KonText – a modern, customizable corpus query interface

Tomáš Machálek (Institute of the Czech National Corpus,
Charles University – Czech Republic)

We present a fully featured corpus query interface based on open-source system NoSketch Engine. The aim is to present a customizable tool providing features that emerged from long-term feedback given by researches and users of the Czech National Corpus (CNC).

To be able to introduce a fully operational service fitting the needs of our present-day users in a reasonable time, we based our approach on our extensive experience with the *NoSketch Engine* (NoSkE) corpus query engine which is maintained and developed by Lexical Computing Ltd. as a simplified, open-source version of their commercial software Sketch Engine (SkE). Although open-source, NoSkE properties and project direction are tightly coupled with its commercial counterpart making the possibility of its direct adaptation to our needs through community-driven development rather unfeasible.

Our solution is based on reusing the core part of NoSkE – the *Manatee-open* library and related utilities which provide essential indexing, searching and analytic functions. Such a decision ensures that KonText can operate on the same indexed data and provide the same analytic results as NoSkE while avoiding development of highly specialized and complex text search software. Our primary *focus is set on the development of a new interface and additional services allowing advanced functions for both querying corpora and result presentation/manipulation*. An integral part of the project is to design the application as a set of interchangeable building blocks communicating with each other via defined interfaces. This allows other institutions and individual developers to adapt KonText to their specific needs without rewriting its core. KonText is a fully operational and mature software deployed at the CNC since 2014 that currently handles more than 1,650 user queries per day (any further actions that operate on query results, e.g. sorting or filtering, are not included in this count).

KonText is a three-layer system. *The first (bottom) layer* is composed by different backend server services. Unmodified *Manatee-open* library plays the primary role here but there are also our custom functions (e.g. database-stored meta-data, calculation control) present there. KonText provides an increased emphasis on scalability which means that in case of growing user base, another server can be added for a better system performance.

The third (highest) layer is the interface itself which is loaded and executed in user's web browser. The source code in this case has been mostly written from the ground up and remaining pieces of the original NoSkE code are gradually being replaced.

The second layer sits between the two and provides a communication between user interface components and backend services. The source code of this layer is derived from NoSkE with some modules intentionally kept with minimum changes (esp. the ones tightly related to Manatee-open library) while others heavily rewritten or added (request processing, custom data access).

From the end-user perspective, we try to keep the core user experience like the one provided by NoSkE. This makes the possible user transition easier as the users are already familiar with basic application operation.

Our additions can be divided into three categories. The first category contains functions related to *query construction and data selection*:

- a module for interactive text selection (based on a combination of criteria) that facilitates user creation of tailor-made subcorpora with the possibility to define also alignment-based subcorpora,
- a function allowing custom text type proportions in subcorpora allowing the users to define a balanced corpus according to their specific requirements,
- a fully re-editable query where any operation (query, filter, sorting, sample) can be changed while keeping the other applied operations intact,
- a local CQL (Corpus Query Language) parser – currently used as a syntax checker but also allowing further future improvements in interactive CQL editing,
- a convenient visual data-driven widget allowing users to select a required tag value (for positional PoS tag formats),
- a storage for recent queries allowing users to review their previous queries for later reuse with data stored on server making it available from any computer and accompanied by additional useful information (date and time, queried corpus, used query type and parameters),
- two alternative modules for finding and selecting a corpus to work with (search by keywords and description or by expanding a category tree).

The second category of KonText's additions and changes is related to *result presentation and manipulation*. These include:

- manually selecting concordance lines and attaching custom numerical labels to them. Such a selection can be further filtered, exported or passed to other users via a URL address,
- support for direct export to Microsoft Excel format,
- an extended support for spoken corpora:
  - regions of texts can to be accompanied by audio chunks for direct playback,
  - intuitive visualization of dialog structure in spoken conversational corpora,
- rendering of dependency syntax trees.

The third category contains *user-transparent changes motivated by improving performance and modularity of the system*. While almost invisible for a user, these changes often belong to the most arduous ones as they intervene the core functionality of the original NoSketch Engine code. These changes include:

- a rewritten asynchronous processing of concordance search which now allows distributing the function to multiple servers,
- an improved caching of frequency distribution and collocation result pages allowing a faster navigation between pages.

KonText is an open-source software licensed under GNU GPL 2 (same as NoSkE) and publicly available on GitHub at https://github.com/czcorpus/kontext. A running production version can be found at https://kontext.korpus.cz/. Its development is maintained by the Institute of the Czech National Corpus but it is open to other contributors as well. Recently, a cooperation on further development of KonText has been established with the Institute of Formal and Applied Linguistics (Faculty of Mathematics and Physics at Charles University) and KonText has been adopted as the query engine at the LINDAT/CLARIN repository. There is also a positive feedback from individuals that examine possibilities of deploying KonText in their (mostly academic) environment. KonText is in daily use by researchers and students across the Czech Republic and it is under an active development with continuous bug fixing and major updates approximately every six months.

References

Rychlý, Pavel. Manatee/Bonito - A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. p. 65-70. ISBN 978-80-210-4471-5.