

## Sketch Engine: The Cutting Edge of Digital Lexicography

Vladimir Benko

Ludovít Štúr Institute

Slovak Academy of Sciences

vladob@juls.savba.sk

While in the not very distant past, one of the main problems in lexicography was the lack of evidence – even for medium-frequency lexical units that could be found in corpora, let alone the rare phenomena that, nonetheless, needed for description in unabridged dictionaries. With the advent of multi-billion-token corpora we face quite the opposite problem: too much data even for medium-frequency lexical units. For example, in the 6-billion-token *Araneum Russicum Maximum* (Russian web corpus), we can find 31,066 occurrences of the adjective *венгерский* (“Hungarian”) – a number that clearly cannot be analysed by “sequential” reading all concordance lines.

To cope with the problem of abundance of data that typically needs to be analysed and lexicographically processed in a very short time, we need tools that could do some sort of aggregation and summarization. Our presentation will introduce the (most likely) best tool of this kind – the *Sketch Engine*. Designed by a prominent British linguist and computational lexicographer *Adam Kilgarriff* and implemented by a team of computer scientists and computational linguists from the *Masaryk University in Brno* led by *Pavel Rychlý*, the *Sketch Engine* is being used to write dictionaries by leading commercial dictionary publishers, as well as by academic lexicographic institutions in many European countries.

An annotation taken from the *Sketch Engine* site (<https://www.sketchengine.co.uk/>) shows the main features of the system.

Sketch Engine is for anyone wanting to research how words behave. It is a **corpus** software interface which works online and offers many corpora in many languages. You can see a **concordance** for any word, phrase or grammatical construction. Also, you can create corpora. Many features allow you specific search with complex parameters. One of them is **word sketches** (automatic one-page corpus-derived summaries of a word's grammatical and collocational behaviour). Finally, check your thinking about words in million sentences of natural language by a few clicks.

In the interface, you find more than 200 corpora in 82 languages. The largest corpora (TenTen corpora family) of main languages contain from 2 to 15 billion words. Similarly, you can create such corpus including automatic disambiguation with tool **corpus architect**. Thanks to various functions, e.g. **term extraction**, **thesaurus**, **word list** and others, you have new possibilities of studying languages. Our services are used by well-known world dictionaries and companies: Oxford University Press, Cambridge University Press, Dictionnaires Le Robert, Shogakukan, and others. Join them.

Sketch Engine is used to write dictionaries by

