Towards automatic detection of morphosyntactic systems from IGT

We present the AGGREGATION project, whose goal is to produce computational linguistic resources from text or transcribed speech corpora with IGT (interlinear glossed text) annotations and the LinGO Grammar Matrix customization system (Bender et al 2002, 2010).

The Grammar Matrix is a toolkit for developing machine-readable grammars which handle morphology, syntax and compositional semantics. Such grammars can facilitate language documentation in many ways: they enhance the precision of analyses, allow for the efficient discovery of exceptions to existing analyses (Baldwin et al 2005) and support the development of treebanks (Oepen et al 2004, Bender et al 2012). Treebanks facilitate the creatation of further computational tools and are a rich source of comparable data for qualitative and quantitative work in typology, grounding higher level linguistic abstractions in actual utterances in a computationally tractable fashion.

The Grammar Matrix consists of a shared core grammar and a range of typologically-grounded "libraries", providing analyses of cross-linguistically variable phenomena. It is accessed through a web-based questionnaire which elicits linguistic descriptions and outputs small but functional grammar fragments able to map between surface strings and semantic representations.

The AGGREGATION project aims to build an automated system for answering the Grammar Matrix questionnaire on the basis of IGT produced in language documentation projects. We build on the work of Xia and Lewis 2007 on enriching IGT by parsing the translation line and projecting that structure through the gloss line to the source line. From enriched IGT, we aim to extract the various information required by the Grammar Matrix quesitonnaire: (i) lexical type definitions and mapping of stems to lexical types, (ii) morphotactics and the morpho-syntactic/-semantic features associated with affixes, and (iii) morphosyntactic systems. In this talk, we will present preliminary results on automatically discovering two aspects of this third category: word order and case alignment.

References

Baldwin, T., Beavers, J., Bender, E. M., Flickinger, D., Kim, A., & Oepen, S. (2005). Beauty and the beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar — and the corpus. In S. Kepser & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives* (pp. 49–69). Berlin: Mouton de Gruyter.

Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., & Saleem, S. (2010). Grammar customization. *Research on Language & Computation*, 1-50.

Bender, E. M., Flickinger, D., & Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In J. Carroll, N. Oostdijk, & R. Sutcliffe (Eds.), *Proceedings of the workshop on grammar engineering and evaluation at COLING'02* (pp. 8–14). Taipei, Taiwan.

Bender, E. M., Ghodke, S., Baldwin, T., & Dridan, R. (2012). From database to treebank: Enhancing hypertext grammars with grammar engineering and treebank search. In S. Nordhoff & K.-L. G. Poggeman (Eds.), *Electronic grammaticography* (pp. 179–206). U. of Hawaii Press.

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Research on Language & Computation*, *2*(4), 575 – 596.

Xia, F., & Lewis, W. (2007). Multilingual structural projection across interlinearized text. In *Naacl-hlt 2007.* Rochester, NY.