A formalism for encoding morphological annotation in HunTools

András Kornai[*], Péter Rebrus[**], Péter Vajda[*] Péter Halácsy[***], András Rung[**], Viktor Trón[†]

The MorphBase library contains all word-oriented routines of the HunTools NLP toolkit currently under development at the Budapest Institute of Technology Media Education and Research Center (see `http://www.szoszablya.hu`). Morphological generation requires paradigmatic information as input, e.g. that we wish to generate the accusative form of a noun or the 3sg past form of a verb, and morphological analysis will provide paradigmatic (inflectional) and possibly deeper (derivational) output. There are some standards for encoding morphological structure, such as MSD, but we have not found any that met our requirements for consistency, informativeness, and maintainability.

The paper motivates and describes in detail the encoding used in MorphBase. The treatment of derivational affixes is standard. The encoding of the inflection affixes is based on arranging paradigmatic dimensions in a simple tree structure, where only positive (marked) nodes can branch. This way, markedness conventions can trivially supply a full feature matrix, so that encoding of the unmarked (typically the most frequent) cases can be kept short. Defective paradigms are described by subtrees of the main tree, and a number of "minor categories" such as pronouns are treated quite naturally as being defective instances of their superordinate categories.

[*] MetaCarta Inc., e-mail: `andras@kornai.com`

[**] MTA Nyelvtudományi Intézet, e-mail: `{rebrus,vajda}@nytud.hu`

[***] BME Média Oktató és Kutató Központ `{hp,runga}@mokk.bme.hu`

[†] IGK, Saarland University, University of Edinburgh `v.tron@ed.ac.uk`