Stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un analyseur syntaxique en constituants

Matthieu Constant* — **Anthony Sigogne*** — **Patrick Watrin****

* *Université Paris-Est, LIGM, CNRS – {mconstan,sigogne}@univ-mlv.fr*
** *Knowbel Technologies – patrick.watrin@knowbel.com*

RÉSUMÉ. *Nous proposons deux stratégies discriminantes d'intégration des mots composés dans un processus d'analyse syntaxique en constituants : (i) présegmentation lexicale avant analyse, (ii) postsegmentation lexicale après analyse au moyen d'un réordonnanceur. Le segmenteur de l'approche (i) se fonde sur les champs aléatoires markoviens. Le réordonnanceur de l'approche (ii) repose sur un modèle de maximum d'entropie. Tous ces modèles intègrent des traits dédiés aux mots composés, dont certains sont calculés à partir de ressources lexicales externes. Nous montrons que l'approche par présegmentation atteint des performances dépassant l'état de l'art, alors que celle par postsegmentation est un peu en dessous de nos espérances. Les différentes expériences menées ouvrent de nombreuses pistes de recherche.*

ABSTRACT. *We propose two discriminative strategies to integrate compound word recognition in a parsing context: (i) compound pregrouping with Conditional Random Fields before parsing, (ii) reranking parses with a maximum entropy model after parsing. These discriminative models integrate features dedicated to compounds, some of theme being computed from external lexical resources. We show that the pregrouping strategy largely overtakes current state-of-the-art results, while the reranking strategy is somewhat deceiving. All the experiments conducted in this paper open new interesting research directions.*

MOTS-CLÉS : *mots composés, analyse syntaxique, champs markoviens aléatoires, réordonnanceur.*

KEYWORDS: *multiword expressions, parsing, Conditional Random Fields, reranker.*

TAL. Volume 54 – n° 1/2013, pages 47 à 70

1. Introduction

L'intégration des expressions multimots (EMM) dans des applications réelles comme la traduction automatique ou l'extraction d'information est cruciale car de telles expressions ont la particularité de contenir un certain degré de figement. En particulier, elles forment des unités lexicales complexes qui, si elles sont prises en compte, peuvent non seulement améliorer l'analyse syntaxique, mais aussi faciliter les analyses sémantiques qui en découlent. Leur intégration dans un processus d'analyse syntaxique probabiliste a déjà été envisagée dans quelques études. Toutefois, elles reposent pour la majorité sur un corpus au sein duquel l'ensemble des EMM a été parfaitement identifié au préalable. Bien qu'artificielles, ces études ont montré une amélioration des performances d'analyse : par exemple, Nivre et Nilsson (2004) et Eryigit *et al.* (2011) pour l'analyse en dépendance et Arun et Keller (2005) ainsi que Hogan *et al.* (2011) pour l'analyse en constituants. Plus récemment, Green *et al.* (2011) ont intégré la reconnaissance des EMM au sein de la grammaire et non plus dans une phase préalable. La grammaire est entraînée sur un corpus arboré où les EMM sont annotées avec des nœuds non terminaux spécifiques.

Dans cet article, nous nous intéressons à un type d'EMM : les mots composés. Nous proposons d'évaluer deux stratégies discriminantes d'intégration de leur reconnaissance dans un contexte réel d'analyse syntaxique en constituants : (a) présegmentation lexicale au moyen d'un reconnaisseur *état de l'art* de mots composés fondé sur les champs markoviens aléatoires [CRF] ; (b) analyse fondée sur une grammaire incluant l'identification des mots composés, suivie d'une phase de réordonnancement des analyses à l'aide d'un modèle de maximum d'entropie intégrant des traits dédiés aux mots composés. La stratégie (a) est une implémentation réaliste de l'approche classique de préregroupement des EMM. L'approche (b) est innovante pour la reconnaissance des EMM : nous sélectionnons la segmentation lexicale finale après l'analyse syntaxique afin d'explorer le plus d'analyses possible (contrairement à la méthode (a)). Cette approche ressemble à celle proposée par Wehrli *et al.* (2010) qui classe les hypothèses d'analyses générées par un analyseur symbolique en se basant sur la présence ou non de collocations. Les expériences que nous avons menées ont été réalisées sur le corpus arboré de Paris 7 [FTB] (Abeillé *et al.*, 2003) où les mots composés sont marqués. Nous avons utilisé un analyseur syntaxique probabiliste fondé sur une grammaire probabiliste indépendante du contexte avec annotations latentes [PCFG-LA] (Matsuzaki *et al.*, 2005 ; Petrov *et al.*, 2006). Ce type d'analyseur obtient les meilleurs résultats en français (Seddah *et al.*, 2009 ; Le Roux *et al.*, 2011), même en incorporant l'identification des EMM (Green *et al.*, 2011).

Notre article [1] est organisé comme suit : la section 2 présente les mots composés et les problématiques de leur identification ainsi que leur intégration dans un analyseur syntaxique. La section 3 détaille les ressources utilisées pour nos expériences : corpus et ressources lexicales. La section 4 décrit plus en détail les deux stratégies proposées

1. Cet article reprend les stratégies utilisées dans (Constant *et al.*, 2012b ; Constant *et al.*, 2012a) et affine les expériences et évaluations de Constant *et al.* (2012b).

et les modèles sous-jacents. Nous décrivons ensuite (dans la section 5) l'ensemble des traits dédiés aux mots composés, qui sont intégrés dans nos modèles. Puis, nous présentons l'environnement expérimental (section 6) et les résultats obtenus pour les deux stratégies (section 7). Enfin, nous discutons les résultats et proposons de nouvelles pistes de recherche dans la section 8.

2. État de l'art

2.1. *Les mots composés*

Les expressions multimots, dans le consensus actuel du domaine du traitement automatique des langues, forment des unités linguistiques qui contiennent un certain degré de non compositionalité lexicale, syntaxique, sémantique et/ou pragmatique. Elles regroupent les expressions figées, les collocations, les entités nommées, les verbes à particule, les constructions à verbe support, les termes, etc. Elles apparaissent à différents niveaux de l'analyse linguistique : certaines forment des unités lexicales contiguës à part entière (ex. *eau de vie*, *San Francisco*, *tout à fait*, *par rapport à*), d'autres forment des constituants syntaxiques comme les phrases figées (*N0 prendre le taureau par les cornes* ; *N0 cueillir N1 à froid* ; *N0 boire les paroles de N1*) ou les constructions à verbe support (*N0 donner un avertissement à N1* ; *N0 faire du bruit*).

Les expressions figées sont des combinaisons de plusieurs mots, non compositionnelles du point de vue sémantique : ex. *chaud lapin*, *s'envoyer en l'air*. Les critères linguistiques pour déterminer si une combinaison de mots est une expression figée sont fondés sur des tests syntaxiques et sémantiques comme ceux décrits dans (Gross, 1982 ; Gross, 1986). Par exemple, l'expression *boîte noire* est une expression figée car elle n'accepte pas de variations lexicales (**boîte sombre*,**caisse noire*) et elle n'autorise pas d'insertions (**boîte très noire*). De même, l'expression *casser sa pipe* avec le sens de « mourir » n'accepte pas de variations lexicales comme **(rompre + briser) sa pipe* ou **casser son fume-cigarette*. Elle n'autorise pas de transformations comme la passivation **Sa pipe a été cassée par Max*. L'opacité sémantique de telles expressions est variable : par exemple, *chaud lapin* est totalement opaque ; *vin rouge* est plus transparent [2]. Le degré de figement peut être détecté au moyen de batteries de tests syntaxiques comme dans (Gross, 1996).

Certaines expressions figées forment des séquences contiguës de mots, que l'on appelle des mots composés. C'est ce type d'expressions qui fait l'objet de notre article. On les traite comme des unités lexicales [3], de manière équivalente à des mots simples.

2. Un *vin rouge* est un vin, mais n'est pas vraiment de couleur rouge. L'ensemble dénote un type de vin.

3. De telles expressions acceptent parfois des insertions, souvent limitées à des modifieurs simples *e.g. à l'insu de, à l'insu justement de*. Linguistiquement, de telles expressions devraient être analysées au niveau syntaxique. Mais, par souci de simplification des traitements informatiques, nous les traitons au niveau lexical.

On peut imaginer que l'on remplace les espaces par des soulignés : *chaud lapin* → *chaud_lapin*. Les mots composés peuvent prendre des structures hétérogènes (*poule mouillée, tête en l'air, tour d'ivoire*). Ils forment des unités lexicales auxquelles on peut associer une partie du discours : ex. *tout à fait* est un adverbe, *à cause de* est une préposition, *table ronde* est un nom.

2.2. *Identification*

L'identification des expressions multimots dans les textes est souvent complexe car elles sont difficilement prédictibles automatiquement. Dans cette partie, nous présentons les principales approches pour les identifier.

La plupart du temps, leur reconnaissance est fondée sur la consultation de lexiques. Ces derniers peuvent être acquis soit manuellement, soit par des méthodes automatiques. Les méthodes automatiques sont généralement fondées sur des mesures statistiques qui servent à filtrer les candidats (Dunning, 1993 ; Dias, 2003 ; Pecina, 2010 ; Ramisch *et al.*, 2010b). Elles sont souvent combinées avec des traitements linguistiques comme l'étiquetage morphosyntaxique ou l'analyse syntaxique. Par exemple, certains utilisent des patrons syntaxiques fondés sur un étiquetage grammatical pour restreindre les candidats (Daille, 1995 ; Ramisch *et al.*, 2010a ; Watrin et François, 2011). D'autres, comme le suggérait Heid (1994), se fondent sur une analyse syntaxique afin de capturer aussi les variations syntaxiques de certaines expressions (Seretan *et al.*, 2003) comme les collocations verbe-nom. On observe également l'émergence de classifieurs binaires (Ramisch *et al.*, 2010a ; Tu et Roth, 2011). Etant donné un candidat, le classifieur indique si ce candidat est une expression multimot d'un certain type : les termes pour Ramisch *et al.* (2010a), les constructions à verbe support pour Tu et Roth (2011). Les modèles de classification incorporent différents types de traits. Par exemple, ils peuvent contenir des traits statistiques (ex. mesures associatives), des traits plus contextuels (les mots mis en relation, leurs étiquettes grammaticales, les valeurs des mots) ou des traits provenant de ressources externes comme les classes de Levin (1993) dans (Tu et Roth, 2011). Une autre méthode consiste à se servir de corpus alignés pour extraire des expressions multimots comme dans (Caseli *et al.*, 2010 ; Zarrieß et Kuhn, 2009).

Le plus grand désavantage des stratégies entièrement fondées sur des lexiques est que cette procédure est incapable de découvrir de nouvelles expressions. Certains comme Vincze *et al.* (2011a) ont donc proposé de combiner, par disjonctions et conjonctions, plusieurs critères fondés, entre autres, sur des lexiques, des patrons syntaxiques, des relations syntaxiques, etc. On assiste aussi à l'émergence d'approches probabilistes supervisées. Celles-ci obtiennent des résultats *état de l'art* comme dans (Green *et al.*, 2011 ; Vincze *et al.*, 2011b ; Constant et Tellier, 2012). Certains travaux montrent qu'il est intéressant de combiner la reconnaissance des mots composés avec un analyseur linguistique : un analyseur syntaxique comme dans (Green *et al.*, 2011) et un étiqueteur morphosyntaxique comme dans (Constant et Tellier, 2012). Ces méthodes ont l'avantage d'être capables d'apprendre de nouvelles expressions et de lever

des ambiguïtés grammaticales sur les séquences formant potentiellement des mots composés (ex. *en fait*). Dans cet article, nous montrerons, en particulier, l'intérêt de combiner l'approche probabiliste supervisée avec la consultation de ressources lexicales.

2.3. *Identification des mots composés dans l'analyse syntaxique*

L'approche la plus classique consiste en premier lieu à préreconnaître ces unités et à les considérer comme des blocs (soit des mots simples) en entrée de l'analyseur. Par exemple, Brun (1998) a d'abord réalisé une reconnaissance préalable de termes avant d'appliquer une grammaire lexicale fonctionnelle. Cependant, la majorité des expériences reposent sur un corpus au sein duquel l'ensemble des EMM a été parfaitement identifié au préalable : par exemple, Nivre et Nilsson (2004) et Eryigit *et al.* (2011) pour l'analyse en dépendance et Arun et Keller (2005) ainsi que Hogan *et al.* (2011) pour l'analyse en constituants. Pour l'analyse en constituants, nous pouvons noter les expériences de Cafferkey *et al.* (2007) qui ont essayé de coupler des annotateurs réels de EMM et différents types d'analyseurs probabilistes pour l'anglais. Ils ont travaillé sur un corpus de référence non annoté en EMM. Les EMM sont reconnues et prégroupées automatiquement à l'aide de ressources externes et d'un reconnaisseur d'entités nommées. Ils appliquent, ensuite, un analyseur syntaxique et réinsèrent finalement les sous-arbres correspondants aux EMM pour faire l'évaluation. Ils ont montré des gains faibles mais significatifs. Récemment, les travaux de Finkel et Manning (2009) et Green *et al.* (2011) ont proposé d'intégrer les deux tâches dans le même modèle. Finkel et Manning (2009) couplent analyse syntaxique et reconnaissance des entités nommées dans un modèle discriminant d'analyse syntaxique fondé sur les CRF. Green *et al.* (2011) ont intégré l'identification des mots composés dans la grammaire. En particulier, ils ont montré, pour le français, que le meilleur analyseur syntaxique était un analyseur PCFG-LA fondé sur une stratégie non lexicalisée, bien que l'identification des mots composés soit moins bonne qu'avec un analyseur syntaxique fondé sur une stratégie lexicalisée (une grammaire à substitution d'arbre).

3. Ressources

3.1. *Corpus*

Le corpus arboré de Paris 7[4] [FTB] (Abeillé *et al.*, 2003) est un corpus annoté en constituants syntaxiques. Il est composé d'articles provenant du journal *Le Monde*. Nous avons utilisé la version la plus récente, celle de juin 2010. Elle comporte 15 917 phrases et 473 904 mots graphiques, et utilise 13 catégories syntaxiques pour identifier les constituants. Les mots composés sont marqués et forment au total plus de 5 % des unités lexicales (mots simples et composés). Nous avons réalisé nos

4. http ://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

expériences sur une instance issue du prétraitement de l'équipe Alpage de Paris 7. Cette instance possède un jeu de 28 étiquettes morphosyntaxiques optimisé pour l'analyse syntaxique et donc très adéquat pour nos expériences. Les composants simples de chaque mot composé sont fusionnés en un seul mot : *d'abord* → *d'_abord*. Le partitionnement apprentissage/développement/test (APP/DEV/TEST) correspond au partitionnement officiel : les sections DEV et TEST sont les mêmes que dans (Candito et Crabbé, 2009), avec 1 235 phrases chacune. La section APP comporte 13 347 phrases, soit 3 390 phrases de plus que la version généralement utilisée.

L'un des problèmes du FTB dans sa forme actuelle est l'incohérence dans l'annotation des mots composés (Boudin et Hernandez, 2012), en particulier, celle des noms composés. C'est d'ailleurs l'une des raisons qui a poussé à la création du FTB-UC (Crabbé et Candito, 2008) où les mots composés de schéma régulier comme nom + adjectif ont été déliés. Cette instance du FTB est utilisée comme version de référence pour les expériences d'analyse syntaxique sur le français. Afin d'avoir le plus grand nombre de mots composés, nous avons souhaité travailler sur la version complète du FTB. Afin de corriger ce manque de cohérence, nous avons rétabli celle-ci sur les sections TEST et DEV en utilisant la procédure suivante : (1) projection du lexique interne des mots composés du FTB en vérifiant certaines contraintes syntaxiques ; (2) vérification manuelle. Par manque de moyens humains, nous avons juste appliqué la phase (1) sur la section APP.

3.2. *Ressources lexicales*

Il existe de nombreuses ressources morphosyntaxiques en français incluant les mots composés. Nous avons exploité deux dictionnaires de langue générale : le DELA (Courtois, 2009 ; Courtois *et al.*, 1997) et le Lefff (Sagot, 2010). Le DELA a été manuellement développé dans les années 80-90 par l'équipe de linguistes du LADL à Paris 7. Nous utilisons la version libre intégrée à la plate-forme Unitex [5]. Il est composé de 840 813 entrées lexicales, incluant 104 350 entrées composées (dont 91 030 noms). Les mots composés présents dans la ressource respectent, en général, les critères syntaxiques définis dans (Gross, 1986). Le Lefff [6] est une ressource lexicale qui a été accumulée automatiquement à partir de diverses sources et qui a ensuite été validée manuellement. Nous avons utilisé la version se trouvant dans MElt (Denis et Sagot, 2009). Elle comprend 553 138 entrées lexicales, incluant 26 311 entrées composées (dont 22 673 noms). Leurs différents modes de construction rendent ces deux ressources complémentaires. Pour toutes les deux, les entrées lexicales possèdent une forme fléchie, un lemme et une catégorie grammaticale. Le DELA possède un trait supplémentaire pour la plupart des mots composés : leur struture interne. Par exemple, *pomme de terre* a le code `NDN` car sa structure interne est de la forme nom – préposition *de* – nom. En complément, nous disposons aussi de lexiques spécifiques

5. http ://igm.univ-mlv.fr/~unitex/.
6. http ://atoll.inria.fr/~sagot/lefff.html

comme Prolex (Piton *et al.*, 1999) composé de toponymes (25 190 mots simples et 97 925 mots composés) et d'autres incluant des noms d'organisation et des prénoms (Martineau *et al.*, 2009).

Cet ensemble de dictionnaires est complété par une bibliothèque de grammaires locales (Gross, 1997) qui reconnaissent différents types d'unités multimots comme les entités nommées (dates, noms d'organisation, de personne et de lieu), prépositions locatives, déterminants numériques et nominaux. En pratique, nous avons utilisé une bibliothèque de 211 automates que nous avons développée à partir de la librairie en ligne GraalWeb (Constant et Watrin, 2008).

4. Modèles discriminants

Nous considérons deux stratégies d'intégration des mots composés dans le processus d'analyse syntaxique : (a) une préidentification des mots composés, suivie d'une analyse ; (b) une analyse syntaxique incorporant l'identification des mots composés suivie d'un réordonnanceur.

4.1. *Préidentification des mots composés*

La reconnaissance de mots composés peut être vue comme une tâche d'annotation séquentielle si l'on utilise le schéma d'annotation BIO (Ramshaw et Marcus, 1995). Ceci implique une limitation théorique : les mots composés doivent être continus. Ce schéma est donc théoriquement plus faible que celui proposé par Green *et al.* (2011) qui intègre les mots composés dans la grammaire et autorise des unités polylexicales discontinues. Cependant, en pratique, les mots composés sont très rarement discontinus. Dans cet article, nous utilisons le schéma d'annotation proposé dans (Constant et Tellier, 2012) pour un étiqueteur morphosyntaxique intégrant l'identification des mots composés (*i.e.* un *segmenteur-étiqueteur*). Chaque mot est associé à une étiquette de la forme X-CAT où CAT est la catégorie grammaticale de l'unité lexicale à laquelle il appartient et X détermine la position relative du mot dans l'unité lexicale (soit B pour la position initiale – *Beginning* –, soit I pour les autres positions – *Inside* –). Par exemple, on aurait *Jean/B-NPP adore/B-V les/B-DET faits/B-NC divers/I-NC*. Une fois le texte segmenté et étiqueté par notre système, on ne garde que la segmentation lexicale (on ne considère pas l'étiquetage trouvé). Cette segmentation est donnée en entrée d'un analyseur syntaxique dont la grammaire a été apprise sur le FTB où chaque mot composé est fusionné en une seule unité.

Pour cette tâche, nous utilisons le modèle des champs aléatoires markoviens linéaires (Tellier et Tommasi, 2011) [CRF] introduits par Lafferty *et al.* (2001) pour l'annotation de séquences. Étant donné une séquence de mots (graphiques)[7] en entrée

7. Un mot (graphique) correspond à une unité minimale (ou token).

$x = (x_1, x_2, ..., x_N)$ et une séquence d'étiquettes en sortie $y = (y_1, y_2, ..., y_N)$, le modèle est défini comme suit :

$$P(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{N} \exp\Big(\sum_{k=1}^{K} \lambda_k f_k(t, x, y_t, y_{t-1})\Big)$$

où $Z(x)$ est un vecteur de normalisation dépendant de x. Il est basé sur K traits définis par des fonctions binaires f_k dépendant de la position courante t dans x, l'étiquette courante y_t, l'étiquette précédente y_{t-1} et toute la séquence en entrée. Chaque mot x_i de x intègre non seulement sa propre valeur lexicale mais aussi toutes les propriétés du mot (*e.g.* ses suffixes, ses étiquettes dans un lexique externe, il commence par une majuscule, etc.). Les traits sont activés si une configuration particulière entre t, y_t, y_{t-1} et x est satisfaite (*i.e.* $f_k(t, x, y_t, y_{t-1}) = 1$). Chaque trait est associé à un poids λ_k. Ces poids sont les paramètres du modèle et sont estimés lors de la phase d'apprentissage. Les traits utilisés pour notre tâche sont décrits dans la section 5. Ils sont générés à partir de patrons qui sont instanciés à chaque position dans la séquence à annoter.

4.2. *Réordonnanceur*

Nous nous concentrons maintenant sur l'approche utilisant un réordonnanceur. Celle-ci s'appuie sur un analyseur dont la grammaire inclut l'identification des mots composés. Pour apprendre cette grammaire, l'annotation des mots composés dans le FTB a été modifiée comme dans (Green *et al.*, 2011). Les mots composés sont défaits et annotés à l'aide d'un symbole non terminal spécifique « MWX » où X est la catégorie grammaticale de l'expression. Ils ont une structure plate comme dans la figure 1.

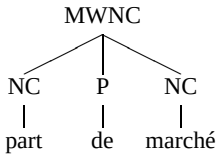

Figure 1. *Représentation des mots composés* part de marché *: le nœud MWNC correspond à un nom composé commun ; il a une structure interne plate NC P NC (nom commun – préposition – nom commun)*

Le réordonnancement discriminant consiste à reclasser les *n* meilleures analyses produites par cet analyseur syntaxique de base. Il utilise un modèle discriminant intégrant des traits associés aux nœuds des analyses candidates. Charniak et Johnson (2005) ont introduit différents traits qui permettent d'améliorer sensiblement les performances d'un analyseur syntaxique. Formellement, étant donné une phrase s, le

réordonnanceur sélectionne la meilleure analyse candidate p parmi l'ensemble de tous les candidats $P(s)$ à l'aide d'une fonction de score V_θ :

$$p^* = argmax_{p \in P(s)} V_\theta(p)$$

L'ensemble des candidats $P(s)$ correspond aux n meilleures analyses produites par l'analyseur de base. La fonction de score V_θ est le produit scalaire d'un vecteur de paramètres θ et d'un vecteur f correspondant à une distribution de traits :

$$V_\theta(p) = \theta.f(p) = \sum_{j=1}^{m} \theta_j.f_j(p)$$

où $f_j(p)$ correspond au nombre d'occurences du trait f_j dans l'analyse p. Selon Charniak et Johnson (2005), le premier trait f_1 est la probabilité de p fournie par l'analyseur de base. Le vecteur θ contient les poids associés aux traits et est estimé lors de la phase d'apprentissage à partir du corpus arboré de référence et des analyses générées par l'analyseur de base.

Dans cet article, l'utilisation du réordonnanceur est légèrement modifiée par rapport à ce qui se fait traditionnellement. En effet, nous y intégrons des traits chargés d'améliorer la reconnaissance des mots composés dans le contexte de l'analyse syntaxique. L'apprentissage du modèle est réalisé à l'aide de l'algorithme de maximum d'entropie utilisé dans (Charniak et Johnson, 2005). Les traits sont décrits dans la section 5 au moyen de patrons qui sont instanciés pour chaque nœud des analyses.

5. Les traits dédiés aux mots composés

Les deux modèles décrits dans la section 4 nécessitent des traits dédiés aux mots composés. Les traits que nous proposons sont générés à partir de patrons. Dans le but de rendre ces modèles comparables, nous avons mis en place deux jeux comparables de patrons de traits inspirés de (Constant *et al.*, 2011) : l'un adapté à l'annotation séquentielle et l'autre adapté au réordonnancement. Les patrons pour l'annotation séquentielle sont instanciés à chaque position de la séquence en entrée. Les patrons pour le réordonnanceur sont seulement instanciés aux feuilles des analyses candidates, qui sont dominées par un nœud de type EMM (c'est-à-dire qui ont un ancêtre de type EMM). Nous définissons un patron T comme suit [8] :

– annotation séquentielle : à chaque position n dans la séquence en entrée x,

$$T = f(x, n)/y_n$$

8. Le symbole / est un simple caractère séparateur.

– réordonnancement : à chaque feuille (à la position n) dominée par un nœud de type EMM m dans l'analyse candidate p,

$$T = f(p, n)/label(m)/pos(p, n)$$

où f est une fonction à définir renvoyant une chaîne de caractères ; y_n est une étiquette de sortie à la position n ; $label(m)$ est l'étiquette du nœud m et $pos(p, n)$ indique la position relative, dans l'unité polylexicale, du mot à l'indice n : B (position initiale), I (autres positions). L'instance d'un patron est une chaîne de caractères qui forme la signature d'un trait (ou une configuration), et donc d'une fonction dans nos modèles discriminants.

5.1. *Traits endogènes*

Les traits endogènes sont des traits extraits directement des mots eux-mêmes ou d'un outil appris sur le corpus d'apprentissage comme un étiqueteur morphosyntaxique.

n-grammes de mots. Nous utilisons les bigrammes et unigrammes de mots pour apprendre les mots composés présents dans le corpus d'entraînement et pour extraire des indices lexicaux afin d'en découvrir de nouveaux. Par exemple, le bigramme *coup de* est souvent le préfixe d'unités polylexicales comme *coup de pied*, *coup de foudre*, *coup de main*, etc.

n-grammes d'étiquettes morphosyntaxiques. Nous utilisons les unigrammes et bigrammes d'étiquettes morphosyntaxiques dans le but d'apprendre des structures syntaxiques irrégulières souvent caractéristiques de présence de mots composés. Par exemple, la séquence *préposition + adverbe* associée à l'adverbe composé *depuis peu* est très inhabituelle. Nous avons aussi intégré des bigrammes mélangeant mots et étiquettes morphosyntaxiques. Par exemple, le bigramme *conseiller + adjectif* correspond souvent à un mot composé dans le corpus d'apprentissage (ex. *conseiller municipal*, *conseiller régional*, etc.).

Traits spécifiques. Chaque type de modèle intègre des traits particuliers car chacun s'attelle à des tâches différentes. On incorpore dans le CRF des traits spécifiques pour gérer les mots inconnus et les mots spéciaux (nombres, traits d'union, etc.) : le mot en lettres minuscules, les préfixes et suffixes de taille 1 à 4, l'information si un mot commence par une majuscule, s'il contient un chiffre, si c'est un trait d'union. Nous ajoutons en plus les bigrammes des étiquettes de sortie. Les modèles liés au réordonnanceur intègrent des traits associés aux nœuds de type EMM, dont les valeurs sont les formes lexicales des mots composés correspondants.

5.2. *Traits exogènes.*

Les traits exogènes sont des traits qui proviennent totalement ou en partie de données externes (dans notre cas, nos ressources lexicales). Les ressources lexicales peuvent être utiles pour découvrir de nouvelles expressions. Généralement, les mots composés, en particulier les noms, suivent un schéma régulier, ce qui les rend très difficilement repérables à partir de traits endogènes uniquement. Pour cela, nous nous sommes initialement inspirés des travaux de Denis et Sagot (2009) sur l'étiquetage morphosyntaxique. Nous avons mis en place une stratégie similaire, cette fois appliquée à un segmenteur-étiqueteur, dans lequel il faut tenir compte de la segmentation contrairement à l'étiquetage pur.

Nos ressources lexicales sont appliquées au corpus à l'aide d'une analyse lexicale qui produit, pour chaque phrase, un automate fini qui représente l'ensemble des analyses possibles. Les traits exogènes sont calculés à partir de cet automate (TFST). La figure 2 donne un exemple de résultat d'une telle analyse [9].

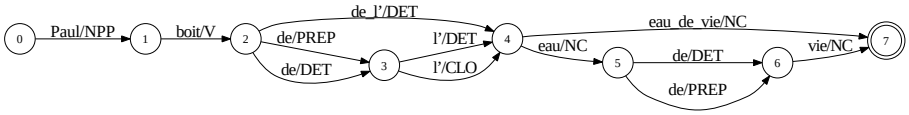

Figure 2. *Analyse lexicale de la phrase* Paul boit de l'eau de vie

Nous convertissons ensuite l'automate TFST dans le schéma *BIO* (Ramshaw et Marcus, 1995). Chaque mot est étiqueté par X-TAG où TAG est l'étiquette de l'unité lexicale à laquelle appartient le mot et X précise la position relative du mot dans l'unité : B au début, I dans les autres positions. L'automate de la figure 2 est transformé en un automate déterministe comme celui de la figure 3.

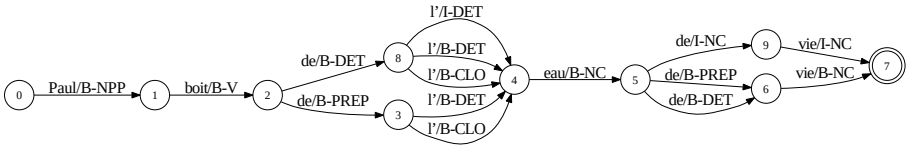

Figure 3. *Analyse lexicale de* Paul boit de l'eau de vie *dans le schéma BIO*

Chaque position dans la séquence peut être associée à un ensemble de transitions, que l'on peut ramener à un ensemble d'étiquettes. Par exemple, pour la position correspondant aux transitions sortant de l'état 5 (le mot *de*), on a l'ensemble d'étiquettes {I-NC, B-DET, B-PREP}. Nos traits se fondent alors sur la concaténation de ces éti-

9. L'étiquette CLO correspond à un clitique objet.

quettes dans un certain ordre (ex. ordre alphabétique), ce qui revient à une classe d'ambiguïté ac : *B-DET/B-PREP/I-NC*.

Nous utilisons également l'automate afin de réaliser une segmentation lexicale préliminaire en appliquant un algorithme du plus court chemin pour favoriser les analyses polylexicales. Cette segmentation préliminaire est source d'indices pour la segmentation finale, donc source de nouveaux traits. On peut associer, à tout mot appartenant à un segment composé, différentes propriétés qui seront utilisées dans les traits : l'étiquette morphosyntaxique mwt du segment, ainsi que sa structure interne mws ; sa position relative $mwpos$ dans le segment (B ou I). Par exemple, si l'on applique l'algorithme du plus court chemin à l'automate de la figure 2, on obtient la segmentation préliminaire suivante : *Paul boit de_l' eau_de_vie*. En se replaçant dans un schéma BIO et en utilisant les étiquettes issues de l'analyse lexicale pour les mots composés, on obtient une séquence *Paul/O boit/O de/B-DET l'/I-DET eau/B-NC de/I-NC vie/I-NC*. Il est alors possible, pour chaque mot de la phrase, d'extraire de nouveaux traits. Ainsi, le mot *vie* peut être associé à l'étiquette du segment figé *eau de vie* auquel il appartient (c'est-à-dire NC). Il peut aussi être associé à sa position relative I dans le segment composé.

Tous les patrons de traits sont donnés dans le tableau 1.

Traits endogènes
$w(n+i), i \in \{-2,-1,0,1,2\}$
$w(n+i)/w(n+i+1), i \in \{-2,-1,0,1\}$
$t(n+i), i \in \{-2,-1,0,1,2\}$
$t(n+i)/t(n+i+1), i \in \{-2,-1,0,1\}$
$w(n+i)/t(n+j), (i,j) \in \{(1,0),(0,1),(-1,0),(0,-1)\}$
Traits exogènes
ac(n)
mwt(n)/mwpos(n)
mws(n)/mwpos(n)

Tableau 1. *Les patrons de traits utilisés à la fois dans l'annotateur séquentiel et le réordonnanceur (n est la position courante dans la phrase) : ils correspondent à la fonction f ; $w(i)$ est le mot à la position i ; $t(i)$ est l'étiquette morphosyntaxique de $w(i)$.*

6. Environnement expérimental

Cette section présente en détail les outils utilisés pour nos expériences, les mesures utilisées pour l'évaluation, ainsi que les différents ajustements réalisés sur le corpus d'apprentissage.

6.1. *Outils*

Pour nos expériences, nous nous sommes basés sur l'analyseur syntaxique de Berkeley dans sa version la plus récente [10] (Petrov, 2010). Nous l'avons optimisé pour le traitement des mots rares et inconnus du français comme dans l'outil Bonsai (Candito et Crabbé, 2009). L'apprentissage [11] de la grammaire se fonde sur un biais calculé à partir de graines aléatoires. Afin de réduire l'effet aléatoire sur les résultats expérimentaux et donc de rendre plus significatifs nos résultats, nous nous fondons sur trois grammaires apprises à partir de trois graines aléatoires différentes [12]. Contrairement à Petrov (2010), nous n'utilisons pas le produit de grammaires car l'analyseur de Berkeley ne permet pas de générer les n meilleures analyses avec cette option. Les résultats obtenus pour chaque expérience sont les moyennes des scores obtenus pour les trois grammaires. Dans le cadre des expériences de postsegmentation, l'analyseur est utilisé comme analyseur de base du réordonnanceur en produisant ses cinquante meilleures analyses. Notre réordonnanceur est une réimplantation maison [13] de celui de Charniak et Johnson (2005) fondé sur un modèle de maximum d'entropie. La méthode de présegmentation s'appuie sur le segmenteur-étiqueteur *lgtagger* [14] fondé sur un modèle CRF linéaire. Cet outil intègre les programmes du logiciel *Wapiti* [15] (Lavergne *et al.*, 2010) qui apprend et applique le modèle CRF. Le logiciel *Unitex* est utilisé pour appliquer les ressources lexicales. Pour extraire les traits liés aux n-grammes de catégories grammaticales, nous avons utilisé *lgtagger* comme étiqueteur simple. Pour ce cas précis, il a été configuré de telle sorte qu'il ne soit pas couplé à des ressources lexicales et qu'il ne segmente pas le texte en unités polylexicales.

6.2. *Protocole d'évaluation*

Les expériences sont évaluées à l'aide de deux mesures classiques : la F_1-mesure [F_1] et la mesure UAS (*Unlabeled Attachment Score*). F_1 [16] prend en compte le parenthésage et l'étiquetage des nœuds. Le score UAS [17] évalue la qualité des liens de

10. Version téléchargée directement sur le dépôt de versionnage.
11. Pour l'apprentissage, nous utilisons tous les paramétrages par défaut.
12. Les graines aléatoires utilisées sont 1, 4 et 8.
13. Il se base sur les bibliothèques C++ PETSC et TAO. Nous avons donné un poids de 2 au trait correspondant au score de l'analyseur de Berkeley. Tous les traits générés sont conservés (il n'y a pas de filtrage par fréquence).
14. Disponible à `http://igm.univ-mlv.fr/~mconstan/research/software/`
15. Wapiti est disponible à `http://wapiti.limsi.fr/`. Nous l'avons configuré de la manière suivante : algorithme 'rprop' et les valeurs par défaut pour les pénalités L1 et L2, ainsi que le critère d'arrêt.
16. Cette mesure est calculée au moyen du programme *Evalb* qui est disponible à `http://nlp.cs.nyu.edu/evalb/`. Nous avons aussi utilisé l'évaluation par catégorie implantée dans la classe `EvalbByCat` de l'analyseur de Stanford.
17. On convertit d'abord automatiquement les analyses en constituants, en analyses en dépendances au moyen du logiciel *Bonsai*. Nous avons adapté ce dernier en ajoutant des règles de

dépendance non typés entre les mots. L'identification des mots composés est évaluée par la F_1-mesure associée aux nœuds de type EMM. Pour que les résultats générés par les méthodes par présegmentation et postsegmentation soient directement comparables, nous avons automatiquement converti les analyses générées avec mots composés fusionnés en leurs analyses équivalentes avec des nœuds non terminaux spécifiques pour les unités polylexicales.

La significativité statistique entre deux expériences d'analyse syntaxique est calculée au moyen du t-test unidirectionnel pour deux échantillons indépendants [18].La significativité statistique entre deux expériences d'identification de mots composés est établie par le test de McNemar (Gillick et Cox, 1989). La différence entre les résultats de deux expériences est considérée comme statistiquement significative si la valeur calculée lors du test est inférieure à 0,01.

6.3. *Corpus d'apprentissage*

Les expériences de référence sur l'analyse du français ont été réalisées avec des grammaires apprises à partir d'une sous-partie (9 957 phrases) de notre corpus d'apprentissage (13 347 phrases). Nous avons donc, dans un premier temps, mesuré sur le corpus DEV la différence entre des analyseurs appris sur ces deux corpus. Nous avons testé deux configurations : (a) la grammaire est apprise sur un corpus parfaitement présegmenté en mots composés [référence] ; (b) la grammaire est apprise sur un corpus où les mots composés sont annotés au moyen d'un symbole non terminal spécifique [baseline]. Le corpus de développement utilisé est la version non corrigée. La configuration (a) correspond aux expériences classiques sur le français où l'on considère en entrée de l'analyseur une segmentation parfaite. Les résultats sont donnés dans le tableau 2 et correspondent aux mesures F_1 observées sur la section DEV. Dans le cas d'une présegmentation parfaite, on observe un faible gain pas très significatif avec le grand corpus d'apprentissage. En revanche, dans le cas où l'on intègre la reconnaissance des mots composés dans la grammaire, le gain est beaucoup plus conséquent (un peu plus de 1 point). L'intérêt d'un plus grand corpus d'apprentissage dans le cadre de notre étude apparaît donc clairement.

#phrases	9 957	13 347
Référence	83,48	83,79
Baseline	78,70	79,78

Tableau 2. *Résultats suivant la taille du corpus d'apprentissage (mesure F_1)*

conversion spécifiques aux mots composés. Puis la mesure est calculée avec l'outil disponible à `http://ilk.uvt.nl/conll/software.html`.

18. La significativité statistique est calculée à l'aide de l'outil de Dan Bikel.

Comme nous l'avons vu dans la section 3, les corpus DEV et TEST ont été corrigés afin de rendre une certaine cohérence à l'annotation des mots composés. Par manque de ressources humaines, nous n'avons pu corriger le corpus d'apprentissage. Cependant, la correction du corpus comprenait deux phases dont une phase automatique qu'il est possible d'appliquer sur le corpus d'apprentissage. Afin de déterminer s'il est préférable d'utiliser le corpus autocorrigé ou le corpus initial lors de l'apprentissage de nos grammaires, nous avons procédé à des expériences préliminaires sur le corpus de développement corrigé, dont les résultats sont indiqués dans le tableau 3. Les scores donnés entre parenthèses correspondent au F_1 sur les mots composés. Au vu de ces résultats, il apparaît clairement que l'utilisation du corpus APP automatiquement corrigé permet d'améliorer l'étiquetage des mots composés (de l'ordre de 2 points pour l'expérience [référence]) et leur reconnaissance (3 points pour l'expérience [baseline]). La qualité de l'analyse syntaxique globale est légèrement améliorée, mais ce n'est pas significatif.

	Apprentissage	
	initial	corrigé
Référence	83,77 (92,5)	84,02 (94,3)
Baseline	79,85 (66,7)	79,92 (69,7)

Tableau 3. *Comparaison entre apprentissage sur le corpus initial ou celui automatiquement corrigé*

Pour le reste des expériences, nous avons utilisé l'ensemble de la section APP pour l'apprentissage dans sa version automatiquement corrigée.

7. Évaluations

Dans cette section, nous décrivons les différentes expériences réalisées et les résultats obtenus pour nos deux approches.

7.1. *Analyse syntaxique avec préreconnaissance des mots composés*

Cette sous-section est dédiée à l'évaluation de l'analyse syntaxique avec préreconnaissance des mots composés. Nous avons réalisé un certain nombre d'expériences en faisant varier l'ensemble des traits utilisés dans le segmenteur-étiqueteur. Nous notons (a) `base` l'ensemble des traits de base d'un étiqueteur grammatical (suffixes, préfixes, valeur lexicale des mots, etc.), (b) `endo` l'ensemble des traits liés aux n-grammes de mots et d'étiquettes grammaticales, et (c) `exo` l'ensemble des traits liés aux ressources lexicales externes. Les grammaires ont été apprises sur le corpus

d'apprentissage présegmenté en mots composés. Les résultats obtenus sont synthétisés dans le tableau 4. Dans l'expérience [référence], l'analyseur prend en entrée un texte parfaitement présegmenté (mais non étiqueté).

On observe tout d'abord que le meilleur analyseur utilise un segmenteur-étiqueteur comprenant tous les traits (base + endo + exo). Les performances générales sont cependant inférieures de l'ordre de 2 points en terme de mesure F_1 et de UAS par rapport à l'analyseur qui prend une présegmentation parfaite en entrée. En terme d'identification des mots composés, on observe que la segmentation réelle coûte à peu près 12 points. Ceci montre clairement que la reconnaissance des mots composés est une tâche difficile qui ne peut être négligée. Si l'on examine en détail les différents traits utilisés dans le segmenteur-étiqueteur, on s'aperçoit que l'utilisation de traits liés aux n-grammes de mots et d'étiquettes grammaticales (traits endogènes) permet de faire gagner environ 2,5 points en terme de F_1 et de UAS, ainsi que près de 15 points en terme de F_1(EMM), par rapport à l'étiqueteur-segmenteur de base. L'utilisation de tels traits facilite la reconnaissance des mots composés déjà dans le corpus d'apprentissage (par les n-grammes de mots) et permet de reconnaître des mots composés aux structures syntaxiques très irrégulières ou avec un ancrage lexical fort (comme *coup de N* = : *coup de foudre* + *coup de poing* + ...). L'exploitation de traits exogènes (traits base + exo) améliore les résultats du même ordre (voire un peu plus) que les traits endogènes en terme de qualité générale d'analyse. Les traits endogènes et exogènes se montrent complémentaires. En effet, en les combinant, on note encore un gain d'un peu moins de 1 point en terme de parenthésage général et de 4 à 6 points en terme d'identification des mots composés (par rapport aux traits endogènes et exogènes pris séparément).

	DEV			TEST		
Traits	F_1	F_1(EMM)	UAS	F_1	F_1(EMM)	UAS
Base	78,08	62,3	85,22	79,03	61,1	85,79
Base + endo	80,65	76,2	87,67	81,83	76,1	88,44
Base + exo	80,96	77,6	88,00	81,87	77,3	88,26
Base + endo + exo	81,79	82,2	88,36	82,69	81,4	88,81
Référence	84,02	94,3	90,36	84,82	93,1	90,76

Tableau 4. *Évaluation de la méthode par préreconnaissance*

Les résultats obtenus dans cet article sont bien plus spectaculaires que ceux obtenus pour des expériences similaires dans (Constant *et al.*, 2012b ; Constant *et al.*, 2012a). Il existe plusieurs facteurs pour expliquer cela. Tout d'abord, on utilise plus de ressources lexicales : lexiques complémentaires (toponymes, prénoms, organisations, etc.) et grammaires locales. De plus, la version de l'analyseur de Berkeley est plus récente. Enfin, on teste trois grammaires au lieu d'une seule. Mais la raison principale provient essentiellement de l'utilisation d'une version corrigée du FTB. En effet, si l'on utilise un segmenteur-étiqueteur intrégant tous les traits, on s'aperçoit,

comme le montre le tableau 5 [19], que l'évaluation sur les sections DEV et TEST corrigées conduit à une hausse spectaculaire des performances (en particulier sur les mots composés) par rapport aux évaluations sur les sections initiales. L'apprentissage sur un corpus automatiquement corrigé améliore encore les performances.

		DEV		TEST	
		initial	corrigé	initial	corrigé
APP	initial	80,98 (75,6)	81,61 (81,1)	82,20 (74,9)	82,50 (79,6)
	autocorrigé	- -	81,79 (82,2)	- -	82,69 (81,4)

Tableau 5. *La méthode de la préreconnaissance dans différents environnements*

7.2. *Analyse syntaxique avec réordonnancement*

Cette section est dédiée à l'évaluation de l'analyse syntaxique suivie d'un réordonnancement. Comme pour les expériences sur l'approche par présegmentation, nous avons fait varier les jeux de traits afin d'évaluer leur influence. Les traits `endo` et `exo` correspondent respectivement aux traits endogènes et exogènes décrits dans la section 5. Les traits `std` correspondent à différents traits généraux [20] décrits dans (Collins, 2000 ; Charniak et Johnson, 2005) et optimisés pour l'analyse syntaxique. La première ligne de résultats correspond à l'analyseur de base (baseline) incluant la reconnaisance des mots composés sans réordonnanceur.

	DEV			TEST		
Traits	F_1	F_1(EMM)	UAS	F_1	F_1(EMM)	UAS
-	79,92	69,7	87,15	81,11	68,4	87,95
endo	80,14	71,8	87,27	81,33	69,9	88,06
exo	80,25	72,5	87,31	81,42	70,6	88,16
endo + exo	80,32	73,1	87,36	81,45	71,2	88,10
std	80,55	71,9	87,45	81,71	70,0	88,23
std + endo + exo	80,86	72,9	87,53	81,85	70,6	88,30

Tableau 6. *Évaluation de la méthode par postreconnaissance*

Globalement, les résultats obtenus par l'approche par postsegmentation sont un peu en deçà de nos espérances. La combinaison des traits endogènes (`endo`) et

19. Les scores donnés entre parenthèses correspondent au F_1 sur les mots composés.

20. Plus précisément, nous avons utilisé les traits BIGRAMS, TRIGRAMS, RULE, NGRAMTREE, EDGES, WORD, WPROJ, HEAVY, HEADS et HEADTREE (Collins, 2000 ; Charniak et Johnson, 2005).

exogènes (`exo`) permet d'améliorer légèrement le parenthésage global, mais pas de manière significative (statistiquement). On observe également un gain d'environ 3 points sur la reconnaissance des mots composés. L'intégration des traits généraux (`std`) augmente de manière conséquente l'analyse globale (0,7 – 0,9 point), mais provoque une légère baisse des performances pour les mots composés. Comme prévu, les traits endogènes et exogènes influencent positivement la reconnaissance des mots composés, mais cette influence est relativement limitée par rapport aux traits généraux.

8. Discussion

8.1. *Préidentification ou postidentification ?*

Nous synthétisons l'ensemble des résultats importants dans le tableau 7. La ligne `baseline` correspond aux résultats d'un analyseur qui utilise une grammaire incluant la reconnaissance des mots composés. Jusqu'à Constant *et al.* (2012a), elle correspondait à la référence en terme d'analyse syntaxique de textes français non segmentés (Green *et al.*, 2011). Les résultats `présegmentation` et `postsegmentation` correspondent aux résultats de chacune des deux méthodes en combinant les traits endogènes et exogènes. Les résultats `postsegmentation`$^+$ correspondent aux résultats incluant, dans le réordonnanceur, les traits endogènes et exogènes, ainsi que les traits généraux (std). Afin de pouvoir comparer les résultats avec l'approche de présegmentation, nous avons ajouté un réordonnanceur intégrant des traits généraux à la suite de l'analyseur utilisant une approche par présegmentation (`présegmentation`$^+$).

	DEV			TEST		
Méthode	F_1	F_1(EMM)	UAS	F_1	F_1(EMM)	UAS
Baseline	79,92	69,7	87,15	81,11	68,4	87,95
Présegmentation	81,79	82,2	88,36	82,69	81,4	88,81
Postsegmentation	80,32	73,1	87,36	81,45	71,2	88,10
Présegmentation$^+$	82,35	82,1	88,68	83,17	81,4	89,11
Postsegmentation$^+$	80,86	72,9	87,53	81,85	70,6	88,30

Tableau 7. *Comparaison de la méthode par présegmentation et celle par postsegmentation*

Contrairement à nos expériences[21] dans (Constant *et al.*, 2012b ; Constant *et al.*, 2012a) qui ne permettaient pas vraiment de tirer de conclusions nettes, il apparaît clairement que l'approche par présegmentation est la meilleure méthode pour analyser un texte non segmenté en mots composés. En effet, on observe entre 1 et 1,5 point d'écart en terme de parenthésage global, et 9 à 10 points de différence en

21. Les expériences dans (Constant *et al.*, 2012b ; Constant *et al.*, 2012a) étaient quelque peu biaisées par l'incohérence des données.

terme de reconnaissance de mots composés. Une des explications possibles est que le segmenteur-étiqueteur se fonde essentiellement sur des informations locales lexicales et morphosyntaxiques, qui sont essentielles pour la reconnaissance des mots composés. Par ailleurs, le préregroupement des mots composés permet de réduire la complexité combinatoire du texte en diminuant le nombre d'unités lexicales par phrase. Le réordonnanceur, quant à lui, doit gérer en plus des informations non locales, ce qui complexifie sa tâche.

Les analyseurs décrits dans cet article constituent l'état de l'art dans le domaine de l'analyse syntaxique de textes non segmentés (pour le français). Cependant, ils pourraient encore être améliorés en utilisant les techniques de regroupement lexical décrites dans (Candito et Crabbé, 2009). En effet, considérer les mots composés comme des mots simples augmente le nombre de mots différents dans le texte et donc la dispersion lexicale. Les techniques proposées de regroupement permettent justement de réduire cette dispersion au moyen de techniques statistiques non supervisées et d'augmenter sensiblement les performances des analyseurs syntaxiques sur des textes parfaitement présegmentés. Par ailleurs, nos analyseurs ne sont pas capables de traiter l'ambiguïté purement sémantique entre une séquence figée et une séquence libre. Par exemple, selon le contexte, une *boîte noire* peut très bien être une boîte qui est de couleur noire. Comme, dans la quasi-totalité des cas de notre corpus, c'est le sens figé que l'on retrouve, nous n'avons pas cherché à résoudre ce problème. Cependant, ces ambiguïtés existent dans l'absolu et il serait utile que nos systèmes puissent aussi les traiter. Une solution possible serait d'intégrer, dans nos modèles, de nouveaux traits dépendant du contexte global dans lequel une telle séquence ambiguë est plongée.

8.2. *Une combinaison des deux approches*

Les deux approches proposées nous paraissent complémentaires. En effet, un segmenteur-étiqueteur se fonde essentiellement sur des informations locales lexicales et morphosyntaxiques. Les noms composés sont donc plutôt bien reconnus. Le réordonnanceur, quant à lui, repose sur des informations non locales (analyse syntaxique) qui permettent de résoudre certaines ambiguïtés entre séquences libres de mots et séquences figées. Par exemple, la séquence *de la* constitue, soit un déterminant partitif composé, soit une séquence compositionnelle (préposition + déterminant). La résolution de cette ambiguïté dépend de la tête lexicale à laquelle est syntaxiquement rattachée cette séquence.

Il nous semble intéressant d'essayer de combiner les deux approches. L'architecture de l'analyseur ressemblerait à ceci. Un étiqueteur-segmenteur produit les n segmentations étiquetées les plus probables. Ces dernières sont alors données en entrée de l'analyseur qui génère les m meilleures analyses. Le réordonnanceur reclasse alors les m analyses en fonction de traits non locaux. Cette piste nous paraît prometteuse comme le montrent quelques expériences préliminaires dans (Constant *et al.*, 2013).

8.3. *Validité des évaluations*

Bien que les résultats de nos expériences soient probants en terme de performances, les évaluations nous semblent cependant sujettes à discussion. Tout d'abord, les mots composés annotés dans le corpus ne correspondent pas forcément à ceux se trouvant dans nos ressources lexicales (Constant *et al.*, 2012a), c'est-à-dire il arrive fréquemment qu'un mot composé des ressources ne soit pas annoté comme tel dans le corpus [22]. On peut légitimement se demander si cela a un sens. En effet, dans les cas pratiques, les unités polylexicales qui se trouvent dans nos ressources construites manuellement sont censées être valides. Ainsi, il nous semble fondamental pour de futures évaluations de tenir compte de cet aspect. Considérer qu'une unité polylexicale de nos ressources ne soit pas valide pourrait, cependant, avoir un sens si nos ressources étaient acquises automatiquement. Par ailleurs, il est clairement admis dans la communauté que les mots composés forment un continuum entre expressions entièrement figées et expressions entièrement libres. Le fait de ne pas reconnaître une expression peu figée comme le terme *police militaire* peut sembler moins grave que de ne pas repérer une expression entièrement figée comme *chaud lapin*. Une piste pourrait consister à pondérer les expressions multimots selon leur degré de figement lors de l'évaluation.

9. Conclusion

Dans cet article, nous avons évalué deux stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un système d'analyse syntaxique probabiliste : (1) préidentification des mots composés ; (2) repérage final des mots composés après réordonnancement des n meilleures analyses. Nous avons, en particulier, mis en place un jeu de traits endogènes et exogènes spécifiques aux unités polylexicales pour les différents modèles utilisés. Les résultats de nos expériences montrent une amélioration sensible des performances en terme d'analyse syntaxique et d'identification des mots composés par rapport à l'état de l'art. L'approche par présegmentation apparaît clairement comme la plus pertinente pour notre tâche. Cependant, il nous semble intéressant, dans le futur, de combiner les deux approches qui nous paraissent complémentaires, tout en exploitant des techniques de regroupement lexical réduisant la dispersion lexicale.

Remerciements

Nous souhaitons remercier Marie Candito pour avoir mis à notre disposition la dernière version du corpus arboré de Paris 7.

22. Cette incompatibilité est essentiellement due à des différences dans les critères de sélection des mots composés mis en œuvre lors de la construction du FTB et des lexiques. Les critères utilisés pour la constitution du lexique apparaissent plus souples que ceux utilisés pour l'annotation du corpus.

10. Bibliographie

Abeillé A., Clément L., Toussenel F., « Building a treebank for French », *in* A. Abeillé (ed.), *Treebanks*, Kluwer, Dordrecht, 2003.

Arun A., Keller F., « Lexicalization in crosslinguistic probabilistic parsing : The case of French », *Proceedings of the Annual Meeting of the Association For Computational Linguistics (ACL'05)*, p. 306-313, 2005.

Boudin F., Hernandez N., « Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'12)*, p. 281-291, 2012.

Brun C., « Terminology finite-state preprocessing for computational LFG », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING'98)*, p. 196-200, 1998.

Cafferkey C., Hogan D., van Genabith J., « Multi-word units in treebank-based probabilistic parsing and generation », *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP'07)*, 2007.

Candito M. H., Crabbé B., « Improving generative statistical parsing with semi-supervised word clustering », *Proceedings of 11th International Conference on Parsing Technologies (IWPT'09)*, p. 138-141, 2009.

Caseli H. d. M., Ramisch C., das Gracas Volpe Nunes M., Villavicencio A., « Alignment-based extraction of multiword expressions », *Language Resources and Evaluation*, vol. 44, nº 1-2, p. 59-77, 2010.

Charniak E., Johnson M., « Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking », *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 173-180, 2005.

Collins M., « Discriminative reranking for natural language parsing », *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, p. 175-182, 2000.

Constant M., Roux J. L., Sigogne A., « Combining Compound Recognition and PCFG-LA Parsing with Word Lattices and Conditional Random Fields », *ACM Transaction in Speech and Language Processing*, 2013.

Constant M., Sigogne A., Watrin P., « Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, p. 204-212, 2012a.

Constant M., Sigogne A., Watrin P., « La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes », *Actes de la conférence sur le Traitement Automatique des Langues Naturelles (TALN'12)*, p. 57-70, 2012b.

Constant M., Tellier I., « Evaluating the Impact of External Lexical Resources into a CRF-based Multiword Segmenter and Part-of-Speech Tagger », *Proceedings of the 8th conference on Language Resources and Evaluation (LREC'12)*, 2012.

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A., Billot S., « Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français », *Actes de Conférence sur le traitement automatique des langues naturelles (TALN'11)*, 2011.

Constant M., Watrin P., « Networking Multiword Units », *in* B. Nordstrom, A. Ranta (eds), *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL'08)*, vol. 5221 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, p. 120-125, 2008.

Courtois B., « Un système de dictionnaires électroniques pour les mots simples du français », *Langue Française*, vol. 87, p. 11-22, 2009.

Courtois B., Garrigues M., Gross G., Gross M., Jung R., Mathieu-Colas M., Monceaux A., Poncet-Montange A., Silberztein M., Vivés R., Dictionnaire électronique DELAC : les mots composés binaires, Technical Report n° 56, University Paris 7, LADL, 1997.

Crabbé B., Candito M. H., « Expériences d'analyse syntaxique statistique du français », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'08)*, 2008.

Daille B., « Repérage et extraction de terminologie par une approche mixte statistique et linguistique », *Traitement Automatique des Langues (TAL)*, vol. 36, n° 1-2, p. 101-118, 1995.

Denis P., Sagot B., « Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort », *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*, p. 110-119, 2009.

Dias G., « Multiword Unit Hybrid Extraction », *Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics (MWE'03)*, p. 41-49, 2003.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Eryigit G., Ilbay T., Arkan Can O., « Multiword Expressions in Statistical Dependency Parsing », *Proceedings of the IWPT Workshop on Statistical Parsing of Morphologically-Rich Languages (SPRML'11)*, p. 45-55, 2011.

Finkel J. R., Manning C. D., « Joint Parsing and Named Entity Recognition », *Proceedings of the annual conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT'09)*, p. 326-334, 2009.

Gillick L., Cox S., « Some statistical issues in the comparison of speech recognition algorithms », *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'89)*, p. 532-535, 1989.

Green S., de Marneffe M.-C., Bauer J., Manning C. D., « Multiword Expression Identification with Tree Substitution Grammars : A Parsing tour de force with French », *Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11)*, p. 725-735, 2011.

Gross G., *Les expressions figées en français. Noms composés et autres locutions*, Orphrys, 1996.

Gross M., « Une classification des phrases "figées" du français », *Revue québécoise de linguistique*, vol. 11, n° 2, p. 151-185, 1982.

Gross M., « Lexicon Grammar. The Representation of Compound Words », *Proceedings of the conference on Computational Linguistics (COLING'86)*, p. 1-6, 1986.

Gross M., « The construction of local grammars », *in* E. Roche, Y. Schabes (eds), *Finite-State Language Processing*, The MIT Press, Cambridge, Mass., p. 329-352, 1997.

Heid U., « On ways words work together », *Research topics in lexical combinatorics. Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX'94)*, p. 226-257, 1994.

Hogan D., Foster J., van Genabith J., « Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities », *Proceedings of ACL Workshop Multiword Expressions : from Parsing and Generation to the Real World (MWE'2011)*, p. 14-19, 2011.

Lafferty J., McCallum A., Pereira F., « Conditional random Fields : Probabilistic models for segmenting and labeling sequence data », *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, p. 282-289, 2001.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, p. 504-513, 2010.

Le Roux J., Favre B., Mirroshandel S. A., Nasr A., « Modèles génératif et discriminant en analyse syntaxique : expériences sur le corpus arboré de Paris 7 », *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'11)*, 2011.

Levin B., *English Verb Classes and Alternations, A Preliminary Investigation*, University of Chicago Press, 1993.

Martineau C., Nakamura T., Varga L., Voyatzi S., « Annotation et normalisation des entités nommées », *Arena Romanistica*, vol. 4, p. 234-243, 2009.

Matsuzaki T., Miyao Y., Tsujii J., « Probabilistic CFG with latent annotations », *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 75-82, 2005.

Nivre J., Nilsson J., « Multiword units in syntactic parsing », *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*, 2004.

Pecina P., « Lexical association measures and collocation extraction », *Language Resources and Evaluation*, vol. 44, p. 137-158, 2010.

Petrov S., « Products of Random Latent Variable Grammars », *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT'10)*, p. 19-27, 2010.

Petrov S., Barrett L., Thibaux R., Klein D., « Learning accurate, compact and interpretable tree annotation », *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'06)*, 2006.

Piton O., Maurel D., Belleil C., « The Prolex Data Base : Toponyms and gentiles for NLP », *Proceedings of the Third International Workshop on Applications of Natural Language to Data Bases (NLDB'99)*, p. 233-237, 1999.

Ramisch C., Villavicencio A., Boitet C., « mwe-toolkit : a framework for multiword expression identification », *Proceedings of the Language Resources and Evaluation Conference (LREC'10)*, 2010a.

Ramisch C., Villavicencio A., Boitet C., « Web-based and combined language models : a case study on noun compound identification », *Proceedings of the Conference on Computational Linguistics (COLING'10)*, p. 1041-1049, 2010b.

Ramshaw L. A., Marcus M. P., « Text chunking using transformation-based learning », *Proceedings of the 3rd Workshop on Very Large Corpora*, p. 88-94, 1995.

Sagot B., « The Lefff, a freely available, accurate and large-coverage lexicon for French », *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

Seddah D., Candito M.-H., Crabbé B., « Cross-parser evaluation and tagset variation : a French treebank study », *Proceedings of International Conference on Parsing Technologies (IWPT'09)*, p. 150-161, 2009.

Seretan V., Nerima L., Wehrli E., « Extraction of Multi-Word Collocations Using Syntactic Bigram Composition », *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*, Borovets, Bulgaria, p. 424-431, 2003.

Tellier I., Tommasi M., « Champs markoviens conditionnels pour l'extraction d'information », *in* Eric Gaussier, François Yvon (eds), *Modèles probabilistes pour l'accès à l'information textuelle*, Hermès, 2011.

Tu Y., Roth D., « Learning English Light Verb Constructions : Contextual or Statistical », *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, p. 31-39, 2011.

Vincze V., Nagy I., Berend G., « Detecting noun compounds and light verb constructions : a contrastive study », *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World (MWE'11)*, p. 116-121, 2011a.

Vincze V., Nagy I., Berend G., « Multiword Expressions and Named Entities in the Wiki50 Corpus », *Proceedings of the conference on Recent Advances in Natural Language Processing (RANLP'11)*, p. 289-295, 2011b.

Watrin P., François T., « N-gram frequency database reference to handle MWE extraction in NLP applications », *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World (MWE'11)*, 2011.

Wehrli E., Seretan V., Nerima L., « Sentence analysis and collocation identification », *Proceedings of the Workshop on Multiword Expression : From Theory to Applications (MWE'10)*, p. 28-36, 2010.

Zarrieß S., Kuhn J., « Exploiting Translational Correspondences for Pattern-Independent MWE Identification », *Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications (MWE'09)*, p. 23-30, 2009.