 3), mais aussi d'étendre les expériences pour des modèles de langue construits pour d'autres domaines.

6. Conclusions et perspectives

Nous avons présenté ici un projet de recherche ayant comme objectif, d'une part, la constitution de ressources linguistiques pour un système de traduction automatique statistique factorisée français - roumain. D'autre part, ce projet vise l'étude de l'influence des informations linguistiques exploitées (lemmes, propriétés morphosyntaxiques), tant dans le processus d'alignement lexical que dans celui de traduction.

Les ressources linguistiques requises par les systèmes factorisés sont des corpus parallèles lemmatisés, étiquetés et alignés au niveau propositionnel et lexical. Ainsi, nous avons tout d'abord prétraité le corpus bilingue parallèle utilisé. Ensuite, nous nous sommes concentrés sur le développement d'une méthode d'alignement lexical adaptée à la paire de langues étudiées car, à notre connaissance, il n'existe pas de systèmes d'alignement disponibles, conçus spécialement pour le français et le roumain.

Nous avons développé une méthode hybride d'alignement lexical qui combine des techniques statistiques et des informations linguistiques (Tiedemann, 2003 ; Cherry et Lin, 2003 ; Tufiș *et al.*, 2006 ; Schrader, 2006 ; Hermjakob, 2009 ; Cendejas *et al*., 2009 ; Pal *et al*., 2013), afin d'améliorer les résultats d'un système de base purement statistique (Brown *et al*., 1993). L'amélioration de l'alignement est une étape importante car un alignement performant améliore la qualité de la traduction automatique. Même si certaines approches (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007) montrent que le rapport entre la qualité de l'alignement et celle de la traduction n'est pas clairement démontré (les résultats de traduction ne sont pas toujours appréciés de manière significative), nous avons amélioré l'alignement lexical français - roumain afin d'étudier son influence sur les résultats de la traduction automatique pour cette paire de langues.

Cette méthode intègre un module de détection et d'alignement automatique de cognats et une base de règles heuristiques linguistiques définie manuellement à partir de l'étude des erreurs fournies par le système de base statistique pur, obtenu à l'aide de *GIZA++* (Och et Ney, 2000, 2003). Ces modules supplémentaires ont contribué de manière significative à l'amélioration de l'alignement lexical par rapport au système de base (le score *AER* a baissé d'environ 7%).

Nous avons développé une nouvelle méthode de détection et d'alignement automatique de cognats qui combine des techniques statistiques (méthodes n-grammes (Simard et *al.*, 1992), calcul des fréquences, etc.) et des informations linguistiques (lemmes, parties du discours, équivalences de catégorie lexicale). De plus, comme un nouvel point, le module applique initialement un ensemble de règles d'ajustements orthographiques sensitives ou non sensitives au contexte phonétique, au niveau des lemmes. Cette étape a apprécié le rappel d'une simple méthode 4-grammes (Simard et *al.*, 1992) d'environ 25%. Concernant l'alignement des cognats, ceux-ci ont amélioré le score *AER* convenablement d'approximativement 2%, par rapport au système de base. L'originalité de cette méthode consiste dans la combinaison de techniques utilisée afin d'améliorer les résultats par rapport à des méthodes traditionnelles statistiques pures comme la méthode 4-grammes (Simard et *al.*, 1992), le coefficient de *DICE* (Adamson et Boreham, 1974) ou la mesure *SCM* (sous-chaîne maximale) (Melamed, 1999).

Toutefois, les résultats fournis ne peuvent pas être généralisés car ils dépendent des langues étudiées, du domaine et de la taille du corpus de test utilisé, ainsi que de la qualité de l'étiquetage du corpus. De ce fait, des études similaires restent encore nécessaires sur d'autres corpus de test. En outre, la méthode rencontre encore des difficultés dans le cas des candidats ambigus, des formes hapax ou des candidats ayant des similarités orthographiques et/ou phonétiques très faibles. Celle-ci a néanmoins l'avantage d'être facilement adaptable à d'autres paires de langues apparentées, puisqu'elle ne requiert pas de ressources externes telles que des lexiques, des dictionnaires ou des listes de cognats validées par des experts.

Comme le module présente l'inconvénient d'être évalué sur un seul corpus de test, afin de rendre compte de sa fiabilité avec plus de précision, nous l'avons aussi testé directement dans les systèmes de traduction factorisés entraînés par *grow-diag-final* et *union*, dans les deux sens du processus de traduction. Notons que la construction manuelle de listes de cognats et de corpus de référence nécessaires pour l'évaluation du module et de l'alignement lexical est une tâche coûteuse en temps et en personnels. De plus, il n'existe pas, à notre connaissance, de corpus de référence (*gold standard*) disponibles pour la paire de langues étudiées.

Dans la direction de traduction roumain -> français, les cognats identifiés ont légèrement déprécié (0,10 points *BLEU 2*) les résultats du système factorisé (+*grow-diag-final*), mais ont apprécie de manière similaire ceux du système factorisé (+*union*) (0,36 points *BLEU 2*).

Par contre, dans la direction français -> roumain, les cognats ont apprécié les résultats du système factorisé (+*grow-diag-final*) de manière intéressante de 0,50 points *BLEU 2* et ceux du système factorisé (+*union*) de presque un point avant son optimisation (0,93 points *BLEU 1*). Une très faible dépréciation du *BLEU 2* est survenue dans le cas du deuxième système (0,03).

Nous pouvons en conclure que les alignements fournis par le module des cognats peuvent améliorer ou même déprécier les résultats d'un système factorisé français <-> roumain, en fonction de l'heuristique utilisée mais aussi du sens de la traduction. De ce fait, il convient de tester toutes les combinaisons pour retenir celle qui se comporte le mieux. Les variations des scores ne se sont néanmoins pas montrées très significatives. En outre, les cognats ont eu une influence plus intéressante pour les systèmes construits vers le roumain.

Le module à base de règles heuristiques linguistiques appliqué après l'identification et l'alignement automatiques des cognats a contribué à l'amélioration de l'alignement lexical de manière significative d'environ 5%. Les résultats obtenus sont pourtant dépendants de la paire de langues étudiées, du volume et du domaine du corpus de test exploité, de la qualité de son étiquetage mais aussi des annotations d'alignement présentes dans le corpus de référence. L'utilisation d'autres corpus de test est ainsi nécessaire afin de pouvoir généraliser.

La base de règles heuristiques constituée présente aussi l'inconvénient d'être limitée a seulement 37 règles d'alignement lexical définies à partir du corpus de test utilisé. Ainsi, afin d'enrichir cette base et de cibler ainsi d'éventuelles améliorations des résultats d'alignement, nous envisageons dans l'avenir d'utiliser d'autres corpus de test. De plus, comme cette tâche de définition manuelle de règles basée sur les connaissances linguistiques de type contrastif s'avère néanmoins fastidieuse, nous projetons également de combiner cette étape avec l'apprentissage automatique des règles (Ozdowska, 2006). Mais, pour ce faire, des corpus étiquetés et alignés lexicalement sont nécessaires. Notons que nous n'avons pas disposé initialement de telles ressources pour la paire de langues étudiées.

Les règles définies manuellement ont toutefois l'avantage d'être précises du point de vue grammatical à la différence des règles apprises automatiquement. Ces dernières sont pourtant utiles car elles sont adaptables et portables entre les paires de langues ou entre les corpus utilisés (Ozdowska, 2006). À part le fait que l'apprentissage automatique offre la possibilité de collecter un ensemble de règles d'alignement plus élevé par rapport à la définition

manuelle, il est capable de modéliser aussi les erreurs systématiques d'étiquetage et donc les règles qui en résultent sont plus conformes avec les corpus utilisés. Nous saisissons ainsi l'utilité de combiner les deux techniques à l'avenir.

L'idée d'allier des méthodes statistiques et des règles heuristiques ayant un fondement linguistique (Schrader, 2006 ; Tufiș *et al.*, 2006 ; Hermjakob, 2009 ; Cendejas *et al.*, 2009 ; Pal *et al.*, 2013) n'est pas originale. Cependant, la base de règles définies est spécifique à la paire de langues traitées et, à notre connaissance, il n'existe pas actuellement de telles ressources qui puissent être exploitées pendant l'alignement lexical pour le français et le roumain. Cette base peut également être utilisée comme guide d'alignement manuel afin de constituer d'autres corpus de référence. Ces règles sont aussi adaptables à d'autres paires de langues incluant le français ou le roumain dès que des corpus alignés sont disponibles. En effet, en connaissant les structures morphosyntaxiques qui posent problème à l'alignement lexical dans l'une de ces deux langues, il est possible de repérer les structures équivalentes dans d'autres langues à partir de corpus alignés, pour le cas où ces structures sont différentes et nécessitent donc la définition d'une règle d'alignement.

La nouveauté de cette étape concerne la façon de constituer cette base de règles linguistiques et notamment à partir de l'analyse linguistique des erreurs d'alignement lexical fournies par l'aligneur utilisé initialement *GIZA++* (Och et Ney, 2000, 2003). En outre, une étude du corpus parallèle au niveau de la traduction humaine a été également effectuée afin de définir des règles d'alignement lexical stylistiques. L'originalité de cette démarche consiste dans la définition de ces règles qui sont basées sur la connaissance du style juridique du corpus utilisé et des procédés usités en traduction humaine. Il s'agit de certaines règles définies au niveau des procédés comme l'étoffement ou le dépouillement, par la collecte des lexèmes spécifiques avec renforcement de sens à valeur stylistique (certains adjectifs qualificatifs, certains verbes et adverbes, etc.).

Notons que pour certains types d'erreurs d'alignement lexical nous n'avons pas pu définir de règles de correction, comme c'est le cas des erreurs apparues au niveau des pronoms adverbiaux *en* et *y* (qui n'ont pas d'équivalents de traduction en roumain) ou au niveau des paraphrases, par exemple. Nous projetons dans l'avenir de nous pencher sur l'étude de ces aspects et de tenter éventuellement la mise en place de techniques de modélisation du phénomène anaphorique au sein des phrases parallèles mais aussi d'identification des paraphrases.

Pour les mêmes raisons que dans le cas des cognats, nous avons évalué l'influence du module à base de règles directement dans les systèmes factorisés développés dans les deux sens de la traduction.

Dans la direction de traduction roumain -> français, les alignements lexicaux donnés par le module à base de règles ont légèrement amélioré les résultats du système factorisé (+cognats +*grow-diag-final*) (0,24 points *BLEU 2*), mais ont déprécié faiblement ceux fournis par le système (+ cognats + *union*) (0,11 points *BLEU 2*).

En sens inverse de la traduction (français -> roumain), ces alignements ont légèrement apprécié les résultats du système factorisé (+cognats +*grow-diag-final*) (0,15 points *BLEU 2*) et de manière intéressante d'exactement 1 point le *BLEU 2* du système (+ cognats + *union*).

Comme dans le cas des cognats, nous pouvons en conclure que les alignements fournis par le module des règles peuvent améliorer ou même déprécier les résultats d'un système factorisé français <-> roumain, en fonction de l'heuristique exploitée mais aussi du sens de la traduction. De ce fait, il convient aussi de tester toutes les configurations afin de choisir celle qui fonctionne le mieux. Dans ce cas, les variations des scores ne se sont avérées non plus très significatives. Ces alignements ont eu cependant un impact plus intéressant pour le système entraîné vers le roumain à l'aide de l'heuristique *union*.

Étant donné que les collocations posent encore problème aux systèmes de traduction automatique statistique (Ren *et al*., 2009 ; Bouamor *et al*., 2012 ; Ramisch *et al*., 2013), nous avons tenté d'intégrer un dictionnaire de collocations *Verbe + Nom* disponible pour le français et le roumain (Todirașcu *et al*., 2008) dans le processus d'alignement. De plus, nous avons enrichi ce dictionnaire avec les collocations nominales correspondantes. Pour ce faire, nous avons mis en place une méthode de génération automatique de ces collocations à partir des entrées du dictionnaire. Cette méthode exploite le lexique *VerbAction* (Hathout *et al*., 2002 ; Tanguy et Hathout, 2002) pour le français et la morphologie des déverbaux pour le roumain. L'originalité de cette étape consiste dans la combinaison de ressources et d'indices morphologiques afin d'automatiser la collecte des collocations nominales bilingues à partir du dictionnaire. Cette démarche a été proposée comme une alternative moins coûteuse en temps et en personnels par rapport à la définition manuelle des équivalences.

Toutefois, l'évaluation n'a pas montré d'améliorations significatives et le dictionnaire n'apparaît plus dans le système d'alignement lexical final. Cela est dû au fait que les collocations communes entre le dictionnaire et le corpus de test utilisé sont peu nombreuses. Ainsi, nous projetons dans l'avenir d'effectuer des expériences sur d'autres corpus de test (de taille plus importante) pour pouvoir étudier plus précisément l'impact du dictionnaire sur les résultats de l'alignement lexical français - roumain. Nous enrichirons aussi le dictionnaire avec l'ensemble des collocations verbo-nominales supplémentaires (environ 230 collocations par langue) que nous avons validées manuellement pendant la constitution du corpus de référence, pour tenter d'éventuelles améliorations des résultats d'alignement. De plus, nous envisageons également d'intégrer les équivalents collocationnels (tels que fournis par le dictionnaire - cf. sous-section 4.4.2.2.) directement dans les tables de traduction, munis de probabilités artificielles, ou dans le corpus d'entraînement (Ren *et al*., 2009 ; Bouamor *et al*., 2012). De même, dans le but d'améliorer encore les résultats de la traduction, nous intégrerons dans le corpus d'entraînement un ensemble de termes bilingues (Tufiș et Dumitrescu, 2012 ; Weller *et al*., 2014) fournis par le thésaurus multilingue *Eurovoc*[143]. En outre, nous tenterons d'extraire et d'exploiter aussi dans le corpus d'apprentissage la partie français - roumain de l'ontologie lexicale *WordNet*.

Le système d'alignement lexical ainsi mis en place présente l'inconvénient d'être dépendant de la paire de langues traitées. Mais, comme il s'agit d'un système modulaire, il est facilement adaptable à d'autres paires de langues apparentées. Les résultats obtenus dépendent aussi du corpus de test utilisé. Ainsi, afin de généraliser nous projetons d'appliquer la méthodologie proposée sur d'autres corpus de test. Mais, pour ce faire, des corpus de référence sont nécessaires. Pour cette étude, nous avons construit seulement un corpus de référence (en deux variantes : avec et sans l'alignement des collocations), puisque le développement de ces ressources s'est avéré une tâche coûteuse en temps et en personnels.

Nous avons intégré l'alignement lexical développé dans les systèmes factorisés construits dans les deux sens du processus de traduction, afin d'étudier son impact sur la qualité des traductions fournies en sortie. Ces systèmes factorisés sont basés sur les lemmes et les propriétés morphosyntaxiques et ils sont entraînés au moyen des heuristiques *grow-diag-final* et *union*.

[143] http://eurovoc.europa.eu/

Les systèmes retenus comprennent les alignements supplémentaires des cognats et des règles heuristiques linguistiques. Le système vers le français est entraîné par *grow-diag-final* et montre un score *BLEU 2* performant (48,48 points). En revanche, le système vers le roumain est construit par *union* et son score *BLEU 2* (26,87 points) s'avère significativement inferieur à celui obtenu par le système développé en sens inverse.

Nous avons étudié tout d'abord l'influence des facteurs linguistiques (lemmes, propriétés morphosyntaxiques) dans le processus de traduction, par rapport à un système de base purement statistique. Les expériences effectuées vers le français ont montré ainsi des résultats améliorés de manière significative (d'environ 18 points pour le *BLEU 2*), par comparaison avec le système de base. Par contre, les résultats obtenus du français vers le roumain, ont été aussi appréciés par rapport au système de base (d'environ 2 points *BLEU 2*), mais se sont avérés moins performants que ceux fournis en sens inverse.

Ensuite, nous avons évalué la contribution des alignements fournis par les cognats et les règles heuristiques à la qualité de la traduction automatique factorisée. Les résultats obtenus ont montré que l'amélioration significative de l'alignement lexical, par l'intégration des informations linguistiques dans le processus d'alignement, peut apprécier légèrement les résultats d'un système factorisé construit du roumain vers le français (de 0,14 points le *BLEU 2*). Dans le sens inverse de la traduction, du français vers le roumain, le système factorisé intégrant à la fois les alignements lexicaux supplémentaires des cognats et des règles linguistiques a néanmoins apprécié de manière intéressante les résultats finaux (de presque un point (0,97) le *BLEU 2*). Ainsi, une appréciation significative de l'alignement lexical ne mène pas toujours à une amélioration très importante de la qualité de la traduction (Ayan et Dorr, 2006 ; Fraser et Marcu, 2007). Notons pourtant que les améliorations obtenues vers le roumain se sont avérées plutôt intéressantes. En tenant compte du fait que la base de règles heuristiques utilisée est néanmoins limitée à seulement 37 règles, définies manuellement à partir d'un seul corpus de test, assez réduit en taille (environ 30 000 tokens par langue), les résultats obtenus se montrent plutôt encourageants, tant en ce qui concerne le système d'alignement lexical que les systèmes de traduction automatique. Ainsi, par l'enrichissement de la base de règles, comme il a été précisé auparavant, nous pourrions espérer des améliorations plus intéressantes à l'avenir.

Les valeurs des scores d'évaluation ont montré, d'une part, que les résultats de traduction sont dépendants de l'heuristique exploitée mais aussi du sens de la traduction. D'autre part, ces

résultats ont révélé que la traduction automatique roumain -> français fonctionne beaucoup mieux que la traduction en sens inverse français -> roumain (un décalage d'environ 21 points s'est observé entre les scores *BLEU 2* des systèmes retenus). Cela pourrait être expliqué par la morphologie plus riche du roumain par comparaison avec le français. En effet, la dispersion des données en roumain est beaucoup plus importante qu'en français et, par conséquent, pour cette direction de traduction, il semblerait que la taille du corpus d'entraînement devrait être plus volumineuse pour résoudre en partie les problèmes liés à ce phénomène bien connu. Ainsi, les résultats dépendent aussi de la taille du corpus d'entraînement, en fonction du sens de la traduction. En outre, ceux-ci sont dépendants de la paire de langues traitées ainsi que du domaine du corpus utilisé (les règles heuristiques bilingues utilisées pendant le processus d'alignement ont été définies à partir des corpus appartenant au domaine juridique et administratif). De plus, les corpus appartenant à ce domaine présentent des caractéristiques particulières (des structures répétitives, un vocabulaire réduit, l'utilisation seulement de la troisième personne du singulier ou du pluriel, etc.) et posent donc moins de difficultés aux systèmes d'alignement et de traduction, que les corpus de la langue générale, par exemple, où une langue donnée est bien représentée. Cela pourrait expliquer le score *BLEU 2* performant obtenu par le système construit vers le français à partir du corpus d'entraînement de taille réduite. Nous avons dû limiter le volume du corpus, d'une part, pour optimiser le temps d'entraînement et, d'autre part, puisque les étapes d'étiquetage et d'alignement lexical se sont montrées coûteuses en temps et en ressources humaines. Nous avons considéré que le volume de données utilisé (environ un million et demi de mots par langue) est suffisant afin d'apprendre les paramètres des systèmes. Toutefois, un prototype fonctionnel de traduction automatique statistique doit exploiter une quantité considérable de données (environ 30 millions de mots) (Ceauşu, 2009).

L'une des limites des systèmes de traduction développés est donc la taille réduite du corpus d'entraînement, surtout en ce qui concerne les systèmes construits vers le roumain. Même si les techniques de factorisation sont censées compenser le déficit des données pour les langues moins dotées en ressources et riches du point de vue morphologique (en tant que langues cibles), il semblerait que la traduction vers le roumain, en combinaison avec le français, nécessite plus de données d'entraînement pour obtenir des scores d'évaluation performants. De ce fait, notre prochaine démarche est l'augmentation de la taille du corpus d'entraînement pour les systèmes construis vers le roumain.

Un autre inconvénient des systèmes développés est qu'ils utilisent une seule configuration de facteurs linguistiques (lemmes et propriétés morphosyntaxiques). Nous avons appliqué initialement cette configuration en prenant en compte les résultats des systèmes factorisés roumain <-> anglais (Ceaușu, 2009) qui montrent les meilleurs scores *BLEU* pour cette combinaison de facteurs linguistiques. Ainsi, nous envisageons dans l'avenir de tester également d'autres configurations basées sur les lemmes, les parties du discours et les propriétés morphosyntaxiques (Ceaușu et Tufiș, 2011 ; Tufiș et Dumitrescu, 2012) (cf. chapitre 2, sous-section 2.2.2.6.1.), mais aussi d'intégrer éventuellement des informations syntaxiques dans le processus de traduction (Birch *et al.*, 2007 ; Avramidis et Koehn, 2008) (cf. chapitre 2, sous-section 2.2.2.6.2.).

Notons que nous n'avons pas pu construire pour le moment de systèmes exploitant aussi les parties du discours, car nous n'avons pas disposé de modèles de langue appropriés pour le roumain. De ce fait, nous projetons dans l'avenir d'entraîner de modèles de langue basés sur ce facteur afin de pouvoir développer les systèmes factorisés correspondants. Pour le moment, nous avons utilisé des modèles de langue du domaine juridique. Des expériences incluant des modèles construits pour d'autres domaines ou combinant des textes de plusieurs domaines sont prévues aussi dans le futur.

De plus, pour tenter d'éventuelles améliorations des résultats de traduction français <-> roumain, nous prenons aussi en considération la possibilité d'utiliser des modèles de ré-ordonnancement plus performants, construits sur les formes des mots ou sur les propriétés morphosyntaxiques (Ceaușu et Tufiș, 2011) (cf. chapitre 2, sous-section 2.2.2.6.1.), pour les deux langues étudiées, en tant que langues cibles. Ces modèles pourraient être utiles surtout pour la traduction vers le français qui présente un ordre de mots plus fixe que le roumain où cet ordre est plutôt libre. Toutefois, les deux langues montrent un ordre de mots plus flexible que l'anglais, par exemple. C'est pour cette raison que pour ces premières expériences de traduction automatique nous avons utilisé seulement le modèle de distorsion par défaut basé sur la distance. Nous avons ainsi donné la priorité à l'amélioration de l'alignement lexical et à la préparation spécifique de ressources linguistiques nécessaires aux systèmes factorisés pour les langues riches du point de vue morphologique.

Après le choix final de la configuration la plus appropriée pour le français et le roumain, nous projetons aussi d'utiliser la méthode de traduction *en cascade* (Tufiș et Dumitrescu, 2012) (cf. chapitre 2, sous-section 2.2.2.6.1.), afin de tenter d'apprécier encore les résultats de la

traduction factorisée. Cette méthode a fourni de très bons scores *BLEU* pour la paire de langues anglais - roumain.

Notre finalité ultime est l'exploitation du corpus entier disponible pour le français et le roumain (environ 30 millions de mots) (cf. chapitre 3, sous-section 3.2.1.). En outre, comme nous sommes conscients des limites du score *BLEU*, nous envisageons d'utiliser aussi les scores *NIST* (Doddington, 2002) et *METEOR* (Banerjee et Lavie, 2005) lors de l'évaluation automatique, mais aussi de réaliser une évaluation manuelle des résultats afin de repérer les erreurs de traduction, les classifier et chercher ainsi à proposer d'éventuelles solutions de correction.

À plus long terme, nous envisageons d'exploiter aussi des corpus comparables (c'est-à-dire des corpus multilingues qui ne sont pas des traductions réciproques, mais qui traitent des mêmes sujets dans plusieurs langues), afin de pallier le problème du déficit des données d'entraînement et des ressources pour la paire de langues étudiées. En effet, la méthode d'identification et d'alignement automatique des cognats mise en place (Navlea et Todirașcu, 2012) peut s'avérer utile afin d'extraire des séquences bilingues de traduction français - roumain, contenant justement des cognats, à partir de corpus comparables.

Nous projetons également d'utiliser et d'adapter cette méthode d'identification automatique des cognats pour la création de ressources linguistiques multilingues utiles dans le domaine de la didactique des langues, pour d'autres paires de langues moins dotées comme, par exemple, le roumain en combinaison avec d'autres langues romanes. D'ailleurs, cette méthode a été intégrée par le projet *COPAL*[144] (*COrpus Parallèles pour l'Alsacien*) qui vise la construction de ressources linguistiques pour des langues peu outillées comme l'alsacien.

La liste des principales ressources et outils ayant résulté à la suite de ce travail de thèse figure dans l'Annexe 8.

[144] http://lilpa.unistra.fr/fdt/projets/projets-en-cours/copal/. Ce projet est porté par Delphine Bernhard dans le cadre de l'*U. R. LiLPa* de l'Université de Strasbourg.

Annexe 1

Les étiquettes utilisées par *Tree Tagger* (Schmid, 1994) pour le français

ABR Abréviation
ADJ Adjectif
ADV Adverbe
DET:ART Article
DET:POS Pronom Possessif (*ma, ta, ...*)
INT Interjection
KON Conjonction
NAM Nom Propre
NOM Nom
NUM Numéral
PRO Pronom
PRO:DEM Pronom Démonstratif
PRO:IND Pronom Indéfini
PRO:PER Pronom Personnel
PRO:POS Pronom Possessif (*mien, tien, ...*)
PRO:REL Pronom Relatif
PRP Préposition
PRP:det Préposition + Article (*au,du,aux,des*)
PUN Ponctuation
PUN:cit Ponctuation de citation
SENT Balise de phrase
SYM Symbole
VER:cond Verbe au conditionnel
VER:futu Verbe au futur
VER:impe Verbe à l'impératif
VER:impf Verbe à l'imparfait
VER:infi Verbe à infinitif
VER:pper Verbe au participe passé
VER:ppre Verbe au participe présent
VER:pres Verbe au présent
VER:simp Verbe au passé simple
VER:subi Verbe à l'imparfait du subjonctif
VER:subp Verbe au présent du subjonctif

Annexe 2

Les étiquettes *MSD* des adjectifs - projet *Multext* [145]- utilisées par *Flemm* (Namer, 2000) pour le français[146]

ADJECTIFS
Étiquette Exemple Signification
Afcfp- *meilleures* Comp. qualif. adjec. fem. plur.
Afcfs- *meilleure* Comp. qualif. adjec. fem. sing.
Afcmp- *meilleurs* Comp. qualif. adjec. masc. plur.
Afcms- *meilleur* Comp. qualif. adjec. masc. sing.
Afpfp- *bonnes* Qualif. adjec. fem. plur.
Afpfs- *bonne* Qualif. adjec. fem. sing.
Afpmp- *bons* Qualif. adjec. masc. plur.
Afpms- *bon* Qualif. adjec. masc. sing.
Ai-fp- *certaines, mêmes* etc. Indef. adjec. fem. plur.
Ai-fs- *certane, même* etc. Indef. adjec. fem. sing.
Ai-mp- *certain, mêmes* etc. Indef. adjec. masc. plur.
Ai-ms- *certain, même* etc. Indef. adjec. masc. sing.
Ac-fp- *deux* Card. adjec. fem. plur.
Ac-fs- *une* Card. adjec. fem. sing.
Ac-mp- *deux* Card. adjec. masc. plur.
Ac-ms- *un* Card. adjec. masc. sing.
Ao-fp- *premières* Ord. adjec. fem. plur.
Ao-fs- *première* Ord. adjec. fem. sing.
Ao-mp- *premiers* Ord. adjec. masc. plur.
Ao-ms- *premier* Ord. adjec. masc. sing.
As-fp- *leurs, miennes* etc. Poss. adjec. fem. plur.
As-fs- *leur, mienne* etc. Poss. adjec. fem. sing.
As-mp- *leurs, miens* etc. Poss. adjec. masc. plur.
As-ms- *leur, mien* etc. Poss. adjec. masc. sing.

Explication des abréviations et des étiquettes : adjec. : adjectif (**A**) ; comp. : comparatif (**c**) ; qualif. : qualificatif (**f**) ; indef. : indéfini (**i**) ; card. : cardinal (**c**) ; ord. : ordinal (**o**) ; poss. : possessif (**s**) ; fem. : féminin (**f**) ; masc. : masculin (**m**) ; sing. : singulier (**s**) ; plur. : pluriel (**p**).

[145] http://aune.lpl.univ-aix.fr/projects/multext/
[146] Pour la liste complète des étiquettes *MSD* utilisées pour le français voir l'adresse suivante : http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX.LangSpec.fr.html.

Annexe 3

Étiquettes *MSD* (Tufiș et Barbu, 1997) - projet *Multext* - utilisées par *TTL* (Ion, 2007) pour le roumain[147]

NOMS :	**VERBES :**	**ADJECTIFS :**
Ncmsrn frate (**r** : cas nominatif - accusatif, **n** : déterminant zéro)	Vmii1s abandonam	Afpmsrn bun
Ncmson frate (**o** : cas datif-génitif)	Vmii2s abandonai	Afpmson bun
Ncmsvn frate (**v** : cas vocatif)	Vmii3s abandona	Afpmsvn bun
Ncmsry fratele (**y** : déterminant défini)	Vmii1p abandonam	Afpmprn buni
Ncmsoy fratelui	Vmii2p abandonați	Afpmpon buni
Ncmprn frați	Vmii3p abandonau	Afpmpvn buni
Ncmpon frați	Vmis1s abandonai	Afpmsry bunul
Ncmpvn frați	Vmis2s abandonași	Afpmsoy bunului
Ncmpry frații	Vmis3s abandonă	Afpmpry bunii
Ncmpoy fraților	Vmis1p abandonarăm	Afpmpoy bunilor
Ncmpvy fraților	Vmis2p abandonarăți	Afpfsrn bună
Ncfsrn soră	Vmis3p abandonă	Afpfsvn bună
Ncfsvn soro	Vmil1s abandonasem	Afpfson bune
Ncfson surori	Vmil2s abandonaseși	Afpfprn bune
Ncfprn surori	Vmil3s abandonase	Afpfpon bune
Ncfpon surori	Vmil1p abandonaserăm	Afpfpvn bune
Ncfpvn surori	Vmil2p abandonaserăți	Afpfsry buna
Ncfsoyy soră(-sii)	Vmil3p abandonaseră	Afpfsoy bunei
Ncfpry surorile	Vmip1s abandonez	Afpfpry bunele
Ncfpoy surorilor	Vmsp1s abandonez	Afpfpoy bunelor
Ncfpvy surorilor	Vmip2s abandonezi	Afcmsrn ulterior
Ncfsryy soră(-sa)	Vmsp2s abandonezi	Afcmson ulterior
Ncmsrn creion	Vmip3s abandonează	Afsmsrn extrem
Ncmson creion	Vmip3p abandonează	Afp gri
Ncmsryy creionu-(i)	Vmsp3s abandoneze	
Ncmsoy creionului	Vmsp3p abandoneze	
Ncfprn creioane	Vmsp1p abandonăm	
Ncfpon creioane	Vmsp2p abandonați	
Ncfpry creioanele	Vmm-2s abandonează	
Ncfpoy creioanelor	Vmm-2p abandonați	
Npfsr Ioana	Vmnp abandona	
Npfso Ioanei	Vmp--sm abandonat	
Npmsrn București	Vmp--sm---y abandonatu	
Npmsry Bucureștiul	Vmp--sf abandonată	
	Vmp--pf abandonate	
	Vmp--pm abandonați	
	Vmg abandonând	
	Vmg-------y abandonându	
	Va--1s as,	
	Voip trebuie	
	Vcip1s sunt	
	Vcip3p sunt	
	Vcip1s----y –s	
	Vcip3p----y –s	

[147] Une description détaillée concernant ces étiquettes figurent dans (Erjavec, 2004).

PRONOMS : Pp1msn--------s eu Pp1msd--------w mi Pp1msd--------s mie Pp1msd--y-----s mi- Pd-msr acesta Pd-mso acestuia Pi-mpr toți Ps1fsrs mea Pw-mso cărui Pn-msr nimeni Ph1msr însumi Ph1fsr însămi Px3msa--------s sine Px3msa--------w se Px3msa--y-----w s-	**DÉTERMINANTS :** Dd-mso---e acestui Dd-mso---o acestuia Di-mpr toți Ds1fsrs mea Dw-msr care Dw-mso cărui Dz-msr nici_un Dh1msr însumi Dh1fsr însămi **ARTICLES :** Tfmso lui Tffso lui Timsr un Tsmpr ai Tdfso celei Timsry -ist Timsoy –ist Tfmsry –istul Tfmsoy –istului	**ADVERBES :** Rgp repede Rgs extraordinar Rgc ulterior Rp mai Rz nicăieri Rm probabil Rw cum **PRÉPOSITIONS :** Spsa în Spsay într- Spsd datorită Spca de_la **CONJONCTIONS :** Crssp _și Ccssp dar Ccrsp fie...fie Csrsz nici...nici Csscp de_vreme_ce **NUMÉRAUX :** Mcmprl doi Mcmpol doi Momsrl doilea Momsol doilea Mlmpr amândoi Momsrlyy primu-i Mffpoly treimilor **INTERJECTIONS :** I oh,ah,au **Résidual (X)** .X show, a+b, retro- **ABRÉVIATIONS (Y)** Ynfsoy d-nei Ynnsry apt. **PARTICULES (Q)** Qz nu Qz-y n- Qn a Qs s_a Qa fi Qf o

Annexe 4

L’explication des abréviations et des signes qui apparaissent dans l’écriture des règles heuristiques morphosyntaxiques et stylistiques définies :

N : nom
ADJ : adjectif
V : verbe
dét. : détermianant
dét. déf. : déterminant défini
dét. indéf. : déterminant indéfini
dét. num. : déterminant numéral
num. card. : numéral cardinal
+ : signifie « suivi de »
( ) : comprennent une propriété significative à prendre en compte pour l’élément respectif
| : signifie « ou »
[…] : marquent des éléments supplémentaires pouvant être présents dans le contexte pour lesquels une variable de distance doit être définie
… : signifie « etc. »

Annexe 5

Exemples de cooccurrences *Nom1 déverbal - (préposition) - Nom2* et de leurs contextes pour le roumain (extraites à partir du corpus *JRC-Acquis* (Steinberger *et al.*, 2006)) selon la méthodologie proposée par Todirașcu *et al.* (2008)

lemme1 lemme2 Distance entre les noms Log-likelihood (LL) Fréquence

nom1/lemme1/etiquette1 mot1/lemme_mot1/étiquette_mot1 nom2/lemme2/etiquette2 fréquence

punere aplicare 2 46467.91933 5377

punerea/punere/nsry în/în/s aplicare/aplicare/nsrn 2938

punere/punere/nsrn în/în/s aplicare/aplicare/nsrn 1558

punerii/punere/nsoy în/în/s aplicare/aplicare/nsrn 864

puneri/punere/nson în/în/s aplicare/aplicare/nsrn 5

punerea/punere/nsry în/în/s aplicarea/aplicare/nsry 5

puneri/punere/npn în/în/s aplicare/aplicare/nsrn 2

punerea/punere/nsry în/în/nsn aplicare/aplicare/nsrn 1

punere/punere/nsrn în/în/nsn aplicare/aplicare/nsrn 1

punere/punere/nsrn de/de/s aplicare/aplicare/nsrn 1

punerilor/punere/npoy în/în/s aplicare/aplicare/nsrn 1

punerii/punere/nsoy lor/lui/ps aplicare/aplicare/nsrn 1

publicare jurnal 2 37158.62315 2823

publicarea/publicare/nsry în/în/s jurnalul/jurnal/nsry 1710

publicării/publicare/nsoy în/în/s jurnalul/jurnal/nsry 1080

publicare/publicare/nsrn în/în/s jurnalul/jurnal/nsry 25

publicarea/publicare/nsry în/în/s jurnalele/jurnal/npry 2

publicări/publicare/nson în/în/s jurnalul/jurnal/nsry 2

publicării/publicare/nsoy în/în/y jurnalul/jurnal/nsry 1

publicare/publicare/nsrn a/al/ts jurnalului/jurnal/nsoy 1

publicarea/publicare/nsry în/în/s jurnalului/jurnal/nsoy 1

publicării/publicare/nsoy în/în/nsn jurnalul/jurnal/nsry 1

Annexe 6

L’explication des signes qui apparaissent dans l’écriture des filtres morphosyntaxiques définis afin d’extraire automatiquement les collocations nominales de type *Nom 1 déverbal + (préposition) + Nom 2* :

- : signifie « suivi de »

[ ] : encadrent les propriétés morphosyntaxiques de l’élément précédent

| : signifie « ou »

( ) : encadrent un élément optionnel

/ : signifie « qui introduit »

Annexe 7

Exemple de règle heuristique simple utilisée par l'aligneur lexical *YAWA* (Tufiș *et al.*, 2006 : 155) :

si *b* est aligné avec *c* et *b* est précédé par *a*, alors aligne *a* avec *c*, à moins qu'il n'existe *d* dans le même *chunk* avec *c* et la catégorie de partie de discours de *d* a une affinité significative avec la catégorie de *a*

Dans l'exemple de la Figure 72 ci-dessous *b* et *c* sont des mots pleins, tandis que *a* est très probablement un déterminant ou un modifieur pour *b* (Tufiș *et al.*, 2006 : 155) :

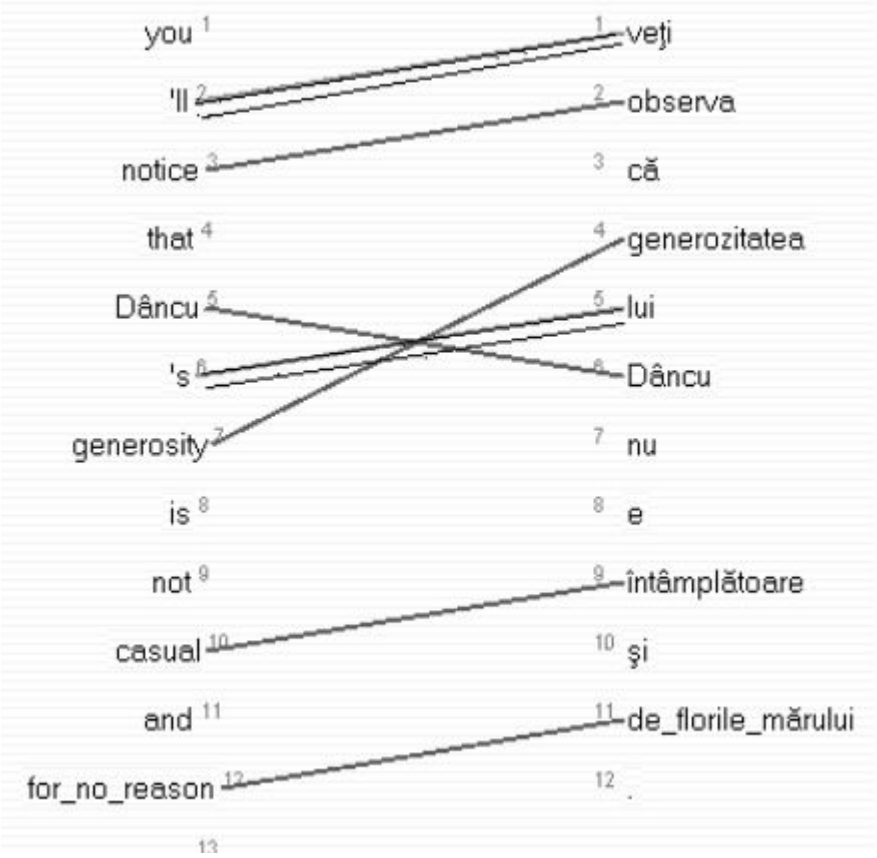


Figure 72. Alignement des mots appartenant aux *chunks* qui sont restés non alignés pendant l'étape précédente d'alignement lexical (Tufiș *et al.*, 2006 : 155)

Annexe 8

La liste des principales ressources et outils ayant résulté à la suite de ce travail de thèse :

1) un corpus parallèle du domaine politique (collecté et nettoyé manuellement à partir du site Web de la Commission Européenne) (environ 200 000 mots par langue) ;

2) un corpus parallèle du domaine politique (constitué et nettoyé manuellement à partir du site Web du Parlement Européen) (environ 130 000 mots par langue) ;

3) un corpus parallèle du domaine aéronautique (construit et nettoyé manuellement à partir des sites Web des compagnies aériennes roumaines *TAROM, Blue Air*) (environ 30 000 mots par langue) ;

4) un corpus parallèle d'entraînement du domaine juridique et administratif (extrait de *DGT-TM* (Steinberger *et al*., 2012)), lemmatisé, étiqueté et aligné lexicalement selon la méthodologie hybride proposée (environ 1 500 000 mots par langue ; 64 923 paires de phrases alignées)[148] ;

5) un corpus parallèle de test (extrait de *DGT-TM* (Steinberger *et al*., 2012)), lemmatisé et étiqueté (environ 30 000 tokens par langue ; 1 000 phrases alignées) ;

6) un corpus de référence aligné manuellement au niveau lexical en deux variantes (avec et sans l'alignement des collocations ; environ 30 000 tokens par langue; 1 000 phrases alignées)[149] ;

7) les ressources linguistiques françaises d'entraînement pour l'étiqueteur *TTL* (Ion, 2007) :

    a. un corpus monolingue français du domaine juridique et administratif, lemmatisé, étiqueté et corrigé (extrait de *JRC-Acquis* (Steinberger *et al.*, 2006)) (498 788 mots) (Todirașcu *et al*., 2011) ;

[148] Avec la collaboration de GONCHAROVA Yuliya (stage de master, 6 mois).
[149] Avec la collaboration de HAVAȘI Sebastian-Flaviu (stage de master, 4 mois).

b. un corpus monolingue français journalistique lemmatisé, étiqueté et corrigé (extrait du *Monde*, 1980-1988)) (488 543 mots) (Todirașcu *et al*., 2011).

8) une base de règles morphosyntaxiques constituée manuellement pour une méthode de correction automatique des erreurs d'étiquetage et/ou de lemmatisation des corpus français prétraités (développée en Perl) (19 règles de correction) ;

9) une base de règles d'ajustements orthographiques au sein des paires bilingues de mots français - roumain (constituée empiriquement à partir du corpus) (17 règles sensitives ou non sensitives au contexte phonétique) ;

10) une base de règles heuristiques d'alignement lexical français - roumain (constituée manuellement) (37 règles morphosyntaxiques et stylistiques) ;

11) un lexique de cognats construit manuellement à partir du corpus parallèle de test français - roumain (environ 2 000 cognats) ;

12) un ensemble de collocations *Verbe + Nom* validées manuellement à partir du corpus de test français - roumain et leurs équivalents de traduction (environ 230 collocations par langue) ;

13) 2 modèles de langue français (un construit sur la forme des mots et un autre développé sur les propriétés morphosyntaxiques) dans le domaine juridique ;

14) un nouveau module hybride d'identification et d'alignement automatiques de cognats (à base de n-grammes (Simard *et al.*, 1992) et d'informations linguistiques), à partir de corpus parallèles français - roumain (développé en Perl) ;

15) 8 systèmes de traduction automatique français - roumain (2 systèmes purement statistiques et 6 factorisés), construits dans les deux directions du processus de traduction (16 au total).

Liste des figures

Liste des tableaux

Index des auteurs

Liste des publications

Articles dans des revues internationales à comité de lecture

1) **Navlea**, **M.**, Todirașcu, A. (2014). A Hybrid Word Alignment System for Statistical Machine Translation, *American Journal of Systems and Software, Special Issue on Multidisciplinary Perspectives of Agent-based Systems*. (à paraître)

2) Todirașcu, A., Grass, T., **Navlea**, **M.**, Longo, L. (2014). La relation de hiérarchie « chef » : une approche translingue français - anglais - allemand. *META*, 59 (2). (à paraître)

Articles dans des revues nationales à comité de lecture

3) Todirașcu, A., Ion, R., **Navlea**, **M.**, Longo, L. (2011). French text preprocessing with TTL. *in Proceedings of the Romanian Academy, Series A: Mathematics, Physics, Technical Sciences and Information Science*, Volume 12, Number 2/2011 (April-June), Bucharest, Romanian Academy, Romanian Academy Publishing House, 151-158. ISSN 1454-9069.

Chapitres d'ouvrages

4) **Navlea**, **M.**, Todirașcu, A. (2013). Experiments with a French - Romanian Word Alignment System. *in* Dan Tufiș, Vasile Rus, and Corina Forăscu (eds). *Towards Multilingual Europe 2020: A Romanian Perspective*, Romanian Academy Publishing House, Bucharest, 2013, 225-240. ISBN 978-973-27-2282-4.

5) Todirașcu, A., **Navlea**, **M.** (2010). CAP: A Hierarchical Lexical Function Related to Proper Nouns: The Case of Romanian and French. *in* Dan Tufiș and Corina Forascu (eds). *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Bucharest, Romania, December 2010, Romanian Academy Publishing House, 317-330. ISBN 978-973-27-1972-5.

Communications à des conférences internationales à comité de lecture

6) **Navlea**, **M.**, Todirașcu, A. (2012). Using Cognates to Improve Lexical Alignment Systems. *in* Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (eds). *Text, Speech and Dialogue (15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings), Lecture Notes in Computer Science*, Volume 7499, Springer Verlag Berlin Heidelberg, 370-377. ISBN: 978-3-642-32789-6 (Print) 978-3-642-32790-2 (Online)

7) **Navlea**, **M.**, Todirașcu, A. (2012). Using Cognates to Improve Lexical Alignment Systems. *in Proceedings of the TSD 2012 Hybrid Machine Translation Workshop,* Brno, Czech Republic, September 3, 2012.

8) **Navlea**, **M.**, Havaşi, S.-F. (2012). A French-Romanian lexical alignment system using legal parallel corpora: Error Classification. *in Proceedings of The International Conference for Young Researchers Classification in Linguistics: Methodology, Problems, Challenges*, June 6-8, 2012, Strasbourg. (à paraître)

9) **Navlea**, **M.**, Todirașcu, A. (2011). Using Cognates in a French - Romanian Lexical Alignment System: A Comparative Study. *in* Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nikolai Nikolov (eds). *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria, September 2011. INCOMA Ltd., Shoumen, Bulgaria, 247-253. ISSN 1313-8502.

10) **Navlea**, **M**., Todirașcu, A. (2011). Cognate Identification for a French - Romanian Lexical Alignment System: Empirical Study. *in* Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste (eds). *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011)*, Leuven, Belgium, May 2011, 145-152. ISBN 9789081486118.

11) **Navlea**, **M.**, Todirașcu, A. (2011). Repérage automatique des équivalences traductionnelles pour un système de traduction automatique statistique français - roumain. *in Actes du Colloque Métiers et technologies de la traduction : quelles convergences pour l'avenir ? (TRALOGY 2011)* [en ligne], 3-4 mars 2011, Paris

(France). Mis à jour le : 11/01/2012, [URL : http://lodel.irevues.inist.fr/tralogy/index.php?id=162].

12) **Navlea**, **M.**, Todirașcu, A. (2011). Ressources linguistiques pour un outil de traduction automatique statistique factorisée français - roumain. *in* Marc Van Campenhoudt, Teresa Lino et Rute Costa (éds). *Actes des 8èmes Journées Scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction : Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face au défi de la diversité (Lisbonne, Portugal, 15-17 octobre 2009)*, Éd. des Archives Contemporaines, Paris, France, et Agence universitaire de la Francophonie (AUF), avril 2011, 375-388. ISBN 9782813000521.

13) **Navlea**, **M.**, Todirașcu, A. (2010). Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems. *in* Stelios Piperidis, Milena Slavcheva, and Cristina Vertan (eds). *Proceedings of the International Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-)Eastern European Languages*, 7th *International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May, 2010, 41-48.

Communications à des conférences nationales à comité de lecture

14) **Navlea**, **M.**, Todirașcu, A. (2011). Identification de cognats à partir de corpus parallèles français - roumain. *in* Mathieu Lafourcade et Violaine Prince (éds). *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Volume 2/2 (session poster), Montpellier, France, juin 2011, 81-86.

15) **Navlea**, **M.**, Todirașcu, A. (2010). Un sistem de traducere automată română - franceză. *in* Adrian Iftene, Horia-Nicolai Teodorescu, Dan Cristea et Dan Tufiș (éds). *Actes de la 6ème Conférence Ressources linguistiques et outils pour le traitement automatique du roumain*, Universitatea « Al. I. Cuza Iaşi », Ed. Universității "Al. I. Cuza", Iaşi, Roumanie, septembre 2010, 165-174. ISSN: 1843-911X.

Bibliographie

Adamson, G.W., Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10(7-8), 253-260.

Allauzen, A., Crego, J. M., El-Kahlout, I. D., Yvon, F. (2010). LIMSI's Statistical Translation Systems for WMT'10. *in Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, juillet 2010, Uppsala (Suède). Stroudsburg (USA, PA): Association for Computational Linguistics, 54-59.

Allauzen, A., Wisniewski, G. (2009). Modèles discriminants pour l'alignement mot à mot. *TAL*, 2009, 50(3), 173-203.

Alp, N. D., Turhan, C. (2008). English to Turkish Example-Based Machine Translation with Synchronous SSTC. *in Fifth International Conference on Information Technology: New Generations (ITNG)*, 2008, Las-Vegas (USA, NV). Washington (USA, DC): IEEE Xplore, 2008, 674 - 679.

Artstein, R., Massimo, (2008). Inter-coder agreement for computational linguistics. *Journal of Computational Linguistic*. Cambridge (USA, MA): MIT Press, 2008, 34(4), 555-596.

Avramidis, E., Koehn, P., (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. *in Proceedings of ACL-08: HLT*, juin 2008, Columbus (USA). Stroudsburg (USA, PA): Association for Computational Linguistics, 763-770.

Ayan, N. F., Dorr, B. J. (2006). Going beyond aer: an extensive analysis of word alignments and their impact on MT. *in Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, 17-21 juillet 2006, Sydney (Australie), 9-16.

Bahl, L. R., Jelinek, F., Mercer, R. L. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, mars 1983, PAMI-5(2), 179-190.

Baldwin, T., Kim, S. N. (2010). Multiword expressions. *in* Indurkhya, N., Damerau, F. J. (eds). *Handbook of Natural Language Processing*, Second edition, Boca Raton (USA, FL): CRC Press, Taylor and Francis Group, 2010, 267-292.

Banerjee, S., Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *in Proceedings of the ACL Workshop on*

Intrinsic and Extrinsic Evaluation Measures for MT and/or summarization, juin 2005, Ann Arbor, Michigan (USA), 65-72.

Bangalore, S., Joshi, A. (1999). Supertagging: An approach to almost parsing. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, juin 1999, 25(2), 237-265.

Banks, D. (2005). *Introduction à la linguistique systémique fonctionnelle de l'anglais*. Paris : L'Harmattan, 128 p.

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3, 1137-1155.

Bergsma, S., Kondrak, G. (2007). Multilingual Cognate Identification using Integer Linear Programming. *in Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP)*, 27-29 septembre 2007, Borovets (Bulgarie), 11-18.

Bertoldi, N., Haddow, B., Fouet, J.-B. (2009). Improved Minimum Error Rate Training in Moses. *Prague Bulletin of Mathematical Linguistics (PBML)*, 2009, 91, 7-16.

Birch, A., Osborne, M., Koehn, (2007). CCG Supertags in factored Statistical Machine Translation. *in* P*roceedings of the Second Workshop on Statistical Machine Translation*. Prague (République Tchèque). Stroudsburg (USA, PA): Association for Computational Linguistics, 9-16.

Blumenthal, (2007). A Usage-based French Dictionary of Collocations. *in* Kawaguchi, Y., Takagaki, T., Tomimori, N., Tsuruga, Y. (eds). *Corpus-Based Perspectives in Linguistics*. Amsterdam (Pays-Bas): Benjamins (Usage-Based Linguistic Informatics 6), 67-83.

Boitet, C. (1993a). La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue. *in* Clas, A., Bouillon, P., (éds). *La traductique*. Montréal (Canada) : Presses de l'Université de Montréal, 109-148.

Boitet, C. (1993b). TA et TAO à Grenoble... 32 ans déjà !. *T.A.L. : Spécial Trentenaire*, 1992. Revue semestrielle. Paris (France) : Association pour le traitement automatique des langues (ATALA), 1993, 33(1-2), 45-84.

Boitet, C. (2008). Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes. *in Actes de TALN 2008*, 9-13 juin 2008, Avignon (France), 12 p.

Boitet, C., Blanchon, H. (1994). Promesses et problèmes de la « TAO pour tous » après LIDIA-1, une première maquette. Revue. *Langages*. Paris (France) : Larousse, janvier 1994, 28(116), 20-47.

Boitet, C., Guillaume, P., Quezel-Ambrunaz, M. (1982). Ariane-78, an integrated environment for automated translation and human revision. *in actes de Coling*, 1982, Prague (République Tchèque), 19-27.

Bolinger, D. (1971). *The phrasal verb in English*. Harvard (USA): Harvard University Press, 1972, 187 p.

Boroș, T., Dumitrescu, Ș. D., Ion, R., Ștefănescu, D., Tufiș, D., (2013). Romanian-English Statistical Translation at RACAI. *in* Mitocariu, E., Moruz, M. A., Cristea, D., Tufiș, D., Clim, M. (eds). *Proceedings of the 9th International Conference: linguistic resources and tools for processing of the romanian language*, 17 mai 2013, Miclăușeni (Roumanie). Iași (Roumanie) : Ed. Universităţii « Al. I. Cuza », 81-98.

Bouamor, D., Semmar, N., Zweigenbaum, (2012). Identifying multi-word expressions in statistical machine translation. *in Proceedings of 8th International Conference on Language Resources and Evaluation (LREC)*, 2012, 23-25 mai Isanbul (Turquie). Paris (France): European Language Resources Association (ELRA), 674-679.

Braasch, A., Olsen, S. (2000). Formalised Representation of Collocations in a Danish Computational Lexicon. *in* U. Heid & *al.* (eds). *Proceedings of the Ninth EURALEX Congress*, Stuttgart (Allemagne), 2, 475-488.

Brew, C., McKelvie, D. (1996). Word-pair extraction for lexicography. *in Proceedings of International Conference on New Methods in Natural Language Processing*, septembre 1996, Bilkent (Turquie), 45-55.

Brill, E. (1992). A simple rule-based part of speech tagger. *in Proceedings of the Third Conference on Applied Natural Language Processing*, 1992, Trento (Italie). Stroudsburg (USA, PA): Association for Computational Linguistics, 1992, 152-155.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, décembre 1995, 21(4), 543-565.

Brown, F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R., Roossin, (1988). A statistical approach to language translation. *in* Vargha, D. (eds). *Proceedings of the 12th Conference on Computational Linguistics: Coling 88*, 1988, Budapest (Hongrie).

Stroudsburg (USA, PA): Association for Computational Linguistics, 1, 71-76. Also: Computational Linguistics, 16, 1990, 79-85.

Brown, F., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., Roossin, S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*. Cambridge (USA, MA): MIT Press, juin 1990, 16 (2), 7 p.

Brown, F., Della Pietra, V. J., Della Pietra, S. A., Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, juin 1993, 19(2), 263-312.

Brown, R. (1996). Example-Based Machine Translation in the Pangloss System. *in Proceedings of the 16th International Conference on Computational Linguistics*: *Coling 96*, 1996, Copenhagen (Danemark). Stroudsburg (USA, PA): Association for Computational Linguistics, 1996, 1, 169-174.

Carreras, X., Marquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *in Proceedings of 9th Conference on Computational Natural Language Learning (CoNLL)*, 29-30 juin 2005, Ann Arbor (Michigan), 152-164.

Casacuberta, F., Vidal, E. (2004). Machine translation with inferred stochastic finitestate transducers. *Journal of Computational Linguistics*. Cambridge (USA, MA) : MIT Press, juin 2004, 30(2), 205-225.

Ceauşu, A. (2008). Colectarea şi procesarea documentelor româneşti ale corpusului JRC Acquis [Collection et traitement des douments roumains du corpus JRC Acquis]. *in Resurse lingvistice şi instrumente pentru prelucrarea limbii române [Ressources linguistiques et outils pour le traitement automatique du roumain]*, novembre 2008, Iaşi (Roumanie), 125-130.

Ceauşu, A. (2009). *Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă [Techniques de traduction automatique et leur application à la langue roumaine comme langue source]*. Thèse de Doctorat. Bucarest (Roumanie) : Académie Roumaine, avril 2009, 123 p.

Ceauşu, A., Tufiş, D. (2011). Addressing SMT Data Sparseness when Translating into Morphologically-Rich Languages. *in* Sharp, B., Zock, M., Carl, M., Lykke Jakobsen, A. (eds). *Proceedings of the 8th international NLPCS workshop*: *Human-machine interaction in translation*, 20-21 août 2011, Copenhagen Business School (Danemark). Copenhage (Danemark): Samfundslitteratur, 2011, 41, 57-68.

Cendejas, E., Barcelo, G., Gelbukh, A., Sidorov, G. (2009). Incorporating Linguistic Information to Statistical Word-Level Alignment. *in* Bayro-Corrochano, E., Eklundh, J.O. (eds). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 14th Iberoamerican Conference on Pattern Recognition, CIARP 2009,* 15-18 novembre 2009, Guadalajara (Méxique). Berlin (Heidelberg): Springer-Verlag, 2009, LNCS 5856, 387–394.

Cettolo, M., Federico, M., Girardi, C. (2012). WIT3: Web inventory of transcribed and translated talks. *in Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, mai 2012, Trento (Italie), 261-268.

Cherry, C., Lin, D. (2003). Proalign: Shared system task description. *in HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 31 mai 2003, Edmonton (Canada). Stroudsburg (USA, PA): Association for Computational Linguistics, 2003, 11-14.

Chiang, D. (2007). Hierarchical phrase-based translation. *Journal of Computational Linguistics.* Cambridge (USA, MA): MIT Press, juin 2007, 33(2), 201-228.

Clark, S. (2002). Supertagging for combinatory categorial grammar. *in Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, 2002, Venise (Italie), 19-24.

Collins, M. (1997). Three generative, lexicalized models for statistical parsing. *in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL'97)*, 7-12 juillet 1997, Madrid (Espagne). Stroudsburg (USA, PA): Association for Computational Linguistics, 1997, 16-23.

Collins, M., Koehn, P., Kucerova I. (2005). Clause Restructuring for Statistical Machine Translation. *in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 25-30 juin 2005, Ann Arbor (USA, MI). Stroudsburg (USA, PA): Association for Computational Linguistics, 2005, 531-540.

Cornu, G. (2005). *Linguistique juridique*. Paris (France) : Montchrestein, 3e édition, 443 p.

Cowie, A. (1981). The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, 2(3), 223-235.

Crego, J. M., Marino, J. B. (2007a). Extending MARIE: a N-gram-based smt decoder. *in Proceedings of the 45th Annual Meeting of the Association for Computational*

Linguistics: Demo and Poster Sessions, juin 2007, Prague (République Tchèque), 213-216.

Crego, J. M., Marino, J. B. (2007b). Improving statistical MT by coupling reordering and decoding. *Machine Translation*. Hingham (USA, MA): Kluwer Academic Publishers, septembre 2006, 20(3), 199-215.

Crego, J. M., Yvon, F. (2010). Improving Reordering with Linguistically Informed Bilingual n-grams. *in Proceedings of the 23rd International Conference on Computational Linguistics: Coling 2010* [Poster], août 2010, Beijing (Chine). Stroudsburg (USA, PA): Association for Computational Linguistics, 2010, 197-205.

Crego, J. M., Yvon, F., Marino J. B. (2011). N-code: an open-source Bilingual N-gram SMT Toolkit. *The Prague Bulletin of Mathematical Linguistics (PBML)*, octobre 2011, 96(1), 49-58.

Cristea, T. (2007). *Stratégies de la traduction*. Bucarest (Roumanie) : Editura Fundației România de Mâine [Edition de la Fondation Roumanie de demain], 3e édition, 196 p.

Delisle, J., Lee-Jahnke, H., Cormier, M. C. (1999). *Terminologie de la traduction/Translation Terminology/Terminología de la traduccción/Terminologie der Übersetzung*. Amsterdam/Philadelphia: John Benjamin Publishing Company, 1999, 433 p.

Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: series B*, 1977, 39, 1-38.

Dimitrova, L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H. J., Tufiş, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *in Proceedings of COLING*, août 10-14, 1998, Montréal: Morgan Kaufmann Publishers / ACL,, 1998, 315-319.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. *in Proceedings of the Second International Conference on Human Language Technology Research (HLT),* 2002, San Diego (USA, CA). San Francisco (USA, CA): Morgan Kaufmann Publishers Inc., 2002, 128-132.

Dumitrescu, Ș. D., Ion, R., Ștefănescu, D., Boroș, T., Tufiș, D. (2012). Romanian to English Automatic MT Experiments at IWSLT12. *in* Eiichiro Sumita, Dekai Wu, Michael Paul, and Chengqing Zong (eds). *Proceedings of the International Workshop on Spoken Language Translation (ISWLT 2012)*, décembre 2012, Hong Kong (Chine), 136-143.

Dumitrescu, Ș. D., Ion, R., Ștefănescu, D., Boroș, T., Tufiș, D. (2013). Experiments on Language and Translation Models Adaptation for Statistical Machine Translation. *in*

Tufiș, D., Rus, V., Forăscu, C. (eds). *Towards Multilingual Europe 2020: A Romanian Perspective,* 2013. Bucarest (Romanie): Editura Academiei, 2013, 205-224.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, mars 1993, 19(1), 61-74.

Dyen, I., Kruskal J. B., Black, (1992). An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*. Philadelphia (USA, PA): American Philosophical Society, 1992, 82(5), 132 p.

Erjavec, T. (2004). *MULTEXT-East morphosyntactic specifications V.3*. Ljubljana (Slovénie): Erjavec, 10 mai 2004, 241 p.

Evert, S. (2005). *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Thèse de Doctorat. Stuttgart (Allemagne) : Institut für maschinelle Sprachverarbeitung, Université de Stuttgart, 30 août 2005, 353 p.

Fraser, A., Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, septembre 2007, 33(3), 293-303.

Gavrilă, M. (2009), SMT experiments for Romanian and German using JRC-Acquis. *in Proceedings of RANLP - associated workshop: Multilingual resources, technologies and evaluation for central and Eastern European languages*, 17 septembre 2009, Borovets (Bulgarie). Stroudsburg (USA, PA): Association for Computational Linguistics, 2009, 14-18.

Gavrilă, M. (2012). Improving Recombination in a Linear EBMT System by Use of Constraints. Thèse de Doctorat, University of Hamburg, Hamburg, 195 p.

Gledhill, C. (2007). La portée : seul dénominateur commun dans les constructions verbo-nominales. *in* Frath, P., Pauchard, J. & Gledhill, C. (éds). *Actes du 1er colloque, Res per nomen, pour une linguistique de la dénomination, de la référence et de l'usage*,, Université de Reims-Champagne-Ardenne, 24-26 mai 2007, 113-125.

Gledhill, C., Todirașcu, A. (2008). Collocations en contexte : extraction et analyse contrastive. « Collocations en contexte : extraction et analyse contrastive », revue électronique *Texte et corpus*, n°3 / août 2008, *Actes des Journées de la linguistique de Corpus 2007*, 137-148.

[http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_todirascu_gledhill.pdf]

Gojun, A. (2010). *Null Subjects in Statistical Machine Translation: A Case Study on Aligning English and Italian Verb Phrases with Pronominal subjects*. Diplomarbeit, Institut fur Maschinelle Sprachverarbeitung, University of Stuttgart, 107 p.

Grass, T. (2010). A quoi sert encore la traduction automatique. *Les Cahiers du GEPE*, N°2/ 2010. *Outils de traduction - outils du traducteur ?* [http://www.cahiersdugepe.fr/index1367.php].

Grossmann, F., Tutin, A. (éds). (2003). Les collocations : analyse et traitement. *Numéro spécial : Travaux et Recherches en Linguistique Appliquée*. Amsterdam (Hollande) : De Werelt, 142 p.

Guţu-Romalo, V. (coord.). (2005). *Gramatica limbii române : Cuvântul, [Grammaire de la langue roumaine, le Mot]*. Bucarest (Roumanie) : Ed. de l'Académie Roumaine, 2005, 1, 712 p.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. London (GB): Arnold, E., 1985, 1e édition, 387 p.

Hathout, N., Namer, F., Dal, G. (2002). An Experimental Constructional Database: The MorTAL Project. *in* P. Boucher, (ed). *Many Morphologies*. Somerville (USA, MA): Cascadilla Press, 2002, 178-209.

Hausmann, F. J. (1979). Un dictionnaire des collocations est-il possible ?. *Travaux de linguistique et de littérature*, 1979, 17(1), 187- 195.

Heid, U., Ritz, J. (2005). Extracting collocations and their contexts from corpora. *in* Kiefer, F., Kiss, G., Pajzs, J. (eds). *Paper in the 8th Conference on Computational Lexicography and Text Research: COMPLEX 2005*, 17-18 juin 2005, Budapest (Hongrie). Budapest (Hongrie): Hungarian Academy of Sciences, 107-121.

Hermjakob, U. (2009). Improved Word Alignment with Statistics and Linguistic Heuristics. *in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 6-7 août 2009, Singapoure. Stroudsburg (USA, PA): Association for Computational Linguistics, 2009, 229-237.

Hervey, S. G. J, Higgins, I. (2002). *Thinking French Translation: a course in translation method: French to English*. London and New York: Routledge, 2nd edition, 2002, 287 p.

Hovy, E. (1993). Pangloss: interlingua vivat?. *in* Brace, C. (eds). *Language Industry Monitor*, mars-avril 1993, 14, 10-11.

Hutchins, J. (1994). Vers une nouvelle époque en traduction automatique. *in* Clas, A., Bouillon, (éds). *Actes des Troisièmes Journées Scientifiques du réseau thématique*

« Lexicologie, Terminologie, Traduction » : TA-TAO : recherches de pointe et applications immédiates, 1993, Montréal (Canada), 3-16.

Hutchins, J., (1993). Latest developments in machine translation technology: beginning a new era in MT research. *in Proceedings of the Fourth Machine Translation Summit: MT Summit IV: International cooperation for global communication*, 20-22 juillet 1993, Kobe (Japon). Tokyo (Japon): AAMT, 1993, 11-34.

Hutchins, J., Somers, H. (1992). *An introduction to machine translation*. Londres (Angleterre) : Academic Press, 1992, 362 p.

Iancu, Ş. (2007). *Dezvoltarea ştiinţei şi tehnologiei informaţiei şi comunicaţiilor în România [Le développement de la science et de la technologie de l'information et des communications en Roumanie]*. Bucarest (Roumanie): Academia Română, 2007, *NOEMA*, vol. VI, 1-27.

Ide, N., Véronis, J. (1994). Multext (multilingual tools and corpora). *in Proceedings of the 15th CoLing*, 5-9 août 1994, Kyoto (Japon), 90-96.

Imamura, K. (2002). Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based mt. *in Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 13-17 mars 2002, Keihanna (Japon), 74-84.

Inkpen, D., Frunză, O., Kondrak, G. (2005). Automatic Identification of Cognates and False Friends in French and English. *in Proceedings of Recent Advances in Natural Language Processing (RANLP-2005)*, 21-23 septembre 2005, Borovets (Bulgarie), 251-257.

Ion, R. (2007). *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română [Méthodes de désambigüisation sémantique automatique. Application pour les langues anglaise et roumaine]*. Thèse de Doctorat. Bucarest (Roumanie) : Académie Roumaine, mai 2007, 148 p.

Irimia, E. (2008). Experimente de Traducere Automată Bazată pe Exemple. *in Actes de l'Atelier : Resurse Lingvistice Româneşti şi Instrumente pentru Prelucrarea Limbii Române [Ressources linguistiques roumaines et outils pour le traitement automatique du roumain]*, 19-21 novembre 2008, Iaşi (Roumanie), 131-140.

Kaji, H., Kida, Y., Morimoto, Y. (1992). Learning translation templates from bilingual text. *in Proceedings of the 14th conference on Computational linguistics (Coling 92)*, 1992, Nantes (France). Stroudsburg (USA, PA): Association for Computational Linguistics, 2, 672-678.

King, M. (1981). Eurotra–a european system for machine translation. *Lebende Sprachen*, 26(1), 12–14.

Kneser, R., Ney H. (1993). Improved clustering techniques for class-based statistical language modeling. *in Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH)*, septembre 1993, Berlin (Allemagne), 2, 973–976.

Knight, K. (1999). A Statistical MT Tutorial Workbook. *prepared in connection with the JHU summer workshop*, 30 avril 1999. [http://ccl.pku.edu.cn/doubtfire/NLP/Parsing/Unification_based_parsing/Kevin%20Knight.htm].

Koehn, (2004). Pharaoh: a Beam Search Decoder for Phrase-Based SMT. *in 6th Association for Machine Translation of the Americas (AMTA)*, 2004, Washington (USA, DC), 115-124.

Koehn, (2005). Europarl: A parallel corpus for statistical machine translation. *in MT Summit*, 5.

Koehn, (2010). *Statistical Machine Translation*. New York (USA): Cambridge University Press, janvier 2010, 433 p.

Koehn, P., Hoang, H. (2007). Factored translation models. *in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, juin 2007, Prague (République Tchèque), 868-876.

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., Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. *in Proceedings of the ACL 2007 Demo and Poster Sessions*, juin 2007, Prague (République Tchèque), 177-180.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical Phrase-Based Translation. *in Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 03)*, mai-juin 2003, Edmonton (Canada). Stroudsburg (USA, PA): Association for Computational Linguistics, 2003, 1, 48-54.

Kondrak, G., Marcu, D., Knight, K. (2003). Cognates can improve statistical translation models. *in Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, companion volume, Edmonton, Alberta, mai 2003., 46-48.

Kraif, O. (1999). Identification des cognats et alignement bi-textuel : une étude empirique. *in Actes de la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN),* 12-17 juillet 1999, Cargèse, 205-214.

Kraif, O. (2001). *Constitution et exploitation de bi-textes pour l'Aide à la traduction.* Thèse de Doctorat. Nice (France) : Université de Nice Sophia Antipolis, 29 juin 2001, 548 p.

Lakoff, G. (1972). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *in* Perantean, M., Levi J. N., and Phares G. C. (eds). *Papers from the 8th Regional Meeting*, Chicago Linguistics Society, 183-228.

Lambert, P., Banchs, R. (2005). Data inferred multi-word expressions for statistical machine translation. *Machine Translation Summit X*, Phuket (Thailand), 396–403.

Langlais, P., Gotti, F. (2006). EBMT by Tree-Phrasing. *Journal of Machine Translation*. Hingham (USA, MA): Kluwer Academic Publishers, mars 2006, 20(1), 1-23.

Lardilleux, A., Lepage, Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. *in Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008),* 2008, Waikiki (USA, Hawaï), 125-132.

Lavecchia, C. (2010). *Les Triggers Inter-langues pour la Traduction Automatique Statistique*. Thèse de Doctorat. Nancy (France) : Université Nancy II, 23 juin 2010, 151 p.

Le, H. S., Lavergne, T., Allauzen, A., Apidianaki, M., Gong, L., Max, A., Sokolov, A., Wisniewski, G., Yvon, F. (2012). Limsi@wmt12. *in Proceedings of the Seventh Workshop on Statistical Machine Translation*, 2012, Montréal (Canada), 330-337.

Le, H. S., Oparin, I., Allauzen, A., Gauvain, J.L., Yvon, F. (2011). Structured output layer neural network language model. *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11),* 22-27 mai 2011, Prague (République Tchèque), 5524-5527.

Léon, J. (2001). Conceptions du mot et débuts de la traduction automatique. *Histoire Épistémologie Langage*. Paris (France) : Société d'Histoire et d'Épistémologie des Sciences du Langage, 2001, 23(1), 81-106.

Léon, J. (2002). Le CNRS et les débuts de la traduction automatique en France. *La revue pour l'histoire du CNRS* [en ligne], 6, 2002. [http://histoire-cnrs.revues.org/3461].

Lepage, Y., Denoual, E. (2005). ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. *in Proceedings of the*

International Workshop on Spoken Language Translation (*IWSLT),* 2005, Pittsburgh (USA), 47-54.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *in Actes de Soviet Physics Doklady*, 1966, 10, 707–710.

L'Homme, M. C., Meynard, I. (1998). Le point d'accès aux combinaisons lexicales spécialisées : présentation de deux modèles informatiques. (sous la dir.) de Chapdelaine, A., *TTR : traduction, terminologie, rédaction* [en ligne et imprimé], 1er semestre 1998, 11(1), 199-227. Association canadienne de traductologie. [http://id.erudit.org/iderudit/037322ar 1708-2188].

Lita, L. V., Ittycheriah, A., Roukos, S., Kambhatla, N. (2003). tRuEcasIng. *in* Hinrichs, E., Roth, D. (eds). *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACl'03),* 7-12 juillet 2003, Sapporo (Japon). Stroudsburg (USA, PA): Association for Computational Linguistics, 2003, 152–159.

Loffler-Laurian, A. M. (1996). *La traduction automatique*. Villeneuve d'Ascq : Presses Universitaires du Septentrion, 156 p.

Maas, H. (1977). The saarbrücken automatic translation system (susy). *in Proceedings of the European Congress on Information Systems and Networks*: *Overcoming the language barrier*, 585–592.

Marcu, D., Munteanu, D. Ș. (2005). State of the Art in Statistical Machine Translation: An English-Romanian Experiment. *in 7th EuroLAN Biennial Summer School (EuroLAN 2005): The Multilingual Web: Resources, Technologies, and Prospects,* 26 juillet 2005, Cluj-Napoca (Roumanie).

Marcu, D., Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* 7-12 juillet 2002, Morristown (USA, NJ). Stroudsburg (USA, PA): Association for Computational Linguistics, 2002, 10, 133-139.

Marino, J. B., Banchs, R. E., Crego, M., de Gispert, A., Lambert, P., Fonollosa J. A. R., Marta, R. C. (2006). N-gram-based machine translation. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, décembre 2006, 32(4), 527-549.

Martin, J., Mihalcea, R., Pedersen, T. (2005). Word Alignment for Languages with Scarce Resources. *in Proceedings of the ACL2005 Workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond*, juin 2005, Ann Arbor (USA,

Michigan). Stroudsburg (USA, PA): Association for Computational Linguistics Association for Computational Linguistics, 2005, 65-74.

Melamed, D. I. (1998). Manual annotation of translational equivalence: The Blinker project. *Cognitive Science Technical Report*, 11 mai 1998, University of Pennsylvania, 13 p.

Melamed, D. I. (1999). Bitext Maps and Alignment via Pattern Recognition. *Journal of Computational Linguistics.* Cambridge (USA, MA): MIT Press, mars 1999, 25(1), 107-130.

Mihalcea, R., Pedersen, T. (2003). An Evaluation Exercise for Word Alignment. *in HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 31 mai 2003, Edmonton (Canada). Stroudsburg (USA, PA): Association for Computational Linguistics, 2003, 1-10.

Molina, L., Albir, A. H. (2002). Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta : journal des traducteurs / Meta: Translators' Journal*, décembre 2002, 47(4), 498-512.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. *in* Elithorn, A., Banerji, R. (eds). *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, 1984, Lyon (France). New-York (USA, NY): Elsevier Science Publishers. B. V., 1984, 173-180.

Nakazawa, T., Yu, K., Kawahara, D., Kurohashi, S. (2006). Example-based machine translation based on deeper NL*In Proceedings of the International. Workshop on Spoken Language Translation (IWSLT' 06): Evaluation Campaign on Spoken Language Translation*, novembre 2006, Kyoto (Japon), 64-70.

Namer, F. (2000). FLEMM : Un analyseur Flexionnel du Français à base de règles. *in* Jacquemin, C. (éd). *Traitement automatique des Langues pour la recherche d'information*. Paris (France) : Association pour le traitement automatique des langues, 2000, 41(2), 523-547.

Navlea, M., Havași, S.-F. (2012). A French-Romanian lexical alignment system using legal parallel corpora: Error Classification. *in Proceedings of The International Conference for Young Researchers Classification in Linguistics: Methodology, Problems, Challenges*, June 6-8, 2012, Strasbourg. (à paraître)

Navlea, M., Todirașcu, A. (2010a). Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems. *in Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010): Workshop on Exploitation of*

Multilingual Resources and Tools for Central and (South) Eastern European Languages, 17-23 mai 2010, Valletta (Malte), 2010, 41-48.

Navlea, M., Todirașcu, A. (2010b). Un sistem de traducere automată română - franceză [Un système de traduction automatique roumain - français]. *in* Iftene, A., Teodorescu, H. N., Cristea, D., Tufiş, D. (éds). *Actes de la 6ième Conférence : Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române [Ressources linguistiques et outils pour le traitement automatique du roumain]*, 6-7 mai 2010, Bucarest (Roumanie). Iași (Roumanie) : Ed. Universităţii « Al. I. Cuza », 2010, 165-174.

Navlea, M., Todirașcu, A. (2011a). Using Cognates in a French - Romanian Lexical Alignment System: A Comparative Study. *in* Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nikolai Nikolov (eds). *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria, September 2011. INCOMA Ltd., Shoumen, Bulgaria, 247-253.

Navlea, M., Todirașcu, A. (2011b). Cognate Identification for a French - Romanian Lexical Alignment System: Empirical Study. *in* Forcada, M. L., Depraetere, H., Vandeghinste, V. (eds). *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 2011)*, 30-31 mai 2011, Louvain (Belgique), 145-152.

Navlea, M., Todirașcu, A. (2011c). Repérage automatique des équivalences traductionnelles pour un système de traduction automatique statistique français - roumain. *in Actes du Colloque Métiers et technologies de la traduction : quelles convergences pour l'avenir ? (TRALOGY 2011)* [en ligne], 3-4 mars 2011, Paris (France). Mis à jour le : 11/01/2012, [http://lodel.irevues.inist.fr/tralogy/index.php?id=162].

Navlea, M., Todirașcu, A. (2011d). Ressources linguistiques pour un outil de traduction automatique statistique factorisée français - roumain. *in* Van Campenhoudt, M., Lino, T., Costa, R. (éds). *Actes des 8èmes Journées Scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction : Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face au défi de la diversité*, 15-17 octobre 2009, Lisbonne (Portugal). Paris (France) : Éditions des Archives Contemporaines et Agence universitaire de la Francophonie (AUF), 2011, 375-388.

Navlea, M., Todirașcu, A. (2011e). Identification de cognats à partir de corpus parallèles français - roumain. *in* Mathieu Lafourcade et Violaine Prince (éds). *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Volume 2/2, Montpellier, France, juin 2011, 81-86.

Navlea, M., Todirașcu, A. (2012). Using Cognates to Improve Lexical Alignment Systems. *in* Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (eds). *Text, Speech and Dialogue (15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings), Lecture Notes in Computer Science*, Volume 7499, Springer Verlag Berlin Heidelberg, 370-377.

Navlea, M., Todirașcu, A. (2013). Experiments with a French - Romanian Word Alignment System. *in* Dan Tufiș, Vasile Rus, and Corina Forăscu (eds). *Towards Multilingual Europe 2020: A Romanian Perspective*, Romanian Academy Publishing House, Bucharest, 2013, 225-240.

Navlea, M., Todirașcu, A. (2014). A Hybrid Word Alignment System for Statistical Machine Translation, *American Journal of Systems and Software, Special Issue on Multidisciplinary Perspectives of Agent-based Systems*. (à paraître)

Niessen, S., Vogel, S., Ney, H, Tillmann, C. (1998). A DP-based search algorithm for statistical machine translation. *in Proceedings* of the *36th Annual Conference Of the Association for Computational Linguistics and 17th Inernational Conference On Computational Linguistics (COLING-ACL'98)*, 10-14 août 1998, Université de Montréal, Montréal (Québec, Canada), 2, 960-967.

Oakes, M. (2000). Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics*, decembre 2000, 7(3), 233-243.

Och F. J., Ney. H. (2000). Improved Statistical Alignment Models. *in Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, octobre 2000, Hong Kong (Chine). Stroudsburg (USA, PA): Association for Computational Linguistics, 2000, 440-447.

Och, F. J. (2003). Minimal Error Rate Training in Statistical Machine Translation. *in Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACl'03)*, 7-12 juillet 2003, Sapporo (Japon). Stroudsburg (USA, PA): Association for Computational Linguistics, 2003, 160-167.

Och, F. J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT Press, mars 2003, 29(1), 19-51.

Och, F. J., Tillmann, C., Ney, H. (1999). Improved alignment models for statistical Machine Translation. *in Proceedings of the Joint Conference of Empirical Methods in Natural*

Language Processing and Very Large Corpora, 1999, University of Maryland, College Park (USA, MD), 20-28.

Olteanu, M., Davis, C., Volosen, I., Moldovan, D. (2006). Phramer - an open source statistical phrased-based translator. *in Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL): Workshop on Statistical Machine Translation*, 4-9 juin 2006, New York (USA, NY), 150-153.

Ozdowska, S. (2006). *ALIBI : un système d'ALIgnement BIlingue à base de règles de propagation syntaxique.* Thèse de Doctorat. Toulouse (France) : Université de Toulouse II-Le Mirail, 14 décembre 2006, 196 p.

Pal, S., Naskar, S. K., Bandyopadhyay, S. (2013). A Hybrid Word Alignment Model for Phrase-Based Statistical Machine Translation. *in Proceedings of the Second Workshop on Hybrid Approaches to Translation*, 8 août 2013, Sofia (Bulgarie), 94-101.

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *in Proceedings of 40th Annual meeting of the Association for Computational Linguistics* (*ACL)*, juillet 2002, Philadelphia (USA, PE). Stroudsburg (USA, PA): Association for Computational Linguistics, 2002, 311-318.

Patry, A., Gotti, F., Langlais, (2006). MOOD: A modular object-oriented decoder for statistical machine translation. *in Proceedings of the 5th Linguistic Resources and Evaluation Conference (LREC)*, 22-28 mai 2006, Gênes (Italie). Paris (France): European Language Ressources Association (ELRA), 2006, 709-714.

Patry, A., Langlais, (2005). Paradocs : un système d'identification automatique de documents parallèles. *in Actes de la 12e conference sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, 6-10 juin 2005, Dourdan (France), 223-232.

Polguère, A. (2003). *Lexicologie et sémantique lexicale : Notions fondamentales*. Montréal (Canada) : Presses de l'Université de Montréal, 2003, 260 p.

Popovic, M., Ney, H. (2006). POS-basedWord Reorderings for Statistical Machine Translation. *in Proceedings of the Linguistic Resources and Evaluation Conference (LREC 2006)*, 22-28 mai 2006, Gênes (Italie). Paris (France) : European Language Ressources Association (ELRA), 2006, 1278-1283.

Quasthoff, U. (1998). Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values. *in Proceedings of the First International Conference on Language Resources & Evaluation (LREC)*, Granada, mai 1998, Vol. II, 853-856.

Ramisch, C., Besacier, L., Kobzar, A. (2013). How hard is it to automatically translate phrasal verbs from English to French?. *in* Monti, J., Mitkov, R., Corpas Pastor, G., Seretan, V. (eds). *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, 3 septembre 2013, Nice (France), 53-61.

Ramisch, C., Villavicencio, A., Boitet, C. (2010). MWEToolkit: A Framework for Multiword Expression Identification. *in Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 10),* 17-23 mai 2010, Valletta (Malte). Paris (France): European Language Resources Association (ELRA), 2010, 662-669.

Ramos, M. A., Nishikawa, A., Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. *in* Granger, S., Paquot, M. (eds). *Proceedings of ELEX2009 : E-lexicography in the 21st century: new challenges, new applications*., 22-24 octobre 2009, Louvain-la-Neuve (Belgique) : Presses Universitaires de Louvain, Cahiers du Cental, 7, 369-374.

Rastier, F., Cavazza, M., Abeille, A. (1994). *Sémantique pour l'analyse, de la linguistique à l'informatique*. Paris (France) : Masson, 1994, 240 Collection Sciences Cognitives.

Ren, Z., Lü, Y., Cao, J., Liu, Q., Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *in* Anastasiou, D., Chikara, H., Preslav, N., Su Nam, K. (eds). *Proceedings of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009),* août 2009, Suntec (Singapore). Stroudsburg (USA, PA): Association for Computational Linguistics, 2009, 47-54.

Riegel, M., Pellat, J.C., Rioul R. (2009). *Grammaire méthodique du français*. Paris (France) : Presses universitaires de France, 1107 p.

Ritz, J., Heid, U. (2006). Extracting tools for collocations and their morphosyntactic specificities. *in Proceedings of the Linguistic Resources and Evaluation Conference (LREC),* 22-28 mai 2006, Gênes (Italie). Paris (France) : European Language Ressources Association (ELRA), 2006, 1925-1930.

Sadat, F., Foster, G., Kuhn, R. (2006). Système de traduction automatique statistique combinant différentes ressources. *in Actes de la 13^{e} édition conférence sur le Traitement Automatique des Langues Naturelles (TALN),* 10-13 avril 2006, Louvain (Belgique). Louvain (Belgique) : Presses Universitaires de Louvain, Cahiers du Cental, 2(1), 590-627.

Schmidt, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *in Meeting of the proceedings of the International Conference on New Methods Language Processing*, 1994, Manchester (UK), 44-49.

Schmidt, H., Schulte im Walde, S. (2000). Robust German noun chunking with a probabilistic context-free grammar. *in Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, 31 juillet - 4 août 2000, Saarbrücken (Allemagne). Stroudsburg (USA, PA): Association for Computational Linguistics, 2, 726-732.

Schrader, B. (2006). ATLAS: a new text alignment architecture. *in Proceedings of the COLING/ACL 2006: Main Conference Poster Sessions*, juillet 2006, Sydney (Australie). Stroudsburg (USA, PA): Association for Computational Linguistics, 2006, 715-722.

Schulze-Furhoff, W., Abbou, A. (1992). Aspects de l'ingénierie linguistique en Allemagne. *La Tribune des Industries de la Langue*, 7-8, 18-24.

Schwenk, H., Abdul Rauf, S., Barrault, L., Senellart, J. (2009). SMT and SPE Machine Translation Systems for WMT'09. *in Proceedings of the Fourth Workshop on Statistical Machine Translation*, mars 2009, Athènes (Grèce). Stroudsburg (USA, PA): Association for Computational Linguistics, 2009, 130-134.

Schwenk, H., Dechelotte, D., Bonneau-Maynard, H., Allauzen, A. (2007). Modèles statistiques enrichis par la syntaxe pour la traduction automatique. *in Actes de TALN 2007*, 5-8 juin 2007, Toulouse (France), 253-262.

Seretan, V. (2008). *Collocation Extraction Based on Syntactic Parsing*. Thèse de Doctorat. Genève (Suisse) : Faculté des Lettres de l'Université de Genève, 9 juin 2008, 249 p.

Seymore, K., Rosenfeld, R. (1997). Using story topics for language model adaptation. *in* Kokkinakis, G., Fakotakis, N., Dermatas, E. (eds). *Proceedings of the Fifth European Conference on Speech Communication and Technology (EUROSPEECH 1997)*: *Language Modelling*, 22-25 septembre 1997, Rhodes (Grèce), 1987-1990.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, juillet 1948, 27, 379-423, 623-656.

Simard, M., Foster, G., Isabelle, (1992). Using cognates to align sentences. *in Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992, Montréal (Canada), 67-81.

Smadja, F., McKeown, K. (1990). Automatically extracting and representing collocations for language generation. *in Proceedings of the 28th annual meeting on Association for*

Computational Linguistics, 1990, Pittsburgh (USA, PE). Stroudsburg (USA, PA): Association for Computational Linguistics, 1990, 252-259.

Smith, A., Cohn, T., Osborne, M. (2005). Logarithmic opinion pools for conditional random fields. *in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 25-30 juin 2005, Ann Arbor (Michigan), 18-25.

Ştefan, I. M., Nicolau, E. (1981). *Scurtă istorie a creaţiei ştiinţifice şi tehnice româneşti, [Courte histoire de la création scientifique et technique roumaine]*. Bucureşti (Roumanie): Albatros, 1981, 1, 286 p.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. *in Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, 21-27 mai 2012, Istanbul (Turquie), 454-459.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 22-28 mai 2006, Gênes (Italie). Paris (France) : European Language Ressources Association (ELRA), 2006, 2142-2147.

Stoichițoiu-Ichim, A. (2000). *Semiotica discursului juridic [Sémiotique du discours juridique]*. Bucarest (Roumanie): Editura Universității din București, 21 juin 2011, 217 p.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. *in Proceedings International Conference Spoken Language Processing*, septembre 2002, Denver (USA, Colorado), 901-904.

Stymne, S. (2011). BLAST: A tool for error analysis of machine translation output. *in Proceedings of the ACL 2011 System Demonstrations*, juin 2011, Portland (USA, OR). Stroudsburg (USA, PA): Association for Computational Linguistics, 2011, 56-61.

Surdeanu, M., Turmo, J. (2005). Semantic Role Labeling Using Complete Syntactic Analysis. *in Proceedings of 9th Conference on Computational Natural Language Learning (CoNLL)*, 29-30 juin 2005, Ann Arbor (Michigan), 221-224.

Tanguy, L., Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. *in Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, 24-27 juin 2002, Nancy (France).

Paris (France) : Association pour le traitement automatique des langues (ATALA), 2002, 1, 245-254.

Tiedemann, J. (1999). Word alignment -step by ste*In Proceedings of the 12th Nordic Conference on Computational Linguistics*, University of Trondheim (Norvège), 216-227.

Tiedemann, J. (2003). Combining Clues for Word Alignment. *in Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 12-17 avril 2003, Budapest (Hongrie), 339-346.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *in* Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R., (eds). *Recent Advances in Natural Language Processing*, 2009, 5, 237-248.

Tillman, C. A. (2004). Unigram Orientation Model for Statistical Machine Translation. *in Proceedings of the HLTNAACL: short paper*, mai 2004, Boston (USA, MA). Stroudsburg (USA, PA): Association for Computational Linguistics, 2004, 101-104.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. (1997). A DP-based search using monotone alignments in statistical translation. *in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL'97)*, 7-12 juillet 1997, Madrid (Espagne). Stroudsburg (USA, PA): Association for Computational Linguistics, 1997, 289-296.

Todirașcu, A., Grass, T., Navlea, M., Longo, L. (2014). La relation de hiérarchie « chef » : une approche translingue français - anglais - allemand. *META*, 59 (2). (à paraître)

Todirașcu, A., Heid, U., Ştefănescu, D., Tufiş, D., Gledhill, C., Weller, M., Rousselot, F. (2008). Vers un dictionnaire de collocations multilingue. *in* Escoda, X.B., L'Homme, M.C., Van Campenhoudt, M. (éds). *Cahiers de Linguistique* : *Lexique, dictionnaire et connaissance dans une société multilingue.* Cortil-Wodon : Éditions Modulaires Européennes EME, août 2008, 33(1), 161-186.

Todirașcu, A., Ion, R., Navlea, M., Longo, L. (2011). French text preprocessing with TTL. *in Proceedings of the Romanian Academy*. Bucarest (Roumanie): The Publishing House of the Romanian Academy, avril-juin 2011, série A, 12(2), 151-158.

Todirașcu, A., Navlea, M. (2010). CAP: A Hierarchical Lexical Function Related to Proper Nouns: The Case of Romanian and French. *in* Tufiş, D., Forăscu, C. (eds). *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Bucarest (Roumanie): The Publishing House of the Romanian Academy, décembre 2010, 317-330.

Trandabăț, D., Irimia, E., Barbu M. V., Cristea, D., Tufiș, D. (2012). The Romanian Language in the Digital Age. Limba română în era digitală. *in* Rehm, G., Uszkoreit, H. (eds). *White Papers Series.* Berlin (Heidelberg): Springer-Verlag, 2012, 93 p.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. *in* Jelinek, F., Nth, E. (eds). *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*. Berlin (Heidelberg): Springer-Verlag, 28-33.

Tufiş, D. (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. *International Conference on Language Resources and Evaluation (LREC'2000)*, 31 mai - 2 juin 2000, Athens (Grèce), 1105-1112.

Tufiş, D., Barbu, A. M. (1997). A Reversible and Reusable Morpho-Lexical Description of Romanian. *in* Tufiş, D., Andersen, (eds). *Recent Advances in Romanian Language Technology*. Bucarest (Roumanie): Editura Academiei Române, 1997, 83-93.

Tufiş, D., Ceauşu, A., (2008). DIAC+: A Professional Diacritics Recovering System. *in Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, mai 2008, Marrakech (Maroc). Paris (France): European Language Ressources Association (ELRA), 2008, 167-174.

Tufiş, D., Ceauşu, A., Ion, R., Ştefănescu, D. (2005a). An integrated platform for high-accuracy word alignment. *in* Proceedings of *JRC Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages*, 26-27 septembre 2005, Arona (Italie).

Tufiş, D., Dumitrescu, Ş. D. (2012). Cascaded Phrase-Based Statistical Machine Translation Systems. *in* Cettolo, M., Federico, M., Specia, L., Way, A. (eds). *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012)*, 28-30 mai 2012, Trento (Italie), 129-136.

Tufiş, D., Ion, R., Ceauşu, A. Ştefănescu, D. (2005b). Combined Aligners. *in Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data- Driven Machine Translation and Beyond*, 29-30 juin 2005, Ann Arbor (USA). Stroudsburg (USA, PA): Association for Computational Linguistics, 2005, 107-110.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. *in* Ishida, T., Fussell, S. R., Vossen T. J. M. (eds). *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, avril 2006, Trento (Italie). Stroudsburg (USA, PA): Association for Computational Linguistics, 2006, 153-160.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008a). RACAI's Linguistic Web Services. *in Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, mai 2008, Marrakech (Maroc). Paris (France): European Language Ressources Association (ELRA), 2008, 327-333.

Tufiș, D., Ion, R., Dumitrescu, Ș. D. (2013a). Wikipedia as an SMT Training Corpus. *in Proceedings of the International Conference on Recent Advances on Language Technology (RANLP 2013)*, 7-13 septembre 2013, Hissar (Bulgarie), 702-709.

Tufiș, D., Ion, R., Dumitrescu, Ș. D. (2013b). Wiki-Translator: Multilingual Experiments for In-Domain Translations. *in Proceedings of the International Conference on Intelligent Systems*, août 2013, Chișinău (Republique de Moldavie), 19-47.

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., Krstev, C. (2008b). Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. *in* Tadić, M., Dimitrova-Vulchanova, M., Koeva, S. (eds). *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, 25-28 septembre 2008, Dubrovnik (Croatie), 145-152.

Tutin, A. (2004). Pour une modélisation dynamique des collocations dans les textes. *in Proceedings of the European Association for Lexicography (Euralex)*, 6-10 juillet 2004, Lorient (France), 207-219.

Utsuro, T. Matsumoto, Y., Nagao, M. (1992). Lexical knowledge acquisition from bilingual corpora. *in Proceedings of the 14th Conference on Computational Linguistics (Coling 92)*, 1992, Nantes (France). Stroudsburg (USA, PA): Association for Computational Linguistics, 1992, 2, 581- 587.

Veale, T., Way, A. (1997). Gaij*in* A Bootstrapping, Template-Driven Approach to Example-Based Machine Translation. *in Proceedings of the International Conference, Recent Advances in Natural Language Processing*, septembre 1997, Sofia (Bulgarie), 239-244.

Vertan, C., Gavrilă, M. (2010). Multilingual applications for rich morphology language pairs, a case study on German Romanian. *in* Tufiș, D., Forăscu, C. (eds). *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Bucarest (Roumanie): The Publishing House of the Romanian Academy, 2010, 448-460.

Vinay, J. P., Darbelnet, J. (1977). *Stylistique comparée du français et de l'anglais : méthode de traduction*. Paris (France) : Didier, 1977, 2e édition, 1, 331 p.

Vogel, S., Ney, H., Tillmann, C. (1996). HMM-based word alignment in statistical translation. *in Proceedings of the The 16th Inernational Conference on Computational Linguistics (COLING '96)*, août 1996, Copenhagen (Danemark). Stroudsburg (USA, PA): Association for Computational Linguistics, 1996, 836-841.

Wagner, R. A., Fischer, M. J. (1974). The String-to-String Correction Problem. *Journal of the ACM*. New-York (USA, NY): ACM, janvier 1974, 21(1), 168-173.

Wang, W., Knight, K., Marcu, D. (2006). Capitalizing machine translation. *in Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 5-9 juin 2006, New-York (USA, NY), 1-8.

Wehrli, E., Seretan, V., Nerima, L., Russo, L. (2009). Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy. *in* Màrquez L. and Somers H. (eds). *Proceedings of the 13th Annual Conference of the European Association for Machine Translation* (*EAMT)*, 14-15 mai 2009, Barcelona, Spain, 128-135.

Weller, M., Fraser, A., Heid, U. (2014). Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT. *in Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 15-18 juin 2014, Dubrovnik (Croatia), 11-18.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Journal of Computational Linguistics*. Cambridge (USA, MA): MIT press, septembre 1997, 23(3), 377-403.

Yamada, K., Knight, K. (2001). A syntax-based statistical translation model. *in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01)*, 9-11 juillet 2001, Toulouse (France). Stroudsburg (USA, PA): Association for Computational Linguistics, 523-530.

Zens, R., Och, F. J., Ney, H. (2002). Phrase-based Machine Translation. *in* Lakemeyer, G., Jarke, M., Koehler, J. (eds). *Proceedings of the 25th Annual German Conference on AI: KI 2002: Advances in Artificial Intelligence*, 16-20 septembre 2002, Aachen (Allemagne). Berlin, Heidelberg: Springer-Verlag, 2002, LNAI 2479, 18-32.

Zhang, Y., Vogel, S., Waibel, A. (2003). Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation. *in Proceedings of International Conference on*

Natural Language Processing and Knowledge Engineering, 26-29 octobre 2003, Beijing (Chine), 567-573.

Dictionnaires

Coteanu, I., Mareş, L. (sous la dir.). *Dicţionarul explicativ al limbii române* (Le Dictionnaire explicatif du roumain), 2ème éd., Ed. Univers enciclopedic, Bucureşti, 1998.

Rey-Debove, J., Rey, A. (sous la dir.). *Le Nouveau Petit Robert*, Dictionnaires Le Robert, Paris, 2002.

Elena-Mirabela NAVLEA

La traduction automatique statistique factorisée :

une application à la paire de langues français - roumain

Résumé : Un premier objectif de cette thèse est la constitution de ressources linguistiques pour un système de traduction automatique statistique factorisée français - roumain. Un deuxième objectif est l'étude de l'impact des informations linguistiques exploitées dans le processus d'alignement lexical et de traduction. Cette étude est motivée, d'une part, par le manque de systèmes de traduction automatique pour la paire de langues étudiées et, d'autre part, par le nombre important d'erreurs générées par les systèmes de traduction automatique actuels. Les ressources linguistiques requises par ce système sont des corpus parallèles alignés au niveau propositionnel et lexical. Ces corpus sont également segmentés lexicalement, lemmatisés et étiquetés au niveau morphosyntaxique.

Mots-clés : traduction automatique statistique, modèles factorisés, modèles de traduction, modèles de langue, corpus parallèles alignés, alignement lexical, systèmes à base de séquences, langues romanes

Abstract: Our first aim is to build linguistic resources for a French-Romanian factored phrase-based statistical machine translation system. Our second aim is to study the impact of exploited linguistic information in the lexical alignment and translation process. On the one hand, this study is motivated by the lack of such systems for the studied languages. On the other hand, it is motivated by the high number of errors provided by the current machine translation systems. The linguistic resources required by the system are tokenized, lemmatized, tagged, word, and sentence-aligned parallel corpora.

Keywords: statistical machine translation, factored models, translation models, language models, aligned parallel corpora, lexical alignment, phrase-based systems, romance languages