Scientific review paper Received: 4. VI. 2021. Accepted: 27. VIII. 2021.

UDK 030(439)]:004.7 811.511.141'374

https://doi.org/10.33604/sl.15.29.4

Towards a new monolingual Hungarian explanatory dictionary: overview of the hungarian explanatory dictionaries

Veronika Lipp
Hungarian Research Centre for Linguistics, Institute for Lexicology, Budapest
lipp.veronika@nytud.mta.hu
László Simon
Hungarian Research Centre for Linguistics, Institute for Lexicology, Budapest
simon.laszlo@nytud.hu

ABSTRACT: The Lexical Knowledge Representation Research Group at the Department of Lexicology is one of the youngest research groups of the Hungarian Research Centre for Linguistics, founded in February 2020. The group is currently working on a new version of a monolingual explanatory dictionary partly based on *The Explanatory Dictionary of the Hungarian Language*. The aim is to compile an up-to-date online dictionary of contemporary Hungarian (2001–2020) by corpus-driven methods.

The present article describes *The Explanatory Dictionary of the Hungarian Language* and the *Comprehensive Dictionary of Hungarian* by presenting their history, the circumstances of their compilation, and the basic editorial guidelines. Then it outlines how the corpus for the planned dictionary is to be set up and how this corpus is to be analysed.

Keywords: lexicography; monolingual; dictionary; corpus; Hungarian

1. Introduction

In lexicography, three paradigms exist with respect to the construction of dictionaries: (i) the traditional; (ii) the corpus-based; and (iii) the corpus-driven approach (Atkins–Rundell 2008; Svensén 2009).

The appearance of electronically available linguistic corpora enabled lexicographers to analyse enormous quantities of linguistic data, which gave rise to a new paradigm in lexicography by the beginning of the 21st century. In this approach, meanings and semantic nodes are determined and mapped based on word use, relying on contexts in which the given word typically and frequently appears (Hanks

2010). Thus, the role of the frequency of words, i.e. their token frequency, became much more important for corpus-based approaches than for traditional ones.

Corpus-driven approaches take another step towards minimising the disturbing effect of the lexicographer's linguistic intuition. As the name of the new paradigm that partly relies on machine learning methods reveals, in this approach the corpus is not simply a collection of examples, but a linguistic source, the frequency list that determines the set of headwords in the dictionary to be compiled. In the ideal case, the exploration and the determination of the semantic sets of words is carried out fully automatically, without human intervention in corpus-driven approaches. One of the advantages of this method is that it provides manageable data series even for extremely large corpora.

Electronic lexicography has fundamentally changed our concepts of dictionaries, and has thus redefined the concept of lexicography itself. More and more often, e-lexicography appears as the topic of conferences and publications, such as the biannual eLex conferences or the publications of Granger and Paquot (2012), Fuertes-Olivera and Bergenholtz (2011) or Tarp (2008). No general modern dictionary can avoid relying on a morphologically annotated corpus for the description of the semantic and morphological properties of the lexicon (Svensén 2009: 45). Our knowledge of language operation had to be reassessed after analysing large corpora with sophisticated query tools, and this had consequences concerning dictionaries and the work of lexicographers (Rundell 2009). Today, the tools of lexicographers include IT solutions such as the GDEX (Good Dictionary Examples) tool in the Sketch Engine corpus query system, which evaluates and ranks relevant search results in a corpus with respect to their suitability to serve as dictionary examples (Kilgarriff et al. 2008); or Word Sketch, which does not only enlist collocations for a given word, but also groups them according to grammatical relationships (Thomas 2015).

2. Current situation

The Hungarian monolingual explanatory dictionaries were all compiled at the Research Institute for Linguistics (today: Hungarian Research Centre for Linguistics), Hungarian Academy of Sciences.

The Hungarian Research Centre for Linguistics has participated in international lexicographic projects for more than a decade; moreover, the centre has been a leading partner in the ELEXIS project. However, owing to the lack of online monolingual explanatory dictionaries, it is impossible for us to take part in serious lexicographic research. When word sense alignment (WSA) and word-sense disambiguation (WSD) tasks emerged – and required the comparison of the dictionary entries and meaning structures of at least two Hungarian explanatory dictionaries – a pro-

blem arose, as there are only two dictionaries available: the one-volume *Magyar értelmező kéziszótár* [Concise Hungarian Explanatory Dictionary] (2003) and the *A magyar nyelv értelmező szótára* [Explanatory Dictionary of the Hungarian Language, EDHL], which was published in seven volumes in the 1950s. We could only take part in international collaboration (Ahmadi et al. 2020) by choosing dictionary entries from the first volume of the larger dictionary (beginning with letters a, b, c, d and e, and comparing those to the entries of the *A magyar nyelv nagyszótára* [Comprehensive Dictionary of Hungarian, CDH] (see Section 3 below for more details), the compilation of which started at the beginning of the 2000s, and is still only approximately 20% ready.

3. Explanatory Dictionary of the Hungarian Language

The seven volumes of the EDHL were published between 1959 and 1962. According to the leading editor, László Országh (1962: 6) this dictionary is more detailed than one or two volume concise or pocket dictionaries, though evidently not as elaborate as comprehensive, full-size dictionaries, which might have more than 20 volumes.

Based on policies and criteria described in articles published before, during and after the completion of the dictionary, the authors – conforming to the contemporary expectations – followed the traditional approach. László Országh and his colleagues were highly experienced in working with bilingual dictionaries. The editors confirmed that they often relied on their own mental lexicons during the work. As the editor-in-chief wrote: »Our own language knowledge and intuition served as natural bases for determining the meaning, the use, and the style of a given word, which we constantly checked by conducting mini-surveys« (Országh 1953: 392, own translation).

The main source of the seven-volume explanatory dictionary, however, was an enormous collection of cards, the first ones of which were written at the end of the 19th century. Although it was a huge task to organise the cards, by the 1950s, the number of cards exceeded 4 million (Országh 1953: 393). The collection is primarily made up of hand-written cards of A6 size, which came to existence as a result of a call for »community collection« (actually, a form of crowdsourcing). This movement was launched by calls published in the journal *Magyar Nyelvőr* in the 1890s (Simonyi 1891: 59; Simonyi 1898: 240), and the *Utasítás* [Directive] given out by the Hungarian Academy of Sciences (MTA Szótári Bizottsága 1899), and went on for almost five decades.

The linguists working on the dictionary continued the extension of the card database in the 1950s: they collected the headwords and collocations from contemporary dictionaries, and also took words from daily and weekly papers, a selection of

fiction, and political and ideological publications. Meanwhile, another project was running with the aim of recording all occurrences of a given headword. Thus, all the words in several poems of the greatest 20th century poets, Endre Ady and Attila József, and also the full text of the Hungarian Constitution, were documented (Országh 1962: 119).

Although such a collection of headword cards can be regarded as a kind of text corpus in present-day terminology, the seven-volume explanatory dictionary is not at all a corpus-based dictionary. The primary function of the cards was to offer a collection of sample sentences out of which the lexicographer could choose the best one to illustrate the given meaning of the headword in question. The *Guidelines* published in the first volume of the dictionary describe what the editors thought about the role of illustration and the citation of examples for use. For example, it stated that, for illustrative purposes, both examples and citations can be applied. A »free example« is a text created for illustrative purposes, while »citations from writers or poets« were taken from the cards themselves (EDHL 1959–1962: XXVI, own translation).

Lexicographers chose 76 authors, and citations were mainly taken from their texts. These authors belong to the most outstanding authors of Hungarian literary history. The overwhelming majority of data was taken from the 19th century classics (EDHL 1959–1962: XXIX). While more than 30 percent of the 115,000 citations in the seven volumes were taken from three authors (whose oeuvre was mainly created in the 19th century), only one or two dozen citations were taken from several of the 76 authors. Only 22 authors were active in the 20th century, all of whose oeuvre could be regarded as closed by the beginning of the data collection process. Only 44,000 (less than 40%) citations belong to these 22 authors.

Citations not present on the cards could be entered into the dictionary only exceptionally, if the lexicographers could find absolutely no suitable text in the collection of cards to illustrate the given meaning. One of the fundamental principles for this was that translations from foreign texts could only be used if both the original author and the translator were outstanding literary authors (EDHL 1959–1962: XXIX).

The planned number of headwords for the EDHL was 45,000 in 1949 (Országh 1953: 394), while the final number of individual entries proved to be 58,000 in the seven volumes. According to the statistics published in the final volume, there are additionally 972 cross-reference entries; furthermore, 550 traditional similes, 1,005 idiomatic sayings and 319 proverbs were included in the relevant entries. The number of words in the lists of »related items« given without meanings at the end of entries is altogether 74,934; the lists of compounds given without meanings in the lexical entry of the first element of the compound amount to 37,625 words, while 21,280 idiomatic phrases were entered with their meanings.

As the originally printed dictionary is now available in XML format (cf. Section 4), it can be stated that the number of »free examples« created by lexicographers exceeds 200,000, i.e. is almost double the number of citations.

4. Comprehensive Dictionary of Hungarian (CDH)[1]

Based on the history of dictionaries in Hungary, it seems that, ever since the beginning of the 19th century, i.e. the Age of Reforms, everyone has aspired to compile this dictionary, i.e. the Comprehensive/Great Hungarian Dictionary. Originally, the seven-volume EDHL was aimed to fulfil this purpose. However, owing to time restrictions and capacities, the authors finally decided to create a more concise dictionary instead of waiting endlessly for the publication of *the* comprehensive dictionary.

In 1821, József Teleki's essay entitled *Egy tökéletes magyar szótár elrendeltetése, készítése módja* [The Aim and Methods for Compiling a Perfect Hungarian Dictionary] was published. Based on this document, the Hungarian Scholarly Association published their concept of a comprehensive dictionary in 1934. Ten years later, in 1944, Gergely Czuczor and János Fogarasi started to write *A magyar nyelv szótára* [Dictionary of the Hungarian Language]. Their aim was to create a shorter, less detailed dictionary focusing on the vocabulary of the first half of the 19th century. The manuscript was completed in 1861, and the six volumes were finally published between 1862 and 1864 (Gerstner 2006: 10).

The compilation of a comprehensive dictionary was a hot topic for decades even after the publication of the last volume of the Dictionary of the Hungarian Language. The preparations started, and a dialect dictionary and an etymological dictionary were thus published; cards were collected in several waves, and a list of headwords was also completed (R. Hutás 1973: 453). In 1953, while the work was already in progress for the EDHL (see Section 2), the Hungarian Academy of Sciences (MTA) declared that the idea of the comprehensive dictionary had not been discarded, and it must also be completed. The relevant committee suggested that the title of this dictionary should be »A magyar irodalmi nyelv nagyszótára a felújulás korától napjainkig« [Comprehensive Dictionary of Literary Hungarian from the Enlightenment to the Present] (MTA Nyelvtudományi Bizottság 1953: 257). The Linguistic Committee also declared that the comprehensive dictionary should heavily rely on literary texts regarded as significant from a literary historical perspective, on excerpts from periodicals (daily papers, journals) and also on academic texts. It was also stated that the collection method of materials should not be revised or modernised (MTA Nyelvtudományi Bizottság 1953: 259).

[1] Both authors of the present article worked as lexicographers on the first eight volumes of this dictionary. Thus, they had personal experiences of the workflow.

Lexicographers regarded it as crucial that the writing process should not take a very long time, that work should be sped up. They estimated that it would be enough to create 1.5 to 2 million cards until 1960 for the presentation of the 20th-century literary and colloquial language. Moreover, the realistic amount of individual entries was regarded to be between 300,000 and 400,000. According to their preliminary calculations, the compilation and publication of the dictionary should have taken ten years, and the last volume should have been published by the beginning of the 1970s (MTA Nyelvtudományi Bizottság 1953: 261).

By the 1980s, the card collection, which was also used by the lexicographers of the Historical-Etymological Dictionary of Hungarian, was transformed into an archive. It seemed to be evident that, at the end of the 20th dictionary, when computational lexicography and corpus-based approaches started to spread, it is impossible to create an up-to-date dictionary based on a collection of text fragments on cards. In 1984, the MTA Presidency passed a resolution on the compilation of an electronic text database and the launch of a project for the computer-based creation of the *A magyar irodalmi és köznyelv nagyszótára (1533–1990)* [Comprehensive Dictionary of Literary and Colloquial Hungarian (1533–1990)]. At that time, the aim was to digitise texts of 13 million words. While the process started in 1986, only a 3-million-word database was ready by 1989 (Kiss–Pajzs 1989).

The first version of the database, which contained texts published between 1772 and 1990, was published at the end of the 20th century (Pajzs et al. 1998). Finally, with further additions, the *Magyar történeti szövegtár 1772–2000* [Hungarian Historical Corpus 1772–2000] was completed by the beginning of the 2000s. In 2014, further 3 million words were added to the corpus, i.e. the time span of the database was extended (Simon 2016: 810). Currently, the *Hungarian Historical Corpus 1772–2010* contains nearly 30 million words from texts spanning almost 240 years.

According to the original ideas, the Hungarian comprehensive dictionary should have primarily relied on this corpus, and this primary source should have been amended with data from the card archive. However, during the pilot project, in which test entries were written, it became evident that the lack of data is a systematic problem. It was impossible to illustrate even the clearly existing meanings present in earlier explanatory dictionaries. As a result, the decision was made to enlarge the corpus of the dictionary by one magnitude, through the addition of 300 million words. The overwhelming majority of these texts were taken from the CD--ROMS and DVDs published by the Arcanum Adatbázis Kft. (see in detail below). These contained the oeuvres of renowned Hungarian writers and poets, various lexicons, Bible translations, full series of journals, etc. (Lipp 2018).

The CDH is to follow a corpus-based approach, as declared by Ittzés (2011: 32, own translation): »...it is not based on mental lexicons or dictionaries published ear-

lier, but is created through the analysis of texts in the corpus used as contexts«. The CDH was to be a printed dictionary, and the first two volumes were published by the Research Institute for Linguistics of the Hungarian Academy of Sciences, i.e. the publisher of the seven-volume CDHL. The seventh volume of the CDH was published in 2018, while the eighth volume is in print at the time of writing this article. The publisher of this latest volume is already the Hungarian Research Centre for Linguistics.

Twenty years ago, the CDH was planned to contain 110,000 dictionary entries. The second volume contains the words beginning with the letter *a*, with more than 5,500 headwords. The next six volumes contain altogether 18,500 entries, and the 8th volume ends with the last entry beginning with the letter *e*. Almost 10 percent of entries are cross-references, so only approximately 20 percent of the planned number of entries have been completed.

The authors of the CDH undertook the task of finding the earliest data for each meaning available in the corpus. Furthermore, owing to the work of the Arcanum Adatbázis Kft., the idea of the »closed corpus« was given up.

The Arcanum Kft. started digitising texts in the 1990s, and they published these on CD-ROMs. However, from the 2000s, their main profile was the digital archiving of printed materials found in public collections, various printed documents, materials found in archives, daily papers and journals using high performance scanners. In the 2010s, they were inspired by the Google Books project, which aims to digitise libraries, and in 2014 the Arcanum Digitheca project (ADT) was launched. Since then, hundreds of thousands, or millions, of scanned printed pages have been made available on their homepage every year, the majority of which is from scientific or professional journals and daily and weekly papers from the 19th, 20th, and 21st centuries. In a similar manner to Google Books, Arcanum uses the two-layer pdf format: along with the photograph of a given page, the corresponding searchable text file is also available.

At the time of writing this article, the counter on the ADT homepage shows 32,945,755 digitised pages, which contain approximately 25–30 billion running words. Already in 2019, when the ADT archive contained only 23 million pages, the director of the company, Sándor Biszak, highlighted a problem: as they were able to digitise as many as 1.5 million pages in a month, the amount of unscanned documents would soon plummet. At that time, he realised that the rate at which the database was growing had slowed down gradually. He thought that, when the threshold of 35–40 million pages is reached, the pace would stagnate (Nagy 2019).

After the ADT homepage was launched, the authors and editors of the CDH decided to incorporate this database into that of the dictionary. As a result, however,

the lexicographers, who were used to manual data processing and could only rely on their own work, faced such a huge amount of data that was impossible to handle using traditional methods.

5. A novel approach: a corpus-driven explanatory dictionary

The Lexical Knowledge Representation Research Group at the Department of Lexicology was founded in February 2020 at the Hungarian Research Centre for Linguistics. The primary aim of the group is to open up new ways and look for new possibilities in lexicography, with significant language technology support.

5.1. Corpus building

According to our experience, a text corpus of appropriate size and quality, which provides a reliable picture of the synchronic state of the language, is an indispensable prerequisite for the majority of modern linguistic research tasks, especially if Hungarian lexicography wishes to join the present European lexicological projects. Concerning their size, the Hungarian National Corpus (HNC) and the Hungarian Gigaword Corpus (HGC) are comparable to the representative corpora for English, German, Dutch, or Slovenian (Klosa 2011; Meyer 2014; Schoonheim–Tempelaars 2010; Holdt et al. 2019); however, they are not suitable for lexicographic research for several reasons.

The criterion system for reference corpora is well-established (Atkins et al. 1992; Biber 1993; Svensén 2009). At present, text databases exist only in electronic formats. It is interesting to see that the first electronic corpus, the Brown Corpus, which was made up of documents published in the USA in 1961 and served as a model for further corpora, dates back to the time when the printed volumes of the EDHL were published (Francis–Kučera 1967).

As the first step of our research, we aim to compile a 1-billion-word corpus, which represents the synchronic state of the Hungarian language well, and is suitable for serving as the source database for an explanatory dictionary. The corpus must meet the following criteria:

1. It should contain texts only from the period between 2001 and 2020.

2. It should be representative, i.e. it should proportionally contain texts typical of the era concerning their register and type, and show the characteristics of both printed documents and texts published online.

3. Concerning metadata, each document included should have a publication date and a source ID as well.

4. The corpus should be annotated, that is, (i) sentence boundaries should be determined; (ii) the text should be tokenised and lemmatised; (iii) the tokens should be morphologically disambiguated. This will be achieved by the state-of-the-art Hungarian NLP system, the *e-magyar* (Indig et al. 2020).

5. The corpus should be parsed, so that the clausal orders of preverbs and verbs can be explored satisfactorily, and that it is possible to identify preverbs separated from their verbs, so that all instances of a given preverb-verb construction can be located (Kalivoda 2018, 2021).

The first step towards the creation of such a corpus is to look at the available Hungarian corpora, and decide which parts of them can be used for our purposes. The new, extended version of the HNC, the Hungarian Gigaword Corpus itself, contains 1 billion words and was designed in a way to serve all sorts of linguistic research purposes (Oravecz et al. 2014). It seems only parts of this corpus will be suitable for our purposes. The same is true for the Hungarian Web Corpus (huTenTen), which is made up of almost 6.5 million documents downloaded from the Internet in 2012, containing nearly 2.5 billion words; or the Webcorpus 2.0, made up of 9 billion words collected by a web crawler (Nemeskey 2020).

In Section 3, the ADT database of the Arcanum Kft. has already been mentioned, a part of which we also wish to use in the new corpus. This database contains documents created by optical character recognition (OCR) software, with the detection, filtering, and correction of specific mistakes, which is a huge task for our computational linguist colleagues. At present, a 10-billion-word database is available from this source, which is currently »cleaned«. Typical OCR-related mistakes include broken words, indecipherable character sequences, and mistakenly identified words, which are correct Hungarian words but do not make any sense in the given context. Our research project also aims to explore which methods of language technology can be used to correct these texts to reach a normative level, if it is possible at all.

5.2. Online lexical database

In our approach, the new dictionary would be nothing more than a lexical database, available only in online format, with dynamic solutions and the possibility of comprehensive changes affecting the whole database. Only those meanings that are validated by the corpus data, or that have been revealed by the analysis of the corpus, will be published in this database.

As the meaning structures of the EDHL will be used as guidelines, the first step was to create an XML version of this dictionary, which meets the requirements

of present-day lexicography and computational linguistics. Consequently, the whole dictionary can be used as a text database, and enables us to update some of the solutions used in printed dictionaries. Furthermore, the meaning structures, which even after decades seem to be well-established both from a grammatical and a semantic perspective, can be used as starting points for collecting new examples from the corpus.

Above all, however, frequency will be the main principle for both the selection of headwords and the order of meanings. The analysis of the whole corpus will be carried out in a similar manner to the methodology followed for the compilation of the *Igei szerkezetek gyakoriségi szótára* [Frequency Dictionary of Verb Phrase Constructions] (Sass 2011).

5.3. Corpus management

The Sketch Engine software (cf. Section 1), which makes sophisticated searches possible, will be used to manage the new corpus. Programs developed for the purposes of lexicographical research can successfully explore the profile of a given lemma even if used on huge corpora. Based on our preliminary studies, in addition to the Word Sketch tool, the Concordance Search and Word Lists tools that generate frequency lists can be of the best use for the present project. The Word Sketch tool provides a quick overview of a word's lexical profile. It looks for collocations containing the given word, and presents these collocations grouped by grammatical relations. For example, if we are looking for the noun *róka* 'fox', we immediately see that its typical modifiers in Hungarian are *vén* 'old', *rutinos* 'experienced' and *ravasz* 'sly'; it is typically a subject of verbs like *elszaporodik* 'breed rapidly', *vadászik* 'hunt' and ólálkodik 'lurk'; its most typical possessor is *A kis herceg* 'The Little Prince'. After that, all collocations can be further studied in a concordance view that shows them in their original context.

We have also tested the Word Sketch Difference tool, which can help identify synonyms that have largely overlapping meaning structures by the analysis of collocations. For example, the word *vörös* 'red' co-occurs typically with nouns like *hadsereg* 'army' and *csillag* 'star', while the word *piros* (also meaning 'red') is the typical attribute of nouns like *lámpa* 'light' and *lap* 'card'. It is important to add these collocations to the relevant dictionary entries, because in these cases the two adjectives (which have very similar meanings) are not interchangeable.

Hopefully, this research proves that the above tools and the data series obtained by them from the corpus can give an objective picture of the lexical profile of a given headword, and that these pieces of information can hasten the creation of a new explanatory dictionary. Additionally, this research also aims to prove that, in

the third decade of the 21st century, the corpus-driven paradigm should be followed by lexicographers.

REFERENCES

Ahmadi, S. et al. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 3232–3242*. https://www.researchgate.net/publication/341567425_A_Multilingual_Evaluation_Dataset_for_Monolingual_Word_Sense_Alignment

Atkins, B. T. S., **Rundell,** M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

Atkins, B. T. S., **Clear,** J., **Ostler,** N. (1992). *Corpus design criteria*. Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing, 7 (1), 1–16.

Biber, D. (1993). *Representativeness in corpus design. Literary and Linguistic Computing*. Journal of the Association for Literary and Linguistic Computing, 8 (4), 243–257.

Francis, W. N., **Kučera,** H. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Fuertes-Olivera, P., **Bergenholtz,** H. (eds.) (2011). *E-lexicography. The Internet, Digital Initiatives and Lexicography*. Continuum, London.

Gerstner, K. (2006). A magyar nyelv nagyszótárának áttekintő története. [A review of the history of the Comprehensive Dictionary of Hungarian] In: Ittzés, N. et al. (eds.): *A magyar nyelv nagyszótára I.* [Comprehensive Dictionary of Hungarian] Budapest, Nyelvtudományi Intézet, 10–17.

Granger, S., **Paquot,** M. (eds.) (2012). *Electronic Lexicography*. Oxford University Press, Oxford.

Hanks, P. (2010). *Compiling a monolingual dictionary for native speakers*. Lexicos, 20, 580–598.

Holdt, A. Š., **Dobrovoljc,** K., **Logar,** N. (2019). *Simplicity matters: user evaluation of the Slovene reference corpus*. Lang Resources & Evaluation 53, 173–190.

Hutás Magdolna, R. (1973). *Az Akadémiai Nagyszótár történetének vázlata (1898–1952)*. [Outline of the history of the Academic Comprehensive Dictionary] Nyelvtudományi Közlemények, 75, 447–465.

Indig, B., **Sass,** B., **Mittelholcz,** I. (2020). The xtsv Framework and the Twelve Virtues of Pipelines. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 7044–7052.

Kalivoda, Á. (2018). Hungarian particle verbs in a corpus-driven approach. In: Gelbukh, A. (ed.): *Computational Linguistics and Intelligent Text Processing: 18th International Conference (CICLing 2017)*. Revised Selected Papers, Part I. Springe, Cham, 123–133.

Kalivoda, Á. (2021). *Igekötős szerkezetek a magyarban*. [Preverb–verb constructions in Hungarian.] PhD dissertation, Pázmány Péter Catholic University, Budapest.

Kilgarriff, A. et al. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In: Bernal, E. & DeCesaris, J. (eds.): *Proceedings of the 13th EURALEX International Congress*. Institut Universitari de Lingüística Aplicada, Barcelona, 425–432.

Kiss, L., **Pajzs,** J. (1989). *A magyar irodalmi és köznyelv nagyszótára (1533–1990)*. [Comprehensive Dictionary of the Hungarian Literary and Colloquial Language] Magyar Nyelv, 85 (2), 129–136.

Klosa, A. (ed.) (2011). elexiko. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (Studien zur deutschen Sprache 55). Narr, Tübingen.

Lipp, V. (2018). Comprehensive Dictionary of Hungarian. In: Bańko, M., Karaś, H. (eds.): *Między teoria a praktyka: Metody współczesnej leksykografii.* Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 145–149.

Meyer, P. (2014). Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von elexiko. In: Abel, A. Lemnitzer & L. (eds.): *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern.* Institut für Deutsche Sprache, Mannheim, 9–21.

MTA Szótári Bizottsága (1899). *Utasítások az új Nagy Szótár adatgyűjtőinek.* [Guidelines for the data collectors of the new Comprehensive Dictionary] Magyar Tudományos Akadémia, Budapest.

MTA Nyelvtudományi Bizottság (1953). *A Nyelvtudományi Bizottság határozata az Akadémiai Nagyszótárról.* [Resolution of the Linguistic Committee about the Academic Comprehensive Dictionary] Magyar Nyelv, 49 (3–4), 257–261.

Nagy, A. K. (2019). *Budai családi házban teszik kereshetővé a múltunkat.* [Our past is made searchable in a Buda detached house] https://index.hu/techtud/2019/09/15/arcanum_digitalizalas_szkenneles_konyvtaerak_folyoiratok_ujsagok_hungaricana (accessed 23 May 2021)

Nemeskey, D. M. (2020). *Natural Language Processing Methods for Language Modeling.* PhD dissertation, Eötvös Loránd University, Budapest.

Oravecz, Cs., **Váradi,** T., **Sass,** B. (2014). The Hungarian Gigaword Corpus. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation,* 1719–1723

Országh, L. (1953). *A magyar nyelv új szótáráról.* [On the new dictionary of the Hungarian Language] Magyar Nyelvőr, 77 (5–6), 387–407.

Országh, L. (ed.) (1962). *A szótárírás elmélete és gyakorlata a Magyar nyelv értelmező szótárában.* [Theory and practice of dictionary writing for the Explanatory Dictionary of the Hungarian Language] (Nyelvtudományi Értekezések, 36) Akadémiai Kiadó, Budapest.

Pajzs, J. et al. (eds.) (1998). *Hungarian Historical Corpus.* Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest. (http://clara.nytud.hu/mtsz)

Rundell, M. (2009). The road to automated lexicography: First banish the drudgery... then the drudges? In: Granger, S. & Paquot, M. (eds.): *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009. Book of abstracts.* Presses universitaires de Louvain, Louvain-la-Neuve, Belgium, 9–10.

Sass, B. (2011). *Igei szerkezetek gyakorisági szótára – Egy automatikus lexikai kinyerő eljárás és alkalmazása.* [Frequency Dictionary of Verb Phrase Constructions – An automatic lexical acquisition method and its applications] PhD dissertation, Pázmány Péter Catholic University, Budapest.

Schoonheim, T., **Tempelaars,** R. (2010). Dutch Lexicography in Progress: The Algemeen Nederlands Woordenboek (ANW). In: Dykstry, A. & Schoonheim, T. (eds.): *Proceeding of the Fourteenth EURALEX International Congress. 6–10 July 2010.* Afûk, Leeuwarden, 718–725.

Simon, L. (2016). *A digitális korszak vívmányainak hasznosulása a lexikográfiában: a nagyszótári projekt informatikai fejlesztéséről.* [Using results if the digital era in lexicography: on the IT developments of the comprehensive dictionary project] Magyar Tudomány, 177 (7), 809–815.

Simonyi, Zs. (1891). *A nyelvujítási vitához.* [To the debate on the language reform] Magyar Nyelvőr, 20, 56–60.

Simonyi, Zs. (1898). *Fölhivás az új Nagy Szótár munkálataiban való részvételre.* [Call for participation in the work on the new Comprehensive Dictionary] Magyar Nyelvőr, 27 (5), 240.

Svensén, B. (2009). *A Handbook of Lexicography.* The Theory and Practice of Dictionary-making. Cambridge University Press, Cambridge

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-knowledge.* Max Niemeyer, Tübingen.

Teleki, J. (1821). Egy tökéletes magyar szótár elrendeltetése, készítése módja. [The aim and methods for compiling a perfect Hungarian dictionary] In: Horvát, I. (ed.): *Jutalom-feleletek a magyar nyelvről.* Pest, 1–72.

Thomas, J. E. (2015). Word Sketches. In: Thomas, J. E.: *Discovering English with Sketch Engine.* Versatile, Brno, 161–176.

DICTIONARIES AND CORPORA

EDHL = Bárczi, G. & Országh, L. (eds.) (1959–1962). *A magyar nyelv értelmező szótára I–VII.* [The Explanatory Dictionary of the Hungarian Language] Akadémiai Kiadó, Budapest.

CDH = Ittzés, N. et al. (2006–2021). *A magyar nyelv nagyszótára I–VIII.* [Comprehensive Dictionary of Hungarian] Nyelvtudományi Kutatóközpont, Budapest.

Benkő, L. (ed.) (1967–1984). *Történeti-etimológiai szótár I–IV.* [Historical-Etymological Dictionary of the Hungarian Language] Akadémiai Kiadó, Budapest.

Czuczor, G., **Fogarasi,** J. (1862–1874). *A magyar nyelv szótára.* [Dictionary of the Hungarian Language]. Emich, Budapest.

Gombocz, Z., **Melich,** J. (1914–1944). *Magyar etymológiai szótár.* [Etymological Dictionary of the Hungarian Language] MTA, Budapest.

Lőrinczy, É. B., **Hosszú,** F. (eds.) (1979–2010). Új magyar tájszótár. [New Dialect Dictionary of the Hungarian Language] Akadémiai Kiadó, Budapest.

Pusztai, F. (ed.) (2003). *Magyar értelmező kéziszótár.* [Concise Hungarian Explanatory Dictionary] Akadémiai Kiadó, Budapest.

Sass, B. et al. 2010. *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára.* [Hungarian Verb Phrase Constructions – Dictionary of the most frequent complements and multiword expressions] Tinta Könyvkiadó, Budapest.

Szinnyei, J. (1893–1901). *Magyar tájszótár.* [Dialect Dictionary of the Hungarian Language] Hornyánszky, Budapest.

Arcanum Digitális Tudománytár [Arcanum Digitheca]: https://adt.arcanum.com/hu/

Magyar nemzeti szövegtár [Hungarian National Corpus, Hungarian Gigaword Corpus]: http://corpus.nytud.hu/mnsz/index_eng.html

Magyar történeti szövegtár [Hungarian Historical Corpus]: http://clara.nytud.hu/mtsz/run.cgi/first_form

PREMA NOVOM JEDNOJEZIČNOM MAĐARSKOM OBJASNIDBENOM RJEČNIKU: PREGLED MAĐARSKIH OBJASNIDBENIH RJEČNIKA

Lipp Veronika
Mađarski istraživački centar za lingvistiku, Institut za leksikologiju, Budimpešta
lipp.veronika@nytud.mta.hu
Simon László
Mađarski istraživački centar za lingvistiku, Institut za leksikologiju, Budimpešta
simon.laszlo@nytud.hu

SAŽETAK: Istraživačka skupina za prikaz leksičkog znanja jedna je od najmlađih istraživačkih skupina Mađarskog istraživačkog centra za lingvistiku, osnovana u veljači 2020. Skupina trenutno radi na novoj inačici jednojezičnoga objasnidbenog rječnika proizišloga iz *Objasnidbenoga rječnika mađarskog jezika*. Cilj joj je kompilirati moderan i ažuriran mrežni rječnik mađarskog jezika (2001–2020) koristeći se korpusom vođenim metodama. Članak opisuje *Objasnidbeni rječnik mađarskog jezika* i *Velikog rječnika mađarskog jezika* predstavljanjem njihove povijesti, okolnosti u kojima su kompilirani, te osnovnih uredničkih načela. Potom skicira kako će se organizirati i analizirati korpus planiranoga rječnika.

Ključne riječi: leksikografija; jednojezični; rječnik; korpus; mađarski