Ludger Paschen

Methods in Automated Register Identification on a German Newspaper Corpus

One of the challenges of computational linguistics is making machines learn things we do not even bother think about in our everyday life. One such case is register identification. We all can tell a crime novel from a recipe or a report about a pop concert by pure intuition. When it comes to programming a machine to learn the difference between such registers, one could think of using content words such as 'cop', 'pot' and 'pop'. However, this would imply a heavy reduction of register to solely lexical criteria. Shouldn't it be possible to implement a register detection based on grammatical features like person or phrase size? Which features can prove useful for this task, and why are some more useful than others? In this talk, these questions will be addressed. This includes the presentation and discussion of a classifier for a German newspaper corpus which the author contributed to.