Partial parsing of spontaneous spoken French

Olivier Blanc[1], **Matthieu Constant**[2], **Anne Dister**[1,3], **Patrick Watrin**[1]

[1]Université de Louvain, Belgium
[2]Université Paris-Est, LIGM & CNRS, France
[3]Facultés universitaires Saint-Louis, Belgium
olivier.blanc@uclouvain.be, mconstan@univ-mlv.fr, anne.dister@uclouvain.be,patrick.watrin@uclouvain.be

Abstract

This paper describes the process and the resources used to automatically annotate a French corpus of spontaneous speech transcriptions in super-chunks. Super-chunks are enhanced chunks that can contain lexical multiword units. This partial parsing is based on a preprocessing stage of the spoken data that consists in reformatting and tagging utterances that break the syntactic structure of the text, such as disfluencies. Spoken specificities were formalized thanks to a systematic linguistic study of a 40-hour-long speech transcription corpus. The chunker uses large-coverage and fine-grained language resources for general written language that have been augmented with resources specific to spoken French. It consists in iteratively applying finite-state lexical and syntactic resources and outputing a finite automaton representing all possible chunk analyses. The best path is then selected thanks to a hybrid disambiguation stage. We show that our system reaches scores that are comparable with state-of-the-art results in the field.

1. Introduction

Large annotated corpora of transcribed spontaneous speech are of great interest for many fields of Natural Language Processing. Nevertheless, their manual construction is painful and requires automatic tools. This paper describes the process and the resources used to automatically annotate a French corpus of spontaneous speech transcriptions in super-chunks. Super-chunks are enhanced chunks that can contain lexical multiword units. This partial parsing is based on a preprocessing stage of the spoken data that consists in reformatting and tagging utterances that break the syntactic structure of the text, such as disfluencies. The chunker uses large-coverage and fine-grained lexical resources for general written language that have been augmented with resources specific to spoken French. In section 2., we describe the corpus used and its specificities. In section 3., we show how we dealt with them during the preprocessing stage. We then describe the architecture of our chunker (section 4.) as well as the language ressources used (section 5.). The last section is dedicated to the evaluation of the whole process. We show that our chunker reaches scores comparable with state-of-the-art for French.

2. Spoken Corpus

The corpus we used is a sub-corpus extracted from the spoken textual data bank of Valibel. It includes 60 texts transcribed from spoken conversations, composed of 443,047 graphical words. They approximately correspond to 40 hours of spontaneous speech, all recorded in the French speaking part of Belgium. The talks are mainly semi-directed interviews and talks between friends, and have the characteristics of not being planned (as opposed to texts written to be spoken). More details about this corpus (speakers, sociolinguistic information, context of talks) can be found in (Dister, 2007).

The transcriptions follow the guidelines that have been developed at the Valibel research Center (Dister et al., 2006). The main principles are, in the most part, similar to those used in other laboratories[1] working on textual transcriptions of recorded speech. Firstly, words are transcribed using their standard spelling. Transcriptions do not contain any punctuation marks because the notion of sentence is not relevant for spoken language (Blanche-Benveniste and Jeanjean, 1987). The sound continuum, that has become linear with the transcription, is divided into speaking turns, defined by the change of speaker. The silent pauses are annotated subjectively by the transcriber with respect to three levels: short pause (/), long pause (//) and silence (///). Texts include phenomena that are specific to spoken language such as disfluencies and overlapping speech segments, as illustrated in the example below:

(1) **blaAD1** avec une / une ba/ une barre qui bah tu es / tu es en l'air et puis tu te laisses glisser |- le long <blaNB1> ouais -| d'une barre
blaAD1 *with a / a ba/ a bar which bah you are / you are in the air and then you let yourself slide |- along <blaNB1> yeah -| a bar*

This transcription indicates that *blaAD1* is speaking. The tag |- (resp. -|) starts (resp. ends) an overlapping segment, where *blaNB1* says *ouais* when *blaAD1* says *le long*. *ba/* indicates a word (starting by *ba*) that has not been completed.

3. Preprocessing of the spoken data

In their present state, the transcriptions cannot be used as is in a chunker without significant modifications of the latter, because of the transcription format and the spoken specificity of the data. The goal of the preprocessing module is to detect any phenomena that are specific to spoken language and normalize them so that they can be processed

[1] cf. for instance, the DELIC corpus (DELIC, 2004) or data from the *Rhapsodie* project `http://rhapsodie.risc.cnrs.fr/fr/index.html`.

more easily by the chunker initially developped for written texts[2]. For instance, disfluencies contain segments that need to be tagged in order not to be taken into account by the annotator. In this section, we briefly describe this preprocessing stage (Dister et al., 2010).

3.1. Preprocessing overlapping fragments

The transcriptions contain thousands of overlapping fragments that are speech segments produced by a person while another person is already speaking. These fragments break the linearity of the reading of the transcribed text because they occur within a speaking turn. The preprocessing module extracts internal overlapping fragments, put them in new speaking turns and insert a tag referencing the new turn at the previous location. The raw transcription in example 1 would then be reformatted as follows:

(2) #23 **blaAD1** avec une / une ba/ une barre
qui bah tu es / tu es en l'air et puis tu te
laisses glisser |- @24 le long -| d'une barre
#24 **blaNB1** ouais

The overlapping fragment enunciated by *blaNB1* is extracted in a new speaking turn (numbered #24) from the turn corresponding to *blaAD1* (#23). A reference to turn #24 (@24) is put at the beginning of the overlapped fragment in #23.

3.2. Insertion of punctuation markers

As explained in section 2., the transcriptions do no contain punctuation marks. Nevertheless, an efficient application of the chunker requires the input string to be divided into units smaller than a speaking turn. We therefore used cues like silence marks (///) and long pauses (//) in order to get a segmentation consistent with the syntactic structure of the statement.

3.3. Preprocessing disfluencies

The disfluencies are standardly seen as a location of the speech flow where the linearity is broken, because it stops for some time at a point on the syntagmatic axis. They are very numerous in spoken texts and are of several different types: (a) repetitions, i.e. sequences of two (or more) contiguous graphically identical forms (e.g. *la la*); (b) immediate self-corrections, i.e. a sequence of two morphemes, the second one having the same part-of-speech as the first one (Candea, 2000) and tending to correct the first one (e.g. *le la*); (c) word fragments, i.e. phenomena that consist of an interruption of the morpheme being enunciated. For instance, one type of word fragments are completed word fragments where the started word is completed after the interruption at the same syntactic location.

(Shriberg, 1994, 7-9), following (Levelt, 1989), divided a disfluency utterance into four distinct elements: (a) **reparandum**, i.e. the part produced by the speaker to be deleted and to be later replaced by the repair (cf. later); (b) **interrupting point**, i.e. the moment just after the end of the reparandum; (c) **interregnum**, i.e. the region that starts at the end of the reparandum and ends at the beginning of the repair; it may contain editing terms (i.e. a silent pause, a filled pause, ...) or several attempts of unachieved reformulation, all to be deleted. (d) **repair**, i.e. correponds to the corrected content of the reparandum.

The automatic preprocessing consists in detecting the disfluencies and annotating the *reparandum* and the *interregnum* parts with tags ($IGN + disf$) such that the chunking process can only take the *repair* into account. Our tool identifies three disfluencies ($IGN + disf$) and two overlapping markers ($IGN + over$) to be ignored by the superchunker as illustrated below:

(3) #23 **blaAD1** avec {**une / une
ba/,.IGN+disf**} une barre qui
{**bah,.IGN+disf**} {**tu es /,.IGN+disf**} tu es
en l'air et puis tu te laisses glisser {|-
@24,.IGN+over} le long {-|**,.IGN+over**}
d'une barre

Note that disfluencies can be combinations of simple disfluencies such as *une / une ba/ une barre* that contains a repetition and a word fragment. Such phenomena complicates their identification.

4. The super-chunker

The annotation process uses an incremental finite-state *super-chunker* (Blanc et al., 2007) that is briefly described in this section.

4.1. Super-chunks

Super-chunks are non-recursive syntactic constituents, like standard chunks (Abney, 1996), but they can contain complex multiword units (MWUs). For instance, *marge d'exploitation* (trading margin) is considered as a standard compound noun, so the utterance *la marge d'exploitation* (the trading margin) is annotated as a nominal super-chunk (XN), while standard chunking would have produced a sequence of a noun phrase (XN) followed by a prepositional phrase (XP)[3]. One (or more) standard prepositional phrase can therefore be integrated into a nominal super-chunk. Considering super-chunks instead of standard chunks has two main interests: (1) it reduces attachment complexity for deep parsing because some of them are resolved with MWU recognition; (2) it allows for the identification of semantic units as MWUs form linguistic units (Copestake et al., 2002).

Several multiword units can be combined into a same super-chunk. For instance, let's consider the following annotated sequence:

> [XN La température] [XP à l'intérieur de beaucoup de maisons] [XP en Moldavie]

[2] A similar approach has been conducted in (Valli and Véronis, 1999) for part-of-speech tagging. (Antoine et al., 2008) propose another approach by including a post-correction stage in order to deal with chunking errors due to disfluencies for instance.

[3] The utterance would be annotated in standard chunks like below:

> [XN la marge XN] [XP d'exploitation XP]
> *(the trading margin)*

([XN the temperature] [XP inside a lot of houses]
[XP in Moldavia])

The whole phrase *à l'intérieur de beaucoup de maisons* is considered to be a simple prepositional super-chunk (XP) because *à l'intérieur de* (inside) is a multiword preposition, *beaucoup de* (a lot of) is a multiword determiner and *maisons* (houses), a simple noun.
Verbal chunks are also very specific because they can include auxiliaries, modal verbs, inserts, clitics and negation. For example, the sentence

Jean n'a pas pu les trouver
(John could not find them)

is annotated:

[XN Jean] [XV n'a pas pu les trouver]
([John] [could not find them])

The discontinuous sequence *n'... pas* (not) is a negation, *a ... pu* is the preterit form of the modal verb *pouvoir* (to can) and *les* is an accusative clitic.

4.2. Tagset

The annotation tagset of the super-chunker is given in table 1. Each tag can have several features. These features can be enumeration, boolean or string. For instance, a verbal chunk (XV) has a 'mood' feature that can be either indicative, subjunctive, infinitive, gerund or past participle. XV also has a 'negation' feature which is a boolean. The 'preposition' feature of an XP contains the lexical value of its introducing preposition. We also use lexical tags like conjunctions or pronouns like relative ones. Other lexical tags (e.g. determiners, prepositions) may be used, for instance, when a speaker enunciate an incomplete sentence ending with a determiner.
The example 1 is then annotated as provided in table 2[4]. For instance, the verbal chunk *tu te laisses* is in the indicative mood (+ind) and contains two clitics: a nominal one (tu,+ppvnom) and a reflexive one (te,+ppvref). Its head is *laisser* (let). In the prepositional chunk *le long d' une barre*, the MWU *le long de* (along) is the preposition and *barre* (bar) is the head noun.

4.3. Segmentation into super-chunks

The chunker is composed of three successive stages as illustrated in the diagram[5] in figure 1: (1) lexical segmentation into simple words and multiword units (MWUs); (2) identification and tagging of super-chunks; (3) disambiguation process. The whole system is mainly driven by linguistic resources in the form of lexicons and grammars (cf. section 5.).
The lexical analysis step takes as input a text segmented into sentences and tokens. First, a dictionary lookup associates each token with all its possible linguistic tags and recognizes compounds. The output of the process is a finite state automaton (TFSA), representing all the lexical ambiguities. Then, lexicalized grammars are directly applied to the TFSA, which is augmented with the analyses of the matching MWUs.

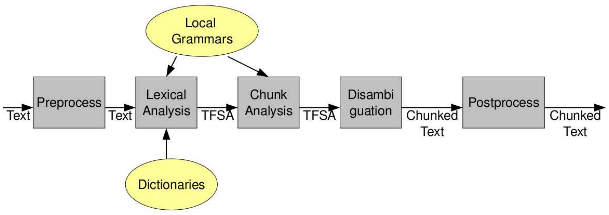

Figure 1: Process diagram

Chunk segmentation is based on a cascade of finite transducers applied to the TFSA, which is augmented each time a new chunk is found. The identified chunks inherit morpho-syntactic properties from their components.
In order to remove ambiguity, the chunker includes an incremental disambiguation module composed of three stages:

- **Disambiguation with Hand-Crafted Rules**: given an ambiguity (a set of possible analyses) and a left and/or right context in the TFSA, a hand-written rule selects one analysis among the ambiguities and removes the others from the TFSA.

- **Shortest Path Heuristic**: it only keeps the shortest paths of the TFSA; it is based on the idea that multiword expression analyses are preferred to sequences of simple word analyses.

- **Simple Stochastic Linearization**: it keeps the path composed of the analyses which are the most frequent in a learning tagged corpus (if not found in the learning corpus, an arbitrary decision is made.)

5. Resources

5.1. Lexical resources

The lexical segmentation module includes large-scale dictionaries developed by linguists. They are lists of lexical entries, each of them being composed of an inflected form, a lemma, a part-of-speech, morphological information (*e.g.* gender, number), syntactic information (*e.g.* transitive or intransitive verbs) and semantic information (*e.g.* human feature for nouns).
The larger dictionary has been developed between the mid-80's and the mid-90's by linguists at the University of Paris 7 (Courtois, 1990; Courtois et al., 1997). It is composed of 746,198 inflected simple forms and 249,929 inflected compounds (including 245,436 compound nouns). Compounds are of the following types :

- nouns: *pomme de terre* (potato), *faux témoignage* (perjury)

- prepositions: *au milieu de* (in the middle of), *afin de* (in order to)

[4] Note that morphological features and the IGN disfluency parts are not specified for readability reasons.

[5] Note that the diagram process not only includes a preprocessing module (cf. section 3.) but also a postprocessing module for the display of all annotations (*i.e.* super-chunks and disfluencies).

TAG	LABEL	FEATURES
XA	adjectival chunk	head, gender, number, person ...
XADV	adverbial chunk	type (date, duration,...)
XN	nominal chunk	head, gender, number, person, ...
XP	prepositional nominal chunk	head, preposition, gender, number, person, ...
XV	verbal chunk	head, mood, person, number, voice, clitics, negation, ...
XVP	prepositional verbal chunk	head, preposition, mood, person, number, voice, clitics, negation, ...

Table 1: Chunker tagset

CHUNK	TAG	TRANSLATION
avec une barre	XP+head=barre+prep=avec	with a bar
qui	PRO+rel	which
tu es	XV+ind+ppvnom+head=être	you are
en l ' air	XP+head=air+prep=en	in the air
et puis	conjc	and then
tu te laisses	XV+head=laisser+ind+ppvnom+ppvref	you let yourself
glisser	XV+inf	slide
le long d' une barre	XP+head=barre+prep=le_long_de	along a bar

Table 2: Chunking result example

- adverbs: *en outre* (in addition), *en pratique* (in practice)
- conjunctions: *bien que* (although), *pendant que* (while)

In addition, we constructed three dictionaries taking into account the double specificity of our corpus of spoken Belgian French. The first one (5,124 entries) is composed of lexical particularities of Belgian French, with forms like *guindailleur* (person who likes parties). The second one (25 entries) is devoted to simple and compound words that could be assigned another part-of-speech specific to spoken French — e.g. *allez* which can be analyzed as an interjection in addition to a inflected form of the verb *aller* (to go). The third one (67 entries) is a dictionary of onomatopoeia, such as *ah* and *nom de dieu* (my god).

Our lexical resources also contain a library of lexicalized local grammars (Gross, 1997) in the form of finite-state transducers. They are Recursive Transition Networks (RTNs) (Woods, 1970) and theoretically recognize algebraic languages. They are mostly used to represent MWUs. They can define syntactic classes such as noun determiners and even syntactico-semantic classes such as time adverbials. Linguistic descriptions are in the form of Finite-State Graphs on an alphabet of terminal and non-terminal symbols. A terminal symbol is a lexical form or a lexical mask. A lexical mask is an underspecified lexical entry (some features are missing) equivalent to a feature structure representing a set of lexical entries: *e.g.* the lexical mask <*avion.noun*> matches all nouns the lemma of which is *avion* (plane). Finally, a non-terminal symbol is a reference to another graph. A graph is a transducer and its output is the annotation assigned to the structures described in the graph. An example of a local grammar is given in figure 2[6]. This grammar describes time adverbials like *en mars 2007* (in March 2007) and *cinq minutes plus tard* (five minutes later). The sequences recognized by this graph are labelled as time adverbs (`ADV+time`). Strings between < and > are lexical masks: for instance, <*minute*> stands for the inflected forms whose lemma is *minute*. Greyed vertices are call to other graphs. For example, `Dnum` and `month` are graphs that respectively recognize numerical determiners and the names of months.

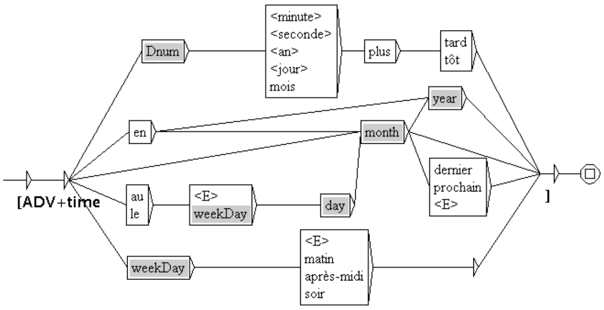

Figure 2: Local grammar of time adverbials

Practically, the lexical module includes a network of 381 graphs recognizing multiword sequences such as function names, locative prepositions, noun determiners or time adverbials.

5.2. Syntactic resources

Most of standard chunks are recognized by means of pure syntactic patterns, factorized in the form of graphs. For instance, graph in figure 3 recognizes XPs. A recognized XP inherits features from its internal consituents. For example, the feature 'prep' is the lemma of its preposition: *+prep=ˆlemma*. The gender of XP is the gender of its head noun +*ˆgender*).

In some cases, the grammar can be heavily lexicalized for example to describe the different semi-auxiliary verbs (in the sense of (Gross, 1999)) that can occur in a chunk XV:

[6] The local grammars are drawn using the graph editor of the Unitex platform (Paumier, 2010).

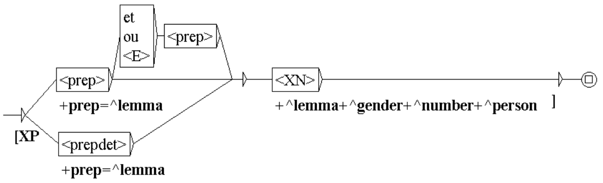

Figure 3: Local grammar of an XP

	SEG	STD	FEAT	FEAT+

with	p=88.7
	r=91.7
	f=90.2
		p=82.8
		r=85.6
		f=84.1
			p=80.7
			r=83.4
			f=82.0
				p=73.6
				r=76.1
				f=74.8

without	p=87.3
	r=89.6
	f=88.4
		p=81.5
		r=83.5
		f=82.5
			p=79.1
			r=81.1
			f=80.1
				p=71.5
				r=73.3
				f=72.4

Table 3: Evaluation (in percentage)

> *Jean [XV a commencé à aimer danser]* (John [did start to like dancing])

In total, the cascade of FSTs for the syntactic phase uses a network of 136 graphs. There is no grammar specific to spoken language at this stage.

6. Evaluation

This section is devoted to the evaluation of our chunker. As there is currently no available corpus for spoken French annotated with super-chunks, it has been necessary to build our own reference annotated corpus. The evaluation experiments were carried out on parts of the spoken French corpus described in section 2. Our corpus was composed of 5,336 graphical words and 2,335 chunks. We applied the chunker twice: once with all resources described in the previous section; once without the additionnal resources specific to spoken language. We also made several evaluations according to the granularity of the annotation tagset :

- segmentation evaluation (SEG): only starting and ending positions of the chunks were taken into account;
- standard evaluation (STD): annotation only included chunk categories (XADV,XA,XN,and so on);
- feature evaluation (FEAT): annotation also included features like the lexical head of a XN, XV, *etc.*, the verbal mood or the preposition value of a XP;
- more feature evaluation (FEAT+): annotation also included morphological information like gender, person and number.

The measures used for the evaluation are precision (p), recall (r) and F_1-measure (f). Precision is the percentage of chunks in the Hypothesis corpus (i.e. the one resulting from the application of the chunker) that are correctly annotated. Recall is the percentage of chunks in the reference corpus that also belong to the hypothesis corpus. F_1-measure[7] is a harmonic mean of recall and precision. The experiments described above ended up with the results in table 3 (given in percentage).

Our chunker reaches a 0.84 f-measure for the STD granularity by including all resources in the system (row WITH in table 3). The results are comparable with state-of-the-art for standard chunking of spoken French (cf. results of the EASY evaluation campaign[8] provided in (Paroubek et al., 2007)). By adding morphological features to the annotation like gender, number and person features (FEAT+), the results suddenly fall down (by around 7 points). This can be partly explained by the fact that there are very few disambiguation rules dealing with these features, especially for gender, number and person ones. We also observe that, without resources specific to spoken French, the score drops by around 2 points in f-measure (cf. row WITHOUT), which shows that such data is useful to bring more accuracy to the system.

By analyzing more carefuly the results, we can see that almost two-third of errors are only due to a bad segmentation in super-chunks (vs. errors in also assigning chunk categories in SEG column). Precision and recall errors are caused by several factors. For instance, some multiword units are missing in the lexical resources: e.g. in our local grammar describing time adverbials, the sequence *à huit heures du soir* (at eight o'clock in the evening) is missing which generates a segmentation error. Characteristics of spontaneous spoken such as disfluencies, ungrammatical or unfinished sentences are infrequent causes of errors because those utterances have been, for the most part, filtered during preprocessing phase.

7. Conclusions and future work

In this paper, we described a process for super-chunking transcriptions of spontaneous French speech. This annotation system includes a preprocessing module based on a linguistic description of spoken specificities. It consisted in reformatting speaking turns with overlapping fragments and in tagging parts of disfluencies to be ignored by the analyzer. The evaluation showed that our resource-based chunker reached scores comparable with state-of-the-art for French. Despite this, it may need an evaluation on a larger corpus. It also requires more improvements such as development of more lexical grammars and integration of Hidden Markov models and Conditional Random Fields in the disambiguation stage.

8. References

Steven P. Abney. 1996. Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4):337–344.

[7] $f = \frac{2.p.r}{p+r}$

[8] The EASY definition of a chunk and the evaluation corpus are different from ours.

Jean-Yves Antoine, Abdenour Mokrane, and Nathalie Friburger. 2008. Automatic rich annotation of large corpus of conversational transcribed speech: the chunking task of the epac project. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Olivier Blanc, Matthieu Constant, and Patrick Watrin. 2007. Segmentation in super-chunks with a finite-state approach. In *Proceedings of FSMNLP 2007 – Finite-State Methods for Natural Language Processing*.

Claire Blanche-Benveniste and Colette Jeanjean. 1987. *Le Franais parlé. Transcription et édition*. Didier Érudition, Paris.

Maria Candea. 2000. *Contribution à l'étude des pauses silencieuses et des phénomnes dits d'hésitation en français oral spontané*. Ph.D. thesis, Université de Paris-3 Sorbonne-nouvelle.

A. Copestake, F. Lambeau, A. Villavicencio, F. Bond, T. Baldwin, I. A. Sag, and D. Flickinger. 2002. Multiword expressions: linguistic precision and reusability. In *Proc. of the Third conference on Language Resources and Evaluation*, pages 1941–1947, Las Palmas, Canary Islands.

B. Courtois, M. Garrigues, G. Gross, M. Gross, R. Jung, M. Mathieu-Colas, A. Monceaux, A. Poncet-Montange, M. Silberztein, and R. Vivés. 1997. Dictionnaire électronique DELAC : les mots composés binaires. Technical Report 56, LADL, University Paris 7.

Blandine Courtois. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87:11–22.

Equipe DELIC. 2004. Présentation du corpus de référence du franais parlé. *Recherches sur le franais parlé*, 18:11–42.

Anne Dister, Michel Francard, Geneviève Geron, Vincent Giroul, Philippe Hambye, Anne Catherine Simon, and Régine Wilmet. 2006. Conventions de transcription régissant la banque de données valibel. Technical report, Université catholique de Louvain. http://valibel.fltr.ucl.ac.be/.

Anne Dister, Matthieu Constant, and Gérald Prunelle. 2010. Normalizing speech transcriptions for natural language processing. In *Proceedings of the international conference on Spoken Communication*.

Anne Dister. 2007. *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel*. Ph.D. thesis, Université catholique de Louvain.

Maurice Gross. 1997. The construction of local grammars. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 329–352. The MIT Press, Cambridge, Mass.

Maurice Gross. 1999. Lemmatization of compound tenses in English. *Lingvisticae Investigationes*, 22.

Willem J.M. Levelt. 1989. *Speaking: from intention to articulation*. The MIT Press, Cambridge.

Patrick Paroubek, Anne Vilnat, Isabelle Robba, and Christelle Ayache. 2007. Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. In *Actes de TALN 2007 (Toulouse)*.

Sébastien Paumier. 2010. Unitex documentation. version 2.1. Technical report. http://igm.univ-mlv.fr/ unitex.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, Université de Berkeley.

André Valli and Jean Véronis. 1999. Étiquetage grammatical des corpus de parole: problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133.

W.A. Woods. 1970. Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10).