Automatic Rich Annotation of Large Corpus of Conversational transcribed speech: the Chunking Task of the EPAC Project

Jean-Yves Antoine, Abdenour Mokrane, Nathalie Friburger

Université François Rabelais Tours
Laboratoire d'Informatique (LI) – Equipe BdTln
3 place Jean Jaurès, 41000 Blois, France
E-mail: {jean-yves.antoine, abdenour.mokrane, nathalie.friburger}@univ-tours.fr

Abstract

This paper describes the use of the CasSys platform in order to achieve the chunking of conversational speech transcripts by means of cascades of Unitex transducers. Our system is involved in the EPAC project of the French National agency of Research (ANR). The aim of this project is to develop robust methods for the annotation of audio/multimedia document collections which contains conversational speech sequences such as TV or radio programs. At first, this paper presents the EPAC project and the adaptation of a former chunking system (Romus) which was developed in the restricted framework of dedicated spoken man-machine dialogue. Then, it describes the problems that are arising due to 1) spontaneous speech disfluencies and 2) errors for the previous stages of processing (automatic speech recognition and POS tagging).

1. Introduction: the EPAC project

With the development of Internet, phone networks, or broadcast media, the amount of digital documents that are accessible to the public follows an impressive increase years after years. As a result, there is a real need for the public but also for professionals to benefit from an efficient access to such multimedia document collections. This requires the achievement of intelligent systems of information retrieval like, for instance, interactive question/answering systems with deep natural language processing. But at first, this huge mass of raw data needs an appropriate indexing to be interrogated by information retrieval tools. Considering their exponential development, these resources can only be indexed by automatic techniques, possibly complemented with human supervision.

The aim of the EPAC project is to develop robust methods for information extraction and indexing of audio/multimedia document collections which contains conversational speech sequences such as TV or radio programs. The project will be carried out on a large database of audio documents (1800 hours of recorded data) which comes mainly from the ESTER corpus of radio programs. The project is founded by the French National agency of Research (ANR) and involves four French laboratories (LIUM, LIA, IRIT, LI). It concerns a wide range of tasks which are necessary to structure and index the corresponding resources. This paper focuses on the question of rich annotation by means of natural language processing (NLP) tools.

2. Rich annotation of conversational speech transcripts

For information retrieval purposes, speech transcripts are only sufficient to achieve a basic "Google-like" search. Indeed, advanced question/answering systems works on objects whose characterization requires a linguistic analysis of the considered documents (and of course of the user request). In particular, robust detection and characterization of named entities (NE) is a crucial issue for researches in information retrieval and automatic indexing of language resources. Consider for instance the following word sequences:

- *John Huston*
- *The director of "Moby Dick" and "The Maltese Falcon"*
- *Angelica's father*
- *This American actor and director, was born in august 2006 […] He died in 1987.*

All of these named entities are referring to the same person. The detection of these co-referential relations requires a good ontology or world knowledge description. But these resources would be useless without previous sophisticated syntactic and contextual processing: syntactic tagging and segmentation, compound noun detection, named-entity recognition, anaphoric co-reference resolution, etc.

One aim of the EPAC project is to provide NLP tools which are the more frequently involved in information retrieval systems: part-of-speech tagging, segmentation into minimal syntactic units (chunks), named-entities detection (see figure 1). Existing tools are usually developed for written texts. The scientific challenge of this work is to adapt these techniques to conversational speech. In addition with the problem of speech recognition errors, the spontaneous nature of conversational speech results in a high rate of spoken disfluencies such as hesitations, self-repairs or false starts (Shriberg 1994). These ungrammatical constructions disturb strongly the application of NLP techniques, and

original approaches must be investigated to obtain a satisfactory robustness.

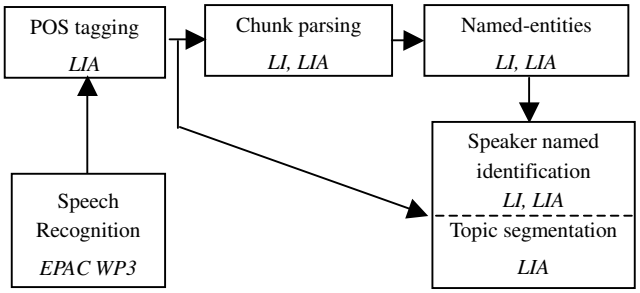

Figure 1 : Rich annotation tasks in the EPAC project

The tools we are currently developing will be freely available at the end of the project. Free annotated corpuses will also be provided. In order to guarantee their reusability, we take a particular care to the current normalization of our corpus annotations (cf. section 4.2).
Generally speaking, part-of-speech tagging constitutes the first stage of sentence syntactic analysis (see figure 1). In the EPAC project, the LIA laboratory is in charge of the achievement of this processing stage, using an adaptation of the platform LIA_TAGG available under a GPL licence[1]. We will now describe the chunking level of conversational speech.

3. Chunk parsing of conversational speech

Chunk parsing is designed to provide a bracketing of a text into minimal non-recursive phrases (Abney 1991, Church 1988). Generally speaking, a chunk is composed of a lexical head (noun, verb, adverb, adjective or a preposition) and its local dependent terms. Consider the sentence you are reading at this moment. It can break up something like this:

(1) $[Consider]_{VC}$ $[the\ sentence]_{NC}$ $[you]_{NC}$ $[are\ reading]_{VC}$ $[at\ this\ moment]_{PC}$

where VC, NC and PC are corresponding to verbal, noun and prepositional chunks respectively, depending on their phrasal head.
For the purpose of conversational speech information retrieval, chunks presents two main interests:

- Chunks identify most of the time a semantic unit which can be related to the world knowledge. As asserted by (Abney 1991) "*when I read a sentence, I read it a chunk at a time*". In particular, named-entities such as "*this American director*" or "*Angelica's father*" correspond each to a chunk.
- Corpus studies on spoken French (Blanche-Benveniste 1990) have shown that chunk is the longer syntactic unit that remains preserved (to a certain extent) by speech repairs and other disfluencies. Chunk segmentation is therefore well fitted as a first parsing step of conversational speech.

Many chunk analyzers that can be found in the literature achieve a robust shallow parsing of large domain written documents or even prepared speech (broadcast news, for instance). On the contrary, works on spontaneous speech usually focus on task-oriented man-machine dialog where the considered vocabulary is restricted (1.000 to 10.000 words). In particular, we have developed in the last years two speech understanding systems (*LOGUS* and *ROMUS*), which were dedicated to tourism information systems. Both systems (Villaneau, Antoine 2004, Goulian and Antoine 2003) involve an incremental strategy to achieve a robust understanding of conversational speech:

1. automatic speech recognition,
2. part-of-speech tagging of the recognised sentence,
3. chunk parsing of the POS sequence,
4. identification of semantic relations between chunk phrasal heads,
5. finally, contextual understanding (resolution of anaphoric co-references)

If part-of-speech tagging is directly affected by the size of the considered lexicon, chunk parsing works on the contrary on a close set of part-of-speech tags. One should then expect that it can be extended to larger domains than dedicated man-machine dialog. This is why one aim of the EPAC project is to generalise our previous works to large vocabulary chunking of conversational speech. The SECARE system which is developed in the EPAC project adopt the principles of the robust parsing and it is not limited to the man-machine dialogue. SECARE is adapted to the general language.

4. Chunking in the EPAC project

4.1 Transducer cascades: CasSys/Unitex

The chunking level in the EPAC project is based on a cascade of finite state transducers (FST). FST cascade is based on a simple idea: apply transducers on the text in a precise order to transform the text or extract patterns from the text. A unique transducer does not aim to cover complete linguistic phenomena but every transducer participates in the coverage of a part of the considered linguistic phenomenon. The recognition of simple patterns reduces the research space. Transducers parse the text in a precise order to first track down the most certain patterns which Abney named "islands of certainties". The uncertain patterns are found next. Every transducer uses the results of the previous ones. As shown by many works on incremental shallow parsing (Abney 1996, Ait-Mokthar and Chanod 1997), parsers using cascades of transducers take advantage of three qualities: robustness, precision and speed brought by transducers.

[1] LIA_TAGG webpage : www.lia.univ-avignon.fr/chercheurs/bechet/download_fred.html

In the EPAC project, the chunk parser uses the UniteX toolkit[2] supplemented by the CasSys extension (Paumier 2003, Friburger and Maurel 2004) that was developed in the LI laboratory. CasSys is a tool for implementing FST cascades using the FST Toolbox of the Unitex system. Unitex represents a transducer as a graph (see figure 2). It is completely generic, what means that it can handle various kind information (phonetic, morpho-syntactic or syntactic knowledge if you consider for instance NLP applications) and may be used for different purposes (syntactic parsing, named entities detection, information extraction a.s.o.). The output of the transducers can be added to the pattern found in the text, or the recognized patterns can be extracted from the texts and replaced by a label. The labels of already found patterns can be used in the transducers that follow the current one in the cascade.
In addition to the Unitex system, CasSys allows to extract patterns which can be enriched by transducers. According to the design methodology in a transducer cascade, CasSys recognizes first the least ambiguous patterns, which are then removed in order to avoid confusions with the patterns recognized by the following transducer. In case of residual ambiguities, CasSys always favours the analysis that leads to the longest sequence of patterns.

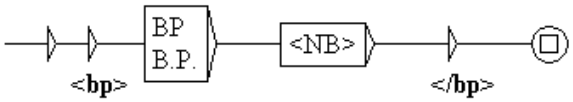

Figure 2: Example of a Unitex Transducer

4.2 Data formats: PEAS segmentation

One aim of the EPAC project is to provide a corpus of 200 hours of transcribed conversational speech. In order to allow an optimal reusability of this resource, we take care of the normalisation of the data formats. All produced data will be encoded in XML. The transcriptions are in the *.trs* Transcriber format (Barras *and al.* 1998) which is widely used in the speech community. In order to add independent level annotations on the transcriptions (POS, chunks, named entities, dialog act a.s.o.), we defined a temporal reference of synchronization which is based on a segmentation of the transcriptions into tokens. This tokenisation takes again the format used in the LUNA[3] European project.

```
<Turn startTime="0" endTime="2.933" speaker="spk1"
mode="spontaneous" fidelity="high"channel="studio">
<sentence id="s0001">
<text> good .</text>
<tokens count="5">
<token id="s0001_t0001" type="sgmltag">
<Sync time="0" />
</token>
<token type="space" id="s0001_t0002" />
<token type="wtoken" id="s0001_t0003">bien</token>
<token type="space" id="s0001_t0004" />
<token type="poncts" id="s0001_t0005">.</token>
</tokens>
</sentence>
<sentence id="s0002">
```

Figure 3 : Example of a tokenized corpus

The figure 2 shows an extract of a tokenized corpus. One can see that a token may correspond to a single word, but also a punctuation sign or even a XML tag. Then, any additional annotation refers to the corresponding tokenized file and not to the original transcription.
The chunks definition is based on the PEAS annotation scheme which was adopted during the French speaking EASy evaluation campaign (Vilnat *et* al. 2003, Paroubek *et* al 2006). The PEAS paradigm includes the following category of chunks:

- Nominal (chunk GN).
- Prepositional (chunk GP).
- Verbal (chunk NV).
- Verbal introduced by a preposition (chunk PV).
- Adjectival (chunk GA).
- Adverbial (chunk GR).

For the purpose of the EPAC project, two kinds of additions to the PEAS format have been proposed:

- specific categories for oral disfluencies, as defined by (Shriberg 1994): REP (reparandum) and ED (edition zone),
- specific categories in order to reach a complete segmentation of the speech transcripts. For instance, we add a chunk COO for the representation of the conjunction of coordination while the latter are not segmented in the PEAS annotation scheme.

Figures 4 and 5 present as an illustration a verbal chunk related to the word sequence « *it describes* ». This annotation refers to the tokens reference, as shown by the XML attributes `token_beg` and `token-end`.

```
<chunk id="s0034_c" token_beg="s0034_t0005"
word_beg="s0034_w0001" token_end="s0034_t0007"
word_end="s0034_w0002" > NV </chunk>
```

Figure 4: Example of chunk annotation related to the "*it describes*" word sequence (NV: verbal chunk)

```
<word id="s0034_w0001" token="s0034_t0005"
pos= "PPERS3s "> it </word>
<word id="s0034_w0002" token="s0034_t0007"
pos="V3Ps"> describes </word>
```

Figure 5: sequence of tokens (synchronization reference) corresponding to the example on figure 4.

4.3 Implementation on CasSys/Unitex

As describe above, we based our chunking stage of analysis on a transducers cascade. More precisely, the identification of chunks is based on two cascades. The first ones identifies all regular chunks, e.g. which are not corrupted. As shown in the result section, the segments that are not identified at the end of the first cascade do not correspond in all cases to disfluencies, but it can result

[2] http://www-igm.univ-mlv.fr/~unitex/manuel.html

[3] LUNA project : http://www.ist-luna.eu/

from errors of the POS tagging or automatic speech recognition components. These chunks will be used as island of certainty for the second cascade, which is in charge of achieving a complete segmentation of the corpus. For the moment being, only the first cascade is completely implemented. It is able to identify all of the PEAS regular chunks. The second cascade is limited for the moment being to the characterization of the complementary kinds of chunks we added to the PEAS annotation scheme (COO, REP, ED, COO, a.s.o.). It is also used for the correction of some POS tagging errors (see section 5). Any sequence of tokens which is not identified by the first cascade is then labelled by a specific CHUNK tag (*UNKnown CHunk*). Figure 6 shows an example of CHUNK which results from an error of POS tagging. Indeed, on this example, the LIA_TAGG has recognized the first name "*Pierre*" but not the last name *Péan*", what prevents the correct identification of the resulting named entity.

```
<word id="s0002_w0003" token="s0002_t0005" pos="XPREM"> Pierre </word>
<word id="s0002_w0004" token="s0002_t0007" pos="UNKWORD"> Péan </word>
<chunk token_beg="s0002_t0005" word_beg="s0002_w0003" token_end="s0002_t0007" word_end="s0002_w0004" id="s0002_c"> CHUNK</chunk>
```

Figure 6: Example CHUNK (pos tagging error)

In practice, a series of transducers is associated to every chunk category. More precisely, we have defined for every category a main transducer and a series of secondary ones. The main transducer describes the regular syntactic structure of the corresponding kind of chunk. For instance, the figure 7 presents the main transducer GN, which describe nominal chunks. On the CasSys cascade, every box is corresponding to a POS tag (or an already characterized chunk sequence) that is supposed to occur in the nominal chunk, while the relations between these boxes model word/POS/chunk successions in the chunk. Secondary transducers have been defined to manage the XML data related to the annotation scheme of the EPAC project (transcripts, tokens, POS for the input and chunk for the ouput). As a result, the main transducers are completely generic and should be re-employed without any modification on resources that follow a different annotation format. We have also re-examined some Unitex pred-definite automata in order to adapt them to the processing of some spoken disfluencies (fragments words, for instance).

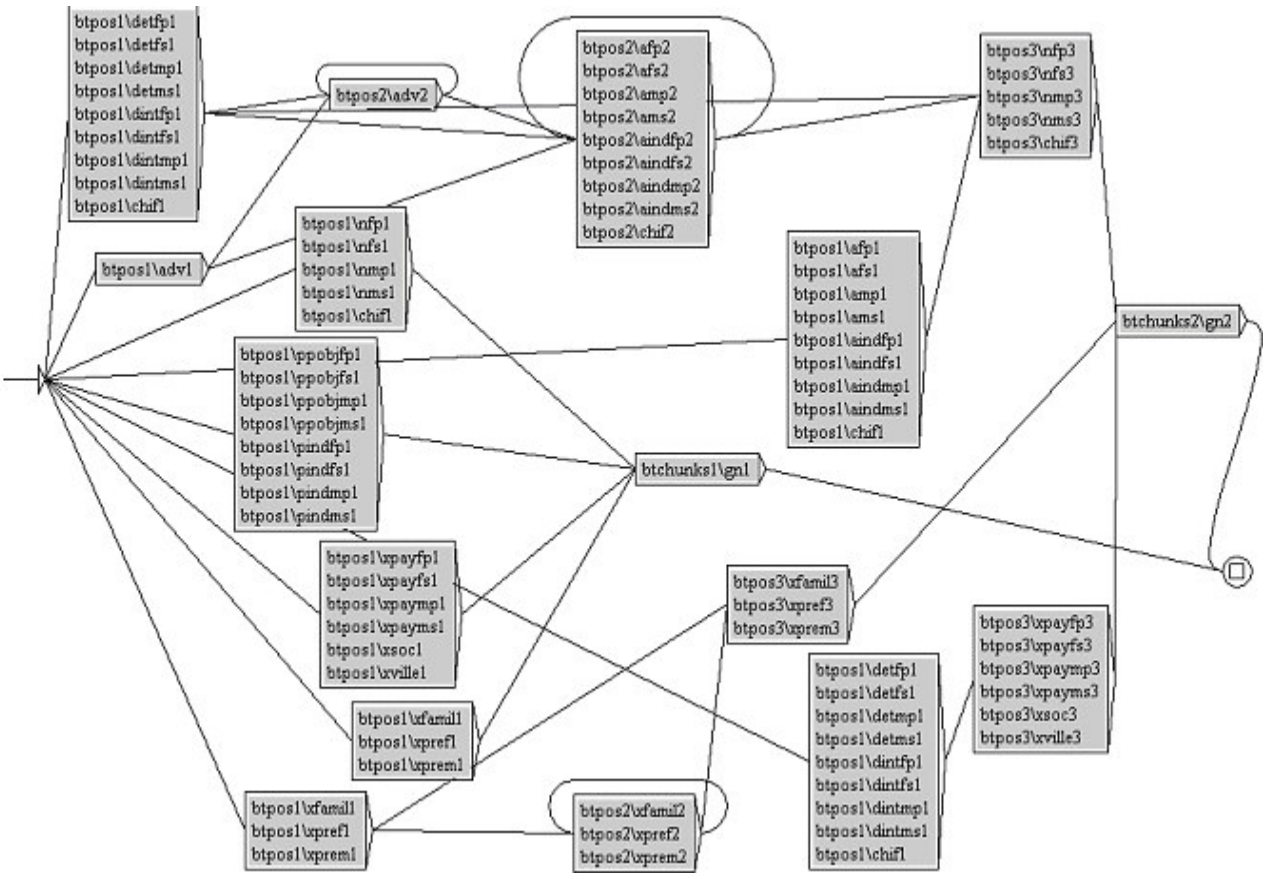

Figure 7: Example of a main transducer (GN) as defined on the CasSys platform

The first cascade is composed of 386 transducers. As described before, these transducers are applied in a precise order. This order must be carefully designed since it is essential for the control of the ambiguity during the chunking process, in particular when some chunk patterns should overlap. For instance, the main transducer GP (related to the prepositional chunk) includes in its definition a GN chunk (nominal chunk), as shown on figure 8. Obviously, it must be applied before the main transducer GN since the application of a main transducer replace the corresponding sequence.
Finally, the transducers of the first cascade are applied in the following order:

1) PV verbal chunk introduced by a preposition
2) NV verbal chunk, with its clitics
3) GP preposition chunk
4) GN nominal chunk
5) GA adjectival chunk
6) GR adverbial chunk
7) CHUNK, which consists only in bracketing the not identified sequences of tokens.

Then, the second cascade will apply only on the CHUNK segments. We will see in the result section that this second stage should be useful to correct errors from the previous stage of processing (POS tagging and, in the future, automatic speech recognition).

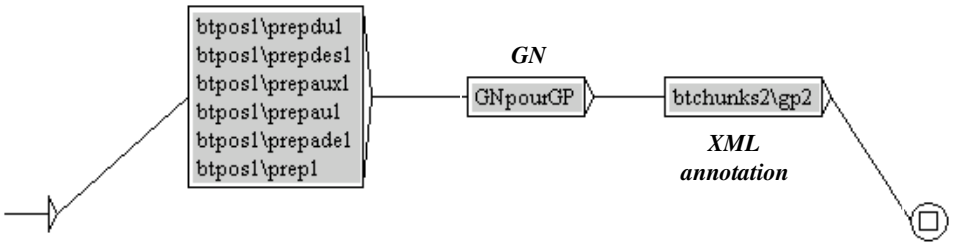

Figure 8: main transducer describing prepositional chunks (GP) on the CasSys plaform

5. Results

SECARE has been evaluated on corpora that are provided in the EPAC project. More precisely, the test corpus is a extract from the LIUM's manual speech transcripts, since automatic transcripts are not yet available. As a result, this section does not study the influence of speech recognition errors on the robustness of the system. On the opposite, the influence of the POS tagging stage is assessed into details.
On the whole, the test corpus involves highly conversational speech turns from a radio program. It gathers 893 chunks. The robustness is evaluated by means of the following measures:

- *Recall* (R): percentage of chunks of the corpus that are correctly detected (temporal frontiers) and labeled,
- *Precision* (P): percentage of correct chunks among all of the decisions of the system
- *Insertions* (I), *Substitutions* (S), *Deletions* (D): percentage of elementary operations (I,S,D) required to transform the segmented text to the correct sequence of chunks. These measures account for errors on chunks delimitations.

In order to asses the influence of the POS tagging. Two versions of the system have been tested. The first one (SECARE) involves only a pure syntactic modeling of chunks encountered in conversational speech. The second one (SECARE+) has an additional cascade which is designed to the correction of the most frequent errors of the POS tagging stage. Table 1 presents the corresponding results on the whole corpus.

System	R	P	I	S	D
SECARE	85.1%	76.3 %	20.0%	3.7%	10.8%
	F-score :	0.805			
SECARE+	90.7%	86.2%	10.2%	3.6%	5.4%
	F-score :	0.884			

Table 1: Performance of SECARE and SECARE+ on the whole test corpus (893 chunks)

System	R	P	F-Score
SECARE	95.3%	92.6 %	0.989
SECARE+	98.%	99.7%	0.987

Table 2: Performance results of SECARE and SECARE+ systems on the test corpus restricted to the chunks without any POS tagging error (816 chunks)

At first glance, the robustness of SECARE seems to be highly improvable. Indeed, its F-score on the test corpus amounts only to 0.805, will a low precision of 76.3%. Insertions are corresponding to 58% of the errors made by SECARE. This means that most of errors results in an over-segmentation of the text string.
However, a detailed analysis of the error cases shows that most of them are the consequences of erroneous POS tags. If we consider only the chunks were the LIA_TAGG didn't make any mistake (see Table 2), then the recall and the precision of the system increase significantly (F-score : 0.989).

One should therefore conclude that SECARE models parses correctly regular chunks and is not really disturbed by the speech disfluencies[4]. This is why we have decided to add a second cascade of transducers whom main aim is to limit the negative influence of the POS tagger. This cascade will try to model and correct the most frequent tagging errors, which result most of the time by the erroneous insertion of CHUNK chunks. As shown on Table 1, this post-correction is very useful, since the resulting F-score increase to the more satisfactory value of 0.884 (recall R = 90.7%; Precision P = 86.2%. The number of insertion decreases of around of 50%, which clearly shows that the influence of POS tagging errors is significantly less important. Besides, we add also a few transducers to model some complex regular structures. From now on, the resulting behaviour of the system on chunks without POS errors is almost perfect (F-score = 0,987). Anyway, we are still working of POS tag errors and our future researches will concern the management of speech recognition errors. For the moment being, SECARE already showed that it is well suited for the automatic chunking of manual speech transcripts.

6. Conclusion

In this paper, we have presented one of the NLP tasks that are involved in the EPAC project: chunk segmentation of large corpuses of conversational speech. We have been able to build a system that present at first glance satisfactory performances on the corpuses that are considered in the EPAC project. The SECARE performances show that it is possible to extend techniques of chunk segmentation to the general language. Now we will complete the second transducer cascade, in order to distinguish the disfluent segments (reparandum and edition zones) and to manage both POS tagging and automatic speech recognition errors. This segmentation can be used to the post-correction of the POS tagging. This capacity of correction would be particularly interesting for the extraction of named entities from transcribed conversational speech.

7. Acknowledgement

This work was financed by the French research agency (ANR-project 06-MDCA-2006). We thank Denis Maurel for his help on the use of the Unitex system.

8. References

Abney S. (1991) Parsing by chunks, In. Berwick, Abney, Tenny (Eds.) *Principle-based parsing. Amsterdam*. Kluwer Academic Publ. Dordrecht, Pays-Bas.

Abney S. (1996) Partial parsing via finite-state cascades. Proc. *ESSLI'1996 Robust Parsing Workshop*.

Aït-Mokhtar S., Chanod J.-P., Roux C. (2003) Robustness beyond shallowness: incremental deep parsing, *Natural Language Enginerring*, Vol. 8 (3-2).

Barras C. *and al.* 1998 (1998) Transcriber : a free tool for segmenting, labeling and transcribing speech. Proc. *LREC'98*, Granada, Spain, 1373-1376.

Blanche-Benveniste C. (1997) Approches de la langue parlée en français, Coll. L'essentiel Français, Ophrys, Paris, France.

Church K. (1988). A stochastic parts program and noun phrase parser for unrestricted text, actes Conference on Applied Natural Language Processing, ACL'1988, Austin, TX, 136-143.

Friburger N., Maurel D. (2004). Finite-state transducer cascades to extract named entities in texts. Theoretical Computer Science, 313(1): 93-104.

Goulian J., Antoine J.-Y., Poirier F. (2003) How NLP techniques can improve speech understanding : ROMUS, a robust chunk based message understanding system using link grammars. Proceedings of *Eurospeech'2003*, Geneva, Switzerland. 2773-2776.

Paroubek P., Robba I.,Vilnat A. Ayache C. (2006) Data Annotations and Measures in EASY the Evaluation Campaign for Parsers of French, Proceedings of the *5th international Conference on Language Resources and Evaluation, LREC 2006*, Genes Italy, pp.315-320.

Paumier S. (2003) De la reconnaissance de formes linguistiques à l'analyse syntaxique, Thèse de Doctorat, Université de Marne-la-Vallée, France.

Shriberg E. (1994). Preliminaries to a theory of speech disfluencies. PhD Thesis. University of California, Berkeley, CA.

Villaneau J., Antoine J-Y., Ridoux O. (2004) Logical approach to natural language understanding in a spoken dialog system. *Proc. 7th International Conference on Text Speech and Dialog, TSD'2004*, Brno, Tchéquie. 637-644

Vilnat A., Paroubek P., Monceaux L., Gendner V., Illouz G., Jardino M. (2003) EASY or How difficult Can It be to define a Reference Treebank for French, *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories* (*TLT*), Vaxjo, Sweden, November 14th-15th, 2003.

[4] This conclusion should however be relativized, since a high part of POS tagging errors occur in speech disfluencies.