Que : la valse des étiquettes

Marie-Paule Jacques

ERSS – Université Toulouse II le Mirail
5, allées Antonio Machado 31058 Toulouse Cedex 9
mpjacques@univ-tlse2.fr

Mots-clés : Analyse syntaxique automatique, étiquetage morphosyntaxique
Keywords: Automatic parsing, tagging

Résumé Nous présentons ici une stratégie d'étiquetage et d'analyse syntaxique de *que*. Cette forme est en effet susceptible d'appartenir à trois catégories différentes et d'avoir de multiples emplois pour chacune de ces catégories. Notre objectif est aussi bien d'en assurer un étiquetage correct que d'annoter les relations de dépendance que *que* entretient avec les autres mots de la phrase. Les deux étapes de l'analyse mobilisent des ressources différentes.

Abstract In this paper I present a method for tagging and parsing the grammatical word *que*. This word is particularly difficult to tag because it may belong to three different categories and may give rise to many constructions for each category. My aim is to assign the correct tag and to annotate dependency relations between *que* and the other words of the sentence.

1 Introduction

Le mot *que* est de ces formes redoutables pour le TAL français. Non seulement peut-il recevoir plusieurs étiquettes différentes selon ses emplois (voir section 2.1), mais encore entre-t-il dans quantité de constructions syntaxiques différentes, qui correspondent à autant de valeurs sémantiques et d'instaurations de dépendances différentes entre les éléments de la phrase. Or, il est bien souvent nécessaire de résoudre les questions posées par la présence d'un *que* dans une phrase pour proposer une analyse syntaxique satisfaisante de celle-ci.

Notre étude poursuit deux objectifs : i. assigner une étiquette morphosyntaxique correcte à *que* ; ii. annoter les relations syntaxiques qu'il entretient avec d'autres éléments de la phrase. Notre propos n'est pas de discuter les avantages de telle ou telle procédure d'étiquetage morphosyntaxique, par règles ou probabiliste (pour un tour d'horizon, Abney, 1996 ; Garside et al., 1997 ; Habert et al., 1997), mais de tester une méthode d'analyse qui **établit l'étiquetage définitif par l'analyse syntaxique.** Alors que bien souvent ces deux tâches se succèdent – étiquetage **puis** analyse syntaxique appuyée sur les résultats du tagger –, nous procédons à un étiquetage définitif de *que* en fonction de l'identification de constructions syntaxiques

caractéristiques de telle ou telle catégorie. Pour tracer un panorama des difficultés d'analyse, un tour d'horizon des diverses valeurs et constructions possibles de *que* s'impose.

2 Polycatégorie et polyfonctionnalité syntaxique

Dans le Trésor de la Langue Française informatisé (TLFi), la description de *que* couvre 14 pages format A4, c'est dire la variété de ses emplois. Selon les cas, il peut s'agir d'un adverbe, d'un pronom ou d'une conjonction de subordination.

2.1 Les diverses catégories de *que*

- ***que* adverbe**

Que se voit classer comme adverbe dans les phrases exclamatives du type :

> ***Que** cela nous semble alors loin !*

En fonction des auteurs, dictionnaires et autres ouvrages de grammaire, il y a désaccord sur la valeur de *que* dans une construction telle que :

> *Les geysers **n'**entrent en activité **que** la nuit et au petit matin.*

Dans cette structure de négation exceptive *ne V que*, *que* est considéré soit comme adverbe, soit comme une conjonction de subordination. Nous avons choisi la première solution, *que* adverbe, qui présente l'avantage d'une cohérence d'ensemble des constructions impliquant le premier morphème de négation *ne* : tous les éléments qui participent à la négation quels qu'ils soient sont étiquetés adverbe[1]. Ainsi, *que* entre dans un paradigme de formes pouvant co-occurrer avec *ne* pour former tous types de négations.

- ***que* pronom relatif**

Comme tous les pronoms relatifs, *que* est susceptible d'introduire une subordonnée dans laquelle il a une fonction syntaxique, généralement celle d'objet direct :

> *Le volcan est connu pour les risques **qu'**il constitue pour les populations avoisinantes.*

Dans certaines constructions clivées, *que* est ce qu'on appelle un relatif sans antécédent :

> *C'est dans les Andes **que** l'on trouve les plus hauts volcans du monde.*

Les clivées dans lesquelles c'est l'objet du verbe de la subordonnée qui est extrait – par exemple *c'est le chocolat que j'aime* – se ramènent au cas de figure évoqué précédemment.

---

[1] Un autre motif de ce choix tient au fait que, dans le corpus CRATER (cf. 3.1) qui nous sert d'étalon pour la mesure des performances de notre module d'analyse, *que* est étiqueté adverbe pour la négation exceptive.

- ***que* conjonction de subordination**

C'est en tant que conjonction de subordination que *que* manifeste la plus grande variété d'emplois. La place nous manque pour être absolument exhaustive, nous indiquons ici les constructions les plus typiques des textes sur lesquels nous avons basé notre inventaire[2].

Dans une part massive de ses emplois, *que* introduit une complétive, c'est-à-dire soit l'objet direct d'un verbe, soit le complément d'un nom ou encore le complément d'un adjectif :

> *Tout le monde est d' accord pour **penser que** le gaz toxique est venu du fond du lac.*
>
> *Les raisons invoquées proviennent **du fait que** la médecine n'est pas une science exacte.*
>
> *Il est **rare qu'**un séisme provoque directement une activité volcanique.*

Tout en continuant à être catégorisé comme conjonction par l'ensemble des grammaires et dictionnaires, *que* entre dans quantité de structures que, par commodité, nous regroupons sous l'appellation de corrélatives. Dans celles-ci, *que* n'introduit pas nécessairement une **phrase** subordonnée, ce qui a pu donner lieu à une analyse en terme d'ellipse (Riegel et al., 1994)[3]. Nous pouvons remarquer que, dans ces emplois, *que* fonctionne nécessairement avec un autre élément, un élément « déclencheur » situé avant lui dans la phrase, pour assurer une mise en relation de deux constituants : celui qui est contigu à (ou qui contient) l'élément déclencheur, celui qui est introduit par *que*. Ces structures corrélatives se caractérisent donc par la présence d'un couple de marqueurs tantôt contigus, tantôt discontinus, comme par exemple *plus/moins... que*, *d'autant plus/d'autant moins... que*, *aussi... que*, *tel... que*, *autre/même... que*, *tant/autant/aussi bien... que*, etc. :

> *Aucun phénomène n'a donné naissance à **autant** de mythes, de symboles, de légendes, de rites ou de superstitions **que** le 'feu de la terre'.*

2.2 Difficultés pour l'analyse

Les difficultés essentielles à surmonter viennent de ce que des constructions analogues sur le plan de la forme ne le sont cependant pas sur la valeur de *que*. Par exemple, nous n'insisterons pas sur la difficulté, bien connue, de distinguer de façon automatique la valeur de pronom relatif de celle de conjonction de subordination :

> *On appelle compétence du courant la **possibilité qu'**il a de transporter des matériaux.*
>
> *J. et M. semblent exclure la **possibilité que** des opportunités demeurent non perçues.*

De la même manière, la valeur d'adverbe ne se laisse pas simplement cerner par la présence de *ne* devant le verbe :

---

[2] Nous laissons de côté diverses collocations qui peuvent sans ambiguïté être étiquetées 'conjonction de subordination', comme *à condition que*, *afin que*, *de sorte que*, *parce que*, etc.

[3] Nous n'avons pas ici la place de discuter du bien-fondé de cette analyse.

> *Cependant on **n'**est plus persuadé **que** cette action rasante soit entièrement responsable des formes de champignons constatées dans les déserts*

Contrairement à ce que l'on pourrait penser, ce n'est pas seulement le morphème *plus* qui fait que l'on n'a pas de négation exceptive, comme on le voit dans la phrase suivante :

> *Le volcan **n'**émet **plus**, par de nombreux points du cratère, **que** de la vapeur d'eau surchauffée*

Autre exemple, la présence d'un marqueur lexical caractéristique d'une corrélative n'est pas un gage d'identification certaine de la valeur de *que*, comme le montrent les deux extraits suivants, dans lesquels, au sein d'une même construction *plus/moins difficile à Vinf que...*, *que* assume une fonction d'opérateur de comparaison aussi bien que d'objet direct du verbe :

> *Une solution française semble dans ce cas **plus difficile à envisager qu'**une reprise par un groupe étranger.*

> *Il n'en reste pas **moins difficile à expliquer que** les crêtes et sillons prélittoraux ne soient pas toujours parallèles à la côte.*

Face à la diversité des constructions et des éléments à prendre en compte pour les identifier, nous avons testé une stratégie qui intervient à deux moments de l'analyse syntaxique et mobilise des informations différentes dans chacun de ces deux moments. Nous indiquons maintenant le cadre de notre expérimentation, la teneur de la méthode et les résultats obtenus.

3 L'analyse de *que* : cadre, méthode et résultat

3.1 Cadre de l'expérimentation

L'expérimentation que nous décrivons prend place dans une analyse syntaxique de type modulaire, celle mise en œuvre par Syntex (voir ici même Bourigault, Frérot, 2005). A partir d'un texte étiqueté par TreeTagger (Schmid, 1994), divers modules se succèdent, chacun effectuant une partie de l'analyse syntaxique, ce qui autorise les modules les plus tardifs à utiliser les relations posées par les modules précédents pour leur propre analyse.

Syntex fournit une annotation de relations de dépendance entre mots. Par exemple, d'un nom pourront dépendre un déterminant, un ou plusieurs adjectifs, une ou plusieurs prépositions… ; d'une préposition pourront dépendre un nom ou un verbe ; d'un verbe pourront dépendre un nom et/ou un pronom en fonction sujet, idem en fonction objet, une ou plusieurs prépositions, etc. Chaque mot de la phrase est donc susceptible d'entrer dans deux types de dépendances – être régi par un autre mot, être recteur d'un autre mot –, mais qui ne sont pas toutes deux obligatoires : par exemple, le verbe d'une principale n'est pas régi.

Pour ce qui concerne *que*, TreeTagger lui attribue l'une des trois étiquettes 'Pronom relatif', 'Adverbe' ou 'Conjonction de subordination'. Notre analyse a pour objet : i. de revenir sur cet étiquetage pour éventuellement le rectifier, ii. de relier *que* aux éléments convenables dans les deux directions, c'est-à-dire vers un recteur et vers un régi (par exemple, le verbe de la subordonnée), et ce par la relation idoine.

Afin de déterminer les indices sur lesquels appuyer l'analyse et d'avoir un aperçu réaliste de la diversité des constructions, nous avons constitué un corpus d'étude en rassemblant, de façon opportuniste, la plus grande variété de textes dont nous pouvions disposer sous format électronique : articles du journal Le Monde, articles scientifiques du domaine de l'ingénierie des connaissances, recettes de cuisine, documents professionnels, textes littéraires, textes spécialisés de domaines aussi divers que la volcanologie, la géomorphologie, la médecine, le vol libre. Il s'agissait pour nous de ne pas nous cantonner à un seul genre, sans toutefois prétendre couvrir tous les genres possibles. Nous pensons tout de même avoir de cette manière recueilli une bonne gamme des possibilités d'emploi de *que*.

L'évaluation n'a pas été menée sur ce corpus, mais sur un corpus de test extrait de la partie française du corpus CRATER[4]. Celui-ci offre l'intérêt d'un étiquetage morphosyntaxique – relativement – fiable parce que vérifié par des annotateurs humains. Il constitue donc un banc d'essai idéal pour mesurer l'efficacité de notre analyse sur l'étiquetage. Nous avons extrait de ce corpus toutes les phrases contenant la forme *que*, puis nous avons annoté manuellement les 1100 premières pour les relations de dépendance.

3.2 Méthode d'analyse

L'analyse de *que* se fait en deux étapes principales intervenant à des moments différents dans le processus global. Son principe essentiel est d'exploiter les acquis des modules précédents pour repérer certaines constructions syntaxiques modélisées à partir du travail sur corpus et décider, en fonction de celles-ci, de la catégorie définitive de *que*. En résumé, **ce n'est que lorsqu'une construction syntaxique est positivement identifiée que *que* est réétiqueté.**

Dans un premier temps, sont repérées certaines constructions qui ne requièrent pour être identifiées que peu d'informations de structure. Puis, en toute fin d'analyse, un module spécifique repère les constructions faisant intervenir des relations à plus grande distance.

3.2.1 Première étape : des constructions locales

Rappelons que Syntex produit une analyse à partir d'un texte préalablement étiqueté. Cela implique que la qualité de l'analyse syntaxique dépend crucialement de la qualité de l'étiquetage. Par exemple, si un *que* adverbe est faussement étiqueté conjonction de subordination, le module de recherche des objets directs est mis en échec : une conjonction de subordination constitue une barrière après laquelle on ne peut avoir un objet direct. Pour éviter ces obstacles liés à des erreurs d'étiquetage, il est apparu nécessaire d'avoir le plus tôt possible dans l'analyse un processus de vérification et de correction de *que*. Mais, « très tôt dans l'analyse » signifie que l'on ne peut s'appuyer que sur très peu d'informations de structure, puisque celle-ci ne sera découverte que par les modules suivants et précisément qu'à condition qu'un mauvais étiquetage ne rende pas la tâche impossible.

Pour l'essentiel, dans cette première étape, sont donc analysés et réétiquetés des *que* dont le contexte proche fournit une information exploitable : des *que* adverbes placés immédiatement

---

[4] Université Lancaster. *UCREL Projects* [en ligne]. http://www.comp.lancs.ac.uk/ucrel/projects.html#crater (page consultée le 4 février 2005)

après le verbe, des *que* objets situés eux aussi à proximité du verbe, des *que* introduisant un complément de nom (*le fait que, l'idée que, l'hypothèse que, …*), des *que* pris dans une structure comparative très locale du type *plus/moins/aussi/autant/si Adj que*. L'objectif lors de cette phase est de découvrir le plus possible de *que* adverbe ou conjonction de subordination en opérant, grâce au contexte, une « déduction affirmative » (Vergne, Giguet, 1998) sur la catégorie de *que*. Dans le même temps, et parce que cette déduction est liée au repérage d'un type de construction syntaxique, sont annotées les relations de *que* avec son recteur. Pour ce qui est des relations avec les régis, seules sont annotées celles des complétives car c'est le seul cas dans lequel on est sûr que l'élément régi soit le verbe conjugué d'une proposition subordonnée. En effet, dans une structure corrélative, *que* n'introduit pas nécessairement une proposition (*davantage* ***que je ne croyais*** *vs. davantage* ***que l'autre jour***).

Les informations mobilisées mêlent des listes lexicales – une vingtaine de noms qui prennent un complément en *que*, une dizaine d'adverbes susceptibles d'entrer dans une structure corrélative mettant en jeu un adjectif ou un participe passé, moins d'une dizaine d'adverbes de négation (*pas, point, guère*, etc.), deux cents verbes qui prennent un objet direct en *que* (*penser que, croire que*, etc.) –, l'exploitation des quelques relations qui ont été posées – rattachement des adverbes autour des verbes et des adjectifs et rattachement des auxiliaires et des modaux aux participes passés pour les premiers, à un verbe à l'infinitif pour les seconds – et enfin, la prise en compte de la catégorie des mots avant et après *que*. Rappelons que la démarche consiste essentiellement à repérer des éléments positifs d'analyse dans le contexte proche (nous donnerons en 3.2.3 un exemple de règle d'analyse). L. Danlos (2005) adopte pour la désambiguïsation de *il* (pronom impersonnel *vs.* pronom anaphorique) une démarche très similaire – modélisation de patrons linguistiques qui constituent les 'marqueurs' d'une valeur donnée – et elle obtient avec cette approche 97,5% de bons résultats, ce qui est plus que satisfaisant.

3.2.2 *Deuxième étape : la structure globale*

Au moment de cette seconde étape, tous les modules ont fait leur travail et ont placé diverses relations qui vont être exploitées pour l'analyse de *que*. Le principe est de remonter vers la gauche de *que* à la recherche d'indices de telle ou telle structure. En fonction de la nature des constituants rencontrés, certains éléments sont recherchés. Par exemple, si dans ce parcours vers la gauche un nom est rencontré, on vérifie si ce nom régit un adjectif tel que *même* ou *autre*, si oui, on estime être face à une structure corrélative et l'analyse s'arrête là, si non, on se déplace sur le recteur du nom, on teste à nouveau une série de contraintes et ainsi de suite. Pour chaque nouveau constituant, une série de contraintes est évaluée et dès qu'une contrainte est satisfaite, l'analyse s'arrête. De cette façon, dans les deux exemples suivants, extraits de CRATER, on analyse successivement divers syntagmes prépositionnels jusqu'à arriver aux éléments qui permettent d'identifier deux structures corrélatives différentes.

> *L'indication de fonction peut ou non contenir le* ***même*** *numéro d'identification de fonction* ***que*** *celui qui se trouvait dans la demande d'activation de fonction d'origine.*
>
> *Les procédures s'appliquent* ***aussi bien*** *à une interface à débit de base* ***qu'****à une interface à débit primaire.*

Les informations mobilisées sont donc les mêmes que pour la première étape, plus – et essentiellement – toutes les informations de structure disponibles. Ce sont celles-là qui sont essentielles car, le principe étant de progresser vers la gauche constituant par constituant, **en**

s'appuyant sur les relations que les modules précédents ont annotées, l'analyse s'interrompt dès qu'un constituant de la chaîne est « orphelin », c'est-à-dire n'est rattaché à aucun autre élément.

Cette seconde étape rajoute quelques règles d'analyse, notamment pour l'analyse des syntagmes nominaux et prépositionnels, pour lesquels on ne disposait d'aucune relation de dépendance lors de la première étape, mais aussi elle permet d'appliquer des règles déjà définies à des éléments de la phrase placés non immédiatement dans le contexte de *que*. Pour l'illustrer, nous prendrons comme exemple la règle qui s'applique à l'analyse d'un verbe conjugué à gauche de *que*.

3.2.3 Un exemple de règle d'analyse

Il s'agit ici d'une règle utilisée aux deux étapes de l'analyse pour repérer certaines constructions verbales. Elle permet de décider dans un certain nombre de cas d'étiqueter *que* comme conjonction de subordination (CSub) ou comme adverbe (Adv), en repérant les éléments d'une corrélative, d'une relation d'objet direct ou d'une négation exceptive.

Lorsque la forme *que* est immédiatement précédée d'un verbe conjugué ou précédée d'un adverbe lui-même précédé d'un verbe conjugué, alors on lance l'analyse du verbe, qui consiste à tester les conditions énumérées ci-dessous et, en cas de test positif, à attribuer à *que* l'étiquette mentionnée après la flèche.

- présence des adverbes : *plus, moins, davantage, autant, mieux, tellement* → CSub ;
- présence de *ne* immédiatement précédé par *rien, nul, personne* (par exemple *Nul ne comprendrait que…* qui n'est pas une négation exceptive) → CSub ;
- présence conjointe de *ne* et de *pas, point* ou un autre *que* adverbe → CSub ;
- présence seule de *ne* : négation exceptive → Adv ;
- appartenance du verbe à la liste de ceux qui prennent *que* comme objet direct → CSub + recherche du verbe de la subordonnée introduite par *que*.

Il est possible aussi qu'on ne se prononce pas, c'est-à-dire qu'on laisse l'étiquette attribuée par TreeTagger en l'état, si on n'a recueilli aucun indice pour faire mieux que lui.

Dans la seconde étape, la même règle est utilisée pour analyser des verbes placés plus loin de *que*, par exemple :

> *Un « champ électrique » implique une activité qui* ***ne*** *peut être exprimée d'une façon univoque* ***qu'****en fonction du temps et de deux ou trois dimensions.*

En suivant les relations de dépendance placées par l'analyse syntaxique, le module se déplace de l'adjectif *univoque* à son recteur *façon*, de celui-ci à la préposition *de*, de cette dernière au participe passé *exprimée*, puis au verbe *être* et enfin au verbe *pouvoir*, auquel s'applique la règle exposée ci-dessus, qui permet de reconnaître un *que* adverbe.

Il s'agit ici d'une partie de la règle qui couvre la plus grande proportion de *que* : dans le corpus CRATER, les *que* objets directs ou adverbes partie prenante d'une négation représentent 52% des relations vers un recteur. Dans la première étape, la règle permet d'identifier 76% des *que* objets et 61% des adverbes, un pourcentage qui est diversement amélioré dans la seconde étape

puisqu'il passe à 78% pour les objets et 82% pour les adverbes. Ces chiffres ne sont qu'une composante des résultats, intéressons-nous maintenant à ceux-ci.

3.3 Résultats

Dans la mesure où la tâche concerne aussi bien l'étiquetage que l'annotation de relations, nous avons évalué notre méthode sur ces deux points.

3.3.1 L'étiquetage

Pour ce qui concerne l'étiquetage, la seule mesure que nous présentons concerne la précision. En effet, le rappel est de 100% : absolument tous les mots reçoivent une et une seule étiquette. La mesure porte sur 1183 *que*, étiquetés initialement par TreeTagger. Le tableau ci-dessous récapitule les résultats :

Etape	Précision	
étiquetage initial	75%	
étape 1	89%	+14%
étape finale	92%	+3%

Tableau 1 : Mesure de la précision de l'étiquetage morphosyntaxique

On voit que l'essentiel du gain est obtenu après la première étape, ce qui tend à montrer que pour la détermination de la catégorie de *que*, des informations très locales suffisent. Il n'en est pas de même pour l'annotation des relations.

3.3.2 Les relations

Les mesures proposées ici sont restreintes aux cas où l'étiquetage de *que* est correct. Le nombre de relations évaluées n'est donc pas le même à chacune des étapes : 1789 pour la première, 1843 pour la seconde, toutes relations confondues. Il faut noter que pour les *que* adverbes, l'analyse n'inscrit qu'une relation, vers le recteur, alors que pour les *que* conjonction ou pronom relatif, deux relations doivent être annotées, l'une vers le recteur, l'autre vers le régi pour les conjonctions, vers un antécédent nominal pour le pronom. Ceci pose, dans le cas des relatifs sans antécédent, tels que ceux qui participent à une structure clivée (voir la section 2.1), de véritables problèmes de représentation : soit on note une relation d'antécédence fictive et arbitraire vers l'un quelconque des éléments du constituant clivé, soit aucune relation d'antécédence n'est mentionnée (c'est la position retenue actuellement).

Le rappel et la précision de l'annotation des relations sont indiqués dans le tableau 2.

Etape	Rappel	Précision
étape 1	60%	97%
étape finale	93% +33%	94% -3%

Tableau 2 : Mesures de rappel et de précision de l'annotation des relations

A titre de comparaison, la précision de Syntex pour le rattachement des prépositions varie de 78 à 87 % selon les corpus (Bourigault, Frérot, 2005).

Il apparaît ici clairement que l'annotation des relations bénéficie considérablement des informations sur les relations existantes au sein de la phrase, notamment parce qu'il devient alors possible de placer des relations à distance. Mais cette annotation de relations se trouve améliorée aussi pour la simple raison que la recherche des régis des conjonctions de subordination dans les corrélatives n'est pas du tout effectuée dans la première étape, et ce parce que l'on a besoin de disposer des dépendances pour, par exemple, ne pas prendre un pronom pour régi et arriver jusqu'au verbe dont il est sujet, en d'autres termes, ne pas analyser *tel que nous l'avons défini* comme *tel que nous*.

A l'heure actuelle, les performances sont limitées par deux types d'obstacles. D'une part, l'ambiguïté de certaines structures : comme nous l'avons déjà mentionné, la présence des marqueurs recherchés ne garantit pas d'être effectivement face à la structure supposée, ce qui est parfaitement illustré par les deux exemples suivants, formellement analogues. Dans le premier, la présence de *tels* conduit l'analyseur à manquer la relation avec le verbe *dire*.

> *On dit couramment de **tels** réseaux mésochrones **qu'**ils sont synchronisés.*
>
> *La nature sociale comporte de **tels** hasards **que** l'imagination des inventeurs est à tout moment dépassée.*

D'autre part, les limites du fichier d'entrée constituent des écueils actuellement infranchissables : si par exemple un verbe n'est pas étiqueté verbe mais nom, si une préposition n'a pas été rattachée ou encore si un adjectif ne régit pas l'adverbe qui manifesterait une comparaison, le processus d'annotation des relations est bloqué. Ce sont donc deux directions vers lesquelles porter les efforts.

Avant de conclure, soulignons que ces résultats sont partiels dans la mesure où ils se limitent à l'analyse de la forme *que*. En fait, il faudrait pouvoir évaluer l'impact d'un meilleur traitement de *que* sur l'analyse globale, ce que nous n'avons pas été en mesure de faire.

4 Conclusion

Nous avons présenté une stratégie qui effectue dans le même mouvement l'analyse et l'étiquetage de *que*, en prenant appui sur les acquis de l'analyse syntaxique au fur et à mesure de son déroulement. La confrontation à un corpus de référence montre que l'étiquetage est grandement amélioré avec peu d'informations de structure, qui sont en revanche indispensables pour l'annotation des relations de dépendance.

Parmi les pistes à continuer à explorer, nous voudrions mentionner la question de la représentation des relations. La représentation de certaines relations ne nous semble pas poser problème : les complétives, les objets directs, les adverbes sont reliés sans difficulté au nom, adjectif, verbe, etc., concerné. Mais les structures corrélatives, par leur diversité, ne se laissent pas facilement appréhender. A l'heure actuelle, nos choix d'annotation impliquent de représenter de la même façon des structures similaires en surface. Considérons par exemple une structure *aussi Adj que…* Il nous importe de reconnaître une corrélative, en identifiant la relation entre l'adjectif et l'adverbe *aussi*, quel que soit l'autre élément de la comparaison. Mais selon la nature de cet élément, **l'interprétation** de la structure n'est pas la même :

Jean est aussi gentil que beau	Jean gentil + beau
Jean est aussi gentil que son frère	Jean gentil + son frère gentil

Dans le premier cas, *que* régit un adjectif, l'interprétation doit être que les deux adjectifs se rapportent au même SN, dans le second cas, *que* régit un SN, on doit alors interpréter que les deux SN « partagent » l'adjectif. Il nous semble qu'il n'est absolument pas trivial de décider comment distinguer ces deux constructions, soit elles sont distinguées au moment de l'annotation, par des relations différenciées, soit on estime que l'annotation doit rester au plus près des structures de surface et que c'est dans l'interprétation qu'elles seront différenciées. Cela mérite réflexion et approfondissement, au-delà du cas particulier de *que*.

Remerciements

Je remercie vivement Cécile Fabre, Didier Bourigault ainsi que les relecteurs de TALN pour toutes leurs remarques et conseils.

Références

ABNEY S. (1996), Part-of-Speech Tagging and Partial Parsing, In K. Church, S. Young et G. Bloothooft (Eds.), *Corpus-Based Methods in Language and Speech*, pp. 118-136, Dordrecht, Kluwer Academic Publishers.

BOURIGAULT D., FREROT C. (2005), Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique, Actes de *TALN'2005*.

DANLOS L. (2005), ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*, Actes de *TALN'2005*.

GARSIDE R., LEECH G., MCENERY T. (1997), *Corpus Annotation: Linguistic Information from Computer*, Londres, Longman.

HABERT B., NAZARENKO A., SALEM A. (1997), *Les linguistiques de corpus*, Paris, Armand Colin.

RIEGEL M., PELLAT J.-C., RIOUL R. (1994), *Grammaire méthodique du français*, Paris, P.U.F.

SCHMID H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Actes de *NEMLAP*.

VERGNE J., GIGUET E. (1998), Regards Théoriques sur le "Tagging", Actes de *TALN'98*, 22-31.