Pre-conference workshop:

Corpus pitfalls: dealing with messy data (and other traps for the unwary)

Convenors: Mark Kaunisto (Tampere University),
Marco Schilk (Universität Hildesheim),
Jukka Tyrkkö (Linnæus University)

Wednesday August 18th, 2021

10:15-10:30 Introduction/Opening discussion

10:30-11:00 Mark Kaunisto: *Proper names as potentially problematic items in corpora*

11:00-11:30 Turo Hiltunen: *Issues in using British Library Newspapers as a corpus*

11:30-12:00 Stefan Hartmann: *Open Corpus Linguistics: Overcoming Rissanen's problems (and others) with open data*

12:00-13:00 **lunch break**

13:00-13:30 Marcus Callies: *Challenges in the annotation and analysis of learner corpora*

13:30-14:00 Daniel Sundberg: *Corpus Categories: What and for whom? When special corpora meet general corpora in comparative studies in literature*

14:00-14:30 Jukka Tyrkkö and Sophie Raineri: *Empirical perspectives on the reliability and accuracy of collaborative pragmatic annotation*

14:30-15:00 **coffee break**

15:00-15:45 discussion on pre-recorded presentations:

- 15:00 Jesse Egbert, Tove Larsson, and Douglas Biber: *Research design in quantitative corpus linguistics: Critical reflections and suggested improvements*
- 15:15 Fabian Vetter: *Register variation & comparability in parallel corpora*
- 15:30 Filip Miletic, Anne Przewozny-Desriaux and Ludovic Tanguy: *Modeling fine-grained sociolinguistic variation: the promises and pitfalls of Twitter corpora and neural word embeddings*

15:45-16:30 Open discussion