Attempts and examples for the discovery of hidden information of Concise explanatory dictionary of Hungarian (2nd edition, 2003)

Mártonfi Attila

MTA–ELTE Research Group of Academic Dictionary of Hungarian
rumci@nytud.hu

Keywords: lexicography, knowledge discovery, etymological statistics, Concise explanatory dictionary of Hungarian

Knowledge discovery and data mining – as its part – are trendy areas of IT, their aim is utilizing characteristically commercial databases. However the goal (namely extracting as much hidden data and unknown patterns as possible by machine) is essentially is the same as the most general goal of scientific research, therefore at least partially its approach and toolkit are applicable to lexicographical databases. (Since the size of lexicographical databases is usually smaller by orders of magnitude than monumental commercial databases occurring with the primer area of data mining, the device requirement of the operations is significantly less and the extractable information is more restricted.)

The first notable lexicographical database of Hungarian is Papp Ferenc's *Reverse-alphabetized dictionary of the Hungarian language* (VégSz.) and its derivative database on PC. The database which is the base of Papp's dictionary had four fields more than the paper-version: the length in characters, the number of senses in ÉrtSz. *(Explanatory dictionary of the Hungarian language),* the etymology based on *Etymological dictionary of Hungarian,* and the usage label given in the head of entries in ÉrtSz. – because of typographical reasons these are omitted from the paper-version and its derivative database.

The new edition of *Concise explanatory dictionary of Hungarian* (ÉKsz.[2]) – as an up-to-date lexicographical project should be – was first prepared as an XML document, and though its grammatical information (constituting the skeleton of VégSz.) is substantially more poor, with suitable conversions a more complete and more modern data tablet can be generated. It is more modern, because ÉKsz.[2] provides up-to-date etymological facts about the widest group of Hungarian words, and it is more complete, since apart from the part-of-speech and usage labels and the numbers of drawn senses all of the entries in this dictionary have the absolute frequency based on *Hungarian National Corpus,* furthermore the word-length in the number of phonemes or syllables can be coded.

With some simple queries the generated relational database gives token and type frequency indices of various etymology, usage label, part-of-speech or number of senses word-groups. Such token frequency indices – for want of a satisfactory database or

corpus background – formerly could not have been calculated; the type frequencies provide the possibility for comparision with Papp's examinations based on former sources.

With the toolkit of data mining more interesting analyses could be performed to discover hidden patterns of the above parameters by means of extracting association rules.