A large-scale corpus-based study of NP-internal word order variation for several languages

Kristina Gulordava, Paola Merlo

University of Geneva

In this talk we will present a corpus-based study of word order variation for a number of languages. We give an accurate qualitative and quantitative description of the NP-internal word order variation based on a large-scale analysis of syntactically gold annotated corpora of 20 languages of Europe and the world, representative of 9 genera. These new data call for a syntagmatic explanation of Greenberg's Universals 18 and 20 (Greenberg, 1963).

The focus of this study are the frequencies of the orders between adjectives, nouns, and numerals. To collect the token frequencies of pre-nominal and post-nominal occurrences of modifiers from our syntactically annotated corpora with a language-independent and fast method, we convert their different annotations to the part-of-speech format defined in Petrov et al. (2012). We produce several types of counts for each language corpus. First, the frequencies of pair-wise word order variants, such as Adj Noun / Noun Adj, Num Noun / Noun Num. Second, the counts of the different placements of Adj and Num when they appear in the same noun phrase, e.g. Num Adj Noun or Num Noun Adj. Moreover, for each word order count in the corpus, we also extract its context characterized by a number of features, such as the complexity or length of the modifier, its lemma, the lemma of the head noun, the presence of punctuation.

Our analysis of the collected data provides new evidence for the debate on word order universals. Quantitative data on languages that freely allow pre-and post-nominal modification (such as Latin and Ancient Greek) suggest a linguistic constraint against Adj Noun Num word order : the observed frequency of the Adj Noun Num word order is smaller than its expected frequency if Adj Noun and Noun Num phrases could freely combine. Indeed, this data is best explained if we posit a phrase-level restriction on the cooccurrence of the two modifiers – adjective and numeral – inside the same noun phrase (cf the derivational account of Universal 20 by Cinque (2005)). Instead, restrictions on the correlation of two typological properties – the order between adjective and noun and the order between numeral and noun – as stated in Universal 18 and studied by Culbertson et al. (2012), would not distinguish between NP-internal cooccurrences of these modifiers from occurrences in separate NPs. This conclusion is further corroborated by qualitative data from languages where adjectives or numerals can be pre- or post-nominal in restricted contexts (such as Russian and Arabic).

Interestingly, using these quantitative data of NP-internal word order variation, we can also model the placement of modifiers in a language probabilistically with a parameter that reflects their complexity. Moreover, this model suggests that the post-nominal placement of complex modifiers can be to some extent predicted based on the word order properties of a language – for instance the ratio of pre- and post-nominal placement for a simple modifier – rather than the language itself.

Bibliography

Guglielmo Cinque, 2005. “Deriving Greenberg's Universal 20 and Its Exceptions”. Linguistic Inquiry, Vol. 36, No. 3. MIT Press, pp. 315–332

Jennifer Culbertson, Paul Smolensky, and Geraldine Legendre, 2012. “Learning biases predict a word order universal”. Cognition, Vol. 122, No. 3, pp. 306–329

Joseph H. Greenberg, 1963. “Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements” In: Joseph H. Greenberg (ed.). Universals of Language. London: MIT Press, pp. 73-113

Slav Petrov, Dipanjan Das, and Ryan McDonald, 2012. “A universal part-of-speech tagset”. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC’12).