TALN 2005, Dourdan, 6–10 juin 2005

Combiner analyse superficielle et profonde : bilan et perspectives

Philippe Blache
Laboratoire Parole et Langage
CNRS & Université de Provence
pb@lpl.univ-aix.fr

Mots-clefs : Analyse syntaxique, analyse superficielle, analyse profonde

Keywords: Parsing, shallow and deep parsing

Résumé L'analyse syntaxique reste un problème complexe au point que nombre d'applications n'ont recours qu'à des analyseurs superficiels. Nous faisons dans cet article le point sur les notions d'analyse superficielles et profondes en proposant une première caractérisation de la notion de complexité opérationnelle pour l'analyse syntaxique automatique permettant de distinguer objets et relations plus ou moins difficiles à identifier. Sur cette base, nous proposons un bilan des différentes techniques permettant de caractériser et combiner analyse superficielle et profonde.

Abstract Deep parsing remains a problem for NLP so that many applications has to use shallow parsers. We propose in this paper a presentation of the different characteristics of shallow and deep parsing techniques relying on the notion of operational complexity. We present different approaches combining these techniques and propose a new approach making it possible to use the output of a shallow parser as the input of a deep one.

1 Introduction

Le problème de l'analyse syntaxique reste une question complexe à la fois du point de vue théorique et computationnel. La solution généralement adoptée pour traiter des masses de données volumineuses ou des entrées non standard consiste à recourir à des analyseurs superficiels, robustes et efficaces, mais ne construisant que des informations partielles. Il existe un certain nombre d'études proposant de combiner les techniques d'analyse superficielle profonde permettant soit d'améliorer l'efficacité des analyseurs profonds en leur offrant un meilleur contrôle des processus, soit de proposer une approche permettant de choisir le type d'analyse désiré en fonction des besoins. Cet article dresse un bilan de ces différentes techniques en caractérisant les notions d'analyse profonde et superficielle. Ces caractéristiques sont données non seulement d'un point de vue opérationnel, mais également en introduisant la notion de complexité des phénomènes syntaxiques à analyser. Il s'agit d'une première tentative de classification distinguant les phénomènes faciles à analyser de ceux plus complexes.

On distingue généralement analyse de surface et analyse profonde en fonction de la précision de l'information linguistique construite par un analyseur. Les techniques utilisées sont habituellement différentes : on retrouve plutôt les techniques probabilistes du côté des analyseurs superficiels tandis que les analyseurs profonds utilisent plutôt des approches symboliques. Cette caractérisation doit être complétée par la prise en compte de la finalité de l'application utilisant l'analyseur. Il convient pour cela d'identifier précisément les besoins en termes morpho-syntaxiques ou sémantiques pour identifier le niveau d'analyse requis (chunks pour les systèmes de synthèse de la parole, repérage d'objets nominaux pour les applications de recherche d'information, etc.). Cependant, certaines applications nécessitent, même ponctuellement, des informations plus détaillées concernant les relations syntaxiques ou les effets de sens pour une construction donnée. Nous avons ainsi d'une part une distinction en termes d'efficacité (les analyseurs superficiels sont plus rapides et plus robustes que les analyseurs profonds) et de l'autre une distinction de finalité.

La question du *déterminisme* est à prendre en compte de façon distincte. Si les analyseurs superficiels sont déterministes, les analyseurs profonds traitent quant à eux l'ambiguïté : toutes les possibilités sont prises en compte pendant l'analyse et le système fournit plusieurs solutions lorsque l'ambiguïté ne peut être levée. Une façon de réduire la complexité d'un analyseur profond sans le ramener à un analyseur superficiel consiste à le rendre déterministe. A un premier niveau, l'entrée elle-même peut être déterminisée par l'utilisation d'un étiqueteur désambiguïsant. La déterminisation de l'analyse consiste alors à éliminer des constructions en cours. Les propriétés de coupure utilisées peuvent être de type très différents : probabilistes (par exemple en utilisant des informations syntaxiques associées à des poids), topologiques (propriétés formelles des structures construites, par exemple profondeur des arbres, taille des constituants, etc.), ou encore cognitives (préférences de catégorisation, de rattachement, etc.). Ces techniques permettent de prendre des décisions de façon incrémentale en cours d'analyse. Elles peuvent être associées à des techniques de retardement consistant à repousser certains choix et maintenir plusieurs solutions en parallèle, par exemple en les factorisant. On peut donc à ce stade donner quelques critères distinctifs entre les deux approches :

- analyseur superficiel : rapide et robuste, il fournit une structuration simple en termes d'unités non récursives ainsi que des relations portant sur ces unités
- analyseur profond : fournit une description couvrante des constructions de la langue en indiquant les relations syntaxiques ou syntactico-sémantiques entre ses constituants.

Il existe plusieurs approches permettant de combiner ces approches, la section suivante en propose une présentation. Nous reviendrons ensuite sur une caractérisation de la complexité des phénomènes à analyser avant de décrire, dans la dernière partie, une technique hybride permettant à un analyseur profond de tirer parti d'une analyse superficielle.

2 Approches combinant analyse superficielle et analyse profonde

Il existe un certain nombre de travaux proposant d'utiliser simultanément les techniques d'analyse superficielle et profonde. Un workshop a été récemment consacré à l'étude de ce problème (cf. [Hinrichs04]) et a permis de faire un tour d'horizon de la situation. Dans la plupart des cas, la technique consiste à utiliser l'analyse superficielle en tant que *pré-traitement* d'une analyse

profonde. Par analyse superficielle, on entend surtout ici formatage de l'entrée visant la désambiguïsation de l'*étiquetage* morpho-syntaxique, le traitement des mots inconnus et pouvant aller jusqu'à l'analyse d'unités entières comme les entités nommées par exemple à l'aide de *grammaires locales*. Ce type d'approche peut s'avérer très efficace et offre l'avantage de réutiliser voire d'adapter des composants différents : on trouve par exemple dans [Grover01] une description de la réutilisation d'outils originellement prévus pour l'analyse en GPSG. Dans ce type d'approche, le contrôle de l'analyse profonde se fait donc en limitant l'espace de recherche de l'analyseur grâce à une réduction du nombre d'étiquettes à prendre en compte. De plus, des parties entières peuvent être pré-analysées, ce qui réduit d'autant le nombre de structures à construire. L'intérêt majeur de ce type d'approche réside dans le fait l'analyseur n'a pas à être modifié : il est donc possible de traiter avec le même système une entrée brute ou pré-traitée.

Un second type d'approche, relativement peu répandu, consiste à utiliser les résultats d'un analyseur syntaxique superficiel. L'entrée de l'analyseur profond est la sortie de l'analyseur superficiel, ce qui nécessite l'adaptation de l'analyseur profond. Une première technique consiste à modifier (on emploie également le terme de *lifter*) les informations construites par l'analyseur superficiel. Cela concerne les unités lexicales comme les groupes syntaxiques. Dans le premier cas, il s'agit de transformer une unité simple en une structure enrichie adaptée au format de l'analyseur profond, par exemple par des patterns de filtrage (cf. [Blache95]). Les chunks ou les unités syntaxiques construites peuvent également être transformés à l'aide de règles adaptées utilisant là encore des patterns. Une telle approche est décrite dans [Marimon02] qui transforme ainsi les unités lexicales et les chunks en structures attribut-valeurs du type HPSG comme décrit dans l'exemple suivant :

$$\texttt{rule(}\left[\begin{array}{l}\text{SYNSEM} \mid \text{LOCAL}\left[\begin{array}{l}\text{MORPH}\left[\begin{array}{l}\text{LEMME } \boxed{2}\\ \text{MORPHEME } \boxed{1}\\ \text{AGR } \textit{fem, sing}\end{array}\right]\\ \text{CAT} \mid \text{HEAD}\left[\text{NCLASS } \textit{common}\right]\end{array}\right]\end{array}\right]\texttt{, [Pos='Ncfs-', Lemma=}\boxed{2}\texttt{],}\boxed{1}\texttt{).}$$

On trouve dans la même perspective une technique consistant à enrichir directement la structure construite à l'aide de techniques spécifiques. C'est le cas de [Johnson02]) qui décrit comment, à partir d'arbres syntaxiques simples, créer des arbres complexes à nœud vide. Il s'agit dans ce cas d'une opération d'adjonction qui s'appuie sur des schémas d'arbre spécifiant les endroits où ces nœuds peuvent être insérés et la valeur des arguments qu'ils doivent prendre. Le contrôle du processus se fait grâce à une hiérarchisation de ces schémas. Un troisième type d'approche, défendu dans [Uszkoreit02], propose l'utilisation en parallèle d'une analyse superficielle et d'une analyse profonde. Cette approche (cf. [Crysmann02] ou [Frank03]) consiste à exploiter les informations de l'analyseur superficiel pour contrôler l'analyseur profond et réduire son espace de recherche. Les informations de contrôle fournies par l'analyseur superficiel portent dans cet exemple sur la structure topologique de la phrase en allemand. L'idée est de repérer les champs topologiques par différentes techniques (cf. [Neumann00]) et guider ainsi la construction de la structure par l'analyseur profond. Le dernier type d'approche repose sur la possibilité de régler la finesse de l'analyse en fonction des objectifs. On peut distinguer deux cas selon que les ressources utilisées sont identiques ou pas. Un premier type d'approche consiste simplement à faire varier la grammaire en entrée. L'utilisation d'une grammaire simple, peu ambiguë et n'utilisant que des constituants de forte granularité permettra d'obtenir une analyse grossière d'un énoncé. Dans ce cas, nous pouvons parler de superficialisation d'un analyseur profond par l'utilisation d'une grammaire superficielle (cf. [Puver04]). Mais il est également

Type	*Caractéristiques*	*Exemple*
Pré-traitement	Etiqueteur désambiguïsant, grammaires locales	[Grover01]
Pré-analyse	Analyse superficielle = input de l'analyseur profond	[Marimon02],[Johnson02]
Contrôle	L'analyseur profond est guidé par l'analyseur superficiel	[Crysmann02], [Frank03]
Granularité variable	Même analyseur, le type de sortie est une option	[Blache02]

Figure 1: Différentes techniques de combinaison d'analyseurs

possible de proposer des techniques permettant d'exploiter des ressources identiques en termes de lexique et de grammaire. Ce type d'approche nécessite la possibilité pour l'analyseur de construire des structures partielles, limitées à un certain type de constituant (par exemple les *SN* dans le cas de systèmes de recherche d'information). De même, ce type de système doit pouvoir construire une segmentation de l'input (par exemple sous la forme de chunks). Mais le même analyseur doit pouvoir à l'autre bout de la chaîne construire également une structure détaillée. Un exemple de ce type d'approche est décrit dans [Blache02]. Il s'appuie sur une représentation décentralisée de l'information sous la forme de contraintes. Le réglage de la granularité d'analyse s'opère en faisant varier la tolérance de l'analyseur par un seuil de contraintes qu'il est possible de relâcher. Le choix de la structure construite en sortie se fait quant à lui en spécifiant le type de contraintes à satisfaire.

3 Les difficultés syntaxiques

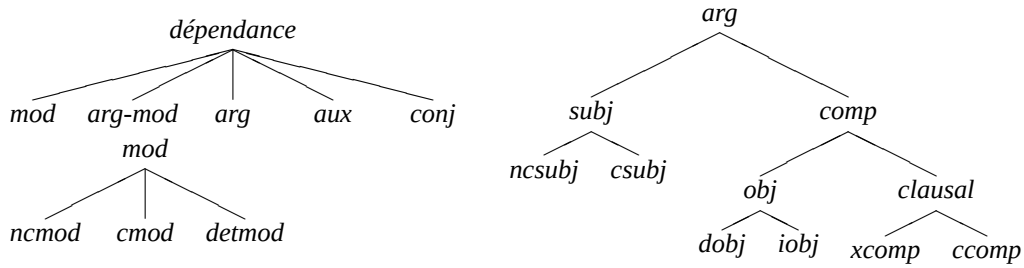
Une analyse précise des difficultés rencontrées par les analyseurs syntaxiques, en dehors des problèmes purement computationnels, reste à établir. Il serait en effet très utile de distinguer les phénomènes faciles à analyser de ceux qui ne le sont pas et d'en expliquer les raisons. Il s'agit d'un exercice difficile, ce problème ne recoupant pas toujours la notion de complexité linguistique : certaines constructions peuvent être facilement interprétables, mais présenter des difficultés en termes d'implantation. C'est le cas par exemple des phénomènes d'extraction qui ne présentent que peu d'ambiguïté d'interprétation mais pour lesquels les systèmes ont des difficultés d'analyse. Réciproquement, les enchâssements de syntagmes peuvent être complexes mais ne présentent pas en soi de difficultés pour un analyseur. Il est intéressant de proposer une ensemble de constructions ou phénomènes qu'un analyseur doit pouvoir traiter. L'article de synthèse [Abeillé00] en fournit une première liste établie de façon tout à fait empirique sur la base d'une analyse des capacités des analyseurs existants au moment de la rédaction de l'article (voici 5 ans, ce domaine a bien entendu beaucoup évolué depuis) :

- dépendances locales: accord, sous-catégorisation des prédicats, expressions semi-figées, restrictions modifieur-modifié, clitiques, etc.
- dépendances moyennes : pronominalisation, contrôle des infinitives, association négative, quantifieurs flottants, etc.
- dépendances à distance : questions, relatives, constructions disloquées, etc.
- alternances syntaxiques : passif, impersonnel, causatives, etc.
- phénomènes de coordination et de comparaison

Cette liste comporte des phénomènes variés et dont la complexité de traitement dépend de la finesse de l'analyse qu'on veut en donner, ainsi que du type de représentation de l'information choisi. Les phénomènes d'accord par exemple sont généralement faciles à traiter pour le français ou l'anglais à condition de disposer dans la grammaire ou le lexique d'un codage

explicite de l'information. Pour ce qui concerne les dépendances à distance, on observe des situations très différentes. Les *relatives* font partie des constructions souvent faciles à traiter, y compris du point de vue de la structure sémantique. Les constructions *disloquées* posent en revanche plus de problèmes. Elles sont assez faciles à repérer, mais la relation sémantique entre l'élément disloqué et le reste de l'énoncé est assez complexe à traiter, même en présence d'un pronom résomptif. Il convient dans ce cas tout d'abord d'identifier ce pronom, celui-ci pouvant apparaître dans des positions très variées, de vérifier les compatibilités morpho-syntaxiques entre l'antécédent et le pronom, mais aussi les compatibilités sémantiques entre l'antécédent et la structure régissant le pronom. Les constructions *clivées* présentent le même type de problème : elles sont en français faciles à identifier, mais le repérage de leur site d'attachement est généralement complexe. Il faut souligner que cette complexité de traitement ne se traduit pas par une difficulté d'interprétation par un humain : les clivées sont au contraire dans la plupart des cas très facile à interpréter (ce type de problème est signalé dans [Puver04]). Pour une même construction, certaines informations sont donc plus complexes à obtenir que d'autres. Si l'on prend en compte le critère de facilité d'analyse pour caractériser un analyseur superficiel, on peut alors dire que l'identification de la présence d'une dépendance à distance peut être obtenue facilement notamment grâce à des marques morphologiques fortes.

Par ailleurs, il faut distinguer d'un côté la structure elle-même (la hiérarchie des objets) et de l'autre les relations existant entre ces objets. Dans le cas d'une approche syntagmatique par exemple, un analyseur devra produire un arbre, mais également indiquer les relations syntaxiques ou sémantiques existant entre les constituants. Un certain nombre de propositions ont été faites pour cela dans le cadre de l'évaluation des analyseurs syntaxiques (cf. [Carroll01], [Briscoe02] ou [Carroll03]). Ce paradigme propose de recenser un certain nombre de relations servant de base à la comparaison et l'évaluation des analyseurs syntaxique. L'ensemble des relations (adapté aux besoins du français par rapport à la proposition de [Carroll01]) est décrit dans la figure 3 et s'organise selon la hiérarchie suivante :



On propose dans le tableau suivant une répartition entre relations faciles et difficiles à identifier. Ce jugement est ici établi sur une base empirique. On essaie de donner quelques arguments justifiant ce classement, mais il conviendrait d'en faire une description plus systématique.

Nom	Description
dépendance	Relation de dépendance générique entre une tête et un dépendant
mod	Relation entre une tête et son modifieur. Le type est le mot introduisant la dépendance
ncmod	Modificateur lexical (non propositionnel)
cmod	Modificateurs propositionnels
detmod	Relation déterminants / noms
arg-mod	Relation tête/argument, celui-ci étant réalisé comme un modifieur (par exemple un SP complément du verbe)
arg	Relation générique tête/argument (plutôt de type complément)
subj	Relation prédicat/sujet
ncsubj	Sujet lexical (non propositionnel)
csubj	Sujets propositionnels (par exemple infinitive sujet)
comp	Relation tête/complément
obj	Relation tête/objet
dobj	Relation prédicat/objet direct (premier complément non propositionnel)
iobj	Relation prédicat/complément non propositionnel introduit par une préposition
clausal	Relation tête/complément propositionnel
xcomp	la proposition complément n'a pas de sujet réalisé
ccomp	la proposition complément a un sujet réalisé

Figure 2: Description des relations

Faciles		**Difficiles**	
Relation	*Caractéristique*	*Relation*	*Caractéristique*
Ncmod	les relations adj/n, sp/n, sp/v, sa/n sont juxtaposées	*n/sp*	séparé par d'autres éléments
Cmod rel	marque morphologique et généralement adjacence	*Cmod*	ambiguïté de rattachement (p. ex. sp/V vs. sp/n)
Detmod	linéarité, adjacence	*Arg-mod*	doit tenir compte de la forme verbale et de la sémantique du mod
Ncsubj	ordre, marque morpho (accord)	*Csubj*	repérage de la proposition, complexité potentielle, identification du verbe tête de la prop sujet, de la tête de la phrase.
Xcomp	marques morphologiques et ordre (eg attribut, infinitif antéposé, etc.)	*Ccomp*	repérage de la subordonnée, identification des têtes verbales
Dobj	forme du complément (nonclausal), ordre (premier) et adjacence avec le verbe Iobj mais ambiguïté rattachement	*Conj*	difficulté de distinguer les conjonctions simples (coordonnés de même type) des autres
Aux	marque morphologique, adjacence	*Contrôle*	Cette relation n'est pas exprimée directement mais par doublement de la relation subj. Dépend du type du verbe et du type du complément

Cette observation rapide permet de dégager quelques éléments de caractérisation de la complexité opérationnelle (et non pas théorique) de l'analyse syntaxique. D'une façon générale en effet, les relations les plus faciles à analyser sont celles profitant d'une conjonction de plusieurs sources d'information stables : faible taux d'ambiguïté des constituants entrant en jeu dans la relation (par exemple seule la relation *Detmod* peut relier un déterminant et un nom), marque morphologique régulière (pronom relatif, conjonctions, construction "*c'est ... que*", etc.), ordre linéaire strict, etc. De plus, le niveau de la relation dans la hiérarchie influe également sur sa complexité : une relation générique sera plus facile à repérer qu'un de ses sous-type (par exemple la relation *Comp* est plus facile à indiquer que *Ccomp*).

A l'inverse, les relations complexes sont celles nécessitant d'accéder à des informations locales spécifiques (par exemple des traits lexicaux), dépendant de la forme des constituants non lexicaux reliés (par exemple type du verbe ou de la préposition dans le *SV* ou le *SP*) ou encore reposant sur des phénomènes sémantiques de restriction. Le niveau sémantique présente

d'ailleurs des caractéristiques similaires : il est par exemple plus facile de traiter le rôle d'un modifieur que la portée de la quantification.

Il n'est donc pas possible de distinguer simplement, comme nous l'avons vu en première partie, un analyseur superficiel d'un analyseur profond sur de simples critères d'efficacité. Mais il ne semble pas non plus pertinent de distinguer les deux approches sur la base du type d'information construit en sortie. Le parenthésage d'un énoncé est une tâche globalement facile si on se contente de constituants non récursifs. Elle devient nettement plus difficile si l'on cherche à décrire le niveau propositionnel ou les dispositifs complexes. De même, comme nous venons de le voir, certaines relations syntaxiques peuvent être plus faciles à identifier que d'autres. Il est donc intéressant d'introduire deux nouveaux critères pour la distinction entre types d'analyse : *niveau opérationnel* (un analyseur profond est non déterministe) et *niveau formel* (un analyseur superficiel ne construit que des informations simples). Les critères de déterminisme et de type d'information peuvent bien entendu être combinés. On pourra par exemple trouver des analyseurs déterministes pouvant construire des informations complexes. Il est possible dans ce cas de parler d'analyseurs intermédiaires.

4 Une stratégie d'analyse hybride

Ainsi que nous venons de le voir, l'analyse profonde a fréquemment recours à des techniques d'analyse superficielle, notamment grâce à une désambiguïsation de l'entrée. Par ailleurs, les résultats obtenus pour la construction d'une analyse superficielle par un analyseur superficiel et un analyseur profond ne sont pas très différents, quelque soit la forme de l'input. La comparaison des résultats obtenus par deux analyseurs sur un même ensemble de corpus dans le cadre de la campagne *Easy* (ces résultats seront présentés lors du workshop *Easy* associé à TALN) montre en effet une forte convergence, aussi bien pour le traitement de corpus de langue écrite que de langue parlée.

Nous avons donc un certain nombre d'arguments qui militent en faveur de systèmes mixtes permettant de fournir comme résultat, en fonction des besoins,aussi bien une analyse superficielle qu'approfondie. Plus précisément, nous proposons une architecture à deux niveaux permettant de réutiliser une analyse superficielle comme entrée d'un analyseur profond. Il ne s'agit pas de modifier la structure superficielle construite (à la différence de l'approche proposée par [Johnson02]), mais bien de construire une représentation plus riche utilisant les objets construits par la superficielle pour construire des objets plus complexes. Il est pour cela nécessaire de définir les objets "superficiels" comme pouvant être des constituants pour l'analyse détaillée. Un parenthésage classique sous forme de chunks ne serait pas pertinente dans cette approche, un chunk ne pouvant être une unité constitutive d'un groupe syntaxique de niveau supérieur.

L'objectif d'une telle approche est tout d'abord de combiner des outils différents, en ne déclenchant éventuellement une analyse détaillée qu'en fonction des besoins. Mais elle permet également d'envisager l'analyse superficielle comme outil de contrôle de l'analyse détaillée. Dans ce cas, toutes les informations construites par l'analyseur superficiel sont susceptibles d'être utilisées par l'analyseur profond. Ces informations sont de deux types : il s'agit d'une part de groupes de mots (donc des informations de parenthésage) et d'autre part des relations entre des formes ou des groupes. Il convient donc de proposer la construction de groupes qui soient à la fois pertinents pour une analyse superficielle, mais également utilisables par un analyseur détaillé. Ces groupes sont nécessairement de premier niveau (i.e. sans constituants emboîtés),

ils ne contiennent que des éléments lexicaux. L'objectif est de définir des groupements très simples et peu ambigus. La grammaire suivante donne une idée du type de groupes pouvant être construits :

```
GV ::= [Adv[neg]] (Clit) [Aux] (Adv) V
GN ::= Det [Adv] [Adj] N[c] | [Det] N[p] | [Det] [Adj] N[p] | Pro[p]
GP ::= Prep Det [Adv] [Adj] N[c] | Prep N[p] | Prep V[ppres] | Prep V[inf]
GA ::= [Adv] Adj | [Adv] V[ppas]
Gadv ::= Adv*
```

Bien entendu, cette grammaire est largement incomplète, et de nombreuses catégories ne sont pas prises en compte, ce qui ne perturbe pas le comportement d'un analyseur superficiel. De même, d'autres règles complétant la description de ce qu'on peut considérer comme étant des syntagmes noyaux peuvent être ajoutées. Enfin, une représentation sous forme syntagmatique ne préjuge pas non plus du formalisme choisi. On peut par exemple décrire cette même information sous forme de dépendances ou de contraintes. Signalons qu'une grammaire de ce type a été utilisée lors de la campagne d'évaluation Easy (cf. [Vilnat04]). De leur côté, les relations pouvant être établies par un analyseur superficiel sont relativement générales et déterminées sur la base d'informations simples, en particulier l'ordre linéaire. Le tableau suivant propose quelques relations avec leur sémantique opérationnelle. Chaque relation est caractérisée par des propriétés qu'il est possible d'extraire de la liste des groupes précédemment construite. Nous utiliserons dans ce qui suit les notations suivantes : GX^+ pour indiquer que le groupe GX fait déjà partie d'une relation et X pour indiquer une suite quelconque d'objets.

Sujet	$(GN \prec X \prec GV) \wedge (\nexists\ GN^+ \in X)$
Aux	$(Aux \prec X \prec V[ppas]) \wedge (\nexists\ V \in X)$
Objet	$(GV \prec X \prec GN) \wedge (\nexists\ GN^+ \in X)$
Conj	$(GX \prec X \prec Conj \prec GX) \wedge (\nexists\ GX \in X)$

Les relations telles qu'elles sont définies ne permettent de spécifier qu'une partie des relations. Par exemple, la relation sujet ne prend pas en compte les inversions, de même que la relation de coordination ne permet que les coordinations simples. Le problème essentiel de ce type de relation est la surgénération. Il est cependant possible d'ajouter un niveau de contrôle spécifique, notamment concernant le type de relation possible pour une catégorie donnée ou encore en ayant recours à des informations lexicales. Une analyse intermédiaire ou détaillée tirant parti d'une analyse superficielle de ce type vise donc la construction de groupes syntaxiques de niveau supérieur ainsi que de relations complexes. Le principe consiste à utiliser en entrée les objets construits par l'analyseur superficiel. Les constituants des unités syntaxiques détaillées sont donc soit des groupes soit des catégories lexicales. Dans les deux cas, l'analyseur superficiel peut associer à ces groupes et catégories des indications en termes de probabilités permettant ainsi de contrôler le processus d'analyse profonde en réduisant l'espace de recherche de l'analyseur. Il est possible de définir les bases d'une analyse intermédiaire. Les règles de la grammaire correspondante utilisent les groupes et les relations construits par l'analyseur superficiel, ce qui permet de préciser ou d'exclure certains constituants en fonction de leur propriétés syntaxiques. Ces relations sont indiquées entre chevrons. Nous obtenons ainsi des règles de la forme :

```
SN ::= GN [GA] [Rel] [GP] <∄ mod(GP,GV)>
SV ::= GV [GN] [GP] <∄ mod(GN,GV)>
Rel ::= Pro[rel] [GN1] GV [GN2] <∄ suj(GN2,GV)>
```

```
SX ← pile_synt[en_cours]
si GC ∈ SX
    ajouter(GC, SX)
finsi sinon
    répéter
        fermer(SX);
        en_cours-;
    tant que ((GC ∉ pile_synt[en_cours]) et (ouvert(pile_synt[en_cours]))
            et (en_cours ≥ 0))
    si en_cours ≠ 0
        ajouter(GC, pile_synt[en_cours])
    finsi
pile_synt[top++] ← GC
```

Figure 3: Algorithme intermédiaire

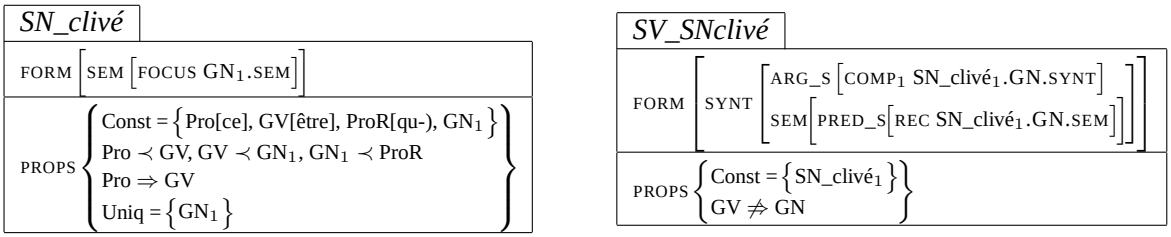


Figure 4: Description du SN clivé en GP

Un algorithme simple (cf. figure 3) consiste pour chaque groupe à vérifier s'il peut appartenir à un syntagme. On utilise pour cela une pile des syntagmes (notée *pile_synt*) construits et deux pointeurs : l'un pointant sur le sommet de la pile (noté *top*), l'autre sur le syntagme en cours (noté *en_cours*). On indique par *GC* le groupe courant, on se dote d'une fonction *ajouter_constituant(Const, SX)* permettant d'ajouter const la liste des constituants de *SX* ainsi que d'une fonction *fermer(SX)* clôturant la liste de constituants de *SX* et d'une fonction booléenne *ouvert(SX)* indiquant si *SX* est ouvert ou fermé.

Il est donc possible d'obtenir à faible coût un analyseur intermédiaire construisant des objets dont les constituants sont des groupes fournis par l'analyse superficielle. La figure (4) présente l'exemple d'une description du clivage du *SN* dans le formalisme des grammaires de propriétés (cf. [Blache05]). Cette analyse s'appuie sur deux constructions : la première décrivant le site de l'extraction, et la seconde les relations avec le site duquel l'élément a été extrait. On y constate, de la même façon que pour l'analyseur intermédiaire, l'utilisation des groupes et des relations en tant que source d'information élémentaire, tout en la complétant avec des informations propres au formalisme choisi.

5 Conclusion

L'analyse syntaxique est un problème dont les facteurs de complexité doivent être précisés. Il est pour cela nécessaire de distinguer précisément les différents types d'analyse (superficielle ou profonde) avant de proposer une caractérisation des phénomènes influant sur cette complexité.

Nous proposons dans cet article une première approche de ce problème qui permet d'envisager une coopération entre ces différentes approches. La technique proposée permet à l'analyse détaillée de s'appuyer sur les résultats de la superficielle, ce qui permet de réduire l'espace de recherche en fournissant en entrée non plus des objets atomiques, mais des informations complexes.

Références

Abeillé A. & P. Blache. (2000). "Grammaires et analyseurs syntaxiques", Traité IC2, Volume Ingénierie des langues, Hermès.

Blache P. & M. Delpui (1995) "Outil d'intégration de bases de connaissances lexicales aux analyseurs syntaxiques", in actes des Journées "Lexicomatique et Dictionnairique".

Blache P., J.-M. Balfourier & T. van Rullen (2002), "From Shallow to Deep Parsing Using Constraint Satisfaction", in proceedings of *COLING-2002*

Blache P. (2005) "Property Grammars: A Fully Constraint-Based Theory", in *Constraint Satisfaction and Language Processing*, H. Christiansen & al. (eds), Springer-Verlag LNAI 3438.

Briscoe, E., J. Carroll, J. Graham & A. Copestake (2002) "Relational evaluation schemes", in proceedings of the *Beyond PARSEVAL Workshop, LREC-02*.

Carroll J. & T. Briscoe (2001) "High Precision Extraction of Grammatical Relations", in proceedings of *IWPT-01*.

Carroll J., G. Minnen & T. Briscoe (2003) "Parser Evaluation. Using a Grammatical Relation Annotation Scheme", in A. Abeillé (ed) *Treebanks: Building and Using Syntactically Annotated Corpora*, Kluwer.

Crysmann B. A. Frank, B. Kiefer, S. Müller, G. Neumann, J. Piskorski, U. Schäfer, M. Siegel, H. Uszkoreit, F. Xu, M. Becker & H. Krieger (2002) "An Integrated Architecture for Shallow and Deep Processing", in proceedings of *ACL-02*.

Frank A., M. Becker, B. Crysmann, B. Kiefer & U. Schäfer (2003) "Integrated Shallow and Deep Parsing: TopP meets HPSG", in proceedings of *ACL-03*.

Grover C. & A. Lascarides (2001) "XML-Based Data Prepartion for Robust Deep Parsing", in proceedings of *ACL/EACL-01*.

Hinrichs E. & K. Simov eds.(2004) Proceedings of the Workshop *"Combining Shallow and Deep Processing for NLP", ESSLLI-04*.

Johnson M. (2002) "A Simple Pattern-Matching Algorithm for Recovering Empty Nodes and their Antecedents", in proceedings of *ACL-02*.

Marimon M. (2002) "Integrating Shallow Linguistic Processing into a Unification-Based Spanish Grammar", in proceedings of *COLING-02*.

Neumann G., C. Braun & J. Piskorski (1999) "A Divide and Conquer Strategy for Shallow Parsing of German Free Texts", in proceedings of *ANLP-00*.

Puver M. & R. Kempson (2004) "Incremental Parsing or Incremental Grammar? ", in proceedings of the workshop *Incremental Parsing: Bringing Engineering and Cognition Together, ACL-04*.

Uszkoreit H. (2002) "New Chances for Deep Linguistic Processing", in proceedings of *COLING-02*.

Vilnat A., L. Monceaux, P. Paroubek, I. Robba, V. Gendner, G. Illouz & M. Jardino (2004) "Annoter en constituants pour évlauer des analyseurs syntaxiques", in actes de *TALN-04*.