This talk presents research from two projects that deal with processing source code and natural language respectively. In the case of source code, the problem domain is authorship attribution [1]. In the case of natural language, the problem domain is paraphrase acquisition [2].

Topic 1: To attribute authorship means to identify the true author among many candidates for samples of work of unknown or contentious authorship. Authorship attribution is a prolific research area for natural language, but much less so for source code, with nine research groups having published empirical results concerning the accuracy of their approaches to date. This research includes an information retrieval approach with indexed representations of source code samples and the use of similarity measures for evaluating author style and authorship. In this talk, the focus is given to our investigation of how timestamps affect authorship attribution accuracy. These results clearly show that author style improves over time, and they suggest that it takes one semester for author style to mature.

Topic 2: To paraphrase means to rewrite content whilst preserving the original meaning. This paper contributes to paraphrase acquisition and focuses on two aspects that are not addressed by current research: (1) acquisition via crowdsourcing, and (2) acquisition of passage-level samples. We present the Webis Crowd Paraphrase Corpus 2011 (Webis-CPC-11), which recently formed one part of the PAN 2010 international plagiarism detection competition, and was developed using Amazon's Mechanical Turk. This research includes machine learning experiments to explore if passage-level paraphrases can be identified in a two-class classification problem using paraphrase similarity features. The results for several cost-analysis scenarios show attractive time and money savings when using our machine learning approach compared to the manual one.

Bibliography:
Kurzlebenslauf:

Dr. Steven Burrows has been working as a postdoctoral researcher at the Bauhaus-Universität Weimar since the start of 2011. His current work is in digital engineering, data mining, machine learning, paraphrase acquisition, and plagiarism detection. He was previously with the RMIT University in Melbourne, Australia, from 2000 to 2010 as a researcher, sessional lecturer, and student. His research down under had an emphasis on information retrieval and culminated in the award of his PhD entitled “Source Code Authorship Attribution”.

Ansprechpartner:

Prof. Dr. Iryna Gurevych (gurevych@tk.informatik.tu-darmstadt.de, Tel.: 06151/16-5411)