Mining corpora for form-meaning associations: perspectives for corpus-driven typology

It has been suggested, by typologists and grammar theorists alike, that it is more useful to think of languages in terms of probabilistic patterns rather than absolute universals (Dryer, 1998; Bickel, 2007; Bresnan, 2007). At the same time, typology operates by compressing information: what is known about a language is reduced to a collection of categorical statements. These statements are derived from language data via a long chain of abstractions performed by human researchers. While very succesfull, this approach suffers from a number of limitations: a) true variation is often underrepresented, b) it often abstracts away too much, and c) the abstractions often lack transparency.

This study attempts to approach these issues by adopting an information-theoretic approach. Here, abstractions are derived directly from annotated corpus data, using a knowledge discovery algorithm of our design. To evaluate viability of the approach, we focus on clause linkage as a particularly complex area of grammar that is closely tied to discourse patterns.

We have annotated a sample of spoken corpora (6000 predicates in total) for three languages: English, Latin and Chintang. Particular care was taken to properly isolate two levels of annotation: form (what analysable structural devices does the language use to convey a message?) and meaning (what is the actual conveyed message by the given linguistic expression in the given context?). To accomplish this, we have designed descriptive models that are able to capture elements of meaning without having to refer to linguistic notions. Our annotations consisten of identity and relative status (e.g. agent-like or patient-like) of event participants, temporal properties, as well as status of the information (e.g. belief, inquiry, assumed common knowledge). The algorithm then examined predicate pairs, identifying statistically conspicuous patterns of associations between form and meaning.

The investigation of the resulting patterns reveals two primary results. First, in every of the investigated languages, the variables showed the same order of importance (in terms of relevance to pattern discovery):

> Morphosyntactic features > event time > assertion/presupposition/coreference > temporal structure > conjunctions > propositional attitude

That is, for all three languages, there were more patterns which were at least partially defined in terms of time specification than patterns defined in terms of propositional attitude or presence of conjunctions and so on. This is surprising, given the substantial differences between the languages. This result suggests that while the actual patterns can be very different between languages, the *information* that is relevant to establishing them is not.

The second result is that the algorithm was able to successfully identify meaningful patterns. The patterns included highly abstract discourse organisation devices (e.g. same-topic-information vs. new-topic-information), constructions with particular function (e.g. communication of intent and/or purpose), as well as syntactic patterns (e.g. converb constructions). Many of the discovered patterns involved non-obvious associations between the annotated variables and did not merely reflect what was annotated in the first place. This leads us to suggest that the language signal encodes much more information than what is commonly assumed.

References

Bickel, Balthasar. 2007. Typology in the 21st century: major current developments. *Linguistic Typology* 11. 239–251.

Bresnan, Joan. 2007. Is syntactic knowledge pobabilistic? experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, Berlin: Mouton de Gruyter.

Dryer, Matthew S. 1998. Why statistical universals are better than absolute universals. *Papers from the 33rd Annual Meeting of the Chicago Linguistic Society* .