While in the not very distant past, one 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 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.