Corpus-based annotation of argument marking in Daakaka

Kilu von Prince, Rainer Osswald, Laura Kallmeyer, Simon Petitjean, David Arps, Natalia Moors
Heinrich Heine University Düsseldorf

The work presented here is part of an ongoing project towards creating syntactic RRG-annotations for a large corpus of Daakaka, an Oceanic language of Vanuatu, spoken by about one thousand speakers on the island of Ambrym. The corpus consists of 59k tokens and is annotated with interlinear morpheme-by-morpheme glosses [6], POS tags and translations to English. The basic word order of Daakaka is SVO, case is marked by subject agreement, in a nominative-accusative alignment. As is typical for the languages of the region, finite clauses are characterized by a tight-knit verbal complex which minimally includes subject-agreement marking (SUBJ), tense/aspect/mood marking (TAM) and the verb root [4]. Between the TAM marker and the verb root, it is possible to include an aspectual auxiliary (AUX), and the verb root can be reduplicated. The verb root can be followed by a resultative suffix, serial verbs(SV) and/or a transitivizer (TRANS).

(1) [SUBJ] [TAM] ([AUX]) (REDUP-)[verb root] ((REDUP-)[RES]) ([SV]) ([TRANS])

The annotation process takes advantage of the English translations and the interlinear glosses, similar to Xia and Lewis [7] and Bender et al. [2]. As in [3], the system proposes a tree to the annotator that is automatically generated. The annotator then checks the candidate annotation and modifies it accordingly. After a sufficiently large part of the corpus has been annotated in this way, we plan to automatize the process via data-driven parsing that exploits the manually annotated syntactic trees. For this talk, we focus on the annotation of subject-agreement markers and their syntactic status. Subject markers preceding the TAM marking correspond to pronominal elements in that they can be syntactically quite distant from the verb root and form their own phonological words. However, they still behave like agreement markers from other languages in crucial ways. In particular, there is a separate set of pronouns that are used as topics or objects, and subject markers can be optionally preceded by a full, complex NP. And while there is a position for topics before the verbal complex, this is not the position that subject NPs appear to occupy, as can be seen in the following example, where a left-dislocated topic NP precedes a subject NP (and subject marker):

(2)	[$_{\text{TOPIC}}$ bwili	wye	en=te] [$_{\text{SUBJECT}}$	vyanten	nyoo] [$_{\text{AGREEMENT}}$	ya]=m		du	tas	kyu
	hole.of	water	DEF=MED	person	3P	3P=REAL		stay	sit	surround
	‘this pond, people were sitting around it’									

A similar challenge for RRG has been discussed in [1], where an AGX node at nuclear level is stipulated (Figure 1). Alternatively, Daakaka could be treated like head-marking languages as described in [5] (Figures 2 and 3). We discuss the differences between Daakaka on the one hand and both Spanish and typical head-marking languages on the other hand, including, for example, the absence of object marking on the verb. Moreover, we illustrate how the qualitative and quantitative evaluation of the evolving Daakaka treebank can provide feedback on the analysis of the phenomena in question.

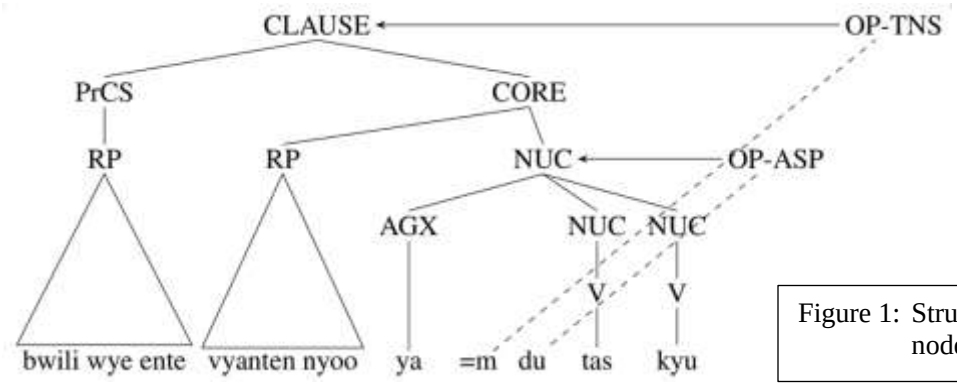

Figure 1: Structure of (2) with AGX node

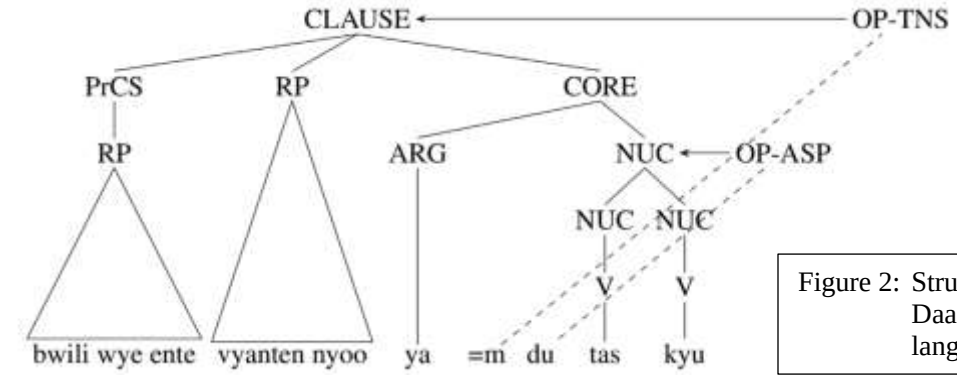

Figure 2: Structure of (2), with Daakaka as head-marking language (syntax version)

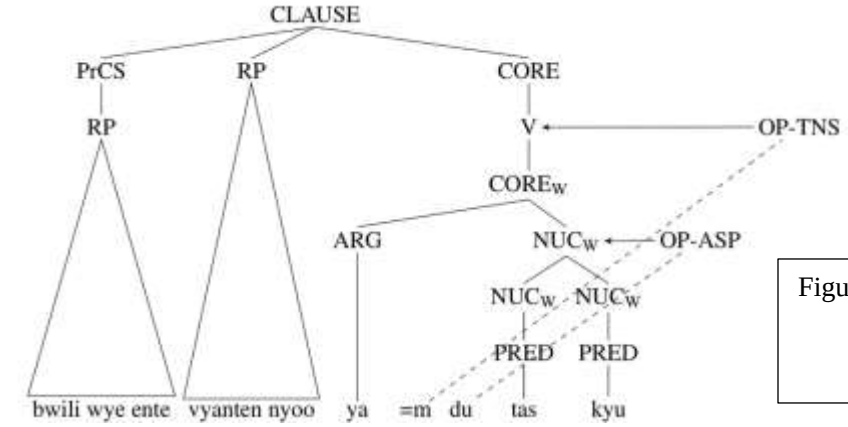

Figure 3: Structure of (2), with Daakaka as head-marking language (morphology version)

References:

[1] Valeria A Belloro. Spanish clitic doubling: A study of the syntax-pragmatics interface. PhD thesis. State University of New York at Buffalo, 2007. [2] Emily M. Bender et al. Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Properties. In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. 2013, pp. 74–83. [3] Tatiana Bladieret al. RRGbank: a Role and Reference Grammar Corpus of Syntactic Structures Extracted from the Penn Treebank. In: Proceedings of TLT17. 2018. [4] Kilu von Prince. A Grammar of Daakaka. Berlin, Boston: De Gruyter Mouton, 2015. [5] Robert D. Van Valin Jr. Head-marking languages and linguistic theory. In: Language Typology and Historical Contingency. Ed. by Balthasar Bickel et al. Amsterdam: John Benjamins, 2013, pp. 91–123. [6] Kilu von Prince. Daakaka. Nijmegen: TLA, 2013.URL: hdl.handle.net/1839/00- 0000- 0000- 000F- 4E20- B. [7] Fei Xia and William Lewis. Multilingualstructural projection across interlinear text. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Associationfor Computational Linguistics; Proceedings of the Main Conference. 2007, pp. 452–459.