RECITAL 2004, Fès, 21 avril 2004

Analyse syntaxique et granularité variable

Tristan VanRullen
LPL-CNRS UMR 6057 - Université de Provence
29 avenue Robert Schuman - 13100 Aix-en-Provence
tristan.vanrullen@lpl.univ-aix.fr
date de soutenance prévue : 2004

Mots-clefs – Keywords

Analyse syntaxique, granularité variable, grammaire de propriétés, shallow parsing, deep parsing, densité de satisfaction
Parsing, variable granularity, property grammars, shallow parsing, deep parsing, satisfaction density

Résumé - Abstract

Il est souhaitable qu'une analyse syntaxique -en traitement automatique des langues naturelles- soit réalisée avec plus ou moins de précision en fonction du contexte, c'est-à-dire que sa granularité soit réglable. Afin d'atteindre cet objectif, nous présentons ici des études préliminaires permettant d'appréhender les contextes technique et scientifique qui soulèvent ce problème. Nous établissons un cadre pour les développements à réaliser. Plusieurs types de granularité sont définis. Puis nous décrivons une technique basée sur la densité de satisfaction, développée dans ce cadre avec des algorithmes basés sur un formalisme de satisfaction de contraintes (celui des Grammaires de Propriétés) ayant l'avantage de permettre l'utilisation des mêmes ressources linguistiques avec un degré de précision réglable. Enfin, nous envisageons les développements ultérieurs pour une analyse syntaxique à granularité variable.

It is gainful for a syntactic analysis - in Natural Language Processing- to be carried out with more or less accuracy depending on the context, i.e. its granularity should be adjustable. In order to reach this objective, we present here preliminary studies allowing, first of all, to understand the technical and scientific contexts which raise this problem. We establish a framework within which developments can be carried out. Several kinds of variable granularity are defined. We then describe a technic developed within this framework using satisfaction density, on algorithms based on a constraints satisfaction formalism (Property Grammars) and allowing the use of the same linguistic resources with an adjustable degree of accuracy. Lastly, we further consider developments towards a syntactic analysis with variable granularity.

1 Introduction

Certaines applications en Traitement Automatique des Langues Naturelles s'appuient sur des techniques d'analyse syntaxique superficielle (typiquement celles traitant des données volumineuses), d'autres comptent sur l'analyse profonde (par exemple pour la traduction automatique). Les techniques employées dans ces deux cas sont tout à fait différentes. La première se fonde habituellement sur des méthodes stochastiques comme dans (Liberman & Church, 1992), la seconde sur des techniques symboliques (voir (Chanod, 2000)). Cependant, ceci peut constituer un problème pour des applications qui auraient besoin d'une analyse peu profonde la plupart du temps et dans certaines situations d'une information plus détaillée. C'est typiquement le cas pour les systèmes de synthèse vocale. De telles applications se fondent habituellement sur des analyseurs peu profonds afin de calculer les groupes intonatifs sur la base d'unités syntaxiques, ou plus précisément sur des 'chunks': c'est l'approche suivie par Allen et Abney dans (Allen *et al.*, 1979), (Abney, 1991) et (Abney, 1996). Mais dans certains cas, une information si superficielle n'est pas assez précise. Une solution consisterait alors à employer une analyse profonde pour quelques constructions. Un système mettant en oeuvre ces deux niveaux d'analyse exige habituellement deux traitements différents (deux passages de l'analyseur sur l'entrée), le second refaisant en fait la totalité du travail. De fait, il est difficile d'imaginer dans le cadre génératif classique de mettre en application une technique d'analyse capable de calculer en un seul passage des 'chunks', et dans certains cas, des expressions plus complexes avec une organisation hiérarchique.

Nous présentons ici un formalisme fondé sur les contraintes, qui constitue une réponse possible à ce problème. Cette approche permet l'utilisation d'une même ressource linguistique (c.-à-d. une grammaire unique) qui puisse être employée entièrement ou partiellement par l'analyseur. Cette approche se fonde sur le fait que (1) toute l'information linguistique est représentée au moyen de contraintes et (2) les contraintes sont des type régulier. L'idée consiste alors à mettre en application une technique qui puisse se servir de quelques contraintes dans le cas de l'analyse peu profonde, et de leur ensemble entier pour l'analyse profonde. Dans notre formalisme, les contraintes sont organisées en différents types. Régler la granularité de l'analyse consiste alors dans (1) le choix des types de contrainte devant être satisfaits, (2) le choix du seuil de tolérance à la violation des contraintes.

La première partie de cet article propose une réflexion sur la position de ce problème au sein du Traitement Automatique des Langues Naturelles. Nous déterminons un cadre d'étude pour positionner la notion de *granularité variable* dans l'analyse syntaxique, que nous définissons de plusieurs manières. Puis nous exposons dans une seconde partie le choix d'utiliser le formalisme des Grammaires de Propriétés, ses avantages principaux en termes de représentation et d'exécution pour l'implantation des algorithmes. Nous proposons enfin une technique en cours d'implantation sur un analyseur, dans la perspective d'une analyse syntaxique à granularité variable. Il s'agira là d'une nouvelle approche fondée sur la mesure de *densité de satisfaction* que nous définissons dont nous indiquons l'utilité pour le *réglage de la granularité*.

2 Le problème de la granularité

Nous formulons ici les contraintes qui orientent notre travail, en s'inspirant des recommandations d' Allen (Allen *et al.*, 1979): un système de synthèse vocale demande des choses plutôt

inhabituelles à un analyseur. “ Il doit avoir une large (quoique légère autant que possible) couverture des textes, sans restriction, plutôt qu'une analyse profonde d'un domaine restreint. L'échec de l'analyse est inacceptable dans un système TTS ”

Notre problème est celui-ci : comment développer un outil d'analyse tantôt superficiel, tantôt profond, basé sur les mêmes ressources linguistiques, pouvant s'adapter à de nouvelles ressources et pouvant intégrer plusieurs interprétations pour une même entrée ?

Découpons ce problème en plusieurs parties :

- réglage de la hiérarchie des structures caractérisées (emboîtement profond ou structures plates): il nous faut un outil capable de réaliser une analyse plus ou moins profonde en fonction du contexte,
- réunion des modalités (réunion sélective d'informations de plusieurs domaines -syntaxe, prosodie, gestuelle etc. . . -): cet outil permettra de gérer des interprétations concurrentes pour un même texte (par exemple un contour syntaxique et un contour prosodique),
- tolérance à l'agrammaticalité: lorsqu'un contexte résiste à une interprétation grammaticale stricte, il doit être possible de le réinterpréter localement en étant plus tolérant sur la grammaticalité,
- modulation du déterminisme des analyses: enfin, lorsque plusieurs interprétations coexistent localement, en fonction des nécessités et des finalités (liées au contexte et à la visée applicative), il doit être possible de préférer une interprétation déterministe ou *a contrario* de laisser l'ambiguïté présente.

Ce cadre englobe des notions usuelles dans le domaine de l'analyse syntaxique. Unifier ces notions n'est utile que dans l'optique d'une prise en compte simultanée de leurs intérêts. C'est ce que nous tentons de faire dans la seconde partie de cette étude.

3 Choix des grammaires de propriétés

La notion de contraintes a une importance cruciale en linguistique, voir par exemple (Maruyama, 1990), (Pollard, 1996), (Bouma *et al.*, 2001). Les théories récentes (du paradigme de l'analyse par contraintes à celui des Principes et Paramètres) se fondent sur cette notion. Un des intérêts principaux de l'emploi des contraintes réside dans la possibilité de représenter n'importe quel genre d'information (très générale aussi bien que la locale ou contextuelle) à l'aide d'un dispositif unique. Nous adoptons pour la technique présentée plus bas le formalisme des Grammaires de Propriétés, décrit en (Bès & Blache, 1999) ou (Blache, 2001), qui permet de concevoir et de représenter toute l'information linguistique en termes de contraintes sur des objets linguistiques. Dans cette approche, les contraintes sont vues comme des relations entre deux (ou plus) objets. En adaptant une proposition de Bès (1999), l'ensemble des contraintes suivantes participe à la description d'un syntagme: linéarité, dépendance, obligation, exclusion, exigence et unicité.

La figure suivante esquisse leurs rôles respectifs, illustrés avec quelques exemples pour le SN.

À la différence d'autres approches comme la théorie d'optimalité, présentée dans (Prince & Smolensky, 1993), il n'existe aucune hiérarchie entre elles et on peut choisir, selon les besoins,

Contrainte	**Definition**	**Exemple pour le SN**
Linéarité (<)	Précédence linéaire	*Det < N; Det < AP;* *AP < N; N < PP*
Dépendance (→)	Dépendance entre catégories	*Det → N; AP → N; PP → N*
Obligation (*Oblig*)	Ensemble de catégories obligatoires et uniques. Une seule de ces catégories doit être réalisée dans un syntagme.	*Oblig(NP) = {N, Pro, AP}*
Exclusion (≠)	Restriction de cooccurrence entre des ensembles de catégories	*N ≠ Pro; N[prop] ≠ Det*
Exigence (⇒)	cooccurrence obligatoire entre des ensembles de catégories	*N[com] ⇒ Det*
Unicité (*Uniq*)	Catégories uniques dans un syntagme	*Uniq(NP) = {N, Det}*

Table 1: Définition des propriétés

de vérifier l'ensemble entier de contraintes ou seulement un sous-ensemble. Ce qui est intéressant est que certaines contraintes comme la linéarité fournissent des indications en termes de frontières, comme décrit par exemple dans (Blache, 2001). Il s'ensuit que la vérification de ce sous-ensemble de contraintes peut constituer une technique de parenthésage. La vérification de plus de contraintes en addition à la linéarité permet de raffiner l'analyse. Enfin, la même technique d'analyse (satisfaction de contraintes) peut être employée pour des analyses peu profondes et des analyses profondes. Pour être plus précis, en utilisant les mêmes ressources linguistiques (lexique et grammaire), nous proposons une technique qui permette de choisir la granularité de l'analyse. Une première approche de la sélection de granularité a fait l'objet des travaux détaillés dans (Balfourier *et al.*, 2002), (Blache & Van Rullen, 2002). Il s'agissait d'une technique de multiplexage permettant la fusion des résultats de plusieurs analyses obtenues grâce aux grammaires de propriétés. L'inconvénient de cette méthode résidait dans l'obligation de réaliser plusieurs passages sur la même entrée. La technique que nous introduisons ci-dessous permet de s'en affranchir.

4 Mesure de densité de satisfaction

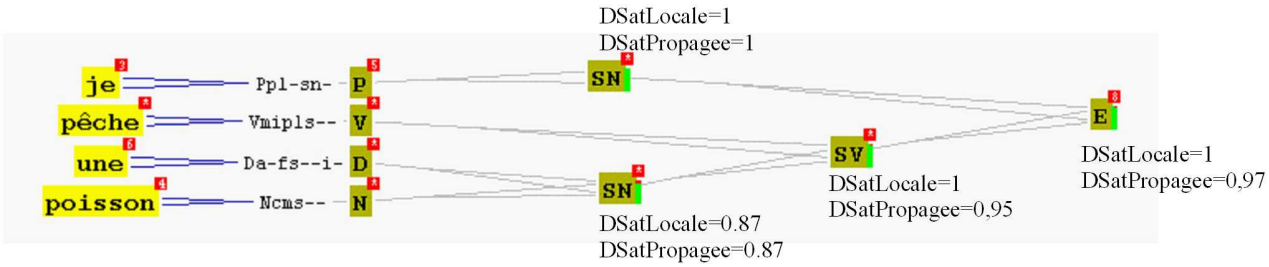
L'analyseur basé sur les contraintes dans le formalisme des Grammaires de Propriétés, dont nous parlons à présent, est en cours de développement. Nous ne pouvons donc pas encore en donner une évaluation. Toutefois, les idées qui y sont introduites se positionnent parfaitement dans le cadre que nous mettons en place. Au prix d'une formalisation complète de la sémantique des grammaires de propriétés (voir à ce propos (Van Rullen *et al.*, 2003)), il a été possible d'écrire un analyseur capable d'une part d'analyser un texte au crible de plusieurs grammaires simultanément et d'autre part d'indiquer le degré de grammaticalité des structures construites. Cet outil se place autant dans le cadre élaboré ici que dans la perspective cognitive décrite dans (Blache, 2003). Nous ne décrivons ici qu'un des principes en oeuvre dans cet analyseur, la densité de satisfaction, qui permet de régler la granularité des constructions en contexte: Chaque catégorie caractérisée au cours d'une analyse porte deux valeurs qui mesurent sa *densité de satisfaction locale* et sa *densité de satisfaction propagée*.

- au cours de l'analyse, pour chaque catégorie caractérisée, la quantité de propriétés satisfaites et la quantité de propriétés enfreintes est comptabilisée, ce qui définit une mesure de la **densité de satisfaction locale** d'une catégorie.

- lorsqu'une catégorie peut à son tour caractériser un groupe de catégories, elle hérite de la

moyenne des densités de ses constituants immédiats, ce qui définit la notion de **densité de satisfaction propagée**.

L'exemple donné ci-dessous illustre ce procédé. Dans cet exemple, une propriété de dépen-



dance en genre est enfreinte dans le second *SN*. Cette violation n'interdit pas de considérer un SN satisfait à 87%. L'information est propagée au *SV* et à la phrase *E*, qui du point de vue de ses constituants immédiats est satisfaite à 100%, mais seulement à 97% du point de vue de l'ensemble de ses sous-constituants. En autorisant ou en interdisant des constructions au dessous d'un certain seuil de satisfaction, cet analyseur recouvre donc les quatre possibilités de réglage de granularité présentées dans notre réflexion initiale:

- réglage de la hiérarchie des structures caractérisées, en jouant sur le seuil de propagation de la densité de satisfaction.
- réunion des modalités [1],
- tolérance à l'agrammaticalité en étant permissif sur les seuils de densité propagée et locale,
- modulation du déterminisme des analyses [2].

Des perspectives en relation avec l'importance relative des seuils à attribuer aux différents types de propriétés sont offertes par les domaines psycho-linguistiques et les sciences cognitives, qui mettent en avant l'importance de certains types de contraintes par rapport à d'autres dans le processus de compréhension des textes. On se référera par exemple à (Winograd, 1983).

5 Perspectives et conclusions

Le formalisme basé sur la résolution de contraintes proposé ici rend possible le réglage de granularité, depuis une simple détection des frontières jusqu'à une analyse profonde et non déterministe, via une analyse superficielle et déterministe. Nous définissons ainsi la notion de *technique à granularité variable*.

[1] cette idée, qui ne fait pas intervenir la densité de satisfaction, est débattue dans un article à paraître (Guénot & Bellengier, 2004)

[2] une heuristique générale portant sur ce point est incluse dans l'analyseur, mais ne peut pour des raisons de place, être présentée ici

La technique basée sur la densité de satisfaction est prometteuse: on peut imaginer à court terme une automatisation du réglage de granularité en fonction des besoins guidés par le contexte textuel.

La possibilité de régler le niveau de granularité en fonction des données à traiter ou de l'application concernée par ce traitement se présente comme un réel progrès, mais il reste encore à déterminer dans quel contexte et de quelle manière le contexte peut dicter, diriger, cette sélection de granularité. Des éléments de réponse sont déjà sensibles dans les travaux réalisés, mais une formalisation systématique de ce problème reste à développer.

Références

ABNEY S. (1991). Parsing by chunks. In R. BERWICK, S. ABNEY & C. TENNY, Eds., *Principle-based parsing*, p. 257–278. Dordrecht: Kluwer Academic Publishers.

ABNEY S. (1996). Partial parsing via finite-state calculus. In *ESSLLI'96 Robust Parsing Workshop*.

ALLEN J., HUNNINCUTT S., CARLSON R. & GRANSTRÖM B. (1979). Mitalk-79: The 1979 mit text-to-speech system. In WOLF & KLATT, Eds., *Speech Communications Papers Presented at the 97th Meeting of the ASA*, p. 507–510.

BALFOURIER J.-M., BLACHE P. & VAN RULLEN T. (2002). From shallow to deep parsing using constraint satisfaction. In *COLING'2002*, p. 36–42.

BÈS G. & BLACHE P. (1999). Propriétés et analyse d'un langage. In *TALN'99*.

BLACHE P. (2001). *Les Grammaires de Propriétés : des contraintes pour le traitement automatique des langues naturelles*. Paris: Hermès Sciences Publications.

BLACHE P. (2003). Vers une théorie cognitive de la langue basée sur les contraintes. In *in actes de TALN-2003*.

BLACHE P. & VAN RULLEN T. (2002). An evaluation of different symbolic shallow parsing techniques. In *in proceedings of LREC-02*.

BOUMA G., MALOUF R. & SAG I. (2001). *Natural Language and Linguistic Theory*, volume 19:1, chapter Satisfying Constraints on Extraction and Adjunction. Kluwer.

CHANOD J.-P. (2000). *Robustness in Language Technology*, chapter Robust Parsing and Beyond. Kluwer.

GUÉNOT M. & BELLENGIER E. (2004). Quelques principes pour une grammaire multimodale non-modulaire du français. In *soumis à RECITAL 2004*.

LIBERMAN M. & CHURCH K. (1992). *Advances in Speech Signal Processing*, chapter Text analysis and word pronunciation in text-to- speech synthesis, p. 791–831. Furui, s., sondhi, m.m. (eds) edition.

MARUYAMA H. (1990). Structural disambiguation with constraint propagation. In *ACL'90*.

POLLARD K. (1996). The nature of constraint-based grammar. In *PACLIC Conference, reprinted in Constructions: an HPSG Perspective, ESSLLI'98 Lecture Notes*.

PRINCE A. & SMOLENSKY P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Rapport interne, Rutgers University Centre for Cognitive Science.

VAN RULLEN T., GUÉNOT M. & BELLENGIER E. (2003). Formal representation of property grammars,. In *proceedings of ESSLLI 2003 Student Session*.

WINOGRAD T. (1983). *Language as a cognitive processs — Vol. 1 : Syntax*. Reading (Mass.): Addison Wesley.