

Workshop (III)

Corpus Linguistics on the Web: Introducing WebCorp Linguist's Search Engine

Convenor: Antoinette Renouf (Birmingham City University, UK)

Time: Wed 26th May, 2.00 - 6.15 pm

Room: Aachen/Bremen

14.00 ANTOINETTE RENOUF

Opening

14.15 ANDREW KEHOE (Birmingham City University, UK)

WebCorpLSE Background and Design

14.35 MATT GEE (Birmingham City University, UK)

WebCorpLSE Features and Operation

14.55 Discussion and Questions

15.15 Coffee break

15.45 ANTOINETTE RENOUF (Birmingham City University, UK)

WebCorpLSE and neology

16.15 NATALIE KÜBLER (Université Paris Diderot, France)

WebCorp, Collocations, Specialised Translation: How not to google the right word?

17.00 CHRISTIAN MAIR (University of Freiburg, Germany)

The role linguistics in the emerging "science of the web": reflections prompted by WebCorpLSE

17.45 Discussion and questions

18.15 ANTOINETTE RENOUF

Closing

Workshop description

This ICAME 2010 pre-conference workshop will introduce the WebCorp Linguist's Search Engine (WebCorpLSE) and the new possibilities it opens up for web-scale corpus-based study. The current publicly-available version of WebCorp was first launched a decade ago (<http://www.webcorp.org.uk>), a system relying on standard web search engines such as Google, adding layers of refinement specifically for linguistic analysis.

WebCorpLSE is designed to bypass the commercial search engines. It crawls and processes the web to build a 10 billion word (7 terabyte) corpus, including a multi-terabyte 'mini-web', designed to approximate a microcosm of the web itself. In addition, WebCorpLSE has built a newspaper sub-corpus of UK broadsheets from 1984 to the present, and recent issues of other UK and international newspapers. We also work with School colleagues to build easily searchable collections to assist in their research and teaching, including sub-corpora e.g. of blogs, science fiction and major English literary works.

The new architecture has allowed us to enhance the sentence boundary detection, date identification, 'junk' (or 'boilerplate') removal, as well as collocational and other statistical analysis options currently available in WebCorp. Additional pre-processing includes grammatical tagging, language detection, full pattern matching and wildcard search.

In this workshop, the developers of WebCorpLSE will introduce its new features and demonstrate how these can be used. Papers dealing with applications and issues, from specific to general, will be offered by Christian Mair, Natalie Kübler and Antoinette Renouf.