A Deep-Parsing Approach to Natural Language Understanding in Dialogue System: Results of a Corpus-Based Evaluation

Alexandre Denis, Matthieu Quignard, Guillaume Pitel

UMR 7503 LORIA (Université Henri Poincaré, CNRS, INRIA)
BP 239 F-54506 Vandoeuvre-lès-Nancy
Alexandre.Denis@loria.fr

Abstract

This paper presents an approach to dialogue understanding based on a deep parsing and rule-based semantic analysis. Its performance in the semantic evaluation performed in the framework of the EVALDA/MEDIA campaign is encouraging. The MEDIA project aims to evaluate natural language understanding systems for French on a hotel reservation task (Devillers et al., 2004). For the evaluation, five participating teams had to produce an annotated version of the input utterances in compliance with a commonly agreed format (the MEDIA formalism). An approach based on symbolic processing was not straightforward given the conditions of the evaluation but we achieved a score close to that of statistical systems, without needing an annotated corpus. Despite the architecture has been designed for this campaign, exclusively dedicated to spoken dialogue understanding, we believe that our approach based on a LTAG parser and two ontologies can be used in real dialogue systems, providing quite robust speech understanding and facilities for interfacing with a dialogue manager and the application itself.

1. Introduction

This paper presents a symbolic-oriented system and its evaluation in the framework of the EVALDA/MEDIA campaign. The MEDIA project aims to evaluate natural language understanding systems for French on a hotel reservation task (Devillers *et al.*, 2004). For the evaluation, five participating teams had to produce an annotated version of the input utterances in compliance with a commonly elaborated format (the MEDIA formalism). Our approach can be summarized as follows:

- a deep LTAG parser is used to produce a syntactic analysis,
- a compositional semantic builder à la Montague produces a conceptual graph from the syntactic analysis, and
- a projection module flattens the graph and constructs the target representation format.

What is worth taking note of is that most of the characteristics of the MEDIA evaluation make it more suitable for statistical approaches, particularly since there was almost no adaptation required for output of a statistical annotation. Given these conditions, the good performance of our system was a surprise.

2. Task

In the EVALDA/MEDIA project, two aspects of understanding are evaluated, a context independent semantic annotation and a context dependant one. The context independent semantic evaluation considers each utterance independently. The context dependant one takes anaphora and sense specification into account. Only the first level will be presented in this paper since the second aspect has not been evaluated yet.

In order to measure the performance of the systems, a common output format has been proposed for the semantic annotation, and all systems are expected to produce annotations within this format.

In a first phase, a separate team of annotators, using the Semantizer tool (Bonneau-Maynard *et al.*, 2005), produced a manual annotation in this format of a finalized dialogue corpus. Participants and annotators collectively agreed on a guide for annotation while the first phase was running and problems were arising.

In the evaluation phase, participants ran their systems on the raw data after having trained their system on a subpart of annotated data.

2.1. Target Format

In the MEDIA representation format shared by all participants, each utterance is segmented into different meaningful chunks and each chunk is associated with a single semantic feature. The features could have two forms depending on the segment: a triplet <*mode*, *attribute*, *value*> if the segment has a meaning in the task, or by convention `<+, null>` if it has not[1]. What is important to notice is that each chunk is annotated with only one feature, which is an important constraint.

The *mode* element describes the modality of the chunk: positive (+), negative (-), interrogative (?) or optional (~).

The *attribute* element is defined by the semantic category of the information conveyed by the chunk. It is composed of two parts: a primitive attribute and a list of specifiers which refines its sense. For example, the chunk "*two rooms*" will be annotated by `<+, number-room, 2>` where `number` is the primitive attribute which is specified by `room`.

Finally, the *value* element is either a string, an integer or a constrained value in a list associated with the *attribute*.

For instance (* indicates plural definite determinant) :

1. **"est-ce qu' i(l) y a un parking privé"**
   is it that there-is a parking-lot private
   Is there / a private car park?

[1] The semantic features are in fact 5-tuples, but we do not present here the reference and dialogue act elements since they are not evaluated in the context-independent phase, see (Bonneau-Maynard *et al.*, 2006)

```
<+, null>
<?, hotel-carpark, private>
```

2. **"vous avez les concernant les chambres…**
you have the about the* rooms*
Do you about / the / rooms
```
<+, null>
<+, refLink-coRef, plural>
<+, object, room>
```

est-ce que vous avez de disponibles douze chambres simples ou non…
is it that you have of available twelve rooms simple or not
do you have / twelve available / single rooms / or not
```
<+, null>
<?, number-room-available, 12>
<?, room-type, single>
<+, null>
```

pour les trois premiers jours de mai"
for the three first days of May*
for the three / first / days / of May.
```
<+, number-time, 3>
<+, time-timeAxis, first>
<+, time-unit, day>
<+, time-month, 5>
```

The segmentation and the annotation itself (that is: how do we annotate what?) is described in a manual, jointly written by the participants.

2.2. Semantic dictionary

The set of allowed triplets (the set of attributes, their allowed specifiers and values) are defined in a semantic dictionary and can be classified into two subsets: the subset of task specific attributes or specifiers (for example `reservation`, `room`, or `hotel`), and the subset of generic reusable ones. The generic ones describe either raw semantic categories (like `number` or `name`), logical connectors (like `connectProp`), dialog oriented categories (like `response` or `dialog-command`) or referential annotation categories (like `refLink`). See (Bonneau-Maynard *et al*., 2005).

2.3. Corpus elaboration

The corpus which is composed of 1,257 transcribed spoken dialogues in French (18,801 client utterances) recorded using the Wizard of Oz technique, has been manually annotated[2] using this formalism by ELDA. The dialogues follow scenarios in the hotel reservation task with varying complexities taking into account simulated misunderstandings or speech recognition errors. As usual in evaluation, a first subset of these dialogues (12,292 utterances) has been used for systems training and a second subset has been used for the test itself.

[2] Annotators could also refer to the speech recordings, for example to better annotate the interrogative modes. Such prosodic information was not available to systems.

3. System

As stated above, our system is based on a deep parsing, rule-based method of analysis. The use of deep-parsing in a context such as the MEDIA project may seem problematic at least for two reasons. The first one is that the variety of speech phenomena (hesitations, reformulations, interpolated clauses, etc.) and the ambiguities of large-scale grammars would dramatically affect the efficiency of the parser, usually designed for well-written sentences.

Secondly, the annotation framework requires the resulting derivation trees to be deeply processed to match the expected ontology and the strict linear alignment of semantic segments on the utterance itself. Therefore, statistical approaches would probably be preferred since many of these specificities can be learned, though a large amount of annotated dialogues would be needed.

Despite the above difficulties, we based our approached on a Lexicalized Tree Adjoining Grammar (LTAG) parsing for syntax (Joshi & Schabes, 1992; Crabbé, Gaiffe & Roussalany, 2003), and a description logic representation for semantics. A deep syntactical representation is important for semantic representation and also for anaphora resolution (which is the second stage of the MEDIA evaluation campaign). Our approach faces several difficulties, but does not require any training on an annotated corpus.

3.1. System Components

As the system is evaluated on the basis of its semantic understanding capabilities, we focused on three modules: a *parser* and a *semantic builder* that construct an internal representation of an utterance, and a *projection module* that translates this representation into the MEDIA format (fig. 1).

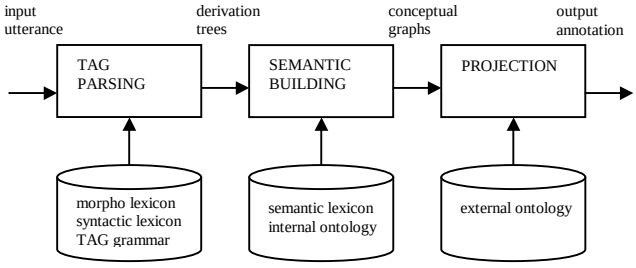

Figure 1. Overall architecture

3.1.1. LTAG Chart Parser

The parser (Lopez, 1999) performs deep syntactical analysis. We decided to work on partial derivations in order to be more robust for speech, since it is more flexible on grammaticality than written language. All the partial derivations resulting from the analysis have been gathered and sorted according to the length of their coverage of the input utterance. Only the longest non-overlapping partial derivations are kept for semantic construction, the other ones are disregarded. Indeed, our goal was not to obtain complex analyses that would fully cover the utterance, but instead, as many subtrees as possible, with the wider coverage. To some extent, this approach is close to the spirit of chunk parsing, excepted that we are still willing to obtain, when possible, a unique deep analysis for the whole sentence. The idea is to

produce small semantic sub-graphs from a set of partial analyses, and then potentially reconnect them using ontological information.

The parser is based on three types of grammatical resources in the XML-compliant TAGML format (Tree Adjoining Grammars Markup Language): a morphological lexicon, a syntactical lexicon and a tree library. The morphological lexicon (5,400 words and 3,000 lemmas) has been extracted from of the Multext lexicon (Armstrong, 1996) and manually revised to avoid lexical ambiguity. The syntactical lexicon anchors the trees with simple manually designed heuristics like all the nouns anchor the noun trees. The grammar has also been written by hand, and tries to find a balance between two opposing principles:

- a *reusability principle*: the grammar should not be specific to the hotel reservation task and should take into account general linguistic phenomena
- a *minimal-size principle*: the grammar should be small to avoid combinatorial explosion.

Therefore the grammar contains only 80 TAG trees while standard TAG grammar may contain thousands of trees. It covers only a small fragment of the French language but it is sufficient for our needs. An average of 63% of the trees concern noun phrases parsing and 37% concern verb phrases parsing.

The semantic minimality principle of TAG grammar asserts that each tree of the grammar should represent one semantic predicate. In our context it means that each tree should be associated with a semantic feature in the MEDIA formalism. However the grammar would then be tied to the hotel reservation task, and would not fit our reusability principle. Therefore the semantic minimality cannot be directly applied to the MEDIA predicates. Instead each tree denotes a predicate in an internal ontology in accordance with the semantic minimality principle and each predicate in the internal ontology is mapped onto an external ontology.

3.1.2. Semantic Builder

The semantic builder constructs from a partial derivation a conceptual graph in the MultiModal Interface Language (MMIL), (Landragin *et al.*, 2004). This language was primarily designed to represent multimodal events in a dialogue system (linguistic, gestural or haptic) but we only focused here on its linguistic/semantic facet. It allows an utterance to be represented as a conceptual graph of entities (events and participants) each described by a feature structure. MMIL provides a library of generic features both linguistic (gender, number...) and semantic (objType, evtType, modifier...). They are used in conjunction with task-dependant features (relative to the hotel reservation task).

The semantic resource used here is a semantic lexicon composed of 150 schemes defined as couples of a grammatical anchor (a lemma plus a tree) and a semantic typed fragment. The fragments of the conceptual graph could be an entity, an attribute or a relation. They are related to an OWL ontology (called *internal ontology*): the entities and attribute values are typed as concepts, the attribute names and relations are typed as roles. The internal ontology is composed of 220 concepts related to the hotel reservation task.

The construction algorithm is based on derivation trees as it indicates the dependencies between semantic fragments. The conceptual graph is built iteratively by passing through the derivation tree: each derivation node is passed through and combined with its children according to the TAG operation (substitution or adjunction) and to the semantic type.

With the ontology it is then possible to refine or to correct an erroneous conceptual graph whose faults come from an incorrect parsing. Two transformations can be done: a completion of missing relationships, or an elimination of incorrect ones. As the completion was not crucial for the evaluation we just eliminated inconsistent relationships and allowed non-connected graphs in this case.

For example in "*I reserve a room for the 10th January*", any potential relationship between the entities "*the 10th January*" and "*a room*" would be dismissed as it is not consistent relative to the internal ontology. Rather, our ontology gives a link between "*the 10th January*" and the reservation event.

3.1.3. Projection Module

The projection transforms the conceptual graph into the semantic annotation format using description logics.

The projection is done by mapping the internal ontology (the concepts resulting from the semantic builder) to an external ontology (the concepts which have to be produced for the evaluation) close to the MEDIA semantic dictionary. All the concepts of the external ontology are defined in the terms of the concepts of the internal ontology. We divided the 130 concepts of the external ontology into three namespaces corresponding to different types of information in the triplets : a namespace for the primitive attributes (`base#`), one for the specifiers (`spec#`) and one for the mode (`mode#`) (see 2.1). For example, the MEDIA attribute `name-hotel` is defined to be the `Name` of an `Hotel`.

The conceptual graph is first translated as a set of instances and roles assertions of the *internal* ontology in description logics ALCHI-D. Then each instance is scanned to retrieve its most specific instantiator concepts in the *external* ontology using the RACER inference engine (Haarslev and Möller, 2003). Any `Name` which is the name of an `Hotel` is also a `name-hotel` and will be taken for the annotation.

These concepts are associated directly with a MEDIA feature : the mode is defined by the concept found in the `mode#` namespace, the attribute name is defined by the concepts found in the namespaces `base#` and `spec#`. The value is found according to association rules between the concepts in the external ontology and a specific way to compute the value: for example, if the concept `name-hotel` is found, then the value of the feature can be retrieved by extracting the filler of the RACER attribute `name`.

Finally, since the MEDIA formalism is flat and ordered on the chunks of the utterance, we need to produce the MEDIA triplets in the correct order. This is done by keeping the index of the words from the parsing analysis through the semantic construction. Then we can retrieve the position of the found features because they are related to a particular instance which comes from an indexed semantic element.

3.2. Processing example

For example, the utterance "*I want a room between Paris and Versailles*" would be analyzed into the MMIL graph illustrated figure 2.

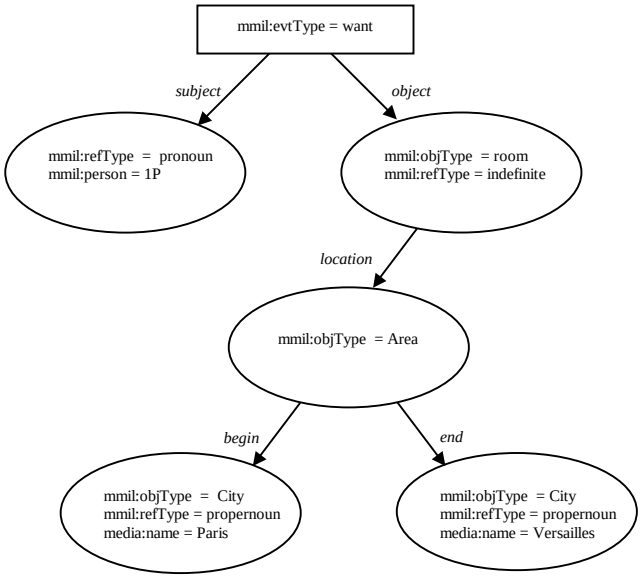

Figure 2. The MMIL graph representing the utterance "*I want a room between Paris and Versailles*"

The translation in RACER would be the following list of ABox assertions:

```
(instance c1 City)
(instance r1 ProperNoun)
(related c1 r1 refType)
(constrained c1 c1-name name)
(constraints (string= c1-name "Paris"))

(instance a1 Area)
(related a1 c1 begin)
...
```

We assume we have the following declarations in the external ontology (toprole is a symmetric role, the highest in the role hierarchy) :

```
City ⊑ base#location-city
∃begin-1.Area ⊑ spec#begin
∃toprole.∃toprole.Room ⊑ spec#hotel
```

And the following value association rule:

```
base#location-city ~ name
```

The most specific instantiators of each instance (`c1`, `r1`, `a1`...) in the base and spec namespaces are retrieved. For instance `c1`, it would then retrieve:

```
{base#location-city, spec#begin, spec#hotel}
```

The two concepts are combined (using specifiers rules given by the MEDIA semantic dictionary) into `location-city-hotel-begin`. Then the value is retrieved thanks to the value association rule which extracts the attribute `name` of the instance `c1` (Paris). Since the default mode is positive, we finally obtain the following triplet (for `c1`):

```
<+, location-city-hotel-begin, Paris>
```

4. Evaluation

Like every other participant in this campaign, we had to conform to a common evaluation protocol, based on a flat semantic representation (the MEDIA formalism), an annotated corpus of transcribed utterances, and some evaluation measures (Devillers *et al.*, 2004; Bonneau-Maynard *et al.*, 2005). This evaluation protocol is described below.

4.1. Evaluation and Comparison Methodology

Each system has been evaluated for their context independant semantic understanding on 3,003 transcribed utterances of the corpus, and the distance between their production and the manually annotated fragment has been quantified using precision/recall and Levenshtein distance. Several fuzzy measures have been proposed, taking or not the *value* field of the triplet into account, or allowing taking less specific *attribute* fields, for instance.

In this paper, three methods are presented, all based on the Levenshtein distance, which minimizes the distance between two *ordered* list of triplets, one given by our system, the other by annotators.

- *Precision* gives the ratio of correct triplets (at the right place) over the number of triplets given by our system. This measures the fiability of our understanding;
- *Recall* gives the ratio of correct triplets over the number of triplets given by annotators.
- The *MEDIA error rate* gives the ratio of all errors (missing, added or substituted triplets) over the number of triplets given by annotators. This ratio may be greater than 1.

4.2. Results

The results are presented at two different times : the 29th April 2005 preliminary test (1,005 utterances) and the 23th June final test (3,003 utterances). Table 1 shows the total number of triplets in the test corpus, the number of the triplets found by our system, and the number of correct found triplets. The precision, recall and the MEDIA error rates are indicated.

	preliminary test	**final test**
Total triplets	3125	8383
Found triplets	2630	8517
Correct triplets	1291	6098
Precision	49.1%	71.6%
Recall	41.3%	72.7%
MEDIA error rate	0.695	0.289

Table 1. Results for tests

The results are satisfactory: using the simplified evaluation measure[3], the error rate is 0.289, which ranked us third. Compared to the best systems scoring .232 and 0.238 with a statistical approach, the difference is significant but far from outstanding, especially if we take into account the time we spent for the design of the system and the necessity to comply to the MEDIA flat formalism. As a reference, the inter-annotator agreement is 0.891 (Kappa measure).

[3] The simplified-mode evaluation measure does not consider *interrogative* or *optional* mode of the annotation triplets. Such modes are interpreted as *positive* ones. This simplification is due to the fact that systems could not access to prosodic information (see note 2) which enables annotators to discriminate interrogative chunks in affirmative utterances.

The difference between the preliminary test and the final test scores (only two months) shows how adding resources may improve the performance of our system. We produced much more triplets than in the preliminary test (by inferring more information during the projection process) and even we produce a little too much triplets than annotators, we reach a good precision score, quite close to the recall score.

4.3. Discussion

4.3.1. Concerning results

Although the results are good, they should not be abstracted from the evaluation context. Since the targeted annotation scheme only takes into account local specifications, we did not focus our efforts on establishing correct dependencies between the chunks, but rather on the chunks annotation itself (63% of the grammar covers noun phrases). Therefore our best results come from the resolution of noun phrases. Thus the main errors relates to propositional coordinations represented in the MEDIA formalism as `<+, connectProp, entail>` or `<+, connectProp, explain>`.

The score we got for the preliminary test was very low and we gained an average of 0.3 points within two months. This remarkable improvement is due to the added resources, and we suppose the system could do even better if we add new resources such as finer annotation rules for analyzing logical connectors.

4.3.2. Concerning the evaluation

We were able to satisfy successfully the two main constraints of the annotation format : one single feature for a meaning chunk, and an ordered list of features for an utterance. Even if our internal representation as conceptual graphs was far from the expected output, we adapted the system in compliance with the annotation guidelines. We are nevertheless very dependant on the coherence between the corpus and the manual itself. The inter-annotator agreement indicates the correlation between the annotators, but not the correlation between the annotators and the manual.

The corpus-based evaluation has some advantages and shortcomings. The interesting point is the readiness of the evaluation: the production of each system is compared to a human annotation using a flat semantic formalism. It is then very easy to provide some quantitative measures on the distance between the human and the machine annotations. But we do not decorrelate the ability to produce a correct inner semantic representation from the ability to translate this semantic form into the evaluation formalism. The system whose internal representation is closer to the evaluation formalism is then at an advantage.

This quantitative technique is opposed to qualitative evaluations of the understanding abilities of a dialog system. (Antoine et al., 2000) Specific phenomena are tested with a client utterance (called *the declaration*), another utterance (called *the control*) which supervises the particular phenomenon evaluated in the declaration and a boolean value (called *the reference*) which accounts for the coherence of the two previous utterances. The systems are evaluated on their ability to find whether the declaration and the control are semantically compatible. This technique is more difficult to use because it requires hand-designed tests and it supposes that each system is able to compare two utterances. On the other hand it is independent of the internal formalism and therefore does not evaluate the ability to project.

4.3.3. Concerning the system

The two main flaws of our system are its slowness and the need of human craft for the resources. In fact, the runtime of the system is average 5 seconds for annotating one utterance and 1.6 second for producing one triplet. It is much slower than statistical approaches. The system also needs a well-skilled person for creating the resources. Nevertheless, the main advantage of our system is that the approach does not require any training on a costly annotated corpus. It only needs well specified annotation guidelines for the resources creation. Moreover, this approach can handle any non-frequent phenomenon, even those that are not present in the training corpus.

An important feature of our system is fully based on symbolic resources and processes which are understandable by human beings. This is not always the case for statistical systems. For example, HMM models cannot be easily debugged. In our symbolic approach the origin of problems in the process can be found more easily. This ability is also very useful for real dialog systems when they need to understand what is wrong in their interpretation process: knowing the source of the error is in fact necessary to conduct a correction subdialogue.

The resources of our system are sufficient for the evaluated phenomena. However some improvements could be done: particularly a better grammar for questions, relatives and complex propositions. The current resources are not exclusively tied to the hotel reservation task, and then could be reused for another one. We estimate that the resources, grammar and ontologies will be a starting point for considering new applications. For example, embedding our system into a dialogue system for interfacing with an application would not expect to fully redevelop the grammar, but rather design the internal ontology to cover the task domain then specify the semantic building rules for mapping internal concepts on TAG grammar. In a second stage, one would then develop an external ontology to model the application world and elaborate the rules for translating an internal representation into the application language. Having an internal ontology may divide the complexity of the resources development (between application and linguistic grammar) and enable the dialogue system to manage linguistic or dialogic misunderstanding, what the application is not designed for.

5. Conclusion

We have presented a system which has been evaluated on a semantic annotation task. The task consisted of automatically annotating an oral speech corpus with a flat semantic formalism. It is the first time that our system is evaluated and it ranks third. Its error rate is 0.289 close that of statistically trained system (0.232 and 0.238). Our system relies on purely symbolic hand-written resources: a TAG grammar for deep-parsing and an OWL ontology for description logics.

Its main advantage is that it does not require an annotated corpora but well specified guidelines for

creating the resources. Through this evaluation we have proved that using a partial deep-parsing was feasible on spoken language. It also validated our MMIL conceptual graph which gave us total satisfaction for semantic representation. We hope that its usefulness will be demonstrated for anaphora resolution in the next evaluation phase of the MEDIA campaign.

6. References

Antoine, J.-Y., Siroux, J., Caelen, J., Villaneau, J., Goulian, J., Ahafhaf, M. (2000). Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm. In *Proceedings of LREC 2000*, Athena.

Armstrong, S. (1996). "Multext: Multilingual Text Tools and Corpora", in *Lexikon und Text*, H. Feldweg ; W. Hinrichs (ed.), Tübingen: Niemeyer, pp. 107–119.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A. and Mostefa, D. (2005). Semantic Annotation of the French Media Dialog Corpus. In *Proceedings of InterSpeech*, Lisbon, September 2005.

Bonneau-Maynard, H., Ayache, C., Béchet, F., Denis, A., Khun, A., Lefevre, F., Mostefa, D., Quignard, M., Rosset, S., Servan, C., Villaneau, J. (2006). Results of the French Evalda-Media evaluation campaign for literal understanding. In *Proceedings of LREC 2006*, Genoa, Italy.

Crabbé, B., Gaiffe, B.**,** and Roussalany, A. (2003). Une plateforme de conception et d'exploitation de grammaire d'arbres adjoints lexicalisés. In *Actes de TALN 2003*, Batz-sur-mer.

Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Bousquet, C., Charnay, L., Vigouroux, N., Béchet, F., Romary, L., Antoine, J.-Y., Villaneau, J., Vergnes, M., and Goulian, J. (2004). The French MEDIA/EVALDA Project : the Evaluation of the Understanding Capability of Spoken Language Dialog System. In *Proceedings of LREC 2004*, Lisbon, Portugal.

Landragin, F., Denis, A. Ricci, A. and Romary, L. (2004). Multimodal Meaning Representation for Generic Dialogue Systems Architectures. In *Proceedings of LREC 2004*, Lisbon, Portugal.

Lopez, P. (1999). *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres*. PhD thesis. Université de Nancy 1

Joshi, A. K. and Schabes Y. (1992). Tree adjoining grammars and lexicalized grammars. In M. Nivat & A. Podelsky, Eds. *Tree Automata and Languages*. Amsterdam, The Netherlands: Elsevier.

Haarslev, V. and Möller, R. (2003). Racer: A Core Inference Engine for the Semantic Web. In *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools* (EON2003), located at the 2nd International Semantic Web Conference ISWC 2003, Sanibel Island, Florida, USA, October 20, 2003, pp. 27-36.