Building and exploiting a dependency treebank for French radio broadcasts

Christophe Cerisara, Claire Gardent and Corinna Anderson

CNRS/LORIA, Nancy
Firstname.Lastname@loria.fr

Abstract

We describe the construction of a dependency treebank for French radio broadcasts and present some results on how genre-specific phenomena affect parsing. Preliminary experimental results realized on one hour of speech suggest in particular that not only disfluencies but also radio headers and guest speech have a negative impact on parsing accuracy.

1 Introduction

Much work in recent years has focused on developing treebanks for transcribed speech. For English, the most well known is the Switchboard corpus [7]. For French however, the only efforts made in this direction we are aware of concerns the syntactic annotations of the European parliament debates [2]. Although this treebank is very large, it has only been automatically analyzed so far. We thus initiate in this work some efforts to manually analyze a more common type of speech, broadcast news transcripts.

The ESTER Corpus contains transcripts of broadcast news while the Media corpus contains about 70 hours of dialogs which were manually transcribed and semantically annotated with a set of 80 basic concepts. However, neither of these two corpora is syntactically annotated. Similarly, the Paris 7 treebank (P7TB, 12 500 sentences, 325 000 words, [1]) consists of articles from *Le Monde* newspaper semi-automatically enriched with phrase structure annotations and manually verified. The P7 dependency treebank (P7Dep, [6]) was automatically derived from it by conversion. Neither of these treebanks however contain a sizable portion of speech.

In this paper, we report on the construction and exploitation of a dependency treebank for spoken French. We start by presenting the annotation schema used and relate it to the schema used for written French in the P7Dep treebank (Section 2). We then describe the tools and methodology used to construct the treebank (Section 3). Finally, we exploit the speech treebank for training a parser and present

some first results concerning the impact of various speech-specific constructs on speech parsing. In particular, we show that for broadcast news, disfluency might not be the main factor for performance degradation, and that genre-specific constructs such as headlines and guest speech, which tends to be characteristic of natural unplanned discourse (as opposed to the prepared speech of professional radio journalists), play an important role in performance loss (Section 4).

2 Corpus and Annotation schema

To develop a treebank of spoken French, we build on existing work and take as a starting point the ESTER corpus of French radio broadcasts and the P7Dep and Syntex [3] annotation schema for dependency structures.

2.1 Corpus

The corpus used to develop a treebank of spoken French (the ESTER treebank henceforth, ETB) is the ESTER corpus of manual transcripts for French radio news (1998 - 1999 and 2003).

Because these transcripts were developed with the aim to evaluate speech recognisers, only complete words were transcribed. Hesitations "euh" were considered as words and were transcribed but noise, starts, laughs, jingles indications, etc. were not. For parsing, all punctuation information was removed so as to simulate the output of a speech recogniser, which typically does not produce such information.

Further, words are grouped by the transcribers into prosodic segments which do not necessarily coincide with sentences. During the annotation however, the annotators can join segments together whenever the prosodic segments form incomplete constituents.

2.2 The ESTER Annotation schema and its relation to the P7Dep annotation schema

The annotation schema (The ESTER Annotation schema) we define to annotate the ESTER corpus is derived from our previous works with the Syntex parser [3] and with the P7Dep corpus. It comprises 15 dependency relations: SUJ (subject), OBJ (object), POBJ (prepositional object), ATTS (subject attribute), ATTO (object attribute), MOD (modifier), COMP (complementizer), AUX (auxiliary), DET (determiner), CC (coordination), REF (reflexive pronoun), JUXT (juxtaposition), APPOS (apposition), DUMMY (syntactically governed but semantically empty dependent e.g. expletive subject "il / it" in "il pleut / it rains"), DISFL (disfluency).

As shown in Table 1, this schema has a direct partial mapping to the annotation schema used to annotate the P7 treebank of newspaper text. The differences between the two annotation schemes relate to prepositional objects, auxiliaries, mod-

ifiers and speech or written text specific constructs such as disfluencies (speech) or punctuation (text).

We thus defined and implemented a rule-based converter from the ETB to the P7Dep formats as follows. Prepositional objects are differentiated in the P7Dep annotation schema, between a_obj, de_obj and p_obj , while in the ETB, all prepositional objects are marked as POBJ. To convert from ETB to P7Dep, we systematically convert prepositional objects headed with the "à" and "de" preposition into a_obj and de_obj respectively. For clitics (which do not contain a preposition and can be ambiguous), we use default mappings and exception lists for non-default cases. For instance, the clitic "*en*" usually indicates a de_obj (e.g., *en rêver / rêver de Paris*) but can also pronominalise the object NP (e.g., *Jean donne des pommes à Marie / Jean en donne à Marie*). We use the information that *donne* is a ditransitive and *rêver* a de_obj verb to appropriately map the clitics *en* to obj and de_obj respectively.

To differentiate between the various types of auxiliaries (passive, causative or temporal), we use hand written rules and lists of e.g., passivisable verbs, causative verbs and temporal auxiliaries that have been compiled over the years in our team. For instance, we use the list of passivisable verbs and pattern matching rules to decide whether the auxiliary "be" is used as a passive (*Il est aimé / He is loved*) or as present perfect (*Il est venu / He has arrived*) auxiliary.

Similarly, modifiers are differentiated into mod_rel (a relative clause modifying a noun), dep (a prepositional phrase modifying something else than a verb) and mod (all other types of modifiers) using hand-written pattern matching rules describing these three configurations. Further rules are used to convert coordination constructs (coord and dep_coord P7Dep relations) and reflexive clitics (REF ETB relation). Relations that have an onto mapping in the P7Dep format (SUJ, ATTS, ATTO, OBJ, COMP, DET, DUMMY) are mapped to the corresponding ETB relation. Relations that exists only in the ETB (DISFL,JUXT,APPOS,MULTIMOTS) are all mapped to the mod relation.

We assessed the accuracy of these conversion rules on the ESTER test corpus manually annotated in the P7Dep format: the conversion labelled attachment score (percentage of tokens with correct predictor governor and dependency type, LAS) is 92.6% and unlabelled attachment score (the ratio of words with a correct head, UAS) is 98.5%.

2.2.1 Annotation of speech-specific constructs.

Some constructs which are very frequent in speech are either absent in the P7Dep schema (disfluencies) or not differentiated from one another (juxtaposition and apposition treated as modification). To allow for a detailed study of these constructs, the ESTER schema labels them separately and annotates them as described below. Additionnally, sentence-level annotations were introduced in order to support a finer-grained analysis of the impact of speech constructs on parsing.

ETB	Label description	P7Dep
MOD	modifier	mod mod_rel dep
COMP	complementizer	obj
DET	determiner	det
SUJ	subject	suj
OBJ	object	obj aux_caus
DISFL	disfluency	mod
CC	coordination	coord dep_coord
POBJ	prepositional object	a_obj de_obj p_obj
ATTS	subject attribute	ats
JUXT	juxtaposition	mod
MultiMots	multi-word expression	mod
AUX	auxiliary	aux_tps aux_pass aux_caus
DUMMY	empty dependent	aff
REF	reflexive pronoun	obj a_obj de_obj
APPOS	apposition	mod
ATTO	object attribute	ato

Table 1: The mapping between the ESTER and the P7 dependency relations

Juxtaposition, apposition and disfluencies While apposition and juxtaposition occur in written text as well as in speech, both relations are labeled as *mod* in the P7 treebank. Because they occur very frequently in transcribed speech, we created specific labels (*appos* and *juxt*) for them in the ESTER corpus.

The apposition relation is used to relate two adjacent constituents denoting the same referent (1a)[1].

(1) a. Jean Tiberi le maire de la capitale *maire = appos (Jean)*
Jean Tiberi the mayor of the capital city

b. la seconde modification de la constitution celle qui concerne la réforme législative *celle = appos (modification)*
The second modification of the constitution that which concerns the legal reform

[1]The notation *D = rel (G)* indicates that the dependent D is related to its governor G by the dependency relation *rel*.

In contrast, the juxtaposition relation is used to relate constituents which occur in the same prosodic group but are not syntactically related: juxtaposed constituents (2a), enumeration (2b) and incidents (2c).

(2) a. le téléphone sonne bonsoir *bonsoir = juxt (téléphone)*
Le téléphone sonne good evening[2]

b. des gens qui roulent à 70 80 *80 = juxt (70)*
People who drive at 70 80 mph

c. Un voyage j'espère agréable *espère = juxt (voyage)*
A trip I hope pleasant

Disfluencies include hesitations, repairs and missing words. Hesitations (euh, 3a) are attached as a dependent to the head of the right adjacent constituent. When a word is missing (3b), the incomplete constituent is linked to the intended governor. Finally, in case of repairs, if the disfluency forms a constituent (3c), its head is linked to the head of the repair. Otherwise, the tokens forming the disfluency are each attached to the right adjacent token (3d).

(3) a. des conditions euh identiques *euh = disfl (identiques)*
Conditions hum identical

b. on n' attend pas euh euh 6 mois lui dire ce que tu as fait n' est pas bien
dire = disfl (attend)
You dont wait hum hum 6 months tell him that what you did was not good

c. Nous en relevons nous le relevons
nous = suj (relevons) en = mod (relevons) relevons = disfl (nous)
We notice them we notice it

d. et bien je ... *et = disfl (je)* *ben = disfl (je)*
and so I ...

Sentence level annotations As mentioned above, we introduced sentence-level annotations in addition to the dependency annotations in order to facilitate data analysis. These sentence-level annotations are the following:

- GUEST / SPEAKER: differentiates guests utterances from radio speaker ones
- ELLIPSIS: indicates that the sentence contains an ellipsis
- HEADER: indicates a radio header e.g.,(4)

(4) Michel à Aix-en-Provence en ligne bonsoir Michel
Michel from Aix-en-Provence on line good evening Michel

[2] *Le téléphone sonne* is the name of the show.

3 Treebank construction

Manual annotation of a full raw textual corpus with dependencies is time consuming, error prone and cognitively demanding for the human annotators. We have therefore opted for an iterative annotation procedure alternating training, parsing and manual correction. The cognitive effort required by the human annotator per sentence is thus greatly reduced, as she only has to check the proposed dependencies and possibly to modify some of them. The corpus produced in this way is further validated by an expert linguist. This validation phase involves several meetings between the annotators and the expert linguist where ambiguous and difficult examples are discussed and solved.

3.1 Methodology

The annotation procedure is as follows:

1. The manual transcription of a contiguous session of one hour-length is extracted from the development set of the broadcast news ESTER corpus [4]. This session constitutes the raw corpus that is annotated next.

2. This full raw unlabeled text corpus is split into 17 sub-corpora $(C_1, \cdots, C_{17})$ of about 630 words each. These sequences of words are manually segmented into utterances during the course of the following annotation procedure.

3. At iteration t, the unlabeled sub-corpus C_t is first automatically tagged with POS tags with the TreeTagger configured for French.

4. C_t is then automatically parsed with the Malt Parser [8] (cf. section 3.2) and the previous models λ_{t-1}. Note that the initial models λ_0 have been trained on another corpus of 20000 words that has been previously annotated with dependencies, but which is not used in the work described here [3].

5. The annotator (a linguistics student) then loads C_t into the edition software J-Safran, segments it into utterances, checks the proposed dependency labels and may modify them according to the annotation guide.

6. The malt parser models are retrained on C_t, leading to the new parameters λ_t.

7. This process is iterated from step 4 until the whole corpus is annotated.

This iterative process is interleaved with meetings between the annotators and the expert linguist whose aim is to discuss outstanding issues and to validate the annotations produced. The treebank thus obtained contains 1 hour of speech, i.e., 10654 words and 594 segments. The distribution of dependencies is shown in table 2.

[3] Because of a minor mismatches in the annotation schemas and because of the absence of sentence-level annotations

Dependency	Description	Number of occurrences	Proportion
MOD	modifier	2707	27%
COMP	complementizer	1723	17%
DET	determiner	1346	13%
SUJ	subject	1073	11%
OBJ	object	843	8%
DISFL	disfluency	580	6%
CC	coordination	495	5%
POBJ	prepositional object	312	3%
ATTS	subject attribute	198	2%
JUXT	juxtaposition	180	2%
MultiMots	multi-word expression	179	2%
AUX	auxiliary	161	2%
DUMMY	empty dependent	91	<1%
REF	reflexive pronoun	75	<1%
APPOS	apposition	62	<1%
ATTO	object attribute	18	<1%
	total:	10043	

Table 2: Distribution of the different types of dependencies in the ETB corpus

3.2 Software environment

The J-Safran platform was developed with the aim to facilitate the iterative annotation procedure described in the previous section. One important motivation for developing yet another annotation platform was the need for easy use, installation and portability. Because the annotators were linguistics students working from home, it was necessary to have a platform that could be easily installed and used under different operating systems. Another important motivation was the need for easy modification and extension. For instance, we recently extended J-Safran to support joint syntactic-semantic annotation in view of adding semantic role labels and training a semantic role labeller for French.

Implemented in Java and available in open source on the web[4], J-Safran (Java Syntaxico-semantic French Analyser) integrates the following modules:

- The Malt Parser: a deterministic shift-reduce parser with a machine learning approach for computing local decisions and actions. The version used in this work exploits a Support Vector Machine (SVM) for this purpose, and integrates an interface module that facilitates the control of the parser from the Graphical User Interface (GUI).

- A part-of-speech (POS) tagger: the TreeTagger [9] with its associated Java Wrapper TT4J [5];

[4] http://www.loria.fr/~cerisara/jsafran/index.html
[5] http://www.annolab.org/tt4j

- The evaluation scripts derived from the standard CoNLL evaluation campaign [10].
- A Java GUI that provides most common vizualisation, editing and search functionalities for dependency annotations.

A screenshot of the J-Safran GUI is shown in Figure 1.

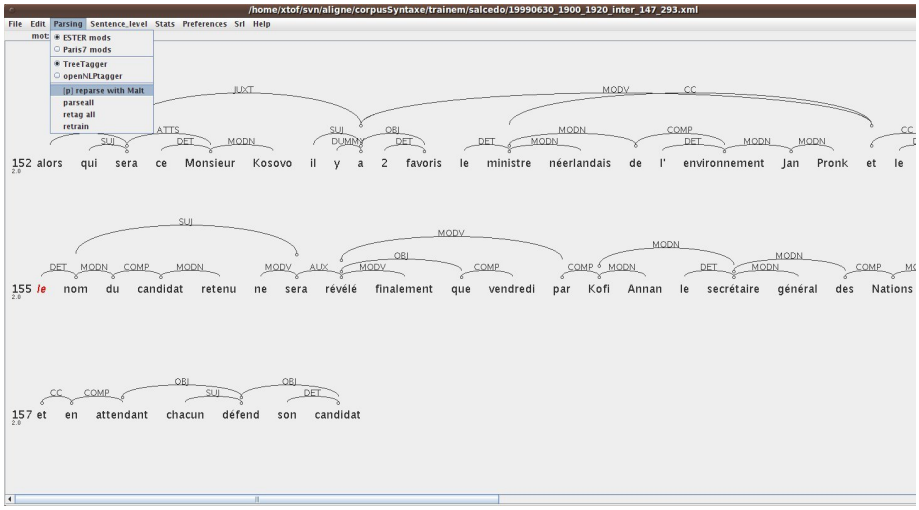

Figure 1: Screenshot of the J-Safran GUI for dependency tree edition

4 The impact of spoken data on parsing accuracy

We used the ESTER treebank, annotated in dependencies, to train and test the Malt parser. Using 8544 words for training and 1747 words for testing, we obtained a LAS (labelled attachment score i.e., percentage of tokens with correct predictor governor and dependency type) of 63.6% . Unsurprisingly, training a parser on such a small quantity of annotated speech transcripts yields results well below the state of the art both for written and for spoken data. Training on a larger speech corpus would obviously help improve performance. However, syntactic annotation is costly, and it would be both useful and interesting to have a better understanding of which phenomena in speech most affect parsing performance. Such an understanding could be used for instance, either to manually enrich the training corpus with annotated data for these phenomena or, within an active learning approach, to guide the automated selection of the data to be annotated.

4.1 Impact of disfluencies

We start by investigating the impact of disfluencies. A characteristic feature of spoken language is that, because there is no possibility of deleting what has been said,

all editing is performed online so that utterances often contain disfluencies i.e., false start, filled pauses, word fragments, repetitions, corrections and interruptions. To determine the impact of disfluencies on parsing, we manually remove disfluencies in the test corpus so as to isolate the impact of disfluencies from other factors, such as sentence length. Indeed, preliminary experiments indicate that disfluencies occur more frequently in longer sentences. The results are shown in Table 3. On disfluent sentences, disfluencies degrade the Labeled Attachment Score (LAS) by 4.1 points. The two other CoNLL scores given are the Unlabeled Attachment Score (UAS) and the Label Accuracy score (LAC).

	W/o disfl	With disfl	Δ(w,w/o)
LAS	70.2%	66.1%	+4.1
UAS	77.2%	73.5%	+3.7
LAC	76.5%	72.7%	+3.8

Table 3: Comparing parsing scores on test sentences with disfluencies (*with*) and after manual deletion of disfluencies (*without*). Only that part of the test corpus that contains disfluencies is considered here.

In the test corpus, 41% of the sentences are sentences with disfluencies and 59% sentences without disfluencies. To evaluate the impact of disfluencies on the whole test corpus, we compare the parsing scores on the raw test corpus and on the same corpus after manual correction of disfluencies. The results are shown in Table 4. As expected, the impact of disfluencies is qualitatively the same than in the previous test, but it is quantitavely lower, because the majority of sentences do not have disfluencies at all in this corpus. This experiment gives a better idea of the actual, effective impact of disfluencies on parsing in real conditions. Roughly, disfluencies account for a decrease in performance of 1.6 points.

	W/o disfl	With disfl	Δ(w,w/o)
LAS	67.3%	65.7%	+1.6
UAS	74.2%	73.0%	+1.2
LAC	74.2%	72.6%	+1.6

Table 4: Parsing scores on test sentences with disfluencies (*with*) and after manual deletion of disfluencies (*without*). Scores are computed on the whole test corpus.

4.2 Impact of speaking style

A marked characteristics of the ETB corpus, and more generally of broadcast news, is that it mixes professional radio speech with freer, less prepared, guest speech in interviews. While the utterances of radio announcers are prepared utterances

spoken by professional journalists, guest utterances are less polished and embody unplanned speech, closer to an every day setting.

To evaluate the impact of the journalist/guest style difference, we additionally annotated each utterance with an annotation indicating whether the utterance was that of the radio announcer or of a guest. 72% of the corpus is tagged as journalist style, and 28% as guest style. For the next experiment, 60% of the whole corpus is reserved for training, and 40% for testing. This train/test division is carefully made so that the global proportion of journalist/guest speech is preserved in both the training and test corpus. The test corpus is further split into two smaller test corpora, containing respectively only journalist and guest speaker utterances.

	Journalist	Guest	Δ(S,G)
LAS	70.8%	65.2%	-5.6
UAS	76.5%	71.8%	-4.7
LAC	77.5%	72.0%	-5.5

Table 5: Parsing scores of guest and journalist utterances with disfluencies kept.

As expected, the professional radio speaker style is easier to parse. Obviously though, disfluencies are more frequent in guest than in speaker speech. In order to remove the effect of disfluencies, and so only evaluate the impact of the other speaking style factors (lexicon used, syntactic structures, etc.), we manually fix the disfluencies in both parts of the corpus (speaker and guest) and redo the experiment.

	Speaker	Guest	Δ(S,G)
LAS	71.2%	67.8%	-3.4
UAS	77.2%	74.1%	-3.1
LAC	78.2%	74.5%	-3.7

Table 6: Parsing scores for guest and journalist styles with disfluencies removed.

Two remarks can be made. First, correcting disfluencies improves parsing more on guest speech (+2.6%) than on radio journalist speech (+0.4%), which conforms to intuition, because disfluencies are much more frequent in guest speech than in prepared speech. Second, there is still a significant difference in parsing performance between both styles, which is not due to disfluencies but results from other factors, most probably the lexicon and syntactic patterns typical of spontaneous speech. More precisely, disfluencies explain about 40% of the additional parsing errors in guest speech, while these other factors explain 60% of these errors.

4.3 Impact of headers

Finally, we consider the impact on parsing of a construct typical of radio announcers, namely headline utterances (headers) which structure the news by preparing or announcing the forthcoming subject, often in a telgraphic style with missing functional elements.

This next experiment is realized in 10 fold-cross-validation because there are few "headers" in the ETB (14% of all utterances). Guest utterances are removed and training is performed on journalist utterances including both headers and normal utterances. Two tests are then realized, one on all journalist utterances and the other only on headers. This test shows that headers are much more difficult to parse, which is probably due to the unbalanced training corpus, which largely favors common, non-header utterances. This suggests that parsing of radio speech could be improved by training models that are dedicated to parsing headers and explicitly detecting this speaking style. Such an experiment, however, would require collecting enough examples of header, and thus recognizing them automatically based on their general patterns.

	Normal	Headers	Δ(-H,+H)
LAS	70.6%	61.7%	-8.9
UAS	76.2%	69.7%	-6.5
LAC	77.4%	67.5%	-9.9

Table 7: Comparing performance on headers vs. common speaker utterances.

Although we have shown that headers have a clear impact within the radio speaker style, it is worth noting that their impact on the global baseline performances is not significant, because the relative proportion of headers is quite low.

5 Conclusion

One mid-term objective of the work presented here is the study and comparison of the impact on parsing accuracy, of specific oral constructs in broadcast news. Note however that for now, the finalized treebank is quite small and the conclusions derived from experimental results should be interpreted with care. We distinguish between the prepared speech of professional journalists and the more spontaneous speech of guest speakers, and we show that there are about 20% more parsing errors in the latter than in the former. Obviously, disfluencies occur more frequently in the more spontaneous speech, but interestingly the data shows that disfluencies only account for 40% of these additional errors. This suggests that, in order to improve spontaneous speech parsing, it is not sufficient to treat only disfluencies. Phenomena typical of this kind of speech, such as dislocations and ellipses also need to be considered and better handled. Furthermore, within the subcorpus composed of professional journalist utterances, we distinguish between normal journalistic speech and "header" constructs, whose purpose is to manage and structure the dialog and radio transitions. We show that there are about 30% more parsing errors in header constructs and that these additional errors are mainly due to two factors: the specific structures of these headers, which are often non-verbal utterances with several juxtapositions, and their relatively low number of occurrences in the corpus. Yet, it does not seem superfluous to specifically detect and improve

the analysis of these relatively rare segments. Indeed, headers play an important structuring function in radio news and adequately parsing them might markedly benefit interpretation. They could be used, for instance, to infer the identity of the next or of the previous speaker in the audio stream [5].

References

[1] A. Abeille, L. Clément, and F. Toussenel. *Building a treebank for French*. Kluwer, Dordrecht, 2003.

[2] E. Bick. Parsing and evaluating the french europarl corpus. In *Méthodes et Outils pour L'évaluation Des Analyseurs Syntaxiques (Journée ATALA)*, pages 4–9, Paris, May 2004.

[3] Didier Bourigault. Un analyseur syntaxique opérationnel : Syntex. Mémoire d'habilitation, Université Toulouse-Le Mirail, June 2007.

[4] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[5] V. Jousse, S. Petitrenaud, S. Meignier, Y. Estève, and C. Jacquin. Automatic named identification of speakers using diarization and ASR systems. In *Proc. of ICASSP*, 2009.

[6] Candito M.-H., Crabbé B., and Denis P. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of LREC'2010*, La Valletta, Malta, 2010.

[7] M. Meteer, R. Taylor, and R. Iver MacIntyre. Dysfluency annotation stylebook for the switchboard corpus. Technical report, Distributed by LDC, 1995.

[8] Joakim Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*, pages 149–160, 2003.

[9] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *Proc. of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, 1995.

[10] Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, pages 159–177, Manchester, United Kingdom, 2008.