POLISH ACADEMY OF SCIENCES GREAT DICTIONARY OF POLISH [WIELKI SŁOWNIK JĘZYKA POLSKIEGO PAN][1]

Piotr ŻMIGRODZKI
Institute of Polish Language at the Polish Academy of Sciences, Kraków

Żmigrodzki, P. (2014): Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. Slovenščina 2.0, 2 (2): 37–52.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_04.pdf.

The paper describes a lexicographical project involving the development of the newest general dictionary of the Polish language: the Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. The project is coordinated by the Institute of Polish Language at the Polish Academy of Sciences and carried out in collaboration with linguists and lexicographers from several other Polish academic centres. The paper offers a brief description of the genesis of the project and the scope of information included in the dictionary, the organisation of work, the life of the dictionary on the Web as well as the plans for the future.

Keywords: Polish language, electronic lexicography, general dictionary of Polish, online dictionary

1 INTRODUCTION

The Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN] has been in the process of being created for 10 years. Researchers and lexicographers employed at the Polish Academy of Sciences, with the help of people from other seats of learning, are working hard to finish

[1] This is a scientific work financed in the framework of a Polish Ministry of Science and Higher Education programme called "National Programme of Development of Humanities" in the years 2013–2018 [Praca naukowa finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą Narodowy Program Rozwoju Humanistyki" w latach 2013–2018].

it. It is a general dictionary of the Polish language, which is being published and shared exclusively online under the address http://wsjp.pl/. The present paper will describe: the history of the project, basic features of the dictionary, general rules of its creation, and the division of labour, workflow, as well as our plans for the future.

2 THE HISTORY OF THE PROJECT

2.1 The Beginning

The project started in the summer of 2005, when the Committee on Linguistics of the Polish Academy of Sciences announced a call for tenders for the project of a new great dictionary of the contemporary Polish language. The aim was to replace the already outdated dictionary by Witold Doroszewski (Doroszewski, ed. 1958–1969) as the main lexicographical source. Among others, a team representing the Polish Academy of Sciences Institute of Polish Language, consisting of Piotr Żmigrodzki, Renata Przybylska and Bogusław Dunaj, took part in this call for tenders. They introduced an initial version of the idea for a new dictionary at the meeting of the Committee on Linguistics in October 2005. In January 2006, a lexicographical workshop was held at the Institute, in which more sophisticated preparatory work was conducted. In December 2006, during the meeting of the Committee on Linguistics of the Polish Academy of Sciences, an extended version of the project was presented and accepted by the Committee. Preparatory work continued throughout 2007. It was financed by the so-called statutory funds of the Institute until the grant from the Ministry of Science and Higher Education was finally acquired. The actual lexicographical work on the dictionary started in January 2008.

2.2 Stages of the dictionary creation

The first proper stage of the dictionary creation took place between January 2008 and December 2012. Its aim was a preparation of 15,000 entries describing the most frequently used words of the Polish language (the list was

compiled on the basis of Polish language corpora from that period, especially the corpus of Publishing House PWN and the Corpus of PAS Institute of Computer Science). In order to reach that goal, other tasks had to be undertaken, most importantly designing the dictionary microstructure, its transposition on the database, designing the database itself and the whole digital system of the dictionary; this was done by the IT specialists who were cooperating with the team. (A more detailed description of these tasks can be found in Żmigrodzki 2011.) Between January and August 2013 the project was again financed only by the statutory funds of the Polish Academy of Sciences and was limited to the current expenditure of keeping the dictionary online and the necessary corrections of the existing entries. The acquisition of an honorary auspice of the Senate of the Republic of Poland was a very important event, which raised the status of the project and became a significant argument that could have an impact on its future funding. The second stage of lexicographical work began in September 2013, and is financed by the Ministry of Science and Higher Education with the program called "The National Programme of Development of Humanities" (Narodowy Program Rozwoju Humanistyki). The current stage is planned for the period up to 2018 and has the following objectives:

1. Compilation of 35,000 entries, among these:

- lexemes that were already included in the dictionary (according to the rule of compilation) in meaning relations with the words which were compiled in 2007–2012 (these are mainly: synonyms, antonyms, superordinates, near-synonyms, and near-antonyms);
- formative derivatives from the words already described, especially verbs;
- the most recent vocabulary items, which have not yet been recorded in any general Polish language dictionary.

2. Enrichment of entries compiled previously, especially:

- extension of etymological information to all entries;

- as far as chronology is concerned, instead of listing the older dictionaries in which the word was described, the earliest date of the first usage of the word may be presented, on the basis of information in catalogues, documents from digital libraries, documents from historical dictionaries, etymological dictionaries, etc.;
- as far as adding new meanings is concerned, new meanings that have appeared since the creation of the first entries will be added.

The current stage is obviously not the last one. After its completion, the dictionary will reach 50,000 main headwords (not counting entries describing idioms and proverbs). Ultimately, the plan is to describe "nearly all the lexical units in the Polish language". It is therefore safe to say that the work will continue past 2018. The nature of this work, however, cannot be known as of yet.

3 GENERAL CHARACTERISTICS OF THE DICTIONARY

In some respects, the character of the Great Dictionary of Polish follows the Polish tradition, which is quite different than the western-European ones. At the same time, many of the ideas and solutions used in it can be perceived as innovative, at least when compared to other Polish dictionaries. The main aspects that should be highlighted are:

- in principle synchronic: although the year 1945 was accepted as the beginning of the time span covered, due to the nature of the sources (to which we return later on), the overwhelming majority of the material will belong to the last decades of the 20th and the beginning of the 21st century.
- in principle descriptive: the authors are not going to exclude from description any lexicographical facts deemed incorrect[2] or – for whatever

[2] This tendency dominated Polish lexicography of the 20th century, and because of that some contemporary readers may expect this.

reasons – unworthy of being noted in a dictionary, as long as these facts are well attested in the sources. The authors will only point out the normative unacceptability of a given fact, drawing on the Normative Dictionary of Polish [Słownik poprawnej polszczyzny] (for a more detailed overview, cf. Żmigrodzki 2008), and mark the stylistic qualification of non-standard units.

- an academic dictionary in which the authors aim to employ wherever possible the achievements of Polish 20th-century linguistics, especially in the field of semantic, inflectional and syntactic description of lexical units, at the same time keeping in mind that the description must be accessible to a very diverse group of Polish language users.

The format of description that is prevalent in the Great Dictionary of Polish can be defined as structuralist. It develops the theoretical ideas contained in the academic "Grammar of the Contemporary Polish Language" (Grzegorczykowa et al., ed. 1984) and the works of Polish scientists such as Maciej Grochowski and Andrzej Bogusławski (semantics), and Mirosław Bańko and Zygmunt Saloni (inflection and syntax). Some of the data such as the information on pronunciation, etymology, and chronology is introduced on the basis of lexicographical research of other Polish dictionaries. However, the way of presenting the lexicographical data is simplified, adjusted to the perceived abilities of a wider audience, not only linguists or philologists. This is because the dictionary is aimed to reach as many users as possible.

4 THE CORPUS BASIS FOR THE DICTIONARY

The main source of linguistic data for the dictionary is the National Corpus of Polish [Narodowy Korpus Języka Polskiego, NKJP], a collective undertaking of several academic units (including PAN IPL), carried out as a development project parallel to WSJP and available for free on the Internet (http://nkjp.pl). The second most important source inventory is an auxiliary corpus created at the PAN IPL specifically to serve the needs of the emerging dictionary; it

comprises texts which for various reasons were not (and are not going to be) included in the NKJP. Polish Internet sites constitute the third source. Finally, the authors of particular entries may also rely on citations they have found themselves. Although we are quite aware that this set of sources is not perfect and might be criticized especially by philologists and lexicographers representing more traditional approaches, we believe that a better corpus of sources for WSJP would not be feasible within the foreseeable time. The question of how best to export corpus data into the dictionary database is yet to be resolved. Until now, the editors were able to use the tools available in the corpus search engines, but they can only transfer the data with the help of the Windows Clipboard.

In the process of creating specific entries, the editors also use other lexicographical sources. One resource worth mentioning here is "The Grammatical Dictionary of Polish Language" [Słownik gramatyczny języka polskiego] (cf. Saloni et al. 2007). Its creators have approved sharing the inflectional paradigms pertaining to the entries of the Great Dictionary of Polish and the creation of a tool to import them directly to our dictionary's entries. Other dictionaries are sometimes also used, but to a lesser degree. For example when it comes to the description of syntax, "the Syntactic-Generative Dictionary of Polish Verbs" [Słownik syntaktyczno-generatywny czasowników polskich] (Polański, ed. 1980-1992). For the insertion of information about etymology and chronology, as mentioned before, etymological and historical dictionaries will be consulted.

5 THE METHOD OF PRESENTATION AND THE SCOPE OF INFORMATION IN THE DICTIONARY

The microstructure of the dictionary covers all elements which can be found in a typical general dictionary, in other words:

- headword form (with variants);
- information about the 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[3]);
- 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 (based on the NKJP);
- 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 is the set of information present in the description of the two most numerous types of language unit in the Great Dictionary of Polish, namely single lexemes and idiomatic expressions. The most frequently used abbreviations, acronyms and proper names are also included in the dictionary. Their information content is a little different than the aforementioned set.[4] The description of functional words (in Polish tradition this term is used to describe prepositions, conjunctions etc.) is more distinct. To fully describe this topic a separate article would be needed.

[3] A guideword is a type of indicator used in polysemous words which points to the meaning that the reader wants to find. The idea of a guideword was taken from English pedagogical dictionaries (e.g. LDOCE).

[4] More information on the contents of these entries can be found in Żmigrodzki 2011.

6 THE DIGITAL STRUCTURE OF THE DICTIONARY

When it comes to the digital structure of the dictionary, it has a form of a MySQL database. The integral part of it are also two modules: editing panel (see Figure 1), in the form of an online form, which the lexicographers fill in, and the so-called presentation panel (see Figures 2 and 3), which displays the contents of the entry on the end-user's computer. Since the entry displayed to the end-user is generated every time directly from the database, this enables all changes made by the editors to be immediately available to users. From the point of view of the end-user, there are two possible modes of viewing the entry article:

- the so-called bookmark view, in which the contents are structured and sorted, the user can switch between different tabs and access the specific segments of the entry article;
- the consolidated view, in which all the elements of the entry article are extended and set in a linear order (as in a traditional, paper dictionary). After choosing this view, it is also possible to print the entry article.

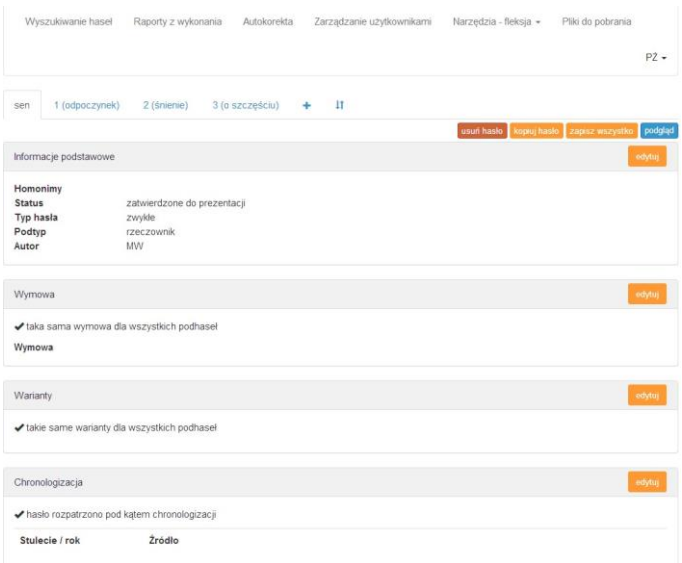

Figure 1: Editing Panel of the dictionary, general view of the entry *sen* 'sleep; dream'.

Figure 2: Presentation panel of the dictionary, entry *sen*, bookmark view (showing meaning 1).

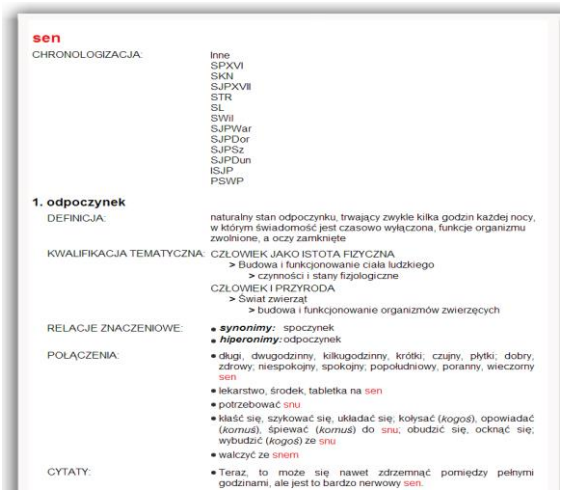

Figure 3: Presentation panel of the dictionary, entry *sen*, consolidated view.

The dictionary offers different ways of accessing entries, such as:

- an alphabetical list of entries, alphabetical (a fronte) and reverse dictionary (a tergo) orders are available;
- simple search: the user enters one or more words into the search bar and the program lists all the entries containing them;
- advanced search: various criteria can be used, such as any combination of signs, any inflectional form, a type of entry, a part of speech, thematic classification, etymology, labels (stylistic, typical user-group (e.g. criminal, teenage) and chronological features (e.g. archaic)).

There are also links to other entries if needed – for example to superordinates, synonyms, antonyms, aspectual oppositions – and it is possible to instantly access the related entry by clicking the displayed word. In the future, after the compilation of more entries and the implementation of complete inflectional data, there are plans to make it possible to reach entries by clicking on any of the words contained anywhere on the entry page.

Apart from the information included in the entries, on the dictionary page you can also access materials that were traditionally considered a part of the so-called *front matter*, namely: introduction, outline of the history of the project, a comprehensive description of the rules by which the dictionary has been created, information about the authors (with photos and biographical notes), and the list of articles concerning the project (with active links to the online version, if available).

7 PROJECT TEAM

The number of people involved in the process of creating the dictionary has changed in relation to the stage of the work and the financial conditions. In 2005, only 3 people worked on the original concept without salary. In 2006, when the dictionary workshop was set up, 10 people were working on the dictionary, but only two of them full-time, the others being involved in other

projects as well. Between the years 2007-2012 and currently, the number of people actively working on the dictionary is about 40. They are mostly lexicographers and linguists: scholars, PhD students; some of them began working on the project as MA students. Most of them (25 people) are affiliated with PAN Institute of Polish Language, but there are also employees of Jagiellonian University of Krakow, University of Warsaw, University of Silesia (Katowice), and Nicolaus Copernicus University (Toruń). 20 people are employed full-time, the rest have contracts for specific tasks. There is also a group of IT specialists (employed in computer companies). The administrative duties are performed by one person who is supported by the administrative personnel of the institute. It is worth mentioning that the total number of people involved in the process of creating the dictionary is over 70. The staff was changing very frequently during the first stages of the project, but now it has stabilized.

8 THE DIVISION OF WORK AND ITS ORGANIZATION

The tasks and assignments of the team are based on a certain hierarchy amongst the members of the team. The main categories of editors (and their competences) are as follows:

- Editors:
  - create their own entries and edit them;
  - can view the entries created by other editors but not modify them;
  - fill in all the entry fields except Etymology, Chronology, Thematic Class.
- Supervising editor:
  - reviews the entries created by editors, and makes changes as necessary.
- Supercoordinator (leader of the project):

  - proofreads the entries after they have been revised by editors;
  - controls the adequacy of information in fields fulfilled by specialists;
  - accepts the entries for presentation.

- Specialists:
  - fill in only the specified field (Etymology, Thematic Classification or Chronology); or
  - create and edit entries of a particular type (functional lexemes, idioms).

The creation of entries is carried out online, so it is possible to work on the dictionary from any place in the world, using a simple web browser.

The sequence of actions required for entry creation is as follows:

- Editor: creates the initial version of the entry, filling in all the information except the one reserved for specialists.
- Supervising Editor: reviews the entry created by the editor, adds comments and then, after introducing adjustments and discussing any controversial elements with the editor, forwards the entry for further improvements.
- Specialists dealing with etymology, thematic classification and chronology fill in the information connected to these areas.
- Supercoordinator (presently this job is only performed by the leader of the project and the chief editor of the dictionary, prof. Piotr Żmigrodzki) reviews the entry one more time, this time including the information added by the specialists, and suggests any necessary corrections and then checks if they were implemented properly. Then he green-lights the entry for publication, and gives it a proper status in the database.

In order for the entry to be accepted, automatic control of entry completeness takes place (all fields have to be filled), the correctness of the spelling and the

correctness of cross-references between the current entry and the others. Immediately after an entry is given a status of "accepted for presentation", the entry becomes visible in the presentation panel, so that it is available for the end-users. This kind of entry is then blocked from further editing.

It is worth noting that confirmation of entries is reversible, in other words it can be undone – for example if there is a need to introduce adjustments or additional information, such as adding a new meaning. The supercoordinator is the only person who is allowed to unblock the entry and forward it for further editing.

9 THE LIFE OF THE DICTIONARY

The first dictionary entries were published in 2009. At the time of writing (15th October 2014) 21,130 entries are available. As mentioned before, the number of entries will be systematically increased. In 2018, it is going to reach 50,000 according to the plans. Using the dictionary is completely free and does not require any additional software. In the last 12 months the dictionary has had more than a million views and over 356,000 users from more than 140 countries. Of course most of the users connect from Poland, but there is also a significant number of visitors from the USA, Great Britain, Germany, Russia, Ukraine and other countries of the former USSR. The Polish diaspora is occupying these territories in great numbers. The team of editors is trying to promote the dictionary on their Facebook fanpage, and organizes many public presentations for scholars, teachers, students and learners. The dictionary has also been promoted in various radio programmes in which the head of the project took part. Attracting the attention of mass media is quite difficult when it comes to linguistic issues in general and even more so in the case of the Great Dictionary of Polish. Maybe in the future when the dictionary expands it will be easier.

10 PLANS FOR THE FUTURE

The current, second stage of work on the dictionary is not of course the final one. The basic task of the team for the future is obviously adding more words and expressions. Apart from that, the development of the dictionary should include:

- enrichment of the existing entries: especially adding new meanings if they emerge; there are also plans to include information about pronunciation, in the form of transcription and voice files, possibly also multimedia files complementing the definitions;
- improvements of the technological aspects, when it comes to editing (i.e. permitting automatic import of the data from other lexicographical sources), as well as from the end-user's point of view (improvements of the presentation panel, making it compatible with the newest devices, maybe integration with other digital lexicographical sources).

There is also a need for other work pertaining to the contemporary Polish language, indirectly connected with the Great Dictionary of Polish, such as extending and bringing corpus sources and other materials on the Polish language up to date.

Independently of the circumstances, the dictionary team is determined to continue with their work, until the moment their original goal, which is describing almost all lexical units of Polish, is reached.

REFERENCES

Doroszewski, W., ed. (1958-1969): *Słownik języka polskiego PAN*, vol. 1-11. Warszawa: Państwowe Wydawnctwo Naukowe.

Grzegorczykowa, R., et al., ed. (1984): *Gramatyka współczesnego języka polskiego*. *Składnia*. *Morfologia*. Warszawa: Państwowe Wydawnctwo Naukowe.

Saloni, Z., Gruszczynski, W., Woliński, R. and Wołosz, R. (2007): Grammatical Dictionary of Polish. Presentation by the Authors. *Studies in Polish Linguistics*, 4, 5-26.

Żmigrodzki, P. (2008): Nowy *Wielki słownik języka polskiego* a problemy poprawności językowej. In M. Święcicka (ed.): *Siła słów i ludzi*: 88-100; Bydgoszcz: Wydawnictwo Uniwersytetu Kazimierza Wielkiego.

Żmigrodzki, P. (2011): Polish Academy of Sciences Great Dictionary of Polish: history, presence, prospects. *Studies in Polish Linguistics*, 6, 7-26.

VELIKI SLOVAR POLJSKEGA JEZIKA POLJSKE AKADEMIJE ZNANOSTI

Prispevek opisuje izdelavo najnovejšega splošnega slovarja poljskega jezika: Veliki slovar poljskega jezika Poljske akademije znanosti (Wielki słownik języka polskiego PAN). Vodilni partner projekta je Inštitut za poljski jezik, pri projektu pa sodelujejo tudi jezikoslovci in leksikografi s številnih poljskih akademskih ustanov. Prispevek na kratko predstavlja zgodovino projekta in glavne značilnosti slovarja. Opisana je tudi organizacija dela in naloge, ki jih opravljajo posamezni člani ekipe. Prispevek dalje prikazuje spletno zasnovo slovarja in dele spletnega vmesnika, npr. različne načine iskanj. V zaključku so navedeni načrti za prihodnost, ki vključujejo tako dodajanje novih iztočnic oziroma novih informacij obstoječim iztočnicam kot tudi izboljšavo leksikografskih postopkov z uporabo sodobnih jezikovnotehnoloških orodij.

Ključne besede: poljski jezik, elektronska leksikografija, splošni slovar poljskega jezika, spletni slovar