# Analyse automatique *vs* analyse interactive : un cercle vertueux pour la voyellation, l'étiquetage et la lemmatisation de l'arabe

Fathi DEBILI[1], Zied BEN TAHAR[1], Emna SOUISSI[2]
[1] LLACAN, INALCO, CNRS
7, rue Guy Môquet, 94801 Villejuif cedex, France
[2] ESSTT, 5, Avenue Taha Hussein – 1008 Tunis
fathi.debili@wanadoo.fr, bentaharzied@gmail.com,
emna.souissi@planet.tn

**Résumé.** Comment produire de façon massive des textes annotés dans des conditions d'efficacité, de reproductibilité et de coût optimales ? Plutôt que de corriger les sorties d'analyse automatique moyennant des outils d'éditions éventuellement dédiés, ainsi qu'il est communément préconisé, nous proposons de recourir à des outils d'analyse interactive où la correction manuelle est au fur et à mesure prise en compte par l'analyse automatique. Posant le problème de l'évaluation de ces outils interactifs et du rendement de leur ergonomie linguistique, et proposant pour cela une métrique fondée sur le calcul du coût qu'exigent ces corrections exprimé en nombre de manipulations (frappe au clavier, clic de souris, etc.), nous montrons, au travers d'un protocole expérimental simple orienté vers la voyellation, l'étiquetage et la lemmatisation de l'arabe, que paradoxalement, les meilleures performances interactives d'un système ne sont pas toujours corrélées à ses meilleures performances automatiques. Autrement dit, que le comportement linguistique automatique le plus performant n'est pas toujours celui qui assure, dès lors qu'il y a contributions manuelles, le meilleur rendement interactif.

**Abstract.** How can we massively produce annotated texts, with optimal efficiency, reproducibility and cost? Rather than correcting the output of automatic analysis by means of possibly dedicated tools, as is currently suggested, we find it more advisable to use interactive tools for analysis, where manual editing is fed in real time into automatic analysis. We address the issue of evaluating these tools, along with their performance in terms of linguistic ergonomy, and propose a metric for calculating the cost of editing as a number of keystrokes and mouse clicks. We show, by way of a simple protocol addressing Arabic vowellation, tagging and lemmatization, that, surprisingly, the best interactive performance of a system is not always correlated to its best automatic performance. In other words, the most performing automatic linguistic behavior of a system is not always yielding the best interactive behavior, when manual editing is involved.

**Mots-clés :** analyse automatique *vs* interactive ; annotation séquentielle, parallèle ; voyellation, lemmatisation, étiquetage de l'arabe ; métrique pour l'évaluation de l'analyse interactive.

**Keywords:** automatic versus interactive analysis of Arabic, proposal of metrics for evaluating the interactive analysis, design and implementation of software for interactive vowellation, lemmatisation and POS-tagging of Arabic, evaluation.

# 1 Introduction

L'analyse automatique semble avoir précédé l'analyse interactive, laquelle signifie intervention manuelle. Elle a été la préoccupation première des chercheurs, pour la plupart d'entre eux et dès le départ[1], et sans doute restera-t-elle longtemps encore le but à atteindre, la dimension à parfaire. L'analyse manuelle que nous dirons « artisanale » a, elle aussi, été pratiquée d'emblée, avec des objectifs divers, en particulier celui de la confection de corpus annotés orientés vers l'apprentissage ou l'évaluation. Même si l'on s'est très vite rendu compte de la difficulté matérielle qu'il y avait à produire de l'analyse manuelle, ce n'est que tardivement, sous la pression d'une double exigence, de performances et de plus large couverture, que l'on y a consacré des efforts soutenus. Avec la rédaction de guides d'annotation pour rendre l'opération autant que faire se peut reproductible (cf. l'action GRACE par exemple, Adda et al. 1999, Véronis, 1999, Abeillé et Clément, 2003). Puis avec la confection d'outils informatiques dédiés où la part de l'automatique au service du manuel a été peu à peu introduite et amplifiée (Habert, 2005). Le présent travail s'inscrit dans cette dynamique qu'il prolonge. Nous abordons les problèmes que pose l'annotation massive, manuellement vérifiée et corrigée, de corpus arabes. Autrement dit, de l'analyse morpho-grammaticale interactive de l'arabe.

Au travers des difficultés que présentent la voyellation, l'étiquetage et la lemmatisation de l'arabe, cf. partie 2, des coûts prohibitifs qu'elles engendrent sous l'angle de la vérification et saisie manuelle, cf. partie 3, nous décrivons, partie 4, les spécifications qui nous ont amenés à développer une analyse interactive vue non pas comme indépendante de l'analyse automatique, même si elle en utilise les résultats qu'elle est sensée lui renvoyer éventuellement corrigés, mais bien comme une extension rétroactive de celle-ci.

Soulevant le problème que pose l'évaluation des performances de l'annotation interactive, nous montrons qu'il y a intrication entre les deux processus, automatique et interactif, où le service rendu mutuel va au-delà du simple échange de données annotées. L'on constate en effet que l'exigence de meilleures performances pour les procédures d'analyse interactive, qui passent par la définition de diverses ergonomies linguistiques intuitives et efficaces, amène à reconsidérer la conception même des algorithmes de la dimension automatique.

Deux ergonomies linguistiques se dégagent. La première, séquentielle, est celle, classique, qui vient naturellement à l'esprit. Elle est liée au fait que les vérifications annotations manuelles nécessitent en général que soit consulté le contexte du mot en cours de vérification. Cette ergonomie s'avère très lente, et donc peu productive, cf. partie 5. La seconde, parallèle, essaie de parer à cette lenteur en mettant à profit le fait que bon nombre de mots apparaissent souvent, le gain projeté étant alors que l'on puisse tous les vérifier et annoter en même temps. Cette ergonomie s'avère plus productive mais est plus difficile à mettre en œuvre, cf. partie 6.

Faisant ainsi converger nos préoccupations vers la réalisation d'un système intégré, comment évaluer les performances des ergonomies linguistiques et interactives qui en constituent l'interface, performances qui relèvent *a priori* du qualitatif, et qui en même temps restent dépendantes des performances des traitements d'analyse automatique qui, eux, constituent le cœur du système ? Une métrique et un protocole expérimental sont proposés pour la mesure

[1] Les diverses applications *assistées par ordinateur* (*x*. A. O.) ne visaient pas la confection massive de données dictionnairiques ou textuelles annotées.

des performances de l'analyse interactive, lesquelles ne se calculent pas de la même façon que celles de l'analyse automatique.

Résultats d'expérimentations et commentaires sont livrés parties 5 et suivantes.

## 2 Des niveaux d'ambiguïté élevés

Le mot arabe, tel qu'on le rencontre dans les textes, c'est-à-dire sous sa forme fléchie, simple ou agglutinée (*proclitique+forme simple+enclitique*, que nous conviendrons d'appeler hyper-forme), présente des niveaux d'ambiguïté segmentale, vocalique, casuelle, lemmatique, et grammaticale relativement élevés. Le tableau 1 donne à titre indicatif les valeurs moyennes mesurées en définition (comptages effectués sur des données dictionnairiques : un dictionnaire de 66 millions d'entrées non voyellées obtenues par synthèse lexico-syntagmatique, un autre de 157 mille entrées issues d'un corpus de 2 millions d'occurrences), et en usage (comptages effectués sur des données textuelles : ici, sur les 2 millions d'occurrences du corpus précité).

| Ambiguïté | Segmentale | Vocalique et Casuelle | Lemmatique | Grammaticale |
|---|---|---|---|---|
| Dictionnaire $66.10^6$ | 1,08 | 2,17 | 1,68 | 2,99 |
| Sous lexique 157 031 | 1,26 | 6,40 | 2,65 | 9,16 |
| En usage | 1,32 | 7,84 | 3,66 | 10,76 |

Tableau 1 : Niveaux d'ambiguïté de l'hyper-forme arabe

Ces valeurs placent l'arabe à des niveaux d'ambiguïté sensiblement plus élevés que ceux du français. Elles se rapportent en effet, ainsi que nous venons de le dire, non pas aux formes simples de l'arabe, dont les niveaux d'ambiguïté sont plus élevés encore (Debili et al. 2002), mais aux formes simples et agglutinées. Une autre mesure, plus globale, a pu être effectuée. Elle se rapporte au niveau d'ambiguïté composée, c'est-à-dire toutes ambiguïtés segmentales, vocaliques, casuelles, lemmatiques, et grammaticales confondues. La synthèse lexico-syntagmatique donne en effet pour 500 mille formes fléchies simples non voyellées arabes, 305 millions de formes simples et agglutinées, voyellées, lemmatisées et étiquetées, différentes, correspondant à 66 millions de formes simples et agglutinées non voyellées. Le rapport de 305 sur 66 conduit à une ambiguïté moyenne d'environ 4,6 acceptions morpho-grammaticales différentes par entrée. Ce chiffre est de 14,7 si les comptages sont effectués sur le sous lexique de 157 mille entrées. En usage, comptages effectués sur le texte de 2 millions d'occurrences, cette moyenne passe à 16,7 acceptions morpho-grammaticales différentes par occurrence. L'on peut remarquer, incidemment, que la répétition textuelle semble ainsi puiser davantage dans l'ambigu que dans le non ambigu.

Ces niveaux d'ambiguïtés sont relativement importants. Nous ne disposons pas de chiffres équivalents pour le français ou l'anglais. Il nous faudrait pour cela considérer les ambiguïtés liées non pas seulement aux formes fléchies, mais aussi aux syntagmes constitués de ces formes simples et des mots vides (articles, prépositions, pronoms, etc.) qui peuvent leurs être adjoints, afin d'établir le parallèle avec l'arabe où ces mots s'attachent sous forme de proclitiques et d'enclitiques. Dans la terminologie de Lucien Tesnière, considérer les *mots constitutifs* accompagnés de leurs *mots subsidiaires* ou *satellites* (Tesnière, 1969, p. 57, §18).

Dans une perspective d'annotation manuelle, au-delà des difficultés à caractère linguistique (définition des étiquettes, critères de choix, etc.) dont nous admettrons qu'elles puissent être comparables d'une langue à une autre, ces niveaux d'ambiguïté indiquent que l'opération d'annotation sera sans doute comparativement plus coûteuse au plan matériel qu'elle ne peut l'être pour le français par exemple, l'étendue des choix étant plus large. Avec la saisie des voyelles, la situation va être plus critique encore.

## 3 Des coûts d'annotation et de saisie élevés

En effet, en arabe, la plupart des lettres (87% en définition, 77% en usage) demandent pour être voyellées d'être accompagnées d'un signe diacritique dont la saisie coûte 2 frappes au clavier, à l'image du tréma en français. La saisie des lettres voyellées en arabe est donc particulièrement coûteuse : 3 frappes en l'occurrence, soit autant que pour les lettres avec tréma en français. Le tableau 2 donne le coût moyen du caractère exprimé en nombre de frappes, calculé pour différents corpus : français (673 mille mots), anglais (650 mille mots), arabe voyellé (800 mille mots), et arabe non voyellé (2 millions de mots).

| | Coût moyen du caractère | Proportion des signes diacritiques | Proportion dans le coût de la saisie |
|---|---|---|---|
| Anglais | 1,00001 | 0,0005 | 0,001 |
| Français | 1,003 | 3,51 | 3,84 |
| Arabe non voyellé | 1,037 | - | - |
| Arabe voyellé | 1,46 | 43,7% | 59,9% |

Tableau 2 : Coût moyen du caractère en nombre de frappes

Ces chiffres signifient que la saisie d'un texte de N caractères (lettres avec ou sans signe diacritique) coûtera approximativement N×1,00001 frappes au clavier si le texte est en anglais, contre N×1,003 si le texte est en français, N×1,037 si le texte est en arabe non voyellé, mais N×1,46 si le texte est en arabe voyellé ! Si l'on ajoute que la voyellation d'un texte préalablement saisi ne coûte pas moins, mais autant que de le ressaisir entièrement voyellé (Debili et Fluhr, 2006), alors l'annotation vocalique de l'arabe, sans autre précaution, s'avère prohibitivement coûteuse.

Ces caractérisations sont bien entendu liées à la technologie, aux claviers respectivement associés à chacune des trois langues. Elles offrent une sorte d'évaluation *a posteriori* des standards et normes en vigueur qu'elles sont susceptibles de conforter ou d'infléchir[2]. Mais

[2] En incitant à les amender pour un meilleur rendement. Car sous cet angle, la technologie ne semble pas conférer les mêmes avantages aux langues qu'elle prend en charge. Sur un autre plan, ces comptages et observations suggèrent que les systèmes d'écriture qui persistent ou qui s'installent dans l'usage sont ceux dont le coût est proche de 1, tel que celui de l'anglais, du français, ou de l'arabe non voyellé. On peut remarquer que l'arabe voyellé qui présente un coût de 1,46 le caractère est très peu pratiqué. Même si les raisons qui sous tendent ce constat sont sans doute de nature bien plus complexe, n'y a-t-il pas là un seuil au-delà duquel un système d'écriture n'est plus pratiqué ?

elles permettent aussi, en appréhendant les difficultés que pose la confection massive de corpus annotés sous un angle matériel, d'introduire, aux côtés des métriques d'évaluation des procédures d'analyse automatique classiques, une métrique pour l'évaluation quantitative des processus d'analyse interactive, fondée sur le calcul des coûts qu'engendrent précisément les nécessaires interventions manuelles.

# 4 Evaluation de l'annotation interactive

Un système d'annotation automatique est performant à 100% lorsque ses résultats sont jugés totalement conformes à une annotation manuelle. Ce critère ne vaut évidemment pas pour un système d'annotation interactif, puisque par définition la conformité est ici atteinte à la fin du processus. Un système d'annotation interactif est en fait d'autant plus performant que le nombre de manipulations imposées à l'annotateur pour accomplir une tâche donnée est petit. Lorsque les performances de sa composante automatique, ici, d'étiquetage, de lemmatisation, et de voyellation sont totales, cet objectif est évidemment atteint puisque pour chaque occurrence de mot, les trois propositions – de lemmatisation, d'étiquetage, et de vocalisation – classées en tête de leurs listes respectives s'avèreront systématiquement correctes. Dans ces conditions les manipulations de l'annotateur se réduisent à de simples validations qui ne lui coûtent en nombre d'opérations qu'une seule action (frappe au clavier, clic de souris, pointage sur un écran tactile, etc.). C'est une situation idéale, mais que l'on ne parvient pas atteindre pour toutes les occurrences qui constituent un corpus, les performances des programmes d'analyse automatique étant, comme on le sait, en deçà du 100%. Pour ces occurrences, le coût de l'annotation est d'autant plus élevé que les solutions proposées par la composante automatique se trouvent situées loin dans les listes des voyellations, des étiquettes et des lemmes résiduels, c'est-à-dire des solutions potentielles qui n'ont pu être éliminées. L'opération d'annotation interactive la plus coûteuse advient lorsque la résolution est en queue de liste, ou plus grave, lorsqu'elle ne s'y trouve pas du tout.

Les performances de l'analyse interactive dépendent des performances de l'analyse automatique, mais tandis que dans un cas, elles sont évaluées au nombre ou à la proportion des occurrences qui sont correctement annotées ou non, elles sont évaluées dans l'autre cas au nombre ou à la proportion des interventions manuelles effectives nécessaires pour valider le correct, et corriger l'incorrect. En cela, et quoique corrélées aux extrêmes, ces deux performances sont complémentaires et ne renseignent pas de la même façon. L'évaluation sous l'angle interactif jette en fait un autre regard sur les performances de la composante automatique, et peut conduire, ainsi que nous allons le montrer, jusqu'à suggérer d'en modifier la conception ou le comportement interne, aboutissant ainsi à des spécifications d'analyseurs automatiques différents, selon qu'ils sont destinés à un usage interactif, ou à un usage automatique pur, du moins si leurs performances restent en deçà d'un certain seuil.

Il y a cercle vertueux parce que les actions manuelles, dès lors qu'elles sont prises en compte, modifient à leur tour de façon dynamique les performances de l'analyse automatique. En effet, en éliminant les ambiguïtés là où elles résistent, ces actions améliorent les performances locales des règles automatiquement mises en jeu, et donc les performances globales de l'analyse automatique, laquelle, offrant de meilleurs résultats, diminue d'autant la charge manuelle, améliorant ainsi les performances de la partie interactive, et ainsi de suite.

Mais l'enseignement qu'apportent l'évaluation interactive et son impact sur la définition de l'analyse automatique va plus loin encore. L'on s'aperçoit que l'ordre d'application des règles qui conduit aux meilleures performances automatiques n'est pas forcément celui qui conduit

aux meilleures performances interactives, sauf cas extrême d'une annotation automatique totalement réussie où les deux performances se rejoignent alors. Ce point nous paraît important. Nous ne pointons pas le fait que, ayant bénéficié d'une contribution humaine externe, alors l'analyse automatique produit de meilleurs résultats. Cela est entendu. Nous disons que *les meilleures performances interactives d'un système ne sont paradoxalement pas toujours corrélées à ses meilleures performances automatiques*. Autrement dit, que le comportement linguistique automatique le plus performant n'est pas toujours celui qui assure, dès lors qu'il y a interférence manuelle, le meilleur rendement interactif. Nous décrivons dans le paragraphe suivant le protocole expérimental et les résultats qui ont conduit à ce constat contre intuitif.

# 5 Annotation interactive séquentielle

L'ergonomie interactive qui vient en premier à l'esprit est séquentielle. Elle est liée à la nature des ambiguïtés que nous voulons lever, ici les ambiguïtés que pose la voyellation, la lemmatisation, et l'étiquetage de l'arabe, et au fait que pour lever ces ambiguïtés, le recours au contexte s'impose. De sorte que c'est tout naturellement que l'on s'oriente vers une lecture séquentielle lorsque l'on souhaite établir ou vérifier les annotations d'un texte.

Ayant à accomplir pour chaque occurrence trois choix, – de sa voyellation, de son lemme, de son étiquette, – et dans la mesure où ces choix peuvent interférer, c'est-à-dire influer de façon dynamique sur l'ordre selon lequel sont présentées les solutions des annotations non encore fixées, plusieurs (6 au total) séquences ou protocoles d'intervention peuvent être proposés à l'annotateur, selon que l'on commence par l'un ou l'autre de ces trois choix, et que l'on poursuive ainsi. L'arborescence Figure 1 donne les six cas possibles. A ces six séquences ou protocoles, il convient d'ajouter un septième, celui où les choix resteraient indépendants : pas d'interférence ; on ne retient pas que la résolution de l'une des trois valeurs puisse réduire l'ambiguïté qui porte sur les deux autres, puis, en cascade, que la résolution d'une deuxième puisse réduire l'ambiguïté de la dernière.

Deux protocoles sont *a priori* privilégiés : Etiquetage, puis Lemmatisation, puis Voyellation (séquence ELV, à gauche sur la figure 1), et Voyellation, suivie de Lemmatisation, puis Etiquetage (séquence VLE, à droite). Le premier donne l'ordre selon lequel opèrent les traitements automatiques, la machine donc. Le second donne l'ordre selon lequel opèrent préférentiellement les annotateurs, c'est-à-dire selon lequel les traitements manuels sont effectués. Ces deux protocoles sont privilégiés en vertu de considérations qui sont liées à leurs performances attendues d'une façon générale, et supposées être les meilleures par opposition aux performances des autres protocoles.

Dans le premier cas, les meilleures performances d'analyse automatique attendues semblent pouvoir provenir d'une succession Etiquetage, Lemmatisation, puis Voyellation, à l'image par exemple de ce qui est communément retenu pour le français. En effet, dans une approche modulaire, les règles utiles pour lever ces différents types d'ambiguïtés paraissent pouvoir être plus facilement apprises pour le niveau grammatical, que pour les deux autres niveaux. Ce sont donc en premier les ambiguïtés grammaticales qui sont réduites. Les ambiguïtés lemmatiques et vocaliques, pour lesquelles il semble plus difficile ou plus long de rassembler des règles qui leurs soient propres, peuvent néanmoins bénéficier de ces réductions d'ambiguïtés grammaticales : précisément, en écartant les candidats lemmes et/ou voyellations exclusivement liés aux étiquettes éliminées. Par exemple, l'élimination durant la phase d'étiquetage de l'étiquette *nom* permet de ne plus retenir au compte du mot *élève* que le

lemme *élever*. Le lemme *élève* n'étant que *nom*, il est éliminé en même temps ou suite à l'élimination de l'étiquette *nom*.

Dans le second cas, ce sont les performances globales du processus interactif machine-annotateur que l'on essaie de maximiser. Le facteur humain est ici prépondérant. Quel est le protocole ergonomique qui assure la meilleure efficacité, le meilleur rendement ? Il semble raisonnable de supposer que les annotateurs auront plus de facilités à d'abord Voyeller, Lemmatiser, puis Etiqueter (parcours VLE, à droite), ou à Lemmatiser, puis Voyeller, puis Etiqueter, (parcours LVE, au centre), que de commencer par Etiqueter (parcours de gauche).

Ces parcours interactifs induisent des comportements linguistiques machine différents. Du fait que les règles interagissent entre elles, on ne sait pas *a priori* lequel de ces parcours ou comportements est le plus performant sous l'angle automatique, ni lequel est le plus performant sous l'angle interactif.

Pour mesurer ces performances, nous avons imaginé et mis en œuvre le protocole expérimental simple suivant. Partant d'un corpus préalablement annoté de 145 mille hyper-formes (toutes entièrement voyellées, lemmatisées et étiquetées), nous en avons extrait les fréquences relatives : f(étiquette | Mot Non Voyellé) = Nbre(MNV, étiquette)/Nbre(MNV) ;
f(lemme | mot non voyellé) ; f(voyellation | mot non voyellé) ;
f(lemme | mot non voyellé, étiquette) ; etc., voir légende de la figure 1.

MNV

**Traitements automatiques** — **Traitements interactifs**

| | Etiquette | Lemme | Voyellation |
|---|---|---|---|
| **A** | 76,76% | 93,04% | 84,18% |

| | Lemme | Voyellation | Etiquette | Voyellation | Etiquette | Lemme |
|---|---|---|---|---|---|---|
| **B** | 74,06% | 75,96% | 74,04% | 80,99% | 75,87% | 73,78% |
| **C** | 96,39% | 98,91% | 79,68% | 87,18% | 90,68% | 96,28% |

| | Voyellation | Lemme | Voyellation | Etiquette | Lemme | Etiquette |
|---|---|---|---|---|---|---|
| **D** | 73,88% | 73,87% | 73,86% | 73,77% | 73,78% | 73,77% |
| **E** | 99,73% | 97,19% | 99,73% | 91,48% | 97,19% | 91,48% |
| **F** | **0,43** | 0,43 | 0,41 | **0,36** | 0,37 | 0,37 |
| **G** | **0,21** | 0,21 | 0,20 | **0,17** | 0,18 | 0,18 |

Ligne **A** : Performances automatiques, Application des règles f(E|MNV), f(L|MNV), f(V|MNV).
Ligne **B** : Performances automatiques, Application des règles f(L|MNV, E), f(V|MNV, E), f(E|MNV, L), f(V|MNV, L), f(E|MNV, V), f(L|MNV, V).
Ligne **C** : Performances interactives, Application des règles f(L|MNV, E), f(V|MNV, E), f(E|MNV, L), f(V|MNV, L), f(E|MNV, V), f(L|MNV, V). Ici, dans les conditions | MNV, *y*), *y* est correct.
Ligne **D** : Performances automatiques, Application des règles f(V|MNV, E, L), f(L|MNV, E, V), f(V|MNV, L, E), f(E|MNV, L, V), f(L|MNV, V, E), f(E|MNV, V, L).
Ligne **E** : Performances interactives, Application des règles f(V|MNV, E, L), f(L|MNV, E, V), f(V|MNV, L, E), f(E|MNV, L, V), f(L|MNV, V, E), f(E|MNV, V, L). Ici, dans | MNV, *y*, *z*), *y* et *z* sont corrects.
Ligne **F** : Coût ergonomique des interventions manuelles, annotation séquentielle} *exprimé en nombre moyen*
Ligne **G** : Coût ergonomique des interventions manuelles, annotation parallèle } *de frappes ou clics par mot.*
MNV : hyper-forme non voyellée ; E : étiquette grammaticale ; V : voyellation ; L : lemme

Figure 1 : Performances des analyses automatique et interactive liées aux six séquences possibles d'application des règles et d'interventions manuelles

Utilisant ces fréquences comme autant de règles unaires que nous avons réappliquées en cascade le long des six parcours, nous en avons calculé les performances, et de façon rétrospective les coûts qu'auraient engendrés les interventions manuelles nécessaires pour en corriger les écarts.

La figure 1 liste ces résultats. Les lignes A, B, et D donnent les performances de l'étiquetage, lemmatisation et voyellation automatiques mettant en œuvre les séquences de règles f($x$|MNV), f($x$|MNV, $y$), et f($x$|MNV, $y$, $z$), avec, selon les parcours, $x$, $y$, $z$ = E, L ou V. Les lignes C et E donnent les performances issues de l'application de ces mêmes règles, mais avec $y$ et $z$ manuellement corrigées. La ligne F donne les coûts moyens rapportés au mot, selon les parcours, des diverses interventions manuelles, interventions qui consistent à simplement valider si les choix machine sont corrects (coût nul), et à désigner au moyen de la souris les valeurs potentielles E, L, et V qui conviennent étant données l'occurrence MNV et son contexte, si celles-ci ne sont pas proposées en première position dans leurs listes respectives. Le coût partiel est nul si la valeur E, L, ou V classée première par le système est correcte. Il est sinon d'autant plus élevé que la résolution est classée loin dans la liste des solutions potentielles non éliminées. La formule retenue pour calculer le coût global d'une opération d'annotation interactive est simple :

$$\text{Coût d'annotation séquentielle} = \Sigma_{i=1 \text{ à nombre de mots du corpus}} \Sigma_{x=\text{E, L, ou V}} (\text{rang de la résolution } x_i - 1)$$

Nous constatons que les coûts d'annotation rapportés au mot sont tous différents. Mais surtout que coûts d'annotation interactive et performances d'analyse automatique ne sont pas corrélés, comme l'on aurait pu s'attendre. Le coût le plus bas (42 652 clics ou déplacements de curseur, soit 0,36 clic ou frappe en moyenne par mot, séquence LVE) ne correspond pas au comportement automatique le plus performant (73,88% des mots tous correctement annotés, séquence ELV) qui, lui, réclame 50590 interventions manuelles pour en corriger les écarts, soit 0,43 clic en moyenne par mot. Les séquences d'annotation qui donnent les coûts les plus bas s'avèrent être les séquences qui consistent à commencer par la vérification de la lemmatisation ou voyellation, puis respectivement la lemmatisation ou voyellation, puis étiquetage, tandis que les séquences qui donnent les meilleures performances automatiques s'avèrent être celles qui commencent par l'étiquetage. Les premières correspondent aux trois parcours dessinés à droite sur la figure 1, les secondes, au trois parcours de gauche. Cette distribution spatiale qui partage la figure 1 en deux parties selon les niveaux de performances et de coûts (voir fig. 2), révèle qu'il ne faut pas privilégier un seul comportement automatique, le plus performant en l'occurrence. D'autres comportements moins performants peuvent se révéler meilleurs dès lors qu'il y a interaction. Elle confirme aussi le bien fondé des approches *a priori* préconisées, selon qu'elles sont orientées vers l'autonomie, ou vers l'interactivité.

# 6 Annotation interactive parallèle

Si les hapax sont rares (5 à 12% selon les corpus dont nous disposons), et les proportions des mots qui apparaissent deux fois ou plus dans un corpus, importantes, ne pourrait-on factoriser les annotations, c'est-à-dire voyeller, lemmatiser et étiqueter en même temps toutes les occurrences d'un même mot ? Car outre les gains de productivité attendus, ces conditions pourraient aussi assurer une meilleure reproductibilité dans la mesure où, opérant de façon contrastive (toutes les occurrences en contexte d'un même mot sont visibles en même temps), l'annotateur pourrait en effet décider de façon plus homogène. Ces considérations nous ont amené à dessiner les contours d'une ergonomie d'annotation parallèle dont la figure 1, ligne G, donne, pour le même corpus, les performances calculées de façon rétrospective en se fondant sur les mêmes conventions de coût.

La comparaison des lignes F et G révèle un gain de facteur 2 : l'annotation parallèle coûte approximativement deux fois moins cher que l'annotation séquentielle. Mais l'on constate surtout que les observations que nous avons pu faire plus haut restent vraies. Avec une acuité légèrement accrue, nous remarquons en effet que l'annotation parallèle la moins coûteuse n'est pas corrélée au traitement automatique le plus performant. Et que la partition droite gauche observée plus haut, selon les niveaux de performance ou de coût, est confortée.

Dans cette ergonomie, l'annotation interactive n'est plus appliquée à toutes les occurrences du corpus prises une à une, mais aux seules entrées du lexique qui leur correspond. Dans le cas présent, aux seules 24291 différentes hyper-formes non voyellées qui constituent le lexique du corpus, et non aux 117900 occurrences reconnues de ce corpus, même si de fait, il y a bien prise en compte de 117900 contextes potentiellement tous différents. L'annotation retient pour les 24291 entrées non voyellées, 38108 descriptions morpho-grammaticales différentes, sur 334179 descriptions potentielles, c'est-à-dire qu'elle donne lieu à 38108 hyper-formes dûment voyellées, lemmatisées et étiquetées différentes.

# 7 Conclusion

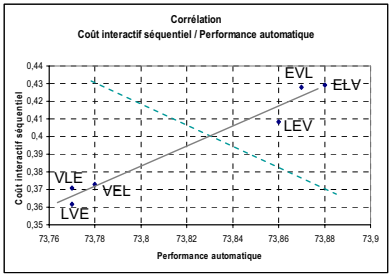


Figure 2.

Ce graphe, qui reprend les résultats lignes D et F de la figure 1, indique que, dans une plage de performances donnée, *le meilleur comportement automatique n'est pas celui qui assure toujours, dès lors qu'il y a intervention manuelle, le meilleur comportement interactif.* Dit autrement, et en soulignant le caractère local de nos observations, nous découvrons en effet que *le comportement autonome le plus performant n'est pas toujours celui qui garantit, dès lors qu'il y a interaction, le comportement coopératif le plus performant.* C'est ce résultat *empirique* qui ne laisse de surprendre – la corrélation *Coût-Performance* attendue était et reste en effet qu'à performance automatique meilleure corresponde coût interactif moindre, et que les points dessinent la courbe de tendance en pointillée, et non celle observée trait continu – qui est devenu, chemin faisant, prépondérant, au-delà de la conception et réalisation d'un système d'analyse morpho-grammaticale interactive de l'arabe et de sa mise en œuvre pour la confection de corpus voyellés, étiquetés, et lemmatisés. Sur le plan méthodologique, il remet en cause les stratégies communément préconisées pour la confection massive de données linguistiques annotées, où l'idée est de corriger les sorties d'analyse automatique au moyen d'éditeurs dédiés, en considérant *a priori* que le rendement optimal est atteint dès lors que l'analyseur automatique qui est mis en œuvre est le plus performant. Nous pressentons, sans l'avoir encore constaté, que cet *a priori* n'est vrai qu'au-delà d'un certain seuil de performance automatique, seuil qu'il conviendrait de déterminer. En deçà de ce seuil critique, nous assisterions à des comportements « erratiques » où précisément, ainsi que nous l'avons observé, la corrélation *meilleure performance automatique* alors *meilleure performance interactive* n'est pas maintenue. Au-delà, au contraire, la corrélation est ou serait rétablie.

Mais il nous semble que les conséquences de ce constat vont plus loin encore. S'il devait être confirmé par d'autres expérimentations menées par nous ou par d'autres, sur d'autres langues et/ou d'autres types de règles, alors nous serions fondés à dire que nos objectifs devraient non plus se focaliser sur la seule dimension automatique, ainsi que nous disions au début de notre introduction, mais devraient aussi, d'emblée, prendre en compte le développement de la

nécessaire dimension interactive et de ce que celle-ci induit dans le développement et l'évaluation de la dimension automatique. Car l'on s'aperçoit qu'introduire parallèlement une dimension interactive plus dynamique, loin de réduire la surface de la composante automatique ou de ce que l'on peut en exiger, conduit au contraire à en multiplier les comportements linguistiques et à en étendre les potentialités, tout en en révélant les insuffisances critiques. L'interactif ramène ainsi à l'automatique.

Sous l'angle de l'évaluation, l'interactif conduit, comme pour les applications qui mettent en œuvre différents composants linguistiques, et où l'on distingue, (cf. par ex. Berthelin et al. 2001), les performances intrinsèques de ces composants d'une part, et les performances globales de ces mêmes composants interagissant ensemble, à une caractérisation tierce de ces composants. Mais là s'arrête l'analogie. Les métriques restent en effet les mêmes dans le premier cas. Qu'il s'agisse de performances locales et directes, ou globales et indirectes, l'on essaie de compter les écarts, les erreurs, les silences. Alors qu'elles sont différentes lorsqu'il s'agit de mesurer les performances de la dimension interactive, ainsi que nous avons essayé de le montrer. L'on rejoint ici l'une des multiples « *dimensions* » de l'évaluation recensées par Chaudiron, en l'occurrence, la notion d'*efficience* vue, pour une tâche donnée, « *comme la possibilité pour un utilisateur d'accomplir cette tâche à moindre coût en terme de charge de travail et d'effort cognitif* » (Chaudiron, 2001, p. 100). Corrélée à l'évaluation de l'analyse automatique, l'évaluation de l'annotation interactive souligne au final qu'il est certes crucial de parfaire les performances de l'automatique, mais qu'il est aussi utile d'octroyer à celui-ci non plus un, mais différents comportements pour être à même de s'adapter de façon optimale à la variabilité des comportements des annotateurs, sous peine de ne pas être retenu.

## Remerciements

Le présent travail a été initié dans le cadre du projet EurADic (Action Technolangue du Ministère de la recherche), et se poursuit dans le cadre du projet MUSCLE (6ème PCRD).
A J.-B. Berthelin, pour la traduction du résumé, et sa disponibilité à aborder ces thématiques.

## Références

ABEILLE A., CLEMENT L. (2003). *Annotation morpho-syntaxique. Les mots simples – Les mots composés. Corpus Le Monde*. Technical report, Paris 7.

ADDA, G., MARIANI, J., PAROUBEK, P., RAJMAN, M., & LECOMTE, J. (1999). L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, 2(1).

BERTHELIN J.-B. (2001). Two levels of evaluation in a complex NL system. Actes d'*ACL'2001*, *Toulouse*.

CHAUDIRON S. (2001). *L'Évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigme*. Habilitation à diriger des recherches, Paris X, Nov. 2001.

DEBILI F., ACHOUR H., SOUISSI E. (2002). La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique. *Correspondances N°71, IRMC,* Tunis, 10-26.

DEBILI F., FLUHR C. (2006). Confection de ressources dictionnairiques et textuelles multilingues. Actes de *TALN'2006, Louvain, Belgique, 10-13 Avril 2006*, 910-917.

DEBILI F., SOUISSI E. (2005). Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage grammatical ? Actes de *TALN'2005, Dourdan, Juin 2005*, 363-372.

HABERT B. (2005). *Instruments et ressources électroniques pour le français*. Paris : Editions Ophrys.

TESNIERE L. (1969). *Eléments de syntaxe structurale*. Paris : Editions Klincksieck.

VERONIS J. (1999). *Guide d'étiquetage Multitag*. Version 3.1, 6 novembre 1999.