L3S-Experts talk at the 4th Workshop for "Internetlinguistik und Korpusanalyse"

L3S presented three of its projects at the Fourth Hanover Workshop on "Internetlinguistik und Korpusanalyse" (Internet Linguistics and Corpus Analysis). The workshop was organised by Prof. Dr Peter Schlobinski at the Department for German at the University of Hanover, with support of the Gesellschaft fuer Deutsche Sprache (Society for the German Language). It attracted a professional audience from all over Germany and German-speaking neighbouring countries with an interdisciplinary program that focused on the analysis of language as displayed in social media and other web-specific genres. The analysis of web language is essential for varied applications such as opinion analysis and forecasting, geolocation of social media users, the analysis of language change and the sociolinguistic analysis of language use, for example among young users.

The speakers of the L3S contributed to the spirited discussion from a computational perspective, presenting the following three projects:

Dr Thomas Risse, Research Group Leader, talked about the collection of large and varied web corpora over time via web archiving and web crawling as conducted in L3S’ Alexandria project. Only temporal collections of language can answer questions about language change or change in user behaviour, making automated web archiving a prerequisite for a wide variety of linguistic analyses.

In contrast, Dr. Katja Markert’s, Senior Researcher at L3S, talked centered on the categorization of documents in already collected corpora or archives to allow focused search or analyses by subdividing a corpus into different genres, such as news, scientific papers etc.

Traditionally this has been a task conducted by linguistic experts, which is, however, not scalable to very large corpora. She showed that crowdsourcing can instead accomplish the task speedily, cheaply and accurately.

Asmelash Teka Hadgu, together with Netaya Lotze (German Linguistics, University of Hanover) and Dr. Saskia Kersten (University of Hertfordshire), use linguistic insights to automatically assign English tweets to British, Australian or US American English users. This is an important step for better automatic geolocation of tweet users when user profiles are incomplete.

More information about the workshop (including copies of all talks) can be found at https://www.mediensprache.net/de/aktionen/2015/workshop/