4D-01

Corpus of East Cushitic Languages (CEC)

Sumiyo Nishiguchi
Center for Language Studies, Otaru University of Commerce

Abstract

This paper introduces a newly built corpus of East Cushitic Languages which include Dhaasanac, Rendille, Burji, Somali, and Afar. The face-to-face data collection has taken place in East Africa with generous support from local communities.

1 Corpus of East Cushitic Languages

Indigenous languages in East Cushitic languages in East Africa tend to be more oral than written. For example, the writing system of Dhaasanac (also written as Dasenach, Daasanach, and so forth) has developed only recently and uses International Phonetic Alphabet (IPA) [cf. 15]. With relatively small number of speakers, there is no dictionary made yet for Dhaasanac and there are few publications to read. Even the Bible translation is only partial [16]. Given such a situation, building corpora of East Cushitic languages contributes to translation, communication, and documentation. I have spent fifteen years working on East Cushitic languages including Dhaasanac, Somali, Afar, Rendille and Burji for theoretical linguistics and have collected data via oral communication with native speakers who include linguists and translators. The sessions generally entailed transcription, glossing, translation by the informants, and some recordings. The published papers contain a lot of examples for argumentation so that the data themselves can be separated as independent corpora. Being the single author of these papers, building corpora out of them does not interfere with authorship of publications.

2 Source of Data

As mentioned, the data in the corpus originates in face-to-face communications with native speakers living in Kenya and Djibouti. Giving pens and paper to the informant, mishearing or transcriptive errors have been reduced. Although some data come from [15] and [13]. the majority of sentences derive from [1, 2, 3, 4, 5, 6, 8, 7, 9, 10, 11].

3 Content

CEC is an aligned parallel corpus with word by word gloss between Cushitic languages and English using Aligned Corpus Reader in *Natural Language Tool Kit* (NLTK) [14], following [12]. (2) is one of the examples. At the moment, transferring data from papers to corpus is an ongoing project and two files on Dhaasanac are included.

(1) a. DasenachWords1.txt (15 words)

b. DasenachSentences1.txt (5 sentences)

(2) Um sugul maa niini dee
child.PL0-0 school1-1 person2-2 none3-3 a word4-4
wodh-e yu arg-e-man.
utter-PAST5-5 I6-6 see-PAST-NEG7-7

"I did not see students speak anything."

4 Challenges with Making a Corpus of East Cushitic Languages

In making a corpus of East Cushitic languages, the following challenges have to be overcome:

- Diacritics and tone marking
- IPA forms
- Long journey
- Languages

Writing a paper with LaTEX eases complications with diacritics and tone marking which coincide with these languages. The writing system uses IPA fonts which have to be represented in the corpus. The areas that speak these languages tend to be isolated and traveling may take time. Communication in English may not be possible in some places and higher educational institutes may not be available in the area.

5 Value

- Rare language
- Few written publications

Even so, there is tremendous value in making language resources of indigenous languages.

Acknoweldgment

I owe the data in corpus to Jackson Achinya, James Korie Haile, Mike, Benedict Lokono, Nyekitala Akol, Ambrose Awasio, Elmi Djima, Marian Sirayon, Hirkena Hargura and other people. The data collection is the by-product of the project "Indexicality and Negation in Dhaasanac and Somali" supported by JSPS KAKENHI Grant Number JP16K02643 between 2016 and 2021.

References

[1] Sumiyo Nishiguchi. Consonant assimilation and sonority: A case study in Daasanach. In Yukio Oba, editor, *OUPEL*, volume 9, pages 39–51. 2005.

[2] Sumiyo Nishiguchi. Bimoraic filter and sonority sensitive syllable contact in Dasenach imperfective. *Research in African Languages and Linguistics (RALL)*, 7:43–58, 2007.

[3] Sumiyo Nishiguchi. Polarity focus in Dhaasanac. *IHAFA: A Journal of African Studies*, 5:244–259, 2009.

[4] Sumiyo Nishiguchi. Prosodic morpheme in Dasenach. In Charles G. Haberl, editor, *Afroasiatic Studies in Memory of Robert Hetzron: Proceedings of the 35th Annual Meeting of the North American Conference of Afroasiatic Lingusitics (NACAL 35)*, pages 296–310, Newcastle, 2009. Cambridge Scholars.

[5] Sumiyo Nishiguchi. Shifty operators in Dhaasanac, 2011.

[6] Sumiyo Nishiguchi. Negation in Dhaasanac. *SULA6: Proceedings of the Sixth Conference on the Semantics of Under-represented Languages in the Americas and SULA-Bar*, 2012.

[7] Sumiyo Nishiguchi. Negative polarity item in Dhaasanac. In *The 147th Meeting of LSJ, Abstracts of Oral Presentations, Workshops and Poster Presentations*, page 7. 2013.

[8] Sumiyo Nishiguchi. Temporal indexicals in Dhaasanac. Attitudinal Semantics, Keio University, September 2013.

[9] Sumiyo Nishiguchi. Indexical shifting in Dhaasanac and Somali. In *Proceedings of Triple A3*. 2017.

[10] Sumiyo Nishiguchi. Shift-together in Burji, 2017.

[11] Sumiyo Nishiguchi. De re indexicals in Rendille. *LSJ159 Proceedings*, pages 339–345, 2019.

[12] Hongwoo Qin, Lei Kong, and Ranran Chu. The role of translation played in the evolution of Mandarin: A corpus-based account. pages 31–58, 2020.

[13] Hans-Jurgen Sasse. Dasenech. In M. Lionel Bender, editor, *The Non-Semitic Languages of Ethiopia*, pages 196–221. African Studies Center, Michigan State University, East Lansing, Michigan, 1976.

[14] Edward Loper Steven Bird, Ewan Klein. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit (English Edition), year=2009, publisher=O'Reilly Media, address=Sebastopol, CA*.

[15] Mauricio Tosco. *The Dhaasanac Language*. Rudiger Koppe Verlag, Köln, 2001.

[16] Bible Translation and Literacy. *Fayam Midhab Yiesu Kiristoká: He Hol Af Idiny Kaanaká (The New Testament in the Daasanach Language of Kenya and Ethiopia)*. Bible Translation and Literacy, 2014.