Constructing a proverb-based audio corpus of Ngie

Cutting-edge work in linguistics and natural language processing often involves the study of large corpora of spontaneous speech (Liberman, 2019). Such corpora are rare for African languages, and practically nonexistent for non-majority languages (Bird, 2020; Joshi et al. 2020). Creation of these resources for 'low-resource' languages is a growing priority in computer science (Nekoto et al. 2020), since they are required for the development of useful and societally beneficial speech technologies for most of the world's thousands of languages (Blasí et al, 2021). But many new resources created as part of this drive (i.e. Gutkin et al, 2019), however useful from an engineering perspective, lack ecological validity and naturalness, mainly consisting of read speech that is not well connected to community desires for documentation.

Here, we consider one way to better balance the needs of community stakeholders, language technologists, and linguists: leveraging written collections of proverbs to rapidly record large amounts of read material, including meta-text in the form of spoken commentary. The "seasoning" of speech with proverbs is an oft-remarked and important aspect of verbal communication in Africa down to the present (Yankah, 1989; Tadi, 2016). The proposed approach stands to generate data well-suited *both* to linguistic research *and* the production of culturally significant research outputs which may better align with community interests.

We model this approach by introducing our effort to create an audio corpus of spontaneous and read speech in Ngie (ISO 639-3 [ngj], Grassfields Bantu). The collection obtains its organizing structure from a written collection of over 500 Ngie proverbs or *iberger* (Abiedu, 2020). Ngie talkers read the organizing proverb and discuss the interpretation of each proverb in a less controlled, more informal register. Data is expected to be mainly monologic, but we plan to add multi-talker discussion. Phonemic and morphosyntactic annotation accompanies each entry (Figure 1). After this demonstration of the structure of the corpus, we conclude by considering possibilities of this model for community auto-documentation of verbal arts more generally.

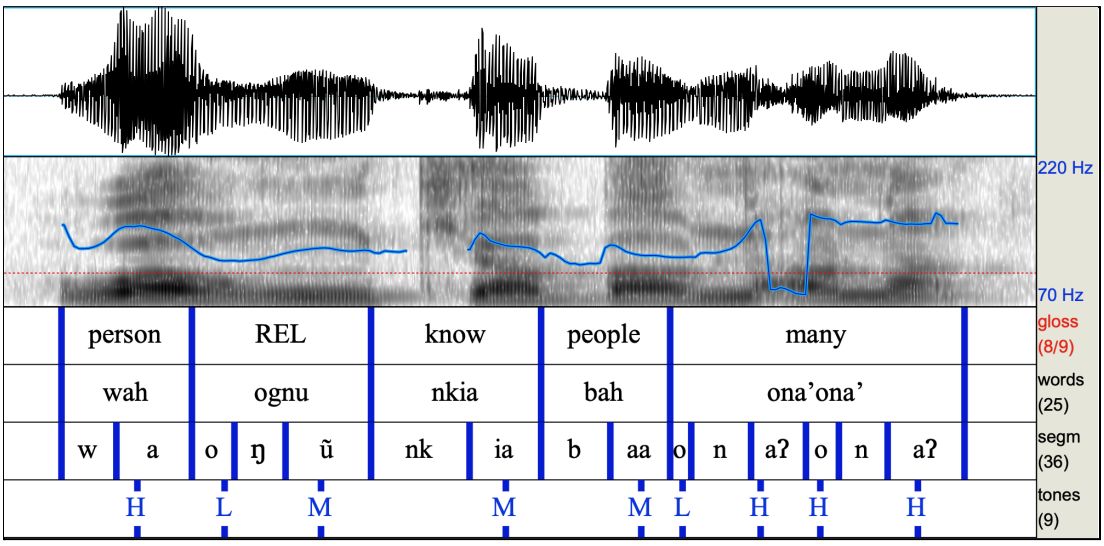

Figure 1. Waveform, spectrogram, and annotations for *oberger* no. 13: *Wah ognu nkia bah ona'ona' amer wah okap* 'the person who knows many people is rich.'

References

Abiedu, J. (2020). *Proverbs of Ngie (African Wisdom from Southern Cameroons)*. Unpubl. ms.

Bird, S. (2020). Decolonising speech and language technology. *Proc of ICCL 28*, 3504–3519.

Blasí, D., Anastasopoulos, A., & Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. arXiv:2110.06733

Gutkin, A., Demirsahin, I., Kjartansson, O., Rivera, C., & Túbọ̀sún, K. (2020). Developing an Open-Source Corpus of Yoruba Speech. *Proc Interspeech 2020*: 404-408.

Joshi, P., Santy, S., Budhiraja, J., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proc Assoc Comp Ling 58*: 6282-6293.

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, 5, 91-107.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., et al. (2020). Participatory research for low-resourced machine translation: A case study in African languages. arXiv:2010.02353.

Tadi, N. (2016). A Daily Dose of Wisdom: Globalization and SMS Proverbs in Nigeria. *Proverbium* 33: 411-430.

Yankah, K. (1989). Proverbs: The Aesthetics of Traditional Communication. *Research in African Literatures* 20(5): 325-346.