Rethinking morphosyntactic annotations with automatic prosodic labelling: Test-driving a novel intonosyntactic treebank format on Nigerian Pidgin

Emmett Strickland
National Institute for Oriental Languages and Civilizations (INALCO)
Paris, France
emmett.strickland@inalco.fr

Relevant UniDive working groups: WG1, WG4

1 Introduction

Nigerian Pidgin, or Naijá, is a low-resource creole language spoken by as many as 110 million people in West Africa (Faraclas, 2021). Historically stigmatized and lacking any official status, Naijá is relatively understudied compared to other languages of its size. This provides fertile ground for new tools to better understand the grammar and phonology of one of the world's fastest-growing languages.

This submission presents a demonstration of a recently published treebank annotation scheme developed for Nigerian Pidgin, which allows for the joint study of syntax and prosody with public-facing tools like GREW-Match (Guillaume, 2021). This NaijaSynCor-Prosody treebank notably combines traditional syntactic dependency annotations with a detailed layer of phonetic annotations describing every syllable of every token.

In this contribution, we will provide an interactive, web-based demo of our annotation scheme, which is fully language-independent and has since been applied to a corpus of French (Botero-Garcia et al., 2025). During this live, software-enhanced poster demonstration, we will use a combination of phonetic and morphosyntactic annotations to reveal distinct prosodic categories within part of speech groups, highlighting how phonetic information can be leveraged to reinterpret morphosyntactic categories.

2 Corpus design

Our work is based on the NaijaSynCor treebank, developed over the course of a 2017-2021 project funded by the French National Research Agency (Caron et al., 2019). The original corpus was primarily encoded as a syntactic treebank of transcribed spontaneous speech, with every utterance represented as a dependency tree in the Surface-Syntactic Universal Dependencies (SUD) annotation scheme (Gerdes et al., 2018). An example of a tree encoded in the original format can be seen in Figure 2. Regrettably, this format excluded the multitude of phonetic information accessible in the original field recordings.

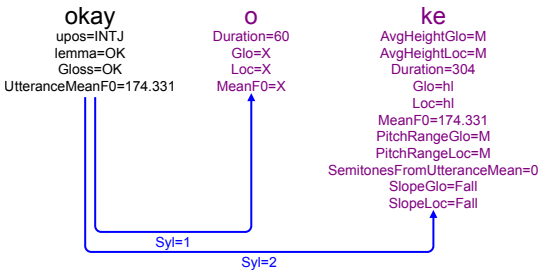

Figure 1: Example of data encoding

Our corpus addresses this gap by associating every orthographically-transcribed token in the original NaijaSynCor treebank with a distinct category of node describing each syllable. These notably carry:

1. A SAMPA phonetic transcription.
2. The shape of the syllable's pitch contour.
3. Mean F0 and various normalizations.
4. Duration and various normalizations.
5. Loudness and various normalizations.

Each of these syllable nodes are connected to their associated tokens using `Syl` edges specifying the syllable's position within that word. For instance, the edge `Syl=2` would connect a word token, carrying traditional morphosyntactic annotations, to that word's second syllable. This syllable node in turn would carry the various phonetic features listed above. These prosody-oriented edges and nodes can be queried and filtered using the same tools used to manipulate their traditional syntactic counterparts.

An example of this encoding is provided in Figure 1. Each of these annotations is encoded in the same `.conllu` tabular data format as the pre-existing syntactic annotations, allowing users to access both syntactic and prosodic labels using the

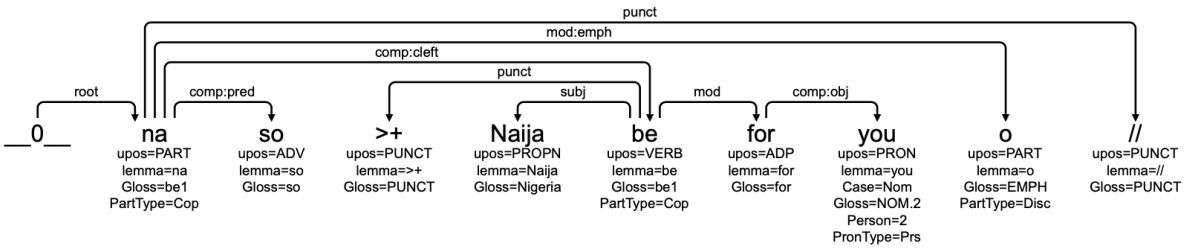

Figure 2: Example of syntactic dependency tree from the original NaijaSynCor treebank

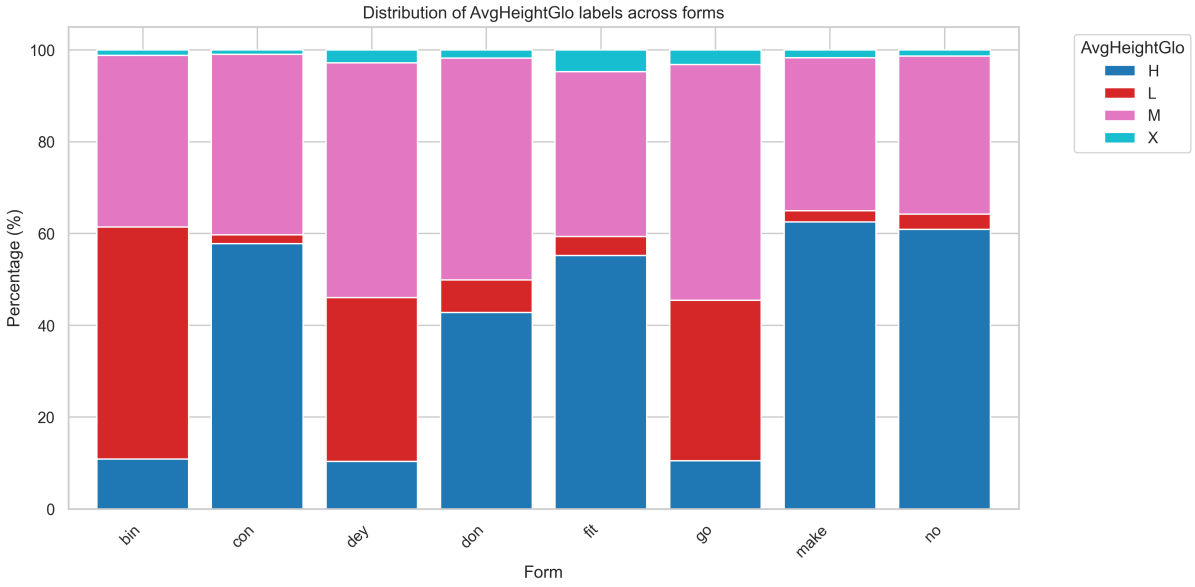

Figure 3: Distribution of categorical pitch labels across various auxiliaries, with high, medium, and low labels represented in blue, fuchsia, and red.

same GREW-Match query interface. Concretely, this allows users to study how prosody and syntax interact by viewing which syntactic labels are most strongly correlated with which prosodic annotations. This format can also be used to study sociolinguistic variation, as each utterance is associated with basic biographical information about its speaker.

3 Using this format to explore intonosyntax

During our demo, we will use these annotations and query tools to revisit longstanding questions about the prosodic typology of Nigerian Pidgin. Naijá has traditionally been described as a tone language in which pitch can be used to distinguish a handful of mostly monosyllabic minimal pairs like *gò* 'FUTURE' and *gó* 'go' (Mafeni, 1971; Faraclas, 1996). Consulting this richly-annotated corpus of spontaneous speech sheds new light on these analyses. In particular, we show that while such minimal pairs exhibit clear pitch differences, they are also associated with differences in duration and intensity reminiscent of stress-accent languages.

The demo also shows how this format can help to optimize corpora more generally. For example, we will see that a single part of speech category can contain markedly divergent prosodic profiles during our exploration of Naijá auxiliaries. Our prosodic labels reveal two broad categories hidden beneath the original syntactic labels: a low-pitched and low-duration group composed of *bin* 'PAST', *dey* 'IMPERFECTIVE', and *go* 'FUTURE'; and a mostly high-pitched and high-duration group composed of *con* 'CONSECUTIVE', *don* 'PERFECTIVE', *make* 'SUBJUNCTIVE', *fit* 'ABILITY', and *no* 'NEGATIVE'.

This prosodic split is clearly observed in the distribution of categorical pitch labels in Figure 3, one of several striking results we will share during our demo. We will also argue that some prosodic groupings may correspond to separate syntactic

categories. One major contribution of this corpus format is therefore allowing annotators to use directly accessible prosodic information to inform their syntactic annotations. We will also discuss how prosodic labels might be used to better annotate otherwise ambiguous constructions where more than one label is possible. For example, both the `compound` and `modifier` relations can link a noun to an adjective depending on how lexicalized the construction is. Our methods reveal a potential prosodic difference between these two constructions, an approach we believe can be used to better annotate serial verbs, multi-word expressions, and other relevant constructions.

4 Summary

We present a language-independent treebank annotation scheme allowing for the joint study of prosody and syntax. Our contribution offers an on-site demo in which we use these annotations to better understand variations within part of speech categories used for a low-resource creole. We will show that automatic phonetic labeling can be useful for reanalyzing syntactic and morphosyntactic labels.

References

Maria Paz Botero-Garcia, Emmett Strickland, Bruno Guillaume, Sylvain Kahane, and Anne Lacheret-Dujour. 2025. An intonosyntactic treebank for spoken French: What is new with rhapsodie? In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 111–118, Ljubljana, Slovenia. Association for Computational Linguistics.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic ud treebank for naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Nicholas Faraclas. 1996. *Nigerian Pidgin*. Routledge.

Nicholas Faraclas. 2021. Naija: A language of the future. *Current trends in Nigerian Pidgin English: A sociolinguistic perspective*, pages 9–38.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal dependencies workshop 2018*.

Bruno Guillaume. 2021. Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175.

Bernard Mafeni. 1971. Nigerian pidgin. *The English Language in West Africa*, pages 95–112.