Annotating Numeral Systems Across Languages

Johann-Mattis List, Arne Rubehn, Christoph Rzymksi, Kellen Parker van I

2024-12-12

Abstract

Numeral systems across the world's languages vary in fascinating ways, both regarding their synchronic structure and the diachronic processes that determined how they evolved in their current shape. For a proper comparison of numeral systems across different languages, however, it is important to code them in a standardized form that allows for the comparison of basic properties. Here, we present a simple but effective coding scheme for numeral annotation, along with a workflow that helps to code numeral systems in a computer-assisted manner. We illustrate the basic aspects of this workflow and provide sample data including numeral systems form 20 language varieties.

1 Introduction

- what makes numeral systems interesting?
  - the fact that they are compositional but also distinctive, so they are built from a fixed set of morphemes, but the resulting word forms cannot "colexify"
  - the fact that they have evolved, the are not built once, but they have evolved into their current shape, and this evolution is interesting, showing how they have developed mnemotechnic characteristics sporadically (contamination) and other processes contrary to sound change
  - the fact that they are easily exchanged across cultures, they are often not a fix immutable part of a language
  - as a result, they are also pretty useful to study methods for morpheme segmentation
- consistent coding of numeral systems is so far lacking, but we think it could enlighten ourselves in multiple ways here, so we present the coding scheme and some tests

2 Background

- how have numeral systems been studied?
  - the google group
  - some work by bender
  - work by Xu et al. (2020)
  - think of other previous work (Comrie 2020)

1

- what are the problems in current coding that we identify?
    - coding not true to the segments (no phonetic transcirption)
    - not consistent coding of similarities / cognacy inside a language

Example for figures is shown in Figure 1.

ID	DOCULECT	CONCEPT	VALUE	TOKENS	MORPHEMES	COGIDS	NOTE
62994	Welsh	finger	bys	b ɪ s	finger	1³	None
62998	Welsh	foot	troed	t r ɔɪ d	foot	2²	None
62997	Welsh	toe	bys troed	b ɪ s t r ɔɪ d	finger + foot	1³ 2²	None

Figure 1: This is my caption.

Run the code to compile by typing:

```
$ sh pandocit.sh draft.md draft.docx
$ sh pandocit.sh draft.md draft.pdf
```

3 A cross-linguistic coding scheme for numeral systems

3.1 Preliminary considerations

3.2 Annotation

- segmented strings
- phonetic transcription in IPA / CLTS
- morpheme glosses
- slash construct (allomorphmic annotation)

3.3 Representation

- we follow lingpy / lingrex / edictor here

3.4 Computer-assisted coding of numeral systems

- edictor
- morpheme segmentation experiments

3.5 Implementation

- morseg (preliminary library or individual functions, decide later)
- edictor 3.0

4 Examples

4.1 Sample data of coded numeral systems

- describe data set and size
- run analysis of ideal allomorphs vs morphs for all languages

4.2 Automated morpheme segmentation

- describe morpheme segmentation experiments

4.3 Examples for numeral codings

- three concrete examples with a few numerals

5 Discussion and outlook

References

Comrie, Bernard. 2020. "Revisiting Greenberg's 'Generalizations about Numeral Systems' (1978)." *Journal of Universal Language* 21 (2): 43–84. https://doi.org/10.22425/jul.2020.21.2.43.

Xu, Yang, Khang Duong, Barbara C. Malt, Jiang Serena, and Mahesh Srinivasan. 2020. "Conceptual Relations Predict Colexification Across Languages." *Cognition* 201 (104280). https://doi.org/https://doi.org/10.1016/j.cognition.2020.104280.