Wide-Coverage French Syntax and Semantics using Grail*

Richard Moot
LaBRI (CNRS, Bordeaux) & SIGNES (INRIA Bordeaux SW)
351 cours de la Libération, 33405 Talence, FRANCE
Richard.Moot@labri.fr

Abstract. The system demo introduces Grail, a general-purpose parser for multimodal categorial grammars, with special emphasis on recent research which makes Grail suitable for wide-coverage French syntax and semantics. These developments have been possible thanks to a categorial grammar which has been extracted semi-automatically from the Paris 7 treebank and a semantic lexicon which maps word, part-of-speech tags and formulas combinations to Discourse Representation Structures.

Résumé. Cette démonstration décrit Grail : un analyseur syntaxique pour grammaires catégorielles. Elle met l'accent sur les recherches récentes qui ont permis à Grail de donner des analyses syntaxiques et sémantiques du Français. Ces développements sont possibles grâce à une grammaire extraite semi-automatiquement du corpus de Paris 7 ainsi qu'un lexique sémantique qui traduit des combinaisons de mots, des étiquettes syntaxiques et des formules en Discourse Representation Structures.

Mots-clés : Discourse Representation Theory, grammaires catégorielles.

Keywords: Categorial grammar, Discourse Representation Theory, type-logical grammar.

1 Introduction

Grail is platform for developing and parsing multimodal categorial grammars (Moortgat, 1997). Up until now Grail has mainly been used for the development of fairly small grammars treating specific linguistic phenomena, often with an emphasis on the syntax-semantics interface[1]. The current system demonstration will showcase a prototype wide-coverage multimodal categorial grammar for French which outputs discourse representation structures in the style of Bos *et al.* (2004) for unseen text (current newspaper articles or sentences proposed by the audience). Though this is work in progress and the semantic lexicon still needs to be significantly extended, early results are promising.

2 The French Grammar

The French grammar used by Grail has been semi-automatically extracted from the Paris 7 treebank (Abeillé *et al.*, 2003). The lexical ambiguity in the resulting grammar, even after a significant amount

* The software described in this system demonstration has benefited from a grant of the conseil régional d'Aquitaine.

[1] Grail can be downloaded under LGPL — together with a number of example grammars — at
`http://www.labri.fr/perso/moot/grail3.html` (source code and grammars)
`http://www.labri.fr/perso/moot/tutorial/` (tutorial)

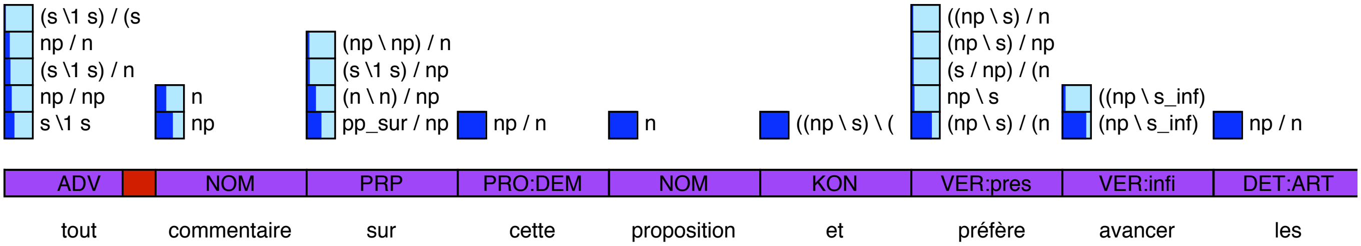

Figure 1: Screenshot of the interface to the supertagger

of lexical cleanup, is enormous. As a consequence, frequent lexical items have many lexical formulas assigned to them; for example, “et” (69), the comma “,” (54) or “est” (57, the other forms of “être” are slightly easier).

Fortunately, we can reliably estimate the most likely formula assignments based on local information only (surrounding words and part-of-speech tags, preceding formula assignments). This approach is called *supertagging* (Bangalore & Joshi, 2010).

The part-of-speech tagger and supertagger of Clark & Curran (2004) have been trained on the extracted grammar and achieve a precision of 98.4% for the part-of-speech tagger and 90.5% for the supertagger. Coverage for the supertagger can be significantly increased by using a multi-tagger, assigning all supertags with a probability of greater than a certain factor of the probability assigned to the best supertag to each word. This factor is a parameter which can be used to trade coverage of the corpus for speed of parsing. When keeping all supertags with a probability of greater than 1% of the probability assigned to the best supertag, the supertagger assigns an average of 4.7 formulas to each word, which includes the correct formula in 98.4% of the cases.

To illustrate the supertagger and give an example indicative of what can be expected during the demo, the following sentence has been taken from one of today’s French newspapers[2].

1. *Le gouvernement refuse tout commentaire sur cette proposition et préfère avancer les chiffres positifs récoltés par la mesure.*

Figure 1 shows part of the output of the part-of-speech tagger and the supertagger for this example sentence. The relative probability of the different supertags is indicated by the percentage of the block next to it which is displayed in a darker blue.

As can be seen, the number of assigned supertags for each word remains quite reasonable. For example, “et” is assigned only a single formula, that of a VP conjunction — taking a VP or $np\backslash s$ to its right as well as its left in order to produce a new $np\backslash s$. Categories about which the supertagger is less sure — the adverbs ADV, prepositions PRP and the different verb forms VER:X — have multiple formula assignments.

So the preposition “sur” has four assigned categories: pp_{sur}/np, the category for a preposition which selects a noun phrase to its right to become a *pp* verb argument, being the supertagger’s first choice, with $(n\backslash n)/np$, the category for a preposition modifying a noun as its second choice. The two other possibilities (adverbial sentence modifier and noun phrase modifier) being considered considerably less likely.

[2] http://www.lepoint.fr/actualites-societe/2010-05-24/solidarite-le-lundi-de-pentecote-ne-fait-pas-recette/920/0/458325, visited 24 May 2010

3 Syntax-Semantics Interface

Categorial grammars come with a simple and transparent semantics: since categorial parses/derivations are a proper subset of derivations in intuitionistic logic, which by the Curry-Howard isomorphish are in a 1-1 correspondance with lambda terms in the simply typed lambda calculus, each parse corresponds to a different lambda term.

The use of the simply typed lambda calculus has been popularized by Richard Montague and many others, though it should be noted that the simply typed lambda calculus is compatible with more modern dynamic semantic theories like Discourse Representation Theory (Kamp & Reyle, 1993) and Montegovian Dynamics (de Groote, 2006) as well.

The current prototype uses DRT as target for its semantic output, though a Montegovian Dynamics semantic lexicon is being developed as well. There are many reasons for having chosen DRT as the semantic language, the principal reasons being:

1. there is an impressive body of research into a large number of semantically interesting phenomena which have been formalized in DRT,
2. while providing a visually simple presentation of its semantics in the form of discourse representation structures (DRSs) there is an easy translation of the DRSs into first-order logic.[3]

A semantic lexicon, in the style of the one used for English by Bos *et al.* (2004), is currently under active development. The lexicon contains two main parts: the first assigns lambda terms to specific word-formula pairs, these are frequently occurring words which require some form of special treatment, eg. the conjunctions such as “et” and relativizers such as “que” and “dont”. The second part contains default rules, which assign a semantics to words not in the first lexicon, based only on the part-of-speech tag, formula and lexical lemma. This lexicon contains the ‘open’ classes such as nouns and verbs. The current lexicon contains entries for over 300 words and has around 150 default rules.

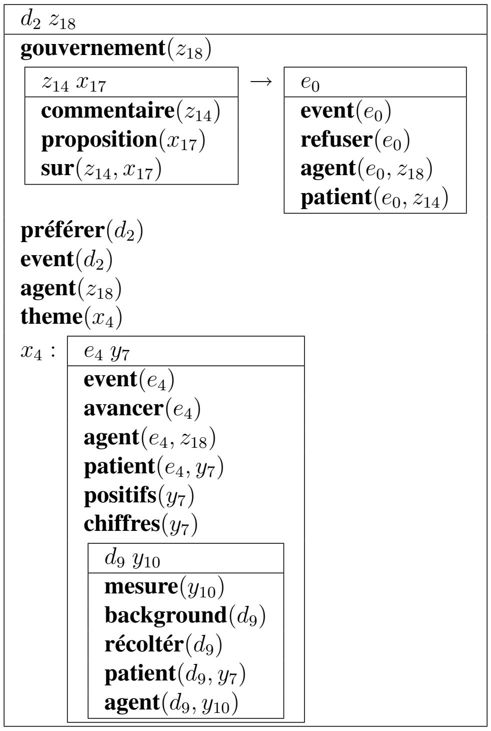

Figure 2: Grail LaTeX output

Figure 2 shows the Grail output for the example sentence of Section 2. The variables d, e and f are event variables and variables x, y and z are entity variables. So d_2 in Figure 2 corresponds to an event of “préférer”, with *agent* (subjet) z_{18}, which is a variable denoting an entity with the property of being “gouvernement”. The *theme* of the event d_2, ie. the situation being preferred, is the embedded DRS with label x_4. Note that z_{18} is agent of this embedded DRS x_4 as well (since “préférer” is a subjet control verb) as well as agent of the top-right embedded DRS at the right-hand side of the implication since, as discussed in the previous section, our example phrase conjoins two verb phrases with “le gouvernement” being the subjet of both.

[3] Of course, there are some drawbacks as well, which include the absence of the higher-order constructions allowed by the simply typed lambda calculus and the complications arising from what essentially amounts to the use of free variables in DRT semantics to handle anaphora resolution.

4 Software Requirements and Licensing Conditions

Grail is programmed in SWI Prolog[4]. All Grail files are free software, licensed under the GNU lesser general public license (LGPL). This includes all Grail source files and libraries, the semantic lexicon as well as the model files for the part-of-speech tagger and for the supertagger. The SWI Prolog kernel is licensed under GNU LGPL and the libraries under GPL.

Grail interfaces with the wide-coverage French grammar using the Clark & Curran (2004) part-of-speech tagger and supertagger[5], available under a license for non-commercial use. For ease of use, a user-interface to Grail, the part-of-speech tagger and the supertagger have been programmed in TclTk[6] (shown in Figure 1). TclTk is licensed using a BSD-style license.

Optionally, but highly recommended, Grail produces LaTeX[7] output of the semantics (shown in Figure 2) and GraphViz[8] output of both the supertagger output and (partial) parse results. LaTeX is licensed under the LaTeX Project Public License. GraphViz is licensed under the Common Public License.

All software has been tested on both UNIX/Linux-based systems and Mac OS X.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks: Building and Using Parsed Corpora*, chapter 10, p. 165–187. Dordrecht: Kluwer.

S. BANGALORE & A. K. JOSHI, Eds. (2010). *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*. MIT Press.

BOS J., CLARK S., STEEDMAN M., CURRAN J. R. & HOCKENMAIER J. (2004). Wide-coverage semantic representation from a CCG parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, p. 1240–1246, Geneva, Switzerland.

CLARK S. & CURRAN J. R. (2004). Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL-2004)*, p. 104–111, Barcelona, Spain.

DE GROOTE P. (2006). Towards a montegovian account of dynamics. In *Proceedings of Semantics and Linguistic Theory XVI*. to appear with CLC publications.

KAMP H. & REYLE U. (1993). *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers.

MOORTGAT M. (1997). Categorial type logics. In J. VAN BENTHEM & A. TER MEULEN, Eds., *Handbook of Logic and Language*, chapter 2, p. 93–177. Elsevier/MIT Press.

[4] `http://www.swi-prolog.org`, note that Grail requires a full install which includes the optional libraries

[5] `http://svn.ask.it.usyd.edu.au/trac/candc/wiki`

[6] `http://www.tcl.tk`

[7] `http://www.latex-project.org`

[8] `http://www.graphviz.org`