Predicative LTAG grammars for Term Analysis

Patrice Lopez* and David Roussel†

*DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
lopez@dfki.de

†Aerospatiale Matra
Centre Commun de Recherches
Suresnes, France
David.Roussel@aeromatra.com

Abstract

In a restricted domain and task, we propose that the elementary tree backbones represent statically the predicative level and the possible distribution of arguments while the syntactic categories and constraints would be only processed dynamically by the way of features. The resulting grammar can be viewed as an intermediate level between the surface syntax of a sentence and its conceptual representation. In addition to possible speed efficiency and robustness relevance, an interesting property is that such a grammar could be tested in a straightforward way to integrate contraints provided by additional trees and to inject progressively semantic and pragmatic constraints during the analysis.

1. Introduction

Considering applications such as spoken annotation of elements into a specific virtual environment, the most important task is first to identify referred objects or terms among several speech hypotheses. Given these expectations, how can be used the assets of a LTAG grammar with robustness? To address this question, we propose a Feature-Based LTAG grammar focusing on the semantic and predicative level while the pure syntactic processing is achieved by the two-step unification mechanism. Before introducing this predicative LTAG grammar, we define the applicative framework.

2. From Terms extraction to spoken annotations

The research project under consideration is based on a virtual platform (which represents an architecture of aeronautical components and a terminological model obtained from technical documents (example : cautions to set on the manipulation of components).

Let us clarify that first the virtual platform (i.e. a 3D scene) is used as an interface between the desing and assembly tasks. The aim of this interface is to let people easily move in a complex architecture, to display or mask related annotations, and to gather vocal synthetic annotations that overlap one or several elements of a scene (example : recommendations for people of a related trade).

Secondly, the terminology of the technical documents is ideally subjected to editorial constraints and is getting close to a controlled language. A terms extraction and clusterization based on statistical criteria supply classes of elements. Then, an expert is efficient to grab the terms in a knowledge base containing ontological and conceptual relationships. Tools of the market are helpful for these tasks (Fig. 1, Fig. 3). The aim of this step is twofold:

- Build a model used to check the cohesion from various technical documents or versions.
- Identify the stable terms and build up various terminological resources (authoring memory, multilingual thesaurus needed for automatic language processing). For example, the

knowledge base designed by the experts is used to categorize various technical documents within an Information Retrieval System. For the spoken annotation purpose, we derived constraints from the knowledge base in order to restrict the combination between technical properties (ex: float valve, needle valve), functionalities (ex: drain valve, directional valve) and the system in which a unit is used (ex: water valve, bleed valve). By this way, terms like *water drain valve*, *electrical drain valve* are well recognized, but some other complex terms are rejected.

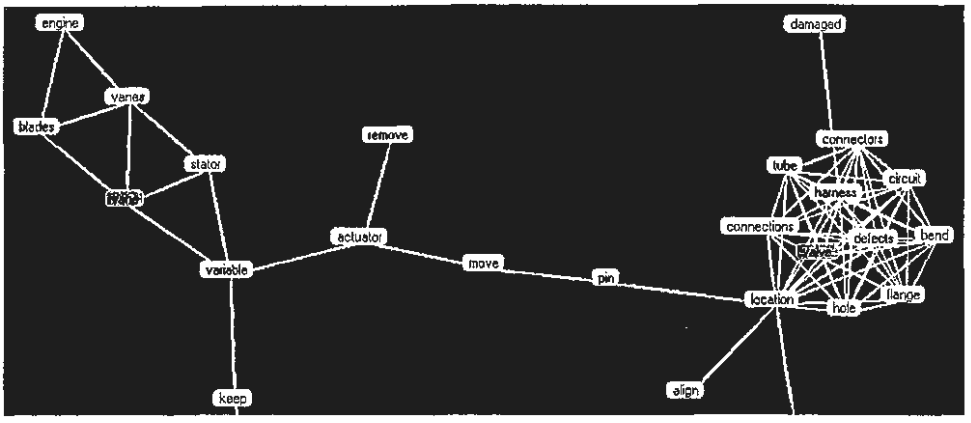

Figure 1: Cluster of words computed for technical documentation extracts. Note that the word *valve* covers at least three notions expressed in French by the terms *valve*, *soupape* and *vanne*

<C.2616> MAKE SURE YOU WILL NOT CAUSE UNWANTED CHANGES TO OTHER SYSTEMS BEFORE YOU PUSH THE ENG 1 (2, 3 OR 4). WHEN YOU PUSH THE ENG 1 (2, 3 OR 4) FIRE PUSHBUTTON SWITCH THESE VALVES CLOSE: THE LP FUEL VALVE. THE HYDRAULIC FIRE VALVE. THE BLEED AIR VALVES. THE ANTI-ICE VALVES. THE AIR CONDITIONING PACK VALVES.

Figure 2: Example of caution integrated in a structured technical documentation.

Taking advantage of lexical resources obtained from technical procedures called "warnings and cautions" (see Fig. 3), the MRTERESA project (Multilocutor speech Recognition, TERms Extraction and Spoken Annotation) consists in the customization of a speech recognizer for vocal annotations, a robust term analysis of speech recognition hypotheses and vocal annotations indexing with regard to components existing in a virtual scene. If necessary, the indexing has to be confirmed by the users. The robust terms analysis relies upon:

- A mapping between lexicalized elementary trees and technical terms. Some category labels in these trees are semantic types that belong to an ontology.
- A representation of the terms variability in the spoken annotations thanks to the TAG substitution and adjonction operations. This variability results from spatial relations between the displayed objects and the spontaneity of the verbalizations.
- Semantic labels compatibility constraints for modification and dependency relations
- If necessary, syntactic constraints are applied to filter out speech recognition hypotheses.

3. Syntactic vs. semantic Dependencies

The *semantic head* is the lexical unit that represents the semantic type of the interpretation of a given phrase structure. We consider that the *syntactic head* is the lexical unit that constraints the

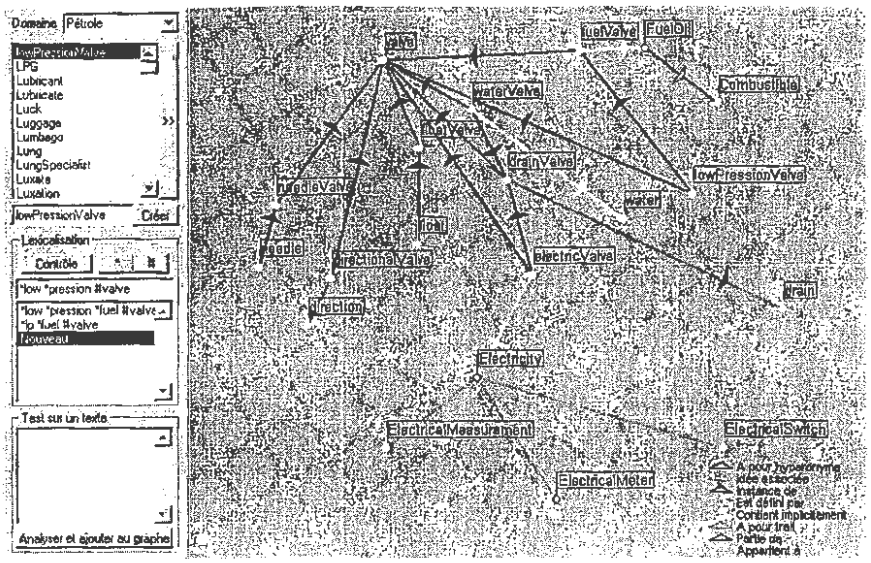

Figure 3: Example of knowledge base from few valves achieved with a tool of the market

morphosyntactic and mode features of the phrase structure it belongs to. The LTAG formalism is well suited to localize semantic dependencies, but is limited to represent syntactic dependencies for very frequent phenomena as object extraction with auxiliary.

When we use the term *localizing semantic dependencies* for a LTAG grammar, we suppose that the elementary trees have been designed properly to capture this kind of dependencies, i.e. that the elementary trees respect the Predicate-Argument (PA) and Semantic Consistency (SC) principles introduced in (Abeillé, 1991). These principles stipulate that a lexicalized elementary tree corresponds to an unique semantic unit (*semanteme*) and that we have a terminal node (substitution or foot node) per argument expected by the corresponding semanteme. In our approach we systemize the localization of semantic dependencies: we drop out from the elementary tree backbones all the aspects which traditionally refer to syntactic categories and replace them dynamically with semantic types.

4. A new definition for the elementary trees

The first point is to capture in an elementary tree a particular word distribution and the corresponding predictive structure under the form of semantic dependencies. Closely to the solution proposed in (Abeillé, 1992) for the representation of this level, we use the following *predicative categories* as node labels of elementary trees:

- Formula (**F**) or proposition representing the association of a relation and its arguments.
- Term (**T**) which corresponds to the non-relational semantic heads.
- Relation (**R**).
- Property (**P**).
- Null (**N**): used for semantically empty nodes (in general preterminal nodes of co-anchors, semantically empty prepositions or auxiliaries).

Top and bottom features are added on this backbone in order to check syntactical constraints at the end of the parsing. The figure 4 gives examples of Feature-Based predicative LTAG elementary trees. During the lexicalization process, semantic types are added to the LTAG tree backbone according to the semanteme that the elementary tree represents and an ontol-

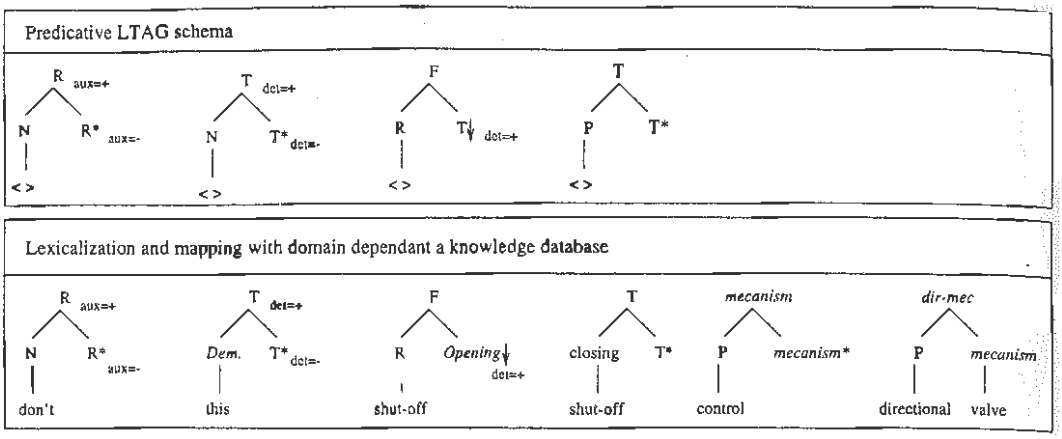

Figure 4: Examples of predicative LTAG elementary trees and their lexicalization

ogy obtained as explained in section 2. This ontology controls the adjunction and substitution operations between the semantic categories.
On the contrary to classical LTAG, the semantic basis for post-parsing processing is here the derived tree and not the derivation tree. For complex cases, semantic features may control the derivation with specific mecanisms as suggested in (Roussel, 1999).

5. Related works and conclusion

Previous works have shown that focusing parsing first on semantics can lead to superior speed efficiency than syntax-first approach, particularly on restricted domain as shown in (Lytinen, 1991), but also for large coverage grammar (Dowding *et al.*, 1994). The trees currently developped for our application and their lexicalization are closed from the semantic grammars paradigm (Seneff, 1992) and works on terminological variability (Jacquemin, 1999). We expect that such a LTAG grammar will allow, in our application, a stronger and an easier integration of different level of constraints. In terms of reusability, the same linguistic representation (the predicative LTAG grammar) could be mapped into concepts of various restricted domains with a domain-dependent semantic module.

References

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

ABEILLÉ A. (1992). Synchronous TAGs and French Pronominal Clitics. In *COLING*, Nantes, France.

DOWDING J., MOORE R., ANDRY F. & MORAN D. (1994). Interleaving syntax and semantics in an efficient bottom-up parser. In *ACL'94.*

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *ACL'99, University of Maryland.*

LYTINEN S. (1991). Semantic-first natural language processing. In *AAAI'91, Anaheim, CA.*

ROUSSEL D. (1999). *Intégration de prédictions linguistiques issues d'applications a partir d'une grammaire d'arbres hors contexte. Contribution a l'analyse de la parole.* PhD thesis, Joseph Fourier University, Grenoble.

SENEFF S. (1992). Tina: A natural language system for spoken language applications. *Computational Linguistics*, **18** (1 p.), 61–86.