Parsing Spoken Naija

Minnie Kabra, Benjamin Lecouteux, Maximin Coavoux
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

MOTS-CLÉS : Analyse syntaxique automatique, parole, naija.

KEYWORDS: Syntactic Parsing, speech, Naija.

1 Introduction

Data-driven linguistics usually rely on linguisticly annotated data. However, spontaneous spoken data, arguably the most realistic type of linguistic data is notoriously hard to acquire and all the more difficult to annotate. Direct dependency parsing of the speech signal has been recently proposed as a task (Pupier *et al.*, 2022) in an attempt to avoid relying on noisy automatic speech recognition (ASR) outputs and to use acoustic features that are a crucial source of information for syntactic segmentation and disambiguation (Price *et al.*, 1991). In this abstract, we present a dependency parsing model for speech, i.e. a system that takes the audio signal of a recording and outputs both (i) a transcription and (ii) its syntactic tree. Our system is refinement of the parser presented by Pupier *et al.* (2022), that is simpler, more parameter-efficient and faster.

We assess the utility of our system in the context of low-resource languages by presenting a set of experiments in Naija, an English-based creole language spoken in Nigeria. In particular, we use the NaijaSynCor treebank (NaijaNSC, Caron *et al.*, 2019), a spoken treebank designed to study the prosody-syntax interface. The NaijaSynCor treebank contains macrosyntactic segmentation annotations materialised by punctuation symbols. We study the impact of these segmentation symbols on speech parsing, as they were found to be useful for parsing transcriptions by Caron *et al.* (2019).

In short, our contributions are : [1]

— a parameter-efficient end-to-end speech parser
— a set of speech parsing experiments in Naija that show that syntactic-prosodic segmentation symbols has no positive impact on parsing.

1. This abstract's content overlaps with an article by the same authors currently under review at ARR. The abstract has been specifically rewritten for LIFT's audience.

2 Dataset : Naija-NSC

We use the NaijaSynCor (Naija-SNC) treebank v2.14 (Caron *et al.*, 2019), released as part of Universal Dependencies (Nivre *et al.*, 2020). Naija-NSC is a SUD (Gerdes *et al.*, 2018) dependency treebank of spoken Naija released with the corresponding audio recordings. It features various kinds of interactions including life stories, speeches, radio programs, free conversations, cooking recipes, comments on current state of affairs. We use the provided train/dev/test split corresponding respectively to 6.8h (7268 trees, 67 speakers), 0.9h (990 trees, 10 speakers), 0.9h (972 trees, 11 speakers).

The dataset contains syntactic and prosodic segmentation annotations materialized by punctuation symbols that are included as tokens in the dependency trees. For exemple, the tree in Figure 1 contains the following symbols :

— `//` delimits illocutionary units (final token of each dependency tree)
— `{ }` indicates a sequence of elements with the same syntactic function ; the disfluency with a false start is represented with the pattern `{ reparandum || repair } .`
— # denotes a short pause.

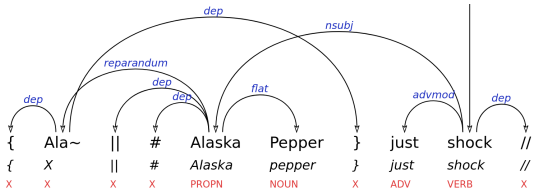
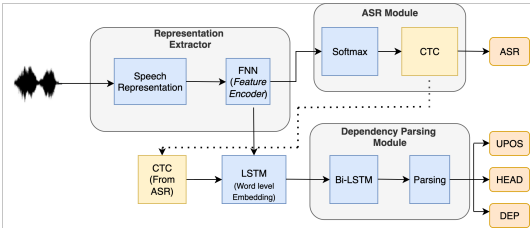
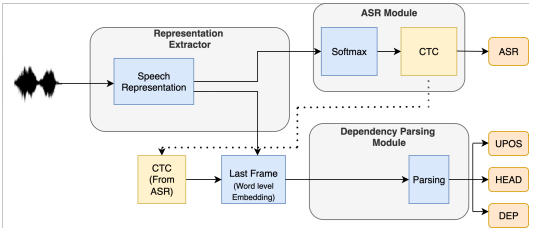


FIGURE 1 – Example (NaijaNSC treebank). EN : Ala-... Alaska Pepper was shocked.



(a) Wav2tree architecture (Pupier *et al.*, 2024).



(b) Simplified Wav2tree (ours).

FIGURE 2 – Comparison original (Pupier *et al.*, 2024) and proposed architecture.

3 Model : From Wav2tree to Light-Weight Wav2tree

Wav2Tree (Pupier *et al.*, 2022) is an end-to-end dependency parsing model whose only input is the raw speech signal. Its architecture is illustrated in Figure 2a, and performs jointly speech recognition and parsing. The Wav2tree architecture is composed of three modules : (i)

Wav2tree	Wav2vec2	FFN + LSTMs	Punctuation presence	WER ↓	CER ↓	UPOS ↑	UAS ↑	LAS ↑	Parameters
Original	24 layers	yes	Train & Eval	39.8	21.4	72.3	58.3	51.9	$315M^{w}$+$13M^{f}$+$4.5M^{p}$
Ours (simplified)	24 layers	no	Train & Eval	**37.6**	**20.5**	**73.6**	**59.1**	53.4	$315M^{w}$+$5M^{p}$
Ours (simplified)	24 layers	no	Train only	35.8	18.2	75.6	65.9	58.1	$315M^{w}$+$5M^{p}$
Ours (simplified)	24 layers	no	None	**34.9**	**17.3**	**76.0**	**66.9**	**59.1**	$315M^{w}$+$5M^{p}$

TABLE 1 – Evaluation on the Naija-NSC treebank. Parameter counts : ^{w}Wav2vec2, ^{f}feedforward + LSTM, ^{p}parsing module.

a representation extractor that computes a vector representation for each frame in the speech signal, (ii) an ASR module that computes both a transcription and the frame boundaries of each predicted token, (iii) a parsing module that uses the word boundaries to compute audio word embeddings and run a classical parsing algorithm on the sequence of embeddings.

The resulting word-level representations are passed to a biaffine parser (Dozat & Manning, 2017). The whole Wav2tree model is trained end-to-end with a multitask objective (ASR and parsing) to reduce error propagation (Figure 2a).

Simplified Wav2tree Our architectural innovations consist in ablating the Wav2tree submodules that we thought redundant. The changes are illustrated in Figure 2 :

1. Wav2tree uses a feedforward network on top of a pretrained acoustic model (Wav2vec2), which we remove given the high quality of wav2vec2 representations.
2. instead of using an LSTM to create an audio word embedding from the sequence of frames that form a word, we simply represent a word by the vector of its last frame (given the frame vectors are already contextualized by the Wav2vec transformer).
3. we do not run an utterance level bi-LSTM on the sequence of audio word embeddings, considering once again the word vectors are already contextualized thanks to the Wav2vec2 transformer.

4 Experiments and Discussion

Our experiments are meant to evaluate the proposed system modifications and assess whether keeping the prosodic-syntactic segmentation symbols in the training set is helpful for parsing. We use `wav2vec2-xls-r-300m`,[2] a multilingual Wav2vec2 pretrained acoustic model (Babu *et al.*, 2022). Its weights are fine-tuned during training.

We evaluate with Word Error Rate (WER) and Character Error Rate (CER) for speech recognition, POS accuracy (POS), Unlabeled Attachment Score (UAS), and Labeled Attachment Score (LAS) for dependency parsing.

Results We report results for all experimental settings in Table 1. The simplified architecture we propose improved the performance of both ASR (-2.2% WER) and parsing (+1.5%

2. https://huggingface.co/facebook/wav2vec2-xls-r-300m

LAS). As shown in the last two lines of Table 1, including punctuation during training slightly degrades both ASR and parsing performance on the evaluation dataset without punctuation (-0.9 WER and +1.1 absolute LAS when removing punctuation from the training set). This result is not in line with Caron *et al.* (2019) who found the same punctuation marks useful when parsing texts. We attribute this difference to the difficulty for the ASR system to predict segmentation symbols accurately.

Références

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J., BAEVSKI A., CONNEAU A. & AULI M. (2022). XLS-R : self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, p. 2278–2282. DOI : 10.21437/Interspeech.2022-143.

CARON B., COURTIN M., GERDES K. & KAHANE S. (2019). A surface-syntactic ud treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, p. 13–24 : Association for Computational Linguistics.

DOZAT T. & MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* : OpenReview.net.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2018). Sud or surface-syntactic universal dependencies : An annotation scheme near-isomorphic to ud. In *Universal dependencies workshop 2018*.

NIVRE J., DE MARNEFFE M.-C., GINTER F., HAJIČ J., MANNING C. D., PYYSALO S., SCHUSTER S., TYERS F. & ZEMAN D. (2020). Universal Dependencies v2 : An evergrowing multilingual treebank collection. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France : European Language Resources Association.

PRICE P. J., OSTENDORF M., SHATTUCK-HUFNAGEL S. & FONG C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, **90**(6), 2956–2970.

PUPIER A., COAVOUX M., GOULIAN J. & LECOUTEUX B. (2024). Growing trees on sounds : Assessing strategies for end-to-end dependency parsing of speech. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 225–233, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-short.22.

PUPIER A., COAVOUX M., LECOUTEUX B. & GOULIAN J. (2022). End-to-end dependency parsing of spoken french. In *Interspeech 2022*, p. 1816–1820. DOI : 10.21437/Interspeech.2022-381.