HAL
archives-ouvertes.fr

ODIL Syntax : a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees

Ilaine Wang, Aurore Pelletier, Jean-Yves Antoine, Anaïs Halftermeyer

▶ **To cite this version:**

Ilaine Wang, Aurore Pelletier, Jean-Yves Antoine, Anaïs Halftermeyer. ODIL Syntax : a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees. Language Resources and Evaluation Conference, LREC, May 2020, Marseille, France. hal-02523141

HAL Id: hal-02523141
https://hal.archives-ouvertes.fr/hal-02523141

Submitted on 28 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ODIL_Syntax : a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees

Ilaine Wang[1,2], Aurore Pelletier[2], Jean-Yves Antoine[1], Anaïs Halftermeyer[2]

[1]LIFAT, Université de Tours, France
[2]LIFO, Université d'Orléans, France

{ilaine.wang, anais.halftermeyer}@univ-orleans.fr,
jean-yves.antoine@univ-tours.fr

Abstract

This paper describes *ODIL_Syntax*, a French treebank built on spontaneous speech transcripts. The syntactic structure of every speech turn is represented by constituent trees, through a procedure which combines an automatic annotation provided by a parser (here, the Stanford Parser) and a manual revision. *ODIL_Syntax* respects the annotation scheme designed for the *French TreeBank* (*FTB*), with the addition of some annotation guidelines that aims at representing specific features of the spoken language such as speech disfluencies. The corpus will be freely distributed by January 2020 under a Creative Commons licence. It will ground a further semantic enrichment dedicated to the representation of temporal entities and temporal relations, as a second phase of the ODIL@Temporal project. The paper details the annotation scheme we followed with a emphasis on the representation of speech disfluencies. We then present the annotation procedure that was carried out on the Contemplata annotation platform. In the last section, we provide some distributional characteristics of the annotated corpus (POS distribution, multiword expressions).

Keywords: treebank, syntactic annotation, spontaneous speech, French language, constituent trees

1. Introduction

The development of large treebanks has provided useful training data for the development of robust and large coverage parsers. While the first treebanks relied on a constituency-based representation of syntactic structures, dependency treebanks have met for more than one decade an irrepressible surge of interest, due to a conjunction of practical and theoretical reasons:

- efficient dependency parsing algorithms (as well as the related machine learning techniques) have been proposed, leading to state-of-the-art performances;
- dependency trees are closer to predicate-argument structures than constituent trees, which prepares in an easier way further semantic-oriented tasks such as information extraction, question answering, etc.;
- dependency structures are assumed to fit better the representation needs of free word order languages, while constituency-based structures must struggle on such languages with serious discontinuity problems.

This explains that current treebanks are based on a dependency representation, as shown for instance by the *Universal Dependencies* initiative[1] or the CoNLL shared tasks[2] from 2006 to 2009. Following a different path, this paper presents a constituency treebank: *ODIL_Syntax*. This seemingly odd choice is motivated by specific annotation needs. Indeed, *ODIL_Syntax* constitutes the first annotation layer of a larger ressource, *ODIL_Temp*, which describes all the temporal entities and the temporal relations that can be found in the ODIL corpus. The temporal annotation of *ODIL_Temp* relies on an extension of the TimeML standard (ISO, 2012). The main originality of this extended annotation scheme is to delimit temporal mentions not by their minimal chunk, but by the range of the constituency subtree that covers the mention (Antoine et al., 2017).

This broader annotation is justified by theoretical reasons. For instance, the resolution of temporal abstract anaphora often needs the consideration of a whole clause or a whole speech turn (Zinsmeister and Dipper, 2010), which cannot be modelled by a lexical head-based annotation such as ISO TimeML.

From a practical point of view, the pilot experiments we conducted have shown that the cognitive load required by the manual annotation of temporality is reduced if the phrase-based structure of the utterances is displayed to the annotator. It seems that temporal annotation requires a syntactic disambiguation in most of the cases. The rationale behind the *ODIL_Syntax* treebank is precisely to provide such a resource.

Despite the specific purpose of the resource, we consider that the *ODIL_Syntax* treebank should present an interest for the NLP and the linguistics community:

- while spoken French is still poorly described by treebanks, the ODIL corpus focuses exclusively on spontaneous speech. The only spoken ressource that can be compared to *ODIL_Syntax* is *Rhapsodie* (Lacheret et al., 2014). *Rhapsodie* is a dependency treebank, conversely to *ODIL_Syntax*. As attempts have been made to convert constituency treebanks to dependency ones (Candito et al., 2010), we may say that both resources are complementary to a certain extent;

[1]UD: `https://universaldependencies.org/`
[2]CoNLL shared tasks: `https://www.conll.org/previous-tasks`

- the ODIL corpus follows the annotation scheme of the *French TreeBank* (*FTB*), the largest constituency treebank available for French (Abeillé et al., 2003; Abeillé and Barrier, 2004). *FTB* was built exclusively on written texts. *ODIL_Syntax* can be considered as an extension of the *FTB* to spoken language, which should be interesting for instance in terms of language model adaptation to spontaneous speech.

The next section of the paper presents in detail the annotation scheme underlying the building of the treebank. In Section 3, we describe the semi-automatic annotation procedure we followed. At last, we give some quantitative and qualitative results that give an account of the resource.

2. Annotation scheme

For compatibility reasons, *ODIL_Syntax* follows the annotation scheme of the *FTB*. Since this resource concerns only written French, annotation guidelines that describe the specific features of spontaneous speech had to be defined. Most of these additions are inspired by, or at least are compatible with the proposals made for the Rhapsodie project. This section describes the main additions we provide to the *FTB* annotation scheme.

2.1. False start

False starts are observed when the speaker suddenly suspends their speech turn, or starts another utterance with no regard to the overall syntactic coherence. Such situations result in the production of incomplete constituents. False starts are explicitly annotated in the *ODIL_Syntax* treebank to offer the possibility to avoid those noisy structures when using machine learning techniques. The annotation scheme considers the lowest subnode of the constituency tree that covers a phrase or clause that is clearly incomplete from a syntactical perspective. This node is labelled with the expected syntactic category of the incomplete constituent, preceded with a `$` specific mark of non-completion.
Consider the example in Fig. 1:

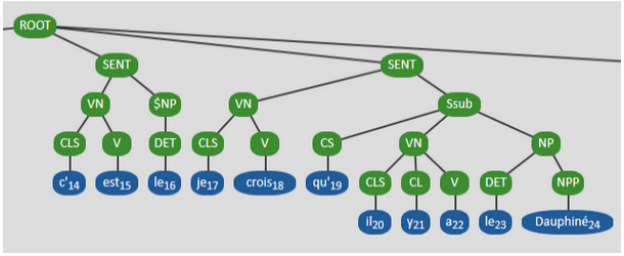

Figure 1: Example of annotation of a false start:
c'est le- je crois qu'il y a le Dauphiné
[transl. 'this is the- I guess there is the Dauphiné' [newspaper]]

In this example, the `$NP` label indicates explicitly that *le-* ('the-') begins a nominal phrase that was interrupted. This annotation corresponds to the `-INA` suffix used by (Abeillé and Crabbé, 2013) in their speech corpus (*ESTER 3*): `$NP` is strictly equivalent to `NP-INA`.

Similarly, as shown in Fig. 2, when a word is truncated but its POS can still be inferred, the `$` mark is used on the phrase level, not on the POS level.

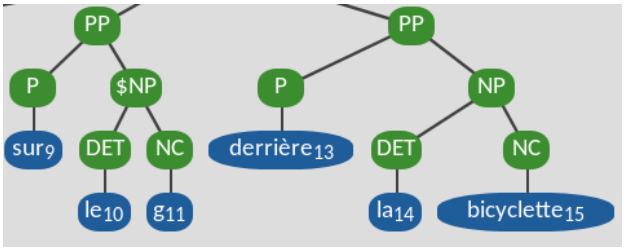

Figure 2: Example of annotation of a truncated word -
sur le g- derrière la bicyclette
[transl. 'on the g- behind the bicycle']

2.2. Repetition and self-repair

Repetitions and self-repairs are ordinary phenomena in spontaneous speech. They result in a syntagmatic accumulation, where several elements fill the same syntactic role in the speech turn, as shown in the following examples.

(1) *C'est c'est madame [Nom]* (repetition)
'This is this is Mrs. [Name]'

(2) *Dans deux minutes trois minutes* (self-repair)
'In two minutes three minutes'

The first elements are called *reparanda*, while the final one is the repair of the corresponding speech disfluency. For instance, in Example (2), 'two minutes' is the *reparandum* while 'three minutes' is the repair.
(Levelt, 1983) proposed a description of these disfluencies that relates to coordination, what has been considered as a too restrictive representation by many authors (Blanche-Benveniste, 1987). To highlight the specificity of self-repairs and repetitions, we propose to label them with a specific tag: `PARA` (which stands for *entassement PARAdigmatique* - in English, syntagmatic accumulation or pile).
Consider the repetition in Example (1) and illustrated in Fig. 3.

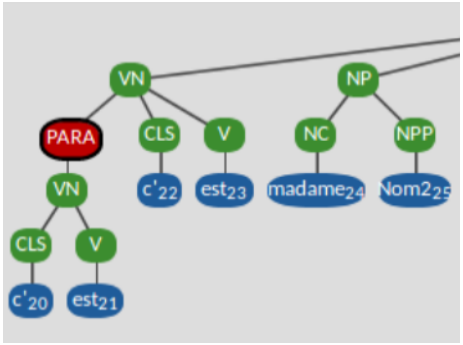

Figure 3: Example of annotation of a repetition -
C'est c'est madame [Nom]
[transl. 'In two minutes three minutes']

As shown in the figure, the *reparandum c'est* ('this is') is annotated as a VN (verbal nucleus). Then, the VN node is dominated by a PARA node to indicate the occurrence of a repetition. The *reparandum* appears finally as a dependent of the VN repair.
This representation is close to the one adopted by (Abeillé and Crabbé, 2013), where the *reparandum* is under a REP or a REV node, respectively to annotate a repetition of a self-repair (revision). However, considering that repetitions and self-repairs cannot be separated from a purely syntactic point of view (Blanche-Benveniste, 1987) because of their semantico-pragmatic nature, we do not make any distinction at this level. For the same reason, we do cover two of the seven types[3] of PARA dependency links (Kahane et al., 2019) considered in Rhapsodie, para_disfl (disfluency) and para_reform (reformulation), but under one single PARA label. This choice makes us a little closer to the Universal Dependencies' framework (version 2), which incorporates a *reparandum* dependency link to label any disfluency overridden in a speech repair.

2.3. Cleft structure, ellipsis and parenthesis

Cleft utterances (3) ellipses (4) and parenthesis (5) are very frequent in spontaneous speech.

(3) *C'est la rouge que j'ai achetée* (cleft utterance)
'It is the red one that I bought'

(4) *Moi aussi* (ellipsis)
'Me too'

(5) *Je l'ai vue je pense oui jeudi* (parenthesis)
'I saw her I guess yes on Thursday'

These structures are not specific to spoken language. As they are actually regularly annotated in the *FTB*, we chose to comply with their annotation scheme to represent them. For instance, the parenthesis *je pense oui* ('I guess yes') is ordinarily labelled as an internal sentence, e.g. it is covered by a subtree anchored by a Sint node.

2.4. Tagset

The tagset of *ODIL_Syntax* is the one used for the French model of the Stanford parser (Green et al., 2011). The full tagset is shown in Table 1 along with examples extracted from *ODIL_Syntax*. We note that there is no example shown for ADJWH, ET and PREF for the simple reason that there is no occurrence of interrogative adjective, foreign word nor prefix in our corpus.
This tagset is almost identical to the one used in (Crabbé and Candito, 2008) for their intermediate treebank. Their version of the *FTB* was created to train a syntactic parser, its tagset was enhanced with selected morphosyntactic features when those were highly discriminating. The only two differences with Crabbé and Candito lie in that contractions of prepositions and determiners (P+D) are simply annotated with the P preposition tag, and that there is an additional tag (CL) for clitics that are neither subject (CLS), nor object (CLO), nor reflective (CLR). It is typically the case with expletive (e.g. non referential) pronouns like in the idiomatic structure *il y a* ('there is'), as seen in the sentence displayed in Fig. 1.
The distribution of each tag is given in Section 4.2, and compared to the distribution of the *FTB*.

[3]The five other types do not concern structures that are specific to spoken language.

3. Annotation procedure

The annotation of *ODIL_Syntax* was carried out through an incremental process: the annotation is divided into five successive stages that combine automatic and manual procedures. All the annotation steps are conducted on Contemplata, a generic platform dedicated to treebank annotation that was developed during the ODIL project (Waszczuk et al., 2020). This incremental strategy relieves the cognitive load and has demonstrated its ability to limit the total workload of the annotation process, through appropriate calls of the automatic parsing. The five stages are described as follows:

1. Automatic pre-processing – the first step consists in a pre-processing of the speech transcripts to ease the parsing. We proceed in the sidelining of noises, interjections, and phatic expressions not carrying patent temporal information (for instance, *oui* ('yes'), *bonjour* ('good morning')) whereas verbal expressions like *excusez-moi* ('excuse-me') or *attendez* ('wait') are kept.

2. Preliminary automatic syntactic annotation – we use the constituency version (French model) of the Stanford parser to supply automatically the annotator with the syntactic structure of every speech turn.

3. Manual revision: utterance segmentation and POS tagging – the French model of the Stanford parser is trained on the *FTB*. Since the *FTB* contains exclusively written texts, the parser inevitably encounters difficulties to parse the spontaneous speech turns of our corpus. In particular, the concept of sentence, which is helpful to the parser, is not operative on spoken language. When the parser completely fails and when appropriate, the annotator is asked to divide the speech turns into several pseudo-sentences that correspond to fully independent clauses with their potential subordinate clauses (Fig. 4). The annotator is also requested to correct all the POS tags of the speech turns.

4. Final automatic syntactic annotation – the first manual revision phase appeared to be very helpful to the parser. This time, we use the parser to proceed to a second round of processing the speech turns but with a constraint: all of the revisions (segmentation and POS tagging) have to be taken into account and remain unchanged. Most of the time, the last revision is sufficient to provide quite satisfactory parse trees.

5. Constituent trees revision – during this last revision phase, the annotator is requested to correct all the residual errors that are found in the constituent trees.

POS	Tag(s)	Example(s)	POS	Tag(s)	Example(s)
Adjective	ADJ	*toute*, *effarant*	Common Noun	NC	*monsieur*, *gens*
Interrogative Adjective	ADJWH		Proper Noun	NPP	*Blanc*, *Loire*
Adverb	ADV	*pas*, *enfin*	Ponctuation	PUNC	*?*
Interrogative Adverb	ADVWH	*quand*, *combien*	Preposition	P	*au*, *à*
Conjunction	C	*parce*, *et*	Prefix	PREF	
Coordinating Conjunction	CC	*ou*, *c'est-à-dire*	Pronoun	PRO	*quelqu'un*, *elle*
Subordinative Conjunction	CS	*si*, *que*	Relative Pronoun	PROREL	*qui*, *lesquelles*
Clitic	CL	*y* (*il y a*)	Interrogative Pronoun	PROWH	*quoi*, *lesquels*
Object Clitic	CLO	*y*	Verb	V	*reviendrez*, *fait*
Reflexive Clitic	CLR	*nous*, *se*	Imperative Verb	VIMP	*écoutez*, *voyons*
Subject Clitic	CLS	*on*, *je*	Infinitive Verb	VINF	*emmener*, *mettre*
Determiner	DET	*quelques*, *les*	Past Participle	VPP	*bouleversée*, *fini*
Interrogative Determiner	DETWH	*quel*	Present Participle	VPR	*pédalant*, *appartenant*
Foreign Words	ET		Subjunctive Verb	VS	*aies*, *suive*
Interjection	I	*euh*, *oh*			

Table 1: POS tagset with examples from the corpus

It is noteworthy that a unique coder has conducted the annotation on the whole corpus, while three supervisors have checked the resulting annotation in a consensus-seeking procedure. We consider that this method guarantees the reliability of the annotation, despite the fact that the computation of some quality metrics such as inter-coder agreement is not operative here.

4. Results: corpus description

4.1. Source corpora: speech transcripts

Although it cannot be considered as a balanced corpus, the *ODIL_Syntax* treebank aims at representing a certain variety of language registers and dialogue situations. It is based on extracts of three corpora of speech transcripts that were built during previous research projects (Table 2). These corpora present different degrees of spontaneity and interactivity. The *ESLO* corpus is a collection of sociolinguistic interviews with a restricted interactivity (Eshkol-Taravella et al., 2011). Conversely, *OTG* and *Accueil_UBS* concern highly interactive Human-Human dialogues (Nicolas et al., 2002). These last two corpora differ by the media of interaction: direct conversation for the first and phone call for the latter. All of these corpora are freely distributed under a Creative Commons license (see Sec. 4.4).

The figures in Table 2 show the word count of the raw corpus, with 12,355 words (for a total duration of a little more than one hour). The sidelining of phatic expressions prior to the syntactic annotation reduced the number of words that are included in the treebank to 10,295.

4.2. POS distribution

As specified in Sec. 2.4, the tagset used for *ODIL_Syntax* is the one used for the French model of the Stanford Parser. Table 3 shows the distribution of the 28 tags

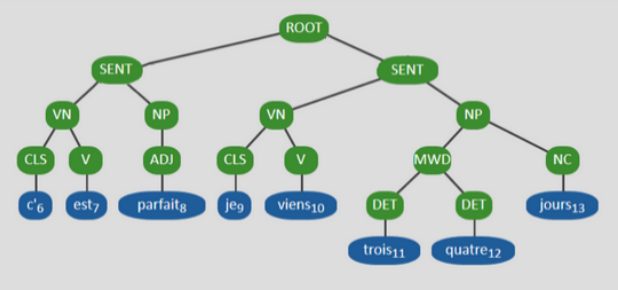

Figure 4: Example of a speech turn divided into several pseudo-sentences (SENT) after manual revision - *c'est parfait / je viens trois quatre jours* [transl. 'that's OK / I'll be there during three or four days']

Corpus	Speech type	Word count	No. of samples	Total duration
ESLO (extract)	Interview	9 663	3	47 min 31
OTG (extract)	Task-oriented conversational speech	705	2	2 min 30
Accueil UBS (extract)	Phone conversational speech	1 987	8	12 min 56

Table 2: Corpora used in ODIL_Syntax

POS tag(s)	ODIL	FTB
ADJ, ADJWH	3,65 % (372)	8,26 % (42 101)
ADV, ADVWH	9,70 % (988)	5,48 % (27 932)
C, CC, CS	7,08 % (721)	3,97 % (20 261)
CL, CLO, CLR, CLS	15,93 % (1 623)	3,20 % (16 321)
DET, DETWH	10,73 % (1 093)	17,15 % (87 449)
I	0,06 % (6)	0.01 % (70)
N, NC, NPP	17,05 % (1 737)	25,36 % (129 324)
P	10,86 % (1 107)	20,91 % (106 628)
PRO, PROREL, PROWH	4,75 % (484)	2,23 % (11 385)
V, VIMP, VINF, VPP, VPR, VS	20,20 % (2 058)	13,44 % (68 522)

Table 3: POS distribution grouped by categories: comparison between ODIL and the FTB

over *ODIL_Syntax* as a corpus. No substantial difference was observed between the samples of the three different sources. This leads us to say that the degree of interactivity of the spoken dialogues has little or no influence on the POS distribution.

On the other hand, sharing a tagset allowed us to compare *ODIL_Syntax* to the *FTB*, leading us to some observations on the differences between spoken and written French of journalistic genre. Unsurprisingly, clitic pronouns (CLS, CLO, CLR, CL) are predominant in our spoken corpus: they appear almost five times as much as in the *FTB* in terms of distribution. This can be easily explained by the predominance of the subject clitic *je* ('I'), which confirms Chafe and Danielewicz's observation (Chafe and Danielewicz, 1987) that oral dialogue is characterised by greater personal engagement (Biber, 1986).

We also note that prepositions, nouns and determiners are significantly more frequent in the *FTB*, whereas verbs are less frequent. This observation may result more from a difference of genre (journalistic genre for the *FTB*) rather than from a difference of modality (spoken vs. written language). Indeed, journalistic writing is characterised by more precise and elaborate utterances (Biber, 1986) with a potentially more frequent use of verbs and adjectives, while our spoken corpora tend to mobilise shorter sentences centred on verbs. In addition, written language is generally considered to be more explicit (Chafe, 1982), which could explain, again, these differences. Finally, the higher frequency of prepositions in the *FTB* is a clear indication of the higher syntactic complexity of most of written language genres (Chafe, 1982).

4.3. Multiword Expressions

Multiword expressions (MWEs) are groups of words whose meaning does not derive from the meaning of their components (no semantic compositionality) and/or from their syntactic structure (no syntactic compositionality) in a regular way. They are annotated MWX, with X being the category of the whole expression (for instance, *un peu* ('a little') is a MWADV, a MultiWord ADVerb).

However, the core purpose of *ODIL_Syntax* is to be a solid ground for semantic annotation, for which a detailed account of MWEs is unnecessary. This is why only MWEs which do not have a regular structure are annotated as such. This is the case of the MWADV *de plus en plus* ('more and more') in Fig. 5 (P + ADV + P + ADV). The underlying structure of the MWADV was kept as a flat subtree, just as in the *FTB* where compounds are represented as flat trees.

Conversely, in Fig. 6, *un numéro de code* ('a digit code') is not identified explicitly as a MWE. Its structure (DET + NC

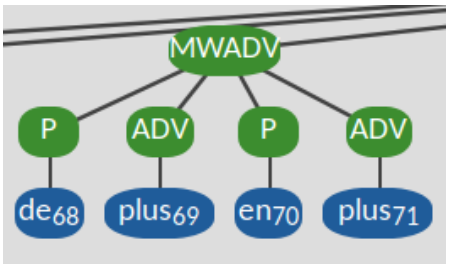

Figure 5: Example of a multiword adverb (MWADV) - *de plus en plus*
[transl. 'more and more', lit. 'of more in more']

+ PP) is not only a regular pattern for a NP but also a very productive pattern in French.

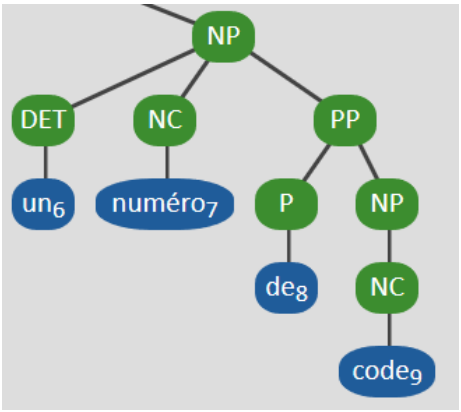

Figure 6: Example of a regular NP construction - *un numéro de code*
[transl. 'a digit code']

This explains why MWN are significantly less frequent in *ODIL_Syntax* compared to a corpus like the *FTB*: only 9 (3,6% of all MWEs) in *ODIL_Syntax* while *FTB* has 14401 (46,5%); most of them are in fact not explicitely annotated in our corpus. Consequently, *ODIL_Syntax* cannot be considered as a relevant resource for the study of multiword expressions.

4.4. Corpus License

ODIL_Syntax is freely distributed on ORTOLANG[4] under two different Creative Commons licenses: CC-BY-SA for OTG and Accueil_UBS samples, but CC-BY-SA-NC for samples that were extracted from ESLO.

5. Conclusion

We presented *ODIL_Syntax*, a free and useful resource for both linguists and NLP scientists who are interested in constituency syntax for spoken language, and more specifically in the adaptation of norms created for the *FTB* to spontaneous speech transcripts. Along with *ESTER 3*, this resource constitutes a first step towards scaling the *FTB* scheme for any type of spontaneous oral phenomena. Incidentally, our annotation choices are compatible: false starts are marked directly with a specific symbol ($ in our case, the suffix -INA in their case), repetition and revisions are labelled (respectively, PARA for us, REP and REV for them). The only notable difference is that (Abeillé and Crabbé, 2013) did annotate marks of hesitation such as *euh* ('er') and discourse markers (labelled HES and DM) while we chose to discard them from our syntactic trees as they are not needed for our subsequent semantic annotation.
Indeed, the production of this resource is rooted in the need to ground the annotation of semantic phenomena on a tree structure, as proposed in Temporal@ODIL (Lefeuvre-Halftermeyer et al., 2016). Our project requires the annotation of semantic (temporal) units (namely, Events, Signals and Timexes) within the framework of the ISO-TimeML standard on *ODIL_Syntax*. However, to ensure the reliability of ouf annotation, we decided to simplify the annotation conventions defined by ISO-TimeML. We are still working on this phase of the project and plan to deliver a corpus of smaller size by the end of 2020. This final version, called *ODIL_Temp*, will expose the articulation of syntax and semantics regarding temporal phenomena in natural language.

[4] https://www.ortolang.fr/market/corpora/odil

6. Acknowledgements

The ODIL project was founded by the council of the Centre Val de Loire Region (APR-IA).
We are also grateful to our colleagues Lotfi Abouda, Emmanuel Schang and Agata Savary for their precious help in not getting lost in the winding roads of the definition of guidelines for syntactic annotation in a spoken corpus.

7. Bibliographical References

Abeillé, A. and Barrier, N. (2004). Enriching a French treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, Lisbon, Portugal.

Abeillé, A. and Crabbé, B. (2013). Vers un treebank du français parlé. In *20ème conférence du Traitement Automatique du Langage Naturel (TALN'13)*, Sables d'Olonne, France.

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. *Treebanks. Text, Speech and Language Technology*, 20:165–187.

Antoine, J.-Y., Waszczuk, J., Lefeuvre-Halftermeyer, A., Abouda, L., Schang, E., and Savary, A. (2017). Temporal@ODIL Project: Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech. In *Thirteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-13), 12th International Conference on Computational Semantics (IWCS'2017)*, Montpellier, France.

Biber, D. (1986). Spoken and written textual dimensions in english: Resolving the contradictory findings. *Language*, 62(2):384–414.

Blanche-Benveniste, C. (1987). Syntaxe, choix de lexique, et lieux de bafouillage. *DRLAV. Documentation et Recherche en Linguistique Allemande Vincennes*, 36(1):123–157.

Candito, M., Crabbé, B., and Denis, P. (2010). Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, pages 1840–1847, La Valletta, Malta.

Chafe, W. and Danielewicz, J. (1987). Properties of Spoken and Written Language. In Rosalind Horowitz et al., editors, *Comprehending Oral and Written Language*, pages 83–113.

Chafe, W. (1982). Integration and Involvement in Speaking, Writing, and Oral Literature. In Deborah Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, pages 35–54.

Crabbé, B. and Candito, M. (2008). Expériences d'analyse syntaxique statistique du français. In *15ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'08*, pages 44-54, Avignon, France.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., and Tellier, I. (2011). Un grand corpus oral "disponible" : le corpus d'Orléans 1 1968-2012. *Traitement Automatique des Langues*, 53(2):17–46.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 725–735, Stroudsburg, PA, USA.

ISO. (2012). Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and Events. ISO 24617-1:2012. *International Organization for Standardization*.

Kahane, S., Pietrandrea, P., and Gerdes, K. (2019). The annotation of pile structures. In Anne Lacheret-Dujour, et al., editors, *Rhapsodie: A prosodic and syntactic treebank for spoken French*, pages 65–95. John Benjamins.

Lacheret, A., Kahane, S., Belião, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In *Proceedings of the 9rd International Conference on Language Resources and Evaluation (LREC'2014)*, pages 295-301, Reykjavik, Iceland.

Lefeuvre-Halftermeyer, A., Antoine, J.-Y., Couillault, A., Schang, E., Abouda, L., Savary, A., Maurel, D., Eshkol-Taravella, I., and Battistelli, D. (2016). Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'2016)*, Portoroz, Slovenia.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Nicolas, P., Letellier-Zarshenas, S., Schadle, I., Antoine, J.-Y., and Caelen, J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the "Parole Publique" project and its first realisations. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Canary Islands.

Waszczuk, J., Wang, I., Antoine, J.-Y., and Halftermeyer, A. (2020). Contemplata, a Free Platform for Constituency Treebank Annotation. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'2020)*.

Zinsmeister, H. and Dipper, S. (2010). Towards a Standard for Annotating Abstract Anaphora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, pages 54–59, Valletta, Malta.