Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage grammatical ?

Fathi Debili (1), Emna Souissi (2)

(1) CNRS – ICAR – ENS LSH
15, Parvis René Descartes
69342 Lyon Cedex 07
France
fathi.debili@wanadoo.fr

(2) ISG – Université de Sousse
BP 763 Sousse 4000 - Tunisie
emna.souissi@isgs.rnu.tn

Mots-clés : Etiquetage grammatical, règle de succession, taille des règles, chaînage de règles, règle attestée, règle simulée, discriminance, couverture, évaluation en usage *vs* évaluation en définition d'un ensemble de règles.

Keywords: Part-of-speech tagging, tag sequences, rule length, rule composition, attested rule, simulated rule, evaluation of generation *vs* evaluation of analysis.

Résumé

La quasi-totalité des étiqueteurs grammaticaux mettent en œuvre des règles qui portent sur les successions ou collocations permises de deux ou trois catégories grammaticales. Leurs performances s'établissent à hauteur de 96% de mots correctement étiquetés, et à moins de 57% de phrases correctement étiquetées. Ces règles binaires et ternaires ne représentent qu'une fraction du total des règles de succession que l'on peut extraire à partir des phrases d'un corpus d'apprentissage, alors même que la majeure partie des phrases (plus de 98% d'entre elles) ont une taille supérieure à 3 mots. Cela signifie que la plupart des phrases sont analysées au moyen de règles reconstituées ou simulées à partir de règles plus courtes, ternaires en l'occurrence dans le meilleur des cas. Nous montrons que ces règles simulées sont majoritairement agrammaticales, et que l'avantage inférentiel qu'apporte le chaînage de règles courtes pour parer au manque d'apprentissage, plus marqué pour les règles plus longues, est largement neutralisé par la permissivité de ce processus dont toutes sortes de poids, scores ou probabilités ne réussissent pas à en hiérarchiser la production afin d'y distinguer le grammatical de l'agrammatical. Force est donc de reconsidérer les règles de taille supérieure à 3, lesquelles, il y a une trentaine d'années, avaient été d'emblée écartées pour des raisons essentiellement liées à la puissance des machines d'alors, et à l'insuffisance des corpus d'apprentissage. Mais si l'on admet qu'il faille désormais étendre la taille des règles de succession, la question se pose de savoir jusqu'à quelle limite, et pour quel bénéfice. Car l'on ne saurait non plus plaider pour une portée des règles aussi longue que les plus longues phrases auxquelles elles sont susceptibles d'être appliquées. Autrement dit, y a-t-il une taille optimale des règles qui soit suffisamment petite pour que leur apprentissage puisse converger, mais suffisamment longue pour que tout chaînage de telles règles

pour embrasser les phrases de taille supérieure soit grammatical. La conséquence heureuse étant que poids, scores et probabilités ne seraient plus invoqués que pour choisir entre successions d'étiquettes toutes également grammaticales, et non pour éliminer en outre les successions agrammaticales.
Cette taille semble exister. Nous montrons qu'au moyen d'algorithmes relativement simples l'on peut assez précisément la déterminer. Qu'elle se situe, compte tenu de nos corpus, aux alentours de 12 pour le français, de 10 pour l'arabe, et de 10 pour l'anglais. Qu'elle est donc en particulier inférieure à la taille moyenne des phrases, quelle que soit la langue considérée.

Abstract

Is there an optimal n for n-grams used in part-of-speech tagging?
Almost all part-of-speech taggers apply rules about permitted successions and collocations of two or three grammatical categories. Their performance amounts to 96 percent of correctly tagged words, and to less than 57 percent of correctly tagged sentences. These binary and ternary succession rules represent a small fraction of succession rules one can extract from sentences in a learning corpus, where most sentences (more than 98 percent of them) have a length of more than three words. In other words, most sentences are processed by rules that are reconstructed, or simulated, from shorter ones, here ternary at best. We show that most such simulated rules are agrammatical, and that, if some inferential benefit comes from the chaining of short rules to compensate inexistent learning, mainly in the case of long rules, this benefit is nullified by the permissive behaviour of this process, in which a variety of weights, scores or probability are ineffective in hierarchizing its production and yield a separation between grammatical and agrammatical rules. So we feel forced to look again at larger-than-ternary rules. However, if we admit a necessity of enlarging succession rules, we must ask the question "up to which limit, and for what profit". For we also decline to argue for rules as long as the longest sentences upon which they might apply. So the real question is, can we define an optimal size for rules, short enough for learning to converge, and long enough for any chaining of rules to deal with larger sentences to be grammatical? A positive result would be that weights, scores or probability would then be invoked only to decide between equally grammatical successions of tags, and no longer to eliminate agrammatical ones.
This optimal size apparently exists. We show that the use of rather simple algorithms leads to its determination. And its value, according to our corpora, is near 12 for French, 10 for Arabic and 10 for English. Therefore, it is less than the average length of sentences, for each of these three languages.

1 Introduction

Cet article rend compte d'une étude critique des règles les plus couramment mises en œuvre dans les étiqueteurs grammaticaux. Il souligne de façon quantitative le caractère infondé de l'emploi généralisé, du moins dans la plupart des modèles probabilistes, des seules règles de succession (ou de précédence) d'ordre deux ou trois que renferment les expressions : *P(étiquette de rang i | étiquette de rang i-1) ; P(étiquette de rang i | étiquette de rang i-2, étiquette de rang i-1).*

La sanction, s'il en est, de cet état de fait est bien connue : un niveau de performance – 96% de mots correctement étiquetés – rapidement atteint (par exemple Debili, 1977 ; DeRose, 1988), mais que l'on peine à dépasser de façon substantielle et reproductible, en dépit d'efforts qui ne se sont point relâchés depuis plus de trente ans (par exemple Andreewsky et Fluhr, 1973 ; Mérialdo, 1994 ; Leech et al., 1994 ; Adda et al., 1999 ; Valli et Véronis, 1999 ; Nasr et Volanschi, 2004). Ce niveau de performance est relativement faible. Il signifie que si la taille moyenne de la phrase est de 20 mots, alors en moyenne 4 phrases sur 5 sont mal étiquetées. Dans la pratique, la situation est sans doute meilleure, mais demeurant à des niveaux qui restent très faibles, les performances au niveau de la phrase sont rarement affichées. 57% de phrases correctement étiquetées est semble-t-il le meilleur résultat publié que l'on ait pu atteindre (Toutanova et al., 2003). Erigés en barrière depuis une trentaine d'années, ces niveaux de performances, ajoutés aux difficultés réelles que pose leur évaluation comparative, ont pu amener jusqu'à douter du statut de l'étiquetage tel que préconisé (Fairon et Sennelart, 1999).

Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage grammatical ?

Nous voulons ici revenir sur une hypothèse non toujours formulée liée à la portée des règles mises en œuvre dans les étiqueteurs grammaticaux. Leur extrême localité – examen de contextes très proches portant sur les deux positions qui précèdent, entourent ou suivent l'ambiguïté étudiée – est souvent pointée pour expliquer les échecs constatés, mais elle est également souvent assumée au nom de contraintes techniques liées autant à la puissance des machines qu'à la taille des corpus d'apprentissage, et de fait en partie justifiée sur le plan expérimental, puisque aussi bien l'on a pu en effet enregistrer des résultats meilleurs avec des règles de portées paradoxalement plus courtes.

Il ne s'agit pas évidemment de rejeter ces règles qui restent utiles. Il s'agit de s'attaquer au problème que leur chaînage soulève dès lors que l'on essaie de les appliquer à des phrases dont la taille est supérieure à 2 ou à 3 mots, ce qui arrive dans plus de 98% des cas. Le schéma suivant où les m_i sont les mots de la phrase, et les t_{ij}, leurs diverses étiquettes grammaticales potentielles respectives, rappelle ce qui se passe :

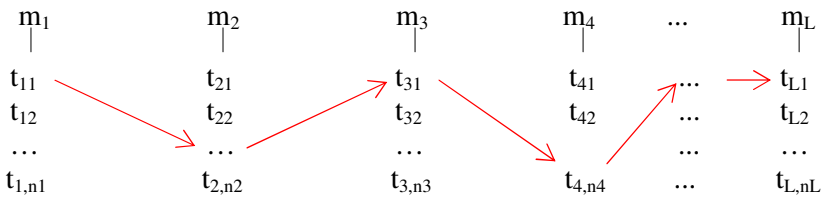

Figure 1

L'étiquetage consiste à reconnaître parmi les N^L chemins combinatoirement possibles (N étant le nombre moyen d'étiquettes pour un mot, et L le nombre de mots de la phrase), le chemin qui permet d'attribuer à chacun des mots l'étiquette grammaticale qui lui sied dans la phrase. La situation est plus complexe en fait car la segmentation en mots n'est pas toujours unique. Mais nous ne nous en préoccuperons pas ici, car cela n'a pas d'incidence sur ce que nous essayons de montrer. Nous ne nous préoccuperons pas non plus de la nature des étiquettes que nous recevrons comme telles.

Appliquer des règles *binaires* pour construire ces chemins, c'est joindre bout à bout des successions *attestées* (c'est-à-dire qui ont été relevées dans un corpus d'apprentissage) de 2 étiquettes, de façon telle que la seconde étiquette de la première règle soit identique à la première de la seconde règle. Ainsi, si *a*, *b*, *c*, …, *z*, sont des étiquettes grammaticales, et (*a b*), (*b c*) des successions attestées, alors (*a b c*) est une succession d'ordre 3 potentiellement valide ou, dirons-nous, potentiellement *attestable*.

Appliquer des règles *ternaires*, c'est chaîner de la même façon des successions attestées de trois étiquettes de façon telle que les deux dernières du premier triplet soient identiques aux deux premières du second triplet. Ainsi, si (*a b c*) et (*b c d*) sont des successions ternaires attestées, alors on peut former la succession quaternaire (*a b c d*). Et de proche en proche (*a b c d e*) si (*c d e*) est une succession attestée, et ainsi de suite. Nous remarquons qu'il y a là production de successions qui à leur tour peuvent être assimilées à des règles d'ordre supérieur. La différence est que les premières sont attestées, alors que les secondes ne le sont pas. Pour les distinguer, nous les appellerons *simulées* dans la suite de l'exposé.

C'est ainsi qu'opèrent basiquement la plupart des étiqueteurs qui essaient d'établir ou de reconnaître des continuités syntaxiques. Le problème réside dans le fait que les successions ainsi obtenues et que l'on peut en effet assimiler à des règles de succession de taille aussi longue que les phrases auxquelles elles s'appliquent, ne se révèlent pas toujours, loin s'en faut, attestables, bien qu'établies à partir de règles binaires ou ternaires attestées. L'on s'aperçoit que ces successions ou règles simulées sont au contraire souvent agrammaticales. Cela est bien connu. Mais à notre connaissance aucune étude à caractère quantitatif n'a été menée pour mesurer la proportion du potentiellement attestable ou grammatical par opposition à ce qui demeurera non attestable ou agrammatical.

La question qui surgit est alors : y a-t-il corrélation entre ces proportions grammaticales *vs* agrammaticales de règles simulées et la taille des règles attestées leur ayant donné naissance ? Sans doute oui, mais là aussi nous n'en connaissons pas la nature. L'on imagine toutefois que la proportion des règles simulées agrammaticales devrait aller en diminuant à mesure que la taille des règles

attestées génératrices irait croissant. Si tel est le cas, quelle est la valeur minimale de cette taille, taille à partir de laquelle l'on n'obtiendrait plus que des successions simulées attestables ? Ne pourrait-on pas élaborer un protocole expérimental pour déterminer de meilleure façon, empirique et non plus apriorique, la taille optimale des règles de succession devant intervenir dans l'étiquetage grammatical ?

C'est à ces questions que nous allons essayer de répondre à partir d'observations et de calculs effectués sur trois corpus étiquetés : • français (MULTITAG du CNRS-LIMSI, un ensemble de textes du *Monde* d'environ 754 000 mots regroupés en 31 points et utilisant 337 étiquettes grammaticales, Paroubek et Rajman, 2000) ; • arabe (un ensemble de 53 textes du *Monde Diplomatique* d'environ 92 000 mots manuellement voyellés et étiquetés au CNRS-ICAR, utilisant 558 étiquettes) ; • et anglais (SUZANNE Corpus, 64 textes d'environ 149 000 mots, utilisant 308 étiquettes).

2 Notations

Ayant à considérer pour les besoins de nos expérimentations différentes tailles de corpus ou de règles :

- $T_{1...x}$: désignera le corpus constitué de l'ensemble de textes $\{T_1, T_2, \ldots, T_x\}$
- $RA_p(T_{1...x})$ ou plus simplement $RA_p(T)$: désignera l'ensemble des règles d'ordre *p* attestées ou apprises à partir du corpus $T_{1...x}$, ou plus simplement T.
- $RS_q[RA_p(T)]$: désignera l'ensemble des règles simulées d'ordre *q* engendrées à partir des règles attestées d'ordre *p* ($p < q$) du corpus T.

Rappelons qu'une règle simulée d'ordre *q* est obtenue en chaînant $(q - p + 1)$ règles attestées d'ordre *p*. Exemple avec $q = 7$ et $p = 5$:

		1	2	3	4	5	6	7
Règle attestée	1	*a*	*b*	*c*	*d*	*e*		
Règle attestée	2		*b*	*c*	*d*	*e*	*f*	
Règle attestée	3			*c*	*d*	*e*	*f*	*g*
Règle simulée		*a*	*b*	*c*	*d*	*e*	*f*	*g*

Figure 2 : Règle simulée d'ordre 7 engendrée au moyen de règles attestées d'ordre 5.

3 Règles de succession attestées

Les règles attestées sont les successions d'étiquettes grammaticales de diverses longueurs que l'on peut extraire des phrases prises une à une dans des textes préalablement étiquetés. Une phrase de trois mots respectivement étiquetés (*a b c*) donnera lieu aux six successions suivantes : • unaires : (*a*), (*b*), (*c*) ; • binaires : (*a b*), (*b c*) ; • et ternaire : (*a b c*). Plus généralement, une phrase donnée de longueur L donnera naissance à L(L + 1)/2 successions différentes, de taille allant de une à L étiquettes.

La figure 3 donne les effectifs associés aux trois corpus français, arabe et anglais. Dans les cas d'espèce, si l'on s'intéresse aux règles binaires et ternaires qui sont les seules mises en œuvre dans la quasi totalité des étiqueteurs probabilistes, ces histogrammes révèlent qu'en fait celles-ci cumulées ne représentent au plus que 3,57% du total des règles de succession potentielles que l'on peut extraire d'un texte dûment étiqueté, cas de l'anglais. Et moins d'un pour cent (0,75%) lorsque la taille du corpus est plus importante, cas du français, l'arabe étant à 3,10%. Plus de 96%, voire 99%, des règles potentiellement utiles pour une analyse plus discriminante de ces mêmes corpus ou d'autres ne sont pas retenues, tandis que plus de 98% des phrases ont d'une façon générale plus de trois mots. Ces proportions soulignent que dans une perspective d'analyse, plus de 98% des phrases sont étiquetées au moyen de moins de 4%, voire moins d'un pour cent, du total des règles que l'on peut extraire des corpus. Autrement dit, 98% des phrases ou davantage sont étiquetées aux moyen de règles simulées.

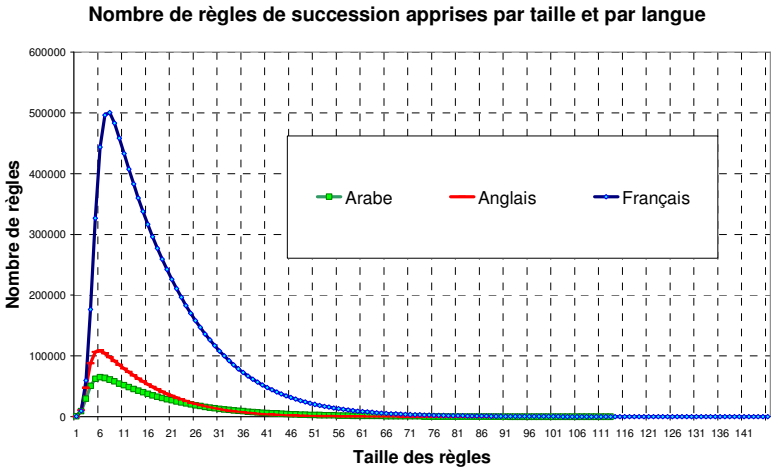

Figure 3 : Histogramme des successions attestées selon leur taille

Ce constat ne signifie pas bien entendu qu'il faille opter d'emblée pour l'apprentissage de toutes les règles de succession, y compris de celles qui seraient aussi longues que les plus longs tronçons ou phrases dont elles sont issues. Quelle valeur aurait une règle hapax qui aurait toutes les chances de ne pouvoir jamais être appliquée qu'au fragment ou à la phrase d'où elle provient ? L'idée n'est donc pas de sauvegarder toutes les règles quelle que soit leur taille, d'autant qu'il s'avère que celles-ci sont toutes précisément systématiquement hapax dès lors que leur taille dépasse un certain seuil, ici respectivement pour les trois corpus : 16 pour l'anglais, 19 pour l'arabe, et 60 pour le français.

Cependant, dans la perspective qui est la nôtre d'argumenter en faveur d'une extension de la taille des règles apprises, ces valeurs peuvent être considérées dès maintenant comme des limites qu'il ne sera pas utile de dépasser. Nous verrons, moyennant d'autres observations, qu'il ne sera pas non plus utile d'aller aussi loin.

Il nous faut auparavant argumenter davantage et en particulier tenter d'évaluer, – en amont des performances d'étiquetage déjà signalées, et que nous qualifierons d'extrinsèques (ou exogènes), dans la mesure où elles renseignent sur la couverture et la discriminance *en usage* des règles binaires et ternaires, – de ces mêmes propriétés mais *en définition*. C'est-à-dire de la propension combinatoire d'un ensemble de règles à n'engendrer intrinsèquement (ou de façon endogène) par chaînage, que des règles simulées attestables, ou en tout cas à minimiser la proportion du généré non attestable. A l'inverse de l'évaluation *en usage*, qui sans doute au final seule compte, l'évaluation *en définition* d'un ensemble de règles a une valeur prédictive. En introduisant des critères et des indices permettant de comparer des ensembles de règles différents sous l'angle de leur fonctionnement interne et en dehors de toute confrontation aux textes à étiqueter, on se dote de moyens permettant d'anticiper précisément de la qualité de l'étiquetage. L'on imagine en effet que de deux ensembles de règles engendrant les mêmes règles simulées attestables par ailleurs, celui qui produirait proportionnellement le moins de règles agrammaticales concomitamment, devrait conduire incontestablement à de meilleures performances *en usage*, c'est-à-dire de étiquetage proprement dit.

4 Règles de succession simulées

L'argument souvent avancé en faveur des règles binaires ou ternaires est lié à leur capacité inférentielle ou « générative » et à la rapidité de leur apprentissage. Les règles de plus longue taille sont par opposition moins productives, et l'on met plus de temps à les acquérir, temps signifiant ici volume des corpus nécessaires à leur apprentissage.

Dans la perspective d'une extension de la taille des règles apprises, ces avantages subsistent, puisqu'il ne s'agit nullement de se départir des règles binaires et ternaires, mais seulement de leur adjoindre les règles de plus longue taille. L'on peut donc s'interroger sur l'utilité de cette extension, d'autant que ces règles de plus longue taille peuvent être totalement simulées à partir des règles binaires ou ternaires, et qu'en outre, cette simulation produira même potentiellement plus encore de règles attestables de longue taille que nous ne pourrions en extraire directement à partir des mêmes corpus d'apprentissage. La figure suivante illustre ces imbrications.

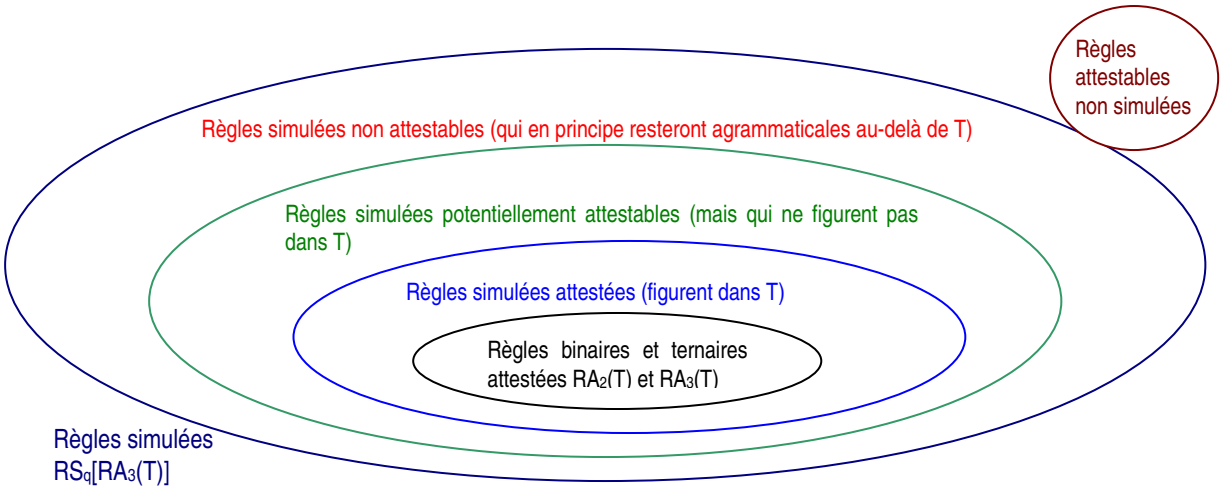

Figure 4 : Imbrication des ensembles de règles simulées $RS_q[RA_3(T)]$ avec $3<q\leq L$

Or, ces capacités productives des règles binaires ou ternaires n'ont à notre connaissance jamais été évaluées autrement qu'au travers de l'utilisation qui en est faite, c'est-à-dire des performances des outils d'étiquetage mettant en œuvre ces règles. Une évaluation en usage donc, qui au demeurant reste à des niveaux peu satisfaisants comme nous avons pu le voir, et non en définition. En particulier, nous n'avons aucune connaissance des proportions relatives des diverses règles simulées : – attestées, – potentiellement attestables, – et non attestables, rapportées au total des règles simulées. Ni *a fortiori* de l'évolution des ces mêmes proportions en fonction de la taille des règles ou des corpus d'apprentissage. Ces proportions et leurs évolutions respectives semblent pourtant être de nature à nous renseigner de façon apriorique sur la couverture et la discriminance d'un ensemble de règles par opposition à un autre. Idéalement, à couverture ou à capacité d'engendrement équivalente, celui des deux ensembles qui produira en proportion le plus de règles attestables sera considéré plus prometteur, car, surgénérant moins, il est plus discriminant. Dans la réalité, concilier couverture et discriminance s'avère contradictoire. C'est pourquoi la comparaison reste difficile, et qu'il nous faudra rechercher non le meilleur, mais un optimum qui ne correspondra qu'à un meilleur local.

Problème : comment mesurer ces proportions et leurs évolutions ?

5 Discriminance, couverture, évaluation d'un système de règles

Nous proposons d'appeler *discriminance en génération* de rang q d'un ensemble de règles attestées d'ordre p, $q>p$, le rapport :

$$\delta_{q,p} = \text{Card}\ \{\ RS_q[RA_p(T)] \cap RA_q(T)\ \} \ /\ \text{Card}\ \{\ RS_q[RA_p(T)]\ \}$$

Ce qui simplement correspond à la proportion des règles simulées d'ordre q, et attestées par le corpus T, rapportée au total des règles simulées d'ordre q, les règles simulées étant engendrées à partir des règles d'ordres p issues du même corpus T.

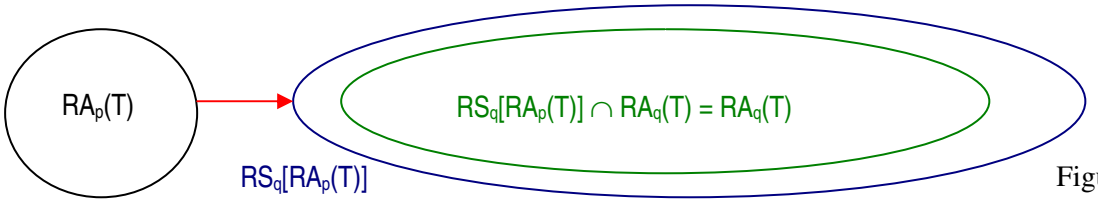

Figure 5a

Nous dirons que cette discriminance est *intrinsèque*, et nous désignerons par :

$$\Delta_{q,p} = \text{Card}\ \{\ RS_q[RA_p(T_a)] \cap RA_q(T_{1\ldots x})\ \} \ /\ \text{Card}\ \{\ RS_q[RA_p(T_a)]\ \} \qquad \text{avec } T_a \subset T_{1\ldots x}$$

la discriminance *extrinsèque*, laquelle correspond à la proportion des règles simulées d'ordre q, et attestées par le corpus $T_{1...x}$, rapportée au total des règles simulées d'ordre q, les règles simulées étant engendrées à partir des règles d'ordres p issues du corpus T_a, T_a pouvant être inclus, ou éventuellement, partiellement inclus, ou non inclus dans $T_{1...x}$.

Figure 5b : Intersection des règles simulées $RS_q[RA_p(T_a)]$ avec les règles attestées $RA_q(T_{1...x})$

De même, nous proposons d'appeler *couverture en génération* de rang q d'un ensemble de règles attestées d'ordre p, $q>p$, le rapport :

$$\kappa_{q,p} = \text{Card } \{ RS_q[RA_p(T_a)] \cap RA_q(T_{1...x}) \} \ / \ \text{Card } \{ RA_q(T_{1...x}) \}$$

Cela correspond à la proportion des règles simulées d'ordre q, engendrées à partir des règles d'ordres p issues du corpus T_a, et attestées par le corpus $T_{1...x}$, rapportée au total des règles attestées d'ordre q de ce même corpus $T_{1...x}$. Nous remarquons que dans le cas particulier où T_a est identique à $T_{1...x}$, alors $\kappa_{q,p} = 1$.

D'une façon plus générale et par extension, nous appellerons *discriminance en génération* d'un ensemble de règles attestées de diverses longueurs issues d'un corpus d'apprentissage T_a, la proportion des règles engendrées à partir de ces règles, et attestées par le sur corpus $T_{1...x} \supset T_a$, rapportée au total des règles engendrées ; et *couverture en génération* d'un ensemble de règles attestées de diverses longueurs issues du sous corpus $T_a \subset T_{1...x}$, la proportion des règles engendrées à partir de ces règles, et attestées par $T_{1...x}$, rapportée au total des règles attestées de $T_{1...x}$.

Les protocoles expérimentaux permettant de calculer les rapports $\delta_{q,p}$, $\Delta_{q,p}$ et $\kappa_{q,p}$ sont relativement simples. Ils exigent néanmoins des temps de calcul relativement long, d'autant plus long que p est petit, et $(q-p)$ grand. C'est pourquoi nous ne pouvons donner ici, en particulier pour p=3, que les évolutions liées à de faibles écarts, mais que nous continuons d'élargir. Les tendances que nous voulons faire valoir à l'appui d'une extension des règles de succession d'une part, et pour la détermination d'une limite à cette extension d'autre part, sont cependant présentes.

En faisant varier les paramètres p, q, et x (taille du corpus), plusieurs familles de courbes peuvent être tracées. Ces courbes donnent les évolutions des discriminances (intrinsèque $\delta_{q,p}$ et extrinsèque $\Delta_{q,p}$), et de la couverture ($\kappa_{q,p}$) d'un ensemble de règles, soit en fonction de la taille des corpus (x), soit en fonction de la taille des règles génératrices (p), soit en fonction de la taille des règles engendrées (q).

L'observation de l'évolution en interne d'un système de règles au travers de ces diverses courbes et caractérisations recèle ce qui pourrait permettre de comparer entre eux des systèmes de règles différents, et dès lors de procéder à une certaine forme d'évaluation de ces systèmes de règles.

6 Discriminance en génération

Nous focaliserons notre attention ici sur l'évolution des indices de discriminance en génération d'ensembles de règles en fonction de p et de q. Et sur deux constats pressentis, mais désormais chiffrés : 1°) la décroissance, d'autant plus rapide que p est petit, de la discriminance, aussi bien intrinsèque qu'extrinsèque, à mesure que q croit ; 2°) et à l'inverse, la croissance rapide de cette même discriminance à mesure que p croit.

Les courbes ou tableaux $\delta_{q,p}$ et $\Delta_{q,p}$ (voir figure 6 pour le français) montrent qu'en passant simplement de q=4 à q=5 pour p=3, alors la proportion des règles attestées passe de 13,35% à 1,30% pour le

français, de 11,08% à 1,11% pour l'arabe, et de 10,73% à 0,8% pour l'anglais. Et que si l'on tient compte en outre des règles potentiellement attestables, alors l'on passe de 50,13% à 14,78% pour le français, de 50,49% à 18,71% pour l'arabe, et de 56,09% à 18,92% pour l'anglais.

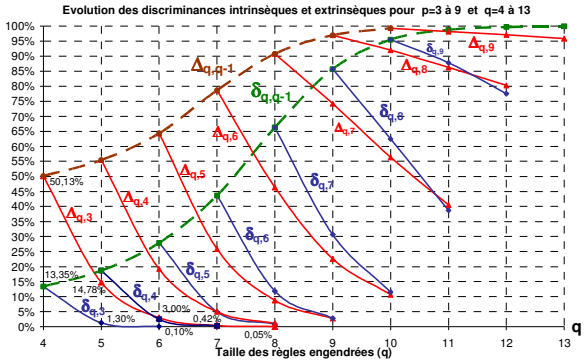

Figure 6 : La discriminance décroît à mesure que *q* croît ($\delta_{q,p}$ et $\Delta_{q,p}$ du français)

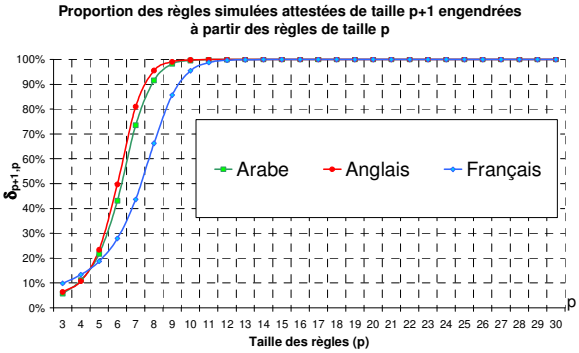

Figure 7 : La discriminance croît à mesure que *p* croît ($\delta_{p+1,p}$ de l'arabe, de l'anglais et du français)

Une décroissance qui signifie que, dans le meilleur des cas, si pour étiqueter une phrase de quatre mots l'on peut encore compter sur 50,13% de règles attestées ou potentiellement attestables, pour étiqueter une phrase de cinq mots l'on ne peut plus compter que sur 14,78%. Et beaucoup moins si l'on passe à l'étiquetage des phrases de six mots (3%) ou de sept (0,42%), et moins encore au-delà ($\Delta_{8,3}$ = 0,05% !).

Si maintenant nous observons l'évolution de la discriminance en fonction de *p*, (voir figure 7), nous constatons que ce rapport tend rapidement vers 1 à mesure que *p* augmente. Ce qui suggère que nous puissions limiter l'apprentissage aux seules règles de taille inférieure ou égale à une certaine valeur de *p*, valeur au-delà de laquelle les règles simulées se révèlent pour la plupart attestées. La figure 7 donne l'évolution de la proportion des règles simulées attestées rapportées au total des règles engendrées à partir des règles attestées d'ordre immédiatement inférieur, c'est-à-dire de $\delta_{p+1,p}$.

Nous observons que plus de 99% des règles simulées se révèlent attestées à partir de p = 10 pour l'anglais et l'arabe, et à partir de p=12 pour le français (corpus 5 à 7 fois plus important que celui de l'anglais ou de l'arabe). Moins d'un pour cent des règles simulées restent non attestées, sans que l'on puisse affirmer qu'elles sont toutes non attestables.

Nous retrouvons là, de façon empirique, ce que nous savions déjà : plus le contexte avant (ou après) est long, plus le choix sur ce qui suit (ou précède) est contraint. Ce que nous ne savions pas, pour ce qui est de l'étiquetage, c'est la taille de ce contexte, taille minimale à partir de laquelle nous constatons que le choix sur ce qui suit devient si contraint qu'il laisse peu de place à la production de successions qui ne soient pas très probablement attestables.

Nous n'avons pas fini d'exploiter les potentialités qu'offrent ces calculs. Nous pensons en particulier que les évolutions croisées de certaines courbes (discriminance et couverture en fonction de la taille des corpus), que nous n'avons pas pu mentionner ici, pourraient renseigner sur l'état de convergence de l'apprentissage, indépendamment de toute application des règles à l'étiquetage d'un texte nouveau, ainsi qu'il est traditionnellement fait.

7 Couverture en génération d'ensembles de règles de succession

Si la discriminance est à l'avantage des règles de succession longues, la couverture est à l'avantage des règles courtes. Les courbes de la figure 8 le rappellent, même si elles sont incomplètes. Associées aux courbes des figures 6 et 7, elles matérialisent ce que l'on peut intuitivement pressentir : à savoir que plus une règle ou un ensemble de règles est inférent ou productif, moins il est discriminant, et inversement. Mais par leurs formes ces courbes révèlent en même temps que l'avantage inférentiel doit être relativisé, et combien au final la taille des corpus doit être plus importante encore. L'on

constate en effet que si à *q* constant, la couverture est bien meilleure pour *p*=3 que pour *p*=4 ou plus (courbes en rouge), cet avantage ne semble se réaliser de façon notable que pour des valeurs de *p* petites (inférieures à 5), mais surtout ne se maintenir que pour des valeurs de *q*, ou plus exactement des différences (*q*-*p*) relativement faibles, n'excédant pas quatre en l'occurrence. Au-delà, pour *p* constant, l'on observe que la couverture, après avoir un temps progressée, décroît à partir de *q*=*p*+4 environ (courbes en bleu discontinu). En particulier pour *p*=3, on note une décroissance à partir de *q*=8. Décroissance qui indique que les effets inférentiels du chaînage de règles ne semblent assurer une meilleure couverture que pour les phrases dont la taille est légèrement supérieure à *p*.

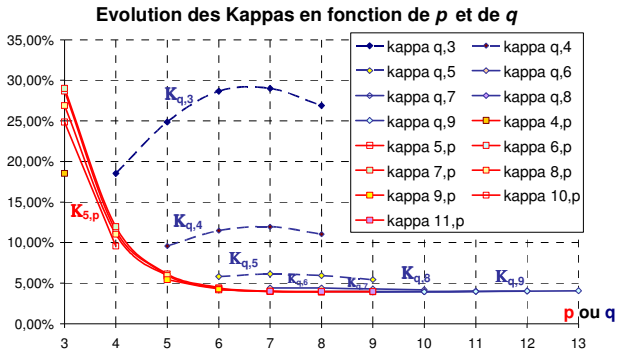

Figure 8 : La couverture de rang *q* des règles d'ordre *p* ($\kappa_{q,p}$) décroît à mesure que *p* croît. Elle croît légèrement puis décroît à mesure que *q* croît.

Pour les phrases plus longues, les effets semblent s'estomper rapidement, à mesure que *q* croît, ainsi que les tendances des courbes l'annoncent. L'espoir d'une meilleure couverture au prix d'une moins bonne discriminance se perd, puisque plutôt que d'évoluer en sens inverse, discriminance et couverture semblent, au-delà d'un certain seuil, évoluer dans un même sens, celui d'une dégradation à mesure que *q* croît. Ce résultat souligne en fait moins l'importance de la capacité inférentielle des règles de succession courtes, que l'insuffisance récurrente des corpus d'apprentissage.

8 Conclusion

Dans le résumé et l'introduction nous avons argumenté en faveur et pour l'extension de la taille des règles de succession couramment mises en œuvre dans les étiqueteurs. Les arguments, détaillés dans les sections 2 à 7, ont été de montrer que ne mettre en œuvre que les règles binaires et ternaires, c'est, dans plus de 98% des cas en situation d'étiquetage, très rapidement surgénérer un très grand nombre de règles non attestées, que l'avantage inférentiel ne permet pas de résorber, la proportion des règles engendrées attestées ou attestables restant à moins d'une fraction de pour cent du total des règles engendrées, dès que l'on dépasse la taille 5 ou 6.

La seconde partie de l'argumentation a été de montrer qu'il n'était pas déraisonnable non plus d'envisager cette extension, et d'ajouter aux règles binaires et ternaires, des règles de tailles supérieures. Pourquoi ? Parce que nous avons pu *découvrir* que cette extension pouvait être limitée, sans perte ou pratiquement, et que cette limite se situait aux environs de 12. Une taille de règle au delà de laquelle l'on constate, pour trois langues différentes, le français, l'arabe, et l'anglais, que l'on n'engendre plus, – ou très peu, c'est-à-dire bien en deçà du pour cent – de règles non attestables. Une taille qui pourrait trouver ses fondements dans la notion de proposition ou de phrase simple, dont la taille moyenne est précisément inférieure à la taille moyenne de la phrase en général, (soit 27 pour le corpus du Monde).

Les protocoles expérimentaux proposés sont simples. Ils n'exigent de disposer que d'un corpus aussi grand que possible, dûment étiqueté et découpé en phrases. En fait, seules les séquences d'étiquettes associées aux phrases sont utiles. C'est pourquoi, disponibilité des corpus aidant, nous espérons voir s'étendre ces observations à d'autres langues d'une part, et pour des volumes de corpus plus grands, d'autre part.

Cette étude des règles en soi, sous l'angle de leur fonctionnement interne, c'est-à-dire sous l'angle de leur assemblage ou aspect génératif, indépendamment de leur utilisation dans l'étiquetage, en analyse donc, nous a amené à poser le problème d'une évaluation intrinsèque, *en définition* avons-nous proposé de dire, d'un système de règles d'une façon générale. Le meilleur argument en faveur de règles ayant une portée supérieure à deux ou à trois est encore d'en montrer l'efficacité en comparant

les performances auxquelles ces règles conduisent, au regard des performances obtenues en utilisant des règles plus classiques, binaires ou ternaires. Bien entendu, et nous aurons à effectuer ces évaluations *en usage*. Mais nous voulons insister sur la complémentarité de ces deux visions de l'évaluation, « *en usage* » *vs* « *en définition* », et des potentialités de la seconde, dans une perspective prédictive ou comparative, comme nous l'avons signalé.

Cette distinction évaluation *en usage vs* évaluation *en définition* semble bien s'appliquer à notre cas qui est un cas d'analyse. Mais qu'en est-il de cette distinction si l'application englobante n'est pas, à l'image de l'étiquetage, une application d'analyse, mais une application de synthèse. Comment évaluer *en définition* un système de règles intervenant dans un programme de génération de phrases par exemple ?

Faut-il voir dans cette dichotomie « en usage » *vs* « en définition », la dichotomie « analyse » *vs* « synthèse » ? Ou encore la distinction « performance » *vs* « compétence » ? Ce qui conduirait à parler d'une évaluation *en analyse* (ou *en performance*) par opposition à une évaluation *en synthèse* (ou *en compétence*). La question est ouverte. Mais bien que nous n'ayons pas pour l'heure une vision claire de ce que pourrait être une évaluation *en définition* d'un ensemble de règles orienté vers un système de génération automatique de phrases comme en traduction par exemple, il nous semble que la distinction « en définition » *vs* « en usage » devrait encore subsister. La génération ou synthèse serait dans ce cas, comme l'est l'analyse, une application particulière qui met en œuvre un ensemble de règles dont l'évaluation indépendante reste pertinente et opératoire. A moins que seule la génération ne puisse faire l'objet d'une grammaire sur laquelle s'appuierait l'analyse. Ce qui ne remet pas en cause l'idée d'une double évaluation d'un système de règles, mais débouche sur une question qui dépasse le cadre de ce papier.

Remerciements

Le présent travail a été réalisé dans le cadre des deux projets EurADic (Action Technolangue du Ministère de la recherche), et MUSCLE (6ème PCRD). Il a en outre bénéficié de l'accueil de la Faculté des Lettres de l'Université de la Manouba, sous les auspices du MRSTDC.

Références

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).

ANDREEWSKY A., FLUHR C. (1973), "Expérience de constitution d'un programme d'apprentissage pour le traitement automatique du langage", Actes de *COLING 1973*, Volume 2.

DEBILI F. (1977), "Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage", *Thèse de Docteur-Ingénieur*, Université Paris VII, U.E.R. de Physique, Septembre 1977.

DEROSE STEVEN J., (1988), "Grammatical Category Disambiguation by Statistical Optimization", *Computational Linguistics*, Vol. 14, Num. 1.

FAIRON C., SENELLART J. (1999), "Réflexions sur la localisation, l'étiquetage, la reconnaissance et la traduction d'expressions linguistiques complexes", Actes de *TALN 1999*.

LEECH G., GARSIDE R., BRYANT M. (1994), "CLAWS4: The tagging of the British National Corpus", *Proceedings of the 15th International Conference on Computational Linguistics, COLING 94,* Kyoto, Japan.

MÉRIALDO B. (1994), "Tagging English text with a probabilistic model", *Computational Linguistics*, 20.2.

NASR A., VOLANSCHI A. (2004), "Couplage d'un étiqueteur morpho-syntaxique et d'un analyseur partiel représentés sous la forme d'automates finis pondérés", Actes de *TALN 2004*.

PAROUBEK P., RAJMAN M. (2000), "MULTITAG, une ressource linguistique produit du paradigme d'évaluation", Actes de *TALN 2000*.

TOUTANOVA K., KLEIN D., MANNING C. D., SINGER Y. (2003), "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", Actes de *HLT-NAACL*, 2003.

VALLI A., VÉRONIS J. (1999), "Etiquetage grammatical des corpus de parole: problèmes et perspectives", *Revue Française de Linguistique Appliquée*.