Polish Academy of Sciences [PAS] Great Dictionary of Polish – lexicographical workflow and summary of the project

Piotr Żmigrodzki
Instytut Języka Polskiego PAN
E-mail: piotr@ijp-pan.krakow.pl

1 Introduction

PAS Great Dictionary of Polish (pol. *Wielki słownik języka polskiego PAN*, WSJP) is an online scholarly dictionary of contemporary Polish, based mostly on National Corpus of Polish (Przepiórkowski et al., 2011). The main objective of the dictionary is to describe Polish vocabulary from 1945 till now. The dictionary is planned to be:

– 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 shall 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 eliminate from the descriptions any lexicographical facts deemed incorrect or – for whatever 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, basing on the *Great Normative Dictionary of Polish* [*Wielki słownik poprawnej polszczyzny*], Markowski, ed. 2003, and mark the stylistic qualification of substandard units.

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

The project started in 2006 and it was financed at first by the Institute of Polish PAS, from 2008 – from a grant of Polish Ministry of Science and Higer Education (the 1st stage – so-called developmental project, 2007-2012t; the 2nd, from 2013 to 2018 – by the *National Programme for Development of Humanistics*; pol. *Narodowy Program Rozwoju Humanistyki*). The 1st stage was completed in December 2012 and it provided a 15 thousand lexical units describing most frequent Polish words. The aim of the second stage is to expand the dictionary to about 50 000 lexical units (not only most frequent, but also some neologisms and derivatives) and also to enrich the lexicographical information in the "old" units. The final aim of the planned work is to describe almost all Polish lexemes and idioms.

2 Preparation and lexical process

The lexicographical process in WSJP is not a typical one, because the entries of our dictionary are published directly after the end of the work, so the number of entries increases systematically day by day. However, before the lexicographical work started, we had to make some plans and preparations.

2.1 Preparation

The very first conception of the dictionary was presented in 2005 on the plenary session of the PAS Committee of Linguistics. Institute of Polish Language had been systematically developing the initial concept since January 1st, 2006; financing the work of the team from the statutory funds. A detailed dictionary project framework was presented at the meeting of the PAN Committee on Linguistics in December 2006. (see Zmigrodzki et al., 2007). The presentation met with a favorable reception, which was reflected in the Committee's resolution. In order to secure additional funding sources, a grant application was submitted to the

Ministry of Science and Higher Education.

The first stage (2006-2007) was the elaboration of the dictionary structure and the problems of the computerization of it. The computer system of the dictionary was designed in 2007 and early 2008 about one hundred sample entries were written in order to test the computer system and to verify some theoretical solutions. Between July, 2008 and December 2011 we prepared 15 000 entries and the last year of the first project was devoted to proof reading and an automatic control of the dictionary. The second stage started in August 2013, we spent the first six months of it on verifying the lexicographical instruction, testing the new (refreshed edition panel) and working with new lexicographers employed. The real lexicographic work started in April 2014.

2.2 Data acquisition

The main (and in the practical sense, the only) source 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 source inventory, used mostly during the first stage of the project, is an auxiliary corpus created at the PAN IPL specifically to serve the needs of the emerging dictionary; it comprises texts which were for various reasons not included in the NKJP. Polish Internet sites constitute the third source. Finally, the authors of particular entries may rely on their own excerption. The main problem is the lack of indirect interface between corpora and the dictionary. The problem is to be solved in the future, but now the only solution is to transfer citations and collocations from NKJP to WSJP via Windows Clipboard. This is because both projects were developed simultaneously and there was no time for thinking on technical methods of transferring data between them.

As concerns secondary sources, lexicographic ones, at the very beginning of the project we decided not to make use either of previous dictionaries of Polish. In our opinion, they were not much reliable and contained many words that are actually not frequently used in contemporary Polish – or not used at all.

However there is one source database, which is imported to our lexical entries automatically, i.e. inflectional data. The inflectional data (paradigms) come from the Grammatical Dictionary of Polish, *(Słownik gramatyczny języka polskiego)* (Saloni et al., 2007) via the web interface called *Kuźnia* (pol. for *forge*), see Woliński et al., 2012. When the lexicographer creates a lexical entry with a specified lemma, the inflectional data are automatically imported to the database of WSJP or he/she is able to import them by clicking a special link in the edition panel.

2.3 Computerisation

As I wrote in the previous paragraphs, the National Corpus of Polish was a separate project led in other PAS institute (namely Institute of Computer Science), so we had less influence on the shape of grammatical description, lemmatization etc. in it. Computer system of our dictionary was projected by Mr. Mateusz Żółtak, former IT specialist from our Institute, (the earlier version of it was implemented in the Electronic Dictionary of 17th and 18th century Polish, see Gruszczyński 2005). From the technical point of view, the dictionary consists of three parts (levels):

- a relational database (MySQL) on a computer server;
- an edition panel (interface), by means of which the editors enter lexicographical data in the database, filling in respective forms reflecting the microstructure of specific types of entries;
- a presentation panel, by means of which the completed dictionary entries are presented to the user.

The main advantage of this solution is that the dictionary entries can be edited without any specialist software; the edition panel is an electronic form, which (after logging in) can be opened in any web browser. Since the text corpus is also accessible online, the dictionary can be developed anywhere and anytime, the only technical requirement being a computer with an Internet connection. All documents, such as editorial manual and guidelines, are uploaded onto a special protected website, so that practically all information is exchanged between the dictionary authors via the Internet. This solution has proven extremely useful in the light of the geographical dispersion of the co-workers and the workspace limitations of the PAN IPL.

The first version of the computer lexicographic system was launched in early 2008, but it is being (especially edition panel and presentation panel) constantly developed in order to make both the lexicographic work and the dictionary use more comfortable.

2.4 The team and workflow in WSJP

The WSJP team now consists of about 40 persons: editors, supervising editors, specialists, IT specialist. About half of them are full-time employed in the Institute of Polish Language at the Polish Academy of Sciences in Krakow and Warsaw. Some other co-operators work at other Polish academic centers: Jagiellonian University in Krakow, Warsaw University, University of Silesia in Katowice, and Nicolas Copernicus University in Toruń.

Editor is a person who creates his/her own entries and edits them, he/she can also view the entries created by other editors but cannot modify them.
Supervising editor is the first proofreader of the entries created by editors. The supervising editor can view all entries but can only modify the entries created by editors who were assigned to supervision by the system.
Specialist - this person fills in only one specified field, but in all dictionary entries. At present, only four fields are under the charge of specialists: etymology, thematic classification, chronology and semantic relations.

The workflow in WSJP looks now as follows.
1. The editor creates a lexical entry and edits it, filling in all fields except those reserved for specialists[1], e.g.:

- subentries devoted to particular meaning of the word in question
- definitions
- semantic identifiers (when necessary)
- stylistic or field labels
- collocations
- quotations

2. The supervising editor checks if the guidelines were followed properly and if the description is adequate; all remarks are entered in a special field.
3. The editor modifies the entry, taking into account the supervising editor's remarks.
4. When the entry is accepted by the supervising editor, the specialists start adding etymological, chronological information, thematic classification and semantic relations (synonyms, antonyms, hyponyms).
5. The leader of the project uses computer system of the dictionary to find the entries which are complete (it means that all the „specialist" fields are edited) and accepted by supervisors, reads each entry again and if he estimates that it is ready for publishing, he accepts it for presentation. Before the status „ready for presentation" can be set, each entry is automatically spell-checked. After setting the status „ready for presentation", the entry becomes visible for external users of the dictionary in the presentation panel.

The entry view in the presentation panel is generated in response to the user's query from the dictionary database in its current state. In this way, every change made to an existing entry by its editor is almost immediately visible in the end form of the dictionary.

[1] Detailed information, see Żmigrodzki 2011

The leader of the project is able to block the presentation of any entry, for example when he finds a mistake in it or when he wants to make some corrections, add new subentries etc.

2.5. **Perspectives**

I hope the dictionary will be developed after 2018, but at the moment there is no time to make more detailed plans. However, I'm sure that the dictionary will stay online on the server of PAS Institute of Polish Language and the Institute will be able to cover all costs of technical updates etc.

3 Time span of the different phases

Phase	**Duration**	
	WSJP stage 1	WSJP stage 2
Preparation	2005-2007	2013
Data acquisition	2007-2010	2013
Elaborating of the computer lexicographic system	2007-2008	2013-2014
Preparing of Dictionary entries	2008-2011	2014-2017
Final proofreading of the dictionary (partly automatically)	2012	2018

Table 1 Process phases of dictionary project and their time span

4 References

Gruszczyński, W. (2005): O przyszłości słownika języka polskiego XVII i 1. połowy XVIII wieku. Poradnik Językowy, (07), 48-61.

Przepiórkowski et al., 2011: Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński i Piotr Pęzik. National Corpus of Polish. W: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, s. 259–263, Poznań,

Markowski A., ed., 2003: Wielki słownik poprawnej polszczyzny PWN, Warszawa.

Wolinski, M., Milkowski, M., Ogrodniczuk, M., & Przepiórkowski, A. (2012): PoliMorf: a (not so) new open morphological dictionary for Polish. In LREC (pp. 860-864).

Saloni, Z., Gruszczyński, W., Woliński, M., & Wołosz, R. (2007): Grammatical Dictionary of Polish. Studies in Polish Linguistics, 4, 5-25.

Żmigrodzki et al., 2007: Żmigrodzki, P.., Bańko, M., Dunaj, B., & Przybylska, R.. (2007): Koncepcja Wielkiego słownika języka polskiego–przybliżenie drugie,[in:] R. Przybylska, P. Żmigrodzki (Hg.), Nowe studia leksykograficzne, Kraków, 9-21.

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