Keyword Spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions

Harald Hammarström
Uppsala University
harald.hammarstrom@lingfil.uu.se

Nearly all global-level databases with structured information about the languages of the world have been constructed manually (see, e.g., the listing at languagegoldmine.com). Manual data collection by humans is a time-expensive enterprise — a database treating a single linguistic topic for some 200 languages is typically the size of a PhD project, whereas the world has 7 000 languages and there is grammatical information for over 4 500 (see glottolog.org, the remaining 2 500 being largely undocumented). Furthermore, human curated data is typically less transparent and allows for inconsistencies.

It is conceivable that at least some features for such databases can be extracted automatically from either raw-text data or linguistic descriptions originally written for humans. The present work addresses the latter case, in its simplest form: extraction of information — more specifically typological features — of language from digitized full-text grammatical descriptions. In particular, we focus on the prospects of keyword extraction, i.e., extracting information which is signalled by a specific keyword. For example, keywords like `classifier`, `suffix`, `preposition` or `inverse` signal the existence of the corresponding grammatical element and the existence of the grammatical element is signalled, perhaps not exclusively, but at least very frequently with the term in question. In contrast, other grammatical features, such as whether the verb agress with the agent in person, may be expressed in a myriad of different ways across grammars and cannot be associated with a specific keyword. Keyword-signalled features are, of course, far simpler to extract, but not completely trivial, and hence the focus of this work.

The data for the experiments in this essay consists of a collection of over 10 000 raw text grammatical descriptions digitally available for computational processing (Virk et al., 2020). The collection can be enumerated using the bibliographical- and metadata is contained in the open-access bibliography of descriptive language data at glottolog.org. The grammar/grammar sketch collection spans no less than 4 527 languages, very close to the total number of languages for which a description exists at all (Hammarström et al., 2018).

At first blush, the problem might seem trivial: simply look for the existence of the keyword and/or its relative frequency in a document, and infer the feature associated with the keyword. Unfortunately, to simply look for the existence of a keyword is too naive. In many grammars, keywords for grammatical features do occur although the language being described, in fact, does not exhibit the feature. For example, the grammar may make the explicit statement that there are "no X" incurring at least one occurrence. Also, what frequently happens is that comments and comparisons are made with other languages — often related languages or other temporal stages — than the main one being described. Furthermore, there's always the possibility that a term occurs in an example sentence or text. However, such "spurious" occurrences will not likely be frequent, at least not as frequent as a keyword for a grammatical feature which actually belongs to the language and thus needs to be described properly. But how frequent is frequent enough? A grammatical description may be modeled as a mix of four classes of terms: genuine keywords, noise keywords, meta-language words and language-specific words. We are interested in the first class, and in particular, to distinguish them from the second class. Except for rare coincidences, the words from these two classes do not overlap with the latter two, so they can be safely ignored when counting linguistic descriptive keywords. Now, a simple model for the frequency distribution of the keywords of a grammar $G(t)$ is that it is simply composed of sample of the "true" underlying descriptive terms according to their functional load $L(t)$ and a "noise" term $N(t)$, with a weight α balancing the two:

$$G(t) = \alpha \cdot L(t) + (1 - \alpha) \cdot N(t)$$

Whenever we have more than one description for the *same* target language, we will show that it is possible to estimate the noise term and thereby obtain a frequency threshold without any further human intervention. Evaluation against manually curated databases yields 85%-90% accuracy for the simple keyword-spotting approach (similar, in fact, to the accuracy of manually curated data).

References

Hammarström, Harald, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg & Bettina Speckmann. 2018. Simultaneous Visualization of Language Endangerment and Language Description. *Language Documentation & Conservation* 12. 359–392.

Virk, Shafqat Mumtaz , Harald Hammarström, Markus Forsberg & Søren Wichmann. 2020. The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World's Languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 871–877. Marseille, France: European Language Resources Association.