Reading concordances with algorithms

Nathan Dykes, Stephanie Evert, Michaela Mahlberg, Alexander Piperski
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Abstract

Concordance analysis has long been central to corpus linguistics and other text-based disciplines, including digital humanities, computational social sciences, and computer-assisted language learning. It gives researchers a systematic lens for observing and interpreting patterns of language use, integrating both quantitative and qualitative perspectives. By focusing on a single search word or phrase in a context-limited display—commonly known as a KWIC (Key Word In Context)—scholars can investigate various aspects of its usage and meaning.

In spite of its wide applications, concordance reading has seen little innovation to date. Popular functions of concordance tools are still the traditional approaches, such as sorting lines alphabetically by the left or right context of the node or filtering for specific words. Another challenge for concordance reading is the documentation of the research process and methods applied, in order to ensure reproducibility. The tutorial addresses these challenges by introducing both a taxonomy of concordance-reading strategies and a set of computational algorithms that build on these strategies to organize large amounts of textual data efficiently and transparently. Through hands-on exercises using the new Python library FlexiConc (https://pypi.org/project/FlexiConc/) integrated into CLiC (https://clic-fiction.com/), the tutorial will demonstrate how to apply robust concordance reading approaches to a variety of research contexts.

The tutorial starts with an introduction to concordance analysis, including its place in the continuum of quantitative and qualitative research. We cover the most common general strategies for concordance analysis: selecting, sorting, and grouping lines, and show how each of them can aid interpretation. Participants will also learn about basic formal definitions and mathematical properties of the computational algorithms that underlie these strategies. We will discuss how algorithms extend beyond simple alphabetical ordering, opening up new possibilities for advanced text analysis.

The tutorial will include practical exercises, in which participants explore the functionalities of the FlexiConc library and its web interface. This library is designed to support a wide range of concordance reading strategies and to document user decisions in a systematic way. The CLiC web interface is designed to be intuitively accessible and to enable convenient interactive exploration. We introduce the concept of an ‘analysis tree’ to ensure the reproducibility and accountability of concordance research. By using a tree structure to trace the decisions taken when selecting lines from concordances, ordering, and grouping them, we can document not only the final results but also the process that led there. This approach fosters transparency, which is crucial for collaborative and interdisciplinary projects, as well as for replicating or extending research.

Tutorial outline

Introduction to concordance analysis: fundamentals and strategies

- Participants are introduced to basic concepts of concordance analysis. After a brief definition of fundamental terms and concepts, we give an overview of concordance software and its functionalities and allow participants to explore selected example concordances. Participants will be encouraged to share their observations on linguistic patterns as they work with existing concordancing tools.
- We introduce strategies for organizing concordances (different types of selecting, ordering, and grouping). Each strategy is discussed with regards to its purpose, and how it may be combined with other strategies. In a hands-on exercise, participants apply different strategies themselves to example data and compare their observations to those from the step before to see how the application of dedicated strategies helps with concordance organisation and enhances systematicity.

Computational algorithms

- Participants are introduced to our algorithmic approach to concordance reading, which extends the basic strategies and enhances their flexibility.
- In a hands-on exercise, participants try out different concordance algorithms, including complex applications such as clustering, which are not widely available in current concordance tools. They can work with the web interface of our library on a public server, so no software installation is required (but advanced participants are welcome to work directly with the Python library via Jupyter notebooks, which enables them to process their own corpus data).

Analysis trees for research documentation (~ 30 minutes)

- We discuss reproducibility as a central challenge for concordance analysis and how this problem can be solved with the help of the ‘analysis tree’. The tree-like display, accessible through the web interface of our library, enables users to trace and illustrate decision-making during concordance analysis.

Summary and outlook (~ 20 minutes)

Target audience

This tutorial targets an interdisciplinary audience, including students and researchers in corpus linguistics, general linguistics, computational linguistics, digital humanities, and computer-assisted language learning. We will keep the technical discussion to a manageable level to accommodate participants from both technical and non-technical backgrounds. Those interested in advanced techniques, such as more low-level concordance processing using Python, will be directed towards additional online resources and are invited to attend a follow-up tutorial at the KONVENS conference.

Technical requirements

Participants are encouraged to bring their own laptops, and technically more advanced participants may want to install Jupyter and FlexiConc on their own computer. However, a modern Web browser is sufficient to follow all hands-on exercises.