La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes

Matthieu Constant[1] Anthony Sigogne[1] Patrick Watrin[2]
(1) université Paris-Est, LIGM, CNRS, 5, bd Descartes 774545 Marne-la-Vallée
(2) Université de Louvain, CENTAL, Louvain-la-Neuve
mconstan@univ-mlv.fr, sigogne@univ-mlv.fr,patrick.watrin@uclouvain.be

RÉSUMÉ

Nous proposons deux stratégies discriminantes d'intégration des mots composés dans un processus réel d'analyse syntaxique : (i) pré-segmentation lexicale avant analyse, (ii) post-segmentation lexicale après analyse au moyen d'un réordonnanceur. Le segmenteur de l'approche (i) se fonde sur un modèle CRF et permet d'obtenir un reconnaisseur de mots composés *état-de-l'art*. Le réordonnanceur de l'approche (ii) repose sur un modèle MaxEnt intégrant des traits dédiés aux mots composés. Nous montrons que les deux approches permettent de combler jusqu'à 18% de l'écart entre un analyseur *baseline* et un analyseur avec segmentation parfaite et jusqu'à 25% pour la reconnaissance des mots composés.

ABSTRACT

Recognition of compound words tested against parsing and vice-versa : evaluation of two discriminative approaches

We propose two discriminative strategies to integrate compound word recognition in a real parsing context : (i) state-of-the-art compound pregrouping with Conditional Random Fields before parsing, (ii) reranking parses with features dedicated to compounds after parsing. We show that these two approaches help reduce up to 18% of the gap between a baseline parser and parser with golden segmentation and up to 25% for compound recognition.

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

KEYWORDS: Multiword expressions, parsing, Conditional random Fields, reranker.

1 Introduction

L'intégration des expressions multi-mots (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 EMMs a

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 57–70,
Grenoble, 4 au 8 juin 2012.

été parfaitement identifié au préalable. Bien qu'artificielles, ces études ont montré une amélioration des performances d'analyse : par exemple, (Nivre et Nilsson, 2004; Eryigit *et al.*, 2011) pour l'analyse en dépendance et (Arun et Keller, 2005; Hogan *et al.*, 2011) pour l'analyse en constituants. Plus récemment, (Green *et al.*, 2011) a intégré la reconnaissance des EMMs au sein de la grammaire et non plus dans une phase préalable. La grammaire est entrainée sur un corpus arboré où les EMMs sont annotées avec des noeuds non-terminaux spécifiques.

Dans cet article, nous nous intéressons à un type d'EMMs : les mots composés. Nous proposons d'évaluer deux stratégies discriminantes d'intégration de ces expressions dans un contexte réel d'analyse syntaxique en constituants : (a) pré-segmentation lexicale au moyen d'un reconnaisseur *état-de-l'art* de mots composés basé sur les champs markoviens aléatoires [CRF] ; (b) analyse basée sur une grammaire incluant l'identification des mots composés, suivie d'une phase de réordonnancement des analyses à l'aide d'un modèle maximum d'entropie intégrant des traits dédiés aux mots composés. (a) est une implémentation réaliste de l'approche classique de pré-regroupement des EMMs. Nous souhaitons évaluer si la reconnaissance automatique des EMMs a toujours un impact positif sur l'analyse syntaxique, c'est-à-dire, si une segmentation lexicale imparfaite ne provoque pas trop d'effets de bord sur les constituants supérieurs. L'approche (b) est innovante pour la reconnaissance des EMMs : nous sélectionnons la segmentation lexicale finale après l'analyse syntaxique afin d'explorer le plus d'analyses possibles (contrairement à la méthode (a)). Cette approche ressemble à celle proposée par (Wehrli *et al.*, 2010) qui reclasse 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 l'analyseur syntaxique de Berkeley (Petrov *et al.*, 2006) qui est fondé sur une stratégie non-lexicalisée et qui obtient les meilleurs résultats en français (Seddah *et al.*, 2009), même en incorporant l'identification des EMMs (Green *et al.*, 2011).

L'article est organisé comme suit : la section 2 présente les problématiques du repérage des EMMs et de leur intégration dans un analyseur syntaxique. La section 3 décrit plus en détail les deux stratégies proposées et les modèles sous-jacents. La section 4 détaille les ressources utilisées pour nos expériences : corpus et lexiques. Nous décrivons ensuite (dans la section 5) l'ensemble des traits dédiés aux EMMs intégrés dans nos deux modèles. Enfin, la section 6 rapporte et analyse les résultats obtenus lors de nos expériences.

2 Les mots composés

Les expressions multi-mots sont des unités lexicales constituées de plusieurs lexèmes qui contiennent un certain degré de figement. Elles couvrent une large gamme de phénomènes linguistiques : les expressions figées et semi-figées, les constructions à verbe support, les entités nommées, les termes, etc. Elles sont souvent divisées en deux classes : les expressions définies au moyen de critères linguistiques et celles définies par des critères purement statistiques (collocations) (Sag *et al.*, 2002). La plupart des critères linguistiques pour déterminer si une combinaison de mots est une EMM sont basés sur des tests syntaxiques et sémantiques comme ceux décrits dans (Gross, 1986). Par exemple, l'expression *caisse noire* est une EMM car elle n'accepte pas de variations lexicales (**caisse sombre*) et elle n'autorise pas d'insertions (**caisse très noire*). De telles expressions définies linguistiquement peuvent coincider en partie avec les collocations

qui constituent des associations habituelles de mots (fondées sur la notion de fréquence). Ces dernières sont souvent identifiées au moyen de mesures statistiques associatives. Dans cet article, nous nous focalisons sur les EMMs continues qui 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. Les variations morphologiques et lexicales sont très limitées – e.g. *caisse noire+caisses noires*, *vin (rouge+blanc+rosé+*orange+...)* – et les variations syntaxiques très souvent interdites [1]. De telles expressions sont généralement analysées au niveau lexical. Par la suite, nous utilisons le terme *mot composé* ou *unité polylexicale*.

2.1 Identification

L'identification des EMMs dans les textes est souvent complexe car leur propriété de figement les rend difficilement prédictibles. Elle repose généralement sur des stratégies lexicalisées. La plus simple est fondée sur la consultation de lexiques comme dans (Silberztein, 2000). Le plus grand désavantage est que cette procédure se base entièrement sur des dictionnaires et est donc incapable de découvrir de nouveaux mots composés. L'utilisation d'extracteurs de collocations peut donc s'avérer utile. Par exemple, (Watrin et François, 2011) calcule à la volée pour chaque collocation candidate dans le texte traité, son score d'association au moyen d'une base externe de n-grammes apprises sur un grand corpus brut. L'expression est ensuite étiquetée comme EMM si son score d'association est plus grand qu'un seuil donné. Ils obtiennent d'excellents résultats dans le cadre d'une tâche d'extraction de mots-clés. Dans le cadre d'une évaluation sur corpus de référence, (Ramisch *et al.*, 2010) a développé un classifieur basé sur un séparateur à vastes marges intégrant des traits correspondant à différentes mesures d'associations des collocations. Les résultats sont plutôt faibles sur le corpus GENIA. (Green *et al.*, 2011) a confirmé ces mauvais résultats sur le corpus arboré de Paris 7. Ceci s'explique par le fait que de telles méthodes ne font aucune distinction entre les différents types de EMMs et que les types de EMMs annotés dans les corpus sont souvent limités. Une approche récente consiste à coupler, dans un même "modèle", l'annotation des mots composés avec un analyseur linguistique : un étiqueteur morphosyntaxique dans (Constant *et al.*, 2011) et un analyseur syntaxique dans (Green *et al.*, 2011). (Constant *et al.*, 2011) apprend un modèle CRF intégrant différents traits classiques de l'étiquetage morphosyntaxique et des traits basés sur des ressources externes. (Green *et al.*, 2011) a proposé que l'identification des mots composés soit intégrée dans la grammaire de l'analyseur, qui est apprise sur un corpus arboré où les mots composés sont annotés au moyen de noeuds non-terminaux spécifiques. Ils ont utilisé, avec succès, une grammaire à substitution d'arbres au lieu d'une grammaire probabiliste indépendante du contexte (avec annotations latentes) afin d'apprendre des règles lexicalisées. Les deux méthodes ont l'avantage d'être capables d'apprendre de nouveaux mots composés. Dans cet article, nous exploitons les avantages des méthodes décrites dans cette section en les intégrant comme traits d'un unique modèle probabiliste.

2.2 Intégration dans l'analyse syntaxique

La majorité des expériences d'intégration des EMMs dans un processus d'analyse syntaxique repose sur des corpus au sein desquels les mots composés ont été parfaitement identifiés au

1. De telles expressions acceptent très rarement des insertions, souvent limitées à des modifieurs simples e.g. *à court terme*, *à très court terme*.

préalable. Bien qu'artificielles, ces études ont montré une amélioration des performances d'analyse : par exemple, (Nivre et Nilsson, 2004; Eryigit *et al.*, 2011) pour l'analyse en dépendance et (Arun et Keller, 2005; 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 EMMs et différents types d'analyseurs probabilistes pour l'anglais. Ils ont travaillé sur un corpus de référence non annoté en EMMs. Les EMMs sont reconnues et pré-groupées 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 EMMs 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) couple analyse syntaxique et reconnaissance des entités nommées dans un modèle discriminant d'analyse syntaxique basé sur les CRF. (Green *et al.*, 2011) a intégré l'identification des mots composés dans la grammaire. Ils ont, en particulier, montré, pour le français, que le meilleur analyseur syntaxique était toujours l'analyseur de Berkeley (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. Enfin, il existe les travaux de (Wehrli *et al.*, 2010) qui reclasse les hypothèses d'analyses générées par un analyseur symbolique en se basant sur la présence ou non de collocations.

3 Modèles discriminants

Nous considérons deux stratégies d'intégration des mots composés dans le processus d'analyse syntaxique : (a) une pré-identification des mots composés, suivie d'une analyse ; et (b) une analyse syntaxique incorporant l'identification des mots composés suivie d'un réordonnanceur intégrant des traits dédiés aux EMMs.

3.1 Pré-identification 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 IOB (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 très rarement discontinus et dans la majorité des cas, la discontinuité est due à l'insertion d'un simple modifieur qui peut être incorporé dans la séquence figée : *à court terme*, *à très court terme*. Dans cet article, nous proposons d'associer les composants simples des unités polylexicales à une étiquette de la forme CAT+X où CAT est la catégorie grammaticale du mot composé et X détermine la position relative du token dans le mot composé (soit B pour le début – Beginning–, soit I pour les autres positions –Inside–). Les mots simples sont étiquetés O (outside) : *Jean/O adore/O les/O faits/N+B divers/N+I*.

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.

Etant donné une séquence de mots (graphiques) [2] en entrée $x = (x_1, x_2, ..., x_N)$ et une séquence d'étiquettes en sortie $y = (y_1, y_2, ..., y_N)$, le modèle est défini comme suit :

$$P_\lambda(y|x) = \frac{1}{Z(x)}.\sum_{t}^{N}\sum_{k}^{K} log\,\lambda_k.f_k(t, y_t, y_{t-1}, x)$$

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} and x est satisfaite (i.e. $f_k(t, y_t, y_{t-1}, x) = 1$). 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. Chaque instance x correspond à une fonction binaire f_k qui retourne 1 si l'instance à la position courante correspond à x, 0 sinon.

3.2 Réordonnanceur

Le réordonnancement discriminant consiste à reclasser les n meilleures analyses produites par un analyseur syntaxique de base et à sélectionner la meilleure. Il utilise un modèle discriminant intégrant des traits associés aux noeuds des analyses candidates. (Charniak et Johnson, 2005) a 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 de traits f :

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

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 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

2. Un mot (graphique) correspond à un token.

des mots composés dans le contexte de l'analyse syntaxique. Ces traits sont décrits dans la section 5 au moyen de patrons qui sont instanciés pour chaque noeud des analyses. L'apprentissage du modèle est réalisé à l'aide de l'algorithme de maximum d'entropie utilisé dans (Charniak et Johnson, 2005).

4 Ressources

4.1 Corpus

Le corpus arboré de Paris 7 [3] [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 expériences sur deux instances différentes provenant de cette même version : l'instance issue du prétraitement décrit dans (Green *et al.*, 2011) [FTB-STF] et l'instance issue du prétraitement réalisé par la chaîne de traitement de l'équipe Alpage de Paris 7 [FTB-P7]. FTB-STF possède un jeu de 14 étiquettes morphosyntaxiques et a été utilisé pour avoir des résultats comparables avec (Green *et al.*, 2011) en terme d'identification des mots composés. 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. Il existe 11 symboles de type EMM. FTB-P7 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. Pour pouvoir réaliser nos expériences, il a été nécessaire de défaire tous les mots composés et les représenter comme dans l'instance FTB-STF. Les étiquettes morphosyntaxiques des composants simples des unités polylexicales ont été ajoutées à l'aide de l'étiqueteur morphosyntaxique *lgtagger* (Constant et Sigogne, 2011) appris sur la version du FTB où les mots composés ne sont pas défaits. Le partionnement *entraînement/développement/test* correspond au partitionnement officiel : les sections *développement* et *test* sont les mêmes que dans (Candito et Crabbé, 2009), avec 1 235 phrases chacune. La section entraînement comporte 13 347 phrases, soit 3 390 phases en plus que la version généralement utilisée.

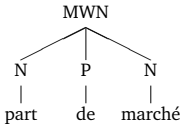

FIGURE 1 – Représentation des mots composés *part de marché* : le noeud MWN correspond à un nom composé ; il a une structure interne plate N P N (nom – préposition – nom)

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

4.2 Ressources lexicales

Il existe de nombreuses resources morphologiques 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 plateforme Unitex[4]. 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[5] 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, *eau de vie* a le code `NDN` car sa structure interne est de la forme nom – préposition *de* – nom.

En terme de collocations, (Watrin et François, 2011) a présenté un système retournant, pour toute phrase, la liste des collocations nominales potentielles accompagnées de leur mesure d'association. Pour le FTB, nous obtenons 17,315 collocations nominales candidates associées à leur log-vraisemblance et leur structure interne.

5 Les traits dédiés aux mots composés

Les deux modèles décrits dans la section 3 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 noeud de type EMM (c'est-à-dire qui ont un ancêtre de type EMM). Nous définissons un patron T comme suit :

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

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

– Réordonnancement : à chaque feuille (à la position n) dominée par un noeud 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 ; y_n est une étiquette de sortie à la position n ; $label(m)$ est l'étiquette du noeud 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).

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

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'entrainement 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.

Traits spécifiques. Chaque type de modèle intègre des traits particuliers car chacun s'attèle 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 noeuds 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. 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.

Les traits basés sur un lexique. Nous associons à chaque mot l'ensemble des étiquettes morphosyntaxiques trouvées dans notre lexique morphologique externe. Cet ensemble forme une classe d'ambiguité ac. Si un mot appartient potentiellement à une unité polylexicale dans son contexte d'occurrence, l'étiquette correspondante au mot composé est aussi intégrée à la classe d'ambiguité. Par exemple, le mot *de* dans le contexte *eau de vie* est associé à la classe *det_nom+I_prep*. En effet, le mot simple *de* est soit un déterminant (det) soit une préposition (prep). Par ailleurs, il se trouve dans une position interne (I) du nom *eau de vie*. Ce trait est directement calculé à partir de l'automate généré par l'analyse lexicale. Nous utilisons également cet 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 : l'étiquette morphosyntaxique mwt du segment, ainsi que sa structure interne mws ; sa position relative $mwpos$ dans le segment ('B' ou 'I').

Traits basés sur les collocations. Dans notre ressource de collocations, chaque candidat du FTB est accompagné de sa structure syntaxique interne et de son score d'association (log-vraisemblance). Nous avons divisé ces candidats en deux classes : ceux qui ont un score supérieur à un certain seuil et les autres. Ainsi, tout mot du corpus peut être associé à un certain nombre de propriétés s'il appartient à une collocation candidate : la classe de la collocation c ainsi que sa structure interne cs, la position relative $cpos$ du mot dans la collocation ('B' ou 'I'). Nous avons manuellement fixé le seuil à une valeur de 150 après une phase de réglage sur le corpus de développement.

Tous les patrons de traits sont donnés dans la table 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)
c(n)/cs(n)/cpos(n)

TABLE 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.

6 Evaluation

6.1 Préliminaires

L'ensemble des expériences décrites ci-dessous ont été réalisées avec l'analyseur syntaxique de Berkeley [6]. Nous notons BKYc l'analyseur dont la grammaire [7] a été apprise sur le FTB où les mots composés sont fusionnés ; BKY l'analyseur dont la grammaire a été apprise sur le FTB où les mots composés sont défaits et annotés par un symbole non-terminal spécial.

Les expériences sont évaluées à l'aide de plusieurs mesures classiques : la F-mesure [F], la mesure UAS (*Unlabeled Attachment Score*) et la mesure LA (*Leaf Ancestors*). F [8] prend en compte le parenthésage et l'étiquetage des noeuds. Le score UAS [9] évalue la qualité des liens de dépendance non typés entre les mots. Finalement, la mesure LA [10] (Sampson et Babarczy, 2003) calcule la similarité entre les chemins allant des noeuds terminaux à la racine de l'arbre et les chemins de référence correspondants. L'identification des mots composés est évaluée par la F-mesure

6. Nous avons utilisé la version adaptée au Français pour la gestion des mots inconnus qui se trouve dans le logiciel Bonsai (Candito et Crabbé, 2009) : `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`.

7. Les grammaires sont apprises avec 6 cycles et une graine aléatoire de 8.

8. 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.

9. On convertit d'abord automatiquement les analyses en consituants, en analyses en dépendances au moyen du logiciel *Bonsai*. Puis la mesure est calculée avec l'outil disponible à `http://ilk.uvt.nl/conll/software.html`.

10. Nous utilisons l'outil disponible à `http://www.grsampson.net/Resources.html`

F(EMM) associée aux noeuds de type EMM. La significativité statistique entre deux expériences d'analyse syntaxique est calculée au moyen du t-test unidirectionnel pour deux échantillons indépendants [11]. La significativité statistique entre deux expériences d'identification de mots composés est établie par le test de McNemar (Gillick et Cox, 1989). Les résultats de deux expériences sont considérés comme statistiquement significatifs si la valeur calculée lors du test est inférieure à 0.01.

6.2 Analyse syntaxique avec pré-identification des mots composés

Nous avons tout d'abord testé l'analyseur BKYc prenant en entrée un texte segmenté par notre reconnaisseur CRF de mots composés (sans les étiquettes). Ce dernier se base sur le logiciel *Wapiti* [12] (Lavergne *et al.*, 2010) qui apprend et applique le modèle CRF. Le logiciel *Unitex* est utilisé pour appliquer les ressources lexicales. L'étiqueteur morphosyntaxique *lgtagger* [13] sert à extraire les traits liés aux *n*-grammes de catégories grammaticales. Notre reconnaisseur intégrant tous les traits atteint 75.9% de F(EMM) sur FTB-P7 (79.1% sans tenir compte des étiquettes). Il est, en pratique, meilleur que celui proposé par (Green *et al.*, 2011) qui a une F(EMM) de 71.1% sur les phrases inférieures à 40 mots de FTB-STF [14] : notre reconnaisseur atteint, sur ce même corpus, 74% pour les traits endogènes (soit un gain absolu de +2.9%) et 77.3% pour tous les traits (soit un gain absolu de +6.2%).

Pour rendre comparables les analyses générées par BKYc couplé au reconnaisseur, avec celles de l'analyseur BKY, nous avons automatiquement transformé les analyses avec mots composés fusionnés en leurs analyses équivalentes avec des noeuds non-terminaux spécifiques pour les unités polylexicales. Les catégories grammaticales des composants internes ont été déterminées à l'aide de l'étiqueteur morphosyntaxique *lgtagger* appris sur notre corpus d'apprentissage sans intégrer de ressources lexicales externes. Les résultats sont synthétisées dans la table 3.

Traits	F	UAS	LA	F(EMM)
BKY	81.13	83.88	92.96	69.3
-	75.85	77.68	91.42	0.0
endo	81.07*	85.01	93.10	73.5
exo+endo	81.14*	85.22	93.11	75.3
gold	84.17	91.29	94.05	93.2

TABLE 2 – Intégration des mots composés dans l'analyse syntaxique par identification préalable. *endo* et *exo* indiquent que le modèle CRF incorpore respectivement les traits endogènes et les traits exogènes. *gold* signifie que la segmentation lexicale avant analyse syntaxique est parfaite. * indique que le résultat n'est pas significatif par rapport à BKY. Les tests sont réalisés sur FTB-P7.

Les résultats montrent un très grand écart de performance entre un analyseur ne tenant pas compte des mots composés [trait -] et un analyseur avec une segmentation lexicale parfaite [gold] : on a $\Delta F = 8.32$, $\Delta UAS = 13.61$, $\Delta LA = 2.69$ et $\Delta F(EMM) = 93.2$. L'analyseur *baseline* BKY permet, en partie, de combler cet écart : 63% de ΔF, 46% de ΔUAS, 59% de ΔLA et 74% de $\Delta F(EMM)$. On constate qu'une reconnaissance préalable des mots composés n'améliore pas

11. Nous utilisons l'outil de Dan Bikel disponible à `http://www.cis.upenn.edu/~dbikel/software.html`.
12. Wapiti est disponible à `http://wapiti.limsi.fr/`. Nous l'avons configuré de la manière suivante : algorithme 'rprop' et valeurs par défaut pour les pénalités L1 et L2, ainsi que le critère d'arrêt.
13. Disponible à `http://igm.univ-mlv.fr/~mconstan/research/software/`
14. (Green *et al.*, 2011) ont évalué leur système sur les phrases inférieures à 40 mots uniquement.

le parenthésage général des analyses (F-mesure) [15]. Par contre, on observe un gain significatif de +1.34 en UAS, soit une réduction relative de 18% de l'écart avec l'analyseur gold pour le système intégrant tous les traits. On remarque également une amélioration significative de la reconnaissance des mots composés de +6.0 en F(EMM), soit une réduction relative de l'écart de +25%. Si l'on analyse les résultats de F-mesure par catégorie, on s'aperçoit que le pré-repérage des EMMs provoque des effets de bord sur les constituants supérieurs comme les relatives et les subordonnées, et même les groupes nominaux. L'un des principaux problèmes vient de l'identification des verbes composés à l'indicatif ou au subjonctif qui est dramatique (F-mesure de l'ordre de 20%). Dans une moindre mesure, le repérage des noms communs composés et des conjonctions de subordination composées pose également des problèmes.

6.3 Analyse syntaxique avec réordonnancement

Nous avons ensuite évalué l'intégration d'un réordonnanceur après l'analyseur BKY. Comme dans (Charniak et Johnson, 2005), le réordonnanceur se base sur un modèle maximum d'entropie dont les paramètres sont déterminés par un algorithme d'optimisation de second ordre appelé Limited Memory Variable Metric. Concrètement, nous utilisons une implémentation de cet algorithme disponible dans les bibliotèques mathématiques PETSc [16] et TAO42 [17]. Dans un premier temps, nous avons appliqué un modèle incorporant uniquement les traits dédiés aux mots composés (cf. section 5). Nous avons ensuite comparé avec un modèle intégrant aussi les traits généraux décrits dans (Charniak et Johnson, 2005) ou (Collins, 20O0) par les patrons suivants : *Rule*, *Word*, *Heavy*, *HeadTree*, *Bigrams*, *Trigrams*, *Edges*, *WordEdges*, *Heads*, *WProj*, *NGramTree* et *Score*. Pour chaque expérience, le réordonnanceur prend, en entrée, les 50 meilleures analyses de Berkeley. Les résultats sont synthétisés dans la table 3.

Analyseur	Traits	F	UAS	LA	F(EMM)
BKY	-	81.13	83.88	92.96	69.3
BKY	endo	81.35^{*}	84.48^{+}	93.03	70.7^{+}
BKY	endo+exo	81.64	84.98	93.12	74.2
BKY	std	81.98	84.40	93.41	70.8
BKY	tous	82.05^{+}	84.45^{+}	93.42	70.2^{+}
BKYc^{+}	std	81.66^{+}	85.70	93.41	74.8

TABLE 3 – Intégration d'un réordonnanceur dans l'analyse syntaxique. Les notations *std* et *tous* correspondent respectivement aux traits généraux et à tous les traits décrits. BKYc^{+} correspond à l'analyseur BKYc couplé au reconnaisseur de mots composés avec tous les traits endogènes et exogènes. * et $^{+}$ indiquent que le résultat n'est pas significatif respectivement par rapport à l'analyseur baseline BKY et à l'analyseur BKY couplé au réordonnanceur avec les traits *std*. Les tests sont réalisés sur FTB-P7.

L'utilisation de tous les traits dédiés aux mots composés permet d'augmenter toutes les mesures par rapport à BKY : +0.51 en F, +1.10 en UAS, +0.16 en LA et +4.9 en F(EMM). Sur la reconnaissance des mots composés, on constate une relative faiblesse par rapport à la méthode par pré-identification : en analysant les analyses oracles selon F, on s'aperçoit que F(EMM) a un

15. Ces résultats sont cependant à mettre en perspective par rapport aux résultats sur le corpus de dévelopement où l'on observe des gains absolus significatifs : entre +0.2 et +0.7.
16. `http://www.mcs.anl.gov/petsc/`.
17. `http://www.mcs.anl.gov/research/projects/tao/`.

potentiel maximum de 76.9% ce qui n'est pas très élevé. Par ailleurs, les traits généraux seuls sont plus efficaces que les traits dédiés aux mots composés pour ce qui concerne le parenthésage (81.98% vs. 81.64%) et le LA (93.41% vs. 93.12%). Par contre, ils dégradent l'UAS (84.40% vs. 84.98%) et la reconnaissance des mots composés (70.8% vs. 74.2%) Le mélange des deux types de traits (*tous*) n'est pas très concluant car on n'observe aucune variation significative de l'écart par rapport à l'analyseur avec les traits généraux, quelle que soit la mesure. Ces résultats montrent qu'il est nécessaire de trouver un autre moyen de combiner ces deux types de traits.

7 Conclusions et Perspectives

Dans cet article, nous avons évalué deux stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un système d'analyse syntaxique probabiliste : pré-identification *état-de-l'art* des mots composés ; repérage final des mots composés après réordonnancement des *n* meilleures analyses. Les différents modèles comprennaient des traits spécifiques aux unités polylexicales. Nous avons montré que le pré-repérage permettait d'améliorer grandement la reconnaissance des mots composés et la qualité des liens de dépendance non typés, alors que la F-mesure tend à se stabiliser. Le réordonnanceur augmente légèrement tous les scores, mais déçoit en terme d'identification de mots composés par rapport à la première méthode. Par ailleurs, l'intégration des traits généraux de (Charniak et Johnson, 2005) rend caducs les traits dédiés aux unités polylexicales et dégrade la qualité des liens de dépendance non typés. Il semble qu'aucune des deux méthodes ne soit entièrement satisfaisante. Mais ces expériences ouvrent de nouvelles perspectives intéressantes. Nous pourrions combiner efficacement ces deux stratégies en permettant au pre-segmenteur de générer l'automate pondéré des segmentations lexicales possibles et de combiner ce dernier avec l'analyseur syntaxique. Nous pourrions également transposer ces deux solutions à l'analyse en dépendance.

Remerciements

Nous souhaitons remercier Marie Candito et Spence Green pour nous avoir mis à disposition leurs versions du corpus arboré de Paris 7.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

Arun, A. et Keller, F. (2005). Lexicalization in crosslinguistic probabilistic parsing : The case of french. *In Actes de ACL*.

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

Candito, M. H. et Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. *In Proceedings of IWPT 2009*.

CHARNIAK, E. et JOHNSON, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

COLLINS, M. (20O0). Discriminative reranking for natural language parsing. *In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*.

CONSTANT, M. et SIGOGNE, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. *In Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World (MWE'11)*.

CONSTANT, M., TELLIER, I., DUCHIER, D., DUPONT, Y., SIGOGNE, A. et BILLOT, S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de Conférence sur le traitement automatique des langues naturelles (TALN'11)*.

COURTOIS, B. (2009). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87:1941 – 1947.

COURTOIS, B., GARRIGUES, M., GROSS, G., GROSS, M., JUNG, R., MATHIEU-COLAS, M., MONCEAUX, A., PONCET-MONTANGE, A., SILBERZTEIN, M. et VIVÉS, R. (1997). Dictionnaire électronique DELAC : les mots composés binaires. Rapport technique 56, University Paris 7, LADL.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*.

ERYIGIT, G., ILBAY, T. et ARKAN CAN, O. (2011). Multiword expressions in statistical dependency parsing. *In Proceedings of the IWPT Workshop on Statistical Parsing of Morphologically-Rich Languages (SPRML'11)*.

FINKEL, J. R. et MANNING, C. D. (2009). Joint parsing and named entity recognition. *In Proceedings of NAACL*.

GILLICK, L. et COX, S. (1989). Some statistical issues in the comparison of speech recognition algorithms. *In Proceedings of ICASSP'89*.

GREEN, S., de MARNEFFE, M.-C., BAUER, J. et MANNING, C. D. (2011). Multiword expression identification with tree substitution grammars : A parsing tour de force with french. *In Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11)*.

GROSS, M. (1986). Lexicon grammar. the representation of compound words. *In Proceedings of Computational Linguistics (COLING'86)*.

HOGAN, D., FOSTER, J. et van GENABITH, J. (2011). Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities. *In Proceedings of ACL Workshop Multiword Expressions : from Parsing and Generation to the Real World (MWE'2011)*.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

NIVRE, J. et NILSSON, J. (2004). Multiword units in syntactic parsing. *In Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.

PETROV, S., BARRETT, L., THIBAUX, R. et KLEIN, D. (2006). Learning accurate, compact and interpretable tree annotation. *In Proceedings of ACL*.

RAMISCH, C., VILLAVICENCIO, A. et BOITET, C. (2010). mwe-toolkit : a framework for multiword expression identification. *In Proceedings of LREC*.

RAMSHAW, L. A. et MARCUS, M. P. (1995). Text chunking using transformation-based learning. *In Proceedings of the 3rd Workshop on Very Large Corpora*, pages 88 – 94.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A. et FLICKINGER, D. (2002). Multiword expressions : A pain in the neck for nlp. *In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*, pages 1–15, London, UK. Springer-Verlag.

SAGOT, B. (2010). The lefff, a freely available, accurate and large-coverage lexicon for french. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

SAMPSON, G. et BABARCZY, A. (2003). A test of the leaf-ancestor metric for parsing accuracy. *Natural Language Engineering*, 9(4).

SEDDAH, D., CANDITO, M.-H. et CRABBÉ, B. (2009). Cross-parser evaluation and tagset variation : a french treebank study. *In Proceedings of International Workshop on Parsing Technologies (IWPT'09)*.

SILBERZTEIN, M. (2000). Intex : an fst toolbox. *Theoretical Computer Science*, 231(1):33–46.

TELLIER, I. et TOMMASI, M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. *In* Eric GAUSSIER et François YVON, éditeurs : *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès.

WATRIN, P. et FRANÇOIS, T. (2011). N-gram frequency database reference to handle mwe extraction in nlp applications. *In Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World (MWE'11)*.

WEHRLI, E., SERETAN, V. et NERIMA, L. (2010). Sentence analysis and collocation identification. *In Proceedings of the Workshop on Multiword Expression : From Theory to Applications (MWE'10)*.