

Authentic speech data and linguistic theory: chances and challenges

Ingo Plag

Heinrich-Heine-Universität Düsseldorf

Speech corpora have been widely used in natural language processing and the IT industry to develop speech recognition algorithms and dialogue systems. But speech corpora have also been used in linguistics to investigate questions concerning the nature of language and language production. In this talk I will show that the use of authentic speech as represented in such corpora can help us discover unexpected and important facts about language which would otherwise go unnoticed, and which often present formidable problems for established linguistic theories. Or we can use speech corpora to empirically test predictions arising from particular theories. The analysis of speech corpus data, however, also comes with enormous methodological challenges that can make the use of speech corpora tedious, difficult, undesirable or even nonsensical. The talk will illustrate these chances and challenges using examples from my own lab.