Catogorizing syntactic chunks for marking disfluent speech in French language

Klim Peshkov, Laurent Prévot, Stéphane Rauzy & Berthille Pallaud

[1] Aix-en-Provence Université, Laboratoire Parole et Langage,
5 avenue Pasteur, Aix-en-Provence, France

Abstract

Disfluency is the first phenomenon one has to address when processing spontaneous speech. Efficient systems combining transcription-based and signal-based cues have been created for English. These systems generally use supervised machine learning models, trained over large annotated datasets combining signal and transcription. As for other languages, including French, the situation is complicated by the lack of resources. A few proposals based on filled pauses, truncated words and repetitions have been made for identifying disfluencies in French. In this paper, we propose a transcription-based approach to this task, with high-quality morpho-syntactic tags as input for identifying disfluent areas. Originally, we adopted a transcription-based approach for obtaining an independent way of characterizing disfluencies. This can be later compared and combined with prosodic cues. Our method consists in building syntactic chunks from our tagging and then classify these chunks into several categories, some of them being considered as disfluent. We apply our method to speaker style characterization, discourse genres zoning, as well as to dataset cleaning. Finally, an attempt is made to relate our disfluent chunks to a more standard description of disfluencies in order to open the way of a deeper integration of our work with the one of the *disfluency* community.

Index Terms: tagging, chunking, transcription-based approach, disfluencies, speaking style.

1. Introduction

Disfluencies are not our primary research objective, but we are facing them in various aspects of our research such as syntactic parsing of spontaneous speech, prosodic phrasing or discourse segmentation [1, 2].

There has been a substantial number of works on disfluency in English. First systematic study of the phenomenon was performed by Shriberg [3]. Recently, Besser provided a more fine-grained classification system for the disfluencies [4]. Virtually all disfluency detection systems make use of statistical modeling approaches, although some authors combine machine learning techniques with rule-based approaches [5, 6].

As for the knowledge sources, most of the studies work either exclusively on transcripts (or speech recognition outputs) [7, 8, 9, 10] or on multiple knowledge sources [6, 11, 5]. However, some authors have tested prosody-only approaches [12, 6]. In the comparison of disfluency detection systems presented in [6], the system based only on text-based cues outperforms prosody-only system, while the combination of prosody and text performs even better.

Works on disfluency in French are less numerous. A number of descriptive studies exist [13, 14]. Automatic detection of disfluencies applied to a highly-specialized domain of aeronautical communication was addressed by [15].

We do not offer here a new account on this well studied phenomenon. We are only trying to use all the information we have at the transcription level to identify disfluent areas.

Concerning the data, we work on a conversational corpus of long conversations (1 hour) in which two speakers tell each other personal stories [16]. Given the nature of the task, the fact that the pairs of interlocutors are good friends and the duration of the recording, this corpus is heavily loaded with disfluencies.

These disfluencies complicate most of the basic tasks we perform on this corpus, including, in particular, syntactic parsing, prosodic phrasing, and discourse segmentation. However, we have a high-quality morpho-syntactic annotation. The paper is an exploration of how close we can get to the disfluencies using only transcription-based sources of information, but integrating richer information than filled pauses, truncated words and repetitions as in [14]. The reason of not using signal-based cues is twofold. First, we would prefer to have an orthogonal characterization which we could later compare and combine with prosodic cues. Second, at this stage we do have a high quality transcription and tagging for the whole corpus while our prosodic analyses are partial and not as reliable. Therefore, for the time being, we prefer to rely exclusively on the transcription. We believe that it is a frequent scenario in the development of spoken language resources.

The paper is structured as follows. We will start by presenting how our chunks are created in section 2. Then, in section 3 we will propose a classification of these chunks that we consider to be interesting for approaching disfluent speech. Section 4 will apply our categorization to several sub-tasks related to disfluent speech. Finally before concluding, we will discuss the relations between our analyses and more standard accounts of disfluencies (section 5).

2. Building chunks

2.1. Morpho-syntactic tagging

The morpho-syntactic tags of the transcription have been obtained in three steps. In a first step, the enriched orthographic transcription has been filtered of annotations not containing syntactic content (filled pause, hesitation, truncation, laughter, ...). This filtered input was then proposed to a tagger [1]. This tagger is a stochastic tagger trained on written French texts. The morpho-syntactic information has been organized in an ad-hoc way in 50 tags. On this tagset, the performance of our tagger for written French is good (a F-measure score of 0.975 is obtained for the tagger output, version 2011). The tagger was slightly modified to account for the absence of punctuation marks in the input transcription. It was therefore allowed to the tagger to insert punctuation marks when appropriate (i.e. when this insertion increases the probability of the sequence of tags treated).

In a second step, an error analysis was performed on the output tags. Unknown words (i.e. the words not present in our initial lexicon) were listed and added to a lexicon specific to spontaneous speech. In our french corpus, the

more frequent phenomena are word reductions (e.g. *appart* for *appartment*), regional version or foreign words, and onomatopoeia. By comparing the lexical frequencies of our corpus versus their written French counterparts, we also established a list of potentially problematic words for the tagging task. Among those, oral discourse markers (e.g. *quoi*, (*what*), *enfin* (*in the end*), *bon* (*well*), *en fait* (*in fact*),…) were identified and their entries in the lexicon were modified. We associated the morpho-syntactic tag *Interjection* when these words are used with their discourse marker function.[1] This last modification concerns almost 10% of the tokens of the whole corpus.

In a final step, the new version of the tagger was applied to the filtered transcription inputs. From this new tagged output, a manual correction was performed on the 115, 000 tokens of the whole corpus. This manually corrected version was afterwards used as a gold standard to evaluate the performance of our tagger adapted for spontaneous speech transcription input. We obtained a F-measure of 0.948 which is a very good score, considering that no special treatment had been applied for dealing with disfluency phenomena.

2.2. Creating chunks

Analysis in chunks is an easy-to-implement and robust method for shallow syntactic analysis [17]. The main principle of chunking consists in including in one unit all the constituents situated to the left of each syntactic head. These units may be helpful for the detection of disfluencies in French, because some recurrent kinds of disfluencies will give rise to chunks with unusual morpho-syntactic patterns. We would like to stress that we do not plan to capture all kinds of disfluences using chunks, but only those that give rise to local perturbations of syntactic structure. For example, a very common disfuency in our corpus is a sequence of function words (1) or discourse markers (2) before a lexical word.

(1) les les les gens
the the the people

(2) ouais ben ouais je crois
yeah well yeah I thinks

Chunking our data is performed by a script using 26 rules. The rules are of two types. The first type specifies tags which are always added in the same chunk as the following token as illustrated in (3). Most of the function words belong to this category.

(3) determiner (*D*) + anything
preposition (*S*) + anything
conjunction (*C*) + anything
personal pronoun (*Pp*) + anything
auxiliary verb (*Va*)+ anything

The second type of rules specifies an ordered pair of POS tags which must be in the same chunk (4).

(4) adjective (*A*) + noun (*N*)
demonstrative pronoun (*Pd*) + verb (*V*)
adverb (*R*)+ adjective (*A*)
proper noun (*Np*) + proper noun (*Np*)
verb (*V*) + verb (*V*)

Any number of filled pauses and truncated words can appear inside a chunk if the rules specify that the last word before these elements must be in the same chunk as the first word after them. This means that the sequence *"Determiner (D) – filled pause (F P) – filled pause (F P) – Noun (N)"* will give rise to a single chunk: *DN* . Otherwise, they are attached to the beginning of the next chunk. The same approach is applied to the treatment of pauses inferior to 200 milliseconds. Chunks cannot span across pauses which length is above this threshold.

3. Categorizing chunks

3.1. Comparison spoken vs. written chunks

To classify our chunks we first compare them across corpora type by calculating their distribution both in spoken corpus and in a written corpus [18]. More precisely we computed a "spokenhood" ρ ratio for each chunk type as follows:

$$\rho_i = \frac{Spoken\ Frequency(i)}{Written\ Frequency(i)}$$

We split the list of chunks' types according to this ratio as illustrated in Table 1:

- $\rho \rightarrow \infty$: Appear in spoken data only, many disfluency related patterns should be there (DISFLUENT)
- $\rho > 17$:[2] Majority in spoken data (SPOKEN)
- $\rho < 17$: Majority in written data (CANONICAL)

Table 1: *Chunk comparison Spoken/Written*

Chunk	ρ	Example
I I I Pp V	∞	ouais ben ouais j' imagine yeah well yeah I imagine
I D N	109.75	bon ce truc well that thing
D A N	0.49	un vrai conflit a real conflict

3.2. Morpho-syntactic classification

The next step consisted in refining these categories based on the inner structure of the chunks. We started by simplifying the chunk tag set by applying a few rules (in the order presented below). Some of the rules are simplifications of some phrase structures, others are regular expression-style rules.

- C ⇝ I: assimilate conjunctions to interjection for the sake of simplicity since in this context they play a very similar role
- {V Vi}, {Va V} ⇝ V : simplification of verb structures, (*V* = Verb ; *Vi* = Infinitive verb ; *Va* = auxiliary verb);
- {Pp Px}, {Pp Po} ⇝ P: simplification of pronoun structures, (*Pp* = Personal Subject Pronoun ; *Px* = Reflexive pronoun ; *Po* = Personal Object Pronoun);
- {X X} , {XXX} ⇝ X+: simple regular expression patterning.

[1] The tag *Interjection* was found convenient because discourse markers, similarly to interjections, have no real syntactic function.

[2] The score of 17 has been decided after qualitative evaluation of the patterns.

Once these simplifications done, we manually tagged the 50 most frequent patterns of the SPOKEN and DISFLUENT categories and identified the following subcategories:

1. SPOKEN CANONICAL: It is a spoken form but there is no disfluency there (4.67% of the total number of chunks)
2. HESITATION: I+ Canonical, that is a sequence of discourse markers or conjunctions followed by a CANONI- CAL chunk (4.48%)
3. BC: Backchannels, defined as sequence of discourse markers, surrounded by silent pauses longer than 200 milliseconds (16.26%)
4. DM: Discourse markers [3]: I+ (6%)
5. INCOMPLETE: Chunks without a head (2.44%)
6. EXCESSIVE: Chunks with abnormally long patterns (0.93%)
7. RARE: Other chunks not belonging to any of the above categories. They are considered to be disfluent (0.03%)
8. FP/WF: This category is created automatically, including all chunks containing a filled pause or a word fragment (8.85%)

The morpho-syntactic analysis proposed that 12.24% of the chunks of our corpus were disfluent. For a comparison [3] reports 6,4% of disfluent tokens in the Switchboard corpus of spontaneous speech. It is quite comparable if we remember that our chunk are a little bigger than the tokens. Here we consider a chunk as disfluent if it belongs to one of the following categories: INCOMPLETE, EXCESSIVE, RARE or FP/WF. Non-inclusion of HESITATION, DM and BC into the definition of disfluency is motivated by the fact that from the perspective of our syntactic parser these categories do not break the syntactic structure.

3.3. Evaluation

The first evaluation we propose is a simple F-score measure against a manual annotation realized by one of the authors. The manual annotation was however at the token level while ours is at the chunk level. Therefore, for each proposed disfluent chunk we check whether the reference contains at least one disfluent token. The results are presented in Table 2.

Table 2: *Evaluation against manually annotated data.*

Version	Precision	Recall	F-score
Baseline (FP and WF)	72.5%	47.4%	57.3%
Morphosyntax alone	69.7%	56.2%	62.2%
Together	70.2%	57.7%	63.3%

Our baseline was obtained by taking *F P* and *W F* chunks (chunks including a filled pause or a word fragment) as disfluents.[4] The low recall when we used only morpho-syntactic cues, can be explained with the fact that, as we mentioned above, this method does not allow detecting disfluencies above the chunk level. Adding morpho-syntactic cues results in a significant improvement over the baseline. Moreover, the precision score is more important for our applications, such as removing disfluent zones from prosodic analyses. Indeed, we prefer not to have the entire set of disfluencies detected than to throw away too much clean data, mistakenly labeled as disfluent.

[3] DM includes discourse markers but also conjunctions because of one of our simplification rules.

[4] We also tried to include verbatim repetitions but it decreased the performance (drop of precision).

4. Applications

As said in the introduction, our research topic is not disfluencies *per se*, but rather their applications. In this section we present three applications relevant for our purposes.

4.1. Speaker style characterization

First of all, we are interested in speaker variability for a number of our studies. Although we have intuition about the fluency of all our speakers, we would like to be able to have several independent measures to describe their speaking style. (Dis)fluency is one of these dimensions. Such characterization will be used later for re-investigating results at different levels of analysis (in particular, phonetics, prosody). In Table 3 we present a comparison of our disfluency rate with an intuitive evaluation of the speakers performed by a linguist before our analysis[5] (*Flu-1* and *Flu-2* were characterized as fluent by the expert, *Dis-1* and *Dis-2* as disfluent). The distribution of canonical and disfluent chunks seems to coincide with the intuition of the expert. Therefore, our method may be useful to characterize speaker styles. However a multi-expert characterization and deeper statistical analyses are needed to confirm this.

Table 3: *Disfluency rate by speaker.*

Speaker	Flu-1	Flu-2	Dis-1	Dis-2	Avg
Canonical	60.7%	63.2%	49.8%	51.0%	55.9%
Disfluency	9.6%	6.8%	14.4%	11.4%	12.3%

4.2. Discourse activities segmentation

We also looked at the distribution of our categories across the discourse activities of the corpus (narrative sequences alternating with very open conversation). Our hypothesis was that narrative sequences are a little better planned and therefore should be less disfluent than the free conversation. However, we could not confirm this. Only a few individual speaker strategies seems to surface but not all in the same direction. Some speakers seem to be more disfluent while narrating while for some others it is the free conversation which is more disfluent. Indeed, a deeper qualitative assessment showed us that speakers have various strategies to deal while alternating these two discourse activities. The only tendency, clearly emerging from narrative versus non-narrative comparison, is the difference in BC category rate, which is obviously lower in the narratives: 0.75% in the narratives and 6.19% in the non-narrative zones.

5. From chunks to disfluencies

We would like to take advantage of our chunk categorization for performing a disfluency tagging more in line with standard accounts such as [3, 19, 4] . We present here structure of some detected disfluencies belonging to three big classes of disfluences from [4]: *Uncorrected*, *Deletable* disfluences and *Revision*.

Here in the Table 4, in the cases of *Uncorrected* and *Deletable* disfluencies, the system marked as disfluent the reparandum (RM), which is convenient for subsequent correction. The reparandum of the uncorrected disfluence is the chunk “sur”, labeled as INCOMPLETE.

[5] Unfortunately, only one expert has done this intuitive evaluation forbidding inter-coder agreement. However, the speakers presented in the table are the ones that, according to the expert, have a clear fluent / disfluent speech style.

The reparandum of the *Deletable* disfluence, "tout ça" belongs to DM category. However, sometimes the chunks are bigger than the disfluence and consequently some material before the reparandum and after the reparans (RS) can be marked as disfluent. The example of *Revision* is in fact a single chunk of the type EXCESSIVE.

Table 4: *Examples of detected disfluencies.*

Left context	RM	IM	RS	Right context
Uncorrected y a un truc there's something	sur about			(pause)
Deletable un peu stressée a little nervous	tout ça and all that			
Revision dans in	un a		une a	auberge hostel

6. Conclusion and future work

Although our results on disfluency marking are only preliminary, we believe that it is worth to develop our methodology further. First of all, there are some aspects we did not considered for this first study: repetitions and disfluent pauses. These categories are not very difficult to include in our approach. Overall, the contribution of the morpho-syntactic patterns seem thin, but our conclusion is that if one wants to go beyond pure (shallow) lexical approaches, our approach can be used to improve the results without including signal-based features.

A lot of the missed cases are not detected simply because some disfluencies occur on a longer stretches of text than chunk. One way to deal with this higher-level disfluencies is to detect lexical repetitions with a risk of introducing too much noise. But the repetition is not the only cue that can help in detecting them. Trying to use the syntactic information more efficiently, one can think of a slightly more complex improvement of the detection system. Each chunk has to be categorized according to the part of speech of its head into verbal chunks, noun chunks etc. After that we can try to define potentially disfluent sequences. Verbal chunks may be characterized with the information about verb valency. This would give us information about the well-formedness of the syntactic structure across the verbal chunk and its neighbours.

Finally, we would like to follow [10] and apply a Transformation-Based Learning approach [20] on our data and compare the results with the ones presented here.

7. References

[1] S. Rauzy and P. Blache, "Un point sur les outils du lpl pour l'analyse syntaxique du français", in *Workshop ATALA: Quels analyseurs syntaxiques pour le français ?*, Paris, France, 2009.

[2] K. Peshkov, L. Prévot, R. Bertrand, S. Rauzy, and P. Blache, "Quantitative experiments on prosodic and discourse units in the corpus of interactional data", in *Proceedings of SemDial 2012: The 16th Workshop on the Semantics and Pragmatics of Dialogue*, 2012.

[3] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, 1994.

[4] J. Besser and J. Alexandersson, "A comprehensive disfluency model for multi-party interaction," in *Proceedings of the 8st SIGdial Workshop on Discourse and Dialogue*, 2008, pp. 182–189.

[5] S. Germesin, T. Becker, and P. Poller, "Hybrid multi-step disfluency detection", *Machine Learning for Multimodal Interaction*, pp. 185–195, 2008.

[6] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge sources", in *Proceedings Eurospeech*, vol. 1. Geneva, Switzerland, pp. 957–960, 2003.

[7] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies", in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, pp. 405–408, 1996.

[8] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue", *Computational Linguistics*, vol. 25, no. 4, pp. 527–571, 1999.

[9] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech", in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, pp. 1–9, 2001.

[10] M. Snover, B. Dorr, and R. Schwartz, "A lexically-driven algorithm for disfluency detection", in *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, pp. 157–160, 2004.

[11] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, "Comparing hmm, maximum entropy, and conditional random fields for disfluency detection", in *Proceedings of the INTERSPEECH '05*. Citeseer, 2005.

[12] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. Eurospeech*, vol. 5, 1997, p. 2383–2386.

[13] S. Henry and B. Pallaud, "Word fragments and repeats in spontaneous spoken French", in *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*, pp. 77–80, 2003.

[14] P. B. d. Mareüil, B. Habert, F. Bénard, M. Adda-Decker, C. Barras, G. Adda, and P. Paroubek, "A quantitative study of disfluencies in French broadcast interviews," in *Disfluency in Spontaneous Speech*, pp. 27–32, 2005.

[15] J.-L. M. Bouraoui, "Analyse, modélisation, et détection automatique des disfluences dans le dialogue oral spontané contraint: le cas du contrôle aérien", Ph.D. dissertation, Université Paul Sabatier-Toulouse III, 2008.

[16] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID – corpus of interactional data – annotation et exploitation multimodale de parole conversationnelle," *Traitement Automatique des Langues*, vol. 49, no. 3, pp. 1–30, 2008.

[17] S. Abney, "Parsing by chunks", *Principle-based parsing*, vol. 44, pp. 257–278, 1991.

[18] G. Adda, J. Mariani, P. Paroubek, M. Rajman, and J. Lecomte, "L'action GRACE d'évaluation de l'assignation des parties du discours pour le français", *Langues*, vol. 2, no. 2, pp. 119–129, 1999.

[19] E. E. Shriberg, "Phonetic consequences of speech disfluency," DTIC Document, Tech. Rep., 1999.

[20] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging", *Computational Linguistics*, vol. 21, no. 4, pp. 543–565, 1995.