Etiquetage grammatical de l'arabe voyellé ou non

Fathi DEBILI - Emna SOUISSI
CNRS - CELLMA / IRMC
20, rue Mohamed Ali Tahar - Mutuelleville - Tunis - Tunisie
Tél. (216.1) 584 677 - Fax : (216.1) 797 376
Courrier électronique : debili@ehess.fr

Résumé

Nous abordons le problème de l'étiquetage grammatical de l'arabe en reprenant les méthodes couramment utilisées, lesquelles sont fondées sur des règles de succession de deux ou trois étiquettes grammaticales. Nous montrons que l'on ne peut pas reprendre tels quels les algorithmes préconisés pour le français ou pour l'anglais, la raison étant que l'arabe pose deux problèmes : l'*absence des voyelles* et l'*agglutination des enclinomènes* dont les segmentations potentielles induisent une combinatoire qui conduit à réécrire partiellement ces algorithmes.

Les résultats obtenus pour l'arabe voyellé sont comparables à ce que l'on obtient pour le français ou pour l'anglais. Pour l'arabe non voyellé par contre, les performances chutent assez sensiblement. L'explication réside précisément dans l'absence des voyellations et l'agglutination qui conduisent à une surmultiplication de l'ambiguïté grammaticale de départ. Pour améliorer ces résultats nous définissons un nouveau jeu d'étiquettes grammaticales qui amène à une diminution de l'ambiguïté de départ et à un élargissement de la portée des règles de succession. Ces étiquettes sont associées aux formes non-minimales de l'arabe telles que rencontrées dans les textes. Il y a dans ce cas amélioration sensible, les résultats atteignant des seuils de résolution de 97% pour le voyellé et de 91% pour le non voyellé.

1. Introduction

Dans une très large proportion, les mots sont grammaticalement ambigus. Par exemple *ferme* est hors contexte *substantif, adjectif, verbe* ou *adverbe*. En arabe, كَتَبَ (kataba, *écrire*) est *verbe à la 3ème personne du singulier de l'accompli actif*. La forme non voyellée correspondante كتب (ktb) admet par contre les cinq étiquettes grammaticales potentielles suivantes :

1. *Substantif masculin pluriel* (كُتُبٌ kutubun : *les livres*)
2. *Substantif masculin singulier* (كَتْبٌ katbun: *un écrit*)
3. *Verbe à la 3ème personne masculin singulier de l'accompli actif* (كَتَبَ kataba : *il a écrit* ou كَتَّبَ kattaba : *il a fait écrire*)
4. *Verbe à la 3ème personne masculin singulier de l'accompli passif* (كُتِبَ kutiba : *il a été écrit* ou كُتِّبَ kuttiba, forme factitive correspondante)
5. *Verbe à l'impératif 2ème personne masculin singulier* (كَتِّبْ kattib : *fais écrire*)

auxquelles, en toute rigueur, il conviendrait d'ajouter les étiquettes associées aux deux autres formes voyellées potentielles كَ+تَبٌ ka+tabba (*comme trancher*) et كَ+تَبٌ ka+tabbin (*comme 'tranchement'*).

2. Ambiguïté grammaticale : comptage en définition

Le tableau suivant donne pour l'arabe les proportions de mots grammaticalement ambigus mesurées dans les dictionnaires de formes voyellé et non voyellé.

Dictionnaires arabes	Nombre total d'UL	UL non ambiguës	UL ambiguës	Nbr moyen d'étiquettes/UL
Voyellé	1 047 873	55,64 %	44,36 %	4,30
Non voyellé	502 998	20,60 %	79,40 %	6,42

UL : Unité Lexicale.

Fig. 1. Ambiguïté grammaticale dans les dictionnaires arabes

La lecture de ce tableau est simple. 44,36% des mots voyellés sont ambigus et acceptent 4,3 étiquettes grammaticales en moyenne. Ces chiffres passent respectivement à 79,4% et 6,42 lorsque mesurés sur le non voyellé. Retenons pour l'instant que la différence

est notable ; et voyons ce que ces mesures donnent sur les sous-lexiques associés à un texte voyellé d'abord, puis dévoyellé, les informations grammaticales provenant dans le premier cas du dictionnaire voyellé, et dans le second cas, du dictionnaire non voyellé.

Sous-lexiques arabes	Nombre total d'UL	UL non ambiguës	UL ambiguës	Nbr moyen d'étiquettes/UL
Texte voyellé	8 321	33,54 %	66,46 %	9,14
Texte non voyellé	7 303	4,96 %	95,04 %	12,48

Fig. 2. Ambiguïté grammaticale associée au vocabulaire d'un texte arabe : comptage en définition.

Commentaires :

Nous remarquons là aussi que la version non voyellé est plus ambiguë que la version voyellée. L'on s'y attendait bien sûr. Mais le plus notable est que cette ambiguïté est plus importante encore que celle qui est observée dans les dictionnaires généraux. Pour le voyellé on passe de 44,36% à 66,46% et de 4,3 à 9,14 étiquettes en moyenne. Pour le non voyellé, de 79,4% à 95,04% et de 6,42 à 12,48. En résumé, les sous-lexiques voyellé et non voyellé issus d'un texte donné sont plus ambigus que les dictionnaires généraux voyellé et non voyellé associés à la langue.

Plusieurs facteurs semblent concourir pour expliquer ce constat. Le plus important est l'inversion des proportions noms/verbes que l'on observe lorsque l'on passe des dictionnaires aux sous-lexiques issus de textes, ainsi que les deux tableaux suivants le montrent. L'on passe en effet de la distribution 29% de noms / 71% de verbes dans le dictionnaire voyellé à la distribution 60% / 38% dans le lexique issu du texte voyellé. Pour le non voyellé on passe respectivement de 43% de noms / 60% de verbes à 70% noms / 50% verbes (la somme des proportions n'est pas égale à 100 ; la raison est qu'une même entrée est plusieurs fois comptabilisée lorsqu'elle est ambiguë, c'est à dire lorsqu'elle est à la fois nom, verbe et/ou particule). Il y a donc plus de verbes que de noms dans les dictionnaires généraux, et inversement, plus de noms que de verbes dans les sous-lexiques issus de textes, et ce dans les deux cas voyellé ou non voyellé. Or, précisément, les noms sont en moyenne plus ambigus que les verbes (le dictionnaire voyellé donne 11,63 étiquettes en moyenne pour un nom contre 1,32 étiquettes pour un verbe, et dans sa version non voyellé, 11,68 étiquettes en moyenne pour un nom contre 2,36 en moyenne pour un verbe).

Dictionnaires	UL	Noms	Verbes	Particules
Voyellé	1 047 873	302 260	745 427	186
		29%	**71%**	
Non voyellé	502 998	214 992	304 014	160
		43%	**60%**	

Sous-lexiques	UL	Noms	Verbes	Particules
Texte voyellé	8 321	5 024	3 173	124
		60%	**38%**	**1%**
Texte non voy.	7 303	5 105	3 626	127
		70%	**50%**	**2%**

Ces résultats exhibent d'ores et déjà des niveaux de difficultés bien plus élevés pour l'arabe non voyellé que pour l'arabe voyellé, l'arabe voyellé offrant lui-même des seuils d'ambiguïté supérieur à ceux du français. A titre comparatif, les tableaux suivants donnent les comptages analogues relatifs au français accentué et non accentué[1].

Dictionnaires français	Nombre total d'UL	UL non ambiguës	UL ambiguës	Nbr moyen d'étiquettes/UL
Dict. accentué	293 573	81,26 %	18,74 %	1,20
Dict. non accentué	282 033	80,58 %	19,42 %	1,21

Fig. 3. Ambiguïté grammaticale dans les dictionnaires français

Sous-lexiques français	Nombre total d'UL	UL non ambiguës	UL ambiguës	Nbr moyen d'étiquettes/UL
Texte accentué	15 065	63,75 %	36,25 %	1,31
Texte non accentué	14 235	59,75 %	40,25 %	1,38

Fig. 4. Ambiguïté grammaticale associée au vocabulaire d'un texte français : comptage en définition.

3. Ambiguïté grammaticale: comptage en usage

Pour mieux circonscrire encore les contours du problème de l'étiquetage de l'arabe, considérons maintenant les mêmes mesures effectuées cette fois sur des textes. Ici les comptages tiennent compte de la répétition des diverses unités lexicales qui composent un texte. S'agissant de fréquences en usage, ces comptages offrent une meilleure appréciation du niveau de difficulté de la tâche d'étiquetage.

Le tableau suivant donne précisément les proportions de mots grammaticalement ambigus mesurées dans un texte voyellé et dans sa version dévoyellée.

Texte arabe	Nombre total d'UL	UL non ambiguës	UL ambiguës	Nbr moyen d'étiquettes/UL
Texte voyellé	37 402	37,98 %	62,02 %	5,63
Texte non voyellé	40 485	24,15 %	75,85 %	8,71

Fig. 5. Ambiguïté grammaticale associée au vocabulaire d'un texte arabe : comptage en usage.

Commentaires :

Comparé au tableau donnant les taux d'ambiguïté mesurés sur les sous-lexiques issus de ce même texte d'expérimentation [cf. fig. 2], nous constatons une diminution des proportions des mots ambigus : 62,02% avec répétition (en usage) [respectivement 75,85% pour le non voyellé] contre 66,46% sans répétition (en définition) [respectivement 95,04%], avec en même temps une réduction du nombre moyen d'étiquettes par mot : 5,63 en usage contre 9,14 en définition pour le voyellé, et 8,71 contre 12,48 pour le non voyellé. La répétition textuelle semble donc puiser davantage dans le non ambigu que dans l'ambigu, à l'inverse de ce que nous observons pour le français, ainsi que le tableau suivant le suggère lorsqu'il est comparé au tableau lié au même texte donnant les taux d'ambiguïté mesurés en définition [cf. fig. 4].

Texte français	Nombre total d'UL	UL non ambiguës	UL ambiguës	Nbr moyen d'étiquettes/UL
Texte accentué	427 560	39,19 %	60,81 %	1,86
Texte non accent.	427 560	36,55 %	63,45 %	1,88

Fig. 6. Ambiguïté grammaticale associée au vocabulaire d'un texte français : comptage en usage.

Il reste que même si la répétition textuelle conduit à plus d'ambiguïté dans le cas du français et à moins d'ambiguïté dans le cas de l'arabe, il n'y a pas rapprochement des niveaux de difficulté : l'étiquetage de l'arabe part d'une situation manifestement plus ambiguë, aussi bien en terme de proportion (75,85% des mots sont ambigus dans un texte non voyellé alors qu'ils ne sont que 60,81% à l'être dans un texte accentué), qu'en terme de nombre moyen d'étiquettes par mot (8,71 pour l'arabe contre 1,86 pour le français).

4. Etiquetage grammatical

Les mots qui composent un texte voyellé ou non voyellé sont donc éminemment ambigus. Comment en contexte faire le bon choix. Autrement dit comment associer aux différents mots qui composent un texte l'étiquette qui leur convient compte tenue du contexte où ils occurrent. Tel est le but de l'étiquetage grammatical, problématique posée dès la fin des années 60 [cf. bibliographie].

Le principe de résolution le plus couramment utilisé fait intervenir des règles qui portent sur les successions permises ou non de *deux*, *trois* ou *n* étiquettes grammaticales. Parce que ne permettant pas de résoudre l'ambiguïté dans tout les cas, ces règles se sont vues adjoindre des poids statistiques afin de choisir les résolutions les plus probables. Ces règles peuvent être lues de plusieurs façons : on peut dire par exemple qu'après telle étiquette, ce sont telle ou telle étiquettes qui peuvent suivre ; mais si l'on considère la dernière étiquette on peut également dire qu'elle dépend de celles qui la précèdent. C'est ainsi la formulation probabiliste utilisant les sources de Markov comme modèle qui s'est très vite répandue dès la fin des années 70 [cf. bibliographie].

Force est de constater cependant qu'au plan international, les résultats relatifs à l'étiquetage grammatical n'atteignent que difficilement la barre des 95% de taux de reconnaissance (99% pour l'anglais[2], et 98% pour le français[3] sont des performances obtenues dans des conditions de laboratoire). On ne parvient pas, en effet, à dépasser de façon substantielle et sur de très larges corpus cette barrière de performance. Il ne s'agit pas, bien entendu, de nier les potentialités d'applications auxquelles ont pu conduire les recherches sur l'étiquetage grammatical, quand bien même dans la limite de ces performances. Il reste que cette barrière fini par poser problème. Doit-on faire l'aveu d'un échec : le problème est bien posé mais la solution est difficile à trouver ou n'est que partielle pour l'instant, ou est-ce là le signe d'un problème mal posé, aucune solution complète n'étant à espérer dès lors.

A y regarder de plus près, la situation peut même paraître plus inquiétante encore. En effet, 95% de bonnes reconnaissances correspond en fait à une vision, nous dirons, optimiste des résultats, puisqu'elle occulte le fait que bien des mots dans la langue sont d'emblée non ambigus. Pour le français par exemple, 80% des mots sont non ambigus dans le lexique. Pour

l'arabe voyellé, 55,6% des mots sont non ambigus dans le lexique. Cette proportion passe à 38% lorsque calculée sur des corpus voyellés, elle est de 52% dans un corpus français[3]. Une vision plus précise des performances amène par conséquent à des résultats bien plus sévères puisque les ambiguïtés correctement résolues pour le français ne représentent plus alors que 89,5% (proportion des mots correctement résolus rapporté aux seuls mots ambigus) dans ce cas. C'est donc dire, d'une façon générale, que les résultats obtenus pour l'étiquetage sont pour l'instant assez étonnamment faibles.

Mieux encore. L'on constate que dans la très riche bibliographie relative à l'étiquetage, qu'outre les travaux qui y ont été proprement consacrés, on trouve un très grand nombre de travaux qui se sont fondés sur les résultats de cet étiquetage alors même que celui-ci restait et reste encore non entièrement résolu.
C'est dans ce contexte général et avec ces interrogations qu'est abordé ici le problème de l'étiquetage grammatical de l'arabe voyellé ou non.

Les défis sont multiples : définition d'un jeu d'étiquettes grammaticales pour l'arabe tout d'abord. Voyellation et agglutination qui rendent les algorithmiques développées pour le français ou pour l'anglais inopérantes telles quelles ensuite. Et bien entendu cette fameuse barrière de performance : allons-nous réussir à faire mieux, aussi bien, ou moins bien que 95% de bonnes reconnaissances ?

5. Etiquettes grammaticales : un problème ouvert

Le problème de la définition des étiquettes grammaticales reste en fait ouvert et actuel. C'est que le problème est très difficile. Il suffit, pour s'en rendre compte, de comparer les diverses listes d'étiquettes grammaticales retenues pour le français ou pour l'anglais, pour constater qu'elles sont toutes différentes. Qu'il arrive même qu'au sein d'une même équipe on entretienne, pour une même langue, plusieurs listes d'étiquettes grammaticales[4]. Que de surcroît, dans tous les cas, les critères formels qui y conduisent ne sont nullement entièrement décrits, mais seulement au mieux résumés.

Quoiqu'extrêmement différentes, ces listes ont en commun entre elles qu'elles se fondent néanmoins sur le même héritage : les parties du discours d'une part, et l'hypothèse distributionnelle selon laquelle les mots obéissent à des règles d'agencement d'autre part.

En toute rigueur, il est difficile en fait de trouver deux distributions identiques pour deux mots différents. Il reste que si l'on observe de façon grossière les distributions et que l'on s'attache davantage aux ressemblances qu'aux différences, alors force est d'admettre qu'il y a bien émergence de contextes distributionnels (quasi)identiques, et donc de classes de mots. C'est ainsi que très vite, il se dégage un certain nombre de classes consensuelles comme par exemple la classe des noms, des articles définis, etc. De sorte que, selon les ouvrages scolaires, l'on dénombre pour le français par exemple de dix[5] à quarante classes grammaticales.
Il ne paraît plus dès lors étonnant que les informaticiens linguistes aient construit des ensembles de classes grammaticales dont la cardinalité est variable, allant de la dizaine à quelques centaines. En effet, la nature des algorithmes d'étiquetage d'une part, et la recherche de la performance d'autre part, a conduit les chercheurs à observer avec plus d'acuité les contextes distributionnels. Or, plus cette acuité était grande, plus grand était le nombre d'étiquettes grammaticales définies. C'est ce qui explique que la plupart des systèmes utilisent plus de 100 étiquettes grammaticales.
Si donc pour le français l'on disposait d'une approche: distributionnelle en l'occurrence, et d'un stock de départ : quelques dizaines d'étiquettes grammaticales, de quoi disposions-nous pour l'étiquetage de l'arabe ?
La tradition grammaticale arabe nous lègue en fait un ensemble d'étiquettes morphologiques d'une part [*participe actif, participe passif, nom verbal, ...* / ... اسم الفاعل، اسم المفعول، مصدر،], et un ensemble d'étiquettes syntaxico-sémantiques, d'autre part [*verbe, sujet, complément d'objet, ...* / ... فعل، فاعل، مفعول به،]. Dans le premier cas, c'est la notion de schème qui occupe une place importante, dans le second, les notions de fonction et de cas. Laissons de côté les étiquettes syntaxico-sémantiques dont on peut trouver l'équivalent pour le français ou l'anglais et comparons le reste. Alors que les étiquettes grammaticales du français ou de l'anglais nous viennent de l'approche distributionnelle avec une volonté clairement affichée "*d'écarter toute considération relative au sens*"[5], les étiquettes héritées de l'arabe nous viennent d'une approche où la sémantique côtoie le formel lié à la morphologie du mot, sans référence à la position de ce dernier dans la phrase.

Le fait que l'on ait à faire à des langues à dominance positionnelle d'une part et casuelle d'autre part, n'est sans doute pas étranger à ces différences d'approches ou à cette évolution historique. Il fallait, en effet, se préoccuper d'abord des faits les plus marquants. Cela ne signifie donc pas que l'on ne puisse se préoccuper du cas dans l'analyse du français ou de l'anglais, et de la position dans l'analyse de l'arabe.

C'est en s'inscrivant dans cette perspective qu'une liste d'étiquettes grammaticales a été définie pour l'arabe voyellé. 264 étiquettes ont été ainsi définies en tenant compte des parties du discours (*substantif, verbe, adjectif, ...*), de la flexion casuelle des noms (*nominatif, accusatif, génitif*), de l'état des noms (*déterminé, indéterminé, en annexion*), de l'aspect (*accompli, inaccompli, impératif*), de la modalité (*indicatif, subjonctif, apocopé*), de la voix (*active, passive*), de la personne (*première personne 'locuteur', deuxième personne 'interlocuteur', troisième personne 'absent'*)[6], et bien sûr des relations de position relative qu'entretiennent entre eux les mots et au delà, les étiquettes elles-mêmes. L'idée étant qu'une étiquette nouvelle n'est créée que si elle est discernable.

6. Etiquetage grammatical de l'arabe

Expériences préliminaires : nous donnons ici les résultats de l'étiquetage grammatical d'un texte arabe voyellé d'abord, puis dévoyellé, dans deux conditions expérimentales :

1. avec un jeu de 264 étiquettes grammaticales ne faisant pas intervenir le genre et le nombre (GN) d'une part,
2. et avec un jeu de 606 étiquettes grammaticales faisant intervenir le genre et le nombre d'autre part.

L'étiquetage est fondé sur l'emploi de règles de succession binaires et ternaires apprises sur le texte lui-même. Les conditions d'expérimentation peuvent donc être considérées comme idéales. Ajoutons enfin que l'étiquetage recherché n'est pas déterministe. Si les règles ne suffisent pas à elles seules à résoudre, alors on conserve l'ambiguïté, éventuellement réduite, mais on ne cherche pas à choisir la résolution la plus probable parmi plusieurs.

L'évaluation de la performance de l'étiqueteur est exprimée en termes de résolution d'une part et de réduction de l'ambiguïté d'autre part. Les tableaux suivants donnent les performances mesurées sur un texte comptant 25 410 unités morphologiques (chaînes de caractères comprises entre deux séparateurs forts).

Les comptages portent tantôt sur les unités morphologiques (UM), tantôt sur les unités lexicales (UL) qui composent les unités morphologiques du fait de l'agglutination. Les proclitiques, les formes simples et les enclitiques sont des unités lexicales. Les formes simples lorsque isolées dans le texte et les formes agglutinées sont des unités morphologiques.

Texte voyellé	Nombre d'UL	Etiquettes / UL	Résolution (%)
Après AM	37 097	1,91	56,37
EG sans GN (264)	36 695	1,04	94,84
EG avec GN (606)	36 688	1,02	97,37

AM : Analyse Morphologique. EG : Etiquetage Grammatical.

Texte non voyellé	Nombre d'UL	Etiquettes / UL	Résolution (%)
Après AM	40 121	5,33	27,97
EG sans GN (264)	36 974	1,20	82,31
EG avec GN (606)	36 855	1,10	90,74

Fig. 7. Performances de l'étiquetage vues sous l'angle des UL

Bien que dans des conditions d'expérimentation idéales, nous remarquons que ces résultats atteignent à peine la qualité des résultats affichés pour le français ou l'anglais, alors même que ceux-ci sont obtenus dans des conditions d'expérimentation somme toute plus difficiles, puisque mesurés sur des textes n'ayant pas en principe participé à l'entraînement. De surcroît, ces résultats ne sont comparables que dans le cas de l'étiquetage du texte voyellé. Pour le texte non voyellé, les résultats sont à l'évidence nettement moins bons. Or, c'est l'arabe non voyellé qui est le plus répandu et qui, bien entendu, nous intéresse le plus.

Reprendre tel quel l'étiquetage grammatical fondé sur des règles de succession pour traiter l'arabe non voyellé n'est donc pas acceptable, d'autant plus que ce que nous avons obtenus, nous l'avons obtenus, rappelons-le, dans des conditions d'apprentissage ad-hoc. Essayons de voir les raisons qui ont pu conduire à une telle baisse des performances.

7. De l'absence des voyelles : ambiguïté vocalique

Considérons les expérimentations faites avec le jeu des étiquettes grammaticales sans genre-nombre. Les tableaux de la figure 7 montrent qu'il y a chute des performances lorsque l'on passe de l'étiquetage du texte voyellé à l'étiquetage de sa version non voyellée. Les taux de résolution passent de 94,84% à 82,31%.

Si l'on tient compte du genre et du nombre la dégradation des performances nous fait passer de 97,37% à 90,74%. Sous l'angle de la réduction de l'ambiguïté il y a aussi dégradation. On observe en effet que l'on passe de 1,02 étiquettes en moyenne par unité lexicale après étiquetage à 1,10 dans le cas qui donne les meilleurs résultats, c'est-à-dire avec genre nombre.

L'explication est simple. Elle réside d'abord dans la surmultiplication de l'ambiguïté qu'occasionne la dévoyellation, dévoyellation dont l'effet se manifeste doublement ainsi que les tableaux de la figure 7 l'exhibent assez bien. On remarque en effet qu'à l'entrée de l'étiqueteur, selon que le texte est voyellé ou non, les proportions de mots ambigus ne sont pas les mêmes, ni le nombre moyen des étiquettes potentielles qui leurs sont associées.

Ainsi, le texte voyellé se présente à l'entrée de l'étiqueteur avec 56,37% de mots non ambigus contre seulement 27,97% pour le texte non voyellé. Autrement dit, 43,63% des mots sont ambigus dans le texte voyellé, proportion qui grimpe à 72,03% lorsque le texte est dévoyellé. De surcroît, les mots sont bien plus ambigus dans le texte non voyellé que dans le texte voyellé : 5,33 étiquettes en moyenne pour le texte non voyellé, alors que l'on ne compte que 1,91 étiquettes en moyenne pour le texte voyellé.

8. Agglutination : ambiguïté segmentale

L'explication réside ensuite dans la surmultiplication de l'ambiguïté qu'occasionne l'agglutination. Celle-ci induit en effet pour le non voyellé un nombre de segmentations en

proclitique + forme simple + enclitique (*p + fs + e*)

plus important que pour le voyellé.

Par exemple, le mot ألمهم ('alamuhum *leur douleur*) dans sa forme voyellée n'accepte qu'une seule segmentation : ألـمُ + هُمْ ('alamu+hum)

Dans sa forme non voyellée المهم ('lmhm) le même mot accepte par contre les trois segmentations suivantes :

- أ + لـمّ + هم ('+lmm+hm *les a-t-il ramassés*)
- ألـم + هم ('lm+hm *leur douleur,* 'llm+hm *il les a fait souffrir*)
- الـ + مهم ('l+mhm *l'important*)

Le tableau suivant don't la lecture est analogue à celle des tableaux de la figure 7, donne pour un texte arabe les proportions d'UM acceptant respectivement une seule ou plusieurs segmentations.

	Nombre UM	UM non ambiguës	UM ambiguës	Seg./ UM	Nbr max de seg.
Voyellé	25 410	96,61 %	3,39 %	1,03	4
Non voy.	25 410	78,00 %	22,00 %	1,30	6

Fig. 8. Unités morphologiques donnant lieu à des segmentations en *proclitique + forme simple + enclitique* ambiguës

Sous l'angle de l'agglutination, on remarque donc que la segmentation d'un texte non voyellé est bien plus ambiguë que celle de son correspondant voyellé :

- Le nombre d'unités admettant plus d'une segmentation est d'abord plus important : 22% contre 3,39%.
- De plus, le nombre moyen de segmentations par unité est plus grand pour le non voyellé que pour le voyellé : 1,3 segmentations en moyenne contre 1,03 pour le voyellé. Le tableau indique en outre que le nombre maximal de segmentations observées est de 4 pour le voyellé et de 6 pour le non voyellé.

L'intensification de l'ambiguïté de segmentation s'opère donc selon deux axes :

- en proportion d'abord selon l'axe horizontal (il y a plus d'unités ambiguës dans un texte non voyellé que dans son correspondant voyellé),
- mais aussi en profondeur selon l'axe vertical (il y a plus de segmentations dans le dévoyellé).

La conséquence est que cela introduit une deuxième source de surmultiplication de l'ambiguïté grammaticale qui vient se surajouter à celle qui est due à l'absence de voyellations. La combinatoire concaténative des étiquettes liées aux différentes segmentations introduit en effet au compte d'une unité morphologique un nombre d'*étiquettes apparentes* bien plus grand que le nombre d'étiquettes associées aux unités lexicales dont elle est constituée.

Exemple :

Le mot وِفَـاقٍ (*accord*) tel que voyellé, n'accepte qu'une seule segmentation. La cardinalité de l'ambiguïté est dans ce cas égale à 2 {*substantif génitif indéterminé, complément de nom indéterminé*}. Le même mot non voyellé accepte par contre deux segmentations :

- وفاق (*accord*)
- و + فاق (*et il a dépassé*)

Ces deux segmentations engendrent 9 étiquettes apparentes, toutes associées à l'unité morphologique non voyellée وفاق. La figure suivante montre le processus qui y conduit :

Unité morphologique	وفاق

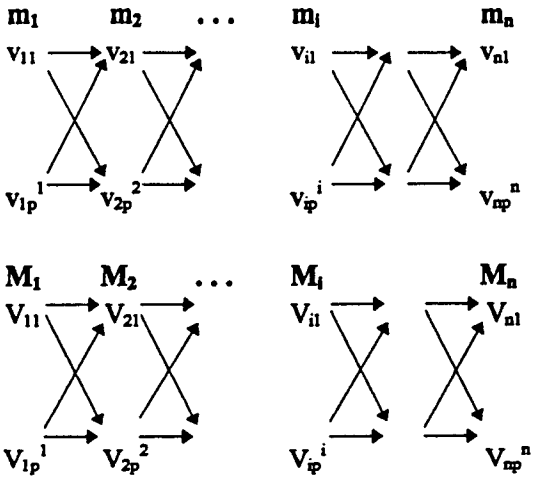
	Encl	forme simple	Proc
Segmentation 1		(*accord*) وِفَاق.	
Segmentation 2		(*dépasser*) فَاقَ	وَ (*et*)

étiquettes grammaticales associées à la segmentation 1		ضَ	سِ	سْ	
		ضَا	ضِ	ضُ	
étiquettes gram. associées à la segmentation 2		فعل ماض معلوم			حرف عطف
					واو رب
					واو المعية

étiquettes grammaticales apparentes associées à l'UM	ضَا	ضِ	ضُ	ضَ	سِ	سْ	
	حرف عطف + فعل ماض معلوم						
	واو رب + فعل ماض معلوم						
	واو المعية + فعل ماض معلوم						

Ainsi, c'est la combinatoire des étiquettes des diverses unités segmentales qui conduit aux 9 étiquettes apparentes associées au mot وفـــاق. Nous les avons qualifiées d'apparentes parce qu'elles sont tantôt simples, tantôt composées, et que ce sont elles précisément qui interviennent dans l'élaboration des différentes continuités grammaticales que tente de construire l'étiqueteur au moyen des règles de succession.

L'ambiguïté vocalique et l'ambiguïté segmentale introduisent donc une surmultiplication de l'ambiguïté grammaticale. Conséquence : une explosion combinatoire au niveau de l'étiquetage bien plus importante encore que celle provoquée par l'analyse d'un texte où le nombre d'étiquettes par mot en moyenne est plus petit en même temps que leur proportion.



Ces deux schémas essaient d'illustrer notre propos. Dans le premier schéma les m_i et les v_{ij} représentent respectivement les unités lexicales et les étiquettes qui leurs sont associées. Les flèches représentent les chemins grammaticaux combinatoirement possibles.

Dans le second schéma les M_i et les V_{ij} représentent respectivement les unités morphologiques et les ensembles de listes d'étiquettes apparentes qui leurs sont associées. Les flèches représentent cette fois non plus une continuité grammaticale simple, mais un ensemble de continuités potentielles puisque les V_{ij} sont non pas des étiquettes simples mais des ensembles de listes d'étiquettes.

Exemple :

الحـــق	بـــيد	صاحبـــه
الـ حـــق	بـ يد	صاحب ه
الحـــق	بـــيد	
أ لحـــق		
أل حـــق		

أن سِ	ح حرّ سِ	ضمو ضُ
أن سِ	ح حرّ ضِ	ضمو ضُ
أن سِ	سْ	ضمو ضِ
أن ضِا	سِ	
ف ماض	سَلا	
ف أمر	ضِا	
ف مضارع		
ف مضارع		
ف مضارع		
استفهام ف ماض		
استفهام ضَ		
استفهام ضُ		
استفهام سْ		
نداء ضَ		
نداء ضُ		
نداء سْ		
استفهام وحر ضِ		
استفهام وحر سِ		
استفهام وتأكيد ضُ		
استفهام وتأكيد سْ		

Fig. 9. Ambiguïté segmentale

9. Performance de l'étiquetage grammatical

Les tableaux suivants donnent les performances de l'étiquetage grammatical comptabilisées sous l'angle des unités morphologiques. Ces tableaux offrent une autre vision des résultats déjà présentés dans les tableaux de la figure 7. Par rapport à ceux-ci, les comptages portent ici non sur les étiquettes associées aux UL, mais sur les étiquettes apparentes associées aux UM, précisément reconstituées à partir des étiquettes que l'analyseur a retenu au compte des UL.

Texte voyellé	Nombre d'UM	Etiquettes apparentes / UM	Résolution (%)
Après AM	25 410	2,65	44,53
EG sans GN	25 410	1,06	92,81
EG avec GN	25 410	1,03	96,28

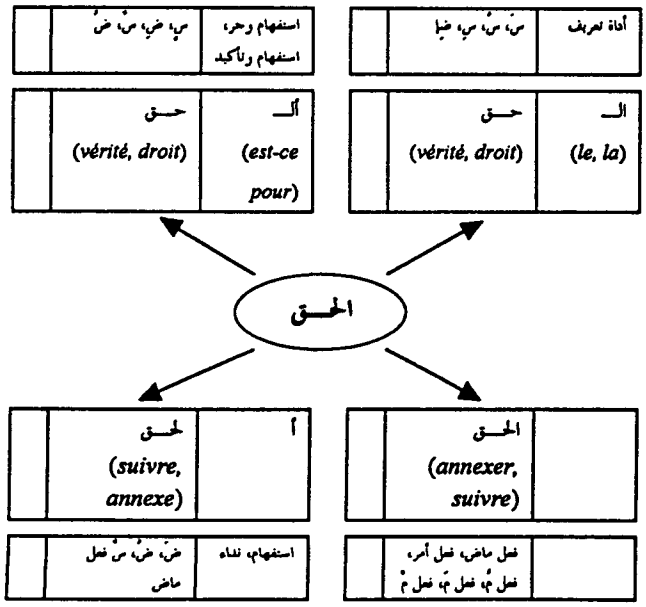
Texte non voyellé	Nombre d'UM	Etiquettes apparentes / UM	Résolution (%)
Après AM	25 410	10,97	18,06
EG sans GN	25 410	1,28	76,86
EG avec GN	25 410	1,14	87,77

Fig. 10. Performances de l'étiquetage vues sous l'angle des UM

L'appréciation des performances du même étiquetage s'avère donc plus sévère lorsque ces performances sont évaluées sous l'angle des UM que sous l'angle des UL. La raison est simple : dès lors qu'une unité lexicale reste ambiguë, elle contamine l'unité morphologique dont elle fait partie.

10. Résolutions locales et étiquettes grammaticales composées

Considérons l'unité morphologique الحق. Non voyellée cette unité donne lieu aux diverses segmentations et ambiguïtés grammaticales suivantes :



La combinatoire concaténative des étiquettes engendre les successions potentielles suivantes, successions que nous avons appelées plus haut *étiquettes apparentes* :

أداة تعريف + س
أداة تعريف + ض
أداة تعريف + ص
أداة تعريف + ضا
استفهام وحر + س
استفهام وحر + ض
استفهام وحر + ص
استفهام وحر + ض
استفهام وتأكيد + س
استفهام وتأكيد + ض
استفهام وتأكيد + ص
استفهام وتأكيد + ض
فعل ماض
فعل أمر
فعل مضارع (أ)
فعل مضارع (ن)
فعل مضارع (ي)
استفهام + ض
استفهام + ض
استفهام + ص
استفهام + فعل ماض
نداء + ض
نداء + ض
نداء + ص
نداء + فعل ماض

La résolution locale conserve les successions licites suivantes :

أداة تعريف + س
أداة تعريف + ض
أداة تعريف + ص
أداة تعريف + ضا
استفهام وحر + س
استفهام وحر + ض
استفهام وتأكيد + س
استفهام وتأكيد + ض
فعل مضارع (أ)
فعل مضارع (ن)
فعل مضارع (ي)
استفهام + ض
استفهام + ض
استفهام + ص
استفهام + فعل ماض
نداء + ض
نداء + ض
نداء + ص

Nous appelons *étiquettes grammaticales composées* ces successions licites.

Pour améliorer les résultats de l'étiquetage l'idée est tout naturellement venue d'utiliser un nouveau jeu d'étiquettes grammaticales constitué par la réunion ensembliste des étiquettes simples et des étiquettes composées. 1730 étiquettes ont ainsi été redéfinies, avec les avantages pressentis suivants :

* utiliser un étiqueteur où l'on n'ait plus à traiter la combinatoire due à l'agglutination, en tout point donc analogue à ceux du français ou de l'anglais ;
* traiter des textes présentant des seuils d'ambiguïté moindre en termes de proportion et de nombre moyen d'étiquettes ;
* enfin, augmenter la portée des règles de successions puisque celles-ci concernent désormais des unités morphologiques et non lexicales. Les deux exemples suivants montrent comment en effet les règles de succession ternaires peuvent embrasser jusqu'à 9 unités lexicales composant précisément 3 unités morphologiques. Dans le premier exemple la règle ternaire embrasse 3 unités lexicales, alors que dans le second elle embrasse 9 unités lexicales.

كـرّاس تلميذ مجتهد ⇒ 3 UL

cahier d'un élève studieux

بـكتابـه وكرّاسـه وقلمـه

بـ+كتاب+ـه و+كرّاس+ـه و+قلم+ـه ⇒ 9 UL

avec son livre et son cahier et son crayon

A titre indicatif, le tableau suivant donne le nombre de successions ternaires avec répétition comptabilisées dans le texte d'expérimentation de 25 410 unités morphologiques. Nous observons que dans 81,5% des cas nous avons à faire à des successions qui mettent en oeuvre des étiquettes composées. C'est dire que dans 81,5% des cas nous avons besoin de règles qui portent sur des unités non-minimales. L'importance de cette proportion justifie donc que l'on essaie d'utiliser des règles à large portée.

Nbr de successions ternaires engendrées	25 408
Nbr de successions mettant en oeuvre minimalement une étiquette composée	20 701 (81,5%)
Nbr de successions mettant en oeuvre des étiquettes simples	4 707 (18,5%)

A l'appui encore de cette remarque le tableau suivant qui donne, pour le même texte d'expérimentation, les proportions des diverses UM considérées sous l'angle de leur formation. 41% des UM sont composées, ce qui signifie que dans le texte plus d'une unité sur trois est composée.

Nature de l'UM	*fs*	*p + fs*	*fs + e*	*p+ fs +e*	*p + e*
Nbr d'UM	14 755	7 188	2 496	620	351
Proportion	59%	28%	10%	2%	1%

11. Performance de l'étiquetage portant sur les UM

L'étiquetage du même texte, dans des conditions d'apprentissage toujours ad-hoc, utilisant ce nouveau jeu d'étiquettes a donné les résultats suivants :

Texte voyellé	Nombre d'UM	Etiquettes composées / UM	Résolution (%)
Après AM	25 410	2,44	45,15
EG étq. composées	25 410	1,02	97,51

Texte non voyellé	Nombre d'UM	Etiquettes composées / UM	Résolution (%)
Après AM	25 410	8,74	18,34
EG étq. composées	25 410	1,13	91,55

Fig. 11. Performances de l'étiquetage utilisant le jeu de 1730 étiquettes simples et composées

Par comparaison avec le tableau de la figure 10, nous enregistrons les améliorations suivantes :

1. pour le voyellé, une augmentation des taux de résolution qui passent de 96,28% à 97,51%,
2. pour le non voyellé, une amélioration qui fait passer la résolution de 87,77% à 91,55%.

12. Conclusion

Parce que les seuils d'ambiguïté de l'arabe, non voyellé notamment, étaient d'emblée bien plus élevés que ceux du français par exemple, que par conséquent les niveaux de difficulté pressentis étaient supérieurs, nous avons voulu tester les potentialités de l'étiquetage grammatical tel que traditionnellement pratiqué. Dans cette perspective, nous avons choisi dans un premier temps de mener des expérimentations dans des conditions d'apprentissage ad-hoc. Trois jeux d'étiquettes grammaticales ont été définies, mais seul le dernier a donné des résultats satisfaisants.

Les tests effectués sur des textes n'ayant pas participés à l'apprentissage ont donné des résultats

bien moins satisfaisants. Les tableaux suivants sont relatifs aux performances obtenues sur le livre de '*Kalila et Dimna*' comptant 39800 UM.

Kalila et Dimna Texte voyellé	Nombre d'UM	Etiquettes apparentes / UM	Résolution (%)
Après AM	39 800	2,51	49,48
EG sans GN	39 800	1,05	84,89
EG avec GN	39 800	1,08	85,42
EG étq. composées	39 800	1,09	85,57

Kalila et Dimna Texte non voyellé	Nombre d'UM	Etiquettes apparentes / UM	Résolution (%)
Après AM	39 800	11,60	16,64
EG sans GN	39 800	1,14	62,89
EG avec GN	39 800	1,21	64,51
EG étq. composées	39 800	1,19	63,03

Fig. 12. Performances de l'étiquetage sur des nouveaux textes

Il y a chute des performances ainsi que nous pouvons le constater. Mais parce que davantage liée au manque d'apprentissage, nous choisissons pour l'instant de ne point interpréter ces résultats, même si a priori ils tendent en fait à conforter nos conclusions.

Car il reste que même dans des conditions d'entraînement ad-hoc, les résultats obtenus ne sont satisfaisants et prometteurs que dans une seule perspective, à savoir l'utilisation d'un jeu de plus de 1700 étiquettes grammaticales associées aux formes non-minimales de l'arabe. Conclusion difficile donc puisque se pose immédiatement le problème de l'entraînement, qui plus est, dans des conditions relativement nouvelles. En effet, nous ne connaissons pas d'expérimentations qui se soient effectuées avec autant d'étiquettes grammaticales, 100 à 250 étant le nombre d'étiquettes couramment mis en oeuvre. Nous croyons voir dans ces résultats et dans ces nécessités une mesure des difficultés que reste poser finalement l'étiquetage de l'arabe non voyellé et, au delà, l'étiquetage fondé sur la seule utilisation de règles de succession ternaires.

Bibliographie

Alexandre ANDREEWSKY, Christian FLUHR
A learning method for natural language processing and application to information retrieval, IFIP Congress, August 1974.

Fathi DEBILI
Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage, Thèse de Docteur-Ingénieur, Paris VII, Septembre 1977.

Christian FLUHR
Algorithmes à apprentissage, Thèse de doctorat d'état, Paris XI, 1979.

Marc EL-BEZE
Les modèles de langage probabilistes: Quelques domaines d'application, Habilitation à diriger des recherches, décembre 1992, Université de Paris 13.

Bernard MERIALDO
Tagging English Text with Probabilistic Model, Computational Linguistics, June 1994, Volume 20, Number 2.

Kenneth W. CHURCH and Robert L. MERCER
Introduction to the Special Issue on Computational Linguistics Using Large Corpora, Computational Linguistics, March 1993, Volume 19, Number 1.

[1] **Marc EL-BEZE, Bernard MERIALDO, Bénédicte ROZERON, Anne-Marie DEROUAULT**
Accentuation automatique de textes par des méthodes probabilistes, Technique et science informatique N°6/1994.

[2] **Atro VOUTILAINEN**
A syntax-based part-of-speech analyser, Research Unit for Multilingual Language Technology, Finland. 7th conference of the European chapter of the Association for Computational Linguistics. 27-31/03/1995 University College, DUBLIN.

[3] **Jean Pierre CHANOD and Pasi TAPANAINEN**
Tagging French comparing statistical and a constraint-based method, Rank Xerox Research Centre, Grenoble. 7th Conference of the European Chapter of the Association for Computational Linguistics. 27-31/03/1995 University College, DUBLIN.

[4] **Emna SOUISSI**
Etiquetage grammatical de l'arabe voyellé ou non, Thèse de doctorat, Université de Paris VII, Octobre 1997.

[5] **J. DUBOIS, L. GUESPIN, M. GIACOMO, C. et J.-B. MARCELLESI, J.-P. MEVEL**
Dictionnaire de linguistique et des sciences du langage, Larousse 1994.

[6] **Djamal Eddine KOULOUGHLI**
Grammaire de l'arabe d'aujourd'hui, Pocket - Langues pour tous, 1994.