Coordination et analyse automatique du français écrit dans le cadre de la Grammaire Lexicale Fonctionnelle

Caroline Brun

Caroline Brun est post-doctorante au Centre de recherche Xerox Europe de Grenoble. Elle a y effectué une thèse sur l'analyse automatique de la coordination dans le formalisme LFG (Lexical Functional Grammar)*, dans le cadre d'une convention Cifre avec le laboratoire Gresec-Cristal. Elle travaille aujourd'hui sur la désambiguïsation sémantique automatique et la génération automatique de documents multilingues.*

Plan

Introduction et cadre de travail
Étude linguistique de la coordination
Modélisation
Implantation
Conclusion
Références bibliographiques

INTRODUCTION ET CADRE DE TRAVAIL

Le travail de thèse [Brun98a] que nous proposons de résumer dans cet article a été effectué dans le cadre d'une collaboration entre le laboratoire Gresec (équipe Cristal) de l'université Stendhal-Grenoble 3 et l'équipe MLTT du centre de recherche Xerox Europe (XRCE).

L'activité de l'équipe Cristal est fondée sur le traitement automatique de la langue écrite (TALN), essentiellement le français. Le but poursuivi est la conception et la réalisation d'analyseurs et de générateurs à large couverture linguistique, donc capables de supporter des affirmations en vraie grandeur. Un autre ensemble de recherches est constitué par l'élaboration et l'implantation d'un système de traitement des connaissances lié à des textes en langue naturelle.

L'équipe de MLTT crée des outils de base pour l'analyse linguistique multilingue, tels que analyseurs morphologiques, étiqueteurs morpho-syntaxiques, plateformes pour le *parsage* (analyse syntaxique automatique) et la génération, ou encore outils d'analyse de corpus. Ces outils sont utilisés pour décrire diverses langues ainsi que les relations d'une langue à l'autre. Les projets en cours incluent des analyseurs syntaxiques à états finis pour le français et l'allemand, une grammaire LFG (*Lexical Functional Grammar*) du français ainsi que des projets en recherche documentaire multilingue, en génération et en traduction.

Ce travail est rattaché au développement d'une grammaire LFG du français, pour l'analyse et la génération de texte, dans le contexte d'un projet plus vaste, le projet ParGram (1) [Butt et al. 99], visant à développer des grammaires LFG les plus parallèles possibles pour le français, l'anglais, et l'allemand, l'application visée étant la traduction automatique entre ces trois langues.

. . . . . .

1. http://www.parc.xerox.com/istl/groups/nltt/pargram/

Dans ce contexte, nous étudions une construction particulière, la coordination. La coordination est un phénomène complexe à analyser quel que soit le modèle choisi, mais est aussi très largement représentée dans les corpus de toutes natures. De plus, la coordination est un phénomène syntaxiquement ambigu susceptible de générer des solutions parasites lors du processus d'analyse. Il s'agit donc de développer un modèle d'analyse permettant une bonne couverture linguistique du phénomène tout en limitant au mieux le nombre de solutions parasites. Pour ces différentes raisons, ce sujet de recherche convenait parfaitement aux objectifs et intérêts communs des deux laboratoires partenaires.

Nous allons tout d'abord présenter l'étude linguistique du phénomène, réalisée sur corpus. Compte tenu de cette étude, un modèle d'analyse syntaxique général est proposé. Nous détaillons pour conclure l'implantation réalisée dans un système dédié à l'analyse automatique par des grammaires LFG, ainsi que l'évaluation sur corpus de cette implantation.

Étude linguistique de la coordination

Corpus d'étude

L'étude linguistique (nous étudions tout particulièrement les coordinations par *et* et *ou*) du phénomène repose sur l'examen de deux corpus en particulier :

- un corpus technique, qui est un manuel d'utilisation de tracteurs. Ce corpus a servi de référence dans le cadre du projet de développement de grammaires auquel nous sommes rattachée. Il s'agit d'un corpus de petite taille (549 phrases ou titres, 6651 mots, 12,11 mots par phrase en moyenne), dont nous avons extrait les phrases contenant des coordinations (124 phrases ou titres, 2344 mots, 18,9 mots par phrase en moyenne).
- Un corpus extrait du journal *Libération*, 1988, qui était également à notre disposition « en ligne ». Nous avons jugé intéressant d'examiner ce corpus car la langue employée est « non restreinte » et plus riche que celle du corpus technique, notre but étant décrire le phénomène le plus généralement possible. Il s'agit d'un corpus de taille importante (59987 phrases et titres, 1358797 mots, 22,65 mots par phrase en moyenne) dont nous avons extrait les phrases contenant des coordinations (17262 phrases ou titres, 510120 mots, 29,55 mots par phrase en moyenne).

Analyse générale

La première observation que l'on peut faire lorsque l'on étudie la coordination est qu'il s'agit d'un phénomène particulièrement fréquent, dans n'importe quel type de corpus. À titre indicatif, la conjonction de coordination *et* serait un des dix mots les plus fréquents de la langue française ([Grevisse 91]). On constate aussi que celle-ci s'applique à tous les niveaux de la phrase, c'est-à-dire qu'elle peut relier deux ou plusieurs propositions (P), mais aussi syntagmes nominaux (SN), des syntagmes prépositionnels (SP), des syntagmes adjectivaux (SA), etc. Nous avons également remarqué que des éléments de nature différente (SP & SA, SN & SP…) peuvent être coordonnés.

Cependant, les fonctions grammaticales des éléments coordonnés sont les mêmes (sujet, attribut du sujet, modifieur du syntagme verbal, complément d'objet, etc.). Dans la mesure où la coordination d'éléments de catégories identiques et fonctions différentes semble proscrite, il nous semble qu'un des critères permettant de vérifier si une coordination est grammaticale est l'identité fonctionnelle des éléments coordonnés, ce qui avait largement été constaté dans la littérature sur le sujet.

> *Page 2*

Nous citerons, entre autres :

> *[Dik72] : « Une coordination est une construction de deux membres ou plus qui ont des fonctions grammaticales semblables, et qui sont liés ensemble au même niveau de hiérarchie structurale par le biais d'un connecteur. »*

Il s'agira donc de rejeter les coordinations de fonctions différentes, mais d'admettre, sous certaines contraintes et en fonction de certains indices syntaxiques, les coordinations d'éléments de fonctions identiques. Une étude quantitative des corpus sur les différents types de coordinations montre néanmoins que la majorité des phénomènes observés (94 %) sont des coordinations d'éléments de même catégorie syntaxique.

Indices syntaxiques et contraintes

L'examen de corpus a également permis d'isoler un ensemble de contraintes grammaticales fortes s'appliquant à la coordination ainsi qu'un ensemble d'indices concernant les fréquences de certains phénomènes. Nous ne détaillerons pas ici l'ensemble de ces contraintes et indices, mais en isolerons quelques exemples. Ainsi, si *a priori* tout syntagme est susceptible de se coordonner avec un syntagme de même catégorie (ou de catégorie différente), certains cas sont soumis à des restrictions.

Par exemple, dans le contexte de la coordination, on doit nécessairement reprendre une conjonction de subordination par elle-même ou par la conjonction *que* (2) :

> *[Lorsque [[le démarreur est mis sous tension] et [l'alimentation du thermostart est maintenue]]]*
>
> **Cela indique [que [le démarreur est engagé et l'alimentation du thermostart est maintenue]]*

La reprise d'une locution prépositive par elle-même ou par la préposition qui lui correspond (*à* ou *de*) est obligatoire :

> *3. [[Grâce à sa conception unique] et [à sa taille réduite]],...*
>
> *4. *[[Grâce à sa conception unique] et [sa taille réduite]],...*

De la même façon, les pronoms clitiques et négations ne se distribuent pas sur les verbes finis :

> *5. *Vous pouvez les [numériser ou copier].*
>
> *6. *Il ne numérise ou copie pas*

La préposition *entre* sélectionne obligatoirement un complément coordonné (ou un complément pluriel) :

> *7. Entre [[les villes] et [les campagnes]]*
>
> *8. *Entre le pays.*

De plus, un certain nombre d'indices, qui ne constituent cependant pas des contraintes fortes sont observables, en particulier en ce qui concerne le parallélisme des structures coordonnées. Ainsi, dans la plupart des cas, les prépositions *à* et *de* se répètent devant chaque terme de la coordination, alors que les autres prépositions ont tendance à ne pas se répéter. Les déterminants se répètent le plus souvent sur la coordination de groupe nominaux. Lorsque des propositions sont coordonnées, le sujet est nécessairement exprimé ou nécessairement omis dans chacune d'elles. Si plusieurs structures coordonnées imbriquées mettent en jeu syntagmes prépositionnels et nominaux, on

. . . . . .

2. Le signe * indique qu'un exemple est incorrect.

> *Page 3*

observe la plupart du temps une similitude de la nature des déterminants (défini, indéfini, possessif, etc.) ou des prépositions.

MODÉLISATION

État de l'art

La coordination a fait l'objet de nombreux travaux, que ce soit dans le cadre de théories linguistiques, ou de développements de systèmes de TALN.

De l'étude du traitement de la coordination dans des formalismes comme la grammaire syntagmatique généralisée [Chomsky 57], les grammaires catégorielles [Steedman85], les formalismes d'unification [Gazard et al. 85], [Pollard et Sag 94], ou encore certains modèles « informatiques », comme les ATN [Woods 70], les grammaires logiques [Dahl et al. 95], les grammaires dynamiques [Milward 94], nous pouvons dégager quatre grands types de traitements :

- Les traitements par réduction, qui réalisent un rétablissement d'ellipse, poussés au maximum pour la grammaire transformationnelle, et qui posent de nombreux problèmes.
- Les traitements exhibant un constituant similaire dans tous les cas, ce que font les grammaires catégorielles. Ce traitement permet de prendre en compte un grand nombre de cas, mais pose des problèmes de surgénération et d'analyses ambiguës pour les applications pratiques.
- Les processus « métagrammaticaux », qui considèrent la phrase comme une suite de transitions, la coordination permettant de réaccéder des points dans l'historique de l'analyse (c'est à dire de « revenir en arrière » vers des mots déjà analysés, et d'utiliser l'information portée par ces mots pour reprendre l'analyse) : ce sont les traitements informatiques, qui semblent donner de bons résultats, mais posent aussi des problèmes de surgénération et d'analyses parasites.
- La coordination au niveau du constituant, proposée, entre autres, par les grammaires d'unification, qui semble linguistiquement justifiée, mais qui restreint les analyses de phénomènes plus complexes comme les coordinations à ellipses. De plus, certains problèmes liés à l'unification se posent.

La couverture du phénomène augmente avec le rejet de la structuration en constituants, observation faite pour les grammaires catégorielles et les processus informatiques. Parallèlement, la surgénération et les problèmes d'ambiguïtés parasites se posent. Il semble donc qu'un compromis entre couverture linguistique et surgénération soit nécessaire, ce qui nous a conduit à la définition d'un modèle général.

Modèle général

L'étude linguistique exposée précédemment a conduit à plusieurs conclusions que nous reprenons ici :

- Toutes (ou presque toutes) les catégories syntaxiques sont susceptibles d'être coordonnées entre elles, mais on trouve des coordinations d'éléments de natures différentes. Les coordinations d'éléments de mêmes catégories sont les plus fréquentes.
- Les éléments coordonnés ont la même fonction grammaticale et attendent des arguments de même fonction.
- Un ensemble de contraintes, du type restriction sur les catégories, contraintes

d'accords, contraintes sur la portée d'éléments distribués, contraintes sur les éléments élidés, etc., s'applique sur les éléments coordonnés, et permet de délimiter la portée de la coordination.

- D'une manière générale, un certain degré de parallélisme entre les éléments coordonnés est requis.

Nous considérerons que la coordination est une opération de type « métagrammatical », reliant des éléments de même type, avec distribution des éléments voisins, qu'ils soient arguments ou prédicats. Afin d'engendrer la structure adéquate, il est nécessaire que la grammaire puisse contenir des règles de la forme :

*C -> C (Conj C)** (3)

Les règles de cette forme entrent dans le cadre des grammaires hors-contexte, et sont adoptées dans la plupart des formalismes syntagmatiques actuels. Ce modèle est utilisé dans un grand nombre de formalismes syntaxiques pour le TALN, et reste général.

Nous proposons de surcroît d'utiliser les contraintes isolées dans l'étude linguistique pour restreindre l'application des règles de coordination en fonction du contexte, et de lexicaliser un ensemble d'expressions figées (prépositions coordonnées complexes, terminologie).

IMPLANTATION

Nous décrivons l'implantation réalisée dans le système XLE (*Xerox Linguistic Environment*). Le processus d'analyse d'une phrase par XLE est le suivant :

- la phrase est tout d'abord segmentée en unités lexicales ;
- chaque unité lexicale subit une analyse morphologique ;
- le résultat de l'analyse morphologique est l'entrée du processus d'analyse syntaxique qui combine règles de grammaire et insertion lexicale à partir du lexique LFG ;
- la phase d'analyse syntaxique à proprement parler est réalisée par l'analyseur, et fournit comme résultat les structures LFG (c-structure, f-structure) associées à la phrase.

Analyse présyntaxique

La segmentation de la phrase en unités lexicales est réalisé par un transducteur à états finis, [Ait 97], [Karttunen 94]. Ce transducteur découpe la chaîne d'entrée en une séquence d'unités lexicales qui peuvent correspondre à une forme fléchie, une marque de ponctuation, etc. Les expressions à mots multiples sont des suites de mots regroupés en unités lexicales, par exemple un adverbe/nom comme *a priori* est composé de plusieurs mots mais forme une unité lexicale. La deuxième étape de la présyntaxe est l'analyse morphologique des unités lexicales produites par la segmentation de la phrase. Cette étape est aussi réalisée par un transducteur qui relie la forme fléchie à la forme lexicale (et vice-versa). La forme lexicale est une séquence comprenant la représentation canonique du mot (le lemme), un ensemble d'étiquettes représentant le comportement morphologique du mot, et sa catégorie syntaxique.

Nous lexicalisons un ensemble d'expressions figées contenant des conjonctions de coordination (*corps et âme*, *au fur et à mesure de*), mais aussi certains termes (*filtre à huile*),

. . . . . .

3. Ici, * est l'étoile de Kleene, symbole mathématique qui permet de répéter un nombre quelconque de fois (y compris 0) un élément donné.

> *Page 5*

en leur associant des procédures de segmentation et d'analyse morphologique adéquates, ce qui a des conséquences intéressantes en terme de précision des analyses obtenues et de réduction du nombre de solutions parasites, [Brun 98b].

Analyse syntaxique

La modélisation syntaxique (4) proposée est appliquée au modèle LFG. Ce formalisme utilise l'unification, et s'appuie sur des représentations lexicales des phénomènes linguistiques [Bresnan et al. 82]. La phrase est décrite par deux types de structures. La structure de surface est décrite par une représentation arborescente, appelée structure de constituants (c-structure). Les relations fonctionnelles (relations prédicats/arguments) sont représentées par des structures de traits codant les différentes fonctions grammaticales, ce sont les structures fonctionnelles (f-structure). Les fonctions grammaticales (sujet, objet, complétive…) peuvent être codées et modifiées dans les entrées lexicales des prédicats qui les gouvernent, c'est pour cette raison que la grammaire est dite lexicale fonctionnelle.

La structure fonctionnelle associée à la coordination dans sa totalité n'est cependant pas une structure de trait classique, mais un ensemble dont les éléments sont les f-structures des différents constituants, [Kaplan et Maxwell 88]. Les propriétés associées à un ensemble sont distribuées à ses éléments. Le comportement syntaxique général de la coordination est décrit à l'aide de méta-règles, qui prennent pour arguments les catégories syntaxiques des constituants de la grammaire :

$$COORD(C) = (CONJ)\ C\ [(,)\ CONJ\ C]+\ \downarrow\in\uparrow\ \downarrow\in\uparrow$$

Chacune des catégories susceptibles de se coordonner, qu'elle soit lexicale ou syntagmatique, peut instancier le paramètre C.

La représentation par ensemble ($\downarrow\in\uparrow$) permet de traiter la coordination de façon transparente du point de vue fonctionnel. En effet, le nombre des constituants n'est pas restreint et ce modèle prend en compte à la fois les coordinations binaires et n-aires.

Nous limitons l'application du schéma de coordination en prenant en compte les contraintes linguistiques isolées précédemment, et ce, en utilisant les mécanismes d'équations contraintes de LFG. Cette modélisation permet de rendre compte de la majorité des cas, mais il est cependant nécessaire d'adapter le modèle LFG pour traiter certains problèmes.

Ainsi, la coordination de groupes nominaux provoque des conflits d'unification, car des groupes nominaux coordonnés au singulier s'accordent le plus souvent avec des prédicats pluriels. Pour éviter les conflits et calculer correctement le genre, le nombre et la personne, nous déclarons ces attributs comme non-distributifs (c'est-à-dire non unifiables) sur les ensembles, et redéfinissons des schémas de coordinations spécifiques qui prennent en compte le calcul de ces attributs en fonction du genre du nombre, de la personne des différents constituants mis en jeu, et du type de conjonction.

De plus, les coordinations de constituants de catégories différentes sont généralement un problème pour les formalismes syntagmatiques, LFG n'échappant pas à cette règle.

9. Ce voyant s'allume lorsque l'imprimante est [[en ligne] et [disponible]].

La représentation de ce genre d'exemple ne pose pas de problèmes en ce qui concerne les structures fonctionnelles. En effet, on peut appliquer la méta-règle de coordination pour des éléments de catégories différentes ayant la même fonction et s'attachant à un

. . . . . .

4. Nous ne détaillons pas ici l'implantation de certains cas de coordination elliptiques.

même nœud dans la structure de constituants. Le problème se situe plutôt dans la structure de constituants : quelle est la catégorie du nœud qui domine une la structure coordonnée ? L'utilisation de catégories sous-spécifiées nous semble adéquate. Par exemple, XP peut être définie comme une catégorie générique, qui se réécrit en une disjonction de catégories :

XP → { *SA* | *SP* | *SN* }.

On pourra ainsi coordonner XP avec elle-même, et obtenir des coordinations de type [SA Conj SP], à la condition que SA et SP aient les même fonctions.

On relève aussi certains exemples, où les éléments coordonnés ont également des fonctions différentes au sens *LFG* du terme.

10. Ils attendent [[ma mort] et [que l'affaire soit réglée]].

Dans le lexique LFG, un verbe comme *attendre* possède plusieurs schémas argumentaux ; en particulier, il régit soit une complétive (fonction COMP), soit un complément d'objet direct nominal (fonction OBJ).

Cet exemple ne peut donc pas recevoir d'analyse, car les arguments du verbe ont des fonctions différentes, rattachées à des entrées lexicales différentes. Les données reliées à la coordination argumentent donc en faveur de la réunion des deux fonctions COMP et OBJ sous une même fonction OBJ. La fonction OBJ se rapprocherait alors de la notion traditionnelle de complément d'objet. Cette redéfinition du modèle fonctionnel de LFG permet de conserver l'identité fonctionnelle comme un critère déterminant pour la coordination, ce que nous cherchons à réaliser compte tenu de notre étude linguistique du phénomène.

Heuristiques de classement

Dans le cadre du développement de grammaires syntagmatiques pour le TALN, on observe des problèmes liés à l'excès de solutions parasites. Le point crucial est que, lorsque l'on obtient des analyses sémantiquement étranges ou peu plausibles, il est extrêmement difficile de les rejeter (5). Appliquer des contraintes syntaxiques risque de rejeter des phrases correctes dans un autre contexte.

Les contraintes implantées jusqu'alors dans notre grammaire sont sures, mais le nombre de solutions parasites reste trop important. Pour pallier ce problème, nous utilisons une méthode fondée sur la théorie de l'optimalité, [Frank et al. 98]. Cette méthode permet d'exprimer des préférences sur les analyses obtenues. Une nouvelle structure (la o-structure) contient des marques d'optimalité (ici des attributs rangés par ordre décroissant selon la préférence et qui sont des marques positives ou négatives). Lorsque les différentes analyses d'une phrase sont produites par le système, la (ou les) structure(s) préférée(s) est celle dont la o-structure contient le moins de marques négatives. Si l'on obtient plusieurs candidats à l'issue de ce premier filtrage, ils sont à nouveau filtrés en fonction du plus grand nombre de marques positives. Prenons un exemple simple mettant en jeu les problèmes d'ambiguïté lexicale. *A priori*, un exemple comme :

Le défaut est corrigé.

. . . . . .

5. [Chanod 93] : « La description informatique des mécanismes langagiers n'est pas neutre. Elle produit ses propres interprétations, hors du champ d'étude de la linguistique traditionnelle. Les descriptions informatiques interagissent et produisent de nombreux effets de bords ou analyses indésirables.... [Le parasitisme computationnel] se manifeste par la production d'analyses indésirables, inappropriées résultant de l'application de règles tout à fait fondées linguistiquement par ailleurs. »

obtient deux analyses, selon que *est* est considéré comme un verbe ou comme un nom modifieur de défaut, ce qui est syntaxiquement possible :

12. [P[SN Le défaut] [SV [AUX est] [V corrigé]

13. [SN[Le [défaut [NMOD est]] [SA corrigé]]]

La système de préférence permet de sélectionner la première analyse en affectant aux noms « rares » comme *est* une marque négative (NOM-RARE). L'analyse optimale sera alors la première. Par contre un exemple comme *Il faut s'orienter vers l'est* ne sera pas rejeté car *est* ne peut être qu'un nom dans ce contexte syntaxique, et comme c'est l'unique possibilité syntaxique, l'analyse est correctement retrouvée malgré la marque négative.

Nous avons utilisé cette technique pour le classement des structures coordonnées produites par notre modèle d'analyse, en implantant diverses heuristiques de filtrage. Le premier principe est d'affecter des poids en fonction du parallélisme des constructions :

- Lorsque des propositions complètes coordonnées ont toutes un sujet exprimé, ou au contraire toutes un sujet omis, nous affectons un poids élevé.
- Lorsque toutes les prépositions sont répétées et identiques dans un syntagme prépositionnel coordonné, nous affectons un poids élevé à la construction.
- Lorsque les déterminants de chaque syntagme nominal coordonné sont de même type (défini, indéfini, possessif, etc.), nous affectons un poids élevé à la construction.
- Nous affectons un poids faible aux syntagmes nominaux coordonnés introduits par *à* ou *de*. Cela exprime le fait que les prépositions sont généralement répétées, mais que certains cas rares présentent une distribution de ces prépositions sur le groupe nominal coordonné.
- Lorsque tous les déterminants sont les mêmes ou tous omis dans un syntagme prépositionnel, nous affectons à nouveau un poids élevé.

Illustrons les conséquences de ce traitement sur un exemple :

14. On trouvera ci-dessous une brève explication des instruments et des témoins lumineux.

Nous ne rejetons pas l'analyse syntaxique où le constituant *une brève explication des instruments* se coordonne avec *des témoins lumineux* (*des* étant un déterminant partitif), mais nous lui préférons l'analyse « plus parallèle » où *des instruments* se coordonne avec *des témoins lumineux*, car il s'agit d'une coordination de syntagmes prépositionnels ayant même préposition et même déterminant (des = de + les).

Le second principe est d'affecter des poids faibles aux constructions rares, c'est-à-dire les coordinations de catégories différentes, d'éléments présentant des ellipses, et les énumérations implicites (sans conjonction exprimée).

Analyse de corpus

Nous allons pour conclure décrire les résultats obtenus sur les deux corpus techniques à notre disposition, le manuel d'utilisation des tracteurs présenté précédemment, et un manuel d'utilisation d'imprimantes. Les phrases testées sont les phrases qui contiennent des phénomènes de coordinations.

Manuel d'utilisation des tracteurs

- Nombre total de phrases : 549
- Nombre de phrases testées : 124 (18,9 mots/phrases)
- Nombre de phrases sans analyse valide : 16 = 13 %

> *Page 8*

- Nombre de phrases ayant au moins une analyse valide : 108 = 87 %
- Nombre moyen d'analyses par phrases : 11,75
- Durée moyenne d'une analyse (CPU secs) : 30,45

Manuel d'utilisation des imprimantes

- Nombre total de phrases : 1076
- Nombre de phrases testées : 260 (17 mots/phrases)
- Nombre de phrases sans analyse valide : 37 = 14 %
- Nombre de phrases ayant au moins une analyse valide : 223 = 86 %
- Nombre moyen d'analyses par phrases : 14
- Durée moyenne d'une analyse (CPU secs) : 35,42

Les résultats de l'analyse des deux corpus sont sensiblement les mêmes : environ 86 % des phrases présentant un phénomène de coordination reçoit une analyse valide. Signalons que certaines phrases ne reçoivent pas d'analyse car elles sont trop longues pour que l'analyseur puisse les traiter, celui-ci stoppant le processus lorsqu'un certain délai est dépassé. Le nombre de solutions parasites est encore important pour les deux corpus, puisque nous recevons en moyenne pour les deux corpus, 13 analyses par phrase. Cependant, les heuristiques de classement et les contraintes linguistiques que nous avons développées permettent un réduction de 5 analyses par phrase en moyenne. En effet, si l'on analyse les mêmes corpus sans appliquer ces méthodes, le nombre moyen d'analyse par phrase contenant au moins une coordination est de 18. De plus, sur ces deux corpus, nous n'avons pas observé de rejet d'analyses valides du fait de l'application des heuristiques.

CONCLUSION

Dans cet article, nous avons tout d'abord décrit une étude linguistique sur corpus de la coordination qui nous a permis d'envisager ses différents aspects : types de catégories susceptibles de se coordonner, cas particuliers, contraintes et fréquence dans les corpus. Il résulte que, compte tenu d'un ensemble de contraintes, les éléments linguistiques de même fonction syntaxique peuvent se coordonner. Cette étude, ainsi que l'examen de différentes propositions de traitements dans des formalismes pour le TALN, a pu faire apparaître qu'un traitement satisfaisant doit réaliser un compromis entre couverture linguistique et limitation des solutions parasites. C'est pour cela que le modèle général que nous proposons traite les cas les plus fréquents tout en tenant compte du contexte linguistique. Ensuite, l'implantation du modèle dans un système de développement de grammaires LFG a été décrite. Nous avons alors validé notre approche par des analyses de corpus techniques. En ce qui concerne la couverture syntaxique, nous pouvons analyser les constructions les plus fréquentes, ainsi que certains cas particuliers. Le système d'analyse syntaxique que nous avons utilisé est couplé à une procédure permettant d'exprimer la préférence sur les analyses. Nous avons utilisé ce mécanisme, qui a donné de bons résultats sur les corpus, et qu'il serait intéressant de développer plus avant afin d'améliorer la précision des analyses.

RÉFÉRENCES BIBLIOGRAPHIQUES

[Ait 97] Aït-Mokhtar (S.), « Du texte Ascii au texte lemmatisé : la présyntaxe en une seule étape », *Actes du colloque TALN 97*, Grenoble, 1997.

[Bresnan et al. 82] Bresnan (J.) et Kaplan (R.), *The mental representation of grammatical relations*, Cambridge, MA, The Mit Press, 1982.

[Brun98a] Brun (C.), *Étude et implantation de la coordination en vue de l'analyse automatique du français écrit dans le cadre de la Grammaire Lexicale Fonctionnelle*, Thèse de doctorat en science de l'information et de la communication, effectuée sous la direction du professeur Jacques Rouault, Université Stendhal-Grenoble 3, 1998.

[Brun 98b] Brun (C.), « Terminology finite-state preprocessing for computational LFG », *Actes de Coling/ACL'98*, Montréal, 1998.

[Butt et al. 99] Butt (M.), King (T.H.), Niño (M.E.), Segond (F.), *A Grammar Writer's CookBook*, CSLI Publications/University of Chicago Press, Stanford University, à paraître.

[Chanod 93] Chanod (J.P.), *Problèmes de robustesse en analyse syntaxique*, Actes du colloque Informatique et Langue Naturelle, Nantes, 1993.

[Chomsky 57] Chomsky (N.), *Syntactic Structures*, Mouton, La Haye, 1957.

[Dahl et al. 95] Dahl (V.), Tarau (P.), Moreno (L.), Palomar (M.), « Treating Coordination with Datalog Grammars », *Actes du workshop Computational Logic for Natural Language Processing*, 1995.

[Dik 72] Dik (S.C.), *Coordination*, Noth-Holland, 1972.

[Frank et al. 98] Frank (A.), King (T.), Kuhn (J.), Maxwell(J.), « Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars », *Actes de LFG98*, Brisbane, Australie, 1998.

[Gazdar et al. 85] Gazdar, Klein, Pullum, et Sag, *Generalized Phrase Structure Grammar*, Harward University Press, 1985.

[Grevisse 91] Grévisse, *Le bon usage*, 12ème édition, refondue par André Goosse, Duculot, 1991.

[Kaplan et Maxwell 88] Kaplan (R.) et Maxwell (J.), « Constituent Coordination in Lexical Functional Grammar », *Actes de Coling 88*, Budapest, 1988.

[Karttunen 94] Karttunen (L.), « Constructing Lexical Transducers », *Actes de Coling94*, Stanford, EU, 1994.

[Milward 94] Milward (D.), « Non-constituent coordination : Theory and Practice », *Actes de Coling 94*, Kyoto, 1994.

[Pollard et Sag 94] Pollard (C.) et Sag (I.), *Head Driven Phrase Structure Grammar*, CSLI, University of Chicago Press, 1994.

[Steedman 85] Steedman (M.), « Dependency and Coordination in the Grammar of Dutch and English », *Language*, 61(2) : 523-568, 1985.

[Woods 70] Woods (W.), « Transition Network Grammars for Natural Language Processing », Communication de ACM13, 1970.