

The linguistics group of IEAS is happy to announce a full-day hands-on workshop on

Doing research with the COW16 corpora

given by **Roland Schäfer (FU Berlin)**

When? 20.2.2018, 9ct – 17.00 **Where? IG 3.201**

In this practical workshop, we very briefly review the structure of the COW16 web corpora (English: 16.5 billion tokens, French: 10.8 billion tokens, German: 19.8 billion tokens, Spanish: 7.1 billion tokens) and the basic access method using the NoSketchEngine front end at www.webcorpora.org. We then proceed to more advanced topics such as:

- (1) making simple queries using Python and the new SeaCOW wrapper around NoSketchEngine's Manatee library, which already allows users to go beyond what the interface can do (for example, sentence-wise de-duplication, export of concordances in a convenient CSV format)
- (2) using the SeaCOW interface to create plug-in objects which filter or modify queried material on the fly (for example, decoding dependency information and refining searches based on dependency trees)
- (3) using the separate document-level databases, possibly including the COREX data for the German DECOW16 created by COW and the IDS Mannheim, which provides extensive Biber-style information about documents.

Simple SeaCOW queries as described in (2) are Python three-liners, and no previous Python knowledge is required. Advanced use such as described in (3) requires some Python skills. However, the simple dependency decoder plugin used for demonstration can be utilised in a limited way without writing any code (almost). As opposed to previous workshops, we use RStudio Server as a convenient Python IDE, and neither Terminal/Bash nor Vim/Emacs skills are required any longer.

The main idea of this workshop is to provide hands-on guidance for researchers and students who are planning to search for specific information in the corpora as part of their project. Therefore, participants are highly encouraged to

1. create an account at www.webcorpora.org ahead of the workshop (or reactivate their old account if they have worked with the corpora before),
2. send a brief message to Roland Schäfer (roland.schaerfer@fu-berlin.de) a few days before the workshop describing the kind of linguistic structures they would like to find in the COW corpora.

This will allow me to determine whether and how these structures can be queried successfully.

Everyone is cordially invited!

Gert Webelhuth, Manfred Sailer, Frank Richter