A Parallel Text Approach to Measuring Forced Expressivity Across Languages

Harald Hammarström and Mikael Parkvall and Sean Roberts

MPI Nijmegen / Stockholm Universty / MPI Nijmegen

Grammars in various languages of the world differ not so much in the amount of information one *can* express but rather in the amount of information one *must* express. For example, some languages require the speaker to mark every noun explicitly as Singular, Dual or Plural whereas other languages let the speaker choose when to specify number. So far, observations on these differences have been largely impressionistic, but a formal measure is needed, e.g., for investigations various questions in psycholinguistics and language contact studies.

Ideally a formalization and comparison of various languages would depart from a large corpus of parallel text glossed according to the specific categories of each individual language. We could then proceed as follows to define a Forced Grammatical Expressivity (FGE) score. With a large number of languages, the full collection of glosses approach the limit of what grammars *can* express. We now want to measure how much of what *can* be expressed *is* expressed in a specific language. Categories across languages are rarely identical, so it is not meaningful to simply count the raw number of categories present. But given the parallelism of the text data, we can easily compute an information theoretic measure of how much categories across languages overlap (cf. Hammarström and O'Connor 2013). For example, if a language L_1 obligatorily marks Sg/Du/Pl, L_2 marks Sg/Pl and L_3 marks none then, by cooccurrence statistics i) Sg and Sg in L_1 and L_2 correspond in both directions, ii) Du and Pl in L_1 always predict Pl in L_2 while Pl in L_2 predicts Du vs Pl in L_1 by some amount between 0 and 1, and iii) the categories of L_1 and L_2 do not correspond to anything in L_3. The total set of categories in this example is five $\{Sg_1, Du_1, Pl_1, Sg_2, Pl_2\}$. The FGE of L_1 is full, as it expresses all the categories fully: the three of its own and the full implications to the two categories of L_2. The FGE of L_2 is less, since it does not predict L_1:s Du_1 and Pl_1, and the forced expressivity of L_3 is zero, as expected.

However, a resource with glossed text is not presently available even for a modest collection of languages and text mass. However, a collection of (partial) bible translations, aligned at the verse-level, is available for some 800 languages[1].

[1] We wish to thank the individual translation organizations and Thomas Mayer and associates at LMU-Marburg for access to these.

Given the verse alignment, we can make a word alignment for every pair of languages using the standard EM alignment algorithm from Machine Translation (implemented in, e.g., GIZA++, see Och and Ney 2003). The word-alignment essentially provides a crude glossing of one language in the other. With this pairwise crude glossing, we can approximate the forced-expressivity measure described in the previous paragraph.

References

Hammarström, Harald & Loretta O'Connor. 2013. Dependency Sensitive Typological Distance. In Lars Borin & Anju Saxena (eds.), *Linguistic Distances*, 337-360. Berlin: Mouton.

Och, Franz Josef & Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models . *Computational Linguistics* 29(1). 19–51.