Exploring the possibility for a multiword expression corpus for Southern Nambikwara

Emilia Roosvall, Stockholm University, Sweden

Objective

The aim of this project is to explore the possibility of creating a multiword expression corpus for Southern Nambikwara, a polysynthetic language spoken in southwestern Brazil by approximately 900 people. Its intricate morphology and non-written tradition makes creating such a resource challenging. However, following my master's thesis (Roosvall 2022), there is a certain amount of newly collected data, segmented and annotated to varying degrees, that could make the base for a corpus that could play a significant role in the continued research on Nambikwaran languages.

Keywords: Southern Nambikwara, corpus creation for low-resource languages, multiword expressions

Background

Southern Nambikwara is a polysynthetic language with a complex grammar and morphology. There have been a few attempts at describing the language (B. Kroeker 1982, Lowe 1999, M. Kroeker 2001, da Silva 2021), but the sources feature considerable inconsistencies regarding segmentation and glossing, both within and between them. In my master thesis (Roosvall 2022), I attempted to combine the data in the previous descriptions with new first-hand data for my analysis. This data could make out the base for the creation of a new corpus, possibly with annotation for multiword expressions. As shown in (a-b) below, sentences in Southern Nambikwara do not necessarily include spacing, but recognised multiword expressions such as "get hurt" are attested in the data, as shown in (c) below (da Silva 2021).

a. ai-Ø-tetahe-nã
   hunt-3SG-Q.P.PNS-NON.SUBJV.W
   'Did he go yesterday?'

b. ũẽha-tehun-ĩnta-ua
   rain-EVAL-DED.D-SUBJV
   'I think it rained (I deduce, because I saw the earth wet).'

c. kanʾahatana tiahla an-ĩton-tiũnsu-na-ɾa
   this.morning 3SG.PRON EMPH.P-fall.ill-PNS-RECP.VIS-NON.SUBJV
   'In the morning, I think he got hurt (I didn't see it).'

Developing a resource for this extremely low resource language would greatly facilitate further research on the Nambikwaran languages as well as strengthen their position for inclusion in future typological research and cross-linguistic comparison. Furthermore, it would be interesting to explore to what extent multiword expressions can be found in the available data and how such a concept can be applied to polysynthetic languages.

The term 'multiword expression' pertains to a varied set of linguistic phenomena that do not conform to the word-phrase dichotomy since they, like phrases, are made up of multiple words, but, like words, have idiosyncrasies that must be learned. While there has been some

discussion of how applicable identified patterns of multiword expressions are to typologically diverse languages (see e.g. Baldwin et al. 2021), their cross-linguistic usefulness has also been questioned (Borin, n.d.) since many languages do not have written forms or word spacing. Nevertheless, Park, Schwartz, and Tyers (2021) used multiword expression annotation mechanisms in their work on the polysynthetic language St. Lawrence Island Yupik to adapt the morpheme-level annotation required by the complex morphology, to the word-level UD annotation guidelines.

Project Methodology

To create a new linguistic resource for a low resource language poses certain challenges. My initial aim will be establishing clear criteria for what constitutes a multiword expression in a polysynthetic language. These criteria will be developed based on the syntactic and morphological features documented in existing grammars, such as Kroeker (2001).

To develop a multiword expression corpus for Southern Nambikwara, this project will utilize existing linguistic data, primarily drawn from my master's thesis. This existing data, already segmented and glossed to varying degrees, provides a good foundation for identifying and documenting multiword expressions within the language.

The second step concerns constructing the corpus, annotated for various morphological properties as well as multiword expressions when or if possible. Following the criteria I will devise, I will identify the multiword expressions in existing data. Preliminary, I am planning to use a web-based annotation tool such as FLAT[1]; however, I am hoping that I will learn new tools and frameworks during the summer school that I can adapt in my project.

Utility of the Training School

The UniDive training would provide essential insights into techniques for corpus annotation in low resource languages. The courses on multiword expressions will be valuable for further insight into how to apply multiword expressions to typologically diverse languages, which is instrumental for the outcome of this project. Moreover, I am hoping to learn new tools and annotation frameworks that I can use in my future studies.

Open Questions

- How can we create corpus resources for languages with low (to no) resources?
- How can multiword expressions be applied to polysynthetic languages?
- Is it possible to create a multiword expression corpus for Southern Nambikwara?

Project Phase: The project is currently in its planning stage, ready to commence identifying the criteria for multiword expressions in polysynthetic languages.

References

Baldwin, Timothy, William Croft, Joakim Nivre, Agata Savary. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics: Report from Dagstuhl Seminar 21351. Dagstuhl Reports, 2021, 11 (7), pp.89–138. DOI: 10.4230/DagRep.11.7.89 hal-03507948

Borin, Lars (n.d.). "Is a cross-linguistic typology of multiword expressions useful or even possible?." https://typo.uni-konstanz.de/parseme/images/WG1-Volume-Outlines/BORIN-outline.pdf

[1] https://flat.readthedocs.io/en/latest/

da Silva, Sivaldo Correia (2021). Uma gramática descritiva do Nambikwara do Campo (Nambikwara do Sul). PhD thesis. Universidade Federal de Pernambuco.

Kroeker, Barbara (1982). Aspectos da língua nambikuara. Cuiabá: Sociedade Internacional de Lingüística-SIL.

Kroeker, Menno (2001). A descriptive grammar of Nambikuara. In: International Journal of American Linguistics 67.1, 1–87.

Lowe, Ivan (1999). Nambiquara. In: The Amazonian languages. Ed. by Robert M.K. Dixon and Alexandra Y. Aikhenvald. Cambridge: Cambridge University Press, 268–291.

Park, Hyunji Hayley, Lane Schwartz, Francis M. Tyers. Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik. Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 131–142. June 11, 2021.

Roosvall, Emilia. 2022. Re-visiting parts of the verb in Southern Nambikwara: Towards a definition of subjectivity as a grammatical category. Master's Thesis, Department of Linguistics, Stockholm University.