2020 PRACE JĘZYKOZNAWCZE XXII/2
ISSN 1509-5304 DOI 10.31648/pj.5231 23–37

Monika Czerepowicka
Uniwersytet Warmińsko-Mazurski w Olsztynie
ORCID: https://orcid.org/0000-0002-4697-7058
e-mail: monika.czerepowicka@uwm.edu.pl

Dictionaries and lexical databases of the Polish language

Słowniki a leksykalne bazy danych języka polskiego

Abstrakt
Przedmiotem badań jest jednojęzyczna leksykografia elektroniczna. Celem artykułu jest ukazanie wpływu technik komputerowych na organizację, rozmiar, przeznaczenie i zawartość słowników. W swych badaniach autorka koncentruje się na elektronicznych bazach danych. Definiuje, czym są, oraz objaśnia, jak ich budowa i sposób organizacji zgromadzonych w nich danych wpływają na postać słowników elektronicznych. W artykule zostały poddane analizie trzy współczesne słowniki języka polskiego: *Uniwersalny słownik języka polskiego PWN*, *Wielki słownik języka polskiego PAN* oraz *Słownik gramatyczny języka polskiego*. Autorka dowodzi, że sposób organizacji i prezentacji wiedzy w omówionych dziełach umożliwia użytkownikom korzystanie z nich w sposób zaawansowany, co oznacza sprawne dotarcie do szczegółowych informacji o jednostkach leksykalnych, grupowanie ich, jak również doraźne kompilowanie „podsłowników", spełniających określone oczekiwania odbiorców.

Słowa kluczowe: leksykografia elektroniczna, baza danych, słownik

Abstract
The subject of the study is monolingual electronic lexicography. The aim of the article is to show the impact of computer techniques on the organization, size, purpose and content of dictionaries. In the present study, the author focuses on electronic databases and defines what they are, explaining how their structure and the data collected in them influence the form of electronic dictionaries. The article analyses three modern Polish dictionaries: *Uniwersalny słownik języka polskiego PWN* [*The Universal Dictionary of the Polish Language PWN*], *Wielki słownik języka polskiego PAN* [*The Great Dictionary of the Polish Language PAN*] and *Słownik gramatyczny języka polskiego* [The *Grammatical Dictionary of Polish*]. The author argues that the way of organizing and presenting knowledge in the discussed reference works allows users to exploit them in an advanced way, which leads to efficient access to detailed information about lexical units and their grouping, as well as ad-hoc compilation of "sublists" that meet specific expectations of recipients.

Key words: electronic lexicography (e-lexicography), database, dictionary, reference work

Contemporary lexicography has become electronic to a significant extent, which is a consequence of the use of computers in linguistic studies, a tendency that has been observed for several decades at different stages of compiling a dictionary. Text corpora[1], which are the source of linguistic units that are recorded and whose meanings are isolated, play a crucial role in the process (Kosek et al. 2018). The effect of technology and computers is also seen in gathering and storing information necessary to produce a dictionary. Moreover, lexicographers use computers for editing entries, and at the final stage of dictionary preparation, i.e. publication. Apart from the above another important feature of electronic lexicography is that it takes into account, more than traditional lexicography, the wide perspective of users. One can mention here accommodating to users' needs in designing and producing lexical tools. Dictionary user studies point out features and expectations of users, and lexicographers make efforts to include in dictionaries exactly those entries and information about them which users look for (cf. Lew 2015, Lew et al. 2013, Wójtowicz 2017). Except obvious limitations related to usage of electronic sources (computer literacy, sometimes access to the Internet), electronic dictionaries and encyclopaedias do not require any particular skills from their users. They are typically characterized by a user-friendly, intuitive interface and an extended network of hyperlinks, which help to navigate between entries efficiently. Additionally, lexicographers increasingly frequently involve users in building and developing electronic dictionaries by making them interactive. The relation of dictionaries to IT tools, as well as the structured knowledge they present have a significant effect on the form of electronic dictionaries. It seems that at the stage of gathering and storing information electronic databases play a major role. Advantages of database data organisation are also noted by authors of printed works. An example can be *Słownik frazeologizmów eponimicznych* [Dictionary of Eponymous Idioms] (Czeszewski, Foremniak 2011), in which the appearance of an entry resembles a database record (cf. Bańko 2013).

According to Bergenholtz and Nielsen (2013), electronic databases provide a foundation for all dictionaries created nowadays. It should be noted that the use of databases is not accompanied by information about them. Although authors of dictionaries discuss in detail the notation methods applied for the units, macro- and microstructure of their work,

[1] In this text, a corpus of texts is technically understood as a set of texts gathered for the purpose of the research which is made available in an electronic version (cf. Korczakowska 2001, Przepiórkowski et al. (eds.) 2012).

research studies presenting the structure and the character of databases used in a given dictionary are rare. This is regrettable, since a properly designed database makes it possible to using a dictionary in an advanced manner. However, it seems justified to indirectly draw conclusions concerning databases, since "each dictionary (…) has the structure of a database, and the development of lexicography was visible, among others, in the fact that over time, the structure of entries became more consistent, increasingly better adapted to the implicit *database* pattern" (Bańko 2013: 14[2]). This study constitutes an attempt to reconstruct the role of databases in preparing selected monolingual electronic dictionaries of the Polish language. Before proceeding to presenting examples, some terminological references will be made.

In the present study, a database, following Bergenholtz and Nielsen (2013: 79), is understood as a structured set of values, based on which language units – in the case of dictionary bases – are described. The main reason to apply databases is to give structure to the data which, in turn, enables their proper management, as well as their presentation. A database presents a specific type of structure, a properly designed container, in which, in the right place, individual pieces of information are located. A database thus defined exists independently of the data, although it should be assumed that their type and character determine its structure. It is believed that databases are extensive – they contain more information of various types than it is presented in a given dictionary. A lexicographer saves in them as much and as detailed information on a given unit as possible, yet decides to present to the user only certain selected information. The type of information excerpted from a database for publication can be determined by the type of dictionary produced with its use. If we assume that the database records the widest possible range of information on linguistic units (pronunciation, spelling, division into syllables, inflection pattern, word-formation structure, meaning, syntax information, pragmatic information, collocations, set phrases, examples), then it can be used to build numerous separate dictionaries using individual types of information, such as a pronunciation dictionary or a collocation dictionary. This feature can be particularly appreciated by publishers and used in commercial dictionary-making, which still has a high share in Polish lexicography (cf. Saloni 2003). Although academic lexicography,

[2] Originally: „każdy słownik […] ma strukturę bazy danych, a rozwój leksykografii wyrażał się m.in. w tym, że z upływem czasu budowa artykułów hasłowych stawała się coraz bardziej konsekwentna, coraz lepiej dopasowana do implicytnego *bazodanowego* wzorca".

i.e. projects financed with public funds (grants of the National Science Centre, the National Humanities Development Program in Poland), has a smaller social reach (this type of lexicons are usually known to a small group of users, mainly specialists), it often exceeds commercial projects as regards the method of applying technical achievements. This article will discuss three electronic human-readable dictionaries of the Polish language, also intended for non-specialists. No account will be taken of studies created only for machine processing of the Polish language, in which the complexity degree with regard to the data presentation structure and the possibility to read it by a non-specialist user would require separate reflections.

Uniwersalny słownik języka polskiego PWN [The Universal Dictionary of the Polish Language of the Polish Scientific Publishers PWN]

The method and the scope of applying databases in lexicographic practice can be highly differentiated. The more elaborate the database structure, the more detailed information the user receives. It seems that databases, due to their nature, provide an ideal method for recording mutual lexical relations between lexemes. An example can be found in *Uniwersalny słownik języka polskiego PWN* [The Universal Dictionary of the Polish Language PWN; USJP] (and in its twin publication: *Wielki słownik języka polskiego PWN* [The Great Dictionary of the Polish Language PWN]). A significant number of lexemes have been included in the network of word relations, along with the relations of synonyms, antonyms, hyponyms, hypernyms, holonyms and meronyms. Information about the network of lexical relations is displayed in the right margin of the main view of the entry or in the table opened by pressing the icon just behind the headword, cf. (Fig. 1).

Each element in the relation network is a hyperlink, which helps the user to easily navigate between entries. USJP, as it can be read in the information on the publisher's website, combines resources of several dictionaries (https://sjp.pwn.pl/oferta/haslo/Uniwersalny-slownik-jezyka-polskiego-serwis-on-line;5725041.html). Including specific dictionaries into a general dictionary enables "reversibility" of the operation, i.e. separation and browsing limited subsets. This makes it possible to search for a given sequence separately among idioms, terms, examples and proverbs, which can be achieved by the so-called advanced search, cf. (Fig. 2).

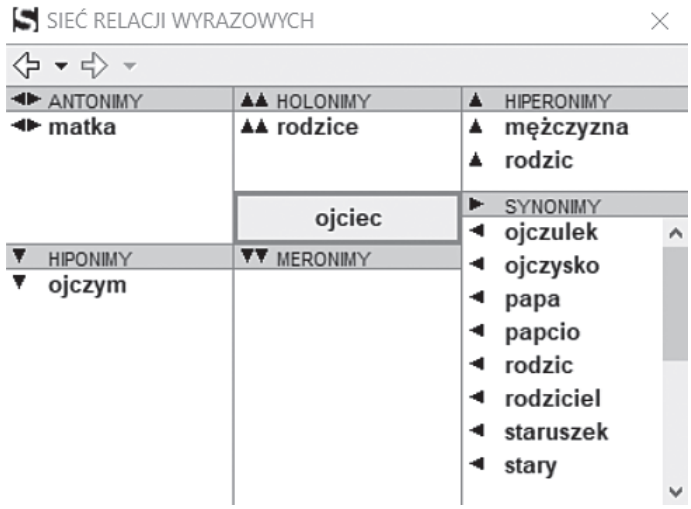

Figure 1. The lexical relation network for the entry OJCIEC [father] in USJP

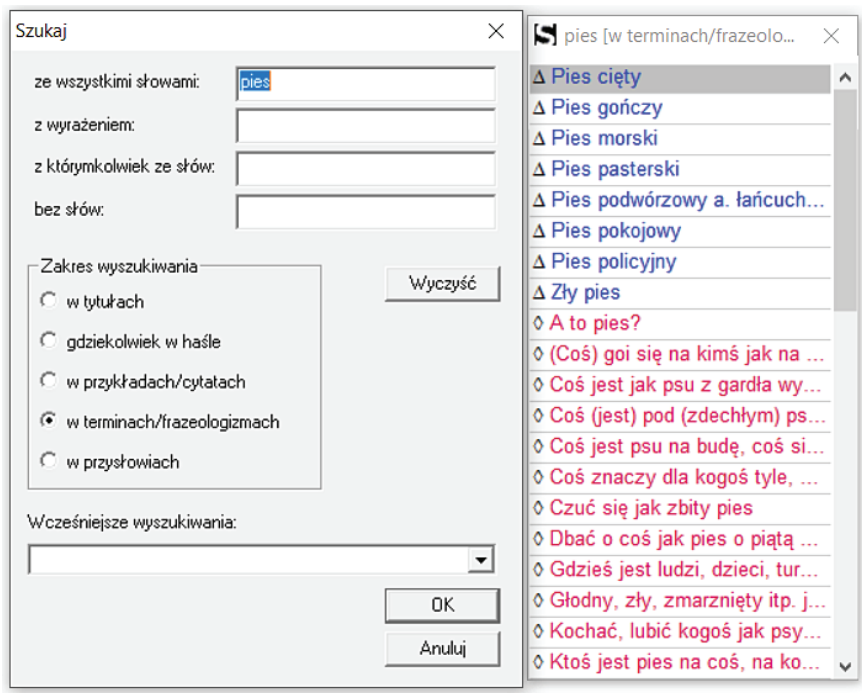

Figure 2. Results of idiom search with the sequence PIES [dog] in USJP

However, the possibilities of selective dictionary browsing, as well as compiling subsets in USJP are quite limited. Contrary to the name, advanced search does not offer specific or specialized options. The dictionary is equipped with a morphological analyzer, thanks to which the search engine interprets the given string bilaterally – not as a sequence of letters, but as a form of the lexeme and filters out entries accordingly. However, search options are limited to three elements of the entry's structure: its head, examples and word combinations (terms, phraseology, proverbs). Although dictionary entries have inflected information, as well as, where necessary, numerous stylistic or domain qualifiers in the search for this type of information cannot be included. Entries with a given qualifier can be reached indirectly only. One can search the contents of dictionary (command: anywhere in the entry) according to a given string of letters. However, this is not enough from the modern user's point of view and the possibilities offered by other available dictionaries. One of the reasons behind that is certainly the age of the tool.

The work was created more than a decade ago[3] and at that time it was available in two versions – traditional (i.e. printed) and electronic. The graphic design has been slightly adjusted in new editions, but the method of using the dictionary has not been significantly altered. The publisher has responded to technological developments only regarding the access to the publication. Originally, the electronic version was attached to the paper edition in the form of a CD-ROM and later on a storage medium (USB flash drive), while nowadays it is possible to buy online access to a virtual disk where the dictionary is located.

Wielki słownik języka polskiego PAN [The Great Dictionary of the Polish Language of the Polish Academy of Sciences]

Much greater possibilities are offered to users by *Wielki słownik języka polskiego PAN* (WSJP PAN) [The Great Dictionary of the Polish Language of the Polish Academy of Sciences], a result of the research project, available online free of charge. The dictionary is a modern work, not only in terms of compiling the dictionary material, but also in terms of technological solutions applied. The electronic form is not secondary or additional, but the only one. The machine-readable dictionary (MRD)

[3] The first edition of the paper dictionary comes from 2003. The electronic version on a CD-ROM was published a year later.

orientation is clearly visible in the work. The information in the dictionary is sorted into relevant categories of knowledge, which correspond to the elements of the entry structure. Lexical data[4] are classified according to types of information, such as the origin of the word, presence in previous or historical dictionaries (chronologisation), meaning(s) (definitions), thematic qualification, the occurrence of a given lexeme or its form in word combinations (collocations), examples (quotations), inflection. Entries are constructed according to a repetitive scheme. Parameters concerning the form of the unit, i.e. the origin of the word and presence in other dictionaries, are displayed just after the headword, which is followed by subsequent meanings. Since the basis for the description is a unit in a given meaning, variable features of the unit are repeated with each definition, i.e. position in a thematic qualification, word combinations, quotations and inflection. Dictionary entries are placed in a non-alphabetical order (cf. Zasady 2018), thus access to the entries is mainly through the search tool. The dictionary can be searched for a set format or features included in the database, which helps to rank the units based on individual parameters efficiently. A set sequence can be searched for in selected places in the entry: headword, definition, collocations or examples. Since the search system makes it possible to combine criteria, it is possible to quickly filter out units of the same origin or similarly defined, to be more precise: units whose definitions employ identical expressions, which in turn suggests that there were strict instructions lexicographers followed defining units in the dictionary, e.g. (Fig. 3).

Apart from a detailed list of quantifiers (chronological, normative, expressive, specialist, domain, geographical, environmental, stylistic), the dictionary applies a thematic classification which is very innovative for a general dictionary of this size. Eight thematic categories have been identified, e.g. man as a psychological and physical being, everyday life of man, man and nature and physical categories. The set of displayed units can be narrowed down as regards the level of main categories and inside them, and consequently, thematic subsets can be compiled, e.g. vocabulary related to various characteristics of matter, human appearance, customs and traditions or family relations. Thanks to the thematic classification used, the dictionary is a very helpful tool in learning Polish vocabulary.

[4] At this point, the focus is on the method of vocabulary compilation, since grammatical (mainly inflectional) information was derived from *Słownik gramatyczny języka polskiego* [*Grammar Dictionary of the Polish Language*], which will be discussed later in the present study.

A well-designed database and the search system mean that the dictionary, which satisfies various expectations of the recipient (*ad hoc* but specific), does not seem to be wishful thinking for the future. Within the database, the dictionary provides tools to model both the network of entries and the type of information browsed by the user. It becomes a multifunctional mother-dictionary, acting as a basis for narrower dictionaries, such as specialised dictionaries or thesauri.

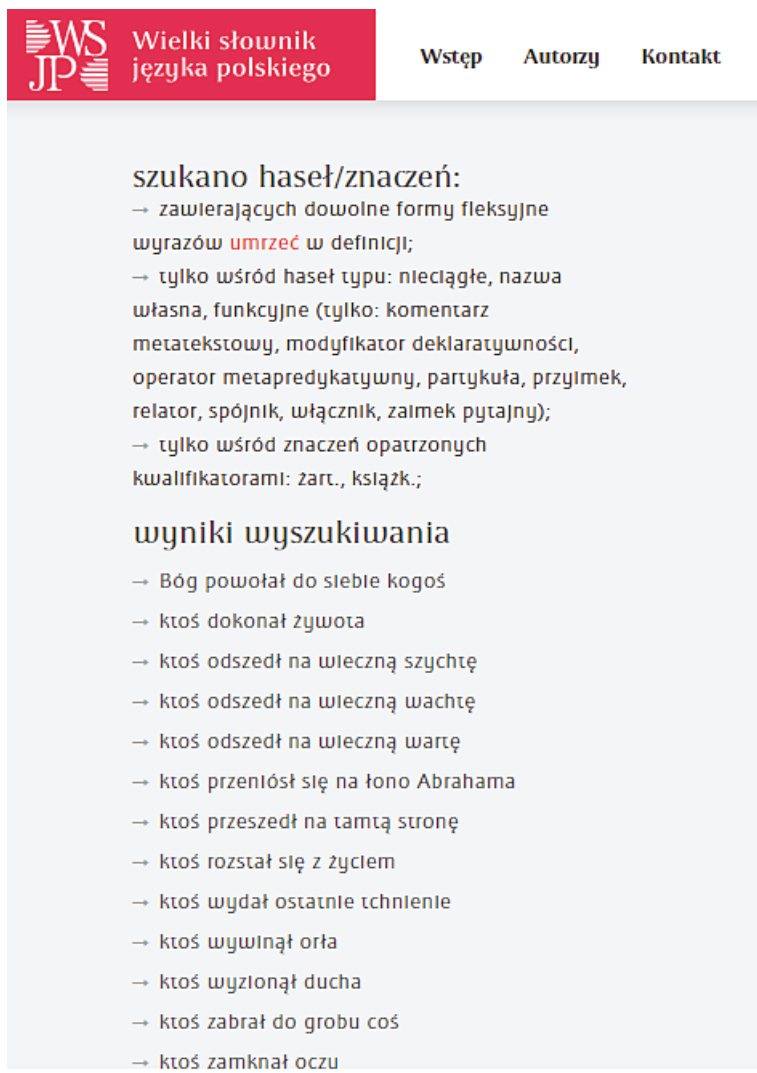

Figure 3. A fragment of a WSJP PAN entry list of multiword units with a sequence UMRZEĆ [die] in the definition and with quantifiers "playful" or "literary"

The lexical material in WSJP PAN is also divided according to grammatical parameters (parts of speech, abbreviations, acronyms) and according to the number of segments (*simple* means lexemes, and *discontinuous* – multiword units). Since filters can be combined, a query results in units not only with specified semantic features or belonging to specific thematic fields, but also with specified grammatical characteristics, e.g. (Fig. 4).

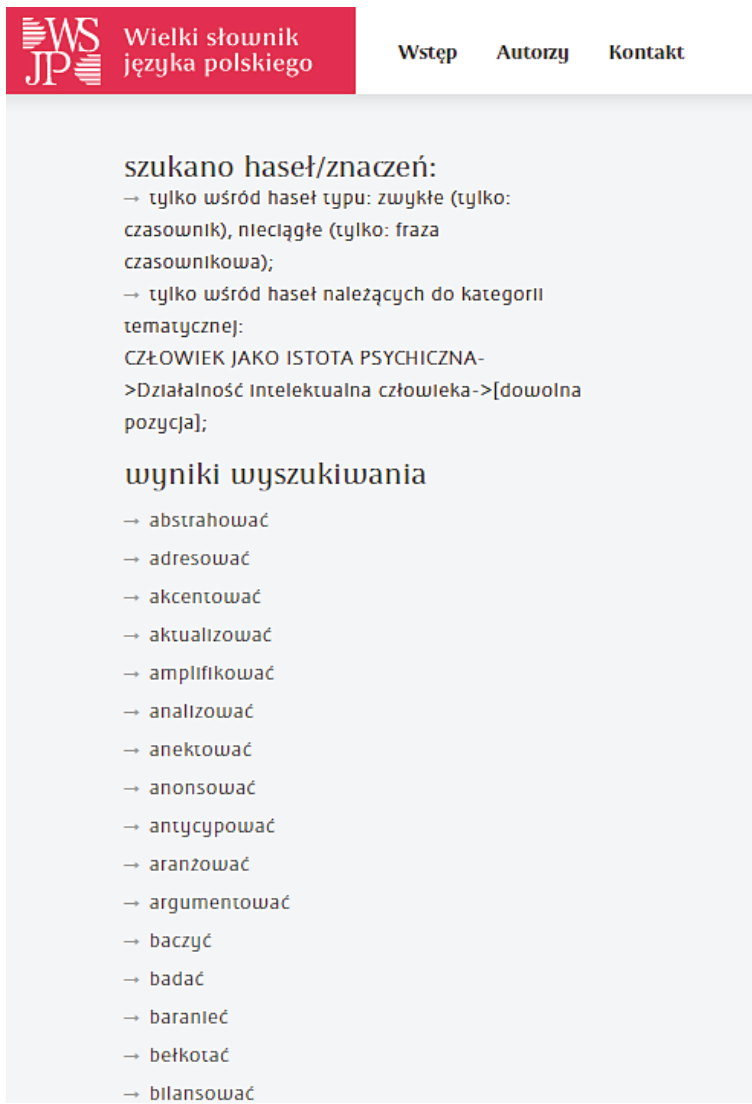

Figure 4. A fragment of a list of verbs related to human intellectual activity in Polish

Designing a database with the use of various values clearly shows the advantages of electronic dictionaries over printed ones. This facilitates extracting from a large set precisely those entries that are required by the user, even when they are quite specific, such as multiword units of Russian origin, technical terms of Finnish origin, or archaic vocabulary (Fig. 5). It would probably be nearly impossible to find analogous entries in traditional (printed) dictionaries, and certainly, it would be extremely laborious.

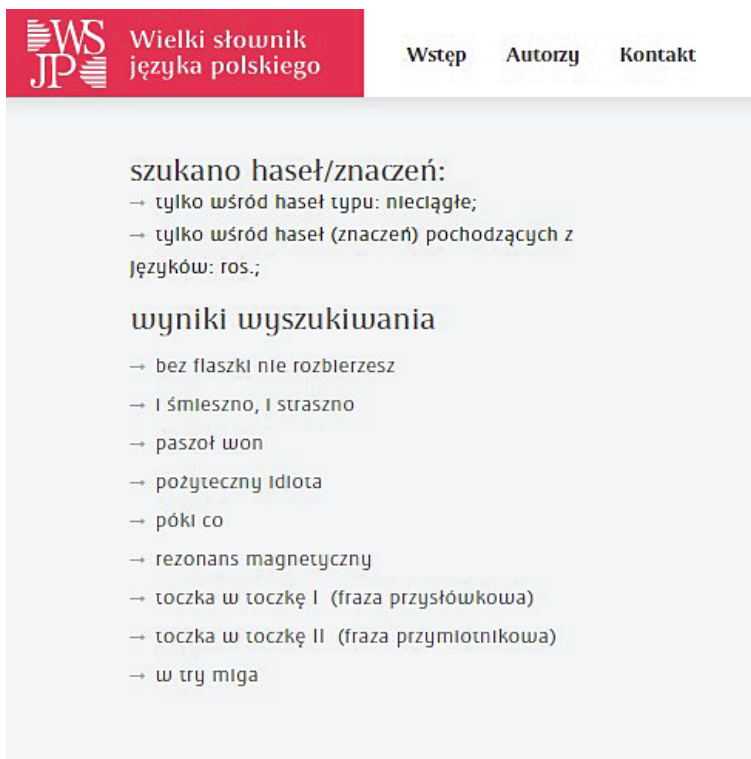

Figure 5. A list of multiword entries originating from the Russian language in WSJP PAN

The richness of the lexical material, its detailed organization and extensive search options are of great value in the dictionary. However, one has to honestly admit that not all types of information can be searched in a precise manner. This concerns, for instance, inflectional information, which in principle is not a search criterion. A user interested in Polish inflection should reach for another work – *Słownik gramatyczny języka polskiego* [The Grammatical Dictionary of Polish].

Słownik gramatyczny języka polskiego [The Grammatical Dictionary of Polish]

The possibility of detailed and specialised separation of subsets and temporary modelling of the dictionary for the needs of the user can be clearly seen in *Słownik gramatyczny języka polskiego* (SGJP) [The Grammatical Dictionary of Polish]. Such modelling appears possible through a detailed design of the database and advanced filtering of the dictionary content. SGJP consists of more than 330,000 lexemes found in other dictionaries and texts in the Polish language and of potential forms, i.e. forms belonging systemically to the paradigm, but with zero or close to zero frequency in texts. These are for instance verb forms of 1st and 2nd person singular of a neuter past tense (*lubiłom*, *lubiłoś*; "liked"), or in the case of nouns, the plural nominative form of selected masculine-personal nouns (type of *szulery* or *sługusi*). Thanks to noting both already retrieved and potential forms, the dictionary is the most extensive database of the Polish language.

The work is of a grammatical type. It is worth noting that the presentation of the inflection pattern differs from the method usually applied in dictionaries. The authors do not use the symbolic notation of paradigms, but provide all forms directly, gathered in clear tables. Dictionary entries are arranged in a standard alphabetical order (*a fronte*), but a reversed order (*a tergo)* can be easily obtained. The latter is very useful and appropriate in grammatical research. Entries are accessed via the list displayed on the left-hand side of the window or via the search tool, in which a specific sequence of characters is entered. In the case of homonymic forms, the dictionary suggests all lexemes to which a given form belongs. The lexical material is divided into nine main grammatical classes: noun, adjective, numeral, verb, adverb, preposition, abbreviation, prefix and invariables. Within some classes, a detailed division is applied according to morphological similarity, e.g. gerunds are included with nouns, past and present participles – with adjectives. The class of indeclinable lexemes, which include all indeclinable lexemes apart from pronouns, is divided in the most detailed manner. Based on syntactic criteria, among the invariable lexemes, the following groups were distinguished, among others: conjunction, particle, declarative modifiers, mode operator, as well as elements of set phrases (forms particularly related to the context, not occurring outside the set phrase, e.g. *bździu*, found only

in the *fiu-bździu* combination, cf. Saloni 2012: 141). With a simple query, it is possible to filter out lexemes of any class from the entire set.

However, the grammatical information included in the dictionary is much richer. The fragmentation of the grammatical information in the database not only into grammatical classes and inflectional patterns, but also the type of pattern, the number of patterns and detailed grammatical parameters of lexeme classes, such as type, number of types, aspect or reflexivity, makes it possible to rank entries based on the specified features and parameters. The dictionary is also distinguished by the level of detail in its inflection description. For instance, in verbs, there are more than two hundred identified conjugation patterns (221 for inflected lexemes and one for predicative expressions; Saloni et al. 2015). For the sake of comparison – Jan Tokarski's classification, applied in most lexicographic works, provides 11 main conjugation patters (19 including subgroups).

A detailed isolation of inflection patterns indicates a rigorous, algorithmized description perspective in research. This kind of data gathered in the database, combined with an appropriate filtering system, facilitates compiling grammar sub-dictionaries with specific features. Search criteria, i.e. individual parameters, can be combined either on a conjunction (operator "and") or alternative ("or") basis. Filtering is carried out according to two types of criteria: closed and open. The first criterion consists in preparing the query based on parameters identified in the dictionary. For example, from the entire set of verbs (almost 30,000, it is possible to isolate regular[5] two-aspect units (imperfective/perfective, perfective/imperfective), and there are 255 of them. This set can be narrowed further on, e.g.. in terms of reflexivity, to finally obtain a list of 10 verb lexemes with obligatory reflexivity (Fig. 6).

On the other hand, the open filtering system is equipped with a unique mechanism, not found in other dictionaries – the syntax of regular expressions applied in programming. Regular expressions are patterns used to describe character sequences (words) in a symbolic way. To achieve this, appropriate special characters[6] and brackets are used. Nevertheless, taking into account search results, it is the "cross search"

[5] Filtering provides more variants, i.e. with the facultative value of aspect: impf/(pf), pf/(impf), as for instance, in verbs: DELEGOWAĆ, EKSPEDIOWAĆ, HOSPITALIZOWAĆ and DAROWAĆ, OFIAROWAĆ, STAWIĆ.

[6] For example, individual graphic characters have the following meaning: the dot (.) replaces any character, "^" is the character denoting the beginning of the sequence – a row or a word, "$" – the character denoting the end of the sequence, the pipe (|) signifies an alternative, etc.

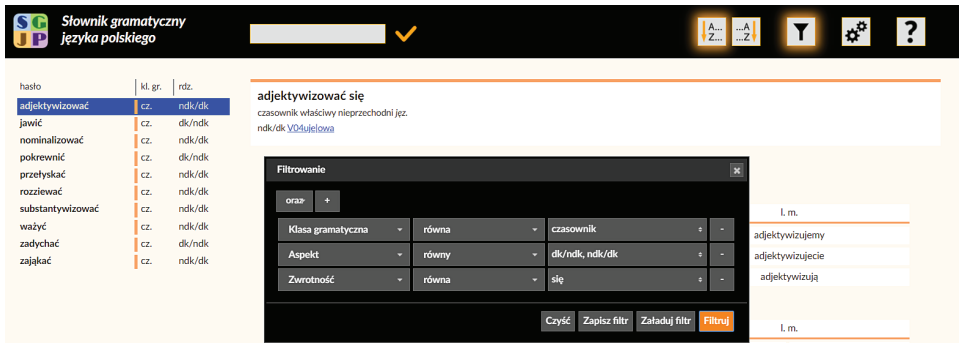

Figure 6. A list of verb entries satisfying specified conditions in SGJP

that seems most important, i.e. the possibility of combining both types of criteria – open and close – and searching through the whole set according to the specified value and its shape. For instance, the dictionary contains more than 170,000 noun lexemes. Even from such an extensive set, it is possible to extract a narrow list of entries with specified parameters: multi-gender nouns, in which one of the genders is masculine sub-gender m3, and the second letter is *d* (Fig. 7). In addition, filtering is carried out not only according to a specific parameter ("equal to something") but also with its exclusion ("different from").

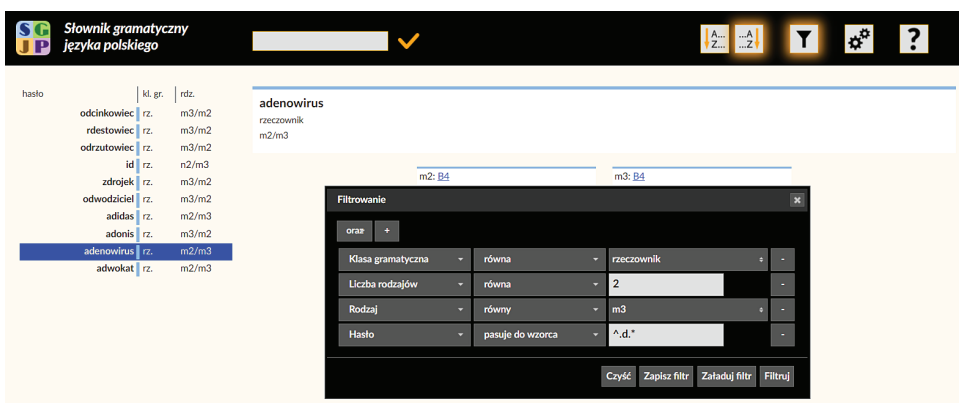

Figure 7. A list of noun entries with specified parameters in the *a tergo* order

The presentation of paradigms *in extenso* and multiple possibilities of narrowing down the content of the dictionary make it a valuable tool for both native users of Polish (including specialists) and learners of Polish as a foreign language.

Conclusions

It can be said that integration with the user is one of the most characteristic features of contemporary electronic lexicography, apart from its technological advancements. The vision of a user who requires comprehensive information located in one place (Żmigrodzki 2008: 97) and from which he can select the required dictionary is no longer a futuristic image. Additionally, the structure and purpose of dictionaries have also been changing. Apart from dictionaries intended for people (*human-readable*), an increasing number of works (computer programs) are being created for machine processing of the Polish language. As a result of using IT tools, aiming at hierarchical arrangement of knowledge, some of them combine features of both types of dictionaries, such as the above-mentioned SGJP, WSJP PAN and other projects not discussed in this study, e.g. Walenty (Przepiórkowski 2014), SEJF (Czerepowicka 2014, Czerepowicka, Savary 2018), Verbel (Czerepowicka et al. 2014). The image of contemporary lexicography has been dynamically changing. It is hoped that the examples discussed in the present article confirm that those changes are mainly positive.

Literature

Bańko M. (2013): *Za mało czy dość? Dylemat leksykografa w dobie rewolucji informacyjnej.* [In:] *Perspektywy współczesnej frazeologii polskiej. Między teorią a praktyką frazeograficzną.* G. Dziamska-Lenart, J. Liberek (eds). Poznań, pp. 9–16.

Bergenholtz H., Nielsen J. S. (2013): *What is a Lexicographical Database?* "Lexicos" 23, pp. 77–87.

Czeszewski M., Foremniak K. (2011): *Ludzie i miejsca w języku. Słownik frazeologizmów eponimicznych*. Warszawa.

Czerepowicka M. (2014): *SEJF – Słownik elektroniczny jednostek frazeologicznych.* „Język Polski" XCIV, z. 2, pp. 116–129.

Czerepowicka M., Savary A. (2018): *SEJF – A Grammatical Lexicon of Polish Multiword Expressions.* [In:] *Human Language Technology. Challenges for Computer Science and Linguistics.* Z. Vetulani i in. (eds). Berlin–Heildeberg, pp. 59–73.

Czerepowicka M., Kosek I., Przybyszewski S. (2014): *O projekcie elektronicznego słownika odmiany frazeologizmów czasownikowych.* „Polonica" XXXIV, pp. 115–123.

Dubisz S. (ed.) (2018): *Wielki słownik języka polskiego PWN*. Warszawa.

Korczakowska M. (2001): *Polskie korpusy tekstów (wybrane zagadnienia).* „Prace Językoznawcze" III, pp. 65–75.

Kosek I., Przybyszewski S., Czerepowicka M. (2018): *Źródła w tworzeniu elektronicznego słownika paradygmatów frazeologizmów werbalnych.* „Biuletyn Polskiego Towarzystwa Językoznawczego" LXXIV, pp. 47–59.

Lew R. (2015): *Dictionaries and Their Users*. [In:] *International Handbook of Modern Lexis and Lexicography.* P. Hanks, G. M. de Schryver (eds.). Berlin–Heildeberg. DOI: https://doi.org/10.1007/978-3-642-45369-4_11-1.

Lew R., Grzelak M., Leszkowicz M. (2013): *How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study*. "Lexicos" 23, pp. 228–254.
Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B. (eds.) (2012): *Narodowy Korpus Języka Polskiego*. Warszawa.
Przepiórkowski A., Skwarski F., Hajnicz E., Patejuk A., Świdziński M., Woliński M. (2014): *Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego*. „Polonica" XXXIII, pp. 159–178.
Saloni Z. (2003): *Głos w dyskusji*. [In:] *Językoznawstwo w Polsce. Stan i perspektywy*. S. Gajda (ed.). Opole, pp. 79–80.
Saloni Z. (2012): *Podstawy teoretyczne „Słownika gramatycznego języka polskiego"*. Warszawa.
Saloni Z., Woliński M., Wołosz R., Gruszczyński W., Skowrońska D. (2015): *O Słowniku*. [In:] *Słownik gramatyczny języka polskiego*. Edition III. Online: http://sgjp.pl/o-slowniku/.
Saloni Z., Woliński M., Wołosz R., Gruszczyński W., Skowrońska D. (2015): *Słownik gramatyczny języka polskiego*. Edition III. Online: Warszawa, http://sgjp.pl/.
Wójtowicz B. (2017): *Revisiting Lemma List in Swahili Dictionaries*. "Lexicos" 27, pp. 561–577.
Zasady opracowania hasła WSJP (2018), https://wsjp.pl/index.php?pokaz=zasady&l=1&ind=0.
Żmigrodzki P. (2008): *Słowo – słownik – rzeczywistość*. Kraków.
Żmigrodzki P. i in. (eds.) (2018): *Wielki słownik języka polskiego PAN. Geneza, koncepcja, zasady opracowania*. Kraków, https://wsjp.pl/.