Repérage automatique des équivalences traductionnelles pour un système de traduction automatique statistique français -roumain. Session 6 - Traduction et traitement automatique des langues (TAL)

Mirabela Navlea, Amalia Todirascu

TRALOGY

Repérage automatique des équivalences traductionnelles pour un système de traduction automatique statistique français - roumain

Mirabela Navlea
Université de Strasbourg, LiLPa (Linguistique, Langues, Parole)

Amalia Todiraşcu
Université de Strasbourg, LiLPa (Linguistique, Langues, Parole)

TRALOGY I - Session 6
Date d'intervention : 04/03/2011

Nous présentons un projet de recherche ayant comme objectif le développement de ressources linguistiques pour un système de traduction automatique statistique factorisée français - roumain. Ce système utilise des corpus parallèles annotés et alignés aux niveaux propositionnel et lexical et une combinaison de facteurs linguistiques (lemmes, catégories lexicales, propriétés morphosyntaxiques, chunks). Ainsi, nous nous sommes concentrés sur l'alignement lexical des corpus parallèles, en exploitant les informations linguistiques associées aux unités lexicales. Nous avons procédé à l'analyse linguistique des résultats du module d'alignement et nous avons repéré plusieurs classes d'erreurs d'alignement lexical, dues principalement aux différences morphosyntaxiques entre le français et le roumain. Ainsi, nous avons relevé des erreurs fréquentes d'alignement au niveau des subordonnées relatives, de l'expression de la relation de possession et du destinataire (le cas datif), de l'infinitif, etc. Pour améliorer les résultats du module d'alignement lexical, nous avons défini un ensemble de règles heuristiques morphosyntaxiques contextuelles. Dans cet article, nous présentons la distribution par classes des erreurs d'alignement lexical et les règles heuristiques proposées pour repérer automatiquement les équivalences traductionnelles d'une langue à l'autre.

lien video : https://webcast.in2p3.fr/video/reperage_automatique_des_equivalences_traductionnellest

Introduction

Notre article présente un projet dont l'objectif est le développement de ressources linguistiques pour un système de traduction automatique statistique factorisée français - roumain. Ainsi, nous étudions l'influence de plusieurs catégories d'informations linguistiques (catégorie lexicale, propriétés morphosyntaxiques) sur la qualité des traductions fournies par le système.

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, vu 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 flexionnaire complexe, comme les langues latines, slaves et balkaniques. Pour ces langues, la constitution de ressources linguistiques (dictionnaires, grammaires, bases de données terminologiques) nécessite du temps et des efforts humains et matériels importants. Concernant le roumain, les ressources linguistiques et les outils disponibles sont peu nombreux. Une description détaillée des outils développés figure dans [Tufiş, D. *et al*., (2008b)]. Notons que la plupart des systèmes de traduction automatique existants disposent des ressources pour la paire de langue anglais - roumain [Marcu, D. & Munteanu, D., S. (2005)], [Irimia, E. (2008)], [Ceauşu, A. (2009)]. D'autres systèmes se concentrent sur la paire de langues allemand - roumain [Gavrilă, M. (2009)], [Vertan, C. & Gavrilă, M. (2010)] ou français - roumain [Navlea, M. & Todiraşcu, A. (2010)].

Les systèmes de traduction automatique commettent un nombre important d'erreurs de traduction dues 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, phraséologie, etc.). [Grass, T. (2009)] identifie treize types d'erreurs fréquentes générées par les outils de traduction automatique : la polysémie et l'homonymie, l'ambigüité (syntaxique, référentielle), les termes vagues (*fuzzy hedges*), les expressions et les métaphores, les néologismes, les noms propres, les mots d'origine étrangère, les emprunts et les calques, les séparateurs, les sigles et les acronymes, la transposition. Pour éviter ces catégories d'erreurs, les systèmes linguistiques de traduction automatique se concentrent sur le développement de ressources linguistiques complexes, telles que : les grammaires, les dictionnaires, les bases de données terminologiques ou de connaissances.

Si les systèmes linguistiques *(Systran*[1]) donnent ainsi des résultats de traduction performants, d'autres systèmes fournissent des résultats comparables en s'appuyant sur des techniques statistiques factorisées (*EuroMatrix*[2], 2009) qui exploitent des corpus parallèles annotés et alignés. Par rapport aux méthodes basées sur les segments de traduction [Koehn, P., Och, F. J. & Marcu, D. (2003)] utilisant seulement la forme d'occurrence des mots, les systèmes factorisés prennent en compte plusieurs facteurs linguistiques associés aux unités lexicales, tels que : lemmes, propriétés morphosyntaxiques, informations syntaxiques, etc. Ainsi, les systèmes factorisés sont modulaires : différents facteurs linguistiques peuvent être utilisés dans le processus de traduction. [Koehn, P. & Hoang, H. (2007)] utilisent les propriétés morphosyntaxiques des unités lexicales, [Avramidis, E. & Koehn, P. (2008)] exploitent l'information syntaxique pour améliorer la qualité des résultats de traduction.

Dans notre projet, nous développons un système de traduction automatique statistique factorisée (implémenté initialement pour l'anglais et le roumain [Ceauşu, A. (2009)], à l'Académie Roumaine de Bucarest[3]) pour la paire de langues français - roumain. Ce système utilise des

(1) http://www.systransoft.com/
(2) http://www.euromatrix.net/
(3) http://www.racai.ro/webservices

corpus parallèles annotés et alignés aux niveaux propositionnel et lexical et une combinaison de facteurs linguistiques (lemmes, catégories lexicales, propriétés morphosyntaxiques, chunks). Nous étudions ainsi l'influence des facteurs linguistiques utilisés sur la qualité des traductions français - roumain, en exploitant des corpus disponibles pour cette paire de langues (domaines : juridique, politique, aviation).

Notre étude adopte la méthodologie proposée dans le projet *SEE-ERA.net* [Tufiş, D. *et al.*, (2008a)] 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 annotés et alignés aux niveaux propositionnel et lexical.

Dans la section suivante, nous présentons l'architecture du système de traduction automatique statistique factorisée et les informations linguistiques exploitées pour construire des modèles de langue et de traduction pour les deux langues étudiées. Nous décrivons les corpus utilisés dans la section 3. Nous présentons la méthode d'alignement lexical développée dans la section 4. Nous discutons la distribution par classes d'erreurs d'alignement lexical repérées, ainsi que les règles heuristiques proposées pour les résoudre, dans la section 5. Nous présentons également nos conclusions et nos perspectives dans la section 6.

1. La méthodologie

Dans notre projet, nous adaptons un système de traduction automatique statistique factorisée [Ceauşu, A. (2009)] pour la paire de langues français - roumain. Ce système utilise un corpus parallèle annoté et aligné aux niveaux propositionnel et lexical. Le corpus est annoté au niveau morphosyntaxique (par l'ensemble d'étiquettes MSD[4] du projet Multext[5] pour le français [Ide, N. & Véronis, J. (1994)] et le roumain [Tufiş, D. & Barbu, A. M. (1997)], lemmatisé et dispose d'une analyse syntaxique partielle, le chunking. Ainsi, nous utilisons une combinaison de facteurs linguistiques associés aux unités lexicales : les lemmes, les étiquettes morphosyntaxiques, les chunks.

Le système initial (construit pour l'anglais et le roumain) utilise le décodeur *MOSES* [Koehn, P. *et al*. (2007)] avec différentes configurations de paramètres linguistiques optimisés (lemmes et descriptions morphosyntaxiques) établies en fonction du sens du processus de traduction. Ce système s'avère efficace principalement pour des textes du domaine juridique et administratif [Ceauşu, A. (2009)].

Pour adapter *MOSES* à une nouvelle paire de langues, dans notre cas pour le français et le roumain, il est nécessaire de construire un modèle de langue de la langue cible, en utilisant un corpus monolingue annoté, et un modèle de traduction factorisé, en utilisant un corpus parallèle annoté et aligné aux niveaux propositionnel et lexical. Ensuite, le décodeur *MOSES* cherche la traduction la plus probable en utilisant les modèles de langue et de traduction construits auparavant.

Afin de prétraiter les corpus utilisés, nous appliquons des outils que nous allons décrire plus loin. Ainsi, pour annoter les corpus, nous appliquons l'étiqueteur *TTL*[6] [Ion, R. (2007)]. Pour effectuer l'alignement lexical des corpus parallèles, nous utilisons *GIZA++* [Och, F. J. & Ney, H. (2000), (2003)]. De plus, afin de corriger les erreurs d'alignement lexical, nous avons implémenté un module supplémentaire utilisant des règles heuristiques morphosyntaxiques contextuelles. Ces règles sont définies à l'issue d'une analyse linguistique détaillée des erreurs générées par le système d'alignement.

(4) Morpho-Syntactic Descriptors
(5) http://aune.lpl.univ-aix.fr/projects/multext/
(6) Tokenizing, Tagging and Lemmatizing free running texts

L'architecture du système de traduction présenté apparaît dans la Figure 1 :

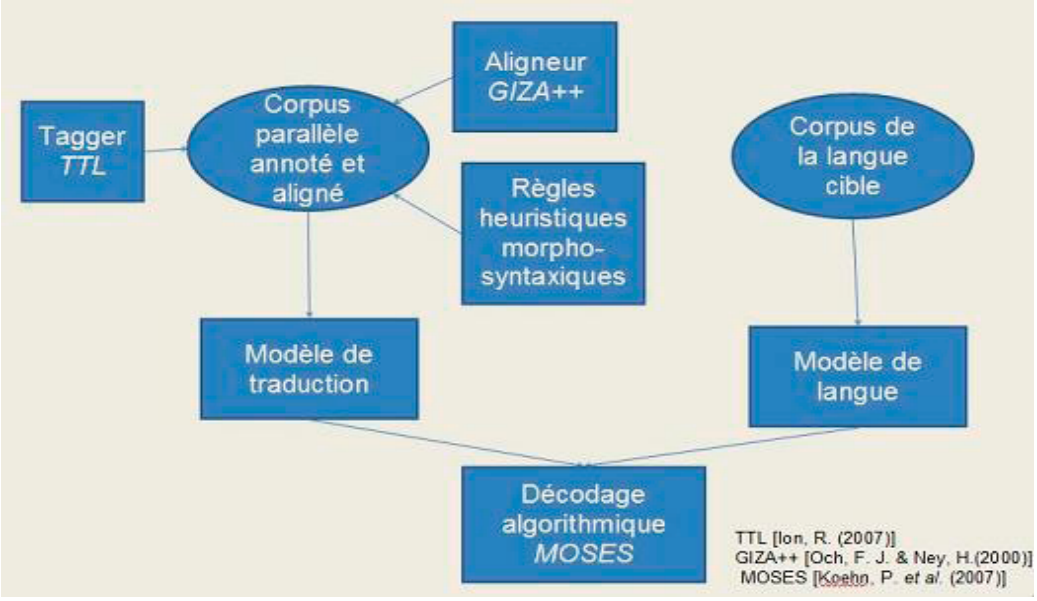

Figure 1 : L'architecture du système de traduction automatique statistique factorisée

Pour développer le système de traduction français - roumain, nous prenons en considération les étapes suivantes :

1) la constitution des corpus parallèles ;
2) le prétraitement des corpus (segmentation lexicale, lemmatisation, annotation morphosyntaxique et au niveau des chunks) ;
3) l'alignement propositionnel et lexical des corpus parallèles ;
4) l'analyse linguistique détaillée des erreurs d'alignement ;
5) l'application des règles de correction des erreurs d'alignement lexical ;
6) la construction de modèles de langue de la langue cible et de modèles de traduction factorisés ;
7) la configuration du système avec les facteurs linguistiques les plus pertinents (l'optimisation du système) ;
8) l'évaluation du système.

Dans la section suivante, nous présentons les corpus monolingues et parallèles utilisés, nécessaires pour construire les modèles de langue et les modèles de traduction français - roumain.

2. Les corpus

Nous utilisons un corpus parallèle extrait de *JRC-Acquis* [Steinberger, R. *et al.* (2006)], qui est basé sur le corpus parallèle multilingue *Acquis Communautaire* composé de la législation de l'UE des années 1950 jusqu'à présent. Ce corpus est disponible pour 231 paires de langues obtenues à partir de 22 langues officielles de l'UE.

JRC-Acquis est aligné au niveau de paragraphes et il est disponible gratuitement en format XML. Pour notre projet, nous en avons extrait un sous-ensemble de 228174 paires de phrases alignées 1 :1, choisies parmi l'ensemble de documents communs en français et en roumain.

Nous utilisons également un corpus parallèle français - roumain extrait de la mémoire de traduction *DGT-TM* (*DGT Translation Memory)*[7] basée aussi sur l'*Acquis Communautaire*. Ce corpus est aligné au niveau propositionnel et a l'avantage que la plupart des alignements sont réalisés manuellement. *DGT-TM* est disponible gratuitement en format TMX. Nous en avons extrait 490962 paires de phrases alignées 1 :1, communes en français et en roumain.

Comme *JRC-Acquis* et *DGT-TM* sont des corpus spécialisés du domaine juridique et administratif, nous avons constitué d'autres corpus parallèles en utilisant des ressources disponibles sur le Web, afin de tester le système de traduction automatique pour d'autres domaines (politique, aviation). Nous avons constitué les corpus parallèles manuellement, en prenant en compte les critères suivants : la disponibilité des textes bilingues, la fiabilité des sources, la qualité des traductions et le domaine. Pour chaque texte, nous avons spécifié l'auteur, la date et sa source indiquée par son URL. Les données collectées ont été nettoyées en éliminant les éléments non-textuels, tels que : les images, les notes de bas de page, les 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+* [Tufiş, D. & Ceauşu, A. (2008c)]. Nous avons aligné ces corpus en utilisant *Alinea* [Kraif, O. (2001)]. Comme cet aligneur n'intègre pas la paire de langues français - roumain, nous avons exploité ses paramètres par défaut que nous avons optimisés pour la paire de langues étudiées.

Ainsi, les textes bilingues disponibles en français et en roumain ont été collectés en utilisant les sites Web du Parlement Européen[8] et de la Commission Européenne[9], ainsi que des sites Web des compagnies aériennes roumaines (*Blue Air*[10], *TAROM*[11]).

Nous décrivons les corpus parallèles utilisés dans le Tableau 1 :

Tableau 1 : Les corpus parallèles français - roumain

Source du corpus	Nombre de mots / français	Nombre de mots / roumain
JRC-Acquis	5 828 169	5 357 017
DGT-TM	9 953 360	9 142 291
Site Web du Parlement Européen	137 422	126 366
Site Web de la Commission Européenne	200 590	185 476
Sites Web des compagnies aériennes roumaines	33 757	29 596

Afin de construire des modèles de langues, nous utilisons les parties monolingues des corpus parallèles, ainsi que d'autres corpus monolingues disponibles par demande aux auteurs. Pour le roumain, nous utilisons les corpus suivants :

- la partie roumaine du corpus journalistique parallèle *NAACL* (800 000 mots) [Martin, J., Mihalcea, R. & Pedersen, T. (2005)] ;
- le corpus *LT4eL* (composé des manuels d'utilisation en informatique, 600 000 mots) [Trandabăş, D*. et al.* (2006)] ;
- le corpus journalistique *RoCo* (7,5 millions de mots) [Tufiş, D. & Irimia, E. (2006)] ;

(7) http://langtech.jrc.it/DGT-TM.html
(8) http://www.europarl.europa.eu/parliament/public/staticDisplay.do?id=146&language=fr
(9) http://ec.europa.eu/index_fr.htm
(10) http://www.blueairweb.com/Page-D-Accueil/
(11) http://www.tarom.ro/

Nous utilisons également pour le français les corpus monolingues ci-dessous :

i. un corpus journalistique extrait du corpus *Le Monde* (1980 - 1988) (488 543 mots) ;
ii. un corpus juridique et administratif extrait de l'*Acquis Communautaire* (498 788 mots).

Pour évaluer le système de traduction automatique, nous utilisons un corpus parallèle français - roumain, aligné au niveau propositionnel et lexical, comprenant 1000 phrases [Todiraşcu, A. *et al.* (2008)]. Ce corpus a été obtenu par un processus de dérivation [Tufiş, D. & Koeva, S. (2007)], en utilisant deux corpus parallèles (anglais - roumain et anglais - français) extraits de *JRC-Acquis* et alignés automatiquement au niveau lexical. Le corpus résultant a été corrigé manuellement.

Comme nous l'avons déjà mentionné dans la section précédente, nous effectuons le prétraitement des corpus en appliquant l'étiqueteur *TTL* [Ion, R. (2007)] disponible en français et en roumain (comme service Web). Nous avons développé les ressources linguistiques pour l'annotation en français, vu que *TTL* était initialement disponible seulement pour l'anglais et le roumain. Les corpus monolingues français présentés auparavant ont été étiquetés, lemmatisés et corrigés manuellement. Ces corpus ont été utilisés pour entrainer *TTL,* qui est un étiqueteur probabiliste. Ainsi, les corpus sont tokenisés, lemmatisés et annotés par des descriptions morphosyntaxiques et au niveau des chunks (découpage en groupes nominaux simples, groupes prépositionnels). Les résultats fournis par *TTL* sont en format XCES et comprennent l'ensemble d'étiquettes du projet Multext pour le français [Ide, N. & Véronis, J. (1994)] et pour le roumain [Tufiş, D. & Barbu, A. M., (1997)].

Nous donnons un exemple d'annotation fournie par *TTL* pour le français dans la Figure 2 ci-dessous. L'attribut *lemma* garde les informations sur les lemmes (avec une probabilité en cas d'ambiguïté), l'attribut *ana* stocke les informations morphosyntaxiques, et l'attribut *chunk* marque l'annotation en groupes nominaux ou prépositionnels :

```
<seg lang="fr"><s id="ttlfr.3">
<w lemma="voir" ana="Vmps-s">vu</w>
<w lemma="le" ana="Da-fs" chunk="Np#1">la</w>
<w lemma="proposition" ana="Ncfs" chunk="Np#1">proposition</w>
<w lemma="de" ana="Spd" chunk="Pp#1">de</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>
<c>;</c>
</s></seg>
```

Figure 2 : Résultats de sortie de *TTL* pour le français

Ces informations sont utilisées par la méthode d'alignement lexical des corpus parallèles exploités, que nous présentons dans la section suivante.

3. L'alignement lexical

La qualité des corpus alignés est essentielle pour construire des modèles de traduction performants. Ainsi, nous nous sommes concentrés sur l'alignement lexical français - roumain en

utilisant les corpus parallèles décrits dans la section précédente, qui sont annotés et alignés au niveau propositionnel.

Notre système d'alignement lexical combine des méthodes statistiques et des règles heuristiques ayant un fondement linguistique. Nous exploitons les informations linguistiques associées aux unités lexicales, afin d'adapter l'algorithme d'alignement lexical à la paire de langues étudiées. Ainsi, nous évaluons l'impact des informations linguistiques sur la qualité de l'alignement lexical français - roumain.

Premièrement, nous utilisons l'aligneur statistique *GIZA++* [Och, F. J. & Ney, H. (2000), (2003)] qui implémente les modèles génératifs IBM [Brown, P, F. *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, un seul ou plusieurs équivalents de traduction dans la langue cible. Comme ces modèles ne réalisent pas d'alignements multiples, nous utilisons aussi des heuristiques [Koehn, P. *et al*. (2003)] [Tufis, D. *et al.* (2005)] pour repérer également des alignements au niveau de segments de traduction, tels que les chunks : syntagmes nominaux, syntagmes prépositionnels, syntagmes verbaux.

Ainsi, nous préparons le corpus parallèle dans le format d'entrée requis par *GIZA++*, en fournissant aussi le lemme suivi des deux premiers caractères de l'étiquette morphosyntaxique, afin de désambigüiser morphologiquement le lemme [Tufis, D. *et al*. (2005)]. Nous donnons un exemple à titre d'illustration : le même lemme peut être un nom commun ou un adjectif participial (qualificatif) *traité_Nc* vs. *traité_Af.*

Afin d'obtenir une précision élevée de la méthode d'alignement, nous réalisons des alignements bidirectionnels (FR-RO et RO-FR) avec *GIZA++* et ensuite, nous obtenons leur intersection [Koehn, P. *et al*. (2003)]. Cette heuristique permet de garder seulement les alignements considérés sûrs, vu qu'ils sont repérés dans les deux sens du processus d'alignement.

Pour obtenir des alignements sûrs, nous utilisons également un ensemble de cognats extraits automatiquement à partir du corpus de travail. En effet, nous filtrons la liste d'équivalents de traduction obtenue à l'issue de l'intersection des alignements bidirectionnels, avec une liste de cognats.

Nous considérons comme cognats les mots remplissant les conditions suivantes :

- ils sont des équivalents de traduction dans deux phrases parallèles ;
- ils ont des lemmes identiques ou présentent des similarités orthographiques au niveau du lemme et possèdent un sens commun.

Pour repérer automatiquement les cognats, nous utilisons un algorithme qui calcule la sous-chaîne maximale commune de caractères entre les mots d'une paire bilingue donnée. Les caractères de la sous-chaîne sont ordonnés, mais pas nécessairement contigus. La taille de la sous-chaîne est rapportée à la longueur du mot le plus long [Melamed, I. D. (1999)], [Kraif, O. (1999)]. Ce score est établi empiriquement à 0,60. Finalement, nous appliquons une série d'heuristiques linguistiques [Tufis, D. *et al.* (2005)], afin d'augmenter le rappel de la méthode d'alignement lexical :

i. la définition des classes d'équivalence de catégorie lexicale (un nom peut être traduit par un nom, un verbe ou un adjectif) ;
ii. l'alignement de noms, d'adjectifs, de verbes et d'adverbes ;
iii. l'alignement de chunks contenant des équivalents de traduction déjà alignés ;
iv. l'alignement des éléments appartenant aux chunks par des heuristiques.

Afin d'améliorer les résultats de l'alignement lexical, nous avons procédé à l'analyse linguistique des erreurs d'alignement. Nous avons repéré plusieurs classes d'erreurs systématiques d'alignement. Pour éviter ces erreurs, nous avons implémenté un module supplémentaire utilisant une base de règles heuristiques morphosyntaxiques contextuelles spécifiques à la paire de langues étudiée.

Nous présentons le corpus aligné, ainsi que la distribution par classes d'erreurs d'alignement lexical identifiées et les solutions proposées, dans la section suivante.

4. La distribution par classes d'erreurs d'alignement lexical

Nous avons aligné lexicalement un corpus parallèle français - roumain du domaine juridique et administratif, extrait de l'*Acquis Communautaire* et comprenant 1000 phrases alignées. Le corpus français contient 33 036 tokens et le corpus roumain 28 645 tokens. Nous avons sélectionné automatiquement des phrases bien formées (commençant par une majuscule et finissant par un signe de ponctuation). Nous avons limité aussi le nombre de mots maximum par phrase à 80.

Les erreurs d'alignement lexical repérées sont dues principalement aux différences morphosyntaxiques entre le français et le roumain. Malgré le fait que les deux langues étudiées sont des langues latines, chacune possède ses caractéristiques morphosyntaxiques spécifiques. Les structures syntaxiques sont similaires, mais les traits morphosyntaxiques présentent des différences importantes. Ainsi, les noms et les nominaux du roumain possèdent le cas (nominatif, accusatif, génitif, datif, vocatif) marqué par des désinences spécifiques (dans le paradigme flexionnaire) et par des morphèmes ou affixes proclitiques présents au niveau syntaxique [*GALR*[12] (2005)]. De plus, ils peuvent présenter le genre neutre. Le déterminant défini du roumain est toujours enclitique et fusionne avec le nom ou l'adjectif antéposé, tandis qu'en français il est toujours proclitique et forme un mot séparé. En roumain, les clitiques en accusatif et en datif s'expriment simultanément avec le complément direct ou indirect. En français, l'utilisation du pronom clitique exclue la réalisation de surface du complément direct ou indirect comme groupe nominal. D'autres différences concernent les structures syntaxiques propres à chaque langue (les subordonnées relatives) et les morphèmes supplémentaires ou différents d'une langue à l'autre (la relation de possession, le destinataire, l'infinitif, la négation, les numéraux ordinaux).

Les erreurs repérées à l'issue de l'analyse linguistique des résultats du module d'alignement lexical sont spécifiques au domaine juridique et administratif du corpus étudié, qui est caractérisé par l'utilisation des constructions impersonnelles, de la troisième personne du singulier ou du pluriel, des termes spécifiques du domaine, etc.

Nous illustrons plus loin le type d'analyse linguistique effectuée, en donnant quelques exemples d'erreurs systématiques d'alignement lexical, repérées dans le corpus étudié [Navlea, M. & Todiraşcu, A. (2010)].

L'une des erreurs fréquentes concerne l'expression de la relation de possession (Figure 3). En français, cette relation est exprimée par le biais de la préposition *de*, tandis qu'en roumain, une des possibilités est l'utilisation d'un morphème supplémentaire de génitif *al, a, ai, ale* suivi d'un nom en génitif. À cause de cette différence, la préposition *de* et le morphème *a* ne sont pas alignés.

(12) Gramatica academică a limbii române (Grammaire académique de la langue roumaine)

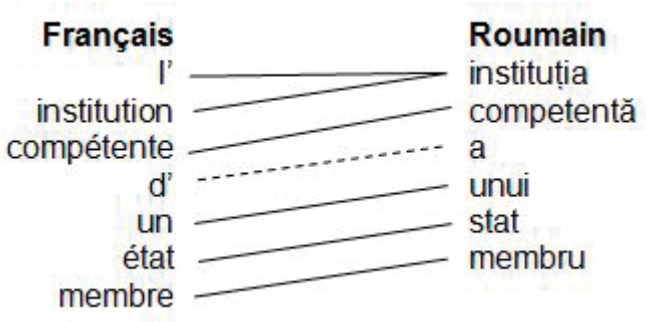

Figure 3 : Relation de possession

Dans le Tableau 2 ci-dessous figure la règle morphosyntaxique contextuelle proposée pour éviter ce type d'erreur :

Tableau 2 : Règle heuristique morphosyntaxique contextuelle pour la possession

Français	déterminant défini + N + ADJ + *de* + déterminant indéfini + N
Roumain	N (déterminant défini) + ADJ + *al* | *a* | *ai* | *ale* + déterminant indéfini (forme de génitif) + N génitif

Une autre classe d'erreurs fréquentes concerne les subordonnées relatives (Figure 4). Dans les subordonnées relatives du roumain, le complément direct est doublement exprimé par le pronom relatif *care* (accusatif), toujours précédé par la préposition *pe*, et le pronom personnel *îl, -l, o, îi, -i, le*, tandis qu'en français, il est exprimé par la pronom relatif *que.* À cause de cette différence, dans l'exemple de la Figure 4, le pronom *que* du français n'est pas aligné avec la préposition *pe* et le pronom personnel *le* du roumain.

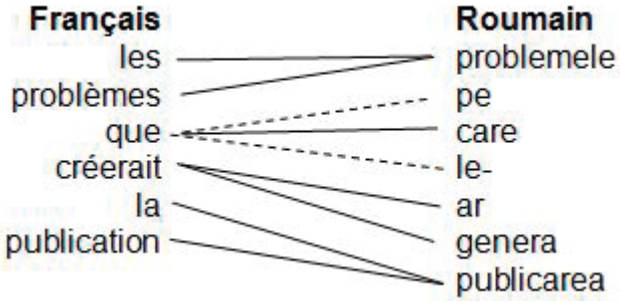

Figure 4 : Subordonnées relatives

La règle morphosyntaxique contextuelle mettant en correspondance le pronom relatif *que* avec *pe* et *le*, figure dans le Tableau 3 :

Tableau 3 : Règle heuristique morphosyntaxique contextuelle pour les subordonnées relatives

Français	N + *que* + V
Roumain	N + *pe* + *care* (accusative) + *îl* | *-l* | *o* | *îi* | *-i* | *le* + V

Nous avons repéré aussi des erreurs fréquentes d'alignement lexical au niveau des collocations qui sont des expressions polylexicales dont les mots entretiennent une relation lexico-syntaxique [Todiraşcu, A*. et al*. (2008)].

Les collocations peuvent avoir comme équivalents de traduction aussi bien des collocations, qu'une seule unité lexicale. Par exemple, une collocation verbo-nominale 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).

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 5, le déterminant indéfini *de* appartenant à la collocation verbo-nominale *ne pas prendre de mesures* (forme négative) n'est pas aligné.

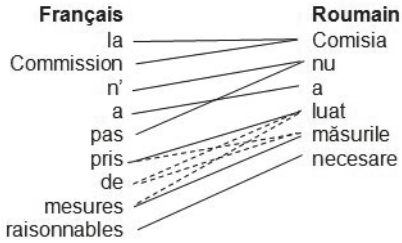

Figure 5 : Collocations

Pour diminuer les erreurs au niveau des collocations verbo-nominales, nous utilisons un dictionnaire de collocations verbo-nominales, munies de leurs propriétés morphosyntaxiques contextuelles [Todiraşcu, A. *et al.* (2008)].

Nous présentons dans le Tableau 4 ci-dessous les classes d'erreurs d'alignement lexical repérées dans le corpus aligné étudié, ainsi que la distribution par classes de ces erreurs. Ces résultats permettent de connaitre les classes d'erreurs prédominantes auxquelles il faut accorder la priorité, afin de diminuer efficacement le taux d'erreurs d'alignement. Toutefois, ces résultats dépendent du domaine et du volume du corpus de travail et peuvent varier en fonction de ces paramètres. De ce fait, nous ne pouvons pas les généraliser. Pour des corpus ayant des domaines différents, des études similaires restent nécessaires afin d'identifier les classes d'erreurs spécifiques. Cependant, ces résultats permettent d'améliorer l'alignement lexical des corpus parallèles juridiques et administratifs français - roumain.

Tableau 4 : Distribution par classes des erreurs d'alignement lexical français - roumain

	Classes d'erreurs d'alignement lexical français – roumain	Distribution par classes des erreurs (%)
1	Déterminants définis	**50,24 %**
2	Termes, collocations	**18,57 %**
3	Possession	**13,50 %**
4	Relatives	**8,51 %**
5	Infinitif	**4,96 %**
6	Négation	**1,86 %**
7	Destinataire	**1,38 %**
8	Numéraux ordinaux	**0,40 %**
9	Autres	**0,58 %**

Nous constatons que les erreurs les plus fréquentes appartiennent à la classe des déterminants définis et représentent approximativement la moitié du nombre total d'erreurs dans le corpus étudié. De ce fait, nous nous concentrons en premier sur la résolution de cette classe d'erreurs, en appliquant les règles heuristiques morphosyntaxiques correspondantes.

Une autre classe d'erreurs significatives est représentée par les termes et les collocations spécifiques au domaine du corpus. Ces structures posent problèmes pour l'alignement lexical, en manque de ressources terminologiques externes pour le français et le roumain. Une solution serait l'exploitation des cognats pour l'extraction de terminologie bilingue à partir du corpus de travail, vu que le roumain a repris la terminologie juridique du français par des calques et des emprunts.

Pour réduire le nombre d'erreurs au niveau des collocations, nous utilisons un dictionnaire multilingue de collocations verbo-nominales [Todiraşcu, A. *et al.* (2008)]. Ce dictionnaire est disponible pour le français, le roumain et l'allemand. Ainsi, nous utilisons les collocations comme indices de base dans le processus d'alignement lexical.

Comme le dictionnaire a été complété avec des données extraites du corpus *JRC-Acquis* et comprend les collocations verbo-nominales les plus fréquentes dans le corpus, cette ressource est efficace pour l'alignement des collocations verbo-nominales, mais elle ne résout pas le problème d'autres classes de collocations (nominales, adverbo-adjectivales, etc).

Des erreurs d'alignement lexical fréquentes apparaissent également au niveau de l'expression de la relation de possession et de subordonnées relatives pour lesquelles des heuristiques morphosyntaxiques contextuelles ont été définies. D'autres classes d'erreurs comme l'infinitif, la négation, le destinataire et les numéraux ordinaux sont moins fréquentes dans le corpus étudié, mais leur nombre peut augmenter significativement si le volume du corpus de travail devient important, vu qu'elles représentent des erreurs systématiques. De ce fait, pour ces classes d'erreurs nous avons défini aussi des heuristiques morphosyntaxiques.

Nous avons inclut dans la catégorie *Autres* des erreurs ne pouvant pas être résolues en utilisant des heuristiques morphosyntaxiques, comme les erreurs produites par les pronoms anaphoriques *en* et *y* du français, n'ayant pas d'équivalent de traduction en roumain, et par les paraphrases. Le taux de ces erreurs est très faible dans le corpus étudié, mais il peut aussi croitre en fonction du domaine et du volume des corpus de travail, vu que les anaphores et les paraphrases sont des procédés très usités en traduction humaine. 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.

L'analyse linguistique détaillée des erreurs de l'alignement lexical français - roumain a permis la définition d'une base de 27 règles heuristiques morphosyntaxiques contextuelles, automatisables dans un module supplémentaire d'alignement dépendant de la paire de langues étudiées. Cette ressource est utilisée pour effectuer l'alignement lexical à l'intérieur des chunks.

Conclusions et perspectives

Nous présentons ici un projet en cours ayant comme objectif le développement des ressources linguistiques pour un système de traduction automatique statistique factorisée, pour deux langues complexes morphologiquement, le français et le roumain. Nous nous sommes concentrés sur l'alignement lexical des corpus parallèles utilisés par un tel système. Nous avons analysé les erreurs du système d'alignement développé et nous avons défini une base de règles heuristiques morphosyntaxiques contextuelles pour les éviter. Les erreurs et leur distribution par classes dépendent du domaine du corpus étudié et aussi du volume des données et ne peuvent pas être

généralisées, mais restent une indication importante pour effectuer des alignements des corpus parallèles juridiques et administratifs français - roumain.

Dans le futur, nous allons construire des modèles de traduction factorisés français - roumain et nous allons évaluer plusieurs combinaisons de facteurs linguistiques, afin d'identifier les paramètres optimaux pour la paire de langues étudiée. Nous allons comparer les résultats ainsi obtenus avec les résultats des systèmes de traduction statistiques pures, afin d'évaluer l'impact des paramètres linguistiques utilisés sur la qualité des traductions français - roumain.

Bibliographie

Avramidis, Eleftherios, Koehn, Philipp (2008), « Enriching Morphologically Poor Languages for Statistical Machine Translation », In *Proceedings of ACL-08: HLT*, Columbus, June 2008, pp. 763-770.

Brown, Peter F., Della Pietra, Vincent J., Della Pietra, Stephen A., Mercer, Robert L. (1993), « The mathematics of statistical machine translation: Parameter estimation », *Computational Linguistics*, 19(2), pp. 263–312.

Ceauşu, Alexandru, Ştefănescu, Dan, Tufiş, Dan (2006), « Acquis Communautaire Sentence Alignment using Support Vector Machines », in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, May 2006, pp. 2134-2137.

Ceauşu, Alexandru (2009), « Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă », Thèse de doctorat, Académie Roumaine, Bucarest, avril 2009, 123 p.

Gavrilă, Monica (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 September 2009, Borovets, Bulgaria, pp. 14-18.

Grass, Thierry (2009), « A quoi sert encore la traduction automatique ? », in *Les Cahiers du GEPE, Outils de traduction - outils du traducteur ?*, n° 3, Strasbourg, 14 p.

Guţu Romalo, Valeria (coord.) (2005), « Gramatica limbii române, Cuvântul », vol. I, Ed. de l'Académie Roumaine, Bucarest, 712 p.

Ide, Nancy, Véronis, Jean (1994), « Multext (multilingual tools and corpora) », in *Proceedings of the 15th CoLing*, Kyoto, August 5-9, pp. 90-96.

Ion, Radu (2007), « Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română », Thèse de doctorat, Académie Roumaine, Bucarest, mai 2007, 148 p.

Irimia, Elena (2008), « Experimente de Traducere Automată Bazată pe Exemple », in *Actes de l'Atelier Resurse Lingvistice Româneşti şi Instrumente pentru Prelucrarea Limbii Române*, Iaşi, 19-21 novembre 2008, pp. 131-140.

Koehn, Philipp, Och, Franz Josef, Marcu, Daniel (2003), « Statistical Phrase-Based Translation », in *Proceedings of HLT-NAACL 2003*, Edmonton, May-June 2003, pp. 48-54.

Koehn, Philipp, Hoang, Hieu (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)*, Prague, June 2007, pp. 868-876.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra, Herbst, Evan (2007), « Moses: Open source toolkit for statistical machine translation », in *Proceedings of the ACL 2007 Demo and Poster Sessions*, Czech Republic, Prague, June 2007, pp. 177-180.

Kraif, Olivier (1999), « Identification des cognats et alignement bi-textuel : une étude empirique », *Actes de la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles, TALN 99*, Cargèse, 12-17 juillet 1999, pp. 205-214

Kraif, Olivier (2001), « Constitution et exploitation de bi-textes pour l'Aide à la traduction », *Thèse de doctorat*, sous la dir. de Henri Zinglé, Université de Nice Sophia Antipolis.

Marcu, Daniel, Munteanu, Dragoş Ştefan (2005), « Statistical Machine Translation: An English-Romanian Experiment », in *EUROLAN 2005*.

Martin, Joel, Mihalcea, Rada, Pedersen, Ted (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*. June, 2005, Ann Arbor, Michigan, Association for Computational Linguistics, 65-74.

Melamed, I. Dan (1999), « Bitext Maps and Alignment via Pattern Recognition », Computational Linguistics, vol. 25, n° 1, pp. 107-130.

Navlea, Mirabela, Todiraşcu, Amalia (2010), « Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems », in *Proceedings of the 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)*, Malta, Valletta, May 2010, pp. 41-48.

Och, Franz Josef, Ney, Hermann (2000), « Improved Statistical Alignment Models », in *Proceedings of the 38th Conference of ACL*, Hong Kong, pp. 440-447.

Och, Franz Josef, Ney, Hermann (2003), « A Systematic Comparison of Various Statistical Alignment Models ». in *Computational Linguistics*, vol. 29, n° 1, March 2003, pp. 19-51.

Steinberger, Ralph, Pouliquen, Bruno, Widiger, Anna, Ignat, Camelia, Erjavec, Tomaž, Tufiş, Dan, Varga, Dániel (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)*, Genoa, May 2006, pp. 2142-2147.

Todiraşcu, Amalia, Heid, Ulrich, Ştefănescu, Dan, Tufiş Dan, Gledhill, Christopher, Weller Marion, Rousselot, François (2008), « Vers un dictionnaire de collocations multilingue », in *Cahiers de Linguistique*, vol. 33, n° 1, Louvain, août 2008, p. 161-186.

Trandabăț, Diana, Iftene, Adrian, Pistol, Ionuț, Forăscu, Corina, Cristea, Dan (2006), « Resurse româneşti în cadrul proiectului LT4eL », in Corina Forăscu, Dan Tufiş, Dan Cristea (eds.) : *Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Editura Universității «Al.I. Cuza» Iaşi, România, ISBN 978-973-703-208-9.

Tufiş, Dan, Barbu, Ana Maria (1997), « A Reversible and Reusable Morpho-Lexical Description of Romanian », in Dan Tufiş and Poul Andersen (eds.), *Recent Advances in Romanian Language Technology*, pp. 83-93, Editura Academiei Române, Bucureşti, 1997. ISBN 973-27-0626-0.

Tufiş, Dan, Ion, Radu, Ceauşu, Alexandru, Ştefănescu, Dan (2005), « Combined Aligners », in *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine*

Translation and Beyond, pp. 107-110, Ann Arbor, USA, Association for Computational Linguistics. ISBN 978-973-703-208-9.

Tufiş, Dan, Irimia, Elena (2006), « RoCo_News - A Hand Validated Journalistic Corpus of Romanian », in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 869-872, Genoa, Italy, May 2006. ELRA - European Language Ressources Association.

Tufiş, Dan, Koeva, Svetla (2007), « Ontology-supported Text Classification based on Cross-lingual Word Sense Disambiguation », in Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, (eds.), *Applications of Fuzzy Sets Theory. 7th International Workshop on Fuzzy Logic and Applications (WILF 2007)*, volume 4578 of *Lecture Notes in Artificial Intelligence*, pp. 447-455, Springer-Verlag, September 2007. ISBN 978-3-540-73399-7.

Tufiş, Dan, Koeva, Svetla, Erjavec, Tomaž, Gavrilidou, Maria, Krstev, Cvetana (2008a), « Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages », in Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.), in *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, Dubrovnik, Croatia, September 25-28, pp. 145-152.

Tufiş, Dan, Ion, Radu, Ceauşu, Alexandru, Ştefănescu, Dan (2008b) « RACAI's Linguistic Web Services », in *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May 2008. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.

Tufiş, Dan, Ceauşu, Alexandru (2008c), « DIAC+: A Professional Diacritics Recovering System », in *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May, 2008, ELRA, ISBN 2-9517408-4-0.

Vertan, Cristina, Gavrilă, Monica (2010), « Multilingual applications for rich morphology language pairs, a case study on German Romanian », in Dan Tufiş and Corina Forăscu (eds.): *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Romanian Academy Publishing House, Bucharest, pp. 448-460, ISBN 978-973-27-1972-5.