Hétérogénéité des corpus: vers un parseur robuste reconfigurable et adaptable

Nuria Gala Pavia

XRCE, 6 chemin de Maupertuis, 38240 Meylan
LIMSI, Bt 508, Université Paris-Sud, 91403 Orsay cedex
Nuria.Gala-Pavia@xrce.xerox.com

Résumé

L'analyse syntaxique robuste est devenue une technique essentielle à toute application qui touche au contenu des documents. Les analyseurs inscrits dans cette approche permettent d'extraire des informations d'ordre linguistique qui peuvent être exploitées postérieurement par des traitements linguistiques plus profonds ou par des systèmes de recherche d'information. Une des caractéristiques principales de ces outils est leur robustesse. Or, cette robustesse est souvent diminuée par la grande hétérogénéité de phénomènes linguistiques et extralinguistiques présents dans les textes tout-venant.

Cet article présente tout d'abord (section 1) la notion de robustesse et caractérise (section 2) les systèmes d'analyse syntaxique robuste. L'article présente par la suite (section 3) un inventaire de phénomènes linguistiques et extralinguistiques non-standard attestés dans divers corpus et, finalement, (section 4) une architecture qui se propose de traiter ces phénomènes.

1. Introduction

Les analyseurs syntaxiques robustes sont conçus pour analyser du texte libre en grandes quantités. Ils se heurtent ainsi à une vaste hétérogénéité de phénomènes dont certains, aussi peu fréquents qu'ils puissent paraître, s'avèrent omniprésents dans certains domaines et formats (Chanod, 2000) : titres et abréviations dans les rapports juridiques, constructions impératives dans les manuels techniques, ellipses dans les textes oraux, etc.. Certains phénomènes sont d'autant plus présents qu'il existe diverses formes d'encodage de l'information : des formes appauvries (textes décapitalisés, sans ponctuation, etc.) ou des formes enrichies (mise en page, typographie, indications de tour de parole, de rires...).

La complexité et la variété des structures présentes dans les textes soulèvent deux questions du point de vue du traitement automatique de la langue. D'une part, quels phénomènes linguistiques sont "traitables" par un analyseur et jusqu'à quel point ; d'autre part, un seul et même parseur peut-il analyser correctement des formats et des domaines différents présentant des structures (linguistiques et/ou extralinguistiques) particulières.

Ces deux problématiques font référence à deux notions généralement utilisées pour évaluer

les systèmes d'analyse automatique : la couverture (*coverage*) et la robustesse (*robustness*). Par "couverture" on comprend l'ensemble de phénomènes linguistiques qui sont pris en compte par la grammaire de l'analyseur. La notion de "robustesse" concerne la capacité de l'analyseur à produire un résultat satisfaisant même devant une situation inattendue, c'est-à-dire, devant un phénomène qui n'a pas été décrit dans la grammaire du système.

Pour certains auteurs, la "robustesse" fait référence conjointement aux deux notions précédentes. Elle implique alors la capacité de produire une analyse pour toute phrase même contenant une structure linguistique ou extralinguistique non décrite dans la grammaire. C'est dans ce sens que nous utiliserons cette notion de robustesse par la suite, tout en faisant la différence (par souci de clarté) entre des aspects plus linguistiques ou plus en relation à la typographie ou à la structure du document.

2. Analyseurs syntaxiques robustes

2.1. Caractéristiques générales

Les outils d'analyse syntaxique robustes sont des systèmes conçus pour pouvoir marquer et/ou extraire des structures syntaxiques prédéfinies et légitimées. Une de leurs caractéristiques principales est l'articulation de l'analyse en deux phases (Ejerhed, 1993) : une première étape de pré-analyse où le but est le repérage de structures minimales de façon à produire une structure préliminaire qui servira de base à l'étape suivante, et une deuxième étape d'analyse syntaxique fondée sur le calcul de structures et/ou relations plus complexes. Ces systèmes sont donc basés sur une approche multi-niveaux (analyse incrémentale des phénomènes linguistiques) et, dans la plupart des cas, ils produisent une seule analyse syntaxique par phrase (déterminisme) [1].

Cette approche s'oriente vers la création de systèmes capables de traiter de grands volumes de textes tout-venant et de produire des analyses utilisables dans d'autres applications : désambiguïsation sémantique (Dini *et al.*, 1999), extraction d'informations (Grefenstette, 1999) etc.

Du point de vue technique, il existe des approches basées sur des grammaires, souvent à états finis (Joshi, 1996) : ici la grammaire est représentée par une cascade de transducteurs créés à partir d'expressions régulières. L'analyse peut être constructiviste (Grefenstette, 1996), et/ou réductionniste (Chanod & Tapanainen, 1996). Une autre type d'approche est statistique (Church, 1988) : les règles de ces systèmes sont acquises automatiquement à partir de corpus annotés.

2.2. Aspects linguistiques

Sur le plan de la représentation de l'information linguistique, quelques analyseurs suivent le modèle formel de la Grammaire Syntagmatique (*Phrase-Structure Grammar*) : ils segmentent la phrase en plusieurs unités (syntagmes ou groupes nominaux, verbaux, etc.). La représentation de l'information syntaxique est en général parenthésée ou arborée et peut être plus ou moins explicite (elle peut contenir de l'information morphologique détaillée ou seulement des étiquettes des parties du discours (*POS tags*)).

Un autre type d'analyseurs se base sur le modèle des Grammaires de Dépendance (*Dependency Grammar*) : ils marquent et extraient des dépendances établies entre les éléments d'une

1. Le déterminisme existe principalement dans la première phase d'analyse, correspondant à la segmentation de la phrase en syntagmes noyau.

phrase. Ces systèmes produisent des analyses qui contiennent de l'information morphologique, des fonctions morpho-syntaxiques et des relations de dépendance qui peuvent intégrer, dans certains cas, de l'information sémantique (Debili, 1982). Un exemple de ce type d'analyseur est le DGP (Tapanainen & Jarvinen, 1997).

Un troisième type de système tient compte des deux modèles linguistiques précédents. Ces analyseurs mixtes produisent comme résultat une structure en constituants et aussi une liste de relations de dépendance. La représentation de cette information peut varier selon les systèmes mais la notion de syntagme noyau ou *chunk* (Abney, 1991) reste commune. Le système IFSP (Aït-Mokhtar & Chanod, 1997) et l'analyseur du GREYC (Giguet & Vergne, 1997) sont représentatifs de cette approche.

3. Phénomènes à caractère "extraordinaire" dans les corpus

L'analyse syntaxique de grandes quantités de texte libre sous format électronique rencontre des phénomènes considérés non-standard : il s'agit de structures non initialement prévues par les grammaires, c'est-à-dire, des séquences d'éléments linguistiques ou extralinguistiques non modélisées.

Nous avons dressé un inventaire de quelques uns de ces phénomènes "extraordinaires", utilisant pour cela des corpus des différents domaines[2].

3.1. Textes écrits

Pour l'écrit, nous avons travaillé[3] sur un corpus total de 126 000 mots, ce qui correspond approximativement à 5 585 phrases, avec une moyenne de 25,6 mots par phrase[4]. Ces résultats sont approximatifs au sens où les limites des phrases ne sont toujours pas faciles d'établir (en raison de l'ambiguïté du point dans les abréviations ou bien au manque de point dans les titres).

Voici quelques phénomènes complexes pour le parseur résultant de la variation typographique : suites de chiffres comme résultats de tennis (*6-4, 6-2*), numéros de téléphone (*04-76-51-23-44*) ; titres de films/chansons (dans une autre langue) ; adresses ; abréviations ; structures combinant chiffres et lettres comme formules (*l'intégrale G1(l,l '), rd=20 mg/cm2*), références bibliographiques (*Vol. 20, Paris, 1992. ISBN 2-9514-0-7*) ; titres et structures sans marque de fin de phrase, etc.

Il est à noter que certains des phénomènes précédents sont propres à un type précis de corpus (abréviations dans les corpus juridiques, symboles et formules dans les scientifico-techniques, etc.) alors que d'autres apparaissent de façon générale dans les différents domaines (mots en langue étrangère, suites de chiffres, titres[5]).

Les chiffres de la table 1 représentent le pourcentage de phrases des corpus contenant ces phénomènes. Par exemple, en moyenne 4,7 % des phrases contiennent des structures chiffre(s)-lettre(s) comme des références bibliographiques, des formules, des symboles, etc. Ceci veut dire

2. Journaux *Le Monde* et *Libération* ; articles financiers et scientifiques ainsi que rapports juridiques de différentes pages web ; manuels techniques de Xerox et corpus oraux provenants du LIMSI.

3. Analyse manuelle afin (1) de repérer les structures complexes qui pourraient poser des problèmes à un parseur et (2) de constituer un corpus de référence qui permettra par la suite d'évaluer la qualité des analyses.

4. En incluant la ponctuation, ou de 22,5 mots par phrase sans en tenir compte. Dans les deux cas nous avons utilisé les corpus après leur analyse morphologique.

5. Les titres et autres structures sans marque de fin de phrase n'ont pas été comptabilisés.

que dans un corpus de 5 000 phrases il y en aura environ 250 présentant cette structure et pour lesquelles l'analyse du parseur pourra engendrer des erreurs.

Corpus/Phénom.	Lgue. diff.	Suite chiff.	Abrévs.	Chiff./Lett.
Journalist.	2,5 %	2,3 %	1,1 %	0,9 %
Juridique	2,5 %	0,3 %	22,7 %	6,6 %
Finances	2,7 %	1,4 %	-	2,9 %
Scient-Techn.	0,4 %	0,5 %	-	8,4 %
Total	2,0 %	1,1 %	5,9 %	4,7 %

Table 1. Présence de phénomènes extralinguistiques complexes dans les corpus écrits, en pourcentages de phrases.

À côté des phénomènes extralinguistiques, les corpus présentent aussi des structures complexes du point de vue linguistique. La table 2 montre ainsi des phénomènes tels que : des propositions interrogatives, exclamatives, à l'impératif, des phénomènes concernant des modifications dans la structure des propositions (vocatives, topicalisation, incises...), des coordinations, des ellipses, des énumérations et des listes (avec ou sans conjonction finale, respectivement), d'autres structures diverses comme les corrélatifs, les coprédicatifs (N + N/Adj), les comparatifs, les distributifs, etc.

Corpus/Phénom	Int/Excl	Imperat.	Coordin.	Enums.	Listes
Journalist.	2,1 %	2,2 %	4,6 %	4,2 %	0,6 %
Juridique	2,2 %	2,5 %	4,9 %	3,1 %	1,4 %
Finances	0,1 %	0,4 %	9,6 %	2,4 %	0,3 %
Scient-Techn	1,5 %	18,4 %	14,5 %	2,7 %	0,1 %
Total	1,5 %	5,9 %	8,4 %	3,1 %	0,8 %

Table 2. Présence de phénomènes linguistiques complexes dans les corpus écrits, en pourcentages de phrases.

Certains phénomènes s'avèrent présents dans tous les domaines (coordinations, énumérations) alors que d'autres sont presque inexistants dans des domaines précis (par exemple les propositions interrogatives, exclamatives et impératives dans les rapports financiers)[6].

Les pourcentages de phrases contenant des structures complexes font preuve de la nécessité de modélisation et de traitement de ces phénomènes dans le cadre de l'analyse syntaxique robuste.

3.2. Textes transcrits de l'oral

Les transcriptions de textes oraux présentent des spécificités telles que nous avons choisi de les présenter séparément. Le corpus utilisé contient des dialogues entre une société et ses clients pour la réservation et l'achat de billets, en tout 61 dialogues. Parmi celles-ci, nous avons comptabilisé un total de 2 161 interventions, ce qui représente une moyenne de 36,2 interventions par dialogue[7].

6. Les phénomènes concernant des modifications d'ordre des éléments dans la structure des propositions ainsi que les coordinations corrélatives et distributives, etc. n'ont pas été comptabilisées.

7. Le nombre total de mots (toute marque extralinguistique exclue) correspond à 16 692, avec 7,7 mots par intervention, et 22 136 mots en incluant toutes les marques.

Du point de vue typographique ou de la structure du corpus, se dégagent des phénomènes comme le marquage des interventions, le marquage de commentaires de transcription ainsi que des formules conventionnelles pour marquer des blancs, des silences courts, des hésitations, etc. Pour les phénomènes complexes du point de vue linguistique nous avons comptabilisé des constructions incomplètes, ellipses, phrases contenant des repétitions, phrases contenant des hésitations, des pauses. etc.

Phén.	Interventions	Commentaires	Silences/Blancs.	Hesit./Pauses	Repetitions
Total	100 %	2,5 %	26,7 %	4,9 %	6,0 %

Table 3. Présence de phénomènes divers dans les corpus oraux, en pourcentages d'interventions.

Les chiffres de la table 3 représentent le pourcentage de phénomènes repertoriés plus haut par rapport aux phrases (interventions). À cause du manque de ponctuation des transcriptions, nous avons considéré chaque intervention d'un locuteur comme une phrase. Si les commentaires de transcription apparaissent dans 2,5 %, les marques de silences correspondent à 26,7 % des phrases. La prise en compte de l'ensemble de ces phénomènes n'est donc pas négligeable.

4. Discussion

Le constat de la présence de phénomènes hétérogènes dans les corpus nous amène à envisager un parseur reconfigurable (modularité de ses composants) et adaptable (apprentissage guidé par le corpus). La figure 1 schématise l'architecture qui s'accorderait à ces caractéristiques.

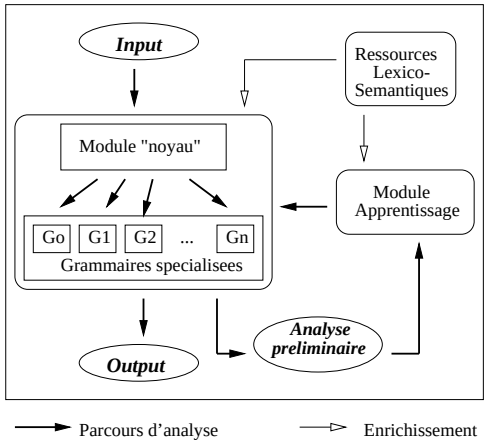

FIG. 1 – *Architecture envisagée.*

D'une part, la notion de modularité implique une meilleure granularité lors de l'analyse syntaxique. Ainsi, dans un premier temps les phrases en entrée sont analysées par un module "noyau" chargé de distinguer phénomènes simples et complexes. Par la suite, des règles ou des grammaires spécialisées pour le traitement d'un phénomène précis s'activent automatiquement. La reconfigurabilité doit être comprise dans le sens où la présence d'un type ou autre de phénomène complexe déclenchera l'application d'un module spécifique de la grammaire (dans la figure *Go, G1, G2,...,Gn*). L'activation d'un module se fera en fonction du corpus.

D'autre part, l'adaptabilité de l'analyseur implique la correction d'une première analyse par apprentissage. Le module de correction par apprentissage acquiert des structures analysées de

façon fiable et modifie, en tenant compte de ces informations, l'application de quelques règles de la grammaire pour obtenir des analyses plus précises (Bourigault, 1992). Ce module, ainsi que l'ensemble de règles de l'analyseur, s'enrichissent avec des ressources lexico-sémantiques dans le but d'aider la résolution d'ambiguïtés (comme l'élagage de relations de dépendance concernant le rattachement prépositionnel).

5. Conclusions

La présence de phénomènes hétérogènes dans des corpus de domaines variés ne peut pas être négligée par des analyseurs robustes. La traitement de ces phénomènes doit mettre en oeuvre une analyse automatique comprenant tout à la fois de la modularité des composants et de l'apprentissage de structures à partir du corpus analysé. Ce type d'architecture devrait pouvoir être à la base d'outils d'analyse performants devant du texte tout-venant.

Références

ABNEY S. (1991). Parsing by chunks. In R. BERWICK, S. ABNEY & C. TENNY, Eds., *Principle-Based Parsing*. Academic Publishers.

AÏT-MOKHTAR S. & CHANOD J.-P. (1997). Incremental Finite-State Parsing. In *Proceedings of ANLP-97*, Washington.

BOURIGAULT D. (1992). Lexter, vers un outil linguistique d'aide à l'acquisition de connaissances. In *Actes des 3èmes Journées d'Acquisition des Connaissances*, Dourdan.

CHANOD J.-P. (2000). Robust Parsing and Beyond. In G. V. NOORD & J. JUNQUA, Eds., *Robustness in Language Technology*. Kluwer (à paraître).

CHANOD J.-P. & TAPANAINEN P. (1996). A Robust Finite-State Parser for French. In *ESSLII'96 Robust Parsing Workshop*, Prague.

CHURCH K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the 2nd Conference on Applied Natural Language Processing*, p. 136–143.

DEBILI F. (1982). *Analyse Syntaxico-Sémantique Fondée sur une Acquisition Automatique de Relations Lexicales-Sémantiques*. PhD thesis, Université Paris XI, France.

DINI L., DI TOMASO V. & SEGOND F. (1999). Ginger ii, an example-driven word sense disambiguator. In *Computers and Humanities, special issue*.

EJERHED E. (1993). Nouveaux courants en analyse syntaxique. *T.A.L.*, **34**(1).

GIGUET E. & VERGNE J. (1997). Syntactic analysis of unrestricted french. In *Proceedings of the International Conference on Recent Advances in NLP, RANLP-97*, Tzigov Chark, Bulgaria.

GREFENSTETTE G. (1996). Light parsing as finite state filtering. In *Workshop on extended finite state models of language, ECAI'96*, Budapest, Hungary.

GREFENSTETTE G. (1999). Shallow parsing techniques applied to medical terminology discovery and normalization. In *Proceedings IMIA WG6, Triennial Conference on Natural Language and Medical Concept Representation*. Phoenix, USA.

JOSHI A. (1996). A parser from antiquity: An early application of finite-state transducers to natural language parsing. In *Proceedings ECAI '96, workshop on extended finite state models of language*, Budapest.

TAPANAINEN P. & JARVINEN T. (1997). A non-projective dependency parser. In *Proceedings of ANLP-97*, Washington.