What Do Users of General Electronic Monolingual Dictionaries Search for? The Most Popular Entries in the Polish Academy of Sciences Great Dictionary of Polish

Ewa Kozioł-Chrzanowska

[1] Institute of the Polish Language, Polish Academy of Sciences, al. Mickiewicza 31, 31-120 Cracow, Poland
E-mail: ewa.koziol-chrzanowska@ijp-pan.pl

Abstract

This article reports on an analysis of the most popular entries in the online general monolingual dictionary, based on the Polish Academy of Sciences Great Dictionary of Polish (GDP). The GDP was created from scratch over 12 years. The given survey aims to present an overview of its users' needs after the completion of the first stage of work (which was the 15,000 most frequently used lexemes) and to draw conclusions which may become useful for other lexicographers facing similar challenges. The analyzed data consist of 500 most popular entries in a four-year time period. The majority (80%) constitutes multi-word expressions: phraseological units (50%), proverbs (29%) and terms (1%). All of the subgroups are varied in style, meaning and form. The remaining 20% of the most popular entries are made up by single lexemes, mostly (15%) by the ones with a low subjective probability factor. Additionally, possible reasons for such results are addressed, considering school needs as well as the content of other online dictionaries.

Keywords: dictionary use; online dictionary; monolingual dictionary

1. Introduction

For the last few decades, there has been a growing interest in the needs and habits of dictionary users. This interest has resulted in most experts now appreciating the necessity to compile dictionaries with the user needs foremost in mind (Lew, 2011: 1). Undoubtedly, intuitions and predictions cannot expand the empirical data to achieve the goal. The given article contributes to the growing literature which seeks to inform lexicographers regarding user needs and expectations.

1.1 Research on Monolingual Electronic Dictionary Use

As electronic dictionaries have been replacing printed counterparts (Lew, 2012: 243) it seems obvious that research into dictionary use is increasingly focussing on the former. However, Töpel (2014), in her paper which was part of the first volume of Lew (2015:

232), focussed on the use of online dictionaries and claims that the "description of the current state of research into electronic dictionaries makes it clear that in several areas there remains much to investigate. On the content side, both research into online dictionaries, in this case particularly monolingual dictionaries, and issues of user-friendly presentation of content have been investigated only a little or not at all" (Töpel 2014: 48). A similar statement was made a year later by Gromann, arguing that most studies empirically evaluate specific learners' dictionaries or specialised translation dictionaries (Gromann, 2015: 55). In 2012, Müller-Spitzer et al. mentioned only three studies that focus solely on monolingual electronic dictionaries. The authors stressed also the fact that most studies focus on multilingual, mainly bilingual, dictionaries or on comparing bilingual with monolingual ones (Müller-Spitzer et al., 2012: 426).

1.2 The "Polish Academy of Sciences Great Dictionary of Polish" Project

The Polish Academy of Sciences Great Dictionary of Polish (GDP) is a general dictionary of the Polish language, published online at http://wsjp.pl/. Access is open and free of charge.

The project has been underway for over 12 years. The initial idea for a new dictionary was presented in 2005. The actual lexicographical work on the dictionary began in January 2008 and continued until the end of 2012. During this period, 15,000 entries were prepared (describing the most frequently used words in the Polish language collected from corpora available from that period). The current stage of lexicographical work began in September 2013 and is expected to continue until 2018. Its aims are to enrich the previously compiled entries as well as to compile an additional 35,000 entries. The latter goal consists of preparing:

- lexemes that were already included in the dictionary (per the rule of compilation) in the meaning relations with previously compiled words;
- formative derivatives from the words already described;
- the most recent vocabulary items, which have not yet been recorded in any general Polish language dictionary.

The current stage of the project is not the last one. After its completion, the dictionary is expected to attain 50,000 main headwords (aside from entries describing idioms and proverbs) (Żmigrodzki, 2014: 37–40).

To provide a background to the current paper it seems indispensable to present the general characteristics of the dictionary:

- Two corpora (the National Corpus of Polish – www.nkjp.pl – and an auxiliary corpus created to serve the needs of the emerging dictionary) and Internet websites

constitute the sources of linguistic data for the dictionary. While preparing specific parts, other lexicographical sources are also used, e.g. the Grammatical Dictionary of Polish Language [Słownik gramatyczny języka polskiego] provides inflectional paradigms.

- The entries are compiled based on contemporary texts only: they include the sources that came into use after 1945.

- The dictionary is, in principle, descriptive, which means that the authors do not exclude from the description any lexicographical facts deemed incorrect. However, the normative unacceptability of a given fact (as per the Normative Dictionary of Polish [WSPP: Wielki słownik poprawnej polszczyzny PWN]) is highlighted.

- The dictionary employs wherever possible the achievements of Polish 20th Century linguistics, especially in the field of semantic, inflectional and syntactic descriptions of lexical units. However, the description is created in a way which is accessible to a very varied group of Polish language users[1].

- The macrostructure consists of single lexemes, idiomatic expressions and so-called functional words (prepositions, conjunctions etc.) as well as the most frequently used proverbs, abbreviations, acronyms and proper names.

- The microstructure covers a headword form (with variants); information about pronunciation (so far only for the words with unpredictable pronunciation, especially recent borrowings); chronology; etymology; description of meaning (in other words definition and, in polysemous entries, an additional guideword); thematic classification; superordinates, synonyms and antonyms of the entry word in the specific meaning; inflection (especially the full paradigm of the word's inflection, its affiliation to a part of speech); syntactic requirements (especially for verbs); collocations; full sentence quotations; abbreviations (if any); normative information (pertaining to some incorrect uses of the word); notes on usage (any other information pertaining to the usage of the word in texts). This set of information is used in the description of the two most numerous types of language units: single lexemes and idiomatic expressions (Żmigrodzki, 2014: 41–43).

1.3 Objectives – Survey Design – Tools

The aim of this article is to identify the types of the most popular entries in the GDP and to share these experiences. The paper outlines one aspect of GDP user behaviour

[1] According to Polish literature regarding this topic, the solutions considered as user-friendly are, for e.g.: the lack of abbreviations and symbols, and grouping all the information about the particular word in one place (like not using the references to inflectional information but joining it with the entry) (Żmigrodzki, 2005: 42).

(what they do, what entries they look up) and draws some conclusions regarding their needs and expectations (what they want, what kinds of entries they are prone to look up, and what are the reasons for such choices). The result of the analysis may indicate solutions for lexicographers facing the challenge of compiling monolingual general dictionaries from scratch, as well as for those continuing such work (including the GDP project itself). Undoubtedly, a dictionary should contain the entries which are needed by its users. This paper attempts to define these needs and identify their possible motivations. The latter issue is important if the conclusions are supposed to be useful for lexicographers participating in projects similar to the GDP, who should thus be able to compare GDP user motivation to that of their own users, as different motivations are reflected through different needs. The paper outlines the general interest in particular groups of entries available in the monolingual general dictionary. In other words, considering the behaviour of different users (e.g. foreign vs native speakers, children vs adults, professionals vs non-professionals) is beyond the scope of the survey. It can be assumed that all these mentioned groups (and many others) have some representation in the large set of those who entered the GDP in the analyzed period. Unfortunately, this approach may lead to an overrepresentation of the needs of those groups of people who use dictionaries more often than the others, e.g. editors, proofreaders, teachers or translators. This problem is well-known and some researchers try to solve it by devising profiles of users (Arhar Holdt et al., 2016). However, if the given analysis is to be useful for lexicographers working on a dictionary from scratch, it seems more effective to provide them with general conclusions. Meeting the expectations of different types of users is, seemingly, the next step in compiling a dictionary. This paper also provides a preliminary attempt to analyze the behaviour of GDP users – this is another reason for choosing a more general perspective for its starting point. Undoubtedly, it is going to be more detailed in the future.

The analyzed data were gathered by Google Analytics. The analysis covers 500 entries which were the most popular between 01.01.2013 and 31.12.2016. This period was chosen since 2013 was the first year of use of the dictionary after the completion of its first stage of preparation. It lasted four years and ended just before the beginning of the data collection which is presented in the current paper.

1.4 Obtaining the Data

Regarding the method of data collection, a few approaches can be distinguished: questionnaires, providing the participants with the task (e.g. a translation of the text), following user behaviour in online dictionaries and via eye-tracking. Collecting unobtrusive[2] data is more reliable in this type of research – that is why the Google Analytics tool was chosen. By "type of research" I consider the above mentioned goal:

[2] "In general, an unobtrusive method can be understood as a method of data collection without the knowledge of the participant [...]" (Müller-Spitzer et al., 2012: 427).

identifying the most popular entry types. According to many authors, analysing log files is not an ideal method of research into dictionary use; its disadvantages are considered by e.g. Müller-Spitzer et al. (2012), Müller-Spitzer et al. (2015), and de Schryver & Joffe (2004). It is probable that some of the problems raised by these authors regarding log files also apply to Google Analytics. However, as has already been mentioned, Google Analytics is a good choice for compiling a list of the most popular entries. Still, log files are used more commonly for studying user behaviour and it can be claimed that they have dominated empirical research in recent years (Lew, 2015: 235). Nonetheless, some researchers (e.g. Lorentzen & Theilgaard, 2012) have also used Google Analytics.

The list of the 500 most popular entries was compiled by making a report using: Behaviour – Site Content – All Pages and adjusting the Date Range (from 01.01.2013 to 31.12.2016). This part was executed automatically by Google Analytics. The most popular pages were ranked by measuring their page view[3] rate. The next step was to exclude those pages which did not refer to the entries, e.g. the search engine of the dictionary, the history of the dictionary, the instruction for users, the page presenting the authors; instead, this stage was completed manually. The ways of entering the sub-sites (e.g. the search engine of the GDP vs the search engine of Google, typing the whole headwords vs typing their parts only) do not fall within the scope of the survey.

2. The Study

It should be emphasised that no assumptions relating to the division of groups have been made in advance, before gathering data. In other words, the criteria for distinguishing groups of entries were prepared after ranking on the basis of character. In the study, two factors are considered: popularity of the given group of entries and its strength. The "popularity" is understood as the percentage of occurrences of the group in question in the ranking of 500 entries. The "strength" is measured in terms of the number of page views.

2.1 Remarkable groups of entries and their popularity

The first conclusion drawn from the observation of the 500 most-popular entries in the GDP is a domination of multi-word expressions over single lexemes. The latter group consists of 99 entries, whereas the former totals 401 entries. Additionally, the popularity of single lexemes is strongly correlated with their position in the rank. Among the 100 most popular entries (i.e. from 1st to 100th position) there are only 12 single lexemes, whereas among the 100 least popular ones (i.e. from 400th to 500th

[3] "A *pageview* is defined as a view of a page on your site that is being tracked by the Analytics tracking code. If a user clicks reload after reaching the page, this is counted as an additional pageview. If a user navigates to a different page and then returns to the original page, a second pageview is recorded as well." (Analytics Help, access: 13.05.2017).

position) there are 33. In the remaining hundreds, the number of single lexemes remains the same amounting to 18.

Having created the two categories (multi-word expressions and single lexemes), we face the problem of dividing them into subcategories as the abovementioned conclusion is far too general. There are many possibilities, e.g. singling out the types of multi-word expressions (MWEs) (verbal, noun, adjectival), contrasting polysemic and monosemic entries, distinguishing the loanwords. As previously mentioned, no criteria for division were given in advance. The analysis of the list led to two surprising conclusions: the proverbs appear to be extremely popular and words that seem to be part of the basic vocabulary scope are rare. The distinction of subcategories was based on these two statements.

2.1.1 Multi-word Expressions (MWEs)

Since among the MWEs proverbs are a distinctive group, the principle of looking for other subcategories was to check if there are any other types of MWEs (e.g. slogans, wing words, phraseological units). Those found in the ranking were: phraseological units and terms. These three subcategories are different in number: the subcategory of proverbs comprises the most popular entries (29%), phraseological units (50%) and terms (1.2%) (all numerical data are provided in Figure 1).

Among terms there is no regularity in form or meaning; they consist of full words as well as abbreviations (one example: *ABS* 1. 'Anti-lock Braking System', 2. 'Avalanche Airbag System'[4]) and concern different topics, e.g. *capital letter* [*drukowana litera*][5], *sign language* [*język migowy*], *collective responsibility* [*odpowiedzialność zbiorowa*].

A similar situation of ambiguousness can be found in two other groups: proverbs and phraseological units. Among proverbs there are examples of old units, *Guest at home, God at home* [*Gość w dom, Bóg w dom*], as well as quite contemporary ones, *The finger and the head are school excuses* [*Paluszek i główka to szkolna wymówka*]. The former is a proverb which encourages the warm and hospitable welcome of guests. In Polish texts, this proverb was noted for the first time in the 17th century (Krzyżanowski, 1969: 717). The latter example is used to deride the complaint of a minor ailment. This proverb has been noted since the end of 19th century (Krzyżanowski, 1972: 803). Both examples, as well as others, highlight the differentiation of mentioned topics: *Better to be safe than sorry* [*Gdyby / Żeby kózka nie skakała, toby nóżki nie złamała*], *He who is born to be hanged shall never be drowned* [*Co ma wisieć, nie utonie*], *One swallow does not make a summer* [*Jedna jaskółka wiosny nie czyni*], *After the New Year the days become longer very fast* [*Na Nowy Rok przybywa dnia na barani skok*]. It can also be

[4] The numbers mark polysemic entries.

[5] In brackets Polish equivalents are provided.

stated that GDP users were interested in popular, well-known proverbs as well as in those which are rarely used. The information regarding the popularity of proverbs is drawn from the research that established a paremiological minimum of the Polish language.[6] The latest one was completed in 2013 (Szpila, 2014) and it contains, for example, *Where two are fighting the third wins* [*Gdzie dwóch się bije, tam trzeci korzysta*], *What goes around, comes around* [*Co się odwlecze to nie uciecze*], *It is a mixed blessing* [*Każdy kij ma dwa końce*]. These proverbs are present in the paremiological minimum as well as in the ranking of the 500 most popular entries in the GDP. However, there are also units absent from the minimum list, even in its extended version from 2013[7]. Here are some examples: *Corruption starts at the top* [*Ryba psuje się od głowy*], *Humility gets you everywhere* [*Pokorne cielę dwie matki ssie*], *A nobleman at the farm is equal to a palatine* [*Szlachcic na zagrodzie równy wojewodzie*].

The differentiation in origins, meanings, forms and stylistic features is also characteristic for the most popular phraseological units in the GDP. Some of them originate from the Bible, mythology or literature, e.g. *Aesopian language* [*język ezopowy*], *Balzakian age* [*wiek balzakowski*], *in the arms of Morpheus* [*w objęciach Morfeusza*], *thorn in the side* [*cierń w oku*]; whereas others are quite new and originate from colloquial language: e.g., humorous equivalent of alcoholic drink [*napój wyskokowy*], units that can be translated literally as *a warmed up chop* [*odgrzewany kotlet*] 'sth that was known in the past, but then was forgotten and is currently presented falsely as sth new' and *sth gobbled so well and then it croaked* [*tak dobrze żarło i zdechło*] 'sth was going well, but the difficulties occurred'. Some traditional phraseological units referring to the world of nature or traditions passing by can be indicated here as well. One such unit can be literally translated as *a spoon of tar* [*łyżka dziegciu*]. The word used here refers to a kind of tar which is made in a process of distillation of wood. It has antiseptic and antifungal characteristics. The meaning of the unit is 'sth unpleasant in a generally good situation'. The phraseological units are also different in their forms – clause, noun, verb, adjective, adverb, exclamation: *hit the bull's-eye* [*strzał w dziesiątkę*], *abc* [*abc*], *the scales fell from sb's eyes* [*łuski spadają z oczu komuś*], *to loosen sb's tongue* [*język rozwiązał się komuś*], *pitch dark* [*choć/że oko wykol*], *my word* [*masz babo placek*]. A wide variety in style can be observed. The gathered units are bookish, *sb takes sb for a ride* [*ktoś gra znaczonymi kartami*] as well as neutral *light sleep* [*lekki sen*] and informal: *you pay your money and you take your* [*do wyboru, do koloru*], *bullshit* [*o dupie Maryni*].

[6] Paremiological minimum is „a set of proverbs that all members of society know or an average adult is expected to know" (Ďurčo, 2015: 183).

[7] The results of the survey were divided into two parts: the proverbs which were indicated by at least 8% of informants and fewer up to two informants; the latter version includes 254 examples (Szpila, 2014: 91-93).

2.1.2 Single Lexemes

As per previous observations, one preliminary conclusion was that few ranked entries seemed to be part of the basic vocabulary range. This statement was a starting point for distinguishing subcategories of single lexemes: lexemes with low and high frequency.

As the correlation between the corpus frequency of a word and the frequency of look-ups in online dictionaries is a subject of analyses[8], applying this method should probably be the natural choice for a given study. However, extracting the list of frequently used lexemes from the National Corpus of the Polish Language [NKJP] did not seem the best solution because of the dominance of the written texts. The National Corpus of the Polish Language contains about 10% of speech data. However, most are media speech data (transcriptions of TV and radio programs) and transcriptions of parliamentary speeches and discussions. The conversational speech data (transcriptions of dialogues of people of different ages, different education levels and descent) amount to about 900,000 tokens (Pęzik, 2012: 38-39). The problem of overrepresentation of data of the written language was also raised by Janusz Imiołczyk who addressed this problem with reference to frequency dictionaries (1987: 24). One of the aims of the basic frequency dictionary completed by the author (Imiołczyk, 1987) was to cope with this problem. To achieve the goal, Imiołczyk conducted a psychometric experiment in which he prepared a list of about 5,000 lexemes ordered according to the rule of subjective probability. He asked informants to fill in the questionnaires by labelling the given lexemes with numbers from 1 to 7 (where 7 means the word is used constantly and 1 means that the word is unknown or never used). For statistical analysis, he provided each lexeme with a rank (Imiołczyk, 1987: 34–39). The author claims that the frequency of words was not the only criterion used by informants; other important issues were: ordinariness, abstraction, connotative meaning (Imiołczyk, 1987: 48). This approach constitutes the next argument for using Imiołczyk's list instead of the rank list extracted from the corpora. Of course, the fact that the list was prepared 30 years ago cannot be ignored. However, I assume that due to the abovementioned reasons using this list is still the most trustworthy reference point. Additionally, in the list of the most popular single lexemes from the GDP there was no word created or borrowed during last 30 years (of course this fact does not exclude the possibility of new meanings)[9]. Thus, the term "frequency" should be

[8] E.g. (de Schryver & Joffe, 2004), (de Schryver et al., 2006), (Verlinde &Binon, 2010), (Koplenig, Meyer, Müller-Spitzer, 2014), (Müller-Spitzer et al., 2015).

[9] This situation is probably the result of the fact that the most recent vocabulary items are currently being added to the GDP, during the second stage of the project (started in September 2013). At the moment of collecting the analyzed data the entries describing the most frequently used words of the Polish language were available for users (Żmigrodzki, 2014: 39). Therefore, the entries being the newest vocabulary are still being prepared and cannot be fully represented in queries (their popularity can be checked later, after completing the current stage of the project).

abandoned and replaced with "subjective probability".

Comparing the list of the most popular single lexemes in the GDP and the list prepared by Imiołczyk confirms this intuitive assumption: units which are absent from his list dominate in the GDP look-ups. They amount to 14% of all analyzed entries (see Figure 1 – Single Lexemes: Low Subjective Probability), whereas lexemes included in the Imiołczyk list form only 5% of all analyzed entries (see Figure 1 – Single Lexemes: High Subjective Probability). The lexeme which has the highest[10] subjective probability and is present on the list of the most popular the GDP entries is *house/home* [*dom*], with rank 22. Other words from the first thousand entries of Imiołczyk's list include a perfective form of *to slice* [*ukroić*], *patience* [*cierpliwość*], *love* [*miłość*], *problem* [*problem*], *youth* [*młodzież*]. The last 21 words hold different places in Imiołczyk's list (from 1256 to 4808). The words which were not included in the list, which is equal to having low subjective probability, are, to name but a few: *abortively* [*aborcyjnie*], *stocky* [*krępy*], *gully* [*źleb*], *optimal* [*optymalny*], *liberalization* [*liberalizacja*], *absorption* [*absorbcja*], *submission* [*uległość*], *empirical* [*empiryczny*], *to whisper* [*szeptać*].

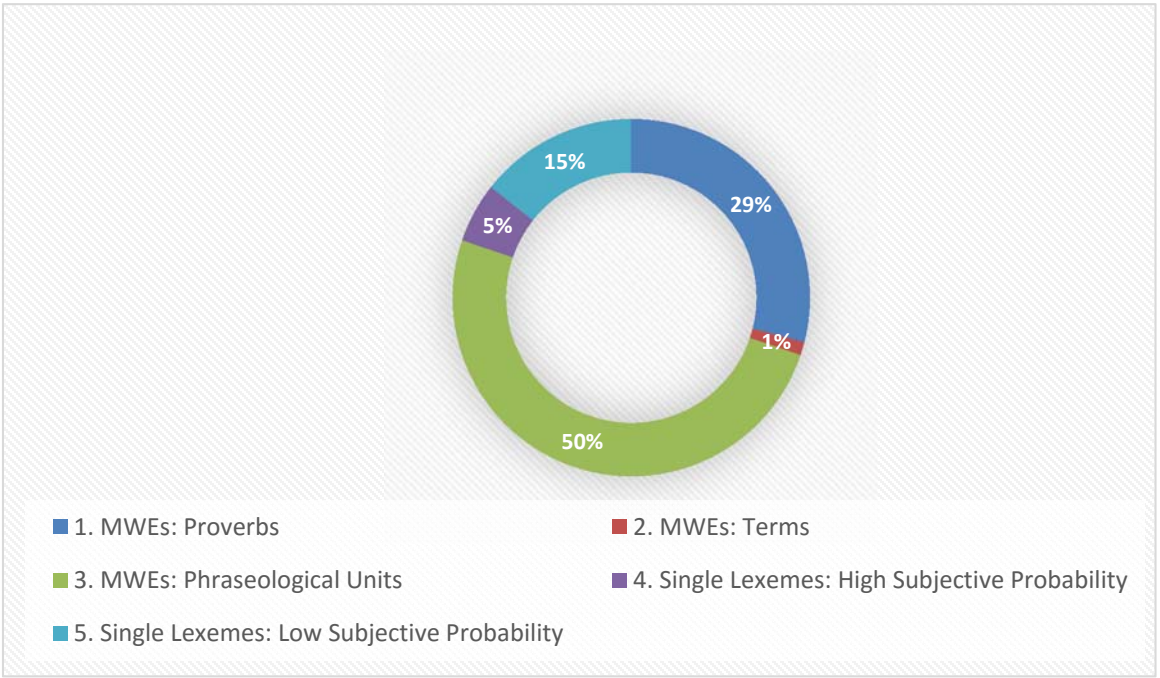

Figure 1: Popularity of Entries

2.2 Remarkable groups of entries and their strength

The attention of GDP users can be measured not only in the number of units representing remarkable groups, but also in their strength (represented by the number

[10] The lower the rank, the higher the subjective probability.

of page views, Figure 2). Its presence on the list of the 500 most popular entries is a distinctive factor. However, it is also important how many times the single entry was viewed.

The analysis of strength of entries (Figure 2) leads to conclusions similar to those drawn regarding their popularity (Figure 1) with reference to the single lexemes and being slightly different in the case of MWEs. The former group differs from the rate of popularity only in 1% with reference to the group of single lexemes with high subjective probability. The latter diverged from popularity in about 10% in both the most numerous subgroups – proverbs and phraseological units. As a matter of fact, the measurement of strength supports the thesis regarding GDP user interests in proverbs. When considering the number of page views, proverbs and phraseological units are almost equal and both constitute groups of entries drawing the most attention from users, despite the fact that the group of phraseological units consists of 250 units, whereas proverbs amount to 145 units.

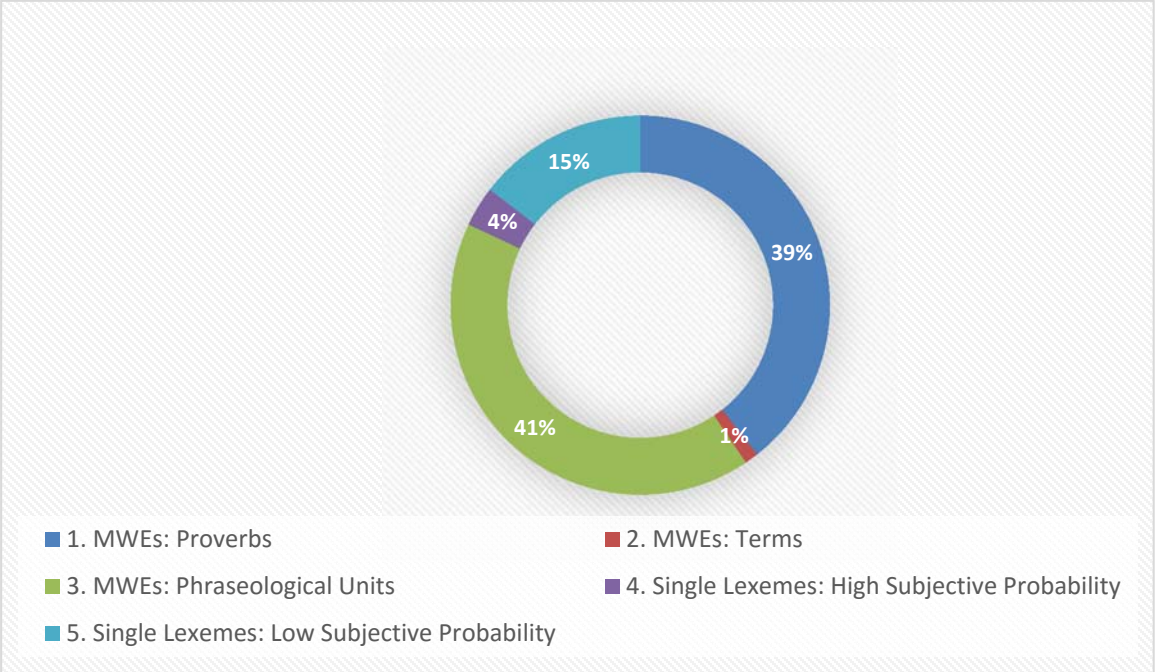

Figure 2: Strength of Entries

3. Findings

The gathered data reveal the following:

- GDP users are mostly interested in multi-word expressions, which constituted 401 of the 500 most popular entries, meaning that single lexemes cover only about 19.8% of the most popular entries.

- Single lexemes are represented in queries mostly by those with a low subjective

probability rate (15% of all most popular entries, 75% of the single lexemes), whereas single lexemes with a high subjective probability rate amount to only 5% of the most popular entries (25% of single lexemes). The popularity (percentage of the group in the rank of 500 most popular entries) and strength (measured in terms of the number of page views) for both subgroups are almost equal (Figure 1 and 2).

- In the MWEs three subgroups can be distinguished: proverbs, phraseological units and terms. Their popularity and strength is usually not equal (Figures 1 and 2) (except for terms). The popularity rank shows that phraseological units account for 50%, whereas proverbs account for 29% of all most popular entries (Figure 1). Considering the number of page views (strength, Figure 2) leads to the conclusion that the two subgroups are almost equal.

- All three subgroups of MWEs are varied in their origins, meanings, forms and stylistic features. No patterns in user needs can be indicated here.

4. Discussion

Knowledge of proverbs in the Polish language is decreasing, according to some authors, for the last 30 years (Buttler, 1989; Szpila, 2000). This observation is supported by empirical research on informants to establish the paremiological minimum of the Polish language. The research conducted in 1998 showed that the minimum consisted of 72 proverbs, whereas the survey from 2013 (using the same method and including as minimum only the units which were indicated by at least 8% of informants) identified only 39 proverbs (Szpila, 2014: 91–93). At the same time, GDP users are mostly interested in MWEs, particularly in proverbs. This surprising fact requires an additional comment.

One possible explanation for the interest in proverbs is school needs. This statement has been raised a few times during discussions among members of the GDP project. What would be the effect of confronting this assumption with school reality? The easiest way to check this is via school textbooks and other widely available sources. The term *proverb* is mentioned only once in the official document, which is currently in force and constitutes the basis of the syllabuses and textbooks of Polish schools (with regards to the subject "Polish language", devoted both to Polish language and literature). The document recommends that pupils from primary school years 4-6 should be able to recognize proverbs as well as stories, legends, novels and so on.[11] A little more attention is given to phraseological units. Pupils from junior high schools should be able to use phraseological dictionaries, understand phraseological units and use them. However, the exercises referring to phraseological units and proverbs often appear in textbooks for Polish language in primary schools and junior high schools.

[11] The document is called *the programme basis* and it is announced by the Minister of Education. The current one has been in force since December 2008.

To check how often Polish pupils face MWEs, four Polish language textbooks were analysed; three from primary schools (in accordance to the previously mentioned document this is the only stage of education which pays attention to proverbs): one chosen at random for each year from the second level of education (i.e. years 4, 5 and 6); and one from junior high school chosen at random from year 2. The scope of the analysis covered only textbooks (without workbooks or any other additional sources) and only those exercises in which pupils were obliged to work with MWEs. The explanations, definitions and texts regarding MWEs were not considered since it was assumed that pupils were encouraged to use dictionaries (e.g. the GDP) only when performing the task. The analysis showed that in the first part of the year 4 textbook there are seven exercises related to MWEs (Michałkiewicz & Mucha, 2011). In year 5 there are 10 exercises (Horwath & Żegleń, 2013), in year 6 19 exercises (Dobrowolska & Dobrowolska, 2014) and in year 2 of junior high school there are 15 exercises (Horwath & Kiełb, 2016). In each textbook, the tasks were mainly related to phraseological units (proverbs were in minority). The exercises comprise tasks such as: explain the meaning of MWE, check the meaning of MWE, create a sentence with MWE, find in a dictionary examples of MWEs containing a particular word and so on. The popularity of the topic is visible not only in textbooks but also in online educational webpages, e.g. *It is a mixed blessing* [*Każdy kij ma dwa końce*] present in www.sciaga.pl (in the part prepared by the website authors), www.zaliczaj.pl, www.zapytaj.onet.pl (as user questions).

On the other hand, exercises in which pupils were obliged to deal with single words were rarer. This fact partially confirms the assumption about the impact of school needs on GDP queries. Table 1 presents the exact MWEs and single words used in exercises for one of the analyzed school years, the fourth year of primary school. The table provides information on the presence or absence of the given MWEs and single words on the list of the 500 most popular entries of the GDP. It should be stated that the number of exercises (mentioned in the last paragraph) is not equal to the number of MWEs and single words. This is because a lot of exercises concern more than one lexical unit. Pupils are also obliged to find some MWEs and single words on their own (instructions such as: find the examples of MWEs containing the given word, give the examples of MWEs linked to the given topic, and so on).

Table 1 shows that in the fourth year of primary school MWEs were more numerous in the exercises than single words (29 MWEs vs 18 single words). Most are not present on the list of the 500 most popular entries from the GDP. The situation was similar in other analyzed textbooks – most lexical units in the exercises requiring meaning checks, finding synonyms or antonyms, or using the units in sentences were MWEs and not single words. Few of these lexical units were present on the 500 most common list.

MWEs	The presence of the MWE on the 500 most popular list	Single words	The presence of the single word on the 500 most popular list
to have a sour look on one's face [*ma skwaszoną minę*]	no	*popular* [*popularny*]	No
to put on a brave face [*nadrabia miną*]	no	*famous* [*sławny*]	No
his face fell [*zrzedła jej mina*]	yes	*scallywag* [*ziółko*]	No
looks askance at sb [*patrzy krzywym okiem*]	no	*fairytale* [*baśniowy*]	No
looks at sb piercingly [*przeszywa kogoś wzrokiem*]	no	vocabulary connected with theatre (chosen by pupils from the given text)	
looks on with a fixed stare [*postawiła oczy w słup*]	no	*tradition* [*tradycja*]	No
truth hurts [*prawda w oczy kole*]	no	*scholar* [*uczony*]	No
'very distant relative' [*dziesiąta woda po kisielu*]	no	*doctor* [*doktor*]	No
'a complete stranger' [*ani brat, ani swat*]	no	*associate professor* [*docent*]	No
as alike as two peas in a pod [*kubek w kubek podobny do*]	no	*house* [*dom*]	Yes
as alike as two peas in a pod [*kropka w kropkę podobny do*]	no	*cottage* [*chałupa*]	No

as alike as two peas in a pod [*podobni jak dwie krople wody*]	no	*small hut* [*chatka*]	No
'the spitting image of one's father/mother' [*wykapany tata, wykapana mama*]	no	*flat* [*mieszkanie*]	no
'talk man to man' [*porozmawiać z kimś po męsku*]	no	*apartment* [*apartament*]	No
'make a quick and firm decision' [*podjąć męską decyzję*]	no	*ruin* [*rudera*]	No
'severe rules' [*ojcowska ręka*]	no	*tenement* [*kamienica*]	No
'done in a way a woman would do' [*znać w czymś kobiecą rękę*]	no	*villa* [*willa*]	No
'woman's intuition' [*mieć kobiecą intuicję*]	no	*residence* [*rezydencja*]	No
'motherly heart' [*matczyne serce*]	no		
radiant smile [*promienny uśmiech*]	no		
glimmer of hope [*promyk nadziei*]	no		
glimmer of joy [*promyk radości*]	no		
glimmer of happiness [*promyk szczęścia*]	no		
feel at home [*czuć się jak u siebie w domu*]	no		
host [*pan/pani domu*]	no		

do the honours [*czynić honory domu*]	no		
'establish a family' [*założyć dom*]	no		
a friend of the family [*przyjaciel domu*]	no		
live out of a suitcase [*życie na walizkach*]	no		

Table 1: The MWEs and Single Words Used in Exercises from a Chosen Textbook for Polish Language for 4th Year Primary School Children

The conclusions of the given analysis are ambiguous. On one hand, the MWEs undoubtedly constitute an important part of school practice. On the other hand, it is clear that most MWEs (as well as single words) found in the textbook exercises were not present on the list of the 500 most popular entries in the GDP. Additionally, other relevant factors can be indicated here. One has already been mentioned: a lot of exercises oblige pupils to find MWEs not mentioned in the exercises. This fact excludes the possibility of preparing the list of MWEs (or single words) taught at school and checking their popularity in the GDP. Unfortunately, it is also impossible to combine the school activities related to the GDP queries with time periods. For example, at the moment of preparing the article there are five textbooks series which can be used for the Polish language subject in schools: in 2012 there were 10 (for years 4–6 and for junior high school)[12]. Additionally, teachers are not obliged to work through all textbook chapters nor to complete all exercises, but instead might set different exercises. Therefore, although this method would likely provide the most convincing evidence of the relation between the growing interest in MWEs and school needs, it is not a feasible analysis.

To sum up, it can be stated that pupils' needs are at least partially responsible for a big popularity of MWEs, especially proverbs. However, it is not the only reason. It is evident that some of the aforementioned examples of entries are not part of the school teaching program (e.g. colloquialisms). In seeking other reasons for the phenomena, the scope of other online dictionaries should be considered. It seems probable that users cannot find answers to their questions elsewhere and therefore turn to the GDP which results in the overrepresentation of the MWE queries.

[12] According to the official website of the Ministry of National Education related to textbooks (www.podreczniki.men.gov.pl).

When looking for sources like the GDP, the website www.sjp.pwn.pl should be considered. This is the source shared by one of the biggest Polish publishing houses, PWN. Under this address, one search engine enables the look-up of words and expressions in two general dictionaries, a spelling dictionary, a corpus and the answers given to questions which have been sent in by users over the past few years. Although this resource is vast, the overwhelming majority of the MWEs which were popular in the GDP cannot be found in dictionaries (some however appear in the user questions). Only 10 of 250 phraseological units which were most popular in the GDP are present in dictionaries provided by PWN publishing house, e.g.: *Aesopian language* [*język ezopowy*], *Balzakian age* [*wiek balzakowski*], *sb leads the way* [*ktoś wiedzie prym*]. Additionally, some are a part of the spelling dictionary, which means that the only available information is regarding their spelling.

5. Concluding Remarks

It has been shown that research on dictionary user behaviour should concern their typology. If not, results will over-represent the needs of the groups which use dictionaries more often than others (Arhar Holdt et al., 2016). The current study on GDP users does not overcome this obstacle; however, even when assuming that the gathered data are not fully representative, the study clearly shows that users are very interested in MWEs. This statement sheds new light on the previous analysis focused mainly on single lexemes.

Generally, the most important answer to the question regarding popular entries in the general monolingual dictionary (on the basis of the GDP) is that users look for MWEs, especially phraseological units and proverbs, and for single lexemes which are not well-known to them (i.e. having low subjective probability). Of course, this statement is not an absolute truth. When considering candidates for inclusion in the dictionary, one should think about additional circumstances which may influence user behaviour. The study demonstrates that this may be school needs or the content of the other dictionaries.

6. Acknowledgements

This scientific work was financed under the programme of the Ministry of Science and Higher Education entitled "National Programme for the Development of the Humanities" in the years 2013-2018, Project No.: 0016/NPRH2/H11/81/2013.

Praca naukowa finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Narodowy Program Rozwoju Humanistyki" w latach 2013-2018, nr projektu: 0016/NPRH2/H11/81/2013.

7. References

Arhar Holdt, A., Kosem, I. & Gantar, P. (2016). Dictionary User Typology: The Slovenian Case. In Margalitadze, T. & Meladze, G. (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 179–187.

Buttler, D. (1989). Dlaczego zanikają przysłowia w dwudziestowiecznej polszczyźnie?. Poradnik Językowy (5), pp. 332–337.

De Schryver, G.-M. & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 187–196.

Dobrowolska, H. & Dobrowolska, U. (2014). *Jutro pójdę w świat*, Warszawa: WSiP.

Ďurčo, P. (2015). Empirical Research and Paremiological Minimum. In H. Hrisztova-Gotthardt & M. A. Varga (eds.) *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies.* Warsaw/Berlin: De Gruyter Open Ltd., pp.183–205.

Horwath, E. & Kiełb, G. (2016). *Bliżej słowa (podręcznik do gimnazjum, kl. II)*, Warszawa: WSiP.

Horwath, E. & Żegleń, A. (2013). *Słowa z uśmiechem. Literatura i kultura.* Warszawa: WSiP.

Imiołczyk, J. (1987). *Prawdopodobieństwo subiektywne wyrazów. Podstawowy słownik frekwencyjny języka polskiego.* Warszawa: Państwowe Wydawnictwo Naukowe.

Kernerman, L. (1996). English Learners' Dictionaries: How Much do we Know about their Use? In Margalitadze, T. & Meladze, G. (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 405–411.

Koplenig, A., Meyer, P. & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In: C. Müller-Spitzer (eds.) *Using Online Dictionaries.* Mannheim: De Gruyter Mouton, pp. 229-250.

Krzyżanowski, J. (eds.), (1969). *Nowa księga przysłów i wyrażeń przysłowiowych polskich.* Warszawa: Państwowy Instytut Wydawniczy.

Krzyżanowski, J. (eds.), (1972). *Nowa księga przysłów i wyrażeń przysłowiowych polskich.* Warszawa: Państwowy Instytut Wydawniczy.

Lew, R. (2011). Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography*, 24 (1), pp. 1–4.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic Lexicography.* Oxford: Oxford University Press, pp. 343-362.

Lew, R. (2015). Research into the Use of Online Dictionaries. *International Journal of Lexicography*, 28 (2), pp. 232–253.

Michałkiewicz, T. & Mucha, K. (2011). *O to chodzi!*, vol. 1. Warszawa: Wydawnictwo

Stentor.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2015). Online dictionary use: Key findings from an empirical research project. In *Electronic Lexicography*. Oxford: Oxford University Press, pp. 425–457.

Müller-Spitzer, C., Wolfer, S. & Koplenig, A. (2015). Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography*, 28 (1), pp. 1–26.

Pęzik, P. (2012). Język mówiony w NKJP. In A. Przepiórkowski & M. Bańko & R.L. Górski & B. Lewandowska-Tomaszczyk (eds.) *Narodowy Korpus Języka Polskiego*. Warszawa: PWN, pp. 37–49.

Schryver de, G.M. & Joffe, D. & Joffe, P. & Hillewaert S. (2006). Do Dictionary Users Really Look Up Frequent Words? – On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* (16), pp. 67–83.

Szpila, G. (2000). Skamielina czy żywy organizm – przysłowie w prasie polskiej. In G. Szpila (eds.) Język trzeciego tysiąclecia: zbiór referatów z konferencji Kraków, 2–4 marca 2000. Kraków: Krakowskie Towarzystwo Popularyzowania Wiedzy o Komunikacji Językowej "Tertium", pp. 215–224.

Szpila, G. (2014). Znajomość przysłów wśród polskich studentów: minimum paremiologiczne. *Literatura Ludowa*, 58 (4-5), pp. 87–101.

Töpel, A. (2014). Review of research into the use of electronic dictionaries. In: C. Müller-Spitzer (eds.) *Using Online Dictionaries*. Mannheim: De Gruyter Mouton, pp. 13-54.

Verlinde, S. & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden/Ljouwert: Fryske Akademy – Afûk, pp. 1144–1151.

Żmigrodzki, P. (2005). *Wprowadzenie do leksykografii polskiej*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Żmigrodzki, P. (2014). Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. *Slovenščina 2.0*, 2 (2), pp. 37-52.

Dictionaries & Websites:

Analytics Help. Accessed at: Analytics Help; https://support.google.com/analytics/answer/1257084#pageviews_vs_unique_views. (13 May 2017)

nkjp.pl. Accessed at: www.nkjp.pl. (24 May 2017)

podreczniki.men.gov.pl. Accessed at: https://podreczniki.men.gov.pl/dopuszczone_lista5.php?file=szko%C5%82a%20podstawowa%20(kl.%204-8) ; https://podreczniki.men.gov.pl/dopuszczone_lista3.php?file=szko%C5%82a%20podstawowa%20(kl.%20IV-VI) ; https://podreczniki.men.gov.pl/dopuszczone_lista3.php?file=gimnazjum. (4 July 2017)

sciaga.pl. Accessed at: www.sciaga.pl. (2 May 2017)
SGJP: *Słownik gramatyczny języka polskiego.* (2017). [Grammatical Dictionary of Polish Language], available at: www.sgjp.pl. (24 May 2017)
sjp.pwn.pl. Accessed at: www.sjp.pwn.pl. (14 May 2017)
wsjp.pl. Accessed at: www.wsjp.pl. (24 May 2017)
WSPP: *Wielki słownik poprawnej polszczyzny PWN.* (2010). Warszawa: PWN. [Normative Dictionary of Polish]
zaliczaj.pl. Accessed at: www.zaliczaj.pl. (2 May 2017)
zaliczaj.pl. Accessed at: www.zapytaj.onet.pl. (2 May 2017)