Natural Language Processing for language documentation: a progress report for Japhug and Na

Guillaume Wisniewski, Cécile Macaire, Benjamin Galliot, Oliver Adams, Nicholas Lambourne, Ben Foley, Janet Wiles, Alexis Michaud, Séverine Guillaume, Guillaume Jacques, et al.

Natural Language Processing for language documentation: a progress report for Japhug and Na

Guillaume Wisniewski [c], **Cécile Macaire** [a], **Benjamin Galliot** [a], **Oliver Adams** [b], **Nicholas Lambourne** [d e], **Ben Foley** [d e], **Janet Wiles** [d e], **Alexis Michaud** [a], **Séverine Guillaume** [a], **Guillaume Jacques** [f], **Nathan Hill** [g]

[a] Langues et Civilisations à Tradition Orale (LACITO), CNRS-Sorbonne Nouvelle, France
[b] Atos zData, United States of America
[c] Université Paris, Laboratoire de Linguistique Formelle (LLF), CNRS, France
[d] The University of Queensland, Brisbane, Australia
[e] ARC Centre of Excellence for the Dynamics of Language (CoEDL), Australia
[f] Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), CNRS-EHESS, France
[g] School of Oriental and African Studies, University of London, United Kingdom
guillaume.wisniewski@u-paris.fr, cecile.macairee@gmail.com, b.g01lyon@gmail.com, oliver.adams@gmail.com, {n.lambourne|b.foley|j.wiles}@uq.edu.au, {alexis.michaud|severine.guillaume}@cnrs.fr, rgyalrongskad@gmail.com, nh36@soas.ac.uk

Abstract

This document is an abstract accepted for presentation at the 6th Workshop on Sino-Tibetan Languages of Southwest China (STLS-2021) held at Kobe City University of Foreign Studies on September 7-11, 2021: https://sinotibetan-japan.com/stls/
We report on progress in interdisciplinary work using state-of-the-art technology for language documentation.

Language documentation is a key aspect of work on languages of Southwest China. Linguists typically produce 'Three Treasures' over the years: texts, a dictionary, and a grammar (Sun, 2007). In the Digital Age, the Three Treasures can be hyperlinked to one another, as well as to the original media files (audio and video). The ultimate goal is to allow seamless navigation between grammars, texts and dictionaries. This highly desirable goal is now within technical reach, as demonstrated by recent work (Musgrave and Thieberger, 2021).

But the journey towards this ultimate goal remains a long and lonely one. Although Natural Language Processing has potential for helping out with time-consuming tasks of corpus transcription and annotation (Adams et al., 2018; van Esch et al., 2019; Partanen et al., 2020), speech recognition technology has not yet been widely harnessed to aid linguists. We will report on work conducted on Japhug and Na using state-of-the-art computational tools to produce transcriptions and enrich the annotation of data sets.

Japhug will be familiar to workshop participants as having a rich system of consonant clusters, as well as flamboyant morphology (Jacques, 2021). In Japhug, syllables can have initial clusters containing at most three consonants, and at most one coda (Jacques, 2019). Japhug does not have lexical tones. The language's phonological profile is thus very different from Na (about which see Michaud, 2017).

The Japhug data set comprises a total of about 30 hours of transcribed recordings of narratives, time-aligned at the level of the sentence (Macaire, 2020), which is a huge amount in a language documentation context. The Na corpus comprises about four hours of transcribed narratives. The recordings were made in the course of field trips from the first years of the century until now, in a quiet environment, and almost all of a single speaker for each of the two languages. Our tests on various data sets so far suggest that these settings (one speaker – hence no speaker overlap – and clean audio) are those in which performance is most likely to be good when one happens to be training an acoustic model from scratch.

The full data sets are openly accessible online from the Pangloss Collection (an open language archive), under a Creative Commons license, allowing visitors to browse the texts, and computer scientists to try their hand at the data sets. The data collectors' approach to data sharing puts into practice some principles which gather increasing support, but which are not yet systematically translated into institutional and editorial policies (Garellek et al., 2020).

Our aim is to give researchers and language workers a clear account of what we did, and to convey a feel for the potential of the technology: to clarify the prospects for integrating Automatic Speech Recognition tools into language documentation workflows. The general perspective adopted consists in setting up pipelines for bringing a wide range of advances in speech recognition to a broad group of users.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365, Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, Honolulu, Hawai'i. https://computel-workshop.org/wp-content/uploads/2019/02/CEL3_book_papers_draft.pdf#page=26.

Marc Garellek, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, Daniel Recasens, Timo Roettger, Adrian Simpson, and Kristine M. Yu. 2020. Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science*, 9(1). https://halshs.archives-ouvertes.fr/halshs-02894375.

Guillaume Jacques. 2019. Japhug. *Journal of the International Phonetic Association*, 49(3):427–450.

Guillaume Jacques. 2021. *A grammar of Japhug*. Language Science Press, Berlin. https://langsci-press.org/catalog/book/295.

Cécile Macaire. 2020. Alignement temporel entre transcriptions et audio de données de langue japhug. In *Actes des Journées scientifiques du Groupement de Recherche "Linguistique informatique, formelle et de terrain" (LIFT)*, Paris. https://hal.archives-ouvertes.fr/hal-03047146.

Alexis Michaud. 2017. *Tone in Yongning Na: lexical tones and morphotonology*. Number 13 in Studies in Diversity Linguistics. Language Science Press, Berlin. http://langsci-press.org/catalog/book/109.

Simon Musgrave and Nick Thieberger. 2021. The language documentation quartet. In *Proceedings of ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Hawai'i. https://computel-workshop.org/wp-content/uploads/2021/02/2021.computel-1.2.pdf.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Jackson Tianshin Sun. 2007. Zàng-Miǎnyǔ de diàochá. *Yuyanxue Luncong*, 36:98–107.