Veronika Lipp
Department of Lexicology and Lexicography
Research Institute for Linguistics,
Hungarian Academy of Sciences, Budapest
e-mail: lipp.veronika@nytud.mta.hu

Comprehensive Dictionary of Hungarian

KEYWORDS: Hungarian, dictionary, historical corpus
SŁOWA KLUCZOWE: język węgierski, słownik, korpus historyczny

Wielki słownik języka węgierskiego

STRESZCZENIE: Celem artykułu jest zwięzły opis *Wielkiego słownika języka węgierskiego* (*A magyar nyelv nagyszótára*). Słownik ten bazuje na danych korpusowych i tworzony jest jako baza danych w formacie XML. Artykuł przedstawia ogólną charakterystykę leksykograficzną słownika, materiał źródłowy wykorzystany przy jego tworzeniu, możliwości przeszukiwania korpusu historycznego oraz różne zastosowane typy haseł.

The Department of Lexicology and Lexicography's main project at present is to compile the Comprehensive Dictionary of Hungarian *(A magyar nyelv nagyszótára,* henceforth *Nszt.)*. This dictionary has a long and complicated history: the original idea emerged in the middle of the 19th century, data collection was started, but we had to wait more than 150 years until the real lexicographic work was begun. In 1984 the Hungarian Academy of Sciences decided to start a new dictionary project which was based on computer and on a historical corpus. The compilation of the list of the headwords as well as the grammatical and semantic description of the lexemes is based on corpus evidence. Therefore, the corpus building started, and from 2002 we started to compile the first entries on the basis of our new editorial manual.

The first two volumes were published in 2006, followed by the third and fourth volumes in 2011. The fifth volume was published in 2013, and in January 2017 the sixth volume (*Di–Ek*) came out.

The dictionary will principally contain the literary and common vocabulary of the Hungarian language in one hundred and ten thousand entries. It is in XML database format and will appear in both printed and electronic form. It is a paper dictionary originally, but from January 2017 it is also available online (http://nagyszotar.nytud.hu/index.html). In September 2015

the department started to use the widely spread, open-source version control system, the SVN (Subversion), which makes it possible for the work group to have access to the files, stored on a central server, at any time and any place through the Internet (Simon 2016, pp. 813–814).

We use three different sources for our dictionary: the historical corpus, 6 million dictionary cards (mostly from the 19th century, we have managed to digitize 1.5 million slips until today) and a huge CD-ROM collection with 360 million words.

One of the main sources is the Hungarian Historical Corpus (http://clara.nytud.hu/mtsz/run.cgi/first_form). In 2015 we completed the expansion of this corpus, its size was increased by more than 10 percent, it was completed with texts of nearly 3 million words, from the fields of literature, journalism, sciences, personal and official language. So now it is a representative corpus of about 30 million words covering texts from 1772 (traditionally it is the beginning of the Enlightenment in Hungary) to 2010.

At the moment we use the Folio View query system. The 4.2 version we use was released in 1998, and generally it works very well, but it has some weak points: for example, you cannot edit the nfo file formats, it has some problems with finding very short words (like the '*a*' Hungarian indefinite article), it cannot make any distinction between capital and small letters, and it does not allow one to find punctuation-marks. Now we are testing another query system: the latest version of the historical corpus was made by the NoSketchEngine, a free corpus management system, and this tool seems very efficient, as one can find everything that is written in CQL, frequency lists can be made and so on.

The list of the dictionary slips' headwords is also available online (http://nszt.nytud.hu/cszj.html). This list is the richest selection of Hungarian vocabulary between 1750 and 1960, it contains more than 600,000 headwords.

The *Nszt.* explains the headwords in three different entry types. A simple entry consists of defined meanings with examples, cf. Fig. 1.

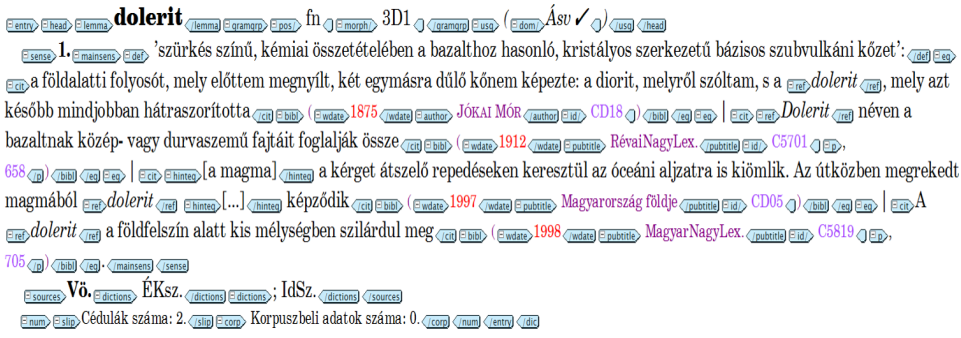

Fig. 1. Example of a simple entry

Complex entries consist of subentries of compound words that share common stems. In the previous Hungarian dictionaries one cannot find these types of entries. There are two types of such entries, depending on the situation of the stem in the compound *(diák-* 'student' and *-állomás* 'station'), cf. Fig. 2 and 3. *Nszt.* uses the complex entries for explaining co-verbs and in the case of certain words with common derivational or inflectional suffix-like posterior constituents.

Fig. 2. Example of a complex entry, type 1

Fig. 2. Example of a complex entry, type 2

The third type of entries are the reference entries, cf. Fig. 4. They contain cross-reference to other entries, and have two main types: reference entries of variations of other headwords and reference entries consisting only of an entry head with a headword and the part of speech indication, followed by references to compounds.

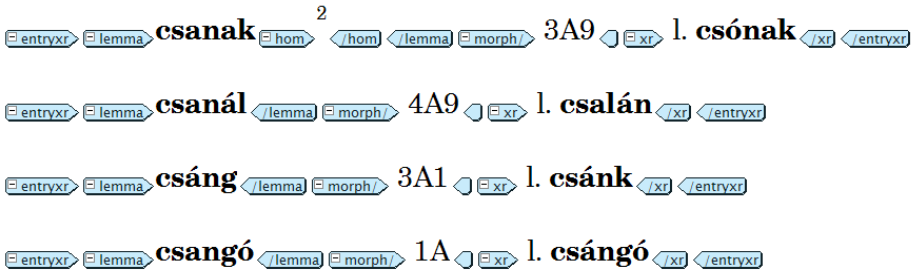

Fig. 4. Example of a reference entry

The *Nszt*. applies four methods in the lexicographical description: grammatical analysis, usage indications, semantic analysis, and examples (Ittzés 2012, pp. 36–40). In the analysis of the grammatical and lexical categories of headwords, the *Nszt.* draws on cutting-edge grammatical research. As a result, the part-of-speech structures of the *Nszt.*'s entries are notably different from those of earlier dictionaries. The *Nszt.* applies a new part-of-speech category called '*partikula*' (particle), similarly to the category of modifiers ('*módosítószó*', a subcategory of the main Hungarian category of pragmatic part of speech called '*mondatszó*'). Another novelty is that the *Nszt.* represents a code of one letter and one or two numbers which indicates the headword's paradigmatic type. These codes and paradigms are based on László Elekfi's *Dictionary of Hungarian Inflections* (Elekfi 1994). The usage indications are qualifying-categorising labels which indicate the position of the lexical elements in the lexicon and refer to certain lexicalized peculiarities of their usage. The *Nszt.* applies these types of labels: usage frequency, temporal or spatial aspects of the usage, usage in specific fields or special types of texts or styles, and the stylistic value and emotional elements related to semantic, morphological and phonetic peculiarities. The largest unit in the entries is the section of definitions and examples. Since most of the lexemes are polysemous, this unit is divided into shorter sections: the data ordered into units of meanings grouped in larger blocks of parts of speech. The order and the structure of meanings depend in most cases on their semantic relations, or sometimes on the chronology of their appearances. It is important that we give a description of a meaning of

a word only if it can be detected in any text of the corpus. The example sentences illustrate the meaning(s) and the usage of the given headword, and (with the date of their sources indicated) they give an idea of the historical development of the meanings. The first example always has to be the first occurrence of the word in the given meaning within the processed period. Examples appear with the indication of bibliographic data: the date of writing, the name of the author and an attributed numeric code which refers to the item of the corpus' list of sources.

The *Nszt.* deals with phraseological units in a different way than in earlier dictionaries: they appear in the entries of their heads, in the related meaning. Three types of phraseological units are distinguished: nouns with adjectival modifiers, noun-verb phrases, and phrases consisting of nominals in adverbial relation.

The *Nszt.* uses a consistent, formalized and controlled system of references that makes the use of the dictionary easier (Ittzés 2012, p. 41). References occur in most cases at the end of the entry, but sometimes (when *Nszt.* illustrates a certain meaning or part of speech of the headword in a different entry) also within the body of the entry. The corpora, the bibliographic data are stored electronically, and the text of the dictionary itself is also created in database format, making it possible to update the dictionary regularly.

References:

Elekfi, L. (1994). *Magyar ragozási szótár. Dictionary of Hungarian Inflections*, Budapest: MTA Nyelvtudományi Intézet.

Ittzés, N. (2012). *The Comprehensive Dictionary of Hungarian*. In Fábián, Zs. (Ed.), Hungarian lexicography II. Monolingual and special dictionaries. Lexikográfiai füzetek 6. (31–43), Budapest: Akadémiai Kiadó.

Simon, L. (2016). A digitális korszak vívmányainak hasznosulása a lexikográfiában: A nagyszótári projekt informatikai fejlesztéseiről. *Magyar Tudomány*, 7, 809–815.