What has LeBenchmark Learnt about French Syntax?

Zdravko Dugonjić,[1,2]* **Adrien Pupier,**[3] **Benjamin Lecouteux,**[3] **Maximin Coavoux**[3]
[1]LASR Lab, Technische Universität Dresden, Germany
[2]School of Embedded and Composite AI (SECAI)
first.last@tu-dresden.de
[3]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
first.last@univ-grenoble-alpes.fr

Abstract
The paper reports on a series of experiments aiming at probing LeBenchmark, a pretrained acoustic model trained on 7k hours of spoken French, for syntactic information. Pretrained acoustic models are increasingly used for downstream speech tasks such as automatic speech recognition, speech translation, spoken language understanding or speech parsing. They are trained on very low level information (the raw speech signal), and do not have explicit lexical knowledge. Despite that, they obtained reasonable results on tasks that requires higher level linguistic knowledge. As a result, an emerging question is whether these models encode syntactic information. We probe each representation layer of LeBenchmark for syntax, using the Orféo treebank, and observe that it has learnt some syntactic information. Our results show that syntactic information is more easily extractable from the middle layers of the network, after which a very sharp decrease is observed.

Keywords: Lebenchmark, wav2vec, probing, syntax, POS

1. Introduction

The analysis of large pretrained models have emerged as a Natural Language Processing (NLP) subfield aiming at understanding their inner workings, their strengths and weaknesses, as well as interpreting their predictions.

Probing (see Belinkov and Glass, 2019, for a general survey) consists in assessing whether some properties of a model's textual input can be predicted from the intermediate representations of the model. Probing has been first proposed under the names 'auxiliary prediction task' (Adi et al., 2017), or 'diagnostic classifiers' (Veldhoen et al., 2016), as a way to analyze deep learning systems, and in particular understand whether they implicitly learn some knowledge they were not trained on. For example, despite being trained on raw texts, the various layers of BERT (Devlin et al., 2019) contain a lot of information about the POS tags of its input (Tenney et al., 2019; Lin et al., 2019; Rogers et al., 2020), that can be extracted with a simple linear classifier.

In this paper, we apply probing to a pretrained **acoustic** model for French, LeBenchmark (Evain et al., 2021), and focus on assessing whether it has implicitly learned some syntactic knowledge. LeBenchmark is a Wav2vec2.0-style (Baevski et al., 2020) pretrained acoustic model, trained on the raw speech signal, i.e. very low level information. As a result, probing it for high-level information such as syntax is of important significance: syntactic information may be considered several abstractions away from the raw speech signal.

Specifically, we probe LeBenchmark for 2 tasks: part-of-speech tagging and unlabeled dependency parsing. We frame both tasks as sequence tagging tasks by reducing dependency parsing as a token-level task: predicting the relative position of the head of a token. We carry out probing using Orféo (Benzitoun et al., 2016), a treebank of spontaneous French spoken in realistic interactions. After probing each of the 24 layers of LeBenchmark, we found that syntactic information is most present in the middle layers of the model, and is much less accessible in the last layers where it seems to almost disappear.

In summary our contributions are as follows:

- we carry out a probing study on LeBenchmark for syntactic information. This is to the best of our knowledge the first study of this type (i) on French and more generally on a language that is not English (ii) on spontaneous speech (rather than read speech).
- we report on a finding: syntax is most extractable from the middle layers of the model and almost disappears in the final layers.

2. Related Work

Self-supervised learning (SSL) consists in acquiring robust representations from extensive unlabeled data (referred to as pretraining) in order to better recognize and understand patterns for other

*Work done during an internship at Laboratoire d'Informatique de Grenoble.

LREC-COLING 2024, pages 17493–17499

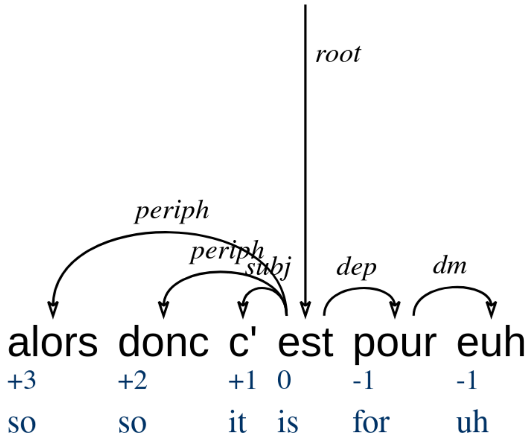

Figure 1: Illustration of the relative head position annotation scheme.

problems (referred to as fine-tuning). Recent research with a focus on speech data has demonstrated remarkable results in representation learning. On French Evain et al. (2021) trained and released self-supervised acoustic models called LeBenchmark to address a variety of speech tasks: spoken language understanding, speech-to-text translation, emotion recognition, speech recognition. From an acoustic signal, LeBenchmark computes a sequence of vectors, each of which corresponds roughly to a 25ms sound window. It was pretrained with a constrastive learning objective. No information about the notion of words or word boundaries is used during the pretraining.

The interest of the community in understanding what information was captured by these models grew recently. In particular, Singla et al. (2022) probe a pretrained acoustic model for English, but focus on audio features, the only syntactic features being the depth of the syntactic tree, and the number of occurrences of some parts of speech, where both of these are predicted from the whole sentence representation. In contrast, we focus on token-level syntactic information. Shen et al. (2023) is the study closest to our own: they probe English acoustic models for syntactic information (tree depth, correlations between the continuous representations of the model and the discrete tree representations). They found that syntax is best represented in the middle layers of networks, and more obviously in models with larger parameter sets. Unlike this work, we focus on French, *spontaneous* speech (rather than read speech), and use a different methodology: we focus on simple linear probes.

3. Data

We use the Orféo treebank (Benzitoun et al., 2016), a corpus of spoken French annotated in dependency trees, and distributed with audio recordings.[1] The Orféo treebank is an aggregation of multiple spoken French corpora, namely CFPP (CLESTHIA, 2018), Clapi (ICAR, 2017), TCOF (ATILF, 2020), OFROM (Avanzi et al., 2012-2020), Fleuron (André, 2016), French Oral Narrative (Carruthers, 2013), c-oral-rom (Cresti et al., 2004), *Corpus de référence du Français parlé* (DELIC et al., 2004), Valibel (Francard et al., 2009), TUFS (Kawaguchi et al., 2006), a professional meetings corpus (Husianycia, 2011), as well as an unpublished corpus provided by Orféo's designers. Most of the subcorpora contain French spoken in spontaneous interactions, except for French Oral Narrative that consists of stories read by narrators.

The total length of recordings is around 196 hours, among which 9 hours have gold syntactic annotations. The rest of the corpus was annotated with good quality silver syntactic trees (Nasr et al., 2020). The corpus has around 3.5 million tokens (among which 170k have gold syntactic information. Finally, the corpus is provided with token timecodes automatically predicted by a forced alignment system. In other words, we have the start and end time of each token. We discard all sentences that contain obvious timecode mistakes (e.g. when the start time of a token is higher than its end time) or annotation mistakes (e.g. POS tags that are not listed in the official documentation and correpond to typos in the manual annotation process).

4. Probing Tasks

We use two different tasks to assess the amount of syntactic information contained in the pre-trained representations: part-of-speech tagging and unlabelled dependency parsing. We cast both tasks as word-level prediction tasks. Though dependency parsing is not typically addressed with sequence tagging methods, recent approaches have shown that it is a viable method (Strzyz et al., 2019). In our case, addressing both tasks as word-level prediction tasks has the advantages of keeping the probes and the decoding algorithms fairly simple.

4.1. POS tagging

The part of speech tagging task is a classification task where the model classifies each word representation into 20 different parts of speech. The tagset is described by Benzitoun et al. (2016), It is somewhat finer-grained than the Universal Dependency (Nivre et al., 2020) tagset. The speech transcriptions in the Orféo treebank do not contain

[1] `https://www.ortolang.fr/market/corpora/cefc-orfeo`

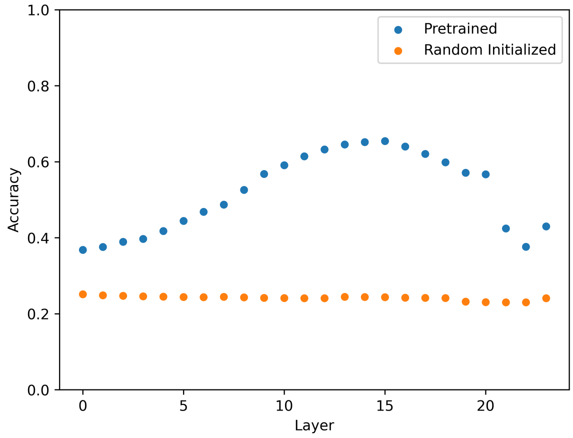

Figure 2: POS tagging task accuracy per layer.

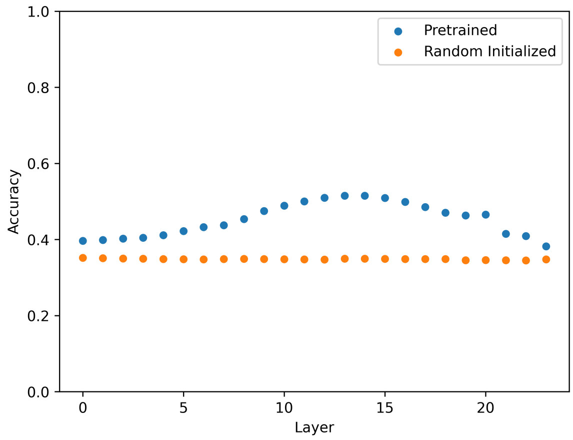

Figure 3: Relative head distance prediction task accuracy (UAS) per layer.

punctuation characters. Hence, the tagset does not include punctuation tags.

4.2. Unlabeled dependency parsing

In order to cast dependency parsing as a token classification task, we use a simple way of encoding a dependency tree into token labels: relative position encoding (Strzyz et al., 2019). With this encoding method, the label of a token is an integer corresponding to the relative position of its head word. This label is computed subtracting the index of the current word to the index of the head word. In the sentence illustrated in Figure 1 : "Alors donc c' est pour euh"[2] ("So so it is for uh"), the root of the sentence is "est" and is annotated with the relative label 0, the head of "Alors" is "est" and thus is annotated with the label +3 ("Alors" is the first word, "est" is the fourth one, $4 - 1 = 3$).

[2]Sentence from CEFC-Orféo id: `cefc-cfpb-1000-5-1`.

5. Experiments

This section presents the probes we use, outline experimental settings and discuss the experiment results.

5.1. Probes

Each layer of the pretrained acoustic model provides a vector representation for each acoustic *frame*. However, both probing tasks require *word* representations. In order to construct them, we rely on the *start* and *end* timecodes available for each token in the treebank (they were obtained automatically through forced alignement, and sometimes contain mistakes). We run the pretrained model (whose weights are frozen) on the speech signal, select the vector representations at a specific layer (1-24) and extract, for each token the list of corresponding frames (i.e. 1024-coefficient feature vectors) that are within the time span of the token. Then we aggregate the frames of a token into a single vector with a mean pooling operation.

Finally, we feed token representations to a simple Softmax classifier (one linear projection with bias followed by the Softmax activation). We repeat the process for each of the 24 layers of the pretrained model in order to probe each of them independently and analyse the dynamics of information across layers.

5.2. Experimental details

The self-supervised speech model that we probed in this study is LeBenchmark large[3] (Evain et al., 2021) based on wav2vec2 (Baevski et al., 2020) architecture, pre-trained on French datasets containing 7k hours of spontaneous, read, and broadcast speech.

We implemented the probing classifiers in Python, relying on PyTorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2020) libraries. All experiments and preprocessing steps are run on an Nvidia GeForce GTX 980 GPU.

We measured the performance of our probes in both tasks with the accuracy metric, which is a widespread measure in the probing literature. For the relative head distance prediction, the accuracy measure corresponds to the unlabeled attachment score (UAS), a metric classically used in dependency parsing.

During the training, we used a batch size of 1024. The classifier was trained using an early stopping mechanism, which stops the training process if there is no accuracy increase larger

[3]`https://huggingface.co/LeBenchmark/wav2vec2-FR-7K-large`

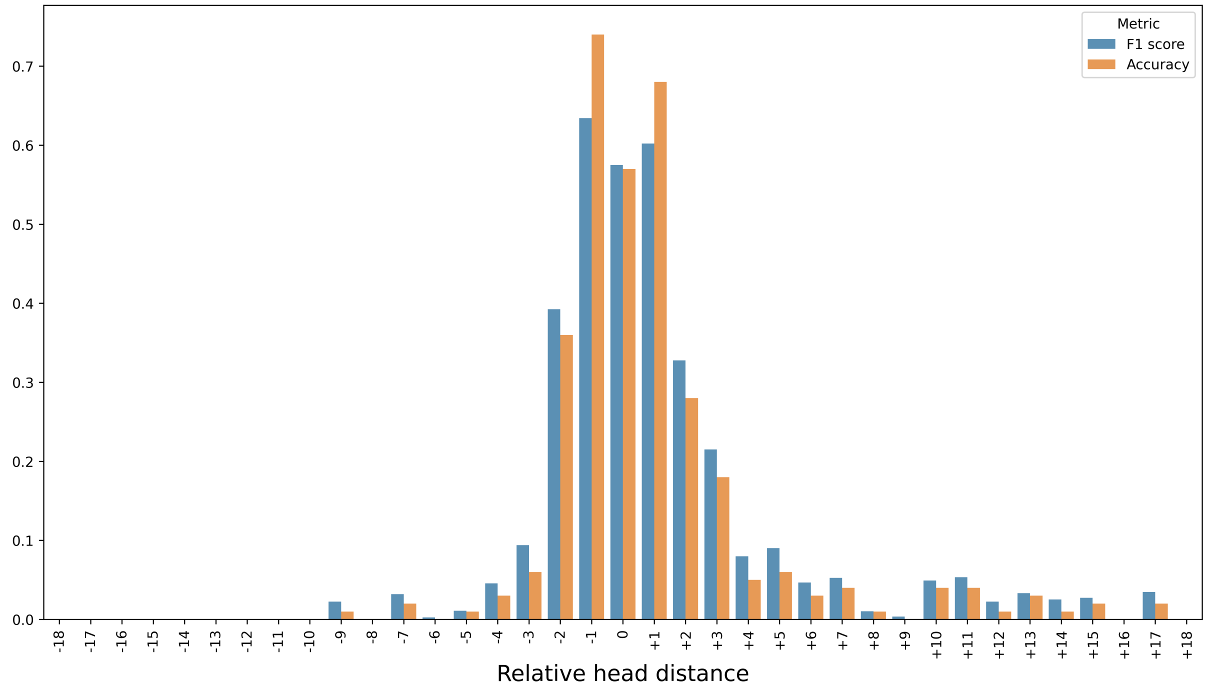

Figure 4: Per-category evaluation of the best layer (layer 14) on the relative head distance prediction task with two evaluation metrics: accuracy (UAS) and F-score.

than 0.0001 on the validation set over a span of 10 epochs.

For the POS tagging task, we use a learning rate of 0.005 and for the relative head distance prediction task we use a learning rate of 0.001, based on preliminary experiments. For both tasks, we optimize models with *stochastic gradient descent* (SGD) with Nesterov accelerator and momentum 0.99. The target optimization function is the negative log-likelihood of the gold labels. We use a a random 80%/10%/10% ratio split to construct the training, development and test sets.

Baseline As a control experiment, we use a baseline that consists of the exact same architecture as the wav2vec2 large model, but where trained parameters are replaced by randomly reinitialized parameters, a classical baseline in the probing literature (Arps et al., 2022).

5.3. Results

We present the accuracy of the probe on POS tagging in Figure 2, while taking as input each of the 24 transformer layers of the pretrained acoustic model. The accuracy ranges from 36.8% for the first layer (layer 0) to 65.5% for layer 15. The accuracy decreases after layer 15 and plummets for the last three layers. In contrast, the random initialisation baseline is about constant across all layers with an average accuracy of 24%, confirming that LeBenchmark encodes indeed POS information.

We observe a similar behavior on the unlabeled dependency parsing task as shown in Figure 3. The accuracy of the classifier reaches the peak of 52% at the 14th layer and then degrades towards the ends. The average random initialized baseline accuracy is 35%. The overall accuracy of the classifier on pre-trained representations is lower than in the POS tagging task, which is expected since the task is harder.

Our results are in line with those of Shen et al. (2023), who observed a similar pattern on English (using different models, data and methods). Compared to the literature on probing BERT models for syntax, a remarkable difference is that for both tasks, the accuracy for the final layers plummets to match the accuracy of the first layers (and even that of the baseline in the unlabeled dependency parsing task). In contrast, Hewitt and Manning (2019) also observe that the middle layers of BERT encode the most syntactic information, but the decrease after the middle layers is not as sharp as what we observe for LeBenchmark.

We now consider the results on the unlabeled dependency parsing task using the best layer as input (layer 14) and present its results in terms of accuracy and F_1 score broken down by label in Figure 4. As expected, we observe that the model fares better on the most frequent classes, i.e. -1, $+1$, and 0 (0 is the label for the root token of an utterance), with F_1 scores above 55%. Overall, we conclude that the model better encodes **local** syntactic information. However, it still has nonzero F_1 scores even on longer distance dependency, despite the corpus having no punctuation marks (whose attachement would be more easily predictable).

6. Conclusion

We have presented a series of experiments aiming at probing each layer of a French pretrained acoustic model (LeBenchmark) for 2 types of syntactic information: parts of speech, and unlabeled dependency arcs. We show that the wav2vec2 architecture encode some information about French syntax, in particular local attachments, despite having been pretrained only on raw speech signals. We show that the middle layers of the model are those from which syntactic information is the more easily extracted, a result in line with recently published results on English (Shen et al., 2023). Finally, the accuracy pattern across layers exhibits a sharp decrease in the last layers, a pattern that is not observed in BERT-style text-trained models (where the decrease is much softer).

7. Limitations

We acknowledge two limitations of our study. First, due to time constraints we performed all experiments with a single pretrained model, whereas using another model (LeBenchmark base 7k or a multilingual model) would have strengthen the robustness of our findings. Secondly, a limitation of the type of probe that we use is that the target information might be present in the representations but not extractable with a simple linear classifier.

8. Ethical Considerations

To the best of our knowledge, we do not see any potential ethical limitations of our work.

9. Acknowledgements

This work is part of the PROPICTO project (French acronym standing for PRojection du langage Oral vers des unités PICTOgraphiques), funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005). MC gratefully acknowledges the support of the French National Research Agency (grant ANR-23-CE23-0017-01). ZD gratefully acknowledges the support of BMBF in DAAD grant 57616814 (`https://secai.org/`).

10. Bibliographical References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for constituency structure in neural language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6738–6757, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. LeBenchmark: a reproducible framework for assessing self-supervised representation learning from speech. In *Proc. Interspeech 2021*, pages 1439–1443.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Alexis Nasr, Franck Dary, Frederic Béchet, and Benoit Fabre. 2020. Annotation syntaxique automatique de la partie orale du ORFÉO. In *Langages*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to syntax: Probing spoken language models for syntax. *arXiv preprint arXiv:2305.18957*.

Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah. 2022. What do audio transformers hear? probing their representations for language delivery & structure. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 910–925.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Viable dependency parsing as sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 717–723, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Sara Veldhoen, Dieuwke Hupkes, Willem H Zuidema, et al. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@ NIPS*, pages 69–77. Barcelona.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

11. Language Resource References

Virginie André. 2016. Fleuron: Français langue Étrangère universitaire–ressources et outils numériques.

ATILF. 2020. Tcof : Traitement de corpus oraux en français. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Mathieu Avanzi, Marie-José Béguelin, Gilles Corminboeuf, Federica Diémoz, and Laure Anne Johnsen. 2012-2020. Corpus ofrom – corpus oral de français de suisse romande. Université de Neuchâtel.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet ORFÉO: un corpus d'étude pour le français contemporain. *Corpus*, (15).

Janice Carruthers. 2013. French oral narrative corpus. Commissioning Body / Publisher: Oxford Text Archive.

CLESTHIA. 2018. Cfpp2000. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

Emanuela Cresti, Fernanda Bacelar do Nascimento, Antonio Moreno Sandoval, Jean Veronis, Philippe Martin, and Khalid Choukri. 2004. The C-ORAL-ROM CORPUS. a multilingual resource of spontaneous speech for Romance languages. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Equipe DELIC, Sandra Teston-Bonnard, and Jean Véronis. 2004. Présentation du corpus de référence du français parlé. *Recherches sur le français parlé*, 18:11–42. Equipe DELIC.

Michel Francard, Philippe Hambye, Anne-Catherine Simon, and Anne Dister. 2009. Du corpus à la banque de données.: Du son, des textes et des métadonnées. l'évolution de banque de données textuelles orales valibel (1989-2009). *Cahiers de l'Institut de linguistique de Louvain-CILL*, 33(2):113.

Magali Husianycia. 2011. *Caractérisation de types de discours dans des situations de travail*. Theses, Université Nancy 2.

ICAR. 2017. Clapi. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Yuji Kawaguchi, Susumu Zaima, and Toshihiro Takagaki, editors. 2006. *Spoken Language Corpus and Linguistic Informatics*. John Benjamins.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.