Interactive practical workshop

Automatic identification of formulaic sequences in (fairly) big data: Practical introduction to a procedure

Andreas Buerki
Cardiff University

In this workshop, I will present an automatic procedure for extracting formulaic sequences from corpus data and guide participants through its practical implementation using example data and software tools. By the end of the workshop, participants will be able to use the N-Gram Processor (Buerki 2013) and the software SubString (Buerki 2011) to extract formulaic sequences from corpus data of their own. Participants will also be aware of some of the strengths and weaknesses of the procedure and its theoretical underpinnings. The workshop is divided into three parts.

The first part addresses the question of how (or even whether) extraction procedures relate to theoretical understandings of formulaic sequences. While the procedure presented takes as its starting point a constructionist view of formulaic sequences, which identifies them as units of form and associated meaning that are conventional in a speech community, this understanding is briefly located within a broader context of thinking on the nature of formulaic sequences. Implications for identification procedures, including of views based on psycholinguistic processing, the traditional phraseological criterion triplet of polylexicality, idiomaticity and fixedness or the frequency-only approach that produces lexical bundles will also be discussed.

In part two of the workshop, participants are invited to work through a hands-on example of how formulaic sequences are automatically extracted from corpus materials following the five-stage extraction procedure outlined in Buerki (2012):

- Data preparation (normalisations, formatting)
- N-gram extraction using the N-Gram Processor (including the use of stop-lists)
- Consolidation of different length n-grams to derive a unified list using SubString
- Filtering (application of frequency thresholds and a lexico-structural filter)
- Assessment of accuracy and recall

This includes an introduction to the installation and use of the necessary open-source software tools. A corpus of Wikipedia texts will be provided as example data.

In the final part of the workshop, strengths and limitations of the procedure will be discussed as well as potential alternatives. Strengths include the methodological transparency of the procedure and the ability to process large amounts of corpus data (subject to sufficiently powerful hardware); the limitations consist mainly of the flipside of this, namely that it is less accurate as an automatic procedure when applied to small amounts of data (< 1 million words). In a final discussion section, participants are invited to share their views on any aspect of the workshop topic including how remaining challenges might be overcome.

References

Buerki, A. (2013). *N-Gram processor 0.4* [Computer Software]. Available at http://buerki.github.io/ngramprocessor/

Buerki, A. (2011). *SubString* [Computer Software]. Available at http://buerki.github.com/SubString/

Buerki, A. (2012). Korpusgeleitete Extraktion von Mehrwortsequenzen aus (diachronen) Korpora. In N. Filatkina, A. Kleine-Engel, M. Dräger, & H. Burger (Eds.), *Aspekte der historischen Phraseologie und Phraseographie* (pp. 263-92). Heidelberg: Universitätsverlag Winter.