Seeing the forest or the trees: lessons for grammar writing from Grambank, a large-scale typological database

Grambank is a typological database project that aims to compile morphosyntactic information from all described languages (see https://glottobank.org/). As of June 2021, the dataset contains around 360,000 data points covering 195 morphosyntactic questions for over 2,000 languages. It is compiled by a team of over 75 coders, with work ongoing. This makes it the largest comparative grammatical database to date. Apart from facts about languages, Grambank also contains facts about grammatical descriptions, making it a useful tool for exploring the use of descriptive grammars in typology.

The data in Grambank is collected by coders and language specialists reading and interpreting reference grammars. Most data points are answers to yes-no questions, for example, "Are there definite or specific articles?" or "Is there an associative plural marker for nouns?". Each of these can be answered with 1 ("yes"), 0 ("no") or, if the answer is unclear or unavailable, a question mark ("?").

Coders are encouraged to provide comments when a case is not straightforward, requires reinterpretation of the data, or if the answer "no" is given because a source does not provide an explicit answer. A "no" in the value field and "not mentioned" in the comment field, for instance, reflect a coder's judgment that in a particular case, absence of evidence constitutes evidence of absence. The comments reflect the complexity of some grammatical phenomena and how difficult it is to strike a balance between fully describing a structure in a particular language and reducing it to a discrete typological feature. By exploring (1) which features are most often answered with a question mark, (2) which features receive the most comments, and (3) which features are most often marked as "not mentioned", we provide some insights into the following questions:

- On which topics are we most likely to lose nuance when translating grammatical descriptions into values of typological variables? And conversely, which topics lend themselves more straightforwardly to typological characterizations?
- Which topics are often left implicit in grammars? Where do typologists need to make an educated guess about the presence or absence of a feature? Does this vary across language families, regions, or time?
- Which topics are under-described across grammars, from a general typological perspective?
- Are more recent reference grammars more exhaustive and explicit, and thus easier to process for a typological questionnaire, than earlier grammars?

To address these questions, we explore the comment field entries and data points in Grambank. We analyze the occurrence of comments and question marks across grammatical subject areas and source grammar characteristics (e.g. publication date). Adopting the perspective of grammar users and typologists, this survey will contribute to the discussion of the role of typology in grammar writing, highlight some challenges to align the priorities of both endeavors, and provide some historical perspective on the development of grammar writing.