Typologie des subordonnées et des connecteurs en vue de la détection automatique des propositions syntaxiques du français

Yayoi Nakamura-Delloye[1]
LATTICE – Université Paris 7

Abstract

The clause identification is a useful preprocessing step for many applications. In order to define a grammar for clauses recognition, we carried out a typology of subordinate clauses and connectors, appropriate for our application. It is based on Le Goffic's linguistics research on French connectors *qu-*. From these studies of subordinate clauses and connectors, we created a grammar, which was then used to develop a Clause Identification System. The results obtained with our system are very promising. It confirms that it is useful to question some traditional practices and to use pointed linguistic works.

Keywords : syntactic clause, typology of subordinate clauses, connectors, shallow parsing, automatic clause identification.

Résumé

La détection des propositions est une opération intéressante pour de nombreuses applications. Afin de définir une grammaire pour la reconnaissance des propositions, nous avons tout d'abord réalisé une typologie des subordonnées et des connecteurs, adaptée à notre traitement, en nous appuyant pleinement sur les travaux linguistiques de Le Goffic sur les connecteurs en *qu-*. Un système de détection a ensuite été réalisé avec cette grammaire mettant en œuvre les études de typologie des subordonnées et des connecteurs. Les résultats obtenus avec notre système sont très prometteurs. Cela semble confirmer l'utilité de la remise en question de certaines habitudes classiques et d'un recours à des travaux linguistiques pointus.

Mots-clés : proposition syntaxique, typologie des subordonnées, connecteurs, analyse syntaxique partielle, détection automatique des propositions.

1. Introduction

Notre typologie des subordonnées et des connecteurs a été conçue en vue de la détection automatique des propositions. La détection des propositions est une opération intéressante pour un grand nombre d'applications. Elle est par exemple utilisée dans la documentation technique pour transformer les phrases trop longues considérées comme mauvaises, en plusieurs propositions plus brèves. Par ailleurs,

[1] LATTICE, Université Paris 7, École Doctorale de Sciences du Langage, yayoi@free.fr

comme la proposition est souvent considérée comme unité discursive élémentaire, la détection des propositions peut être une opération préparatoire à l'analyse discursive.

Pour développer notre système de détection, nous nous sommes tout d'abord attachée à définir une grammaire pour la reconnaissance des propositions. Toutefois, à peine avons-nous commencé à l'écrire, que nous nous sommes rendu compte que les étiquettes « classiques » attribuées aux connecteurs (ou marqueurs) de subordonnées n'étaient pas forcément adaptées à nos traitements. Nous avons alors étudié les travaux linguistiques, notamment ceux de Le Goffic, sur les connecteurs en *qu-* pour aboutir à une typologie des subordonnées et des connecteurs adéquate pour notre grammaire de détection des propositions.

Nous allons tout d'abord aborder le contexte des travaux (section 2) pour présenter les conditions exactes de définition de notre grammaire et la notion centrale de « proposition » (section 3). Nous présenterons également un bref état de l'art (section 4) des études sur la typologie des subordonnées et des connecteurs pour rendre claire la problématique, avant de proposer notre typologie des subordonnées (section 5) et des connecteurs (section 6). Notre exposé traitera ensuite de la réalisation d'un système de détection des propositions basé pleinement sur ces études linguistiques (section 7), pour se terminer par une discussion sur les pistes d'amélioration (section 8).

2. Contexte de la définition de notre grammaire : conditions et besoins

2.1. Conditions

Nous utilisons comme entrée des textes préalablement traités par deux systèmes extérieurs : un *tagger* et un *chunker* développés à Paris 7. Un *tagger* attribue aux occurrences (*tokens*) des étiquettes de catégorie lexicale et un *chunker* réalise à partir d'un résultat de *tagger* un *chunking*, c'est-à-dire le regroupement d'un certain nombre de *tokens* de manière à constituer des *chunks*. Cinq types de *chunks* sont définis : adverbiaux, adjectivaux, nominaux, prépositionnels et verbaux. Les terminaux de notre grammaire seront donc ces cinq catégories de *chunks* et éventuellement les catégories attribuées à des *tokens* par le *tagger*, qui n'ont pas été traités par le *chunker*.

Notre défi est donc non pas de réaliser une analyse fine, mais plutôt d'obtenir l'analyse la plus précise possible avec des informations limitées, donc avec peu de calcul.

2.2. Besoins

Nos travaux sur la détection des propositions se situent plus particulièrement dans le cadre de l'alignement des textes parallèles, opération qui consiste à mettre en correspondance les unités de plusieurs textes dans différentes langues en relation de traduction. Les corpus alignés peuvent être utilisés en tant que ressources linguistiques

aussi bien dans les travaux de linguistique contrastive que dans la traduction automatique ou la traduction assistée par ordinateur.

Une des hypothèses sur lesquelles sont basées la plupart des méthodes d'alignement des phrases et des propositions est le parallélisme de l'ordre des unités dans les textes traduits. Cependant, dans le cas de l'alignement des propositions français-japonais, on constate beaucoup de croisements des alignements dus à leur différence structurale, et l'automatisation de cette tâche nécessite un algorithme qui ne présuppose pas le parallélisme et qui utilise une structure à deux dimensions telle que les graphes. Nous avons donc besoin d'informations sur les relations entre les propositions pour réaliser l'alignement à l'aide de graphes. Dans ce cadre particulier de l'alignement, l'opération de détection des propositions consiste non seulement en la détection automatique de leurs frontières, mais aussi en la mise en relation de ces unités.

3. Notion centrale : proposition

Ce qu'on appelle ici « propositions » sont des unités syntaxiques constituées d'un sujet et d'un prédicat. Les propositions que nous essayons d'extraire sont : racine (ou principale), coordonnée, subordonnée et détachée-insérée. La finalité de nos travaux étant une application en traitement automatique, nous avons choisi de nous appuyer sur deux critères uniquement formels : type de connecteur et position d'apparition dans la phrase. Enfin, nous traitons également les éléments dits extra-prédicatifs – qui, sans appartenir à la proposition, ne sont pas pour autant un type de proposition –, détachés en tête, afin d'être isolés car considérés comme extérieurs à la proposition. Les définitions que nous proposons sont les suivantes.

3.1. Racine ou principale[2]

Dans une phrase contenant au moins une autre construction phrastique enchâssée à l'aide d'un connecteur ou d'une virgule, la structure phrastique racine – qui ne dépend syntaxiquement d'aucun élément – dans laquelle cette/ces sous-structures sont extraites et représentées par des symboles, est appelée proposition racine ou principale. On appelle également proposition racine ou principale la proposition indépendante constituant toute seule une phrase simple.

[2] Certains, comme Le Goffic (1993a) – ou Delaveau (2001) –, renoncent à la notion de « proposition principale » en dénonçant l'impertinence de la conception classique strictement linéaire. Nous compensons ce défaut avec une représentation à l'aide de symboles indiquant les propositions subordonnées enchâssées extraites. Certes, appeler *proposition* une structure telle que « [A] montre [B] » peut quand même être contestable, mais cet emboîtement des éléments n'est en fait pas un phénomène propre à la principale : une subordonnée peut également être une structure de ce type. Nous distinguons les propositions non pas selon leur structure, mais selon leur niveau dans la phrase et nous conservons la notion de proposition racine ou principale.

3.2. Coordonnée

Dans la phrase graphique constituée de plus de deux unités équivalant à des phrases, l'unité, éventuellement indépendante et autonome, faisant partie de la phrase et reliée par une conjonction de coordination ou une virgule à la proposition qui la précède, est appelée proposition coordonnée.

- Mon père est professeur et ma mère travaille dans une banque.

Cette définition, basée uniquement sur un critère formel, entraîne également l'inclusion de propositions non coordonnées mais subordonnées, et regroupe des propositions classées dans des catégories différentes dans les travaux linguistiques présentés dans l'état de l'art. Les phrases ci-dessous sont considérées avec notre définition comme coordonnées (nous indiquons la catégorisation selon Le Goffic entre parenthèses, à titre d'exemple) :

- Mon père est professeur, ma mère travaille dans une banque.
- J'accepte, dit-il. (*Incise*)
- Vous m'auriez appelé, je serais venu tout de suite. (*Subordonnée paratactique*)
- Plus il gagne de l'argent, plus il en veut. (*Subordonnée paratactique*)
- Paul a beau crier, on ne l'écoute pas. (*Subordonnée paratactique*)
- À peine était-il arrivé, il prenait les choses en main. (*Subordonnée paratactique*)

Nous avons cependant gardé le terme « coordonnée » pour éviter au maximum un néologisme, position cependant discutable.

3.3. Subordonnée

La phrase peut contenir d'autres phrases : une structure de phrase non autonome, intégrée à l'aide d'un connecteur de subordination[3] dans une structure de phrase supérieure, est une proposition subordonnée. Sa position dans la phrase est bien définie selon son type (*cf.* section 5).

- Il était déjà rentré quand je suis arrivé.

3.4. Détachée-insérée

Nous appelons proposition détachée-insérée une construction phrastique sans connecteur, entourée et détachée par deux symboles de ponctuation de même type – virgules, parenthèses ou tirets – et insérée dans une autre phrase. Elle est caractérisée en ce qu'elle peut apparaître en différents endroits de la phrase.

- Il s'en est, <u>me semble-t-il</u>, bien sorti.

Notons tout de même que nous ne considérons comme propositions détachées que les structures constituées d'un sujet et d'un prédicat dont le verbe est bien présent. Ne sont pas traitées d'autres structures dégradées au niveau du verbe, notamment les constructions détachées de Combettes (1998) – que nous aborderons dans la section

[3] Les connecteurs de subordination seront étudiés et définis dans les sections 4 et 5.

suivante –, du fait de la distinction difficile entre les éléments réellement extérieurs à la proposition et ceux intérieurs, tels que les éléments coordonnés pouvant parfois apparaître entourés de virgules.

3.5. Éléments de phrase extérieurs à la proposition

Étant donné que les éléments extra-prédicatifs sont en relation avec le reste de la phrase – constitué d'une ou plusieurs propositions – et ce sur un pied d'égalité, il nous paraît plus cohérent d'en isoler ces constituants, quelle que soit leur structure interne. Les compléments extra-prédicatifs sont typiquement des éléments détachés, situés en particulier en début de phrase. Ce sont par exemple des thèmes (ou *topics*), des introducteurs du cadre du discours (Charolles 1997) ou des constructions détachées de Combettes (1998). Nous extrayons donc de la proposition ces syntagmes détachés en tête, non que nous souhaitions leur accorder un statut de proposition, mais du fait de leur extériorité par rapport à la proposition[4].

4. État de l'art : définitions et examens critiques

4.1. Définition des connecteurs

4.1.1. Définition classique

Wagner et Pinchon (1991) distinguent quatre types de « mots dont le rôle consiste à marquer le caractère dépendant de la proposition qu'ils ouvrent » :

- des conjonctions (*que, comme, quand, si*) et des locutions conjonctives construites au moyen de *que* (*afin que, alors que, de peur que, du moment que, lorsque, pour que*, etc.), de *où* (*du moment où, là où*) ;
- des adverbes interrogatifs *quand ? comment ? où ? pourquoi ?* et des pronoms interrogatifs ;
- des pronoms relatifs représentants ;
- des adverbes de quantité simples (*tant, tellement*) (Wagner et Pinchon 1991 : 541).

Beaucoup de grammaires telles que Riegel *et al.* (1994) proposent une définition des connecteurs de subordination plus ou moins semblable à celle-ci. Mais on peut se demander si toutes ces distinctions de catégories sont pertinentes pour la détection des propositions. En effet, cette définition pose un problème crucial pour les travaux à caractère appliqué : la difficulté d'étiquetage.

Certains de ces mots sont très ambigus et un étiquetage erroné provoque des erreurs dans l'opération suivante. Dans le tableau 1, nous avons représenté les différentes étiquettes que peuvent recevoir les connecteurs selon la catégorisation adoptée pour le

[4] À noter que nous ne séparons pas les éléments extra-prédicatifs apparaissant à une position autre que la position initiale du fait de leur extériorité beaucoup moins nette, surtout pour les éléments situés en fin de phrase.

corpus de Paris 7 (Abeillé et Clément 2003). Nous pouvons y constater la forte ambiguïté de ces connecteurs. La détermination de la pertinence de ces distinctions pour notre opération est d'autant plus utile que leur étiquetage correct est loin d'être simple. Le choix d'une étiquette adéquate nécessite souvent une analyse syntaxique.

	pronom		det.	adverbe			conjonction		autres
	rel.	inter.		inter.	excl.	autre	sub.	crd.	
dont	√								
qui	√	√							
que (qu')	√	√			√	√	√		
quoi	√	√							
lequel*	√	√							
où	√			√					
quel*			√	√					
comment				√					
pourquoi				√					
combien				√	√				
quand				√			√		
comme					√		√	√	prép.
si						√	√		note de musique ou affirmation
s'							√		clitique

* ainsi que toutes leurs formes fléchies

Tableau 1. Ambiguïtés des connecteurs

À cette question de difficulté d'étiquetage s'ajoute le problème lié à l'incohérence entre des mots appartenant à une même catégorie : deux termes qui appartiennent à la même catégorie peuvent avoir des comportements syntaxiques différents. Par exemple, le pronom relatif *dont* peut introduire non seulement une proposition (1), mais aussi un syntagme (2)[5], alors qu'un *que* relatif n'introduit qu'une proposition[6].

1. Le gouvernement a retiré sa proposition *dont* la conformité à la Constitution avait été remise en cause.
2. À cette occasion, se sont réunis huit représentants *dont* notre Président.

Dans le cas de la définition (Abeillé et Clément 2003), le même problème se pose entre les conjonctions de subordination *comme* et *que*. Définir des règles de grammaire

[5] On entend ici par « proposition » et « syntagme », des unités purement de surface. Nous n'entrons pas dans la discussion sur la véritable nature de ces unités introduites par ces connecteurs, que certaines théories linguistiques traitent comme un phénomène d'ellipse.

[6] Nous trouvons tout de même, dans Grevisse (1993), deux types d'exemples – bien que qualifiés de rares – de *que* relatif introduisant une structure non phrastique : suivi d'un gérondif *ce que voyant* (en voyant cela) d'une part, et dans le style juridique *tout ce que dessus sera fait de suite* (Code civil, art. 976) de l'autre.

conforme à ces cas est impossible, car nous n'avons aucun moyen de différencier ces éléments une fois qu'ils sont étiquetés.

4.1.2. Définition de Le Goffic

Dans ses travaux systématiques sur les termes en *qu-*, Le Goffic définit un peu différemment les connecteurs. D'après lui, les termes en *qu-* – d'une vieille famille indo-européenne en *kw-*, fondamentalement indéfinis – sont, avec *si*, les seuls connecteurs du français. On en distingue trois types : pronoms, adjectif et adverbes.

- pronoms : *qui, que, quoi, lequel* ;
- adjectif : *quel* ;
- adverbes : *où, quand, comme, comment, combien, que, dont, pourquoi*.

La principale particularité de cette définition réside dans l'absence de catégorie de conjonction. Le Goffic renonce également à la notion de locution conjonctive, en dénonçant le « caractère peu satisfaisant » de leur liste traditionnelle et l'absence de véritable analyse des propositions considérées comme introduites par ces locutions conjonctives (Le Goffic 1993b : 75). Dans ses travaux, les unités introduites par une locution conjonctive sont analysées comme des groupes adverbiaux ou des groupes prépositionnels comprenant une subordonnée introduite par un véritable connecteur en *qu-*. Par exemple, *pour que P* est analysé non pas comme une subordonnée, mais comme un groupe prépositionnel constitué de la préposition *pour* suivi d'une complétive introduite par *que* ; *du moment où P*[7] est analysé comme un groupe prépositionnel contenant une relative introduite par *où* ; *aussitôt que P* est analysé comme un groupe adverbial constitué de l'adverbe *aussitôt* suivi d'une intégrative corrélative[8].

Cette analyse permet un traitement unifié et homogène des unités « propositions ». Mais son plus grand atout pour notre finalité est d'annuler, par exclusion de la catégorie de conjonction, le caractère polycatégoriel de la plupart des connecteurs, facilitant ainsi considérablement l'étiquetage automatique.

4.2. Typologie des subordonnées

Nous analysons ici différents types de typologies existantes des subordonnées. Avant d'entrer dans l'examen de ces travaux, clarifions nos besoins spécifiques pour la typologie des subordonnées, liés à la nature appliquée de notre opération.

---

[7] La définition des locutions conjonctives varie selon les grammaires. Par exemple, dans Riegel *et al.* (1994 : 504), les expressions pour lesquelles « il est possible de faire une analyse plus fine » ne sont pas considérées comme des locutions conjonctives. C'est le cas de cet exemple *du moment où*, alors que dans beaucoup d'autres grammaires telles que Wagner et Pinchon (1991), cette expression est classée comme locution conjonctive (voir section 4.1.1).

[8] Pour les types de subordonnées définies par Le Goffic, voir section 4.2.3, et pour la description détaillée de l'analyse, voir Le Goffic (1993a et 1993b).

4.2.1. Prérequis spécifiques à nos travaux

Nous avons deux points à considérer pour le choix de la typologie. Premièrement, comme pour tous les travaux de traitement automatique, la description doit être systématique et précise. Deuxièmement, du fait de la difficulté d'étiquetage que nous avons abordée dans la section précédente (4.1), il est préférable que la typologie ne présuppose pas une analyse correcte des connecteurs selon la catégorisation classique.

4.2.2. Typologies classiques

Dans les éditions postérieures à la 11e édition du *Bon usage*, refondue par Goosse, telles que Grevisse (1993), les propositions sont divisées en trois catégories, selon la nature du connecteur.

1. Propositions relatives : commençant par un pronom relatif (*qui, que, quoi, dont, où, lequel, quiconque*) ou par un syntagme contenant le pronom relatif ou parfois par un nom accompagné d'un déterminant relatif.
   (a) Relatives sans antécédent
   (b) Relatives avec antécédent
2. Propositions conjonctives : commençant par une conjonction ou une locution conjonctive de subordination.
   (a) Propositions conjonctives essentielles
   (b) Propositions corrélatives
   (c) Propositions adverbiales
3. Propositions d'interrogation et d'exclamation indirectes[9] : ne sont rattachées à la phrase par aucun mot particulier, à l'exception de l'interrogation globale, qui est rattachée à la phrase par la conjonction de subordination *si*.

Beaucoup de grammaires (Bescherelle 1990 ; Gardes-Tamine 1998 ; Wagner et Pinchon 1991) proposent une typologie comparable et le tableau 2 est la synthèse de ces dernières.

Grevisse-Goosse	relatives		conjonctives essent./corrél.	conjonctives adverbiales	interr./exclam. indirectes
Bescherelle	relatives		complétives	circonstancielles	interrogations indirectes
Gardes-Tamine	relatives substantives	relatives adjectives	conjonctives pures	conjonctives circonstancielles	interrogatives indirectes
Wagner et Pichon	relatives		conjonctives	circonstancielles	interrogations indirectes

Tableau 2. Correspondance des classes de subordonnées

Néanmoins, sans parler de la question théorique liée à la notion discutable de relative sans antécédent, la définition de Grevisse-Goosse présentée ci-dessus annonce elle-même les inconvénients de cette typologie pour notre tâche : elle se base sur une

[9] L'auteur appelle ce type de propositions soit propositions d'interrogation/exclamation indirectes (Grevisse 1993 : 1580), soit propositions interrogatives/exclamatives (Grevisse 1993 : 1682).

analyse correcte de la nature du connecteur, très difficile à réaliser de manière automatique.

4.2.3. Typologie proposée par Le Goffic

Comme nous l'avons déjà mentionné dans la section 4.1., Le Goffic décrit d'une manière systématique tout emploi des mots en *qu-*. Dans ses travaux, il propose un classement en quatre types de subordonnées introduites par ces connecteurs : complétive, relative, intégrative et percontative.

1. Complétive
   Je crois qu'il va pleuvoir.
2. Relative (relative avec antécédent)
   Le médecin qui est venu / La maison où je suis né
3. Intégrative
   - Pronominale (relative sans antécédent)
     Qui dort dîne / Embrassez qui vous voulez.
   - Adverbiale (circonstancielle en *qu-* ou *si*) :
     Quand on veut, on peut. / Si vous avez fini, vous pouvez sortir. / Il est à peine sorti qu'il a commencé à pleuvoir.
4. Percontative (interrogative/exclamative indirecte)
   Je sais qui a gagné / où il est allé / comment il l'a fait.

Malgré tout l'intérêt théorique qu'elle présente, cette typologie ne permet pas pour autant la conception d'un système simple de détection automatique des propositions. En effet, le problème est que, dans cette théorie, les connecteurs possèdent différents emplois dans lesquels chaque connecteur introduit différents types de subordonnées. Autrement dit, tout en facilitant l'opération d'étiquetage, la catégorisation des connecteurs ne permet pas directement de repérer chaque type de propositions ainsi définie et l'identification des subordonnées nécessite une étape supplémentaire dédiée à l'analyse de l'emploi exact du connecteur dans le contexte où il est utilisé. Ce qui représente, finalement, une tâche aussi délicate que l'étiquetage avec la catégorisation classique des connecteurs.

4.2.4. Autres typologies

Il existe dans la littérature deux autres types de classements : classement par catégorie du mot simple équivalent et typologie selon la fonction de la subordonnée dans la principale.

- Typologie selon la catégorie du mot simple équivalent (*Le bon usage* 11e éd. ; Biskri et Desclés 2005) :
  1. Substantive : Je pense qu'il viendra. / Que tu m'aimes me réjouit
  2. Adjective : La femme que tu vois / La ville où j'habite
  3. Adverbiale : Il était déjà rentré quand je suis arrivé.

- Typologie selon la fonction de la subordonnée dans la principale (Chevalier *et al.* 1964 ; Grevisse 1969 ; Wilmet 1997) :

1. Sujet : Que je sois malade ne l'a jamais effleuré.
2. Attribut : La triste vérité est qu'il est fou.
3. Objet : Marie sait que Paul viendra.
4. Circonstancielle : Il était déjà rentré quand je suis arrivé.
5. Complément de nom : la certitude que son but était atteint

Etc.

Ces typologies, qui ne se fondent pas sur les types de connecteurs qui introduisent les subordonnées, semblent mieux adaptées à la définition d'une grammaire pour la détection des propositions. Cependant, chacune possède d'autres problèmes.

Pour la typologie selon la catégorie, comme il l'a été signalé dans la grammaire de Riegel *et al.* (1994 : 476), le parallélisme des catégories n'est que partiel : les relatives ne peuvent, par exemple, pas assurer la fonction d'attribut en dépit de leur apparente équivalence à l'adjectif. Par ailleurs, sur le plan pratique, dans le cadre de nos travaux, cette typologie qui classe les subordonnées dans seulement trois catégories risque de multiplier le nombre d'analyses possibles d'une phrase. Par exemple, une subordonnée en *que* peut être substantive, adjective ou adverbiale[10], et avec ces trois possibilités, le nombre d'analyses possibles d'une phrase qui en contient une risque d'être très important, surtout avec la quantité restreinte d'information dont nous disposons pour l'analyse. La discrimination d'un type par rapport aux autres selon la fréquence est impossible, car il n'existe pas d'homogénéité même à l'intérieur d'un type : une subordonnée substantive en *que*, par exemple, est extrêmement fréquente à la fonction de complément, mais elle l'est moins à la fonction de sujet. Ainsi, une fois tous les candidats calculés, une étape supplémentaire serait nécessaire pour choisir la réponse la plus probable. Mais il nous semble possible, en définissant une typologie tenant compte d'autres critères, de contrôler plus efficacement le nombre d'analyses possibles et d'obtenir la réponse la plus probable sans ajout d'étapes supplémentaires.

Quant aux typologies selon la fonction, elles divisent les subordonnées en un nombre plus ou moins important. Néanmoins, une économie des règles de grammaire serait sans doute envisageable par une restriction des types non pertinents pour notre opération de détection des propositions, qui réduirait ainsi les calculs nécessaires. Mais le plus grand défaut de ces typologies est que les structures de subordination n'y sont décrites que partiellement avec seulement des exemples triviaux, ne nous fournissant pas suffisamment d'informations. En effet, nous ne pouvons pas savoir exactement quelles sont les subordonnées de fonction sujet, complément, etc. Une description précise permettrait sans doute de mieux rendre compte des points communs et des divergences entre les types et, avec cette étude, de réorganiser la typologie afin d'en obtenir une plus économique et suffisamment efficace.

[10] Pour les exemples de chaque cas, voir la section 5.

5. Notre typologie des subordonnées selon la position

Face à ces problèmes des typologies existantes, nous avons défini une typologie selon la position d'apparition. À cette fin, nous avons d'abord étudié pour chaque position de la phrase les catégories de subordonnées susceptibles d'y apparaître. Ensuite, nous nous sommes appuyée sur les travaux de Le Goffic afin d'élaborer la description de chaque type de manière à obtenir un caractère suffisamment complet pour fournir une base pour la définition d'une grammaire globale et formelle. Notre typologie présente tout d'abord comme avantage l'indépendance vis-à-vis de la qualité d'analyse des connecteurs dans les catégories classiques ou de leur emploi exact. De plus, le critère de position a permis d'obtenir une division plus optimisée pour définir une grammaire qu'avec les deux critères traditionnellement utilisés, la catégorie et la fonction.

Nous distinguons quatre types de subordonnées selon leur position dans la phrase :

1. Proposition en position post-verbale (fonction de complément) : concerne les propositions substantives ;
2. Proposition à une autre position pouvant être occupée par un SN (syntagme nominal) (fonction de sujet et autres) : concerne les propositions substantives ;
3. Proposition en positions initiale et finale (fonction accessoire) : concerne les propositions adverbiales ;
4. Proposition en position post-nominale (fonction secondaire[11]) : concerne les propositions adjectives et adverbiales.

Chaque type de subordonnée à une position donnée est caractérisé par sa fréquence, afin de pouvoir favoriser l'interprétation comme subordonnée courante par rapport aux subordonnées rares. Faute de données permettant d'obtenir des statistiques représentatives, la définition de ces fréquences est réalisée de manière empirique. La justesse de ces hypothèses est examinée dans l'évaluation (section 7.2.).

Avant de présenter de manière plus détaillée chaque type de subordonnée, faisons un point sur ce que nous appelons connecteur de subordination. Nous adoptons, dans un premier temps, la catégorisation des connecteurs sans classe de conjonction définie par Le Goffic (une autre plus adaptée à nos travaux sera présentée dans section 6). Toutefois, pour des raisons pratiques – notamment par souci d'utilité pour l'alignement –, nous conservons le statut de connecteur locutionnel que nous attribuons aux locutions conjonctives, regroupées et étiquetées comme CS (conjonction de subordination) par notre *tagger*.

[11] Le terme fonction secondaire est emprunté à (Le Goffic 1993a : 71) : « Les fonctions primaires se situent au niveau de la phrase (exemples : sujet, attribut, circonstant), et les fonctions secondaires au niveau (interne) des constituants de la phrase (exemples : complément de nom, compléments d'adjectif). »

5.1. Position post-verbale : subordonnée complément en *qu-* (subQ)

À cette position, apparaissent les propositions substantives : complétives, intégratives pronominales, percontatives.

- S**ubstantives**

(a) Complétives

Je pense qu'il viendra.

(b) Intégratives pronominales

Embrassez qui vous voulez.

(c) Percontatives

Je me demande s'il est parti.

Il ne m'a pas dit quand il rentrerait.

Voyez comme c'est facile.

5.2. Autres positions SN : subordonnée SN (subSN)

Les propositions substantives apparaissent également, bien qu'assez rarement, à d'autres positions où un syntagme nominal peut apparaître : position sujet, après une préposition, position initiale (termes en prolepse).

1. Position sujet : substantives (rare)

(a) Intégrative

Qui dort dîne.

(b) Complétive

Que vous ayez menti me déçoit.

(c) Percontative

Qui a commis ce crime n'a jamais été établi.

Comment il a commis ce crime n'a jamais été établi.

Pourquoi il a commis ce crime n'a jamais été établi.

2. Après une préposition : substantives

(a) Intégrative (rare)[12]

Le pouvoir est seulement entre les mains de qui détient des armes à feu, de qui possède les richesses.

Pour qui appartient aux classes moyennes, le fait de partir de chez soi chaque matin est un combat.

(b) Percontative (rare)[13]

Il faudra se poser la question de pourquoi nous avons été choisis.

Plus récemment se pose la question de comment l'État doit considérer les groupes et minorités défavorisés, s'il souscrit à l'idéal de traiter tous les citoyens et citoyennes comme égaux, indépendamment de leur appartenance sexuelle, religieuse ou ethnique.

[12] Les exemples d'intégrative sont tirés du *Monde diplomatique*.

[13] Le premier exemple est emprunté à un article publié sur *Yahoo ! France*. Les autres exemples sont des résultats de requêtes dans *Google*.

Ce n'était plus une question de « si » mais bien une question de « quand » une telle échéance allait se produire.

(c) Complétive (fréquente dans les structures à locution conjonctive, voir l'explication ci-dessous)

après que, avant que, depuis que, dès que, malgré que, pendant que, pour que, sans que, sauf que, selon que, etc.

3. Position initiale en prolepse[14] : substantives

(a) Intégrative pronominale

Qui ferait cela, il agirait sagement. (Repris de Le Goffic (1993a), obsolète)

(b) Percontative

Comment il a fait, je vous le demande ! (Repris de Le Goffic (1993a))

(c) Complétive

Qu'il y eût en tout être, et en lui d'abord, un paranoïaque, il en était assuré depuis longtemps. (Repris de Chevalier *et al.* (1964))

Après une préposition (cas 2), les intégratives et les percontatives sont rares. Les complétives y sont utilisées très fréquemment, mais l'analyse classique considère qu'elles constituent avec les prépositions qui les précèdent des subordonnées circonstancielles à locution conjonctive et non des complétives. Dans nos travaux, ces locutions étant regroupées et étiquetées par le *tagger* comme conjonctions de subordination, nous ne devrions pas rencontrer à cette position de complétives non regroupées avec des prépositions. Or, la liste sur laquelle se base l'étiqueteur peut être incomplète. Nous conservons donc la possibilité d'avoir après une préposition une complétive – avec comme indication de fréquence rare – afin de pouvoir détecter la proposition introduite par la locution conjonctive que l'étiqueteur n'a pas réussi à regrouper.

L'extériorité des propositions substantives en prolepse (cas 3) est si forte que la percontative en prolepse, en particulier, « peut aussi être interprétée comme une interrogation indépendante » (Le Goffic 1993a : 381)[15]. Par ailleurs, comme le signale Le Goffic (1993a : 381), « le français a perdu depuis l'époque classique l'usage des intégratives pronominales en prolepse ».

5.3. Positions initiale et finale : subordonnée circonstancielle ou périphérique (subP)

Ces positions concernent l'ensemble des propositions à locution conjonctive et les intégratives – les seules que nous étudions. Par ailleurs, apparaissent également les propositions en *que* analysées souvent comme subordonnées paratactiques.

[14] Un élément en prolepse est « jeté en avant, posé pour lui-même, hors fonction et hors structure, comme si l'énonciateur commençait par indiquer le ou les objet(s) de son discours, avant même d'avoir arrêté un projet de phrase syntaxique. » (Le Goffic 1993a : 380)

[15] Même remarque dans Chevalier *et al.* (1964 : 120).

1. Adverbiales : intégratives adverbiales
- Position initiale

Quand je suis arrivé, il était déjà rentré.
Si tu ne manges pas, tu ne guériras pas.
Comme elle est écrite en chinois, il n'a pas pu lire cette lettre.
Où il y a de la gène, il n'y a pas de plaisir. (Repris de Le Goffic (1993a), rare)

- Position finale

Il était déjà rentré quand je suis arrivé.
Tu ne guériras pas si tu ne manges pas.
Il n'a pas pu lire cette lettre comme sa mère l'avait deviné.
Tu peux poser ton manteau où tu veux.
Il aurait bu que je n'en serais pas surpris. (Repris de Le Goffic (1993a))
Viens ici, que je t'embrasse. (Repris de Le Goffic (2003a))
Le crocodile n'eut pas le temps de se demander ce que lui voulait ce lourdaud, que Gropopotin s'était déjà assis sur son dos[16].
La maison est restée aussi conviviale qu'elle l'était avant.
La nouvelle l'a tellement surprise qu'elle s'est mise à pleurer.

2. Adverbiales/substantives ? : intégratives/complétives ?

Que le gouvernement propose une nouvelle loi, l'opposition crie au scandale.
Je pars, que cela vous plaise ou non.

Les propositions intégratives adverbiales (cas 1) sont celles que Le Goffic considère comme les seules subordonnées méritant le nom de « circonstancielles » (Le Goffic 1993b : 69-70). Elles peuvent être caractérisées – à l'exception des intégratives en *que* qui n'apparaissent qu'en position finale – par leur liberté liée à la position de leur occurrence : elles peuvent apparaître non seulement aux positions initiale et finale, mais elles peuvent aussi être insérées, sous forme détachée, entre différents éléments de la phrase. Comme le remarque Le Goffic (1993a : 393), l'intégratif en *où* est rare. Bien que *comme* apparaisse aussi bien en position initiale que finale, son interprétation diffère dans les deux cas (Le Goffic 1993a : 482-484).

Les subordonnées paratactiques en *que* (cas 2) sont des propositions très délicates à analyser. Le Goffic analyse ce *que* comme complétif « dans un emploi de caractère nominal et paratactique » (Le Goffic 1993b : 74). Or ces propositions paratactiques se déplaçant librement nous donnent une impression plus proche de celle des adverbes circonstanciels, et sont souvent considérées comme circonstancielles (Guimier 1993). Afin d'éviter tout jugement prématuré et non suffisamment étudié, nous laissons en suspens l'analyse exacte de cette proposition et du connecteur *que* apparaissant ici.

[16] Repris de *Gropopotin l'hippopotame*, « Wakou », numéro 206, mai 2006.

5.4. Position post-nominale : subordonnée déterminante (subD)

À cette position apparaissent non seulement les propositions adjectives (relatives, complétives), mais aussi, quoique rarement, les propositions adverbiales (intégratives adverbiales), et percontatives en *si*.

1. Adjectives
(a) Relative
La peinture qui m'a fascinée
La peinture dans laquelle notre maison était reproduite
Ce à quoi je m'attendais
(b) Complétive
L'idée que tout est fini
2. Adverbiales : intégratives adverbiales (rare)
La déception du père quand il a entendu cette nouvelle
La fille comme je l'avais imaginée
3. Percontatives en *si* (rare)
Son incertitude s'il devait obéir (Repris de Le Goffic (1993a))

5.5. Autres positions : post-adjective et post-adverbiale

Ces positions ne concernent que les propositions en *que*, complétif, corrélatif ou relatif. Elles sont généralement analysées comme des propositions introduites par une locution conjonctive. Notre description suit l'analyse faite par Le Goffic (*cf.* section 4.1.2.).

- Post-adjective
1. Intégrative (corrélative) : *de même que* ;

- Post-adverbiale
1. Complétive : *à moins que, loin que, cependant que, bien que, déjà que, encore que*[17]*, même que, non que, sinon que, surtout que* ;
2. Relative : *alors que, aujourd'hui que, dès lors que, maintenant que* ;
3. Intégrative (corrélative) : *ainsi que, aussi longtemps que, aussitôt que, d'autant plus/moins que , d'autant que, plutôt que, si bien que, sitôt que, tant que*.

Tout comme les complétives suivant une préposition, ces subordonnées apparaissent rarement à ces positions sans être constituant d'une proposition à locution conjonctive. Nous gardons donc, principalement pour les propositions introduites par une locution conjonctive que l'étiqueteur n'a pas réussi à regrouper et pour quelques autres cas, la possibilité d'une subordonnée en *que* après un adverbe ou un adjectif – avec, comme indication de fréquence, *rare*.

[17] Il est à noter que l'article de Fuchs (1992) signale l'existence de *encore que* avec *que* non complétif mais corrélatif.

6. Notre typologie des connecteurs

6.1. Classement basé sur la typologie des subordonnées selon la position

En nous basant sur l'étude des subordonnées selon les positions d'apparition décrites dans la section précédente, nous avons réalisé une classification des connecteurs. Le tableau 3 présente la synthèse de cette étude.

✔ = fréquent
△ = rare

	Position	post-V				post-N				Int /Fin				pos. SN				Autres			
		I	C	P	R	I	C	P	R	I	C	P	R	I	C	P	R	I	C	P	R
C. isolés	qui	△		✔					✔					△		△					
	que		✔				✔		✔	✔					△			△	△		△
	dont								✔												
	où			✔		?			✔	△						△					
C.amb.	quand			✔		△				✔						△					
	comme			✔		△				✔						△					
	si			✔				△		✔						△					
C. rel.	quoi			✔					✔							△					
	lequel			✔					✔							△					
Ident. prop.	quel			✔												△					
	combien			✔												△					
	comment			✔												△					
	pourquoi			✔												△					

I = Intégrative, C = Complétive, P = Percontative, R = relative

Tableau 3. Typologie des connecteurs

De ce constat, nous avons défini les quatre types de connecteurs suivants.

1. Connecteurs isolés
   qui, que, dont, où
   ayant un comportement particulier et dont les positions d'occurrence ne sont comparables avec aucun des autres connecteurs ;
2. Connecteurs ambigus
   quand, comme, si
   apparaissant fréquemment aux deux positions (post-V, Int/Fin) et rarement aux deux positions (post-N, SN) ;
3. Connecteurs relatifs
   quoi, lequel (et ses formes fléchies)
   apparaissant fréquemment aux deux positions (post-V, post-N) et rarement aux positions SN ;
4. Indicateurs de propositions
   quel (et ses formes fléchies)**,** *combien, comment, pourquoi*
   apparaissant fréquemment seulement en position post-V et rarement aux positions SN.

6.2. Connecteurs composés

Les connecteurs particuliers à caractère de déterminant tels que *quel*, *combien*, ainsi que le connecteur *lequel* pouvant être suivi d'un complément secondaire, doivent être traités différemment des autres. Lorsque ces connecteurs constituent un syntagme avec le substantif qui les suit – le premier type en tant que complément secondaire et le second en tant que tête du syntagme –, nous considérons qu'ils constituent avec le syntagme nominal qui les suit un connecteur composé. Par exemple, dans la phrase *combien d'habitants compte Tokyo*, *combien de* est un connecteur déterminant qui constitue avec le syntagme nominal qui le suit, *habitants*, un connecteur composé *combien d'habitants*. Dans la phrase *lequel de ces romans n'est pas de Voltaire*, *lequel*, connecteur tête, constitue avec son complément secondaire *de ces romans* le connecteur composé. En revanche, lorsqu'ils fonctionnent tout seuls comme dans la phrase *combien coûtait cette bêtise* ou *quel était son intérêt*, nous les traitons comme des connecteurs simples.

7. Réalisation

Nous ne présentons ici la réalisation informatique et les résultats d'évaluation que très brièvement : les détails sont déjà présentés dans notre précédente communication (Nakamura-Delloye 2006) consacrée à cet aspect de nos travaux.

7.1. Développement informatique et résultats obtenus

Les études linguistiques présentées jusqu'ici nous ont permis de définir une grammaire pour la détection des propositions de type CFG (*Context-Free Grammar*). Cette grammaire a ensuite été réécrite selon le formalisme DCG (*Definite Clause Grammar*) et incluse dans un module principal développé en Prolog. Pour favoriser l'interprétation des subordonnées fréquentes plutôt que des subordonnées rares, les règles définissant les subordonnées rares ne sont insérées dans la grammaire que dynamiquement lorsque l'analyse avec seulement les règles des subordonnées fréquentes a échoué.

Nous avons obtenu des résultats assez satisfaisants avec des taux de rappel[18] de 81 à 96 % et des taux de précision[19] de 87 à 98 %. Les erreurs sur la détection des frontières de propositions, à part celles provenant des traitements antérieurs, se limitent essentiellement aux phrases contenant plusieurs virgules dans les structures de coordination. Les erreurs sur la détermination des relations entre les propositions détectées se distinguent en trois types, mais tous les trois montrent bien la limite de l'analyse avec des informations très restreintes. Cependant, l'enrichissement des informations se traduit directement par un calcul très coûteux, ce qui risque de rendre

[18] Le rappel est défini comme la proportion du nombre de phrases dont l'analyse a abouti sur le nombre total de phrases.

[19] La précision correspond au taux de phrases dont la détection des frontières et l'analyse des relations sont correctes sur le nombre total d'analyses de phrases ayant abouti.

le système peu opérationnel. L'introduction de l'opposition fréquent/rare des subordonnées a permis une amélioration de la précision de manière économique et assez efficace.

7.2. Fréquence des subordonnées

Pour vérifier nos hypothèses concernant l'opposition fréquent/rare des subordonnées, nous avons compté manuellement les occurrences de chaque type de subordonnées dans le résultat de deux corpus[20] : LMD (qui contenait au total 501 connecteurs de la famille *qu*-) et du corpus Zadig (516 connecteurs).

Occurrence = %

		Int/Fin			post-V			post-N			Autres SN		
		HYP	LMD	ZDG	HYP	LMD	ZDG	HYP	LMD	ZDG	HYP	LMD	ZDG
Sub.	Intégrative pro.				△	0	0				△	0,4	0,8
Sub.	Percontative				✔	2	4				△	0	0
Sub. / Adj.	Complétive				✔	20	27	✔	2	0,3	△	0,2	0
Adj.	Relative							✔	60	57			
Adv.	Intégrative adv.	✔	15	11				△	0	0			

✔ = fréquent ; △ = moins fréquent / rare

Tableau 4. Fréquence des subordonnées

Le tableau 4 est un tableau comparatif présentant nos hypothèses (colonne HYP) et le résultat de comptage (colonnes LMD et ZDG). Le résultat de ce comptage a confirmé à peu près notre définition des qualificatifs *rare/fréquent* des subordonnées.

En dépit de ce à quoi nous nous attendions, nous n'avons pas constaté de très grandes différences entre ces deux corpus de nature différente. La différence est constatée à un niveau plus précis entre les connecteurs employés dans chaque catégorie.

8. Pistes d'amélioration

Nous pouvons imaginer deux grandes pistes d'amélioration : perfectionnement d'analyse par l'introduction de plus d'informations et affinement des étiquettes par la réalisation d'une analyse sémantique des connecteurs.

[20] Le corpus Zadig est constitué du texte intégral de *Zadig* de Voltaire, réalisé à partir de la version électronique distribuée par Olivier Tableau (disponible sur Internet). Le corpus LMD est constitué de 15 articles du *Monde diplomatique*, tirés des numéros de janvier, février et mars 2004 (édition informatique).

8.1. Exploitation de plus d'informations

Les résultats pourraient être améliorés par l'utilisation de plus d'informations. Mais l'introduction d'informations supplémentaires signifie une augmentation des calculs nécessaires. L'important est de savoir tracer la limite afin de ne pas perdre l'intérêt d'un petit outil opérationnel. Par ailleurs est également envisageable une amélioration par l'utilisation d'autres outils extérieurs fournissant certaines informations précises très fiables, tels que ILIMP (Danlos 2005), étiqueteur du pronom impersonnel *il*, qui permettrait à notre système d'améliorer le résultat de la détection des complétives suivant la proposition principale à *il* impersonnel.

8.2. Affinement des étiquettes

Le second type d'amélioration est l'affinement des étiquettes pour les propositions détectées. En effet, comme on peut le constater en se référant au tableau 3, après la détection des propositions selon la position, on peut déterminer finalement une nature de proposition plus usuelle de caractère syntactico-sémantique. Par exemple, dans le cas où la subordonnée détectée contient le connecteur *quand*, il est possible de déduire sans aucune ambiguïté qu'elle est intégrative à valeur temporelle, si elle est en position post-nominale ou en position initiale/finale, ou qu'elle est percontative (interrogative à valeur temporelle), si le connecteur est en position post-verbale ou dans une autre position SN.

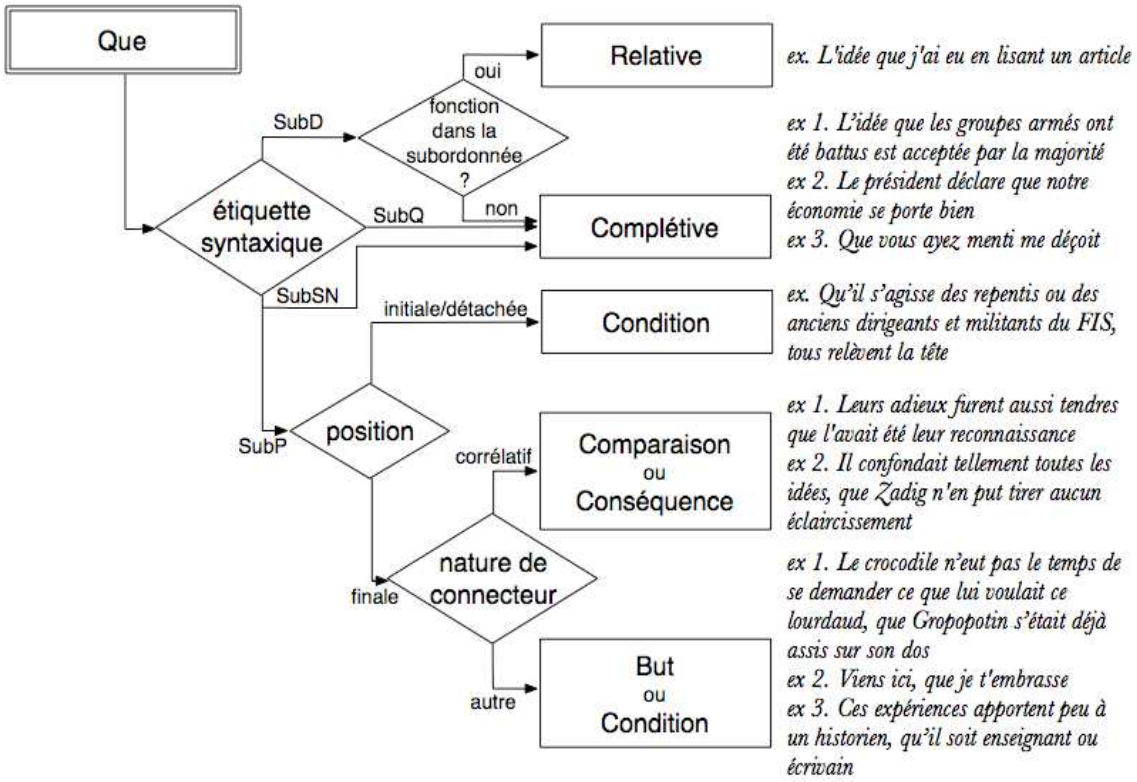

Figure 5. Subordonnées en « que »

Examinons le cas du connecteur le plus ambigu *que* (*cf.* figure 5). Une subordonnée en *que* reçoit l'étiquette syntactico-sémantique « relative » ou « complétive » si son étiquette syntaxique est subD (subordonnée déterminante), ou complétive si son étiquette syntaxique est subQ (subordonnée complément) ou subSN. Dans le cas où son étiquette syntaxique est subP (subordonnée circonstancielle), elle a une valeur de condition si elle est en position initiale ou détachée, et elle a une valeur de comparaison, de conséquence, de but ou de condition, si elle est en position finale.

L'attribution des étiquettes syntactico-sémantiques favoriserait sans doute l'alignement et permettrait peut-être d'élargir le champ d'application. Néanmoins, la définition de la valeur sémantique des subordonnées circonstancielles est très délicate et l'analyse varie selon les linguistes. Il nous serait donc nécessaire de réaliser une étude beaucoup plus approfondie et une évaluation du rapport entre faisabilité et apport réel.

9. Conclusion

Les résultats de notre système semblent confirmer que sont utiles voire indispensables la remise en cause des habitudes classiques ainsi que le recours aux travaux linguistiques pointus sur des sujets connexes. L'utilisation des connaissances linguistiques favorise par ailleurs le développement non seulement d'un grand système complet, mais aussi d'un petit outil qui peut fournir une analyse assez détaillée sur un sujet précis avec des informations très limitées.

Avec ces résultats de détection des propositions, nous travaillons actuellement sur l'alignement des propositions des textes parallèles français-japonais. Nous espérons avec ces travaux fournir des données intéressantes pour les études contrastives de ces deux langues très différentes au niveau aussi bien lexical que syntaxique, et contribuer au développement des traitements multilingues entre les langues à dominance opposition sujet-prédicat[21] et les langues à dominance opposition thème-rhème.

Références

ABEILLÉ A. et CLÉMENT L. (2003), *Annotation morphosyntaxique, Les mots simples – les mots composés, corpus Le Monde.* www.llf.cnrs.fr/fr/Abeille/guide-morpho-synt.02.pdf.

BISKRI I. et DESCLES J. P. (2005), « Analyse de la coordination et de la subordination au moyen de la grammaire catégorielle combinatoire applicative », in *Colloque Subordination-Coordination*, http://www.cavi.univ-paris3.fr/ilpga/colloque-coord-subord-2005/pre-textes/index.html.

CHARROLES M. (1997), « L'encadrement du discours : univers, champs, domaines et espaces », in *Cahier de recherche linguistique* 6 : 1-73.

CHEVALIER J.-Cl., BLANCHE-BENVENISTE Cl., ARRIVÉ M. et PEYTARD J. (1964), *Grammaire du français contemporain*, Larousse, Paris.

[21] Voir par exemple Li et Thompson (1976).

COMBETTES B. (1998), *Les constructions détachées en français*, Ophrys, Paris.

DANLOS L. (2005), « ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il* », in *Actes de TALN 2005*, vol. 1 : 123-132.

DELAVEAU A. (2001), *Syntaxe. La phrase et la subordination*, Armand Colin, Paris.

FUCHS C. (1992), « Les subordonnées introduites par encore *que* en français », in *Subordination (Travaux Linguistiques du CERLICO)* 5 : 89-110.

GARDES-TAMINE J. (1998), *La Grammaire 2. Syntaxe*, Colin, Paris.

GREVISSE M. (1969), *Cours d'analyse grammaticale*, De Boeck Duculot, Paris, sixième édition.

GREVISSE M. (1993), *Le bon usage, grammaire française*, Duculot, Paris, treizième édition. Édition par André GOOSSE.

GUIMIER C. (1993), « L'établissement d'un corpus de circonstants », in Guimier C. (éd.), *1001 circonstants*, Presses Universitaires de Caen, Caen : 11-45.

HATIER (éds) (1990), *La grammaire pour tous*, Hatier, Paris (Bescherelle 3).

LE GOFFIC P. (1993a), *Grammaire de la phrase française*, Hachette, Paris.

LE GOFFIC P. (1993b), « Les subordonnées circonstancielles et le classement formel des subordonnées », in GUIMIER C. (éd.), *1001 circonstants*, Presses Universitaires de Caen, Caen : 69-102.

LI C.N. et THOMPSON S.A. (1976), « Subject and topic : a new typology of language », in Li C.N. (éd.), *Subject and topic*, Academic Press Inc, New York : 457-491.

NAKAMURA-DELLOYE Y. (2006), « Détection des propositions syntaxiques du français », in *Actes de TALN 2006* vol. 1 : 551-560.

RIEGEL M., PELLAT J.-C et RIOUL R. (2004 [1994]), *Grammaire méthodique du français*, PUF, Paris.

WAGNER R.L. et PINCHON J. (1991), *Grammaire du français*, Hachette supérieur, Paris.

WILMET M. (1997), *Grammaire critique du français*, Hachette supérieur, Duculot, Paris.