Computer-aided extraction of grammatical morphemes and constructions from a parallel corpus

Östen Dahl

The possibility of automatic or semi-automatic methods for extracting grammatical morphemes and constructions from a parallel corpus rests on the assumption that elements with similar meanings and/or grammatical functions will have similar distributions in texts which are translational equivalents of each other. For a typologist, it is a somewhat frustrating fact that most parallel texts are restricted to European and a few major non-European languages. One exception to this is the Bible. In recent years, the availability of Bible translations in electronic form has increased dramatically, with more than 200 languages from Papua New Guinea alone. Although Bible translations are far from ideal representations of spoken languages, the information that can be obtained by analyzing them may be no less reliable and often goes far beyond what can be found in grammatical descriptions of the same languages. In the talk, advantages and drawbacks of the parallel corpus approach to grammatical typology will be discssed and some results from probing a parallel corpus of translations of the New Testament from around 800 languages will be reported.