Améliorer l'étiquetage de "que" par les descripteurs ciblés et les règles

Assaf Urieli[1, 2]
(1) CLLE-ERSS: CNRS & Université de Toulouse, Toulouse, France
(2) Joliciel Informatique SARL, 2 avenue du Cardié, 09000 Foix, France
assaf.urieli@univ-tlse2.fr

Résumé. Les outils TAL statistiques robustes, et en particulier les étiqueteurs morphosyntaxiques, utilisent souvent des descripteurs "pauvres", qui peuvent être appliqués facilement à n'importe quelle langue, mais qui ne regarde pas plus loin que 1 ou 2 tokens à droite et à gauche et ne prennent pas en compte des classes d'équivalence syntaxiques. Bien que l'étiquetage morphosyntaxique atteint des niveaux élevés d'exactitude (autour de 97 %), les 3 % d'erreurs qui subsistent induisent systématiquement une baisse de 3 % dans l'exactitude du parseur. Parmi les phénomènes les plus faciles à cibler à l'aide de l'injection de connaissances linguistiques plus riches sont les mots fonctionnels ambigus, tels que le mot "que" en français. Dans cette étude, nous cherchons à améliorer l'étiquetage morphosyntaxique de "que" par l'utilisation de descripteurs ciblés et riches lors de l'entraînement, et par l'utilisation de règles symboliques qui contournent le modèle statistique lors de l'analyse. Nous atteignons une réduction du taux d'erreur de 45 % par les descripteurs riches, et de 55 % si on ajoute des règles.

Abstract. Robust statistical NLP tools, and in particular pos-taggers, often use knowledge-poor features, which are easily applicable to any language but do not look beyond 1 or 2 tokens to the right and left and do not make use of syntactic equivalence classes. Although pos-tagging tends to get high accuracy scores (around 97%), the remaining 3% errors systematically result in a 3% loss in parsing accuracy. Some of the easiest phenomena to target via the injection of richer linguistic knowledge are ambiguous function words, such as "que" in French. In this study, we attempt to improve the pos-tagging of "que" through the use of targeted knowledge-rich features during training, and symbolic rules which override the statistical model during analysis. We reduce the error rate by 45% using targeted knowledge-rich features, and 55% if we add rules.

Mots-clés : étiquetage morphosyntaxique, apprentissage automatique supervisé, descripteurs riches, systèmes statistiques robustes.

Keywords: pos-tagging, supervised machine learning, knowledge-rich features, robust statistical systems.

1 Introduction

Les outils TAL statistiques robustes sont relativement faciles à construire : il suffit de disposer d'un corpus d'apprentissage annoté, d'un classifieur robuste (ex. SVM linéaire), d'un algorithme d'analyse (ex. le parsing par transitions pour l'analyse syntaxique) et de quelques descripteurs. La plupart de ces systèmes utilisent des descripteurs linguistiquement pauvres, limités aux bigrammes ou trigrammes des tokens ou des étiquettes morphosyntaxiques, à quelques informations de base tirées d'un lexique à large couverture, et, au niveau du parsing, à un examen superficiel de la tête ou du dépendant le plus à droite ou à gauche d'un token donné. Même les études qui parlent de descripteurs "riches" (Zhang & Nivre, 2011) se limitent à des descripteurs génériques, qui prennent en compte des informations de surface telles que la valence d'un token (nombre de dépendants) ou la distance entre deux tokens, mais ne cherchent pas à coder les phénomènes spécifiques d'une langue donnée. Cela présente l'avantage d'une application facile à beaucoup de langues, mais nous empêche d'injecter des connaissances linguistiques spécifiques, et limite donc les gains d'exactitude possibles. Notre but principal ici est de trouver des moyens d'améliorer les analyses des systèmes statistiques par l'introduction d'informations plus riches.

L'analyseur syntaxique Talismane [1] a été développé dans l'optique de permettre à l'utilisateur d'injecter le maximum d'informations linguistiques, dans un système qui reste statistique et robuste (Urieli, 2013). Il comprend quatre modules

1. http://redac.univ-tlse2.fr/applications/talismane.html

statistiques enchaînés : la segmentation en phrases, la segmentation en mots (tokenisation), l'étiquetage morphosyntaxique (pos-tagging) et l'analyse syntaxique en dépendances par transitions (parsing), dont l'algorithme de base est décrit dans Kübler *et al.* (2009). Nous nous sommes intéressés tout particulièrement à l'interaction entre les différents modules. Dans une étude précédente, nous avons exploré la propagation des ambiguïtés de l'étiqueteur morphosyntaxique vers le parseur, afin que ce dernier puisse les corriger (Urieli & Tanguy, 2013). Dans cette étude, nous cherchons plutôt à améliorer l'étiquetage morphosyntaxique en amont du parseur, par l'injection des connaissances linguistiques spécifiques pour certains phénomènes particulièrement importants pour le parsing, se concentrant ici sur l'étiquetage du mot *que*.

En effet, nous avons remarqué que des erreurs d'étiquetage de certains mots fonctionnels ambigus induisent systématiquement de multiples erreurs de parsing. Le cas de *que* est particulièrement intéressant car très ambigu, mais la méthodologie présentée ici pourrait être appliquée à d'autres mots fonctionnels ainsi qu'à d'autres classes de mots facilement identifiables (ex. les nombres cardinaux). Nous examinons ici l'injection des connaissances linguistiques par deux moyens complémentaires : l'ajout de descripteurs riches lors de l'entraînement, et l'ajout de règles symboliques lors de l'analyse, qui imposent ou interdisent des décisions locales, contournant ainsi le modèle statistique. Notre approche ici, de correction de phénomènes spécifiques par l'injection d'informations symboliques, est similaire à certaines études précédentes, telles que Danlos (2005) pour le *il* impersonnel, et Jacques (2005) pour *que*. A la différence de ces études, qui considèrent des systèmes uniquement à base de règles, nous mettons l'accent ici sur les descripteurs riches, qui s'insèrent naturellement dans un système statistique robuste. Les règles symboliques sont utilisées uniquement en complément des descripteurs, pour des cas très précis et non ambigus.

2 Etiquetage morphosyntaxique

Dans Talismane, l'algorithme d'étiquetage morphosyntaxique fonctionne de gauche à droite. Ainsi, les descripteurs peuvent prendre en compte tous les tokens qui se trouvent à gauche et à droite du token à étiqueter, ainsi que les étiquettes déjà attribuées à sa gauche. Comme descripteurs de base, nous utilisons des descripteurs similaires à ceux décrits par Denis & Sagot (2012), faisant un usage massif d'un lexique, en l'occurrence le LeFFF (Sagot, 2010). En particulier, nous utilisons les descripteurs de base suivants : *W* la forme lexicale exacte, *P* l'étiquette attribuée au token (si son index < celui du token actuel) ou les étiquettes trouvées dans le lexique pour ce token (si son index $\geq$ celui du token actuel), *L* le lemme de ce token, pour une étiquette donnée, *U* si le token est inconnu dans le lexique, Sfx_n les n dernières lettres de la forme, $Pref_n$ les n premières lettres de la forme, *1st* si le token est le premier de la phrase, $Last$ si le token est le dernier de la phrase. Ces briques de base sont combinées en bigrammes et trigrammes pour les tokens à position -2, -1, 0, +1, +2 par rapport au token actuel. En vue de ce jeu de descripteurs, un descripteur plus "riche" est n'importe quel descripteur qui regarde plus loin que 2 tokens à gauche ou à droite, ou qui regroupe les tokens en classes d'équivalence à un niveau qui se trouve entre le lemme et l'étiquette morphosyntaxique. Dans la pratique, nous avons utilisé des descripteurs bien plus sophistiqués (décrits ci-après), qui mettent en oeuvre des combinaisons logiques complexes des informations de base.

Pour cette étude, notre corpus d'apprentissage est la partie française du corpus SPMRL (Seddah *et al.*, 2013), un corpus disponible en dépendances et construit à partir du French Treebank (FTB) (Abeillé *et al.*, 2003). Nous avons appliqué un pré-traitement aux mots composés, conservant uniquement les mots composés qui ne représentent pas une régularité syntaxique. Pour le corpus d'évaluation, en plus des parties *dev* et *test* du corpus SPMRL, nous utilisons les corpus Sequoia (Candito *et al.*, 2012) et un corpus des pages de discussion du Wikipedia français, FrWikiDisc, décrit dans Urieli (2013). Nous utilisons le jeu d'étiquettes décrit dans Crabbé & Candito (2008). Pour un modèle SVM linéaire construit avec les descripteurs ci-dessus, $\epsilon = 0,01$, $C = 0.5$ et un cutoff de 3 (nombre de fois qu'un descripteur doit apparaître pour être pris en compte), on a une exactitude de 96,58 sur SPMRL-dev, et 96,55 sur SPMRL-test. Toutes les données et les modèles sont disponibles sur simple demande, à l'exception du SPMRL français, qui nécessite un accord préalable.

3 Le cas de "que"

Les difficultés d'étiqueter le mot *que* ont déjà été explorées dans d'autres études, en particulier Jacques (2005), qui décrit les différents contextes dans lequel *que* est utilisé, et qui propose une méthode pour corriger l'étiquetage par un mélange de règles de surface et de corrections appliquées pendant l'analyse syntaxique.

Pour résumer, il y a six options principales pour le token *que* (et son équivalent abrégé *qu'*), annotées selon les normes d'annotation du corpus FTB avec 4 étiquettes différentes, comme illustré par les exemples suivants :

1. Conjonction de subordination (CS) : *Je pense qu'il a trop bu.*
2. Pronom relatif (PROREL) : *Il boit le vin que j'ai acheté.*
3. Pronom interrogatif (PROWH) : *Que buvez-vous ?*
4. Adverbe négatif (ADV) : *Je n'ai bu que trois verres.*
5. Adverbe exclamatif (ADV) : *Qu'il est bon, ce vin !*
6. Construction comparatif (CS) : *Il est plus bourré que moi.*

Le *que* d'une clivée est étiqueté PROREL pour un focus nominal argument du verbe. Quand le focus est un syntagme prépositionnel ou nominal circonstant, l'étiquetage du FTB est assez incohérent entre PROREL et CS.

	ADV	**CS**	**PROREL**	**PROWH**	**Total**	**Erreurs**
ADV	90	44	4	1	139	49
CS	37	1097	61	0	1195	98
PROREL	0	69	244	0	313	69
PROWH	0	4	2	23	29	6

TABLE 1 – Matrice de confusion de base pour *que*

Avec le modèle de base décrit dans le paragraphe précédent, la table 1 montre la matrice de confusion pour le mot *que* dans l'ensemble des corpus d'évaluation, où les lignes représentent la bonne étiquette et les colonnes représentent l'étiquette devinée. Nous avons donc au total 222 erreurs pour 1 676 occurrences, donc une exactitude de 86,75 %. Il est à noter que la confusion se trouve principalement entre CS and ADV d'une part, et entre CS et PROREL d'autre part. Nous traiterons, ci-dessous, chacun de ces cas séparément.

4 Des descripteurs ou des règles ?

Un **descripteur** (*feature* en anglais) spécifie l'information à extraire d'un contexte donné, qui pourra aider le classifieur à choisir la bonne étiquette du token dans ce contexte. Dans Talismane, un descripteur est défini par une expression qui combine des informations de base, soit par concatenation (pour les chaînes de caractères), soit par des opérations mathématiques (pour les nombres) ou logiques (pour les résultats booléens de type vrai/faux). Une **règle** est une expression booléene définie avec la même grammaire que les descripteurs. Si l'expression s'évalue à *vrai* dans un contexte donné, la règle peut soit imposer le choix d'une certaine étiquette pendant l'analyse, soit empêcher le système de choisir cette étiquette.

Par exemple, l'étiquette attribuée au mot précédant le *que* peut être utilisée comme descripteur. Si ce mot est un verbe à l'indicatif (V), e.g. *"il **faut** que…"*, alors on observe certaines tendances sur l'étiquetage du *que* : dans notre corpus d'apprentissage, sur 523 cas, 77 % sont des CS et 23 % des ADV. Un autre descripteur peut porter sur les étiquettes possibles du mot suivant le *que* dans un lexique externe de référence. Si ce mot est listé dans le lexique comme verbe à l'indicatif, e.g. *"l'exemple que **fournit** Dupont…"*, alors dans notre corpus d'apprentissage, sur 205 cas, 1 % des *que* sont des ADV, 20 % des CS, 72 % des PROREL et 7 % des PROWH. Ces informations vont être combinées avec des dizaines d'autres descripteurs pour aider le modèle probabiliste à construire une distribution de probabilités des étiquettes pour un cas donné de *que*.

On peut aussi être amené à définir une règle déterministe : e.g. si on a une structure de type *"ne V que"*, alors on oblige le système à attribuer l'étiquette ADV. Cette règle prend priorité sur le modèle probabiliste qui ne sera même pas consulté. Un autre type de règle est la règle négative : e.g. si *que* est le premier mot d'une phrase, alors on empêche le système d'attribuer l'étiquette PROREL. Dans ce cas, le modèle probabiliste va utiliser tous les descripteurs pour définir une distribution de probabilités des étiquettes, mais l'étiquette PROREL sera supprimée de cette distribution avant que le système ne choisisse l'étiquette la plus probable.

Un descripteur cherche, par nature, à capter des régularités dans le corpus d'apprentissage qui peuvent se généraliser à d'autres corpus. Il est donc limité aux régularités qui se trouvent dans ce corpus, même si elles peuvent être décrites à l'aide de ressources externes pour les rendre plus généralisables (ex. un lexique qui remplace la forme lexicale par son lemme). Par contre, une règle cherche à traduire directement les connaissances linguistiques du concepteur du système, surtout pour des phénomènes sous représentés dans le corpus d'apprentissage. Elle permet donc au système statistique

d'aller au delà des informations qui lui sont directement accessibles. Puisque celle-ci est appliquée uniquement au moment de l'analyse, elle peut aussi traduire des connaissances spécifiques au corpus qu'on est en train d'analyser.

Les descripteurs serviront à alimenter le classifieur (ex. SVM), qui va appliquer sa "magie noire" statistique pour donner plus ou moins de poids à chaque descripteur pour chaque étiquette, selon les occurrences trouvées dans le corpus d'apprentissage. Les descripteurs peuvent donc se contredire et se chevaucher. Ils décrivent des tendances : si X est vrai, alors l'étiquette sera plus probablement Y que Y'. Les règles cherchent, par contre, à décrire des vérités absolues : si X est vrai, alors l'étiquette doit être (ou ne peut pas être) Y. Les descripteurs sont par conséquent beaucoup plus puissants que les règles, car à la différence de celles-ci, ils ne sont pas contraints à viser un phénomène très spécifique et non ambigu. Néanmoins, vu le coût de construction d'un corpus annoté, les corpus sont forcément très lacunaires en informations. Ce sont ces informations que les règles vont cibler.

5 "Que" comme adverbe négatif

En termes de descripteurs ciblés, nous traiterons d'abord le cas de *que* en tant qu'adverbe négatif. Notre méthodologie itérative consiste à :

1. Analyser les erreurs dans le corpus *dev* et concevoir des descripteurs utiles.
2. Écrire ces descripteurs dans la syntaxe de Talismane.
3. Projeter ces descripteurs sur le corpus *train*, et examiner les co-occurrences avec chaque étiquette, surtout celles avec une étiquette inattendue. Nous cherchons à inclure le maximum de résultats tout en maximisant le déséquilibre entre les étiquettes. Revenir à l'étape 2 si nécessaire.
4. Entraîner le modèle avec les descripteurs ciblés, et évaluer. Revenir à l'étape 1 si nécessaire.

5.1 Analyse d'erreurs : adverbe négatif

Dans le corpus SPMRL *dev*, la plupart des erreurs ressemblent au cas suivant , ou *que* est étiquetté à tort comme `CS` :

Exemple 5.1 *Mais cela **ne** représente dans cette mouture, pour un couple avec deux enfants, **qu'**/ADV une prime maximale.*

Dans ce cas, reconnaître *que* comme adverbe négatif revient à chercher une occurrence de *ne* plus tôt dans la même phrase. Il n'y a pas de limitation inhérente de distance car, comme on voit dans l'exemple 5.1, plusieurs syntagmes prépositionnels peuvent séparer les deux particules. Par contre, une autre particule négative peut compléter le *ne*, ce qui rend le *que* ambigu, comme dans les deux exemples suivants :

Exemple 5.2 *Pour cela, il **n'**est **pas** question **que**/CS le zloty, la monnaie polonaise, soit "l'ancre de la stabilité" de l'économie polonaise.*

Exemple 5.3 *…qui, faute de volonté politique, **ne** fut **jamais que**/ADV la caricature du système français.*

Cette ambiguïté existe uniquement pour les verbes qui sous catégorisent un objet direct en *que*, tels que *dire* ou *penser*. Cependant, le corpus *train* contient 245 verbes différents qui répondent à ce critère. Les cas ambigus où le *que* suit une autre particule négative étant assez rares, nous avons décidé de ne pas utiliser la sous catégorisation dans nos descripteurs.

5.2 Liste de descripteurs : adverbe négatif

La prochaine étape consiste à écrire ces descripteurs dans la syntaxe de Talismane, et les projeter sur le corpus *train*. Après affinage pour étendre la portée des descripteurs tout en éliminant des cas non voulus, nous avons défini la liste suivante. Les nombres représentent le nombre d'occurrences dans le corpus *train*.

Descripteur 5.1 ***Ne* précédent sans autre particule négative** : le *que* est précédé par un *ne* sans autre particule négative entre les deux. En plus le *ne* n'est pas lui-même précédé par {*personne*, *rien*, *aucun/e*, *nul/le*}, afin d'exclure des phrases comme *"Personne ne sait que je mange ici."*.

Nous avons 345 cas en tout, dont 312 `ADV` : *"Ils **n'**en comprendront le sens **que**/ADV bien plus tard"* ; et 32 `CS`. Parmi les `CS`, il y a beaucoup d'erreurs d'annotation—les autres sont des phrases où la particule *ne* n'est pas completé par une autre particule négative, dans des expressions de type *moins ADJ qu'on ne…* : *"L'Amérique, moins superficielle qu'on **ne** l'imagine parfois, a entrepris une réflexion sur son identité bien avant **que**/CS […]"* ; ou en modifiant le verbe *pouvoir* : *"[…] **ne** peuvent ainsi éviter **que**/CS, en la matière, l'histoire ne se repète"*.

Descripteur 5.2 Pas de *ne* précédent : il n'y a pas de *ne* précédent le *que*.

Nous avons 2608 cas en tout, dont 1941 `CS`, 622 `PROREL`, 26 `PROWH` et 19 `ADV`, dont 10 sont des erreurs d'annotation, où un *que* comparatif est annoté comme adverbe, 5 sont des adverbes exclamatifs : *"Mais pour parvenir à cela, **que**/ADV d'esprits à convaincre en France et plus encore au-dehors !"* ; et 1 est une phrase "informelle" ou l'auteur a laissé tomber le *ne* : *"Il lui manque **que**/ADV le sac à main de Maggie"*.

Nous avons ajoutés deux descripteurs supplémentaires pour aider l'étiqueteur dans des cas où il y a une autre particule négative entre le *ne* et le *que* :

Descripteur 5.3 ***Que* négatif possible** : y a-t-il un *ne*, *ne pas* ou *ne plus* plus tôt dans la phrase, sans prendre en compte d'autres particules négatives. Ce descripteur couvre tous les cas où un *que* négatif est possible. C'est une version moins exclusive du descripteur 5.1. On trouve 363 `ADV`, 218 `CS` et 47 `PROREL`. Pour ces deux derniers, la grande majorité sont des cas où le *ne* est complété par une autre particule négative.

Descripteur 5.4 Combinaison de particules négatives à courte distance : nous avons remarqué que, dans le corpus d'apprentissage, le *que* se combine avec une autre particule négative uniquement si leur distance est petite (≤ 6 tokens). Ce descripteur s'évalue donc à `vrai` si la distance est ≤ 6, à `faux` si la distance est plus grande, et à rien du tout s'il n'y a pas de particule négative entre le *ne* et le *que*.

Pour la distance courte, sur 184 cas, nous avons 119 `CS`, 17 `PROREL` et 47 `ADV`, ce qui représente plus d'un quart des cas : *"Les spéculateurs sont désormais certains que la dévaluation **n'**est **plus qu'**/ADV une question de jours"*. Pour la distance longue, sur 146 cas, nous avons 101 `CS` : *"Si cela **n'**était **pas** possible, les Onze poursuivraient leur chemin sans perdre l'espoir **que**/CS cela se ferait plus tard"* ; 43 `PROREL` et uniquement 2 `ADV` : *"Il **ne** restait **plus** au président du groupe socialiste de l'Assemblée nationale, dans ces conditions, **qu'**/ADV à négocier la fusion de son texte avec celui de MM Jospin et Delebarre"*.

6 "Que" comme pronom relatif

A la différence du *que* en tant qu'adverbe négatif, où la présence d'un *ne* précédent est un indicateur de surface fort, il n'y a pas d'indicateur de surface simple pour distinguer le *que* pronom relatif du *que* conjonction de subordination, étant donné le peu d'informations disponibles à l'étape de l'étiquetage morphosyntaxique.

6.1 Analyse d'erreurs : pronom relatif

Suivant la méthodologie décrite dans le paragraphe 5, nous analysons les erreurs du corpus *dev* pour identifier des descripteurs utiles.

Exemple 6.1 (…) *la Commission des opérations de bourse (COB) a annoncé le 14 janvier **qu'**/CS elle saisit la justice […]*

Cet exemple est annoté PROREL plutôt que CS. Certains descripteurs sautent aux yeux : d'abord *annoncer* est parmi les verbes qui sous catégorisent un objet direct avec *que*. De plus, le verbe transitif *saisir* a déjà un objet direct (*justice*), ce qui exclut généralement un pronom relatif. Finalement, noter que l'ambiguïté entre PROREL et CS existe uniquement quand il y a un nom qui peut servir d'antécédent entre le verbe précédent et le *que*, dans ce cas *janvier*. La nature de ce nom est un indicateur : les expressions de temps, dont les noms des mois, sont très souvent des circonstants. Ils remplissent rarement l'argument d'objet direct, et sont rarement modifiés par une proposition relative.

Exemple 6.2 *Le gouvernement va présenter dans un délai de trois mois les dispositions **qu'**/PROREL il entend retenir [...]*

Cet exemple est annoté CS plutôt que PROREL. C'est le cas contraire de l'exemple précédent : le verbe *présenter* a déjà un objet direct (*dispositions*) et ne sous catégorise pas un objet direct avec *que*, alors que le verbe transitif *retenir* n'a pas d'objet direct qui le suit.

Nous voyons ici l'importance de reconnaître les verbes qui sous catégorisent avec *que*. Comme mentionné précédemment, le corpus *train* en contient 245. Nous avons choisi manuellement 152 de ces verbes qui nous semblaient les plus aptes à préférer cette sous catégorisation.

Exemple 6.3 *Le fait **qu'**/CS ils aient accepté de reprendre les pourparlers est interprété de façon positive.*

Ici nous avons d'autres indicateurs : certains noms introduisent des propositions subordonnées (ex. *fait*), et le subjonctif (*aient*) indique généralement qu'on a affaire à une subordonnée indépendante plutôt que relative.

6.2 Liste de descripteurs : pronom relatif

Après la projection des descripteurs sur le corpus *train* et affinage, nous avons retenu les descripteurs suivants :

Descripteur 6.1 Structure coordonnée : Si le *que* actuel suit directement une conjonction de coordination, chercher l'étiquette du *que* précédent. De même, si le *que* actuel suit une virgule, chercher l'étiquette du *que* précédent, du moment où il existe un *que* plus tard dans la phrase qui suit une conjonction de coordination. Si l'étiquette précédente est CS (total 104 cas), nous avons 102 CS : *"Or chacun est conscient **qu'**/CS il n'y a aucune vérification de ces acquis et **que**/CS la rétribution est automatique"* ; 1 PROREL : *"Encore faudrait-il **que**/CS, pour faire passer la pilule des réformes nécessaires—et **que**/PROREL beaucoup d'Italiens risquent de trouver plus amère que prévu [...]"* ; et 1 *PROWH*. Si l'étiquette précédente est PROREL (total 8 cas), nous avons 2 CS et 6 PROREL.

Descripteur 6.2 Après nom explicatif : Le *que* suit-il un des mots {*assurance, certitude, doute, enseigne, espoir, fait, fois, idée, point, prétexte, preuve, principe*} ? Les cas précédés par la locution *c'est* ont été exclus. Résultat : 38 CS.

Descripteur 6.3 Verbe précédent sous catégorise avec *que* : Chercher le verbe précédent (en faisant attention de sauter les participes passés modifieurs de noms). On s'intéresse uniquement aux cas où il y a un nom entre le verbe et le *que*, qui peut servir d'antécédent. Est-ce que ce verbe sous catégorise avec *que* ? Pour les cas où le verbe précédent sous catégorise avec *que* (total 98 cas), nous avons 50 CS : *"Helmut Kohl a **annoncé** à l'automne **que**/CS des hausses d'impôts seraient nécessaires en 1994"* ; et 48 PROREL : *"Il a toutefois **refusé** à Mr Vernay les 100 000 francs de dommages et intérêts **que**/PROREL celui-ci réclamait"*. Les deux étiquettes sont donc distribuées de façon à peu près égale. Pour le cas contraire (total 126 cas), nous avons 113 PROREL, et uniquement 12 CS, dont 10 erreurs d'annotation.

Descripteur 6.4 Le verbe qui précède a un objet direct : Le verbe précédant le *que* est-il suivi directement d'un déterminant et d'un nom (ou d'un déterminant, d'un adjectif et d'un nom), en dehors des noms représentant les expression de temps (ex. *la semaine dernière*) ? On enlève les cas où le *que* suit directement un nom "explicatif" du descripteur 6.2. Sur 63 cas, nous avons 56 PROREL : *"En revanche, la CGT **dénonce un texte qu'**/PROREL elle juge "décrédibilisé par le manque de moyens""* ; et 7 CS : *"Nous avons **obtenu l'assurance** du premier ministre **que**/CS la suppression du recours [...]"*, dont 6 sont des erreurs d'annotation.

Descripteur 6.5 Le verbe qui précède sous catégorise avec à + personne + que : Le verbe précédent le *que* est-il dans l'ensemble {*annoncer, certifier, ...*} qui sous catégorise des structures comme *"annoncer à ses parents qu'on se marie"*, et est-il suivi de la préposition *à* ? Résultats : 23 CS.

Descripteur 6.6 Le verbe qui précède sous catégorise avec un objet direct + que : Le verbe précédent le *que* est-il dans l'ensemble {*assurer, avertir, ...*} qui sous catégorise des structures comme *"assurer ses parents qu'on se marie"* ? Résultats : 26 CS.

Descripteur 6.7 Le verbe qui suit a un objet direct : Le verbe suivant le *que* est-il suivi directement d'un déterminant et d'un nom (ou d'un déterminant, d'un adjectif et d'un nom), en dehors des noms représentant les expressions de temps ? Résultats : 93 CS.

Descripteur 6.8 *Que* suivi directement d'un verbe : Le *que* est-il suivi directement d'un verbe ? On s'attend ici surtout à des PROREL et des PROWH. Pour 116 cas, nous avons 105 PROREL : *"L'exemple **que/PROREL fournit** Sombart est particulièrement éclairant"* ; 5 PROWH : *"Si vous pouviez changer le monde, **que/PROWH feriez**-vous ?"* ; et 6 CS, tous des verbes intransitifs avec inversion du sujet, dont 4 sont des formes subjonctives du verbe *être* : *"Ne souhaitant **que/CS soit** envisagée l'hypothèse d'une dévaluation du franc [...]"*.

Le même descripteur sans le verbe *être* donne 89 PROREL, 5 PROWH et 2 CS.

Descripteur 6.9 *Que* suivi d'un verbe réfléchi : Le *que* est-il suivi d'un verbe réfléchi à la troisième personne (à l'exception de certains verbes qui prennent un objet direct en plus du clitique réfléchi, tels *"se poser une question"*) ? Résultats : 12 CS.

Descripteur 6.10 *Que* suivi d'un verbe subjonctif : Le *que* est-il suivi d'un verbe d'une forme clairement subjonctive ? Puisqu'on regarde à droite du token actuel, on n'a pas encore les étiquettes morphosyntaxiques, et on compte sur le lexique pour nous indiquer les tokens qui peuvent être des verbes. Du coup, on s'est retrouvé au départ avec le nom *émissions* comme imparfait du subjonctif du verbe *émettre*. Nous avons donc éliminé les cas où le token avait aussi une étiquette non verbale dans le lexique. Résultat : 80 CS : *"Faut-il encore **que/CS** l'ambiance non seulement le **permette** mais aussi le favorise"* ; et 1 PROREL : *"Faut-il en conclure que le mieux **qu'/PROREL** on **puisse** attendre, c'est le chacun-pour-soi ?"*

Descripteur 6.11 Après clivée : Le *que* suit-il l'expression *c'est*, indiquant une clivée potentielle (typiquement annoté par un PROREL) ? Ayant remarqué que plus la locution *c'est* est proche, plus la clivée est probable, nous avons défini 4 descripteurs, qui cherchent le *c'est* à une distance de 5, 10 et 20 tokens avant le *que*, et sans limite. Les résultats sont pour une distance sans limite : 90 CS et 95 PROREL. Pour une distance de 20 : 79 CS et 90 PROREL. Pour une distance de 10 : 56 CS et 82 PROREL. Pour une distance de 5 : 34 CS et 62 PROREL.

Descripteur 6.12 Etre ADJ que : Trouver les structures de type *"il est probable que..."*. Résultats : 47 CS.

Descripteur 6.13 Existence d'un nom antécédent ? Y'a-t-il un nom (en dehors des expressions de temps) entre le verbe qui précède et le *que* ? Si le *que* est clairement un deuxième coordonné, on cherche avant le premier coordonné. Ici on considère les réponses à la fois positives et négatives. Résultats positifs (cas classique d'ambiguïté CS/PROREL) : 485 CS et 297 PROREL. Résultats négatifs : 1217 CS et 7 PROREL (dont la plupart sont des erreurs dans la reconnaissance du premier coordonné).

Descripteur 6.14 Expression de temps ? Le nom le plus proche avant le *que* fait-il partie d'une expression de temps (ex. *la semaine dernière*) ? Résultats : 53 CS : *"M. Tchechinski a estimé **dimanche que/CS** la Russie ne manquerait pas de pain cet hiver"* ; et 1 PROREL : *"Voici la **nuit que/PROREL** nous avons attendue toute l'année"*.

Descripteur 6.15 Phrase débutant par *que* : la phrase commence-t-elle par *que* (eventuellement précédé d'une conjonction de coordination) ? Sur 51 cas, nous avons 4 ADV, tous exclamatifs : *"**Qu'/ADV** il était insouciant, le mois de janvier 1992"* ; 28 CS, presque tous des phrases incomplètes : *"**Qu'/CS** ils sont "prescripteurs", comme disent les professionnels"* ; 1 PROREL dans une phrase incomplète : *"**Qu'/PROREL** elle n'aime guère voir rapprocher de celui des sociétés concurrentes"* ; et 18 PROWH : *"**Que/PROWH** fait-il de l'excédent de ses revenus ?"*

Descripteur 6.16 Pas de verbe avant le *que* : Si le *que* suit un nom, et qu'il n'y a pas de verbe plus tôt dans la phrase. On exclut les mots explicatifs ci-dessus. Résultats : 100 PROREL : *"Une puissance **qu'/PROREL** elle n'entend partager avec nul autre"* ; et 8 CS : *"Mais quelle prestance **que/CS** celle de l'homme-terminal !"*

Descripteur 6.17 Plus que : Le mot *que* suit-il un quantifieur {*plus, moins, davantage, autant, différent, même, tel, ...*} ? Résultats : 118 CS et 4 PROREL (tous des erreurs d'étiquetage avec *tel que*).

Descripteur 6.18 Plus ADJ que : Le mot *que* suit-il un adverbe comparatif {*plus, moins, davantage, autant, ...*} et un adjectif ? Résultats : 139 CS : *"Cela est **plus motivant que/CS** d'avoir des salariés sous-employés en temps et en compétences"* ; et 1 PROREL : *"L'un des problèmes les **plus graves qu'/PROREL** affronte l'université est celui du premier cycle, avec son considérable taux d'échec."*

Descripteur 6.19 Expression comparative complexe : expressions de type *plus de X que de Y* ou *aussi X que Y*. Résultats : 81 CS : *"C'est **moins** une mode **qu'/CS** un uniforme"* ; et 3 PROREL : *"C'est un organisme du **même** genre **que/PROREL** l'on veut créer au bénéfice de l'Europe tout entière."*

7 Résultats pour les descripteurs ciblés

	ADV		CS		PROREL		PROWH		Total	Erreurs	
ADV	133	(+43)	6	(-38)	0	(-4)	0	(-1)	139	6	(-43)
CS	10	(-27)	1135	(+38)	50	(-11)	0	(-1)	1195	60	(-38)
PROREL	0		52	(-17)	261	(+17)	0		313	52	(-17)
PROWH	0		0	(-4)	4	(+2)	25	(+2)	29	4	(-2)

TABLE 2 – Matrice de confusion pour *que* avec les descripteurs ciblés

La table 2 montre la matrice de confusion pour *que* après l'ajout des descripteurs riches. Les résultats sont considérablement améliorés pour toutes les catégories, mais plus particulièrement pour la confusion entre ADV et CS. En tout, nous avons supprimé 45 % des erreurs, passant d'une exactitude de 86,75 % à 92,72 %. Les résultats sont hautement significatifs, avec 139 nouvelles corrections pour 29 nouvelles erreurs (test de McNemar, p-valeur < 0,001). Pourtant, ces gains ont un prix : la vitesse d'analyse. Dans la version de base, on étiquette 1 million de mots en 6m48s. Avec les descripteurs ciblés, cela prend 1,5 fois plus de temps : 10m09s.

8 Les règles

Dans le paragraphe précédent, certains cas n'ont pas été corrigés même quand les descripteurs riches ajoutaient du poids à la bonne étiquette : les descripteurs riches semblaient noyés dans un océan de descripteurs plus pauvres et génériques, ce qui les empêchait de faire pencher la balance en faveur de la bonne étiquette.

En analysant les erreurs restantes, nous avons identifié certaines règles qui nous semblaient généralisables. Vu la rareté des phénomènes qui peuvent être ciblés par des règles, nous avons examiné ici les erreurs dans tous les corpus sauf SPMRL *test* et EMEA *test*, à la différence de l'expérience avec les descripteurs, où seulement les erreurs de SPMRL *dev* avaient été examinées. Nous avons retenu les règles suivantes :

Règle 8.1 Étiqueter PROWH si la phrase se termine par un point d'interrogation, si *que* est le premier mot de la phrase (hors conjonctions de coordination), et si *que* est directement suivi par un verbe à l'indicatif ou l'infinitif (avec ou sans les clitiques *en* ou *se*). Exemple : *"Et que/PROWH se passera-t-il si un seul syndicat signe un accord de ce type contre l'avis des autres ?"*

Règle 8.2 Étiqueter PROWH le premier *que* dans *qu'est-ce que*.

Règle 8.3 Étiqueter CS dans les locutions de type *attendre/veiller/tenir à ce que*, *n'empêche que*, *dommage que*, *avoir honte à ce que*, *le/du/au fait que*, *une fois que*.

Règle 8.4 Étiqueter CS pour toute locution de type *être ADJ que*, tel que "*il est probable que*", sauf si l'expression est précédé de *ne*.

Règle 8.5 Étiqueter PROREL dans *ceux/celui/celle/celles/quoi/qui/quel/quelle/quels/quelles/où que* et dans l'expression *tout ce que*.

Règle 8.6 Étiqueter PROREL dans les locutions *c'est NPP que* ou *c'est DET NC que*

Règle 8.7 Étiqueter ADV pour toute locution verbale de forme *ne V que*, où il n'existe pas le mot *rien, personne, aucun, aucune, ni* ou *jamais* plus tôt dans la phrase.

Règle 8.8 Ne pas étiqueter PROREL si *que* est le premier mot de la phrase.

Règle 8.9 Ne pas étiqueter PROREL si *que* suit un verbe directement, ou s'il est séparé du verbe uniquement par un commentaire entouré par des virgules, des tirets ou des parenthèses. La dernière condition est un exemple d'une règle visant un corpus spécifique, en l'occurrence le corpus Europarl de Séquoia, qui contient beaucoup de phrases du type *"Je sais, Madame la Présidente, **que**/CS vous êtes déjà intervenue [...]"* où le *que* a été étiqueté PROREL.

Règle 8.10 Ne pas étiqueter PROREL si *que* suit les mots *reprises, tous, toutes, toute, tout, soi*

Règle 8.11 Ne pas étiqueter ADV sauf s'il y a un *ne* plus tôt dans la phrase, ou si *que* est le premier mot de la phrase.

	ADV	CS	PROREL	PROWH	Total	Erreurs
ADV	134 (+1)	5 (-1)	0	0	139	5 (-1)
CS	10	1149 (+14)	36 (-14)	0	1195	46 (-14)
PROREL	0	48 (-4)	265 (+4)	0	313	48 (-4)
PROWH	0	0	2 (-2)	27 (+2)	29	2 (-2)

TABLE 3 – Matrice de confusion pour *que* avec les règles

La table 3 montre les résultats après l'ajout des règles pour tous les corpus *dev* et *test*, avec, entre parenthèses, les gains par rapport au modèle des descripteurs riches. Sans surprise, les résultats sont positifs, puisqu'on a visé des erreurs trouvées dans ces mêmes corpus *dev*. Il nous reste 101 sur 222 erreurs, soit une réduction de 55 % du taux d'erreur, avec une exactitude de 93,97 %. Au niveau de la significativité, nous avons 21 nouvelles corrections pour 0 nouvelles erreurs. Pourtant, il y a un risque de suradéquation des règles aux corpus évalués, les corpus *test* étant trop petits pour mesurer l'impact plus global des règles. Du coup, nous avons testé les mêmes règles sur des corpus non annotés, comparant les différences entre les analyses avec et sans règles. Nous avons donc analysés, avec et sans règles, 200 000 mots de chacun des corpus suivants :

– Est Républicain : le journal régional Est Républicain de l'année 2003, disponible sur le site du CNRTL [2]
– Leximedia : une collection d'articles concernant la campagne présidentielle 2007, extraits des journaux nationaux Le Monde, Libération et Le Figaro, et préparé par le laboratoire CLLE-ERSS [3]
– Frantext : des textes littéraires français du 20ème siècle [4]
– Revues.org : une collection d'articles scientifiques dans les sciences sociales [5]

Au total, il y a uniquement une différence tous les 8 500 mots, mais avec un bilan très positif : 46 corrections pour 5 erreurs dans les 51 premières différences. Les erreurs pourraient être éliminées facilement si on affinait les règles. Les corrections les plus intéressantes concernent les commentaires qui séparent le *que* du verbe qui le gouverne. La règle corrige l'étiquette PROREL en CS, comme dans la phrase suivante :

Exemple 8.1 *Je **conteste**, en tant que père de famille, **que**/CS l'on vienne me dire que l'argent est le corollaire du succès.*

2. http://www.cnrtl.fr/corpus/estrepublicain/
3. http://redac.univ-tlse2.fr/applications/leximedia2007.html
4. http://www.frantext.fr
5. http://www.revues.org/

9 Conclusions et perspectives

Nous avons voulu, dans la présente étude, démontrer l'intérêt d'injecter des connaissances linguistiques riches dans un système statistique. En conclusion, il est clair que les descripteurs riches, spécifiques à une langue donnée, permettent de corriger un grand nombre d'erreurs dans le cas de l'étiquetage morphosyntaxique des mots fonctionnels, avec une réduction de 45 % du taux d'erreur pour le mot *que*. L'application supplémentaire de règles très spécifiques permet d'atteindre une réduction totale de 55 %.

Reste la question de la facilité de maintenance des systèmes basés sur ce type d'information plus riche. Par rapport aux systèmes "rationalistes" qui fonctionnent uniquement à base de règles formelles et cherchent à décrire une langue de façon complète, les systèmes "empiriques" statistiques, à base d'apprentissage supervisé, ont le grand avantage de laisser le travail de description linguistique au corpus d'apprentissage. Du coup, la complexité de la langue peut être décrite dans un guide d'annotateur, au lieu d'être codée de façon formelle au sein du logiciel. En plus, on se contente de décrire les cas présents dans un corpus donné, mettant ainsi l'accent sur les cas les plus courants. Cette abstraction de la complexité linguistique rend la maintenance du système beaucoup plus simple. Cela reste-t-il vrai dans la présente étude ? Dans le cas des descripteurs ciblés la réponse est "oui" : autant l'écriture et l'affinage des descripteurs peuvent s'avérer long et complexe, autant la maintenance du système à long terme est simple, puisque le modèle probabiliste ajuste automatiquement le poids de chaque descripteur au fur et à mesure que d'autres descripteurs sont ajoutés, où que d'autres données d'apprentissage deviennent disponibles. Pour les règles, la réponse est plus complexe. Il est important de baser les règles uniquement sur des erreurs effectivement rencontrées dans le corpus de développement, et que l'on peut décrire de façon non ambigüe. Ceci réduit considérablement le nombre de règles, qui sont là uniquement pour compléter le système dans certains cas bien définis : la plupart du travail continue à être fait par les descripteurs.

Les perspectives sont nombreuses : tout d'abord, il y a la question de la vitesse d'analyse. Bien que l'étiquetage morpho-syntaxique de Talismane soit assez rapide après l'ajout des descripteurs et des règles (5 millions de mots / heure), il est plus lent que dans la version de base (9 millions de mots / heure). Nous avons choisi pour l'instant de mettre tous les descripteurs dans un fichier de configuration. L'avantage est de séparer complètement les descripteurs du code source, permettant ainsi à un linguiste d'inventer de nouveaux descripteurs sans l'aide d'un informaticien. L'inconvénient est que certaines opérations (ex. recherche en avant et en arrière) sont répétées de nombreuses fois, alors qu'elles pourraient être codées de façon bien plus efficace dans un langage informatique compilé et plus expressif. Il serait donc intéressant de mesurer le gain de vitesse en codant ces mêmes descripteurs dans un langage compilé.

Il serait aussi souhaitable d'effectuer une analyse plus fine des erreurs qui restent après l'application des descripteurs et des règles : y a-t-il encore une possibilité de diminuer le taux d'erreur ?

Ensuite, nous souhaitons appliquer la même méthodologie à d'autres mots fonctionnels, tels que "de/des/du" et "soit". Finalement, nous souhaitons tester une méthodologie semblable sur les erreurs du parseur, pour voir si les descripteurs riches et les règles peuvent être aussi efficaces dans le contexte plus compliqué du parsing par transitions.

Remerciements

Je tiens à remercier les relecteurs anonymes pour leurs commentaires et suggestions. Je tiens aussi à remercier l'équipe de l'axe CARTEL à CLLE-ERSS pour leur soutien pendant ce travail.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Kluwer.

B. BIGI, Ed. (2014). *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, Marseille. ATALA, LPL.

CANDITO M., SEDDAH D. *et al.* (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

CRABBÉ B. & CANDITO M. (2008). Expériences d'analyses syntaxique statistique du français. In *TALN 2008- conférence sur le Traitement Automatique des Langues Naturelles* : ATALA.

DANLOS L. (2005). Ilimp : Outil pour repérer les occurrences du pronom impersonnel il. In *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005)*, p. 123–132, Dourdan, France.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, **46**(4), 721–736.

JACQUES M.-P. (2005). Que : la valse des étiquettes. In *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005)*, p. 133–142, Dourdan, France.

KÜBLER S., MCDONALD R. & NIVRE J. (2009). *Dependency parsing*. Morgan & Claypool Publishers.

SAGOT B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIORKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & VILLEMONTE DE LA CLÉRGERIE E. (2013). Overview of the spmrl 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages : Shared Task*, Seattle, WA.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 188–201, Les Sables d'Olonne, France.

ZHANG Y. & NIVRE J. (2011). Transition-based dependency parsing with rich non-local features. In *ACL (Short Papers)*, p. 188–193.