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 creation 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 questionnaire: (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


