Definiteness in languages with and without articles

Laura Becker
Universität Leipzig
laura.becker@uni-leipzig.de

This talk presents an empirical approach to compare the expressions for (in)definiteness in languages with and without articles based on parallel movie subtitles. In order to compare coding strategies for (in)definiteness in European languages, parallel subtitles from 5 movies have been used. From those subtitles, 500 referring expressions with sufficient similarity have been extracted for German, Spanish, Romanian, Hungarian (def. and indef. articles), Macedonian, Bulgarian (def. article), Russian, Czech, Estonian, Finnish (no articles). For the annotation, the following parameters have been considered: different types of definite and indefinite contexts; syntactic position, semantic features of the noun, other elements in the noun phrase, pronominal forms. Including uses of pronouns and pro drop allows us to look at a wider range of definiteness-sensitive contexts that are relevant to the cross-linguistic variation of definiteness coding strategies. As for the use of articles, restrictions show a high degree of variation across languages for both indefinite and definite articles, e.g. the occurrence of indefinite articles in predicate position, the use of definite articles vs. demonstratives in anaphoric and deictic contexts, with abstract nouns, or in generic contexts. In languages without indefinite articles, the use of the numeral *one* to mark non-identifiabilty can be tied to "pragmatically specific contexts", i.e. newly introduced referents who are not accessible to the addressee yet, but which will stay relevant to the discourse (2). New referents with less discourse relevance are typically not marked. Random forest models show that languages without articles showed the most relevant factors to be the semantics of the noun, possessive marking, and other elements in the noun phrase, whereas the syntactic position played a less significant role. The variation of the coding strategies in the different languages can also be used to shed more light on the values of definiteness: based on their marking, all languages so far addressed suggest a major three-way distinction between anaphoric definites (pronouns and pro drop play a significant role here), non-anaphoric definites, and indefinites, rather than a binary split into definite vs. indefinite.