

Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks

Philippe Blache, Stéphane Rauzy
Laboratoire Parole et Langage
CNRS & Université de Provence

philippe.blache@lpl-aix.fr, stephane.rauzy@lpl-aix.fr

Résumé. Nous montrons dans cet article qu'il existe une corrélation étroite existant entre la qualité de l'étiquetage morpho-syntaxique et les performances des chunkers. Cette corrélation devient linéaire lorsque la taille des chunks est limitée. Nous appuyons notre démonstration sur la base d'une expérimentation conduite suite à la campagne d'évaluation Passage 2007 (de la Clergerie *et al.*, 2008). Nous analysons pour cela les comportements de deux analyseurs ayant participé à cette campagne. L'interprétation des résultats montre que la tâche de chunking, lorsqu'elle vise des chunks courts, peut être assimilée à une tâche de "super-étiquetage".

Abstract. We show in this paper that a strong correlation exists between the performance of chunk parsers and the quality of the tagging task in input. This dependency becomes linear when the size of the chunks is small. Our demonstration is based on an experiment conducted at the end of the Passage 2007 shared task evaluation initiative (de la Clergerie *et al.*, 2008). The performance of two parsers which took part in this evaluation has been investigated. The results indicate that the chunking task, for sufficiently short chunks, is similar to a super-tagging task.

Mots-clés : Analyse syntaxique, étiquetage morphosyntaxique, analyseur stochastique, analyseur symbolique superficiel, chunker.

Keywords: Parsing, tagging, stochastic parser, symbolic shallow parser, chunker.

1 Introduction

La qualité de l'étiquetage morpho-syntaxique a bien évidemment des conséquences directes sur les résultats des analyseurs syntaxiques. Mais l'importance de cette relation dépend également de la tâche fixée : les performances d'un chunker sont en effet plus directement liées à celle de l'étiqueteur qu'un analyseur syntaxique profond. Nous montrons dans cet article qu'il existe en fait une corrélation linéaire entre la tâche de chunking et celle d'étiquetage et que cette corrélation dépend du type de chunk visé par la tâche.

Nous nous appuyons dans cet article sur les résultats que nous avons obtenus dans le cadre de la campagne d'évaluation Passage 2007¹ (de la Clergerie *et al.*, 2008) qui est une continuation de la campagne Easy (Paroubek *et al.*, 2006). Ces campagnes ont permis une évaluation comparative de plusieurs analyseurs syntaxiques du français en s'appuyant sur un format d'annotation ad hoc (le guide d'annotation PEAS (Gendner *et al.*, 2003)) proposant des unités syntaxiques plates (i.e. sans constituants emboîtés). Deux évaluations distinctes étaient proposées durant ces campagnes, l'une portant sur la tâche de formation et d'identification des groupes syntaxiques, l'autre consistant à établir les relations de dépendance entre les groupes obtenus. Nous nous concentrons dans cet article sur la première tâche qui correspond ici à un problème classique de chunking : il s'agit de repérer les types et frontières des constituants de niveau 1. Nous avons testé dans le cadre de la dernière campagne deux analyseurs, l'un utilisant des techniques numériques (il s'agit d'un des premiers analyseurs stochastiques pour le français), l'autre étant basé sur une approche symbolique. Tous deux ont obtenu des résultats très satisfaisants, l'analyseur stochastique se situant dans le trio de tête des analyseurs évalués durant la campagne Passage 2007. Ces deux analyseurs prennent en entrée un texte étiqueté désambiguïté.

Pour montrer la corrélation entre étiquetage et chunking, nous proposons de faire varier de façon contrôlée la qualité de l'étiquetage fourni en entrée et d'observer les conséquences sur les résultats des analyseurs.

La première partie de cet article présente notre méthode d'étiquetage et les différentes techniques utilisées permettant le contrôle de la variation du résultat. La seconde partie décrit les deux analyseurs étudiés. La troisième partie présente les résultats obtenus par nos analyseurs en fonction de la qualité de l'entrée proposée. La dernière partie est consacrée à l'interprétation des résultats, montrant que la corrélation provient du type de chunk proposé dans le cadre de cette campagne.

2 Etiquetage morphosyntaxique

2.1 Segmentation et lexique

L'objectif de cette tâche est de segmenter en tokens le texte en entrée, puis d'associer à chacun des tokens de l'énoncé la liste des catégories morphosyntaxiques qui lui correspondent. Notre segmenteur permet de repérer les frontières entre les tokens et d'identifier les entités nécessitant un traitement spécial (nombres, dates, heures, noms propres, sigles, ...). Une fois les tokens formés, la liste des catégories morphosyntaxiques correspondant à une graphie donnée est obtenue par accès au lexique.

¹Le lien vers l'action Passage : <http://atoll.inria.fr/passage/>

Nous utilisons dans cette étude le lexique DicoLPL (Vanrullen *et al.*, 2005). Il s'agit d'un lexique assez couvrant du français (440 000 formes) qui est de plus enrichi par la donnée des fréquences lexicales pour chacune des formes (ie. le couple graphie et catégorie morphosyntaxique). Ces fréquences lexicales sont extraites d'un corpus de textes écrits d'environ 150 million de mots préalablement étiqueté, voir (Vanrullen *et al.*, 2005) pour plus de détails. L'information sur les fréquences lexicales permet notamment de préciser la répartition entre catégories qui est spécifique à chaque graphie ambiguë. Ainsi, la graphie *dans* est rencontrée dans le corpus 1 056 924 fois sous forme de préposition contre 195 fois sous forme de nom commun, alors que la graphie *envers* se distribue plus uniformément (5 174 pour la préposition, 2 123 pour le nom commun).

2.2 Désambiguisation

La procédure de désambiguisation consiste à associer à chacun des tokens de l'énoncé une catégorie morphosyntaxique unique. Nous utilisons ici le modèle des patrons (Blache & Rauzy, 2006; Blache & Rauzy, 2007), un modèle de Markov caché (HMM) plus performant que les modèles de type N-grammes. Pour les N-grammes, les états de l'automate sont identifiés par des séquences de catégories de taille identique N . Le modèle des patrons relâche cette contrainte en acceptant des états identifiés par des séquences de longueur variable (voir par exemple (Ron *et al.*, 1996)). Cette caractéristique permet en pratique d'extraire du corpus d'apprentissage un ensemble d'états, les *patrons* du modèle, qui capture de façon optimale l'information contenue dans le corpus.

Le modèle est entraîné sur le corpus Grace/Multitag (Paroubek & Rajman, 2000), un échantillon d'environ 700 000 mots annoté morphosyntaxiquement selon le jeu de traits Multext (Ide & Véronis, 1994). Dans notre étude, l'information morphosyntaxique disponible est groupée en 44 catégories distinctes (2 types de catégories pour les ponctuations, 1 pour les interjections, 2 pour les adjectifs, 2 pour les conjonctions, 1 pour les déterminants, 3 pour les noms, 8 pour les auxiliaires, 4 pour les verbes, 5 pour les prépositions, 3 pour les adverbes et 11 pour les pronoms). Les informations comme les traits d'accords en genre, nombre et personne ou le temps des verbes ne sont pas exploitées dans la présente analyse.

Afin d'étudier l'influence de la qualité de l'étiquetage sur la performance des analyseurs syntaxiques, plusieurs procédures de désambiguisation sont proposées :

- RAW : Aucun modèle n'est dans ce cas appliqué. L'étiquette associée à chaque token est la première entrée apparaissant dans la liste des catégories proposées pour le token.
- UNI : Le modèle unigramme est ici appliqué. Il donne la distribution de probabilité non contextuelle des catégories utilisées. Pour chaque token, l'étiquette retenue est celle possédant la probabilité la plus forte parmi la liste des catégories proposées.
- RAW+F : Aucun modèle n'est appliqué, mais on prend en compte la distribution des fréquences lexicales dans la liste des catégories associées à la graphie du token (cf. section 2.1). L'étiquette retenue est la catégorie présentant la fréquence lexicale maximale.
- BIG : Un modèle de bigrammes est utilisé (information sur les probabilités des catégories conditionnées par la catégorie précédente). Le modèle est alors décrit par 44 patrons, un pour chacune des catégories utilisées. Pour chaque énoncé, l'étiquetage optimal est obtenu par application de l'algorithme de Viterbi.
- BIG+F : Le modèle de bigrammes est appliqué avec un schéma de pondération qui rend compte des fréquences lexicales associées à chaque graphie de l'énoncé.

- PM : Le modèle des patrons complet, composé de 3 053 patrons de taille variable (le plus grand contexte dans la liste des patrons est composé d'une séquence de 6 catégories). L'étiquetage optimal est obtenu par application de l'algorithme de Viterbi.
- PM+F : Le modèle des patrons complet avec prise en compte des fréquences lexicales.

2.3 Evaluation de l'étiquetage

L'évaluation de la qualité de l'étiquetage est ici réalisée en calculant les scores de précision, rappel et F-Mesure sur l'échantillon de référence Grace/Multitag. Les erreurs affectant l'étiquetage de la référence imposent un seuil maximum limite pour les scores d'évaluation obtenus. Elles proviennent de manière générale d'une description incomplète, voire contradictoire, des règles d'annotation ou de fautes d'annotation sur la référence. Ces seuils limites sont fixés pour une référence donnée. La valeur des scores est ensuite dépendante des erreurs provenant de chacun des modules composant la chaîne de traitement :

- Erreurs de segmentation
- Erreurs du lexique (entrées manquantes, incorrectes, incomplètes)
- Erreurs du modèle implanté dans le module de désambiguïsation

Model	Precision	Recall	F-Measure
RAW	0.829	0.696	0.757
UNI	0.769	0.770	0.770
RAW+F	0.900	0.885	0.893
BIG	0.909	0.902	0.906
PM	0.927	0.923	0.925
BIG+F	0.933	0.926	0.929
PM+F	0.943	0.937	0.940

FIG. 1 – La qualité de l'étiquetage morphosyntaxique mesurée sur la référence Grace/Multitag en termes de F-Mesure, précision et rappel pour les sept modèles de désambiguïsation présentés section 2.2.

Nous présentons figure 1 les résultats de l'évaluation des sept modèles proposés section 2.2. Les sept modèles partagent les erreurs de segmentation et du lexique. Les valeurs des scores ne sont pas absolues, elles dépendent de la finesse de description du système de catégorisation adopté (ici 44 catégories). Nous pouvons remarquer l'influence de la prise en compte des fréquences lexicales. Cette information de nature extra-syntaxique apporte un gain significatif quel que soit le modèle considéré. Les F-Mesures de nos différents modèles couvrent l'intervalle [0.75, 0.95], ce qui nous permettra section 4 d'étudier l'influence de la qualité de l'étiquetage sur les performances de nos analyseurs syntaxiques dans cette gamme de valeurs.

3 Analyse syntaxique superficielle

3.1 L'analyseur stochastique StP1

L'analyseur stochastique StP1, comme notre étiqueteur, est basé sur le modèle des patrons (voir section 2.2). La phase d'apprentissage est effectuée sur le gold standard Easy, un corpus annoté

en constituants d'environ 100 000 mots qui a servi de référence pour la campagne d'évaluation Easy (Paroubek *et al.*, 2006). La grammaire Easy compte six constituants (ie. les groupes Easy GN, GP, NV, GA, PV et GR) faiblement hiérarchisés. Le corpus ne fournit pas les étiquettes des tokens composant les énoncés. Une phase préalable d'étiquetage (en utilisant notre modèle le plus performant PM+F, voir section 3.3) a donc été nécessaire pour produire notre échantillon d'apprentissage. Le module d'apprentissage nous permet d'extraire de cet échantillon 1080 patrons de taille variable identifiés par des séquences de catégories terminales (les catégories morphosyntaxiques) ou non-terminales (les groupes Easy).

L'analyseur stochastique StP1 prend en entrée un texte étiqueté et désambigué (c'est cette option qui sera utilisée dans la suite de l'article), ou une liste de catégories associée à chaque token de l'énoncé (ie. la sortie de la phase segmentation plus accès au lexique). Pour chaque énoncé, l'algorithme de Viterbi permet d'insérer les groupes Easy maximisant la probabilité de l'énoncé. Dans le cadre de la campagne Passage 2007, StP1 a obtenu une F-Mesure de 93.03 %.

3.2 L'analyseur superficiel ShP1

Il s'agit d'un analyseur symbolique déterministe. Il repose sur les Grammaires de Propriétés avec une stratégie de coin gauche. La grammaire utilisée est complète en ce sens qu'elle peut être utilisée indifféremment pour une analyse profonde ou superficielle (Balfourier *et al.*, 2005). La particularité de ShP1 est de s'appuyer sur un sous-ensemble de contraintes de la grammaire (en particulier les propriétés de linéarité et de constituance) pour identifier les coins gauches. La stratégie consiste à repérer à partir des coins gauches la frontière droite du chunk sur la base des autres propriétés. Cette heuristique est très efficace et permet à l'analyseur de bénéficier d'une grande rapidité (moins de 4 minutes pour traiter 1M de mots). Dans le cadre de la campagne Passage 2007, cet analyseur a obtenu une F-Mesure de 91.57 %.

3.3 Evaluation des analyseurs

L'évaluation des performances des analyseurs est ici réalisée en calculant la F-Mesure des groupes Easy sur l'ensemble ou une partie du gold standard Easy. L'estimation du score de F-Mesure n'est pas unique lorsqu'il s'agit de comparer deux structures d'arbre (ie. la référence et la sortie de l'analyseur), et dépend de l'heuristique employée. Notre score de F-Mesure est par exemple systématiquement plus bas que celui calculé dans la campagne d'évaluation Passage 2007.

Les erreurs d'annotations sur la référence Easy induisent un seuil maximum limite pour les scores obtenus, ceci indépendamment de l'analyseur évalué. Elles sont dues d'une part aux imprécisions du guide d'annotations Peas (Gendner *et al.*, 2003) rassemblant les consignes fournies aux annotateurs, et d'autre part aux fautes commises par les annotateurs eux-même. Il serait intéressant d'estimer ce taux d'erreurs, par exemple en comparant un passage du corpus annoté par plusieurs annotateurs différents. Il fixe en effet le score maximum pouvant être atteint par un analyseur.

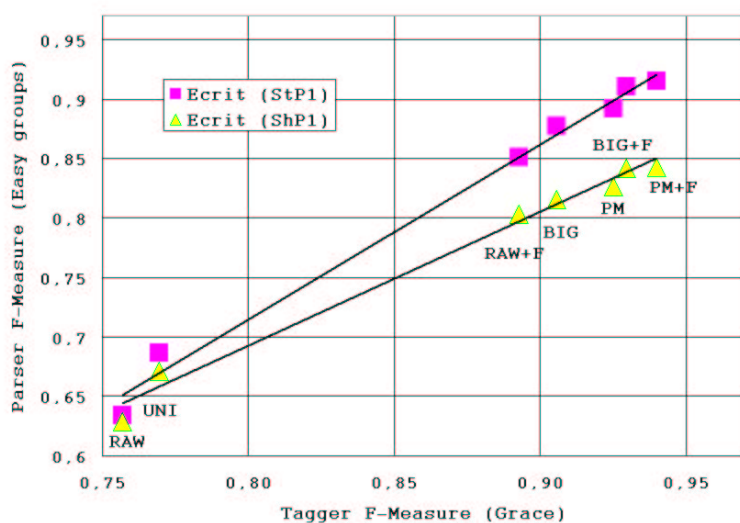


FIG. 2 – En ordonnées, les performances des analyseurs StP1 et ShP1 (pour l’écrit) mesurées sur les groupes du gold standard Easy. En abscisses, la F-Mesure calculée sur le corpus Grace mesurant la performance des modèles utilisés pour étiqueter l’entrée proposée aux analyseurs. Les étiquetages sont obtenus à partir de sept modèles classés par ordre de performance croissant (voir figure 1).

4 Résultats

Nous avons réalisé figure 2 une expérience permettant de tester l’influence de la qualité de l’étiquetage sur la performance de nos analyseurs. Pour chacun des sept modèles de désambiguïsation proposés section 2.2, le gold standard Easy a été étiqueté, puis analysé par nos deux analyseurs. Les scores de F-Mesure des analyseurs sont portés en ordonnées, la F-Mesure mesurant la qualité de l’étiquetage pour chaque modèle en abscisses. Les scores obtenus pour l’étiquetage le plus fiable, celui généré par le modèle des patrons en utilisant les fréquences lexicales (PM+F), sont respectivement de 0.899 pour l’analyseur StP1 et 0.830 pour l’analyseur ShP1 sur tout le corpus². Les scores montrés figure 2 ne concernent que les textes dans le registre de l’écrit (0.915 pour StP1 et 0.842 pour ShP1).

Pour les deux analyseurs, on observe une dépendance linéaire entre les F-Mesures mesurant la performance des analyseurs et les F-Mesures mesurant la qualité de l’étiquetage. Ainsi, dans le domaine de valeurs considérées, la progression de la performance des analyseurs est directement contrôlée par la procédure d’étiquetage adoptée. Nous expliquons section 5 ce phénomène d’un point de vue syntaxique. Nous constatons de plus que l’analyseur symbolique ShP1 présente une pente de progression plus faible que l’analyseur stochastique StP1. L’analyseur ShP1 n’exploite qu’une partie de l’information apportée par l’étiqueteur (les 44 catégories sont en effet groupées en 18 sur-catégories distinctes pour l’analyse ShP1). Nous interprétons la différence de pentes observée comme une manifestation de cet effet.

Une deuxième expérience a été réalisée dans le but de préciser les résultats obtenus. Nous avons sélectionné un sous-échantillon du gold standard Easy (de taille modeste, 5 000 mots pour le registre de l’écrit, 1 000 mots pour l’oral) pour lequel nous avons manuellement corrigé l’étiquetage morphosyntaxique. Nous présentons figure 3 la dépendance observée entre la qualité de l’étiquetage et la performance des analyseurs pour les deux échantillons, en distinguant l’oral de l’écrit. En abscisses, les sept étiquetages proposés correspondent aux modèles RAW, UNI, RAW+F, BIG, PM, PM+F de la section 2.2, et de l’étiquetage manuel de référence (F-Mesure = 1 par définition). La F-Mesure des différents modèles est cette fois-ci calculée par rapport à

²L’heuristique utilisée pour calculer la F-Mesure dans le cadre de la campagne Passage 2007 donne des scores supérieurs, 0.9303 pour StP1 et 0.9157 pour ShP1. Nos deux analyseurs obtiennent de bon résultats lorsque comparés aux scores des analyseurs participants à la campagne.

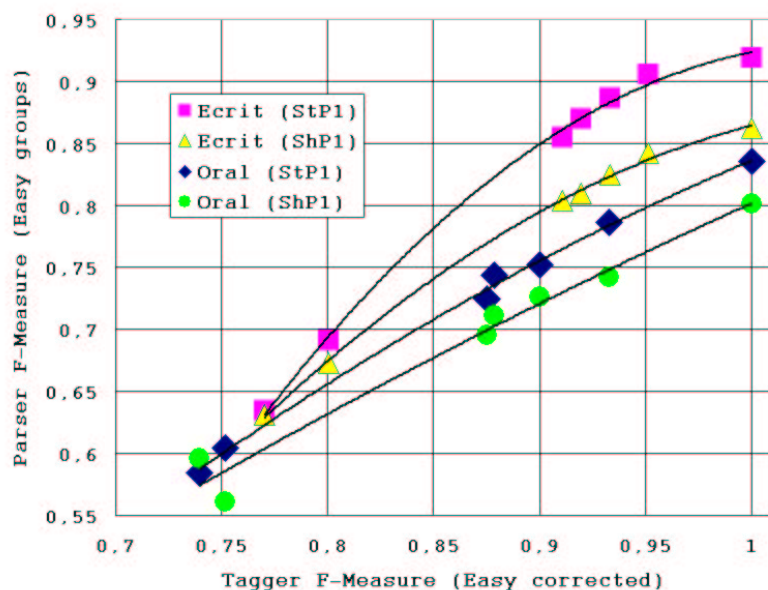


FIG. 3 – En ordonnées, les performances des analyseurs StP1 et ShP1 (pour l’oral et l’écrit) en fonction de la performance de l’étiquetage proposé en entrée. Les mesures sont effectuées sur un sous-échantillon du gold standard Easy pour lequel l’étiquetage morphosyntaxique a été corrigé manuellement. Des courbes de tendance (polynômes de degrés 2) soulignent le comportement des 2 analyseurs.

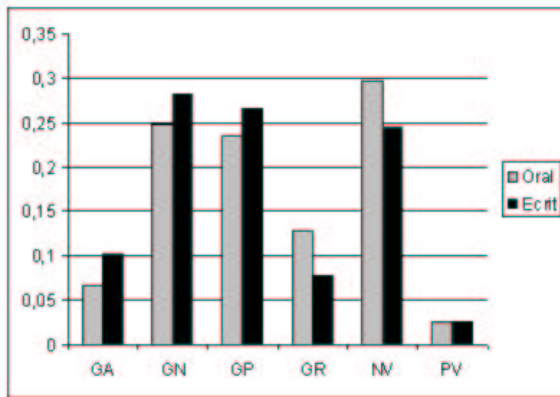
l’échantillon de référence corrigé manuellement. Les F-Mesures mesurant les performances des deux analyseurs sont présentées en ordonnées. Des courbes de tendance ont été ajoutées afin de souligner le comportement de chaque analyseur.

L’analyseur stochastique StP1 montre clairement un plateau lorsque la qualité de l’étiquetage converge vers l’unité alors que cet effet est peu marqué pour l’analyseur symbolique ShP1. Ce comportement peut s’expliquer de la manière suivante. L’analyseur stochastique StP1 a été entraîné sur le gold standard Easy, enrichi par un étiquetage automatique des catégories fourni par notre meilleur modèle de désambiguisation (PM+F). La qualité de cet étiquetage est bonne (F-Mesure de 95 % mesurée par rapport à la référence manuelle), mais pas parfaite. L’analyseur StP1 est ainsi limité par le taux d’erreurs affectant son corpus d’apprentissage. Ce n’est pas le cas pour l’analyseur symbolique ShP1 qui repose sur la spécification de sa grammaire, d’où la quasi-absence d’un plateau au voisinage de l’unité pour ShP1.

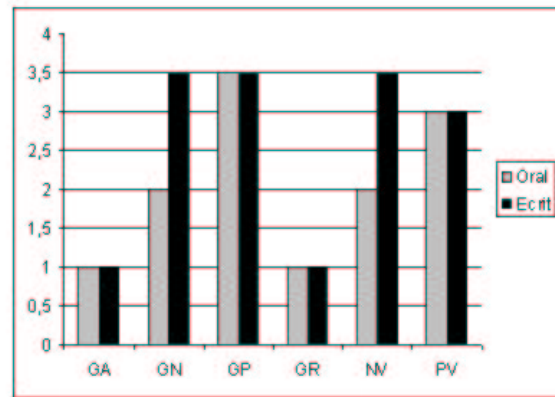
Les résultats pour l’oral, même si l’échantillon est de taille modeste (environ 1 000 mots) et donc sujet à des fluctuations statistiques, nous renseignent sur les différences d’ordre syntaxique entre le registre oral et le registre écrit. A modèle équivalent, l’étiquetage de l’oral est de moins bonne qualité que l’étiquetage de l’écrit (3 % en moyenne pour la F-Mesure). C’est un effet attendu, notre étiqueteur ayant été entraîné sur les textes écrits du corpus Grace/Multitag. Cet effet n’explique néanmoins pas complètement la différence de performances des analyseurs entre les deux registres. A qualité d’étiquetage équivalent, la figure 3 montre que les scores obtenus pour l’oral sont plus faibles de l’ordre de 6 % pour l’analyseur ShP1 et 9 % pour StP1.

5 Interprétation

Le type de chunking proposé dans les campagnes Easy et Passage repose sur la définition de 6 types de groupes : *GA* (Groupe Adjectival), *GN* (Groupe Nominal), *GP* (Groupe Prépositionnel), *GR* (Groupe Adverbial), *NV* (Noyau Verbal) et *PV* (Groupe verbal introduit par une



4.1. Distribution des chunks par type



4.2. Taille des chunks par type

FIG. 4 – Distribution et taille des chunks

préposition). La figure 4.1 donne une indication de la répartition des types sur un corpus de 195 000 mots formé d'extraits de journaux, de textes littéraires et de transcriptions de discours spontané. Nous donnons pour information la distribution des types de chunks dans les sous-ensembles de textes écrits et oraux de ce corpus. La figure 4.2 indique quant à elle la taille moyenne des chunks par type.

Ces figures montrent tout d'abord que les chunks sont globalement très courts. Les groupes les plus nombreux sont aussi les plus longs (*GN*, *GP*, *NV*), mais globalement la taille moyenne globale des chunks reste faible (2,33 mots en moyenne). Cette propriété est essentielle pour la caractérisation du processus de chunking visé dans ces campagnes d'évaluation. Cette tâche peut se faire de nombreuses façons différentes, y compris en effectuant une analyse syntaxique globale. Cependant, si nous nous limitons aux fonctions de base, le chunking consiste à identifier les frontières gauches et droites des groupes (Abney, 1991). Examinons séparément ces deux tâches. La première peut se résumer à l'identification de coins gauches (Rosenkrantz & Lewis, 1970). Il s'agit globalement d'une tâche assez simple consistant à identifier le type d'un chunk à partir de n'importe quelle catégorie (pas nécessairement la tête) pouvant débiter le groupe. Cette tâche peut être contrôlée à l'aide d'un certain nombre d'heuristiques en vue de sa désambiguïsation. L'identification de la frontière droite est plus complexe et repose sur des techniques pouvant s'approcher de l'analyse syntaxique complète.

La taille des chunks joue ici un rôle déterminant. Dans le cas des chunks d'une longueur de 1 mot, il y a évidemment confusion entre frontière droite et frontière gauche. Le cas général quant à lui est également sévèrement contrôlé par la taille : la fenêtre d'analyse nécessaire à l'identification des frontières gauches (que ce soit pour une approche statistique ou symbolique) est limitée, ce qui permet d'obtenir une approche extrêmement efficace. Cette observation a une conséquence directe sur la tâche d'identification de frontière droite. En effet, si les frontières gauches sont déterminées avec une très forte probabilité, l'identification des frontières droites joue donc un rôle secondaire dans le processus. Elle est en tout cas très fortement contrainte au point de devenir quasi sans conséquence (ce qui ne serait pas le cas avec des chunks plus longs). En conclusion, le processus de chunking dans le cadre de cette campagne peut se réduire à un processus d'identification des frontières gauches.

Par ailleurs, l'opération de chunking repose généralement sur un ensemble de types limité. Dans le cas Easy/Passage, 6 types différents sont utilisés, globalement peu ambigus (seuls deux

types ont la même catégorie potentiellement frontière gauche : PV et GP)³. La relation entre les catégories lexicales et leur projection possible vers un type de chunk est également peu ambiguë. Ce phénomène explique donc la corrélation étroite existant entre catégorisation et chunking.

6 Conclusion

Nous avons montré dans cet article la corrélation étroite existant entre la qualité de l'étiquetage et les performances des chunkers. Cette corrélation devient linéaire lorsque la taille des chunks est limitée : on peut dans ce cas réduire la tâche de chunking à celle de l'identification des frontières gauche. Cette fonction, compte tenu du nombre limité du type de chunks, est directement dépendante de la catégorie morpho-syntaxique. Nous montrons ainsi que la qualité du chunking dépend directement de celle de l'étiquetage.

Nous pouvons ainsi conclure que la tâche proposée dans le type de campagne d'évaluation indiqué est plus proche d'un "super-étiquetage" que d'une analyse syntaxique. Ce résultat permet de souligner l'importance du rôle que peut jouer la fonction d'identification de coin gauche. En termes d'évaluation, il constitue un argument en faveur de l'utilisation d'un même corpus étiqueté par tous les analyseurs participant à une campagne : la qualité propre de chaque technique d'analyse peut ainsi être comparée précisément.

Références

- ABNEY S. (1991). Parsing by chunks. In *Principle-Based Parsing*. Kluwer Academic Publishers, p. 257–278.
- BALFOURIER J.-M., BLACHE P., GUÉNOT M.-L. & VANRULLEN T. (2005). Comparaison de trois analyseurs symboliques pour une tâche d'annotation syntaxique. In *Actes de Traitement Automatique des Langues Naturelles*, volume 2, p. 41–48, Dourdan, France.
- BLACHE P. & RAUZY S. (2006). Mécanismes de contrôle pour l'analyse en grammaires de propriétés. In *Actes de Traitement Automatique des Langues Naturelles*, p. 415–424, Leuven, Belgique.
- BLACHE P. & RAUZY S. (2007). Le moteur de prédiction de mots de la plateforme de communication alternative. *Traitement Automatique des Langues*, **48**(2). sous presse.
- DE LA CLERGERIE E., AYACHE C., DE CHALANDAR G., FRANCOPOULO G., GARDENT C. & PAROUBEK P. (2008). Large scale production of syntactic annotations for french. In *Proceedings of the international workshop on Automated Syntactic Annotations for Interoperable Language Resources*, Hong-Kong.
- GENDNER V., ILOUZ G., JARDINO M., MONCEAUX L., PAROUBEK P., ROBBA I. & VILNAT A. (2003). PEAS, the first instantiation of a comparative framework for evaluating parsers of french. In *Research Notes of EACL 2003*, Budapest, Hongrie.

³La difficulté essentielle dans la tâche de repérage de frontière gauche ne se situe donc pas dans l'identification du type, mais dans l'analyse du contexte permettant de déterminer si une catégorie potentiellement frontière gauche est un coin gauche effectif.

- IDE N. & VÉRONIS J. (1994). MULTEXT : Multilingual text tools and corpora. In *Proceedings of the 15th. International Conference on Computational Linguistics (COLING 94)*, volume I, p. 588–592, Kyoto, Japan.
- PAROUBEK P. & RAJMAN M. (2000). Multitag, une ressource linguistique produit du paradigme d'évaluation. In *Actes de Traitement Automatique des Langues Naturelles*, p. 297–306, Lausanne, Suisse.
- PAROUBEK P., ROBBA I., VILNAT A. & AYACHE C. (2006). Data annotations and measures in EASY the evaluation campaign for parsers in french. In *Proceedings of the 5th international Conference on Language Resources and Evaluation*, p. 314–320, Genoa, Italy.
- RON D., SINGER Y. & TISHBY N. (1996). The power of amnesia : Learning probabilistic automata with variable memory length. *Machine Learning*, **25**, 117–149.
- ROSENKRANTZ S. & LEWIS P. (1970). Deterministic left corner parsing. In *Proceedings of the 11th Annual Symposium on Switching and Automata*, p. 139–152.
- VANRULLEN T., BLACHE P., PORTES C., RAUZY S., MAEYHIEUX J.-F., GUÉNOT M.-L., BALFOURIER J.-M. & BELLENGIER E. (2005). Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales. In *Actes de Traitement Automatique des Langues Naturelles*, volume 1, p. 511–516, Dourdan, France.