Rethinking the Role of Digital Author's Dictionaries in Humanities Research

Margit Kiss[1], Tamás Mészáros[2]

[1]Institute for Literary Studies, Hungarian Academy of Sciences, [2]Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics
E-mail: kiss.margit@btk.mta.hu, meszaros@mit.bme.hu

Abstract

Although it is true that computers make one's work significantly easier during some of the work phases of creating an author's dictionary, our goal is to completely revise conventional computer-based dictionary work processes as well as to extend the possible range of applications for author's dictionaries. In our study we examine the role of the digital author's dictionaries in humanities research. We developed the extended digital author's dictionary based on the oeuvre of Kelemen Mikes (1690-1761). We present the advantages of the extended digital author's dictionary and demonstrate its benefits in literary and linguistic research. The main advantages of our system include easily accessible up-to-date encyclopedic information and the improved efficiency of historical text analysis methods. The benefits of the extended functions of the digital author's dictionary help scholars answer more specialized and complex research questions, and reconsider expectations towards the new generation of author's dictionaries.

Keywords: author's digital dictionary, historical text analysis, Kelemen Mikes

1   Introduction

The number of author's dictionaries available today is quite significant, and is continually increasing; still, the literature on lexicography pays little attention to this topic (Karpova 2011; Mattausch 1990). Paper-based author's dictionaries play a limited role in humanities research, their function is essentially limited to contextual interpretation and presenting the linguistic functions of the given word in the text, which makes it difficult to make it a source of comprehensive, complex analyses concerning a larger lexicon. In contrast, the digital environment is ideal for the author's dictionary, since structured data handling and data extraction have a number of unexploited advantages compared to paper-based volumes, and this strengthens and extends the function which the author's dictionary was originally designed to fulfill. We have two aims: on the one hand, we would like to introduce the new generation of digital author's dictionaries, for which we have created the extended digital dictionary through making use of state-of-the-art opportunities in information technology; on the other hand, we would like to sketch the specific application areas to demonstrate the methodological changes and new possibilities this type of dictionary creates in humanities research. In our study, through presenting a digital author's dictionary under preparation based on the oeuvre of an 18th-century Hungarian writer, we describe how we have rethought the role of the digital author's dictionary.

2   The definition and development of the author's dictionary

The main aim of the author's dictionary is to show elements of language use, style, vocabulary and expressions through presenting a significant part of an author's oeuvre, or even its complete vocabulary.

Author's dictionary: a type of reference work which provides information on the vocabulary of a specific author. The material is usually based on a text corpus of one, several or all of the works of the author, and often presented in alphabetical order, with examples or contexts (but not definitions) of the words cited (Hartmann & James 2001: 10).

Many author's dictionaries have been created throughout the world, such as for the works of Johann Wolfgang Goethe, Friedrich Schiller, Thomas Mann, Bertolt Brecht, Jean Racine, Victor Hugo, Dante, Henrik Ibsen, Alexander Pushkin, Mikhail Lermontov, William Faulkner, Mark Twain, Ernest Hemingway, and William Shakespeare. Not only the work of writers of fiction, but also other significant authors' work may constitute the basis of such dictionaries, e.g. the *Kant Dictionary*, the *Hegel Dictionary*, the *Historical Dictionary of Leibniz's Philosophy*, etc. (see Karpova 2011).

The precursors of the author's dictionary as defined today in terms of genre are concordances and glossaries (Karpova 2011: 1-7). While the aim of a concordance is orientation and search within a text (e.g. in the Bible), glossaries help the reader interpret specialist texts (e.g. the writing of ancient Greek and Roman authors). The next milestone in the development of the author's dictionary as a genre happened when the emphasis shifted from sporadic word analyses to complete vocabulary analyses, to processing the whole authorial lexicon (e.g. Clarke 1846). This change in approach marked a generic determination for the author's dictionary that is still valid today.

In 20th-century dictionary literature previous methods used in the processing of authorial oeuvres were replaced by indexing, preparing concordances, and interpretation with stylistic and grammatical qualifications: emphasis shifted from sporadic interpretation to methodical systematization. The most important task of the author's dictionary became the lexicographic processing of the entire oeuvre, by supplying context-dependent interpretation. The function aiding textual understanding became a scientific genre of its own. The complete author's dictionary is also a significant qualitative change, a basic generic characteristic. These functional dictionaries present the words in their real context, based on principles of general grammar, semantics, and stylistics. Being exhaustive implies processing every occurrence of every word. Due to space limitations and because it requires substantial manpower, in the case of paper-based dictionaries this aim is often very difficult to achieve, so there is often a need for solutions that limit size or create a complex, complicated reference system. Influential examples of this interpretative-qualifying complete dictionary type are the Pushkin (Vinogradov 1956-1961), Goethe (GWB 1978-), Ibsen (Ibsen-Ordbok 1958), and Petőfi (J. Soltész et al. 1973-1987) dictionaries. These were typically created slowly, with a detailed entry structure, a rich data structure, providing the occurrence data, with a detailed, sophisticated meaning description, stylistic qualifications, and phraseological references, and all made with virtually no computer support.

With the improvement of technology and the spread of information technology (IT), another type of author's dictionary also appeared (today this has a well-established tradition), which, instead of semantic-stylistic finishing, represented a more formal, grammar- and IT-centered approach. Among the first computer applications in connection with authorial texts was the preparation of concordance lists (Busa 1980). At the dawn of computer-based lexicography (in the 1950s and 1960s), analyzing the works of Thomas Aquinas, Kant, Shakespeare, Goethe, Racine, Baudelaire, Dante, and others on the basis of corpora became the center of attention, in an index and concordance style (Gouws et al. 2013: 974). In the 1970s in Hungary the work of Ferenc Papp also strengthened this line of work through the computer-aided processing of Ady's oeuvre, and by emphasizing the importance of concordance as a raw material of author's dictionaries (Mártonfi 2014). Using the concordance list generated by the machine as a basis, we can create works resembling dictionaries with limited editorial work, which are a transition between concordance lists and author's dictionaries. These works, in contrast with previous ones, can be prepared with much less effort; at the same time, they also provide less in terms of their lexicological-semantic finesse. A concordance list that is the basis of dictionaries

created with computer assistance is often a product on its own and an interesting source of research, but as intermediate textual material between corpus and dictionary it also makes lexicographers' work significantly easier (e.g. OSS 2003-2018).

The IT solutions that are widespread today in the creation of author's dictionaries have made it substantially easier to create dictionaries and thus sped up the process (Čermák & Cvrček 2010). The spread of computers in lexicology work has caused changes in several areas (Gouws 2014). The easy and fast digitalizing of texts, using text recognition processes, creating and using text corpora and databases, producing automatic concordances, and using specialized software for editing entries, are all due to the application of IT, and are now of substantial help in the process of creating a dictionary, with such technologies replacing manual tools. Storing texts and handling data has become easier, and this change has not only made dictionary makers' task easier, but also changed the users' expectations towards dictionaries. These IT innovations have taken over some areas associated with paper-based dictionaries, and acquired new functions (Lew 2013).

IT tools, however, have even more to offer for author's dictionaries than what we are used to today, pointing beyond storing data and handling it in different ways. In addition to their role in the process of dictionary making, such tools can play a part in research and the different ways of processing data. They can integrate and represent the dispersed knowledge elements (linguistic, stylistic, historical, etc.) located in different places, and can make them processable and further applicable. While linguistic corpora have been the main tool in linguistic research for a long time, their capabilities for dictionaries remain underestimated (Apresjan & Mikulin 2016). By using the opportunities of a state-of-the-art set of tools provided by IT, we need to rethink how else we can use digital author's dictionaries beyond the usual applications, which also means a rethinking of their possible roles. In the future this change will result in creating author's dictionaries with greater added value.

3   The extended author's dictionary

Although it is true that computers make one's work significantly easier during some of the work phases of creating an author's dictionary, our goal is to completely revise conventional computer-based dictionary work processes as well as to extend the possible range of applications for such dictionaries. We set out to exploit modern information technology tools and prepare a type of digital author's dictionary which adjusts to the new opportunities and demands, and supports humanities research in a more efficient way than usual. Our main aims cover three major areas.

- In terms of a quantitative change, we are preparing a complete dictionary: we present every single word and sample sentence, without any space limitations.
- We provide much more information content for the individual entries than usual, which can be stored and searched in a structured manner.
- We link the individual entries with external data sources so that we can link the relevant knowledge elements digitally stored elsewhere to the particular entry.

We have implemented this new type of author's dictionary by processing the oeuvre of Kelemen Mikes, who played an influential role in 18th-century Hungarian prose literature. With its 1.5 million words of text, his work stands out among Hungarian authors' oeuvres, and his complete oeuvre is not part of the corpus of the comprehensive dictionary (HHC). The Mikes dictionary is based on the textual material of the complete critical edition, and this is the largest lexicon based on which a Hungarian author's dictionary has been created. The Mikes dictionary is the first complete Hungarian digital author's dictionary (Kiss 2012). It presents all the author's words with all the sample sentences. In a printed format the dictionary would comprise approx. 20,000 pages.

The use of paper-based dictionaries essentially concentrates on interpreting a given authorial text location, and search capabilities are basically limited to individual entries. The structure of the entries is rigid, and word interpretations often contain too much or too little explanation. Publishing in a printed volume makes it necessary to apply a range of limiting functions, which may include limiting the sample material, establishing a complicated reference system, or other space-saving strategies. In a traditional paper-based author's dictionary the typical entry structure of a geographical name is the following:

Figure 1: *Konstantinápoly* in a paper-based author's dictionary.

From this entry we can learn the following pieces of information concerning the given textual location: the headword of the place name, authorial form variations, number of occurrences, paradigmatic forms, a one-sentence interpretation concerning place location, and one or two selected sample sentences for illustration.

However, much more semantic content can fit into the computer-based representation of an author's dictionary than in a paper-based format. In order to provide the dictionary with as much semantic information content as possible, we extended the dictionary in two ways: on the one hand, we allowed the dictionary maker to include lexicographical knowledge about the entries; on the other hand, we enriched the dictionary by linking it to already existing external knowledge sources. We extended the entries of the Mikes dictionary in such a structured way that it contains extra knowledge suitable for computer representation, search, and processing. We completed this extension in the areas listed below. We linked all the sample sentences to each authorial word form, in some cases several thousands of them. From the individual sample sentences we can get to the wider textual context, since the sample sentences are linked with the corpus in the dictionary. We supplemented the word forms with the form variation of Mikes, as well as the contemporary headword. Beside the modern dictionary headwords, we also defined so-called reference headwords where necessary, so we can extend the searchability of the words by adding different headword variations and by presenting their connections to other headwords. Types of these include: reference to headwords within the dictionary based on etymology, e.g. *hívség - hűség*; reference to current Hungarian (or foreign-language) equivalents of words functioning as foreign words and loanwords, e.g. *decembris - december*; presenting the headword variations also existing in contemporary Hungarian together, e.g. *caritas - karitász*; presenting the proper name variations also existing in contemporary Hungarian together, e.g. *Kroiszosz - Krőzus*; presenting the Hungarian equivalents of Latin proper names, e.g. *Casimirus - Kazimír*; associating archaic words with their current form, e.g. *milliom - millió*; presenting connections with foreign words, e.g. *clinicus - klinikus*; presenting the Turkish originals of the Turkish words used by Mikes, e.g. *cház oda - has oda*. Proper names, words rooted in foreign languages, words invented by Mikes (the words which are not included among the headwords of other dictionaries or in the text corpus of the comprehensive Hungarian dictionary), and those headwords which differ from the form used by Mikes, have received further annotations for content. We have attached the part-of-speech category to every occurrence of a word. The dictionary also assigns semantic information to named entities. Geographical names are supplemented by further metadata, for example, geographical coordinates, which, besides interpretation, provide a clear point of reference for determining exact location.

We supplemented the extra knowledge entered by the dictionary maker with a second pillar by linking already existing knowledge elements available in external data sources: adding critical notes, and connecting to external databases such as DBpedia using the LOD technique. We have created links to these data sources, and thus we have made other specialist knowledge that is available in external sources accessible to the users of the dictionary. During the linking we have lifted a narrow set of knowledge from DBpedia and attached it to the relevant dictionary entries with the help of an RDF graph. This form of extension thus does not entail the recording of a specific piece of information content in an entry, as in the first case, but links an external data source to an entry. For example, by extending an entry with a DBpedia identifier we can get from the dictionary entry to a source containing encyclopedic knowledge, where the user can find a significant amount of additional information. This linking of dictionary entries and external databases may be created automatically, but in several cases it requires manual revisions (e.g. to perform semantic disambiguation).

The other element of extending from an external data source relates to linking those information contents to the entry which come from the notes made to the critical edition. The notes of the critical edition provide help during the interpretation of the text. They help the reader with information such as explanations of the less well-known proper and geographical names that appear in the work, as well as those of obsolete and dialectical words, and not commonly known phraseological expressions, among others. It extends to the historical aspects of the given work, points out the sources of the author's views and philosophical influences, and refers to the work's genesis.

Figure 2: Dictionary headword (top right) with full text citation (top left),
the attached critical annotation (bottom left), and related DBpedia entry (bottom right).

Figure 2 shows a sample headword *Konstantinápoly* 'Constantinople'. It contains all its writing variations, word forms, citations from the corpus in their extended textual context. All sample sentences are assigned a part-of-speech designation. The dictionary also marks that it is a geographical name, which is part of a country and region, as well as its exact coordinates. Linking to DBpedia we can access a huge amount of information that would not be possible to include with a traditional, paper-based entry structure. The notes on *Konstantinápoly* (Figure 2, bottom left) are attached to the entry as an external data source, and provide detailed information about Kelemen Mikes' personal attachment to Constantinople.

The result is a digital author's dictionary which supports humanities research by meeting newly emerged opportunities and needs. We exchanged the tools designed to reduce size and content to

exploit opportunities for extension, while the chopped-up content found in different data sources was linked within the entry in order to uncover connections between distant textual locations. We supplemented the traditional entry structure with information from other data sources or databases. With the help of this extended method we have created a structured knowledge base. By being linked to other databases, the dictionary now has new applications that go beyond describing the author's language. Reaching beyond a contextual interpretation that explains a specific textual location, and building a knowledge base, encyclopedic knowledge also becomes easily accessible and analyzable with computer tools.

4   Applications of the extended author's dictionary

4.1   Dictionary as the object of analysis

The extended digital author's dictionary not only constitutes a new type of dictionary, but beyond that the dictionary itself also becomes an object of further analyses. With its help we can choose research directions that have not been covered before, the analysis of which was not possible due to the lack of tools or opportunities. Our system can help researchers perform new kinds of analyses on the author's oeuvre to uncover novel linguistic and literary results. These may include a large-scale analysis of the author's language, a statistical survey of different linguistic features, computing statistics about words with certain kinds of annotations, comparing various author's dictionaries, and analyzing vo-cabulary similarities and influences or changes over time. Since this extended dictionary organizes the corpora in a very different way from the text source, its analysis may also uncover previously hidden connections or similarities between distant pieces of text.

The Mikes dictionary was part of a comprehensive analysis of language history in which a grammat-icalization process that had taken place over approximately 500 years was researched. During this period the complemented participial form *mondván* started to become a conjunction, and through analyzing the morphological and syntactic structures collected from the Mikes dictionary one of the important stages in this change can be found (Dömötör 2013). Creating a digital dictionary also enables us to uncover previously unknown text creation processes in the author's oeuvre. We have uncovered textual connections between different, distant points of the oeuvre, which would be impos-sible to collect manually. In the Mikes corpus we have uncovered approximately 500 textual similar-ities and parallels, which, although they are not a complete textual match, present smaller variations between the different excerpts. We have automated the computer-aided listing of text migration from the historical text corpus, and we can access these without a targeted word search in the 18th-century text. From this we found that some excerpts may even appear seven times in different works of the author, and we have identified which works of the oeuvre are connected in this respect. With the help of the Mikes dictionary we have been able to reevaluate previous findings in the literature concerning his lexicon. In the Mikes oeuvre, translations, as compared to his own texts, did not play an impor-tant role in the literary canon. Analyzing approximately 10,000 entries shows that the lexicon of the translations and that of Mikes' own texts differ substantially, they are a match for a maximum of 30% of cases, and so they cannot be left out of period analyses. From the analysis of the lexicon it has also become apparent that in his translations Mikes took a more modern stylistic approach not only compared to his era, but also compared to his own works, since it is in the lexicon of his translations where we can discover progressive linguistic and stylistic changes. By the 19th century the meaning condensation capability of compounds is stronger, and the perfective function of prefixes becomes determinant, as does its word creating role. These progressive changes can already be detected from the lexicon of Mikes' 18th-century translations (Kiss 2016). While analyzing Mikes' hapax legomena, we uncovered the common etymological roots of two words (*gyaur, kaffer*), and their entire history

of meaning (Kiss 2017). Further results can be expected from using the digital dictionary in the analysis of excerpts of debated authorship, in uncovering the characteristics and reasons for changes in individual style and sets of expressions, as well as exploring parallels among other authors and eras. In the works of Mikes we encountered several unresolved philological-textological problems while creating the dictionary, which due to the size of the oeuvre and use of manual tools had so far been unresolved. We were previously not able to examine these issues with such thoroughness.

4.2   Using the dictionary during corpus analysis

As further added value, the digital dictionary also enhances the efficiency of different IT-supported processes of analysis. The digital author's dictionary is a reliable tool in standardizing the morphological variety of historical texts. A dictionary with a structured setup makes it easier to search for both historical and contemporary word forms, which makes it possible to avoid the difficulties of the morphological analysis of historical texts, and through its use we can carry out the normalization of historical texts with an uneven, unregulated writing style. Whether we search for a historical or a contemporary entry, or any reference entry within those, we will get to all the other forms found in the entry. Beyond this, the search is also aided by semantic information in the corpus. With the help of the dictionary and the knowledge base linked to it, if, for example, we determine as a search parameter that we would like to list the place names connected with Turkey in the dictionary, as a result we will receive all Turkish place names listed in the dictionary, including *Konstantinápoly*. Beyond this, we can also display the geographical places collected according to different parameters on a map – for example, if we want to know where Mikes wrote each of his letters.

As an additional benefit, the dictionary can improve the quality of computational stylometry analysis. In order to compare Mikes' own writing with his translations, and to analyze the author's idiolect based on the similarities and differences between his works, we have conducted stylometry analysis on the corpus. We have used the digital author's dictionary to improve the efficiency of the statistical analytical methods, in which during the lexically-based analysis the dictionary had a lemmatization role by normalizing the different historical form variations. During the computational stylometry analyses we supplemented the statistical analysis with a preparatory (text normalizing) phase using the dictionary, in order to enhance the effectiveness of the analysis. Running a multidimensional analytical process on the entire authorial lexicon, we could determine the similarities and differences between the historical texts more exactly (Kiss & Mészáros 2016). We can thus see the differences between the analyses of the original and the normalized texts.
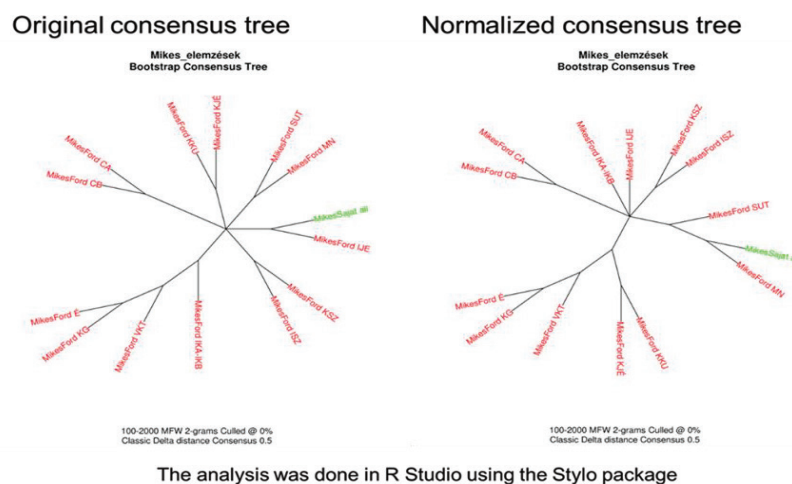

Figure 3: The results of the stylometry analysis: original [left], normalized [right].

5   Summary

If we do not consider the storage capability of large texts as the only opportunity among the advantages provided by information technology, but we also take advantage of the possibilities provided by the computer, we can change more than just the dictionary creation processes. With the extended digital dictionary we have created a type of dictionary which, extended to a knowledge base, makes it possible to access much more semantic content than before. The dictionary itself then becomes the basis of further analyses, as a new data source. With a synoptic, systematic analysis we can conduct discovery analyses for which we did not have the appropriate tools before. The methodology creates a completely new perspective and makes it possible to conduct badly needed tasks in basic research, to get a better awareness of linguistic, literary, historical, and cultural trends and processes.

The advantages that go with the extended functions of the digital author's dictionary help solve some more sophisticated, specific, and complex research problems than before. This leads us to reevaluate our expectations towards the new generation of author's dictionaries. Having rethought the role of digital author's dictionaries in the humanities, we have created an extended version of this type of dictionary, so that easily accessible, encyclopedic information content suitable for machine analysis can become even more readily accessible, and we have shown how we can enhance the efficiency of the analytical methods of historical texts with the help of our dictionary.

Most of the software tools developed during this project are open source, and they are available in the following GitHub repository: https://github.com/mtwebit/dhmine. The Mikes dictionary is also available on-line at http://mikesszotar.iti.mta.hu/, with a revised version at https://dh.mit.bme.hu/mikes/.

References

Apresjan, V., Mikulin, N. (2016). Dictionary as an Instrument of Linguistic Research. In T. Margalitadze & G. Meladze (eds.), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 224-231.

Benkő, L. (1979). *Az írói szótár.* Budapest: Akadémiai Kiadó.

Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. In *Computers and the Humanities*. 14(2), pp. 83-90.

Čermák, F., Cvrček, V. (2010). Author Dictionaries Revisited: Dictionary of Bohumil Hrabal. In A. Dykstra, T. Schoonheim (eds.), *Proceedings of the XIV EURALEX International Congress 6-10 July 2010.* Leeuwarden/Ljouwert: Fryske Akademy –Afûk, pp. 595-598.

Clarke, C. (1846). *The Complete Concordance to Shakespeare.* NewYork: Wiley and Putnam.

Dömötör, A. (2013). Idéző szerkezetből diskurzusjelölő elem: a mondván szerepei és története. In M. Csepregi et al. (eds.) *Grammatika és kontextus. Új szempontok az uráli nyelvek kutatásában III,* Budapest, ELTE, 20-30.

Gouws et al. (2013). *Dictionaries. A International Encyclopedia.* Berlin, Boston: De Gruyter Mouton.

Gouws (2014). Article Structures: Moving from Printed to e-Dictionaries. In *Lexikos 24,* AFRILEX-reeks/series 24, pp. 155-177.

GWB (1978-). *Goethe-Wörterbuch.* Berlin-Brandenburgischen Akademie der Wissenschaften, der Akademie der Wissenschaften zu Göttingen und der Heidelberger Akademie der Wissenschaften (Hrsg.), Stuttgart: W. Kohlhammer.

Hartmann, R.R.K., James, G. (2001). *Dictionary of Lexicography.* London and New York: Routledge.

HHC. *Hungarian Historical Corpus.* Accessed at http://www.nytud.hu/hhc/ [31/03/2018]

Ibsen–Ordbok (1958). *Ibsen–Ordbok.* Ordforradet i Henrik Ibsene samlede Verker. Oslo.

Karpova, O. (2011). *English author dictionaries the (XVIth–XXIst cc.).* Cambridge: Cambridge Scholars Publishing.

Kiss, M. (2012). The Digital Mikes-Dictionary. In G. Tüskés et al. (eds.) *Literaturtransfer und Interkultralität im Exil* [...]. Bern: Peter Lang Verlag, pp. 288-297.

Kiss, M. (2016). „más értelmet adni ezeknek a szoknak": Mikes Kelemen szóhasználatához. In R. Lengyel (eds.) *Nunquam autores, semper interpretes: A magyarországi fordításirodalom a 18. században,* Bp, MTA BTK, pp. 58-68.

Kiss, M. (2017). Hitetlenek: mi köze a gyaur-nak a kaffer-hez? In *Magyar Nyelv,* 113(1), pp. 80-87.

Kiss, M., Mészáros, T. (2016). Creating an extended author's dictionary to support digital literary research, In *DH Benelux 2016*, Luxembourg. Accessed at http://www.dhbenelux.org/wp-content/uploads/2016/05/89_Kiss-Meszaros_FinalAbstract_DHBenelux_2016_long.pdf [31/03/2018]

Lew, R. (2013). From paper to electronic dictionaries: Evolving dictionary skills. In D. A. Kwary et al. (eds.) *Lexicography and dictionaries in the information age*, Selected Papers from the 8th ASIALEX International Conference, [Bali, 20-22 August, 2013], Airlangga University Press, pp. 79-84.

Mártonfi, A. (2014). Számítógép és írói szótár – különös tekintettel a készülő József Attila szótárra. In *Magyar Nyelv,* 110(1), pp. 30-46.

OSS (2003-2018). *Open Source Shakespeare.* Accessed at http://www.opensourceshakespeare.org/concordance/ [28/03/2018]

J. Soltész, K. et al (1973-1987). *Petőfi-szótár.* Budapest: Akadémiai Kiadó.

Vinogradov, V. V. [виноградов, виКтор владимирович] (1956–1961). *Словарь языка Пушкина 1-4.,* Государственное Издательство Иностранных и Национальных Словарей, Москва.