Traduction Automatique à Base de Règles de l'Anglais vers l'Arabe

Sadik BESSOU
Université de Sétif -1-

Résumé :

Dans ce travail, on aborde plusieurs points importants concernant l'analyse morphosyntaxique de la langue arabe appliquée à l'informatique documentaire et particulièrement à la traduction automatique. Tout d'abord, on dresse un aperçu sur les approches de la traduction automatique et particulièrement les approches indirectes de transfert. Puis nous présentons nos contributions pour la résolution des problèmes morphosyntaxiques dans la traduction automatique. Dans une première contribution, nous avons développé un analyseur morphologique pour la langue arabe, et on l'a exploité dans le module de transfert pour la génération des mots dans la langue cible. Dans une seconde contribution, nous avons proposé une liste de règles de transfert morphosyntaxique de l'anglais vers l'arabe, pour une traduction en trois phases : analyse, transfert, génération.

Introduction

La traduction automatique s'est développée autour de trois approches essentielles: les approches symboliques, les approches à base de règles, et les approches statistiques.

Dans cet article nous nous intéressons par les approches linguistiques à base de règles. Ces dernières formalisent des règles de réécritures et proposent une série de transformations des arbres morphologiques, syntaxiques et sémantiques entre langue source et langue cible.

1. Les approches de traduction automatique

Cette section résume plusieurs approches qui ont jalonné la recherche sur la traduction automatique, de 1950 à nos jours [1].

Les premiers programmes d'ordinateurs relatifs à la traduction étaient destinés à servir d'aide à la traduction. Quelques règles et surtout un dictionnaire bilingue composaient le cœur du système. Les années suivantes voient les dictionnaires grandir ; ce qui engendre une augmentation du nombre de règles régissant le ré-ordonnancement des mots. La nécessité d'automatiser l'acquisition des règles et de progresser leur généricité participe au développement de la linguistique informatique.

L'approche à base de règles

Le triangle présenté à la figure 1 est attribué à Vauquois. Il présente de manière synthétique une analyse du processus de traduction encore pleinement pertinente et employée de nos jours.

La traduction peut s'opérer à plusieurs niveaux :

Au niveau le plus bas, on retrouve la traduction directe, qui passe directement des mots de la langue source aux mots de la langue cible.

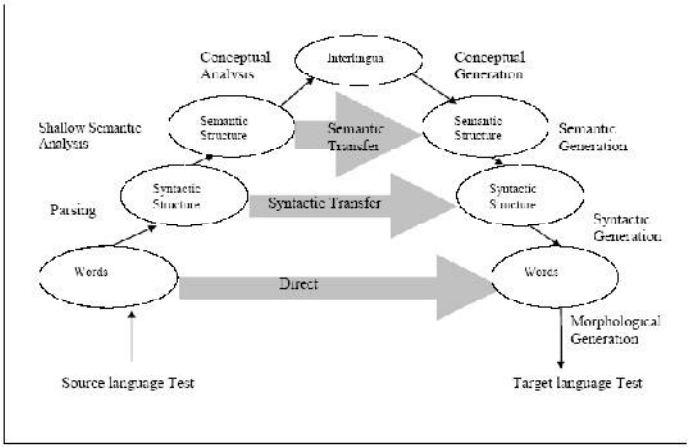


Figure 1 Triangle de Vauquois

Les systèmes semi-directs ont une phase de segmentation ou d'analyse morphologique, voire morphosyntaxique, et une phase de génération morphologique. Si l'on effectue une analyse syntaxique de la phrase source, le transfert à la langue cible devrait être simplifié. À ce niveau, les détails spécifiques à la constitution des groupes nominaux, par exemple, n'ont pas besoin d'être connus des règles régissant le transfert.

Avec une analyse plus approfondie de la phrase source, au niveau sémantique, le transfert devient uniquement sémantique. En revanche, la génération des mots après le transfert est plus complexe qu'au niveau inférieur.

Enfin, une analyse totale de la phrase source peut aboutir à une représentation de son sens dans une « inter-langue » artificielle, de laquelle on dérive ensuite les mots cibles.

Le pivot est un type de représentation utilisant des attributs et relations interlingues, et des unités lexicales de chacune des

langues. Ces systèmes sont à transfert simple, alors qu'on a un double transfert en « pivot ».

L'approche reposant sur une inter-langue est attractive car elle remplace le problème de la traduction par deux problèmes unilingues, d'analyse et de synthèse. L'avantage est que les modules unilingues sont à priori réutilisables. Pour couvrir tous les sens de traduction entre n langues, il suffit de n modules d'analyse et de n modules de synthèse, contre n*(n−1) systèmes de transfert [2].

Des systèmes à véritable langue-pivot, on peut citer : ATLAS-II de Fujitsu ou IVOT/Crossroads de NEC, ou KANT / CATALYST de CMU/Caterpillar, ou UNL, ou MASTOR-1 d'IBM.

Le début des années 1990 voit le développement d'autres types d'approches. Les ordinateurs se répandent et gagnent en puissance, ce qui permet l'émergence de stratégies qui se fondent sur de grandes quantités de données « corpus-basedapproaches ». On distingue en particulier deux grands types d'approches: la traduction automatique à base d'exemples et la traduction automatique par méthodes statistiques.

1.1. L'approche statistique

Elle doit ses origines aux travaux de Brown et al. [3] en 1990 et en particulier au prototype Candide de Berger et al. [4]en 1994, un système de traduction construit à partir de discours disponibles en français et en anglais de parlementaires canadiens. En effet, comme la traduction à base d'exemples, la traduction par méthodes statistiques repose sur un corpus parallèle.

Un modèle statistique de traduction est défini, comprenant une ou plusieurs lois de probabilités. Le corpus est traité afin d'estimer ces lois qui sont souvent constituées de plusieurs milliers, voire millions de paramètres.

1.2. L'approche à base d'exemples

L'approche à base d'exemples (« Example-based machine translation », ou EBMT) repose sur un ensemble « d'exemples » préalablement traduits : un corpus parallèle de phrases traduites l'une de l'autre. Lorsqu'on lui présente une phrase à traduire, le système parcourt sa base d'exemples et produit trivialement une traduction si la phrase s'y trouve. Dans le cas général, la phrase n'apparaît pas dans la base et le système s'emploie alors à rassembler des exemples qui contiennent des fragments communs (des groupes de mots) avec la phrase à traduire. Pour chaque fragment d'exemple dans la langue source, il s'agit ensuite de retrouver sa traduction dans la langue cible : c'est la phase d'alignement. Enfin, la phase de génération assemble les fragments dans la langue cible et produit la traduction. À chacune des trois étapes, il est possible d'utiliser des sources externes de connaissances, telles que des lexiques bilingues, des listes de synonymes, des étiquettes ou des arbres syntaxiques, etc [5].

1.3. Les approches hybrides

Leur idée directrice est qu'une approche unique du problème de la traduction, aussi perfectionnée soit-elle, ne parviendra pas à produire une traduction satisfaisante dans tous les cas. Au contraire, une approche par règles peut s'avérer particulièrement adaptée à certaines phrases, tandis que d'autres phénomènes linguistiques sont correctement traités par une approche reposant sur des corpus.

Les systèmes hybrides sont actuellement envisagés comme des systèmes combinant les méthodes statistiques ou méthodes à base d'exemples avec des méthodes linguistiques (à base de règles), en particulier pour l'analyse morphologique et syntaxique.

Un système hybride pourrait parvenir à tirer profit des forces de chaque approche. Une première stratégie pour mettre en œuvre un système hybride est d'utiliser les différentes approches en parallèle. Enfin, dans un système statistique, il est courant de traiter par un système de règles spécialisées certains fragments de phrases, typiquement les nombres, les dates, etc. Les morceaux de phrases ainsi identifiés et traduits en isolation par le système à base de règles peuvent être transmis au système statistique [5].

2. Règles de transfert morphologique

L'étiqueteur des parties de discours utilisé dans ce travail est le *Stanford Log-linear Part-Of-Speech Tagger* (*POS Tager*) [6] crée par (*The Stanford Natural LanguageProcessing Group*) [7].

Exemple de règles de transfert morphologique

On entend par accord sujet-verbe/ verbe-objet, l'accord intrinsèque dans le mot puisque on est au niveau morphologique.

Si l'ordre des mots dans la phrase est SVO, le verbe s'accorde avec le sujet en genre et en nombre. Si l'ordre est VSO, le verbe s'accorde uniquement en genre.

Exemple

(The boy writes the lesson) → (يكتب الولد الدرس) or (الولد يكتب الدرس)

(The boys write the lesson) → (يكتب الأولاد الدرس) or (الأولاد يكتبون الدرس)

La forme la plus courante en arabe est la forme VSO. L'accord du verbe avec son sujet est expliqué ci-dessous

Le présent

Si NN_c(M): [$DT_sNN_sVBZ_sDT_sNN_s$] → ["يـ"+VB_c …]

Si NN_c(F):[$DT_sNN_sVBZ_sDT_sNN_s$] → ["تـ"+VB_c …]

Exemple

(The boy writes the lesson) → (يكتب الولد الدرس)

(The girl writes the lesson) → (تكتب البنت الدرس)

Le passé

Si NN_c(M): [$DT_sNN_sVBD_sDT_sNN_s$] → [VB_c……..]

Si NN_c(F):[$DT_sNN_sVBD_sDT_sNN_s$]→ [VB_c +"ت".....]

Exemple

(The boy wrote the lesson) → (كتب الولد الدرس)

(The girl wrote the lesson) → (كتبت البنت الدرس)

Le futur

Si NN_c(M): [$DT_sNN_sMD_s$ $VB_sDT_sNN_s$]→ ["سـ"+ "يـ"+VB_c…..]

Si NN_c(F): [$DT_sNN_sMD_s$ $VB_sDT_sNN_s$] → ["سـ"+ "تـ"+VB_c…..]

Exemple

(The boy willwrite the lesson) → (سيكتب الولد الدرس)

(The girl willwrite the lesson) → (ستكتب البنت الدرس)

L'accord sujet-verbe et objet-verbe quand la phrase est réduite en un seul mot est effectué par l'utilisation des pronoms pour remplacer le sujet et l'objet par le biais des préfixes, suffixes et proclitiques, enclitiques.

Exemple

(He writes the lesson)→ (هو يكتب الدرس) ou (يكتب الدرس)

[$PRP_s VBZ_s DT_s NN_s$] → ["pré"+VB_c+"suf" DT_c+ NN_c]

(He writesit)→ (يكتبه)

[$PRP_s VBZ_s PRP_s$]→ ["pré"+VB_c+"suf"+"Enc"]

On constate dans le deuxième exemple que la traduction des trois entités est réduite à un seul mot.

Alors que Systran nous donne une traduction mot à mot (هو يكتب هو) en trois entités, et Google donne (انه يكتب عليه) en trois entités aussi. Donc la traduction mot à mot est valable pour les entités nom (NN) et verbe (VB) et non pas pour les pronoms (PRP) qui plutôt s'agglutinent aux noms et aux verbes.

Le pronom sujet lié à un verbe conjugué au présent
Le pronom (sujet) est collé au verbe conjugué au présent implicitement par un préfixe (أ- نـ - يـ - تـ).

Notons qu'en arabe, le sujet est inclus dans le verbe conjugué (comme trait). Il n'est donc pas nécessaire (comme c'est le cas en anglais) de précéder le verbe conjugué de son pronom.

[$PRP_s VBP_s$] → ["Pré" +VB_c] ou ["Pré" +VB_c +"Suf"]

Si PRP_s={I,YOU(S,M),WE}:[$PRP_s VBP_s$]→["Pré" +VB_c]

Si PRP_s={I}: Pré={أ}:[$PRP_s VBP_s$] → ["أ" +VB_c]

Si PRP_s={YOU(S,M)}: Pré={تـ}:[$PRP_s VBP_s$] → ["تـ" +VB_c]

Si PRP_s={WE}: Pré={ـنـ}:[PRP_sVBP_s] → ["ـنـ" +VB_c]

Si PRP_s = {HE, SHE}: [PRP_sVBZ_s] → ["Pré" +VB_c]

Si PRP_s={HE}: Pré={ـيـ}:[PRP_sVBZ_s] → ["ـيـ" +VB_c]

Si PRP_s={SHE}: Pré={ـتـ}:[PRP_sVBZ_s] → ["ـتـ" +VB_c]

Si PRP_s = {YOU(S,F), YOU(P, M), YOU(B), YOU(P, F), THEY(B,M), THEY(B,F), THEY(M), THEY(F)}: [PRP_sVBP_s] → ["Pré" +VB_c +"Suf"]

Si PRP_s={YOU(S,F)}: pré ={ـتـ} &suf ={ين}:[PRP_sVBP_s] → ["ـتـ" +VB_c +"ين"]

Si PRP_s={YOU(B)}: pré ={ـتـ} &suf ={ان}:[PRP_sVBP_s] → ["ـتـ" +VB_c +"ان"]

Si PRP_s={YOU(P,M)}: Pré={ـتـ} &suf ={ون}:[PRP_sVBP_s] → ["ـتـ" +VB_c +"ون"]

Si PRP_s={YOU(P,F)}: Pré={ـتـ} &suf ={ن}:[PRP_sVBP_s] → ["ـتـ" +VB_c +"ن"]

Si PRP_s={THEY(B,M)}: pré ={ـيـ} &suf ={ان}:[PRP_sVBP_s] → ["ـيـ" +VB_c +"ان"]

Si PRP_s={THEY(B,F)}: pré ={ـتـ} &suf ={ان}:[PRP_sVBP_s] → ["ـتـ" +VB_c +"ان"]

Si PRP_s={THEY(M)}: pré ={ـيـ} &suf ={ون}:[PRP_sVBP_s] → ["ـيـ" +VB_c +"ون"]

Si PRP_s={THEY(F)}: pré ={ـتـ} &suf ={ن}:[PRP_sVBP_s] → ["ـتـ" +VB_c +"ن"]

Transfert syntaxique

La langue arabe est caractérisée par un ordre libre de mots dans la phrase, il n'est pas étrange de trouver plusieurs ordres

possibles comme : VSO, SVO, VOS et parfois même OVS comme dans les phrases suivantes :

VSO → (كتب الولد الدرس)

SVO → (الولد كتب الدرس)

VOS → (كتب الدرس الولد)

OVS → (الدرس كتب الولد)

Toutes ces phrases sont des phrases grammaticalement correctes et ont le même sens.

L'ordre le plus fréquent en langue arabe est l'ordre VSO, les autres sont rarement utilisés, ils servent à des situations pour accentuer des constituants de la phrase[1].

3. Règles de transfert syntaxique

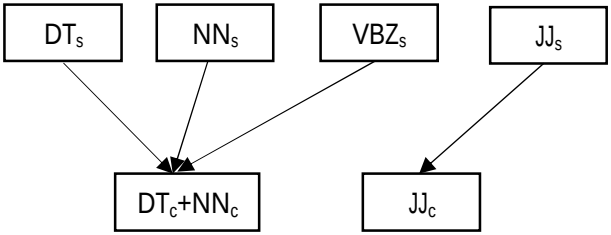
La langue arabe est plus parataxique que l'anglais, plus hypotaxique. Nida et Taber [8] affirment que la traduction devrait minimiser la parataxe ou la transformer en hypotaxe.

L'arabe aussi est fortement flexionnel, alors qu'on trouve des morphèmes avec des caractéristiques syntaxiques mais attachés à d'autres mots comme les conjonctions de coordination, l'article défini, quelques prépositions ainsi les pronoms, ce qui engendre une diminution de nombre de constituants syntaxiques dans la langue cible.

La figure 2 montre un exemple de transfert structurel entre l'anglais et l'arabe.

[1] Aspect emphatique typique des langues flexionnelles, présentant un intérêt rhétorique.

Figure 2 Diminution de nombre de constituants lors d'un transfert syntaxique



Quand on parle de phrase verbale ou nominale, on parle de la structure dans la langue cible, car parfois des phrases verbales en anglais deviennent des phrases nominales en arabe.

Exemple de règles de transfert syntaxique

Si le verbe (to be) est conjugué au futur (MD="will" & VB ="be"), la phrase sera précédée de "سيصبح" ou "ستصبح" selon le genre du sujet (figure 3) et l'attribut sera suffixé. Si MD_s ="will" & VB_s ="be": $[DT_sNN_sMD_s\ VB_sJJ_s]$ → ["سيصبح " $DT_c+NN_cJJ_c$]

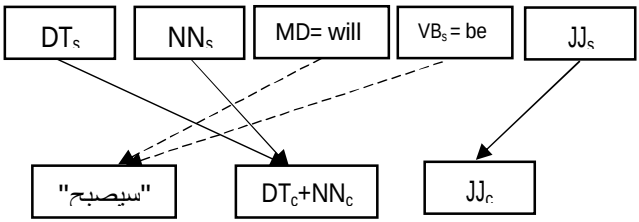


Figure 3 Transfert d'une phrase verbale au futur vers une phrase nominale

Exemple (The book willbeuseful) → (سيصبح الكتاب مفيدا)

4. Expérimentations et Résultats

Dans cette section en présente des échantillons de résultats comprenant les transferts morphologiques et syntaxiques. On montre des phrases en anglais et leurs traductions en arabe par notre système (*Tordjman*). On a choisi les exemples où il y a des problèmes de traduction dans les autres systèmes comme *Googletranslate* et *Systran*.

Interface de Tordjman

L'interface de *Tordjman* comporte un champ de saisie de la phrase source en anglais et un champ d'affichage de résultat de la phrase cible en arabe comme illustré dans la figure 4.

Figure 4 Exemple de Traduction avec Tordjman

4.1. Test de quelques structures

Les résultats sont présentés en mode ligne de commande. La figure 5 montre le résultat de la traduction de la structure (PRP VB DT NN) au présent (VBP ou VBZ), avec des noms masculins définis et indéfinis.

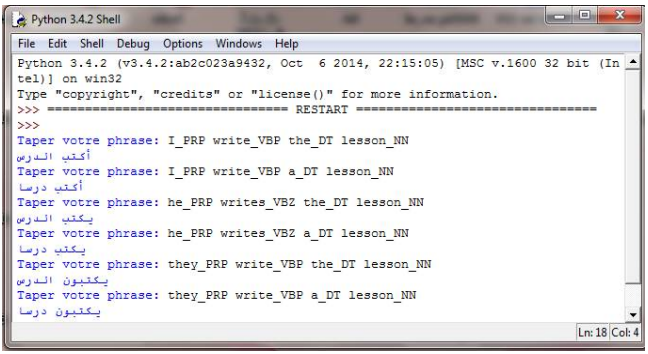


Figure 5 Résultat de la traduction de la structure PRP VB DT NN au présent

La figure 6 montre le résultat de la traduction de la structure (PRP VBD DT NN) Avec des noms masculins, définis et indéfinis.

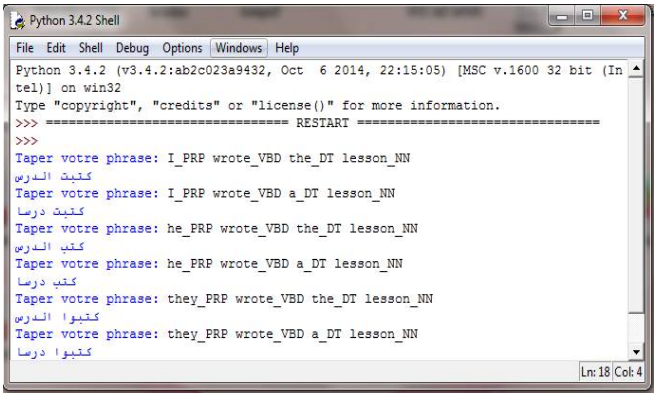
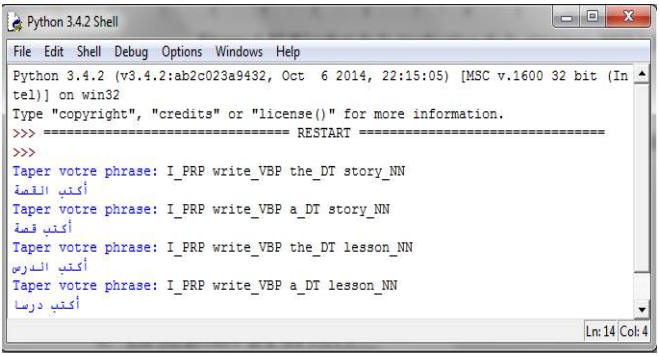


Figure 6 Résultat de la traduction de la structure PRP VB

D DT NN



La figure 7 montre le résultat de la traduction de la structure (PR

P VB DT NN) avec des noms masculins et féminins, définis et indéfinis.

Figure 7 Résultat 2 de la traduction de la structure PRP VB DT NN : NN(F), NN(M)

La figure 8 montre le résultat de la traduction de la structure (DT JJ NN) avec des noms masculins et féminins, définis et indéfinis.

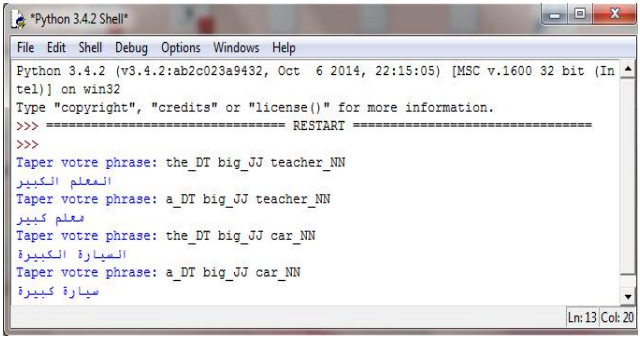


Figure 8 Résultat de la traduction de la structure DT JJ NN

5. **Expérimentations et évaluation**

Dans cette section, on présente les résultats obtenus lors de l'évaluation de l'approche sur des structures partielles de la langue. On doit mentionner ici que notre approche à base de règles est approuvable pour des domaines restreints avec des structures de phrases bien définis. Où elle peut donner de bons résultats par rapport à d'autres systèmes à base de méthodes statistiques. Or si on compare ces derniers comme système complet avec *Tordjman*, leurs résultats sont meilleurs.

Les expériences ont été réalisées on utilisant la métrique BLUE comme moyen d'évaluation automatique. On peut utiliser les codes fournis par Kenneth Heafield [9] ou NitinMadnani [10]. La métrique BLUE exige un fichier test et un fichier référence (traduction humaine), pour cela on a utilisé un corpus parallèle ; les phrases sources en anglais peuvent être utilisé dans les échantillons de tests, les phrases cibles correspondantes en arabe représentent les phrases références pour la comparaison.

Parmi les corpus qui peuvent être utilisé comme traduction référence le corpus de Tatoebatiré du projet OPUS (the open parallel corpus) de l'université d'Uppsala au suède [11]. OPUS propose une collection de corpus dédiée aux chercheurs et aux développeurs de traduction automatique. Pour la langue arabe, les corpus adéquats pour le test sont Tanzil et Tatoeba. Le premier corpus comprend des traductions de Coran, le deuxième comporte des phrases traduites entre différentes langues (plus de 132 langues). Le corpus le plus approprié à notre contribution est celui de Tatoeba. Ce corpus contient plus de 3469809 phrases, celui de la langue arabe plus de 15683 phrases.

Dans ces expériences, 76 phrases sont aléatoirement sélectionnées pour constituer le fichier test de *Tordjman*. Le même fichier est utilisé pour comparer les résultats avec *Google translation* et *Systran*. La procédure d'évaluation est faite phrase par phrase dans le fichier de test. On calcule les scores BLUE (1-gram, 2-grams, 3-grams) pour toutes les phrases résultats. Puis on calcule la moyenne de chaque score n-gram.

Pour toutes les phrases du fichier test notre système a eu un score de 1.00. Selon la table 1, on peut conclure que les résultats de notre approche sont meilleurs que les résultats de *Google* et de *Systran*.

Table 1 Les scores BLUE pour Tordjman, Google et Systran

Score *Tordjman*	Score *Google*	Score *Systran*
100.00	15.19	20.30

Alors notre système est adroit de générer des traductions de phrases simples avec moins d'erreurs. Et ce grâce à l'aspect morphologique dans la génération de mots. *Google* opte pour une approche statistique qui considère le nombre d'apparition d'un mot dans le corpus d'apprentissage pour l'avantager à un autre mot semblable mais avec d'autres flexions ce qui rend le sens incohérent dans la phrase.

La figure 9 présente le résultat détaillé de la métrique BLUE du segment 22, la phrase source (shewillwrite) a eu par *Tordjman* (ستكتب) et par *Google* (وقالت انها سوف أكتب).

La phrase Hypothesis (*Tordjman*) comporte 05 caractères en un seul mot avec un score de [1.00] alors que la phrase Hypothesis (*Google*) comporte 16 caractères en 4 mots avec un score de [0.16].

ID	Segment 22, Document "fakedoc" [Δ_{BLEU}=0.84]
Source	she will write
Reference (fakeref)	**ستكتب**
Hypothesis (*Tordjman*)	**ستكتب** [1.00]
Hypothesis (*Google*)	**وقالت انها سوف أكتب** [0.16]

Figure 9 Résultat BLUE du segment 22

La figure 10 présente le résultat du segment 45, la phrase source (wewrite a lesson) a eu par *Tordjman* (نكتب درسا) et par *Systran* (نحن نكتب درس).

La phrase Hypothesis (*Tordjman*) comporte 8 caractères en 2 mots avec un score de [1.00] alors que la phrase Hypothesis (*Systran*) comporte 10 caractères en 03 mots avec un score de [0.38].

ID	Segment 45, Document "fakedoc" [Δ_{BLEU}=0.62]
Source	we write a lesson
Reference (fakeref)	نكتب درسا
Hypothesis (*Tordjman*)	نكتب درسا [1.00]
Hypothesis (*Systran*)	نحن نكتب درس [0.38]

Figure 10 Résultat BLUE du segment 45

La figure 11 présente le résultat du segment 50, la phrase source (shewrote a lesson) a eu par *Tordjman* (كتبت درسا) et par *Systran* (هو كتب درس).

La phrase Hypothesis (*Tordjman*) comporte 8 caractères en 2 mots avec un score de [1.00] alors que la phrase Hypothesis (*Systran*) comporte 8 caractères en 03 mots avec un score de [0.23].

ID	Segment 50, Document "fakedoc" [Δ_{BLEU}=0.77]
Source	she wrote a lesson
Reference (fakeref)	كتبت درسا
Hypothesis (*Tordjman*)	كتبت درسا [1.00]
Hypothesis (*Systran*)	هو كتب درس[0.23]

Figure 11 Résultat BLUE du segment 50

On peut expliquer ces résultats par deux raisons :

-Notre approche adopte une génération morphologique ce qui diminue le nombre de mots erronés alors augmente les scores BLUE.

-Nous avons testé des phrases simples, ce qui privilège notre approche par rapport à *Google*.

Conclusion

A l'issue de cette expérience, on peut conclure que notre approche est convenable à des domaines restreints où les structures grammaticales sont limitées et invariables. Les résultats montrent une corrélation avec les jugements humains. L'incrémentation de nombre de règles pour plus de couverture de la langue augmente les possibilités d'interférence. A notre avis pour avoir un système robuste et général il faut utiliser ces règles de transfert comme étape de post traitement pour un système purement statistique ce qui élimine les déficiences des

modèles statistiques et généralise les domaines de la langue pour les modèles à base de règles.

Références

-JURAFSKY, D., MARTIN J. H. 2007. *Speech and Language Processing: An introduction to natural language processing, Computational linguistics, and speech recognition.* Martin.

-HIROSHI, U. MEIYING, Z. 1993. *Interlingua for multilingual machine translation.* In Proc. of MT Summit: International Cooperation for Global Communication, Kobe, Japan.

-BROWN, P. F. COCKE, J. Della PIETRA, S. Vincent Della PIETRA, J. JELINEK, F. LAFFERTY, J. D. MERCER, R. L. ROOSSIN, P.S. 1990. *A statistical approach to machine translation.* Computational Linguistics.

-BERGER, A. L. BROWN, P. F., Della PIETRA, S. A., Della PIETRA, V. J ., GILLETT, J. R., LAFFERTY, J. D., MERCER, R. L., PRINTZ, H.,UREŠ, L.. 1994. *The Candide system for machine translation*. In Proc. of the Workshop on Human Language Technology, Plainsboro, NJ.

-DECHELOTTE, D. 2007. *Traduction automatique de la parole par des méthodes statistiques*, thèse de doctorat en sciences, Université Paris-Sud 11 — Faculté des sciences d'Orsay.

-TOUTANOVA, K. KLEIN, D. MANNING, C. SINGER Y. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. In Proceedings of HLT-NAACL.
http://nlp.stanford.edu/index.shtml, consulté en Septembre 2013.

-EUGENE, N. TABER, C. R. 1982. *The Theory and Practice of Translation*. Leiden: E.J. Brill.

https://kheafield.com/code/consulté en Novembre 2014.

-MADNANI. N. 2011. iBLEU: *Interactively Debugging & Scoring Statistical Machine Translation Systems*, 5th IEEE International Conference on Semantic Computing.

-TIEDEMANN, J. 2012. *Parallel Data, Tools and Interfaces in OPUS*,8th International Conference on Language Resources and Evaluation (LREC), Turkey.