TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Évaluation de G-LexAr pour la traduction automatique statistique

Wigdan Mekki(1), Julien Gosme(1), Fathi Debili(2), Yves Lepage(3), Nadine Lucas(1)
(1) GREYC, UMR 6072, CNRS, Université de Caen Basse-Normandie, Caen, France
(2) LLACAN, UMR 8135, CNRS, Villejuif, France
(3) IPS, Université Waseda, Japon

Résumé. G-LexAr est un analyseur morphologique de l'arabe qui a récemment reçu des améliorations substantielles. Cet article propose une évaluation de cet analyseur en tant qu'outil de pré-traitement pour la traduction automatique statistique, ce dont il n'a encore jamais fait l'objet. Nous étudions l'impact des différentes formes proposées par son analyse (voyellation, lemmatisation et segmentation) sur un système de traduction arabe-anglais, ainsi que l'impact de la combinaison de ces formes. Nos expériences montrent que l'utilisation séparée de chacune de ces formes n'a que peu d'influence sur la qualité des traductions obtenues, tandis que leur combinaison y contribue de façon très bénéfique.

Abstract. G-LexAr is an Arabic morphological analyzer that has recently been improved for speed. This paper gives an assessment of this analyzer as a preprocessing tool for statistical machine translation. We study the impact of the use of its possible outputs (vocalized, lemmatized and segmented) through an Arabic-English machine translation system, as well as the impact of the combination of these outputs. Our experiments show that using these outputs separately does not influence much translation quality. However, their combination leads to major improvements.

Mots-clés : traduction automatique statistique, analyse morphologique, pré-traitement de l'arabe.

Keywords: statistical machine translation, morphological analysis, arabic preprocessing.

1 Introduction

L'arabe est une langue à morphologie riche dont la complexité présente des défis pour la traduction automatique (voir (Habash, 2007) pour une description des problèmes morphologiques relatifs à cette tâche).

Des expériences utilisant l'analyse morphologique pour améliorer la traduction automatique ont déjà été menées pour l'allemand (p. ex. Nießen & Ney, 2004) ou le turc (p. ex. Bisazza & Federico, 2009). Ces travaux utilisent diverses sortes de segmentation, lemmatisation et étiquetage grammatical. Dans le cas de l'arabe, les travaux de Lee (2004), utilisant une approche de segmentation en racines et affixes, puis de Habash & Sadat (2006), avec une approche de *tokenization* linguistiquement motivée, ont montré que le pré-traitement morphologique peut être utile à la traduction automatique statistique. D'un autre côté, Diab *et al.* (2007) ont montré que l'utilisation de la voyellation seule ne conduit à aucune amélioration (voyellation partielle), voire à de moins bons résultats (voyellation complète).

Dans cet article, nous évaluons l'analyseur morphologique de l'arabe G-LexAr (Debili *et al.*, 2002) sur des tâches de traduction automatique statistique arabe-anglais, en utilisant l'analyse morphologique comme étape de pré-

traitement. G-LexAr version 3 n'a encore jamais fait l'objet d'une telle évaluation, et a récemment été optimisé.

Cet article est organisé comme suit : la section 2 décrit l'analyseur morphologique G-LexAr et l'analyseur de référence que nous utilisons, BAMA ; la section 3 présente les détails de la conception d'un système de traduction automatique de référence ainsi que les résultats obtenus avec ces deux analyseurs ; la section 4 conclut ces travaux.

2 Analyse morphologique

Cette section présente une vue globale des deux analyseurs comparés : G-LexAr et BAMA. Pour une description détaillée de G-LexAr version 2, voir (Debili *et al.*, 2002) et pour BAMA utilisé comme référence voir (Buckwalter, 2002).

2.1 G-LexAr

G-LexAr est un programme d'analyse morpho-grammaticale de l'arabe (arabe classique et standard moderne), pouvant traiter des textes d'entrée voyellés ou non. Il produit en sortie des analyses où les mots peuvent être indépendamment segmentés, voyellés, lemmatisés ou étiquetés. Il est fondé sur la mise en œuvre d'un grand nombre de dictionnaires et règles qui privilégient la rapidité des traitements à l'espace mémoire.

Il opère en trois étapes. La première segmente le texte d'entrée en unités morphologiques, c'est-à-dire en formes simples et agglutinées de l'arabe (hyper-formes), puis filtre les chaînes de caractères qui ne relèvent pas de l'analyse morphologique de l'arabe proprement dite. La deuxième étape analyse ces hyper-formes indépendamment de leur contexte. À chaque hyper-forme est attribué, sous forme d'un arbre, l'ensemble de ses segmentations, voyellations, lemmatisations et étiquettes grammaticales possibles. Les lemmes résultants de la lemmatisation sont en fait des hyper-lemmes dans la mesure où ils sont associés à des formes simples ou agglutinées. De façon analogue, les hyper-formes correspondent à des hyper-catégories grammaticales, car elles sont elles aussi associées indifféremment à des formes simples ou agglutinées.

L'approche fondée sur l'utilisation de dictionnaires de formes simples et de règles assure une large couverture, mais s'avérait en pratique assez lente dans la version 2. Pour gagner en temps d'analyse, une nouvelle architecture a tout récemment été développée (G-LexAr v. 3). Elle met en œuvre en frontal un dictionnaire d'hyper-formes où chaque entrée est accompagnée de sa propre arborescence lexicale. Dans ces conditions, l'analyse morphologique de l'arabe est comparable à celle du français ou de l'anglais : elle consiste en un simple accès. Ainsi, elle n'est soumise à une analyse traditionnelle en <proclitique + forme simple + enclitique> que lorsque l'unité morphologique n'est pas reconnue, c'est-à-dire lorsqu'elle ne figure pas dans le dictionnaire d'hyper-formes. Le résultat de cette deuxième étape est une succession de mots accompagnés de leurs arborescences lexicales. La figure 1 en donne un exemple, avec le résultat de l'analyse morphologique du mot كتابهم placé à la racine de de l'arbre signifiant *leur livre* "ktAbhm". Dans cette figure, on distingue :

1. les découpages potentiels du mot en <proclitique + forme simple + enclitique>, ici au nombre de deux : <k + tAb + hm> ou <ktAb + hm> (premier niveau après la racine) ;
2. les voyellations potentielles associées produites par la deuxième étape (deuxième niveau, "kitAbihim" *leur livre*, ou "kuttAbuhum" *leurs écrivains* ;
3. les lemmes associés à ces deux voyellations (troisième niveau, un seul lemme pour "kitAbihim" qui est "kitAb" *livre* et deux lemmes possible pour "kuttAbuhum", le premier est "kuttAb" et le second est "kAtib"). Ces lemmes se présentent sous la forme <proclitique voyellé + lemme voyellé + enclitique voyellé> ;

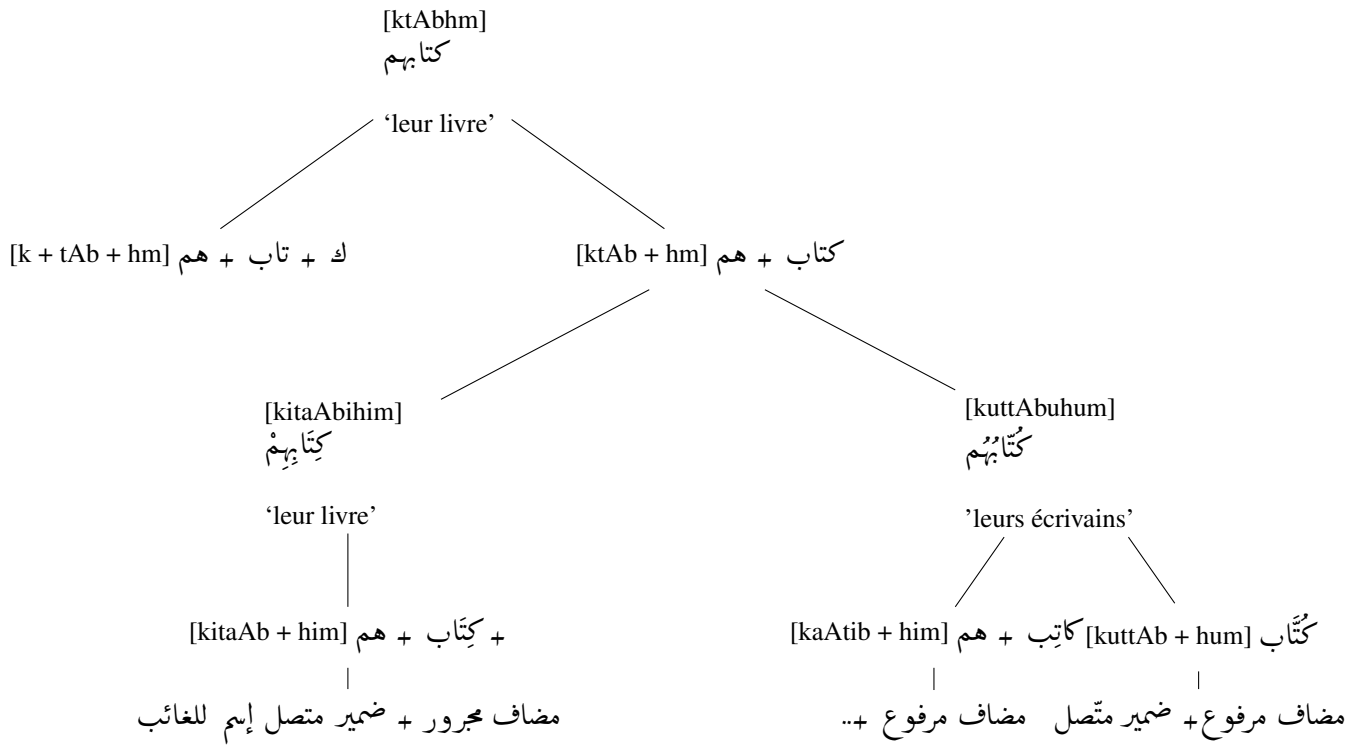

FIGURE 1 – Exemple de représentation arborescente résultant de la deuxième étape avec l'unité morphologique, la forme voyellée, la forme lemmatisée et l'hyper-forme.

4. les étiquettes grammaticales potentielles associées à l'un de ces lemmes (feuilles de l'arbre) ;

Enfin, dans la troisième étape, on procède à l'étiquetage grammatical à proprement parler en élaguant ces arborescences lexicales. L'élagage consiste à ne retenir que les branches dont les feuilles ont des étiquettes vérifiant un certain nombre de règles portant sur la légitimité de la succession de ces étiquettes. Pour rendre ces traitements plus rapides, l'élagage met en œuvre un dictionnaire pré-compilé qui associe aux couples (mot, ensemble d'étiquettes) les arborescences lexicales élaguées qui leur correspondent. À performance linguistique identique et relativement à la version 2 de l'analyseur, cette nouvelle architecture a permis les améliorations suivantes : de 1,3 à 2,2 ko/seconde, la vitesse est passée de 5,7 à 10,8 ko/seconde, soit une analyse trois à cinq fois plus rapide.

2.2 BAMA

Dans les expériences suivantes, nous comparons G-LexAr à l'analyseur morphologique de Buckwalter (BAMA), qui est considéré comme l'un des meilleurs analyseurs de l'arabe, et qui est par conséquent très répandu. Contrairement à G-LexAr, il utilise une approche concaténative du lexique, où les règles morphologiques et orthographiques sont intégrées directement dans le lexique au lieu d'être précisées en fonction des règles générales qui interagissent pour produire la sortie.

Pour chaque chaîne d'entrée, l'analyseur fournit une solution (systématiquement en translittération Buckwalter,

alors que G-LexAr traite le texte arabe sous sa forme brute), comprenant un lemme sous la forme d'un identifiant unique, une ventilation des morphèmes constituants (préfixes, racine, et suffixes), leurs étiquettes grammaticales et la traduction correspondante en anglais. Un exemple est donné dans la figure 2.

```
INPUT STRING: الغاز
  SOLUTION 1: >alogAz
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN
       GLOSS: mysteries/enigmas
  SOLUTION 2: >alogAzu
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN+u/CASE_DEF_NOM
       GLOSS: mysteries/enigmas + [def.nom.]
  SOLUTION 3: >alogAza
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN+a/CASE_DEF_ACC
       GLOSS: mysteries/enigmas + [def.acc.]
  SOLUTION 4: >alogAzi
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN+a/CASE_DEF_GEN
       GLOSS: mysteries/enigmas + [def.gen.]
                    ⋮
```

FIGURE 2 – Exemple de sortie de BAMA : chaque solution consiste ici en un identifiant de lemme (LEMMA_ID), une étiquette grammaticale (POS) et une traduction (GLOSS).

3 Évaluation des analyseurs en traduction automatique statistique

Nous comparons l'analyseur G-LexAr avec BAMA en les utilisant comme outils de pré-traitement sur des tâches de traduction automatique statistique, à l'aide du système *open source* Moses (Koehn *et al.*, 2007). Nous utilisons comme données d'entraînement un échantillon de 251 000 couples de phrases parallèles arabe-anglais extraites d'un corpus constitué d'articles de journaux *(Arabic-English Automatically Extracted Parallel Text)* publié par le LDC *(Linguistic Data Consortium)*. Le corpus ainsi constitué, sous sa forme brute et sans pré-traitement, est appelé « original » par la suite.

Les deux analyseurs produisent une liste de solutions possibles pour chaque hyper-forme, classées selon le score de pertinence. Dans les expériences suivantes, nous ne conservons que la première. Les sorties de l'analyseur G-LexAr comprennent des formes voyellées, lemmatisées et segmentées. BAMA quant à lui ne fournit pas de forme voyellée, aussi nous ne considérons que les deux formes (translittérées) lemmatisées et segmentées. Nous construisons ainsi 6 systèmes de traductions : un en utilisant le corpus original, trois en traitant ce corpus avec chacune des analyses de G-LexAr séparément, et deux avec chacune des analyses de BAMA séparément également.

Les jeux de développement et de test sont constitués respectivement de 500 couples de phrases. Les mesures utilisées pour l'évaluation sont BLEU, TER et mWER. Nous calculons des intervalles de confiance à l'aide de la méthode par ré-échantillonnage par amorce décrite dans (Koehn, 2004) : 1 000 corpus de test de 500 phrases sont constitués par échantillonnage uniforme avec remise à partir des 500 phrases de test mentionnées ci-dessus. Les résultats sont présentés au tableau 1.

Le score de confiance global atteint 95 % dans cette première expérience. Le recours aux analyseurs ne semble pas apporter d'amélioration par rapport au corpus original, et semble même avoir tendance à dégrader légèrement les scores. En particulier, le système ayant recours à la forme voyellée obtient systématiquement de moins bons

TABLE 1 – Scores médians et intervalles de confiance (entre crochets) obtenus par les systèmes de traduction, sur la base de 1 000 corpus de test. Les meilleurs scores selon les mesures TER et mWER sont les plus faibles, les meilleurs selon BLEU sont les plus élevés.

		mWER	BLEU	TER
original		0.4874 [0.4772, 0.4985]	**0.2121** [0.1990, 0.2250]	0.8239 [0.8032, 0.8480]
G-LexAr	voyellée	0.4962 [0.4855, 0.5071]	0.1978 [0.1847, 0.2113]	0.8394 [0.8175, 0.8634]
	lemmatisée	0.5000 [0.4896, 0.5106]	0.1973 [0.1850, 0.2092]	0.8451 [0.8237, 0.8699]
	segmentée	**0.4823** [0.4722, 0.4929]	**0.2066** [0.1850, 0.2092]	**0.8165** [0.7955, 0.8400]
BAMA	lemmatisée	0.4869 [0.4774, 0.4972]	**0.2091** [0.1963, 0.2214]	**0.8111** [0.7905, 0.8332]
	segmentée	**0.4822** [0.4721, 0.4924]	0.1957 [0.1835, 0.2091]	0.8430 [0.8208, 0.8689]
	intersection	[0.4896, 0.4924]	[0.1990, 0.2091]	[0.8237, 0.8332]

résultats, conformément aux expériences de Diab *et al.* (2007). Les différences entre les scores médians selon chacune des trois mesures sont cependant très faibles, et n'évoluent pas toujours dans le même sens d'une mesure à l'autre. En fait, d'après les intervalles de confiance, elles ne sont pas significatives : pour une mesure donnée, les intervalles de scores des systèmes *G-LexAr*, *BAMA* et *original* se chevauchent (les intervalles spécifiés sur la ligne *intersection* du tableau ne sont pas vides). Les deux analyseurs produisent donc des sorties de qualité similaire.

Dans cette première expérience, l'utilisation séparée des formes voyellée, lemmatisée ou segmentée avec l'un ou l'autre des analyseurs n'a pas apporté d'amélioration notable. Par conséquent, dans une deuxième expérience, nous combinons ces formes afin de créer deux nouveaux systèmes que nous appelons « combinés », c'est-à-dire ayant recours à toutes les analyses d'un même analyseur simultanément (voyellée + lemmatisée + segmentée pour G-LexAr, lemmatisée + segmentée pour BAMA). Pour une entrée, toutes les formes correspondantes sont traduites et l'hypothèse de traduction ayant le meilleur score à la sortie de Moses est gardée comme hypothèse finale. Les résultats sont présentés dans le tableau 2.

TABLE 2 – Comparaison des systèmes combinés avec le système original.

	mWER	BLEU	TER
original	0.4876	**0.2121**	0.8244
G-LexAr combiné	**0.4312**	0.2072	**0.7300**
BAMA combiné	**0.4261**	0.2095	**0.7164**

On constate une nette amélioration des scores par rapport au tableau 1. Les deux systèmes combinés sont désormais bien meilleurs que le système original selon TER et mWER (-11 % ou -12 % relativement au système original pour G-LexAr, -13 % pour BAMA sur ces deux mesures), et ne sont que légèrement en retrait selon BLEU (moins d'un-demi point en retrait, soit seulement 2 %). Le gain selon TER et mWER est bien plus important que la légère perte en BLEU. Par conséquent, le recours simultané à toutes les formes produites par un analyseur améliore les résultats d'un système de traduction automatique, alors que le recours à ces formes prises séparément les dégrade, comme l'a montré la première expérience.

4 Conclusion

Cet article a donné un aperçu de l'analyseur G-LexAr, dont la version 3 est plus performante en vitesse de traitement que la version 2. Dans nos expériences en traduction automatique statistique, ses performances se sont

révélées comparables à celles de BAMA, considéré comme la référence en analyse morphologique de l'arabe. G-LexAr a comme avantage indéniable pour les arabisants de traiter directement un texte brut arabe sans nécessiter de translitération intermédiaire.

Les expériences présentées ici confirment les résultats de Diab *et al.* (2007) : la voyellation de l'arabe ne serait pas bénéfique en traduction automatique statistique. Plus généralement, nous avons vu que l'utilisation *séparée* des formes que peut produire un analyseur (voyellation, segmentation et lemmatisation) n'améliore pas les scores, alors qu'une utilisation *combinée* serait bénéfique. Pour aller plus loin, nous envisageons d'étudier plus précisément les contributions positives ou négatives de chacune des formes analysées au sein même des systèmes combinés. L'extension de ces expériences à d'autres domaines ou couples de langues, y compris en utilisant l'arabe en cible, permettra également d'affiner ces résultats.

Enfin, des expériences non rapportées dans cet article ont montré que l'une des faiblesses de l'analyseur G-LexAr est qu'il n'indexe pas encore les mots d'emprunt (voir Gosme *et al.*, 2010). Nous pensons que la résolution de ce problème permettra des résultats encore meilleurs.

Références

BISAZZA A. & FEDERICO M. (2009). Morphological pre-processing for Turkish to English statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, p. 129–135.

BUCKWALTER T. (2002). *Buckwalter Arabic Morphological Analyzer Version 1.0. LDC catalog number LDC2002L49*. Rapport interne, ISBN 1-58563-257-0.

DEBILI F., ACHOUR H. & SOUISSI E. (2002). De l'étiquetage grammatical à la voyellation automatique de l'arabe. *Correspondances*, **71**, 10–28.

DIAB M., GHONEIM M. & HABASH N. (2007). Arabic diacritization in the context of statistical machine translation. In *Proceedings of MT-Summit*.

GOSME J., MEKKI W., LEPAGE Y. & DEBILLI F. (2010). Evaluation of GLexAr through Arabic-English Statistical Machine Translation Systems. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT 2010)*, Paris, France.

HABASH N. (2007). *Arabic Morphological Representations for Machine Translation*, In *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, p. 263–285. Springer Netherlands.

HABASH N. & SADAT F. (2006). Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL*, p. 49–52 : Association for Computational Linguistics.

KOEHN P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, p. 388–395.

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 ACL 2007 demonstration session*.

LEE Y. (2004). Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004*, p. 57–60 : Association for Computational Linguistics.

NIESSEN S. & NEY H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, **30**(2), 181–204.