Anna Dąbrowska / Walery Pisarek

National Corpus of Polish and *Great Dictionary of Polish.* Two leading projects of present-day Polish lexicography

Streszczenie

Tekst poświęcony jest krótkiemu przedstawieniu narodowego korpusu języka polskiego (NKJP) i powstającemu w oparciu o NKJP wielkiemu słownikowi języka polskiego. W pierwszej części wymienione są instytucje współtworzące korpus i narzędzia stworzone do jego obsługi. Omówiona została struktura korpusu, jego zrównoważoność i reprezentatywność, anotacja składniowa, informacja morfo-syntaktyczna oraz anotacja semantyczna. Dodatkowo wspomniane są – wynikające z informacji podawanych w NKJP – słowa dnia. Krótko wymienia się też inne możliwości wykorzystania NKJP.

Część druga artykułu poświęcona jest *Wielkiemu słownikowi języka polskiego* (WSJP) – jednemu z wielkich projektów dotyczących języka polskiego na początku XXI wieku. Nowością w leksykografii polskiej jest to, że powstaje on jedynie w wersji elektronicznej. Konsekwencją tego jest zakładane jego stałe uzupełnianie i modyfikowanie, praktycznie bez końca. WSJP ma w założeniu charakter współczesny i deskryptywny. W tekście wymienia się 19 elementów tworzących hasło słownikowe (m.in. czas pojawienia się danego wyrazu, jego pochodzenie, definicję semantyczną, kwalifikatory, składnię, kolokacje, cytaty, informacje normatywne).

1. National Corpus of Polish

National corpora of various languages have been built since the 1960s (British, American, Chinese Mandarin, Dutch, Czech, Croatian, Russian, Slovak, Slovenian etc.). In Poland, several groups of linguists, computer scientists and lexicographers have been working on corpora of the Polish language since the 1990s. They work in the following institutions:

– IPI PAN: *Institute of Computer Science, Polish Academy of Sciences* (http://nlp.ipipan.waw.pl);
– IJP PAN: *Kraków Institute of Polish Language Polish Academy of Sciences* (http://ijp-pan.krakow.pl);
– PELCRA: *Polish and English Language Corpora for Research and Application, University of Łódź* (http://pelcra.pl) (since 1995);
– *Polish Scientific Publishers* PWN.

Between the years 2001-2004 they created a 250 million word corpus. The researchers involved in the creating of the corpus participate in some European corpus linguistics projects: CLARIN (*Common Language Resources and Technology Infrastructure*, www.clarin.eu), CESAR (*Central and South-East European Resources*, www.meta-net.eu/projects/cesar/) and ATLAS.

After a few years of working separately, the Consortium *National Corpus of Polish* was set up in 2006. Its participant institutions were those mentioned above and their representatives are:

– Adam Przepiórkowski – Institute of Computer Science, Polish Academy of Sciences;
– Rafał Górski – Institute of Polish Language, Polish Academy of Sciences;

– Barbara Lewandowska-Tomaszczyk – University of Łódź;
– Marek Łaziński – University of Warsaw;
– Mirosław Bańko – Polish Scientific Publishers PWN.

Since 2011 NKJP with its largest corpus of Polish has been available free of charge (nkjp.pl) and has become a basis for scientific research and among other things, for the *Great Dictionary of Polish* being currently worked on by a team led by Piotr Żmigrodzki (www.wsjp.pl).

1.1 The usefulness of NKJP in research

1.1.1 Lexicographic and lexicological works

Lexicographers were the first to notice the substantial usefulness of corpus data, allowing them to go far beyond the somewhat inconvenient collections of paper index cards. However, not all linguists were convinced of the usefulness of these corpora – the strong influence of generativism (psychological, introspective, anti-empirical) restrained the use of corpora of various languages built in the 1990s. Lexicographers roused themselves first. The first dictionary in the world based on corpus data was published in 1987: *The Collins Cobuild English Language Dictionary*. It was also shown that the corpus was very useful for the creation of so-called *learner's dictionaries*. Moreover, due to corpora it is possible to use so-called descriptive normativism (based on tradition and corpus data). In Poland, the first dictionary using corpus data to a large extent was *Inny słownik języka polskiego* (*Another dictionary of Polish*) by M. Bańko (2000). *Słownik dobrego stylu* (*The Dictionary of good style*) (2006), i.e. a dictionary of collocations is fully based on the corpus. Currently the *Great Dictionary of Polish* is being worked on as well as other smaller lexicographic works are being created.

1.1.2 Linguistic research

Corpus related linguistics can be characterized by three approaches in research: *corpus illustrated* (corpus – informer)*, corpus based* (verification of an earlier formulated hypothesis) and *corpus driver* (taxonomic construction of grammar on the basis of empirical data from the corpus). In corpus related linguistics, looking for tendencies is more frequent that looking for rules.

The corpus brings, first of all, statistical data. It additionally examines mutual relations between grammar and vocabulary, pragmatics, stylistics etc. It also examines competing grammatical forms (e.g. *pisarzy – pisarzów*[1]). Corpus data can also be used for the purpose of examining collocations (see: *Dictionary of good* style by M. Bańko) or the stylistic diversity of a language.

1.1.3 Corpus in translations

Translators may verify the forms they choose in a corpus. This is necessary both in the translator's work and in didactics when future translators are trained.

It is common nowadays to use Internet browsers in translations and to compare language equivalents in multilingual Wikipedia entries.

[1] 'Writers'; two forms differing from each other by an ending.

1.2 Profile of NKJP

NKJP can be described as representative and balanced. It is assumed that a representative corpus should reflect the language "in its entire variety" (NKJP 2012, 29), thus each type of a text should have its (at least small) contribution to the corpus. Sometimes decisions of this type must be arbitrary (e.g. 7% of Internet texts). Representativeness refers to the reality beyond the corpus. In the subcorpus of spoken texts representativeness refers to the authors of these texts (speakers), whereas the component of writers reflects reception of written Polish (structure of readership in Polish).

Balance means that none of text types covers more than half of the corpus (NKJP 2011, 30).

Table 1 presents the model structure of the corpus.

Type of the text	% contribution to the corpus
Political commentaries and short press releases	50.0%
Belles-lettres	16.0%
Non-fiction	5.5%
Informative and handbook type	5.5%
Scientific and didactic type	2.0%
Other written texts	3.0%
Books other than fiction non-classified	1.0%
Transcripts of conversations, media-based spoken and quasi-spoken – together	10.0%
Internet texts – static and dynamic together	7.0%

Table 1: Model structure of the corpus (source: NKJP 2012, 33)

The chronological selection of texts prefers those created after 1945; an exception was made in relation to literary texts – some of these originate from the beginning of the 20th century.

80% of texts appeared after 1990, 15% were produced in a period 1945-1990, and only 5% appeared before 1945. They are overwhelmingly contemporary texts.

1.2.1 Structure of the corpus

Currently NKJP numbers 1.5 billion words and is not stylistically balanced; however, it has been tagged and made grammatically unambiguous. There is a so-called 1M – one million, manually marked-up (tagged) subcorpus, which remains a very important part of the corpus. It is a basis for programmes servicing the whole corpus. It contains only texts coming from the period after 1945.

Additionally the following can be distinguished:

100M demo – comprising one hundred million words,
300M – balanced corpus covering three hundred million segments (250 million words).

Table 2 presents categories of texts as parts of 1M (proportions shown guarantee its balance).

Category	% contribution to the corpus
Daily newspapers Other periodicals and journals Political commentary books	25.5% 23.5% 1.0%
Belles-lettres	16.0%
Non-fiction	5.5%
Informative and handbook type	5.5%
Scientific and didactic type	2.0%
Internet-related interactive (blogs, forums, Usenet) Internet-related non-interactive (static websites, Wikipedia)	3.5% 3.5%
Quasi-spoken (minutes of Parliament sessions) Media-related spoken Conversational spoken	2.5% 2.5% 5.0%
Other written texts	3.0%
Books other than fiction non-classified	1.0%

Table 2: Categories of texts (source: NKJP 2012, 53)

1.2.2 Mark-up levels for the corpus

a) Segmentation of the text into sentences and words

The inflection classes distinguished in NKJP are slightly different than traditional ones (new classes appeared such as e.g. *burkinostkas, kubliks* or *foreign objects*). Burkinostkas are uninflected forms whose distribution is limited to precisely defined tokens (*omacku*, *trochu*, *naprzeciwka* in *po omacku*, *po trochu*, *z naprzeciwka*),[2] but with the exception of parts of proper names; Kubliks are an incoherent class of uninflected lexemes. They are mainly tokens modifying various classes including nouns (*nawet*, *głównie*, *prawdopodobnie*, *nie*, *-ż*, *również*[3] (adverbs with no comparison forms); foreign objects are foreign tokens which do not directly interact with tokens of the Polish language (*errare*). Thus lemmatization does not agree with the parts of speech appearing in grammars (additionally *kubliks*, *burkinostkas* and *foreign objects* appear).

[2] In translation: 'blindfold', 'little by little, from across'.
[3] In translation: 'even', 'mainly', 'probably', likewise/also/too'.

b) Morpho-syntactic information

It is necessary to follow strict mark-up procedures for annotation processes (each paragraph is annotated by 2 linguists and verified by a super-annotator, as a result of which a final version is approved).

The NKJP tagset is morpho-syntactic in its nature. As compared with the IPI PAN tagset it contains new solutions being developed from earlier ones. The morpho-syntactic annotation undertaken by the Morfeusz (morphological analyser) involved weak dehomonymisation; strong dehomonymisation was undertaken by an annotator. Each token had to be made unambiguous by an annotator, which was not always easy. On the basis of manual annotation of the 1M corpus, a PANTERA (= *Polskiej Akademii Nauk Tager Ekstrahujący Reguły Automatycznie*) tool was developed and trained, with which the morpho-syntactic description of the full corpus was made unambiguous. The Brill algorithm with a small tagset designed for the English language was adopted for this purpose; however, it had to be modified to serve an inflected language such as Polish. Similar procedures have previously been adopted in relation to Czech: tagging was divided into three stages. During the first, the tagger decides on a grammatical class of a given token and on values of chosen grammatical categories (here: case and person). Next it establishes other categories. The distinction of cases, especially Nominative, Accusative and Genitive (syncretism) as well as distinction of certain masculine genders (mainly in relation to pronouns and prepositions) are the issues which cause the biggest problems.

For example – the gender of a noun was established in this way on the basis of diagnostic contexts:

Widzę jednego[4] _____ z tych, których lubię.	m1
Widzę jednego _____ z tych, które lubię.	m2
Widzę jeden _____ .	m3
Widzę jedno _____ .	n
Widzę jedną _____ .	f

c) Semantic annotation

Semantic Word Sense Disambiguation – WSD (http://nlp.ipipan.waw.pl/Anotatornia) was undertaken in order to use automatic disambiguation. Vocabulary of senses means labels for ambiguous words. These are different to those in existing dictionaries, in which they are too detailed. While creating the dictionary of senses, 106 lexemes were chosen (out of the IPI PAN basic word list) – mainly nouns and verbs. Not more than 6 coarse senses were assigned to each lexeme. The frequency of appearance in relation to the use of a given word in a given sense was important here. In the created dictionary one word has 2.85 senses on average. The system automatically decides which of the meanings appeared in a given context. For example:

Zamek$_2$ w drzwiach trzeba było wymienić.	‘a lock in a door’
Okazały zamek$_1$ stał na wzgórzu.	‘a castle’

[4] ‘I see one ..., of those I like’ (in various genders).

There are more than 34 thousand appearances of ambiguous words in the 1M subcorpus. Experiments have been conducted to test these methods – 315 tests have been conducted for each ambiguous word, which gives us 22,790 verified methods. A fragment of a table with statistical data is presented in table 3.

Word	Number of appearances	Number of senses	Distribution of senses	MFS
action (akcja)	369	2	257/112	0,696
near (blisko)	279	2	196/83	0,703
take (brać)	299	7	142/98/30/17/6/4/2	0,475
Go (chodzić)	675	4	401/270/3/1	0,594
body (ciało)	256	3	238/17/1	0,930
member (członek)	368	3	363/3/2	0,986
Feel (czuć)	354	2	352/2	0,994

Table 3: Statistics for manually marked-up appearances of ambiguous words (source: NKJP 2012, 221)

d) Syntactic annotation

This addresses the problem of the distinction of syntactic words and syntactic groups. Syntactic words contain analytical forms (tense, mode, comparison, discontinuous conjunctions) as well as multiword units, e.g. *do czysta* 'clean' is an adverb at the level of syntactic words and an adverb group at the level of syntactic groups.

In syntactic groups a syntactic centre and a semantic centre are distinguished. The majority of groups consist of one (39%) or two (31%) semantic words.

Element	Number
token	1,068,035
sentence	72,944
syntactic word	993,684
syntactic group	305,806

Table 4: Number of elements for the syntactically annotated corpus (source: NKJP 2012, 126)

e) Annotation of proper names

There are names of people, places, institutions and temporal periods.

The annotation of units connected with Europe and the European Union was extremely complicated due to the metonymous nature of use of lexemes *Europe* and *European*, as they hardly ever referred to the continent and most frequently they referred to inhabitants of Europe as a continent or as an organisation (the *European Union* as a block of countries or as an institution). So *Europe is* treated as a geographic or geopolitical name or as a name of an organisation, e.g.

Odnalazłem przylądek św. Wincentego – miejsce, gdzie kończy się Europa [geogName]
Europa i Ameryka miały inne sprawy na głowie [orgName]
Europa powinna popierać ideę, iż wydatki na edukację [...] powinny osiągnąć w ciągu następnych 10 lat przeciętnie 10% [orgName]

For NKJP data, two Internet browsers are available: PELCRA and Poliqarp. PELCRA is available at http://pelcra.ia.uni.lodz.pl. Poliqarp at http://nkjp.pl/poliqarp/help/pl.html. The PELCRA Internet browser is based on the syntax of corpus related inquiries.

As automatic mark-up of the whole corpus is subject to mistakes, work is being done in order to minimise their appearance. This is done under the European project CESAR/ META-NET (www.meta-net.eu/projects/cesar).

f) Words of the day

Words of the day are one of subprojects for NKJP. They are words whose frequency of occurrence is clearly variable in time. They come from four daily newspapers: *Dziennik. Gazeta Prawna, Gazeta Wyborcza, Polska. The Times* and *Rzeczpospolita*. Every morning words from the previous day are published (the sport section is ignored). Words of the day are selected on the basis of the comparison of their absolute and relevant frequency in a given day and in the whole preceding year. This may be a hint as to which of them should be noted in dictionaries or language handbooks. The whole idea of selecting words is taken from the website *Wörter des Tages* of University of Leipzig.

Fig. 1: Words of the day

In other words, this means monitoring the relative popularity of given words. Reflection of political and social realities can be observed by the appearance of certain words highly ranked on a list of words for a given day. In the future, it is planned to look for "key key words" from the key words.

Analogically, words of the month and words of the year have lately been chosen; e.g. *prezydencja* ('presidency'), *kryzys* ('crisis'), *katastrofa* ('catastrophe') and *krzyż* ('cross') have been recognized for the words of 2011.

2. WSJP or Great Dictionary of Polish[5]

Taking into account its status, the amount of financial support granted from the state budget and the number of researchers and universities engaged in its development, 'Wielki słownik języka polskiego PAN' or the 'Polish Academy of Sciences – Great Dictionary of Polish' (the abbreviation of the Polish title is WSJP) deserves to be taken as the second leading project in the Polish linguistics – or rather the linguistics of Polish – of the initial decades of XXI century.

'Wielki słownik języka polskiego PAN' or WSJP is the fifth in the history of Polish linguistics attempt to collect and describe the inventory of the entire Polish vocabulary being contemporarily in use after the following:[6]

1. 'Słownik języka polskiego' (Dictionary of Polish) with ca. 80,000 entries in six volumes, edited by Samuel B. Linde and published in Warsaw 1807-1814, known as the Linde Dictionary;
2. 'Słownik języka polskiego' (Dictionary of Polish) with ca. 280,000 entries in seven volumes, edited by Jan Karłowicz, Antoni Kryński and Władysław Niedźwiecki, published in Warsaw in 1900-1927; known as the Warsaw Dictionary;[7]
3. 'Słownik języka polskiego' (Dictionary of Polish) with ca. 125,000 entries in eleven volumes, edited by Witold Doroszewski, published in Warsaw 1958-1969, known as the Doroszewski Dictionary;[8]

---

[5] This presentation of the WSJP is based mainly on the article by Piotr Żmigrodzki (see Żmigrodzki 2011).

[6] For more information about the latest history of Polish lexicography, see Żmigrodzk (2009).

[7] The Warsaw Dictionary, as the Linde Dictionary, should, in the intention of their editors, comprise the whole stock of Polish words since XVth till the end of XVIIIth century (Linde Dictionary) up to the end of XIXth century (Warsaw dictionary).

[8] The 11-volume Doroszewski Dictionary, according to the intention of its editor and publisher, covered the Polish vocabulary from 1750 to 1950. "Although its approximately 125,000 entries amount to less than a half of the SW [Warsaw Dictionary] entries, the description in SJPD [Doroszewski Dictionary] is much richer, including in particular authentic examples of word usage. Separate sections in the entry word description were devoted to phraseologisms and proverbs. Inflexion tables and a system of reference markers provided detailed information on inflexion. The dictionary played a major role in the Polish lexicography of the second half of the 20th century, becoming the source of material (and the theoretical basis) for many smaller popular dictionaries, especially the 3-volume PWN Dictionary of Polish ['Słownik języka polskiego PWN', called "the Szymczak Dictionary" – SJPSz], which sold in two million copies between 1978 and 2004, and the Little Dictionary of Polish

4. ‘Praktyczny słownik współczesnej polszczyzny’ (Practical Dictionary of Contemporary Polish) with ca. 133,000 entries in 50 volumes, edited by Halina Zgółkowa and published in Poznań in 1994-2005, known as the Zgółkowa dictionary.

But it is also in many ways a dictionary different to all its predecessors. What distinguishes it from them?

Firstly, it was meant, according to the founding resolution, as “a concerted endeavour of Polish humanists, and especially the whole linguistic branch of Polish studies”, and really the WSJP is “a kind of a **linguistic joint venture**”: its editorial team (not counting former or present temporary collaborators) included linguists from the universities of Cracow, Warsaw, Katowice, Toruń, and Olsztyn as well from the Institute of Polish Language of the Polish Academy of Sciences which serves as the leading unit of the whole project.

Secondly, it is developed in an **electronic** version and will be (and partly already is) available online for free. According to the editor, “there will be no printed version of the whole dictionary; in the future, however, the WSJP database may serve as a source for derived dictionaries, which could be published in the printed form” (Żmigrodzki 2011, 10);

Thirdly, the development of the WSJP is conceived as a **permanent** process. Its editor declared: “the dictionary is to be further developed after 2012, and due to the open nature of the project, the work should continue without end” (Żmigrodzki 2011, 11).

Fourthly, the WSJP is **based on two corpora** of Polish texts since 1945: the National Corpus of Polish [‘Narodowy Korpus Języka Polskiego’, NKJP] and supported by an auxiliary corpus created at the PAS IPL specifically to serve the needs of the emerging dictionary. This comprises texts which for various reasons were not (and are not going to be) included in the NKJP. Polish Internet sites constitute a third source.

Fifthly, the editor and authors seek to give to the WSJP an **academic** character, aiming “to employ 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” (Żmigrodzki 2011, 10).

[‘Mały słownik języka polskiego’, MSJP], first published 1968 and re-issued a number of times, in its original form, as well as in various modified versions. Even the 2003 Universal Dictionary of Polish [‘Uniwersalny słownik języka polskiego’] directly draws from the tradition of SJPD and the lexicographical framework developed by Witold Doroszewski. Since the late 1970s, however, the SJPD framework had been criticized by lexicographers of the younger generation. In mid-1980s efforts were made to create a new great dictionary of Polish but due to a number of unfortunate circumstances and also because of the political changes in Poland, this attempt was unsuccessful. After the breakthrough of 1989, it seemed that the emergence of private publishing houses would prompt new lexicographical works. Indeed, there appeared many popular dictionaries (e.g. the Dictionary of Contemporary Polish, edited by Bogusław Dunaj [‘Słownik współczesnego języka polskiego’ – SJPDun], yet the need for a comprehensive academic lexicographical description of the Polish language remained unfulfilled.” (Żmigrodzki 2011, 8-9).

Moreover, the WSJP is presented as "in principle synchronic"[9] and "in principle descriptive".[10]

The editors and authors of the WSJP see its specific value in quality and quantity of information on individual words. The full structure of the entry consists of the following nineteen fields:

1. **Headform**. For nouns, the headword form is the nominative singular (or plural, in the case of plurale tantum), for adjectives and numerals, the nominative singular masculine and for verbs, the infinitive form.
2. **Entry sub-type**. This is a technical field, i.e. it is not visible to the dictionary user; for "regular" entries, the sub-type is related to the lexical category of a given item (noun, verb, adjective, etc.), for "discontinuous" ones – with the structure (clause, verb phrase, noun phrase). The choice of the sub-type determines which forms will be added to other fields to be filled in (e.g. Syntax, Inflexion, Collocations).
3. **Variants**. Information included in this field refers to different phenomena, depending on entry type. For "regular" entries, phonetic-orthographic variants are noted here, that is such cases where a change in spelling is accompanied by a change in pronunciation, e.g. *pośpieszny* and *pospieszny*.
4. **Chronology**. Initially, the editors of the dictionary were planning to include here the exact year of the first appearance of a given headword in Polish texts, yet this plan proved unfeasible. Thus, information on chronology is reduced to information about the appearance of a particular word (or rather its graphic form) in an older dictionary of Polish. This compromise has been criticized by some Polish researchers, but the editors believe that even this kind of information on chronology may be of some help to the dictionary user and it is possible to complete the chronology data in the future.
5. **Origin**. So far this field offers etymological information only on lexemes of foreign origin. Further work on etymological information on all lexemes is planned for the future.
6. **Semantic description**. According to the editors of WSJP, the average user of the dictionary is looking most often for semantic information. Therefore, they treat it with due attention. "Thanks to the fact – the editor explains – that the entries are not created in the alphabetical order but according to a thematic classification, the work on semantic description is made easier in that the authors can start by identifying the problems and strategies of defining lexemes which belong to a given semantic field. Apart from the definition, in the case of headwords with more than one meaning there is one more component in the semantic description field, namely the guide-

[9] It covers first of all the Polish vocabulary of the XXI century "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" (Żmigrodzki 2011, 10).

[10] "The authors are not going to eliminate from description 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 [...] and mark the stylistic qualification of sub-standard units." (Żmigrodzki 2011, 10).

word (or, as we call it, the semantic identifier). Guidewords are single words or short phrases indicating the meaning of the lexeme explained in the particular sub-entry." (Żmigrodzki (2011, 18).

7. **Labels**. The WSJP employs a system of labels developed on the basis of a critical analysis of the choice and use of labels in other Polish dictionaries. The labels are given before the semantic definition.
8. **Thematic classification**. "The WSJP is – according to its editor – the first general dictionary of Polish which makes use of a thematic classification of the vocabulary. We employ a three-tier classification scheme (about 80 categories altogether). [...] In our classification, every separate meaning of a headword is of course categorized independently." (Żmigrodzki (2011, 19).
9. **Semantic relations**. The sub-entries can include lexical units exhibiting the relations of synonymy, antonymy, hyperonymy or incompatibility to the headword. However the WSJP should not be treated as a practical dictionary of synonyms.
10. **Inflexion**. The WSJP is the first general dictionary of Polish to provide direct and exhaustive information about inflexion. Full inflexion paradigms are given for all inflected lexemes, as well as the indication of gender, aspect of verbs, comparison, etc.
11. **Syntax**. The valence of the units is indicated; this information is made available to the user in the shape of symbolic syntactic schema. These might be accompanied – if need be – by a note that a given unit takes an unusual syntactic sequence or that there apply rules of semantic selection (that is the meaning of words determines their ability to form a sequence with a given lexical unit).
12. **Collocations**. Collocations, here understood as statistically frequent combinations of the headword with other lexemes, form the main bulk of the WSJP exemplification material.
13. **Quotations**. A small number (max. five per one headword meaning) of authentic quotations are included. Both quotations and collocations are taken mainly from the NKJP, some also come from other sources previously mentioned in the present paper.
14. **Abbreviations**. The WSJP also notes frequently used abbreviations of a given lexeme, e.g. *dr* from *doktor* [doctor], *zob*. from *zobacz* [see, as in "see above/below", etc.]. These abbreviations are also described in separate entries.
15. **Normative information**. Although the WSJP is a descriptive dictionary it notes in this field that a given form or usage of the headword deviates from the linguistic norm as contained in the latest edition of the 'PWN Wielki słownik poprawnej polszczyzny' (Great Normative Dictionary of Polish). This opinion is not verified or commented by the editors of the WSJP.
16. **Notes on usage**. This field includes any additional information on the headform word that could not be entered in any of the previous fields. For example: that the given word is sometimes confused with another one.
17. **Derivatives** – this field is completed for "proper name" entries only, in the description of the names of towns and states. The sub-fields include: the name of the male inhabitant, the name of the female inhabitant and the derivative adjective.

18. **Expansion**. This field is devoted to information only on entry-types 'abbreviations' (e.g. nr = numer [number], prof. = profesor [professor]) and 'acronyms' (e.g. PIT, for 'Personal Income Tax'.
19. **Lexemes**. This field is used for the entry-type "morphemes" and contains several examples of words in which the headword morpheme is found.

As was said, the full structure of the entry consists of nineteen fields (elements of description). However, the structure of particular entries and the range of information included, depend on the type of the described lexical units. The WSJP distinguishes seven types of entries:

regular (single words);
discontinuous (idioms, proverbs, winged words);
abbreviations;
acronyms;
proper names;
functional lexemes;
morphemes.

The last three (of the nineteen) fields are opened only for some types of entries and in particular: the field Derivatives is opened only for proper names; the field Expansion is opened only for abbreviations and acronyms and the field Lexemes is opened only for morphemes.

The WSJP is – to use the term coined by its editor Piotr Żmigrodzki (2008) – a primarily online dictionary, which means that it has been developed to be presented on the computer screen. As a result, the basic entry view that presents itself to the user is a structured "tab view".

The work to be completed by the end of 2012 is just the beginning of the whole program. The number of entries should be further expanded until practically all lexical units of 21st-century Polish language are described. Due to the electronic form of the work – a form open by its nature – the development of the WSJP can continue without end. On the one hand, new entries can be always added, and on the other hand, the existing descriptions can be extended, new fields included and entries improved.

3. References

NKJP (2012): *Narodowy Korpus Języka Polskiego*. Red. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, w przygotowaniu, kwiecień 2012.

Żmigrodzki, P. (2008): *Słowo – słownik – rzeczywistość. Z zagadnień leksykografii i metaleksykografii*. [Word – Dictionary – Realisty. Problems of Lexicography and Metalexicography]. Kraków: Lewis.

Żmigrodzki, P. (2009): *Wprowadzenie do leksykografii polskiej*. [Introduction to the Polish Lexicography]. Third expanded edition. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

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