# Using ASR-Generated Text for Spoken Language Modeling

**Nicolas Hervé**[1], **Valentin Pelloin**[2], **Benoit Favre**[3], **Franck Dary**[3]
**Antoine Laurent**[2], **Sylvain Meignier**[2], **Laurent Besacier**[4]
[1]Institut National de l'Audiovisuel (INA), France
[2]Laboratoire d'Informatique de l'Université du Mans (LIUM), France
[3]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[4]Naver Labs Europe (NLE), Meylan, France
nherve@ina.fr

## Abstract

This papers aims at improving spoken language modeling (LM) using very large amount of automatically transcribed speech. We leverage the INA (French National Audiovisual Institute[1]) collection and obtain 19GB of text after applying ASR on 350,000 hours of diverse TV shows. From this, spoken language models are trained either by fine-tuning an existing LM (FlauBERT[2]) or through training a LM from scratch. The new models (FlauBERT-Oral) are shared with the community[3] and are evaluated not only in terms of word prediction accuracy but also for two downstream tasks: classification of TV shows and syntactic parsing of speech. Experimental results show that FlauBERT-Oral is better than its initial FlauBERT version demonstrating that, despite its inherent noisy nature, ASR-Generated text can be useful to improve spoken language modeling.

## 1 Introduction

Large language models are trained with massive texts which do not reflect well the specific aspects of spoken language. Hence, modeling spoken language is challenging as crawling 'oral-style' transcripts is a difficult task. To overcome this, our pilot study investigates the use of massive automatic speech recognition (ASR) generated text for spoken language modeling. We believe that this methodology could bring diversity (oral/spontaneous style, different topics) to the language modeling data. This might be also useful for languages with fewer text resources but potential high availability of speech recordings. We also see long-term benefits to using ASR generated text as speech recordings convey potentially useful metadata (ex: male/female speech) that could be leveraged for building LMs from more balanced data. Finally, as speech transcripts are naturally grounded with other modalities (if extracted from videos for instance), ASR could help building large scale multimodal language understanding corpora.

The contributions of this paper are the following:

- we build and share FlauBERT-Oral models from a massive amount (350,000 hours) of French TV shows,
- we evaluate them on word prediction (on both written and spoken corpora), automatic classification of TV shows and speech syntactic parsing,
- we demonstrate that ASR-Generated text can be useful for spoken LM.

## 2 Related Works

We mention here related works to better position our approach: learning LMs from spoken transcripts, multimodal models and using LMs to rescore ASR.

**Learning LMs from spoken transcripts.** Kumar et al. (2021) probes BERT based language models (BERT, RoBERTa) trained on spoken transcripts to investigate their ability to encode properties of spoken language. Their empirical results show that LM is surprisingly good at capturing conversational properties such as pause prediction and overtalk detection from lexical tokens. But their LMs evaluated are mostly trained on clean (non ASR) spoken transcripts except one called ASRoBERTa which is trained on 2000h of transcribed speech only (1k Librispeech + 1k proprietary dataset). As a comparison with this study, we train our models on 175x more ASR data.

**Multimodal models.** While our approach uses ASR to build text-based spoken language models, Chuang et al. (2019) proposed an audio-and-text jointly learned SpeechBERT model for spoken question answering task. They show their model

[1] https://www.ina.fr
[2] https://github.com/getalp/Flaubert
[3] https://huggingface.co/nherve

*Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 17 - 25
May 27, 2022 ©2022 Association for Computational Linguistics

is able to extract information out of audio data that is complementary to (noisy) ASR output text. The architecture proposed by Sundararaman et al. (2021) is different in the sense that it learns a joint language model with phoneme sequence and ASR transcript to learn phonetic-aware representations that are robust to ASR errors (not exactly a multimodal model). While speech or multimodal unsupervised representation learning is an interesting direction, this is out of the scope of this paper which focuses on language modeling from text transcripts only.

**BERT for ASR re-ranking.** We also mention here LMs to rescore ASR as this could be an interesting application of our proposed spoken language models. Chiu and Chen (2021) used BERT models for reranking of N-best hypotheses produced by automatic speech recognition (ASR). Their experiments on the AMI benchmark demonstrate the effectiveness of the approach in comparison to RNN-based re-ranking. A similar idea is introduced by Fohr and Illina (2021) where BERT features are added to the neural re-ranker used to rescore ASR hypotheses. Even more recently, Xu et al. (2022) showed how to train a BERT-based rescoring model to incorporate a discriminative loss into the fine-tuning step of deep bidirectional pretrained models for ASR.

# 3 From FlauBERT to FlauBERT-Oral

## 3.1 ASR system

The speech recognition system used to produce the text transcripts for this study was built using Kaldi (Povey et al., 2011). The acoustic model is based on the lattice-free MMI, so-called "chain" model (Povey et al., 2016). We used a time-delay neural network (Peddinti et al., 2015) and a discriminative training on the top of it using the state-level minimum Bayes risk (sMBR) criterion (Veselỳ et al., 2013).

For the acoustic model training, we used several TV and RADIO corpora (ESTER 1&2 (Galliano et al., 2009), REPERE (Giraudel et al., 2012) and VERA (Goryainova et al., 2014)). A regular backoff n-gram model was estimated using the speech transcripts augmented with several French newspapers (see section 4.2.3 in Deléglise et al. (2009)) using SRILM.

A 2-gram decoding is performed, followed by a 3-gram and a 4-gram rescoring step. The LM interpolation weights between the different data sources were optimized on the REPERE (Giraudel et al., 2012) development corpus. The vocabulary contains the 160k most frequents words in the manually transcribed train corpus. Automatic speech diarization of the INA collection was performed using the open source toolkit LIUMSpkDiarization (Meignier and Merlin, 2010).

Some results on different test corpora can be found in table 1.

| Corpus | WER |
|---|---|
| REPERE test corpus | 12.1 |
| ESTER1 test corpus | 8.8 |
| ESTER2 test corpus | 10.7 |

Table 1: ASR Performances on French TV or Radio corpora

## 3.2 Automatically transcribing 350,000 hours of the INA collection

The transcripts used in these experiments were taken from time slots corresponding to news programmes on French television and radio between 2013 and 2020. We transcribed the continuous news media between 6am and midnight each day (BFMTV, LCI, CNews, France 24, France Info and franceinfo). For radio, the morning news were used (Europe1, RMC, RTL, France Inter) and for generalist television channels we transcribed the evening news (TF1, France 2, France 3, M6). A total of 350,000 hours were automatically transcribed. The system we use provides us with raw text, without punctuation or capitalization. In order to have a pseudo sentence tokenization, we leverage the speaker diarization output to segment our transcriptions into "sentences". We end up with a total of 51M unique speech segments for a total of 3.5G words (19GB of data). The ASR generated text is strongly biased towards news content.

## 3.3 Fine-tuning or re-training FlauBERT-Oral

The initial French language model (**FBU**), trained in 2020 on natural text, is FlauBERT (Le et al., 2020). Models of different sizes were trained using masked language modeling (MLM) following a RoBERTa architecture (Liu et al., 2019) and using the CNRS Jean Zay supercomputer. They were shared on HuggingFace.[4] For comparison, these

[4] https://huggingface.co/flaubert

models were trained on 71GB of natural text.

Following the architecture of Le et al. (2020), we propose several learning configurations in order to observe the impact of different parameters on the performance of the models obtained. Since we only have lowercase transcripts, we consider the *flaubert-base-uncased* model as our reference.[5]

The first configuration, **FlauBERT-O-base_uncased** (**FT**), consists in fine-tuning the public *flaubert-base-uncased* model for some epochs using our ASR transcripts.

The second configuration **FlauBERT-O-mixed** (**MIX**) is a full model re-trained using a mix of ASR text and written text, as training data. Written text comes from two main sources: the French wikipedia dump and press articles captured by the OTMedia research platform (Hervé, 2019) (online press and AFP agency for the same time period). Overall, this learning dataset is also strongly news-oriented. For the written text, we use the same sentence segmentation tool as the one used for FlauBERT. Our dataset is balanced between ASR and written text: we use 94M randomly selected written text sentences representing 13G of data to which we removed the punctuation and capitalization to make it consistent with our ASR data. For this mixed model, we also retrain the BPE tokenizer (50K sub-word units).

The third configuration, **FlauBERT-O-asr**, consists in re-training LMs from scratch using ASR data only. For the first model (**ORAL**), we use the tokenizer provided with the *flaubert-base-uncased* model and for the second one (**ORAL_NB**) we re-train a BPE tokenizer (50K sub-word units). Both tokenizers share 35088 (overlap) out of 67536 (FlauBERT initial) tokens, only 52% overlap.

These different configurations therefore provide us with 4 language models to evaluate. Training was done on a single server with 2 Xeon CPUs of 12 cores each, 256 GB of RAM and 8 Nvidia GeForce RTX 2080 Ti graphics cards with 11 GB of memory. With this hardware, it took us 15 days to train 50 epochs of each model in the flaubert-base configuration (137M parameters) using FlauBERT code.

## 4 Word Prediction Experiments

The first step in evaluating our models is to look at their behaviour for the word prediction task. In addition to the performance on the trained models, we also want to have an idea of the performance on texts of different nature (written style or oral style). We therefore assembled several datasets to measure the word prediction performance of the models we trained.

We make sure that these datasets are not included in the training data of the default FlauBERT model nor in our own. We have a first corpus (**afp2021**) of AFP dispatches from the year 2021, i.e. after the period of our training data collected from the online press. This will allow us to have a measure of performance on written text. Secondly, we want to evaluate our models on oral texts. We use the transcripts of the French National Assembly sessions.[6] We are using the 13th (under Sarkozy **parl_13**) and 15th (currently under Macron **parl_15**) mandates. These texts are a manual transcription of what is said in the hemicycle, which are prepared speeches with some degree of spontaneous style as well. A second corpus is constituted with, once again, the manual transcriptions made for educational videos[7] and interviews[8] that INA makes available via its web studio (**studio_manual**). These transcriptions are of very good quality. We also transcribed these videos from the studio with our ASR system (**studio_asr**) in order to be able to compare the performance on both types of data.

We report in the graphs the accuracy obtained on the different datasets for a word prediction task after a word has been masked. The masking parameters are the same as those used during training with MLM loss.

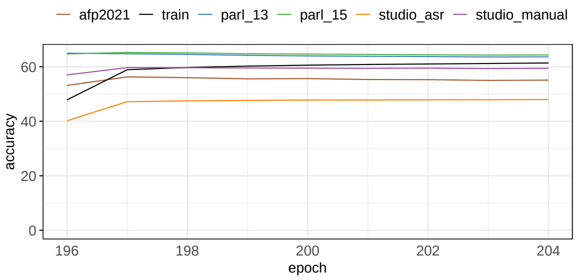

Figure 1: **FT** - Word prediction accuracy of *FlauBERT-O-base_uncased*

Figures 1 to 4 show the results assessed at each epoch. In table 2, we summarise the results for the last epoch and also for the default

[5] https://huggingface.co/flaubert/flaubert_base_uncased

[6] https://data.assemblee-nationale.fr/

[7] https://www.ina.fr/offres-et-services/fresques-numeriques

[8] https://entretiens.ina.fr/

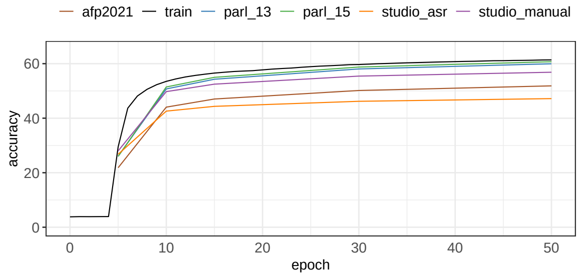

Figure 2: **ORAL** - Word prediction accuracy of *FlauBERT-O-asr*, using the initial *flaubert-base-uncased* BPE tokenizer

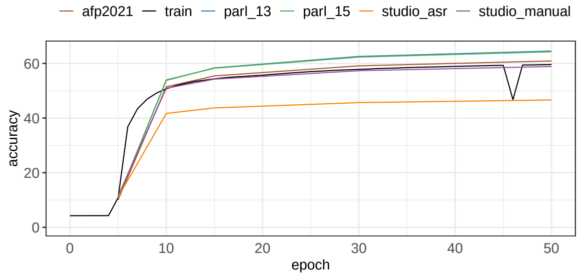

Figure 3: **MIX** - Word prediction accuracy of *FlauBERT-O-mixed*

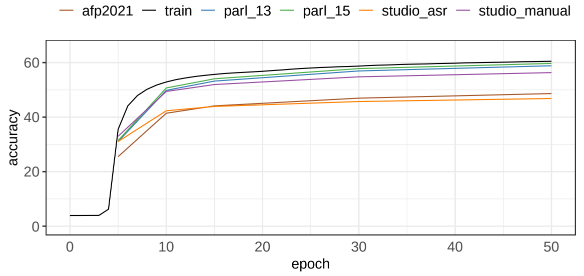

Figure 4: **ORAL_NB** - Word prediction accuracy of *FlauBERT-O-asr_newbpe*, using a new BPE tokenizer trained on ASR data

flaubert_base_uncased model (**FBU**). For the fine-tuned FlauBERT-O-base_uncased model, we notice a slight improvement in performance for afp and studio datasets, obtained from the first epoch, which means that adding ASR generated text improves word prediction task on these datasets. We observe that globally, whatever the model, the datasets of the parliamentary sessions are those for which the best performances are obtained on the word prediction task, even exceeding that of the training dataset for the FlauBERT-O-base_uncased and FlauBERT-O-mixed models. These models are trained on written and spoken texts and it is not surprising that the performance is good since the very nature of the parliamentary data is a mixture of prepared and spontaneous speech. There is no significant difference between parl_13 and parl_15. On these parlementary speeches, there is no significant performance difference between the 3 models that have seen written text during their training (FBU, FT and MIX). As we observed also that our FlauBERT-O models improve also on written text (afp2021), we explain this by the fact that those texts are strongly related to news events, so they are in a similar context to our ASR data which is focused on news slot transcripts. For the last corpus, from the INA web studio, we have educational videos or interviews of personalities which are more distant from news data. There is a great disparity in performance depending on whether we consider manual (studio_manual) or automatic (studio_asr) transcription. We believe that the different sentence segmentation algorithms have a very clear impact on this corpus. Finally, we notice that the ORAL_NB model performs slightly worse than the ORAL model. The BPE tokenizer obviously has an impact on the overall performance of the LMs and it seems, from this result, that using BPE units extracted from clean data (and not noisy ASR data) is beneficial even if the training material is itself ASR generated text.

| Corpus | FBU | FT | MIX | ORAL_NB | ORAL |
|---|---|---|---|---|---|
| afp2021 | 53.1 | 55.1 | **60.9** | 48.6 | 51.9 |
| parl_13 | **64.9** | 63.6 | 64.5 | 58.8 | 60.0 |
| parl_15 | **64.6** | 64.3 | 64.3 | 59.7 | 60.7 |
| studio_asr | 40.2 | **48.0** | 46.6 | 46.8 | 47.2 |
| studio_manual | 57.0 | **59.4** | 58.9 | 56.3 | 56.9 |

Table 2: Word prediction task accuracies

## 5 Downstream Task 1: Automatic Classification of TV Shows

We evaluate our different models on a news classification task. For the main generalist channels, INA's documentalists finely segment the newscasts and annotate them in order to describe their content. This very rich metadata is used in particular to establish quantitative studies on the news in France. The InaStat barometer[9] has set up a stable method-

[9] http://www.inatheque.fr/publications-evenements/ina-stat/

ology over time to classify these news items into 14 categories (such as society, French politics, sport or environment). We use the news items of 4 channels (TF1, France 2, France 3 and M6) for the years 2017, 2018 and 2019, which gives us a total of 47 867 short TV shows. The average length of these shows is 92 seconds.

### 5.1 Standard Learning Setting

The objective is to assess to what extent it is possible to classify these topics into the 14 categories solely on the basis of what is pronounced, i.e. from the ASR transcripts. We establish a baseline using a simple SVM classifier (with a non-parametric triangular kernel) on TF-IDF vectors with two vocabulary sizes of 5K and 20K words. To test the FlauBERT models, we use the HuggingFace Transformers library and the *FlaubertForSequenceClassification* class, which adds a simple dense classification layer on top of our models. To obtain a vector representation of our texts before this classification layer, we use the 'mean' summary type. We do not make any model selection and report the results for all learning epochs. Since the 14 categories are not well-balanced, we use the weighted F1 measure to evaluate the performance. The experiments are systematically performed on 10 different random splits of the dataset, taking into account the cardinality of the 14 categories, so as to have 38K examples for the training set and 5K for the test set. We show the average results and the standard deviation in figure 5.

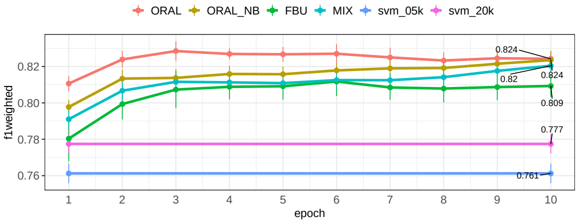

Figure 5: TV news classification - train 38K, test 5K

We can see in this configuration the contribution of the LMs compared to the SVMs along the training epochs of the classifier. If we look at the performance at the first epoch, we can see that the flaubert_base_uncased model has almost equivalent performance to the SVM (0.78). It is only after a few iterations of learning that the model fits the ASR data and reaches 0.81. On the other hand, the models that have already seen ASR data during their training have a better performance from the first epoch. The model trained only on ASR data is the best performing (ORAL). After 10 epochs, the 3 FlauBERT-Oral models converge and are equivalent for this task.

### 5.2 Few Shot Learning Setting

In order to test the LMs under more challenging conditions, we progressively reduce the number of training examples to get closer to few-shot learning conditions. We thus restart the classification with 5K training examples, then 500 and finally 200. Again, we take into account the cardinality of the 14 categories. For the last experiment with only 200 training examples, the vocabulary is too small and we can only test the SVM baseline with a vocabulary of 5K words, but not the version with 20K words. Moreover, we push to 30 epochs in this latter case.

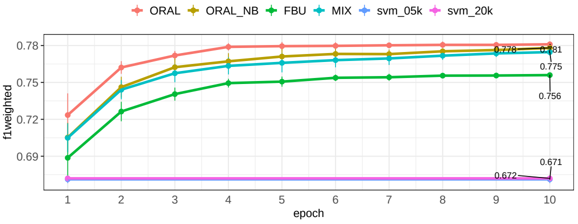

Figure 6: TV news classification - train 5K, test 38K

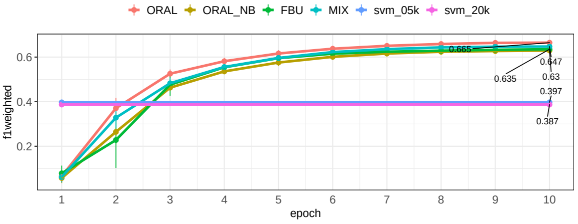

Figure 7: TV news classification - train 500, test 47K

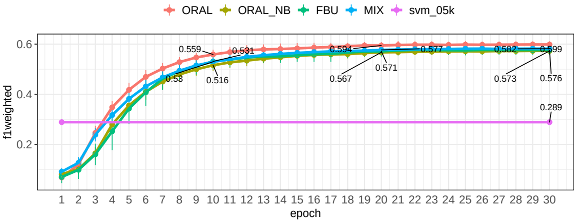

Figure 8: TV news classification - train 200, test 47K

As the number of training examples decreases, the performance gain over SVMs becomes more obvious. This is an expected result. In all cases,

ina-stat-sommaire.html

the models trained on ASR only text (ORAL) are the best of the FlauBERT-O models. Compared to the ORAL_NB model, only the tokenizer is different. This result may appear counter-intuitive in a first place, as one would expect a model entirely learned on ASR data to perform better on a classification task using only ASR data as input. However, this is probably counterbalanced by the fact that using BPE units extracted from clean data is important (as we have seen in the word prediction experiments). This invites us to further investigate the role of the tokenizer in spoken language modeling. As in the previous case, the Flaubert models converge almost with a 2 F1 point difference in favour of the FlauBERT-O models over the initial FlauBERT model.

## 6 Downstream Task 2: Syntactic Analysis of Spoken Conversations

This section is about the downstream task of jointly predicting part of speech tags (POS) and building a labelled dependency tree. The models performing these tasks typically rely on word representations, that are often pretrained, especially when the data is scarce. We will use our different spoken language models to obtain contextual word representations of a syntactically annotated and manually transcribed oral French corpus. For each of these representations, a model will be trained to perform the joint prediction of POS tags and labelled dependencies. We also use as baseline a model trained using non-contextual representations obtained with FastText,[10] and a model learning its own representations without any pretraining.

### 6.1 Data

We used the annotated subset of the speech corpus of the Orfeo project (Benzitoun et al., 2016; Nasr et al., 2020), gathered with the goal of reflecting the contemporary usage of the French language.

The audio extracts on which this corpus is based come from various origins and modalities: from one to multiple speakers, work meetings, family dinner conversations, narration, political meeting, interview, goal-oriented telephone conversations. Their duration varies from four minutes to an hour.

The reference audio transcripts have been obtained after correcting the output of an ASR system. The corpus is annotated in part of speech (POS) tags, lemmas, labeled dependency trees and sentences boundaries. There are 20 possible POS tags and 12 syntactic functions.

We randomly split the corpus into train/dev/test sets of respective sizes 134,716/27,937/29,529 words; we sampled from each source so that the various origins of the audios are equally represented in each split.

### 6.2 Parsing Model

The model is a transition based parser using the arc-eager transition system (Nivre, 2008), which has been extended for the joint prediction of POS tags and parsing transitions (Dary and Nasr, 2021).

It consists of a single classifier, taking as input a numeric representation of the current state of the analysis, called a configuration. The classifier predicts a probability distribution over the set of POS tagging actions or parsing actions, depending of the current state of the configuration. The analysis assume that the text is already tokenized and segmented into sentences; the words of each sentence are considered one by one, in the reading order; a POS action is predicted for the current word, then a sequence of arc-eager actions is predicted until the current word is either attached to a word on its left or shifted to a stack for future attachment to a word on its right. The predictions are greedy: it is always the top scoring action among the allowed ones. We do not use beam search for decoding.

The numeric representation of the current configuration is comprised of:

- The concatenation of the word embeddings, reduced from dimension 768 to dimension 64 by a linear layer, of the following context: the current word, the three preceding ones, the two following ones, the three topmost stack elements and the rightmost and leftmost dependents of the three topmost stack elements,

- The output of three different BiLSTM processing sequences of tags of the same nature. The first one is taking as input the sequence of POS tags and syntactic function of the current word, the three previous ones and the three topmost stack elements. The second one is taking the sequence of the last 10 actions that have been applied to this configuration. The last one is taking the sequence of distances (in number of words) between the current word and the three topmost stack elements. In each case,

[10] https://fasttext.cc

the sequence elements are encoded by learnable and randomly initialized embeddings of size 128, and the output of the BiLSTM is a vector of size 128,

- A learnable and randomly initialized embedding encoding the current state of the configuration (POS tagging or dependency parsing).

A dropout of 50% is applied to the resulting vector; then it passes through two hidden layers of respective sizes 3200 and 1600, both with a dropout of 40% and a ReLU activation. Finally, the network is ended by one of the two decision layers, depending on the current state, which is simply a linear layer of dimension the number of possible actions followed by a softmax.

Each model was trained for 40 epochs; after every epoch the model was evaluated on the dev set and was saved if it was an improvement. After the fourth epoch, the entire train set was decoded using the model that was being trained, in order to generate and integrate novel configurations in the dataset for the epochs to come. This technique allows the model to be more robust, exploring non-optimal configurations during its training. It is based on the dynamical oracle model of Goldberg and Nivre (2012).

### 6.3 Experiments

The first set of experiments compares input representations from the FlauBERT variants (FBU, MIX, ORAL) to uncontextual word embeddings (Fasttext) and randomly initialized embeddings. Except for random embeddings, token representations are frozen when the parsing system is trained.

As pre-processing, we deanonymize the transcripts by replacing masked proper name tokens with non-ambiguous names randomly chosen for each recording. In the fasttext setting, representations are computed for unknown words from their character n-gram factors. Contextual representations are computed at the whole recording level in chunks of 512 tokens without overlap. The parser is applied on the reference transcript and reference segmentation. We use mean pooling for words that are split in multiple tokens by BPE.

Parsing performance is evaluated with Labeled Attachment Score (LAS), the accuracy of predicting the governor of each word and its dependency label, Unlabeled Attachment Score (UAS), which ignores the dependency label, and Part-of-speech tagging accuracy (UPOS). The scoring script is from CoNLL campaigns.

| **Repr.** | **LAS** | **UAS** | **UPOS** |
|---|---|---|---|
| No pretraining | 84.92 | 88.48 | 94.51 |
| Fasttext | 85.36 | 88.76 | 95.12 |
| FBU | 85.55 | 89.02 | 93.36 |
| MIX | 86.33 | 89.79 | 94.43 |
| ORAL | **87.65** | **90.92** | 95.55 |
| ORAL_NB | 87.54 | 90.73 | **95.63** |

Table 3: Main result on syntax prediction. Metrics are Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS) and Part-of-speech tagging accuracy (UPOS). Higher is better, highest figure in bold.

Results presented in Table 3 show that pretraining is valuable for syntactic parsing in that setting and that pretraining on ASR (**MIX** and **ORAL**) leads to a substancial improvement in LAS over the text-only FlauBERT model (**FBU**) even though there is no domain overlap between the TV shows on which the earlier is trained and the data of the Orfeo corpus. There is no benefit from retraining BPE (**ORAL_NB**).

| **Repr.** | **LAS** | **UAS** | **UPOS** |
|---|---|---|---|
| FBU | 85.55 | 89.02 | 93.36 |
| FBU w/ punct | 87.48 | 90.69 | 95.03 |
| ORAL | 87.65 | 90.92 | 95.55 |

Table 4: Effect of repunctuating speech transcripts on syntactic parsing prior to extracting representations. Results from the ORAL representations are given for reference.

As noted earlier, speech recordings do not have punctuation and it is debated whether punctuation is suitable for spontaneous conversations. As punctuation is rather regular in text, it would make sense for LMs trained on text to over-rely on the cues it brings, and representations to be affected by a lack of punctuation. Table 4 shows syntactic parsing results on representations where a simple heuristic is applied to add a period at the end of each sentence prior to extracting representations. This punctuation is stripped before passing the tokens to the syntactic parser and only used at the encoding stage. Results show that most of the difference in performance between the **FBU** and **ORAL** models can be compensated by this use of virtual punctuation. Using accurately predicted punctuation with diverse symbols and intra-sentence marks is

| **Repr.** | **LAS** | | | **UAS** | | | **UPOS** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Global | OOV | Δ | Global | OOV | Δ | Global | OOV | Δ |
| FBU | 85.55 | 74.10 | -11.45 | 89.02 | 82.20 | -6.82 | 93.36 | 79.00 | -14.36 |
| MIX | 86.33 | 74.40 | -11.93 | 89.79 | 82.47 | -7.33 | 94.43 | 80.35 | -14.07 |
| ORAL | 87.65 | 73.68 | -13.97 | 90.92 | 82.81 | -8.11 | 95.55 | 79.00 | -16.55 |

Table 5: Syntactic parsing performance on OOV words according to automatic transcription system. The Δ column contains the difference between the global accuracy and the accuracy on OOVs only.

left as future work, but we conjecture that it will marginally improve over this crude heuristic.

Gauging the impact of speech-to-text errors on representations from LMs trained on such data is difficult since there are no manual references available for large quantities of speech transcripts. Since the system used to transcribe the recordings is closed vocabulary, one way to look at this problem is to compute the accuracy of the syntactic parser on words that are out-of-vocabulary (OOV) for the LM training data. Due to BPE, those words are necesseraly tokenized in smaller units which are pooled prior to passing them to the parser, and might hamper the quality of the associated representations. Table 5 details the performance of the syntactic parser on OOVs. Due to their infrequent nature, OOVs are mainly swear words, proper names, and tokenization artifacts. They are difficult to handle for all models, and suffer from a large performance reduction compared to the global figure, even for the **FBU** model which has seen a much larger variety of texts. The system fed with representations of the model trained on ASR data only (**ORAL**) is the most affected despite its better global performance.

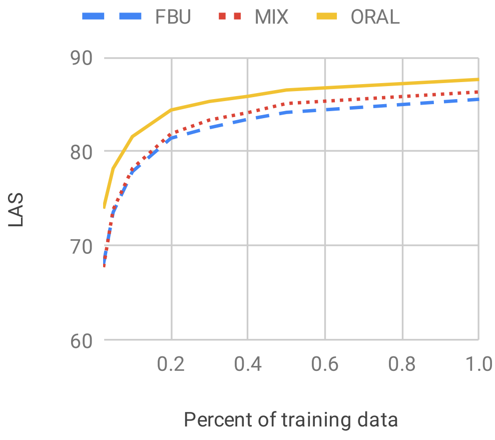

Figure 9: LAS learning curve for syntactic parser according to quantity of training data. Similar shape is obtained for UAS and UPOS.

Finally, Figure 9 shows the learning curve when reducing the training data available to the syntactic parser. For this, we randomly sampled 10 subsets of the training data at the recording level in order to fit a target ratio from 2.5% to 100%. The figure shows that LAS is always better for **ORAL** representations and that **MIX** is closer to **FBU** when less data is available.

### 6.4 Takeaways

It seems that exploiting ASR transcripts for learning LMs is beneficial for syntactic parsing of speech transcripts. Analyses presented show that punctuation plays an important role in representations. Our analysis of parsing performance on OOV words (according to the speech-to-text system) reveals that our FlauBERT-O-asr (**ORAL**) model is more affected than its initial FlauBERT baseline (**FBU**), despite overall better performance.

## 7 Conclusion and future work

We investigated spoken language modeling using ASR generated text (350,000 hours of diverse TV shows). The new models for French (FlauBERT-O) are shared with the community. Experimental results show that FlauBERT-O is generally better than its initial FlauBERT version for the downstream speech tasks we experimented with. However we should also check its performance on text downstream tasks (such as (Le et al., 2020)) and on more downstream speech tasks (SLU or ASR re-scoring).

In this work, all our texts were uncased as our ASR only generates lowercased transcripts. We believe that applying massively re-capitalisation (and restoring punctuation as well) might be beneficial to train stronger spoken LMs. We also plan to analyze more the specificities of our ASR-generated texts (do they contain more oral features such as word repetitions, more interjections?). Finally, some of the results obtained lead us to believe that it is important to further evaluate the impact of BPE units for spoken language modeling.

## References

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

Shih-Hsuan Chiu and Berlin Chen. 2021. Innovative bert-based reranking language models for speech recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*.

Yung-Sung Chuang, Chi-Liang Liu, and Hung-yi Lee. 2019. Speechbert: Cross-modal pre-trained language model for end-to-end spoken question answering. *CoRR*, abs/1910.11559.

Franck Dary and Alexis Nasr. 2021. The reading machine: A versatile framework for studying incremental parsing strategies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 26–37, Online. Association for Computational Linguistics.

Paul Deléglise, Yannick Esteve, Sylvain Meignier, and Teva Merlin. 2009. Improvements to the lium french asr system based on cmu sphinx: what helps to significantly reduce the word error rate? In *Tenth Annual Conference of the International Speech Communication Association*.

Dominique Fohr and Irina Illina. 2021. BERT-based Semantic Model for Rescoring N-best Speech Recognition List. In *INTERSPEECH 2021*, Proceedings of INTERSPEECH 2021, Brno, Czech Republic.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107.

Yoav Goldberg and Joakim Nivre. 2012. A dynamic oracle for arc-eager dependency parsing. In *Proceedings of COLING 2012*, pages 959–976.

Maria Goryainova, Cyril Grouin, Sophie Rosset, and Ioana Vasilescu. 2014. Morpho-syntactic study of errors from speech recognition system. In *LREC*, volume 14, pages 3050–3056.

Nicolas Hervé. 2019. OTMedia, the TransMedia news observatory. In *FIAT/IFTA Media Management Seminar 2019*.

Ayush Kumar, Mukuntha Narayanan Sundararaman, and Jithendra Vepa. 2021. What BERT based language model learns in spoken transcripts: An empirical study. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 322–336, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sylvain Meignier and Teva Merlin. 2010. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*.

Alexis Nasr, Franck Dary, Frédéric Bechet, and Benoît Fabre. 2020. Annotation syntaxique automatique de la partie orale du orféo. *Langages*, (3):87–102.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript.

Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. 2013. Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, pages 2345–2349.

Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2022. Rescorebert: Discriminative speech recognition rescoring with bert.