Représenter et utiliser les contraintes de la langue orale à l'aide d'une grammaire lexicalisée d'arbres adjoints

Patrice Lopez

Représenter et utiliser les contraintes de la langue orale à l'aide d'une Grammaire lexicalisée d'arbres adjoints

Patrice Lopez

LORIA
BP 239 - 54506 Vandœuvre-lès-Nancy, France
lopez@loria.fr

Résumé

Cet article souligne le problème de l'analyse grammaticale des énoncés oraux incomplets en contexte de dialogue homme-machine. Des contraintes minimales de l'oral sont cependant exploitables afin de rester prédictif face aux phénomènes d'ellipses. Nous proposons un enrichissement du formalisme LTAG afin de capter ces contraintes et d'adapter à l'oral une grammaire initialement conçue pour l'écrit.

1. Introduction

Un des enjeux majeurs de l'analyse syntaxique appliquée au dialogue homme-machine est de pouvoir traiter les énoncés incomplets normalement utilisés par les utilisateurs. Les verbalisations employées étant contraintes à un type de dialogue particulier (dialogue de commande, d'assistance, etc.) et au domaine de l'application, on peut attendre de la part des systèmes de dialogue une certaine robustesse et des traitements opportuns lorsque l'énoncé s'écarte d'une normalisation qui n'est souvent caractéristique que de l'écrit. En particulier, des études empiriques (Carbonell, 1983) ont montré que, si l'utilisation d'énoncés très fragmentaires était chose courante dans la communication entre humains, elle l'était tout autant en situation de dialogue homme-machine. De plus, bien que les utilisateurs puissent facilement éviter l'emploi de structures syntaxiques complexes, ils peuvent très difficilement se contraindre à l'emploi d'expressions uniquement complètes.

Afin de rester robuste à ce phénomène et à d'autres spécifiques de l'oral (reprises, répétitions, etc.), des techniques d'analyse superficielle ou stochastique ont été employées visant à extraire avant tout l'information utile. En effet, dans la mesure où ces phénomènes sont difficiles à prédire, il peut sembler inutile de s'attacher à une grammaticalité jugée illusoire en situation de dialogue oral. A l'image de (van Noord *et al.*, 1998), nous tentons de montrer que l'utilisation de formalismes grammaticaux élaborés peut être bénéfique à la fois en terme de robustesse, d'efficacité et de couverture de la langue pour les systèmes de dialogue oral homme-machine. Si des techniques d'analyse robuste restent nécessaires pour extraire des énoncés un maximum d'information même si une analyse complète échoue (par exemple tous les constituants pos-

sibles), un effort de modélisation des contraintes orales portant sur les énoncés de l'utilisateur nous semble nécessaire. La plupart des recherches sur la notion de grammaire de la langue se sont concentrées sur les problèmes de *compétence linguistique* plutôt que de *performance linguistique*, autrement dit il s'agissait de caractériser les énoncés perçus comme corrects plutôt que ceux effectivement utilisés. En se conformant à une théorie grammaticale particulière, on se conforme aux principes de correction qui fondent en fait cette théorie.

Cet article définit des principes additionels en vue d'adapter une grammaire lexicalisée (en l'occurance une Grammaire Lexicalisée d'Arbres Adjoints, LTAG) conçue pour l'écrit à l'analyse d'énoncés oraux et incomplets. Des expérimentations sur corpus permettent d'évaluer l'intérêt de l'approche proposée.

2. Une vue empirique sur les énoncés incomplets

Nous présentons ici le résultat de l'analyse de l'ensemble des énoncés utilisateurs du corpus de commande vocale Gocad (Chapelier *et al.*, 1995), soit 862 énoncés, obtenus grâce à une simulation de type Magicien d'Oz. Nous avons utilisé une LTAG suivant les principes définis dans (Abeillé, 1991), donc une grammaire conçue pour l'analyse de l'écrit et un algorithme d'analyse ascendant par *chart* (Lopez, 1998). On peut considérer que la grammaire employée (529 formes fléchies d'entrée et environ 80 motifs d'arbres élémentaires) approxime le sous-langage d'application au sens de (Deville, 1989). Nous nous sommes d'autres part fondés sur les critères de l'écrit pour obtenir une analyse complète (la catégorie P est l'unique axiome et la solution ne doit pas présenter de nœuds pieds ou de substitution non saturés). La table 1 présente également le nombre moyen d'analyses partielles [1] obtenues en fin d'analyse.

Nb moyen de mots par énoncé	% d'anayse complètes	Nb moyen d' analyses/énoncé	Nb moyen de resultats partiels/énoncé
6,4	64,7	1,5	7,1

TAB. 1 – *Le corpus Gocad et son analyse avec une grammaire LTAG pour l'écrit*

Les résultats présentés table 1 montrent que la proportion de phrases complètement analysées est relativement faible. Une taxonomie des échecs d'analyse montre que 50,9% des erreurs viennent uniquement d'énoncés incomplets au sens de l'écrit, le reste pouvant être attribué à d'autres phénomènes (reprises, répétitions, ...) ou à un phénomène non couvert par la grammaire (énumération par exemple).

3. Contraintes minimales sur les énoncés incomplets

Nous étudions ici les ellipses sous l'angle de leur réalisation syntaxique, bien entendu la résolution d'une ellipse (projection de la structure complète), comme sa prédiction, dépend du contexte : historique du dialogue, entité saillante, etc. (Carberry, 1989) (Sauvage, 1992). Bien que les énoncés de l'utilisateur puissent être très incomplets, des contraintes purement syntaxiques existent cependant dans la langue orale, comme le montre les échanges suivants :

(1) *Q : Où voullez-vous aller?*
R : **à** *Cargèse.*

1. Une analyse partielle correspond ici à l'extension maximale d'un îlot bien reconnu, donc, dans le cadre de l'analyse par *chart* mise en œuvre, à un *item* qui n'est l'origine d'aucun autre *item.*

(2) *Q : Prends-tu ton café avec du sucre?*
R : Non, **sans**.

Dans l'exemple (1) la préposition « à » requiert un argument mais « sans » dans (2) peut être utilisé avec ou sans l'introduction d'argument tout en restant une structure syntaxiquement correcte et autonome. Ces exemples montrent d'une part un degré de contraintes minimales présentées par l'oral que nous souhaitons ici pouvoir prendre en compte, et d'autre part que la capture de ces contraintes demande un formalisme fortement lexicalisé puisque le contexte syntaxique minimal varie par exemple ici d'une préposition à l'autre.

Nous nous intéressons ici aux ellipses rencontrées dans le cadre de dialogue de commande finalisé et de couple question/réponse. Les énoncés présentent alors un nombre important d'ellipses sur la tête prédicative des constituants que l'on peut illustrer avec les exemples suivants :

(3) *système : Le milieu de quoi?*
utilisateur : **du** *carré.*

(4) *système : Quel objet?*
utilisateur : **le petit.**

Au vue de ces exemples, il semble donc possible d'exprimer des contraintes purement syntaxiques concernant la verbalisation d'une ellipse. Cependant les cas particuliers d'*emploi en mention* échappent à toute contrainte syntaxique. Dans la mesure où tout mot est suceptible d'être ainsi utilisé, comme le montre l'exemple (5), les contraintes que nous souhaitons représenter ici excluent donc ce style d'emploi [2].

(5) *Q : Vous avez dit prendre, pendre ou fendre?*
R : prendre.

4. LTAG et structures elliptiques

Nous avons choisi d'utiliser le formalisme des Grammaires Lexicalisées d'Arbres Adjoints (Joshi & Schabes, 1992). Ce formalisme repose sur une définition mathématique et ne constitue donc pas en soi une théorie linguistique mais un outil de représentation structurelle; des principes complémentaires, comme par exemple ceux de (Abeillé, 1991), doivent être définis pour en faire un réel modèle de la langue. Deux principales raisons ont motivé ce choix :

- la lexicalisation et le principe de localité étendu permettent d'exprimer de façon simple les contraintes lexicales dans des arbres partiels d'analyse (arbres élémentaires),
- des analyseurs ascendants robustes, des modèles stochastiques et des précompilations efficaces des grammaires existent pour les Grammaires Lexicalisées d'Arbres.

Des grammaires LTAG à large couverture ont été implantées pour l'anglais (Doran *et al.*, 1994) et le français (Abeillé *et al.*, 1994) mais dans l'optique de l'analyse de textes écrits. Dans la mesure où ces grammaires ne sont pas elliptiques, l'analyse des exemples présentés serait impossible. Cela signifie donc qu'ayant échoués sur le plan syntaxique, leur interprétation sera obtenue uniquement par la sémantique, le risque étant une combinatoire prohibitive. La question de la prédictivité d'une LTAG est importante dans la mesure où cette caractéristique permet de réduire le nombre des hypothèses artificielles et donc la complexité moyenne des analyses.

2. Dans les corpus de DHM simulé, on remarque que les emplois en mention n'apparaissent qu'à l'initiative du système, après des questions précises donnant le choix d'un mot parmi une liste fermée. Ce cas particulier peut être évité par l'emploi d'un rang ou couvert par des heuristiques spécifiques lors d'une première étape d'analyse lexicale en fonction de l'état du dialogue.

Couvrir les ellipses augmente le nombre des contextes syntaxiques à prendre en compte au cours de l'analyse, un des problèmes essentiels est ici de concevoir une grammaire elliptique sans augmenter dramatiquement le nombre d'arbres élémentaires. En particulier, pour le corpus Gocad, les 64,7% d'énoncés non elliptiques correctement analysés verraient leur complexité moyenne d'analyse sensiblement augmenter. Certains principes définis par (Abeillé, 1991) concernent la relation entre un arbre élémentaire et la sémantique, en particlier un arbre élémentaire ne doit correspondre qu'à un unique prédicat (non nul). Considérant ces principes, on peut envisager deux premières solutions pour analyser des énoncés fragmentaires en évitant autant que possible la multiplication des arbres élémentaires :

- modifier le mécanisme d'analyse en autorisant la substitution à ne pas être obligatoire, les nœuds non saturés étant autant de marques d'ellipses. Cependant les exemples (1) et (2) montrent que le lexique doit indiquer clairement les possibilités structurelles d'ellipses afin de contraindre l'aspect compositionnel ;
- adapter une grammaire non elliptique en considérant toute catégorie comme axiome. Mais ceci aboutira à une multiplication des ambiguïtés artificielles, si on veut rester prédictif on ne peut pas considérer toute structure (par exemple un déterminant) comme complète.

En se fondant sur ces deux critères, on constate que la proportion d'analyses complètes obtenues passe de 65,9% à 78.3%, 15 cas d'ellipses (sur les 155 initiaux) échappant à la grammaire mais pouvant être récupérés *a priori* par un mécanisme de rattrapage. Néanmoins, si la couverture est donc bonne, l'aspect prédictif est ignoré entrainant une surgénération des analyses partielles.

4.1. Limite du formalisme

Le formalisme LTAG emploie deux types d'opérations pour la combinaison des arbres élémentaires, l'adjonction et la substitution, chacune d'elle pouvant être associée à deux rôles, considérant les choix linguistiques de (Abeillé, 1991) :

- un rôle formel : les substitutions sont obligatoires et représentent donc les incomplétudes d'une structure, les adjonctions sont optionnelles représentant des rattachements non-obligatoires ;
- un rôle linguistique : le rattachement d'un modifieur se réalise par adjonction, celui des arguments par une substitution.

Si on considère l'analyse de structures incomplètes telles qu'observées en situation de dialogue homme-machine, il est possible d'avoir des arguments non obligatoires, par exemple (2), et des modifieurs apparaissant seuls (4). Ceci fait que le rôle formel devient incompatible avec le rôle linguistique. Nous proposons dans la section suivante un enrichissement du formalisme permettant de préserver le rôle linguistique.

4.2. Enrichissement du formalisme

Afin de préserver l'interprétation linguistique des dérivations, nous utilisons toujours un arbre auxiliaire pour représenter les modifieurs et la substitution pour les rattachements prédicats arguments selon les critères présentés par exemple dans (Candito, 1999). Nous souhaitons de plus représenter que l'occurrence d'un mot implique celle d'une structure syntaxique de façon obligatoire ou optionnelle. Par exemple l'occurrence d'un déterminant implique celle d'un nom, mais la présence d'un nom n'implique pas obligatoirement celle d'un déterminant. En

conséquence nous introduisons les annotations suivantes au niveau des nœuds terminaux :

- $\downarrow$: le nœud peut accueillir de manière *optionnelle* une substitution d'un constituant argument, voir exemples figure 1 (a) and (g) ;
- $\downarrow\uparrow$: substitution *obligatoire* d'un constituant argument (b) ;
- $*$: adjonction *optionnelle* du modifieur, voir (a) et (f) ;
- $* \uparrow$: adjonction *obligatoire* du modifieur, voir (e) ;
- ancre simple : occurrence *obligatoire* d'une ancre dans l'énoncé (cas general) ;
- ancre $\downarrow$: l'occurrence de l'ancre dans l'énoncé est *optionnelle*[3], voir (h).

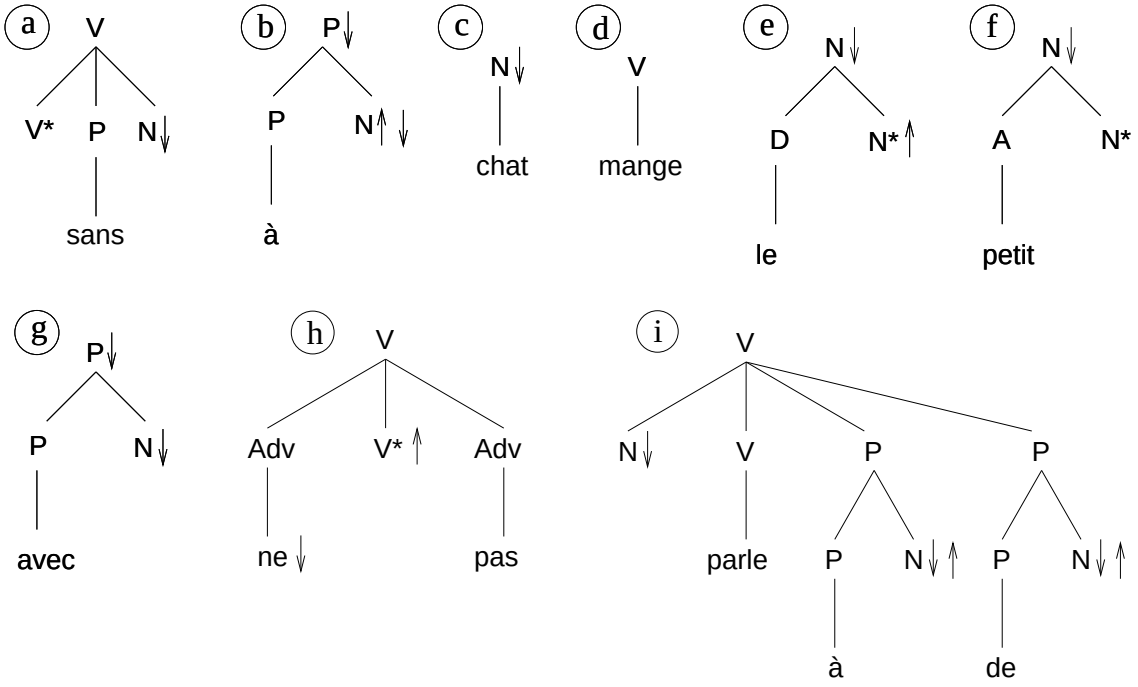

FIG. 1 – *Exemples d'arbres élémentaires enrichis*

Un constituant peut être complet mais toutefois être employé avec une ellipse du prédicat dont il est argument. Afin d'aider des niveaux supérieurs de traitement dans la résolution de ce style d'ellipses nous marquons les nœuds racine de la façon suivante :

- $\downarrow$: le constituant est obligatoirement argument d'un prédicat, voir par exemple 1 (c) ;
- pas de marque : le constituant est argument de manière optionnelle, voir (d).

L'enrichissement proposé permet de construire méthodologiquement un lexique LTAG représentant les contraintes portant sur les énoncés incomplets, ou d'en adapter un existant. Compte tenu de ces annotations, nous pouvons introduire le principe de complétude d'une analyse d'un énoncé oral de la façon suivante : tout arbre dérivé présentant en fin d'analyse au moins une marque $\uparrow$ ne constitue pas une analyse acceptable. Ceci offre une possibilité de filtrage des analyses partielles. La détection et la caractérisation des ellipses sont obtenues avec les nœuds présentant les marques $*$ ou $\downarrow$, facilitant les traitements ultérieurs.

3. Ce cas peut être utile dans un contexte oral pour représenter par exemple la chute du discordantiel *ne* dans le cas d'une négation, comme illustré par (h).

5. Retour au corpus Gocad

Nous avons mené cette fois une expérimentation en tenant compte des principes et annotations présentés ici et du critère de complétude des énoncés oraux. Le taux d'analyse complète passe à 79,1% (plus que huit cas d'ellipse sont à prendre en compte par un mécanisme de réparation). De plus, l'analyse présente un gain en terme de réduction de la combinatoire des analyses partielles. On constate que 39,8% des analyses partielles posent problème et ne forment pas des constituants complets selon les principes introduits. Ceci présente deux intérêts :

- séparer les analyses partielles acceptables des analyses agrammaticales pouvant nécessiter un mécanisme de réparation (répétition, reprises, ...) ;
- limiter le nombre d'hypothèses à examiner pour contraindre un système de reconnaissance de la parole guidé par une grammaire.

Afin de diminuer la combinatoire d'énoncés partiels, la prise en compte au plus tôt des contraintes sémantiques et pragmatiques apparaît nécessaire et peut reposer de manière uniforme sur l'utilisation de grammaires TAG synchrone (Shieber & Schabes, 1994).

6. Remerciements

Je remercie Bertrand Gaiffe, David Roussel et Susanne Salmon-Alt pour leurs enrichissants commentaires sur ce travail, toute erreur et imprecision étant de mon seul fait.

Références

ABEILLÉ A. (1991). *Une grammaire lexicalisée d'arbres adjoints pour le français*. PhD thesis, Paris 7.

ABEILLÉ A., DAILLE B. & HUSSON A. (1994). FTAG : An implemented Tree Adjoining grammar for parsing French sentences. In *TAG+3*, Paris.

CANDITO M.-H. (1999). *Structuration d'une grammaire LTAG : application au français et à l'italien*. PhD thesis, University of Paris 7.

CARBERRY S. (1989). A Pragmatics-Based Approach To Ellipsis Resolution. *Computational Linguistics*, **15**(2), 75–96.

CARBONELL J. G. (1983). Discourse Pragmatics and Ellipsis Resolution in Task-oriented Natural Languages Interfaces. In *ACL'83*, Cambridge.

CHAPELIER L., FAY-VARNIER C. & ROUSSANALY A. (1995). Modelling an Intelligent Help System from a Wizard of Oz Experiment. In *ESCA Workshop on Spoken Dialogue Systems*, Vigso, Danemark.

DEVILLE G. (1989). *Modelization of task-Oriented Utterances in a Man-Machine Dialogue System*. PhD thesis, University of Antwerpen, Belgique.

DORAN C., EGEDI D., HOCKEY B. A., SRINIVAS B. & ZAIDEL M. (1994). XTAG System - A Wide Coverage Grammar for English. In *COLING*, Kyoto, Japan.

JOSHI A. K. & SCHABES Y. (1992). Tree Adjoining Grammars and lexicalized grammars. In M. NIVAT & A. PODELSKI, Eds., *Tree automata and languages*. Elsevier Science.

LOPEZ P. (1998). Analyse guidée par la connexité de TAG lexicalisées. In *Conférence sur le Traitement Automatique du Langage Naturel (TALN'98)*, Paris, France.

SAUVAGE C. (1992). *Parallélisme et traitement automatique des langues, application à l'analyse des énoncés elliptiques*. PhD thesis, Université Paris XI Orsay.

SHIEBER S. & SCHABES Y. (1994). Restricting the weak-generative capacity of synchronous tree-adjoining grammars. *Computational Intelligence*, **10**, 371–385.

VAN NOORD G., BOUMA G., KOELING R. & NEDERHOF M.-J. (1998). Robust Grammatical Analysis for Spoken Dialogue Systems. *Natural Language Engineering*, **1**, 1–48.