UDK 811.16'373.46:81
811.16'374
Stručni rad
Rukopis primljen 2. X. 2019.
Prihvaćen za tisak 18. III. 2020.
doi.org/10.31724/rihjj.46.2.18

Paweł Kowalski
The Institute of Slavic Studies of the Polish Academy of Science
Juliana Bartoszewicza 1B, PL-00337 Warszawa
pawel.kowalski@ispan.waw.pl

MULTILINGUAL DICTIONARY OF KEYWORDS AS A TOOL FOR THE DIGITAL BIBLIOGRAPHIC DATABASE OF WORLD SLAVIC LINGUISTICS

The paper presents the structure of a multilingual dictionary of keywords, which is an integral part of the bibliographic database of Slavic linguistics iSybislaw representing the digital information retrieval system (www.isybislaw.ispan.waw.pl). The lexical units (keywords) of the language of keywords used in the system are represented primarily by linguistic terms. In spite of a different denotation – the keywords directly denote sets of documents, and indirectly the non-documentary reality, while the terms denote elements of linguistic reality – they are formally equal with linguistic terms, which allows them to map the semantic field of a particular discipline, in this case Slavic linguistics. The dictionary is therefore a domain-based online specialist dictionary, which is a tool for users of the bibliographic database of Slavic linguistics. The dictionary is addressed to all those who deal with linguistics and linguistic terminology, first of all to scholar-linguists, Ph.D. students and students of philologies, as well as translators of academic papers in the field of linguistics.

1. Introduction

At the beginning of the 21st century dynamic digitization has a very strong impact on each field and sub-field of linguistics including modern lexicography. Not so long ago one could search for lexical information mostly in printed dictionaries. Computers were used to create dictionaries in the second half of the 20th century giving an impulse for electronic lexicography and to support lexicographers' work, but the electronic dictionaries were not wide-spread (Granger and Paquot 2012: 1–4). Nowadays, the number of digital dictionaries and different databases in open access is growing. This allows us to retrieve sufficiently relevant information. Such international projects as CLARIN or DARIAH that are carried out by experts in the humanities and IT specialists try to integrate the resources and elaborate digital tools for researchers, especially to solve the information-quality control problems and to overcome information overload.

There is no one single homogeneous field of lexicography (Fuertes-Olivera 2017: 1). Instead, we can observe varieties of such a field, i.e. monolingual, bilingual, multilingual lexicography, natural language and special language (terminological) lexicography, etc. so the notion of lexicography has to be considered as an inclusive term (cf. Bergenholtz and Gouws: 2012). The natural environment for academic work is nowadays the digital space (area) so there is a need to create digital tools for linguists that will support the knowledge processing procedures. One such tool is a multilingual dictionary of keywords integrated with the Bibliographic database of Slavic linguistics publications (iSybislaw information retrieval system), which is an initiative of Polish Slavic environment linked institutionally to the Institute of Slavic Studies of the Polish Academy of Sciences.

The aim of the paper is to present the structure of this multilingual dictionary of keywords. In previous works, the multilingualism of keywords in the database has been briefly discussed (cf. Rudnik-Karwatowa 2002, Banasiak 2014, Karpilovska 2014, Łuczków 2014, Stanojević and Kryżan-Stanojević 2014, Kowalski and Banasiak 2017). However, so far the issue has not been broadly acknowledged and addressed from the lexicographical (and e-lexicographical) perspective. Thus, the paper brings new insight into the topic and gives prediction about the further development of the dictionary.

2. iSybislaw – the Bibliographic database of Slavic linguistic publications

The iSybislaw information retrieval system remediates a printed bibliography that was published at the Institute of Slavic Studies PAS in the 90s on the initiative of the Slavic Academic Information Centre founded at the beginning of 1990 by Zofia Rudnik-Karwatowa. Historically, it continues the tradition of bibliography on Slavic linguistics published since 1908 in the journal *Rocznik Slawistyczny*. From the beginning, it was a current bibliography with annotations and abstracts that includes articles printed worldwide, published in the previous year. The iSybislaw system was launched in 2007 superseding the local Sybislaw database dedicated to generating the paper bibliography and only available on the computers at the Institute of Slavic Studies of the PAS. From the beginning, the implementation of the system has been carried out in collaboration with linguists and specialists in the field of IR systems and IR languages from Polish and foreign scientific institutions. The modern iSybislaw system gives (distributed in open access) formal (in natural language) and substantial characteristics (in the form of keyword language and classification language) of the indexed documents. The international team that works with the database index (elaborate) of the well-known and prestigious Slavic linguistic journals and monographs relevant for Slavic linguistics from each of the Slavic countries.

2.1. Keywords and terminology in the bibliographic database

The iSybislaw system uses two information retrieval languages (IRL) that map information contained in the documents and allow to retrieve the information. Linguistic terminology is used in both IRLs, however, their characteristics differ. The linguistic terms in the classification language are constructed within the repetitive classes with a dominant structure: number + language group or the name of the language + linguistic discipline or subdiscipline, e.g. *3. Bulgarian-Macedonian group*; *3.1. Bulgarian language*; *3.1.1. Bulgarian. Modern literary language*; *3.1.1.1. Bulgarian. Phonetics. Phonology*; *8. Lech group*; *8.2. Polish 8.2.1. Polish. Modern literary language* etc. The main IRL of the system is the keyword language that has greater functional potential in comparison with the classification language. In the keyword language that is not so structurally restricted as the classification language, the terminology occurs along with the non-terminological items. These non-terminological items are expressions from natural languages and they reproduce elements of document content relevant for search pragmatics.

Such non-terminological keyword units include: personal names: *Mickiewicz Adam*, *Słowacki Juliusz*, including names of linguists whose scientific achievements are the subject of research reported in the documents: *Sławski Franciszek*, geographical names: *Chorwacja*, *Polesie*, *Zagreb*, and corporate names: *Polska Akademia Nauk*, *Międzynarodowy Komitet Slawistów*, *Uniwersytet Łódzki* (cf. Rudnik-Karwatowa 2002, Kulpina and Tatarinov 2014).

From the perspective of information science, it is important to note the fundamental difference between terminology and keywords. Keywords denote directly sets of documents contained in the iSybislaw and directly connote some of their subsets. The terms denote and connote elements of linguistic (extra-textual) reality. Therefore, the keywords indirectly denote extra-documentary reality objects described in the documents. For example, one may not understand the meaning of the keyword *noun* as a word functioning as the name of a specific object, but may understand it as a unit mapping the set of all publications referring to nouns in the system. As we can observe, semantics of keywords is not given to the user directly. However, using the natural notation in the keyword language, users might decode the content of an individual keyword unit based on their knowledge of the meaning of a particular term (graphically equivalent in the terminological system). All this leads to the further conclusion that linguistic terminology as keywords in the iSybislaw system has two functions: search and meta-information function. The meta-information function consists of mapping the content of the document in the form of search characteristics, while the retrieval function enables to obtain in the system information set the information relevant for the user. (Bojar 2002, Banasiak 2014, Kowalski and Banasiak 2017).

3. Dictionary of keywords in the field of Slavic linguistics

To provide relevant information by indexers and help users to organize bibliographic query in the printed version of Bibliography Zofia Rudnik-Karwatowa and Hanna Karpińska created the first dictionary of keywords in the field of Slavic linguistics. The Dictionary was printed in 1999 (Karwatowa and Karpińska 1999) and contained about 2500 dictionary entries and was made available in the electronic version in 2006. The electronic version was expanded to about 3000 entries. The dictionary was considered a special domain dictionary and a tool that facilitates the usage of the Slavic linguistic bibliographic database iSybislaw, because it was implemented within the system. Primarily, it was very useful for the indexers because it helped to create relevant characteristics of the documents (indexing process), and furthermore it reflected the semantic field (main topics) of modern Slavic linguistics and its publishing production. However, looking back, the monolingual character of the dictionary occurs as the major inconvenience. The dictionary with the biggest sets of keywords in Polish was a tool mainly designed for Polish-speaking users. It contrasted with multilingualism of indexed documents.

3.1. Methodology beyond the printed dictionary

When one acknowledges the methodology of creating a terminological dictionary it is often (in most cases) mentioned that the terms were excerpted from other relevant lexicographic work and from the linguistic texts and it is a standard procedure (see for example Bekisz nad Fontański 1997). In the dictionary of keywords (paper version) the authors followed the theory and applied the induction-deductive method to collect the Dictionary lexis. They used a corpus of texts representative of literature on the specific subject and their documentary descriptions (induction process), and on the other hand – supplementing the created lexical resource based on lexicographic sources, such as encyclopedias, thesauri, as well as terminological dictionaries and academic grammars (deduction process). To give an overview and characteristics of modern Slavic linguistics they are based on linguistic material (all kinds of sources from representative grammars, monographs, studies, and articles on Slavic linguistics since 1990), only with a little support from selected works from the period 1970–1990. As mentioned above, the dictionary includes not only the most common terms but also those that have theoretical value and are used in documents representing a particular school or research method. However, the dictionary did not provide all keywords related to Slavic linguistics, but only those that reflect the topics of a representative number of documents currently being created. It did not provide all equivalent keywords. For example, in the documents from Slavic Linguistic

Bibliography one can find Polish linguistic terms such as *jer twardy*, *mocna pozycja jera*, *jer napięty*, etc. (*hard yer*, *strong position of the yer*), but in the Dictionary one can find only two of them (one with qualificator *zob.* ‘see’): *jer twardy* and *jer napięty zob. jer twardy* (Rudnik-Karwatowa and Karpińska 1999: 51). The authors of the dictionary have chosen in most cases only one keyword unit (most frequent), so the rest of the relevant terms were not presented in the form of keywords (for example *abrewiacja* ‘abbreviation’). From the beginning, the structure of the Dictionary was flat, without any hierarchical relationships between the items. For example, there is no connection between *językoznawstwo diachroniczne* ‘diachronic linguistics’ and *językoznawstwo synchroniczne* ‘synchronic linguistics’ but it might be useful to present them in the term family and in connection with more general term *językoznawstwo* ‘linguistics’. The preferred terms have been identified using cross-references, which represents a whole range of terms in the language of keywords, related in a natural language by a relation of synonymy and proximity. The dictionary takes into account the phenomenon of polysemy resulting from the fact that in linguistic literature the same terms are sometimes used to describe various phenomena. The polysemic entries are presented with numbers, e.g. *przekład 1* (*translation 1*) ‘result of action’ and *przekład 2* (*translation 2*) ‘action’. The selection of preferred keywords is guided by the following criteria: frequency (incidence); terminological (linguistic) correctness; timeliness (current usage); brevity; structural clarity, and whether the term is native or foreign.

4. Digital dictionary of keywords in the iSybislaw system (multilingualism and classes of equivalence)

Remediation of the Bibliography from the paper version to the digital space in 2007 opened new perspectives and allowed to integrate the monolingual dictionary of keywords with the system. However, remediation brought also some practical and theoretical obstacles. Primarily, the problem of interlanguage equivalency (cross-language equivalency) has occurred due to the multilingualism of the documents and indexers. For the dictionary, the digital space allows up-to-date upgrade of the lexis, so Polish units can be continuously expanded and supplemented by the interlanguage equivalents (keywords in other Slavic languages as well as in English). The work on foreign language dictionaries of keywords is not equally advanced and without such integrated and connected dictionaries one cannot retrieve cross-lingual information in the system.

Table 1: Incomplete class of equivalence *językoznawstwo konfrontatywne* ‘contrastive linguistics’

contrastive linguistics	
słowo kluczowe (‘keyword’)	język (‘language’)
językoznawstwo konfrontatywne	polski
сопоставительное языкознание	rosyjski
супастаўляльнае мовазнаўства	białoruski

Dictionary – when fully developed – will provide access to information by every item whether it is in Croatian, Polish, Russian, Slovene, or other Slavic language and English, which was impossible in the previous system and in the printed bibliography. For example, the users who know one or two Slavic languages and are unfamiliar with the English language will be guided by the system to all heterolingual documents, in which the needed information is included. Conversely, English-speaking users, having only language competence in the area of one of the Slavic languages will be able to reach information about the whole equivalent class by searching for only one keyword unit. Ultimately, a multilingual dictionary is to present all multilingual items and the users will be able to switch from one language to another.

4.1. Some remarks on particular problems and methodology

The first stage in constructing equivalence classes is to gather terminological units in one language and then organize them as well as eliminate ambiguity. The elimination of ambiguity of the terminology that is used as the keywords is of cardinal importance. Without taking it into account the user cannot reach relevant information already within one language, and the effects of “projecting” an image from a given language onto another are unpredictable. For example, the ambiguous Polish term, *język* (‘the system of signs’ vs ‘the speech organ’) corresponds to two terms in English, respectively: *language*; *tongue*. The English situation is further complicated by the fact that the term *tongue* is also rarely used in the sense of 'system of signs'. To eliminate these problems and to present basic semantic of the term that is used as a keyword, which may help the user to retrieve relevant information, specific keywords are complemented with basic micro-definitions in parentheses, e.g. *gramatyka 1 (system reguł językowych)* 'grammar 1 (system of language rules)', *gramatyka 2 (dyscyplina)*, 'grammar 2 (discipline)', *gramatyka 3 (dokument)* 'grammar 3 (document)'; *pragmatyka 1 (użycie znaków językowych)* 'pragmatics 1 (using the linguistic signs)', *pragmatyka 2 (dyscyplina)* 'pragmatics 2 (discipline)'; *semantyka 1 (znaczenie znaków językowych)* 'semantics 1 (meaning of the linguistic signs)', *semantyka 2 (dyscyplina)* 'semantics 2 (discipline)'.

In addition to this kind of ambiguity, there are also more regular phenomena in terminology, which can be taken into account and described in terms of specific lexical parameters, cf. e.g. process → result (process → result): pol. *nominacja 1* 'nomination' *(process) → nominacja 2* 'nomination' *(result)*, *przekład 1* 'translation' *(process) → przekład 2* 'translation' *(result)*, *zapożyczenie 1* 'borrowing' *(process) → zapożyczenie 2* 'loanword' *(result)*[1]. From the point of view of access to the relevant information set in the dictionary the importance of micro-definition for the users for such ambiguities *nominacja 1 (proces)* 'nomination 1 (process)' → *nominacja 2 (rezultat)* 'nomination 2 (result)' is smaller than in the case of Polish type pairs *język 1* 'language' vs *język 2* 'tongue', that is more irregular, and thus more informative. In the pairs with parameters: process and result the items are bound by a semantic derivative that reflects the close relationship that occurs in non-linguistic and extra-documentary reality between the indicated processes and their results. Information relevant to the former (*nominacja 1*) is highly relevant to the latter (*nominacja 2*).

One interesting example would be designing a class of terms related to univerbation. In Polish linguistic terminology, the international term *uniwerbacja* has become more common, although it was used for the first time in Polish in the form of *uniwerbalizacja* (Siatkowska 1964). Variant form of *uniwerbalizacja* was not accepted in the Polish linguistic environment and in principle, its use can be described as individual (occasional). For this reason and due to the lack of credentials for this term in a larger number of texts, it has not been included

[1] More on this topic see in Kowalski and Banasiak 2017.

in the iSybislaw system keyword collection. In Slovenian linguistics there are two variants of the international term *univerbizacija* and *univerbacija* (similarly in other Slavic languages) and the native term *poenobesedenje*. The definition ranges of the native and international terms differ. The term *poenobesedenje* covers a wide variety of word-formation processes. As Jože Toporišič writes: "Poenobesedenje is the formation of one word from two or more words, e.g. *seveda* from *se ve*, *da*; or replacement of a two- or multi-word using word formation derivational mechanisms, e.g. *nedovršni glagol* > *nedovršnik*, *cesta za avte* > *avtocesta*" (Toporišič 2002: 187). The term *poenobesedenje* can therefore be regarded as superior to the international term *univerbizacija* / *univerbacija*. However, in many works of Slovenian linguists both terms are used interchangeably. The introduction of two units to the keyword set as separate classes would distort information perception and could affect the receipt of an incomplete set of documents devoted to the broadly understood issues of univerbation in Slovenian works. To increase search pragmatics and create a user-friendly system, all three Slovenian units: *poenobesedenje*, *univerbizacija* and *univerbacija* should be combined into one class of intra-language equivalence (within one language). Then such a monolingual equivalence class must be combined with other language classes that reflect the same semantic field 'univerbation'. Thus, the dictionary presents, although in a simplified manner, the place of individual terms used as keywords in relations and gives an image of the language terminological system.

Table 2: The dictionary entry of the keyword *univerbation*

univerbation 'forming one word from two or more words'	
słowo kluczowe ('keyword')	język ('language')
uniwerbizacja	polski
uniwerbacja	polski
uniwerbalizacja	polski
универбация	rosyjski
універбацыя	białoruski
univerbacija	chorwacki
univerbizacija	chorwacki

univerbizace	czeski
univerbizácia	słowacki
poenobesedenje	słoweński
univerbacija	słoweński
univerbizacija	słoweński
універбація	ukraiński
універбізація	ukraiński

It is worth noticing that in the iSybislaw database the user can search for information (documents) by using each item (keyword) from the multilingual entry. Online character of the dictionary allows also to switch the language of the main keyword in the entry, so the English univerbation might be replaced by Polish, etc.

5. Summary

The primary aim of the Dictionary of keywords is to allow users to retrieve accurate and complete information that will satisfy their information needs. To reach this, it is necessary to extend, build, and develop its lexical resource (set of keywords) with great care and with the integration of other digital tools. The omission of useful topics and hence important items in the dictionary of keywords will result in the loss of information (information silence and blank spaces in the system). The goal is to find a balance between a generalization and redundancy of information provided in the dictionary. At present, there is still a large disproportion between lexical units within particular languages, which limits to some extent the functionality of the system. One solution is the introduction of a list of the first level of keyword language fragmentation and supplementing the equivalence classes with the most important units in the set of keyword classes (just like showed above in the class of univerbation). Unlike in the paper version of the dictionary in the digital dictionary, it is possible to include already available computer and digital infrastructure CLARIN, which is dedicated to analyzing linguistic documents and excerpting linguistic terms. There are such tools as the first version of the thematic Wikipedia K-Nearest Neighbors theme for Polish and English texts ("PELCRA NLP Tools – WiKNN classifier" http://pelcra.clarin-pl.eu/tools/classifier/); a tool for determining keywords in the text ("ReSpa" https://ws.clarin-pl.eu/respa.shtml); a tool for detecting terms in the text ("TermoPL" http://zil.ipipan.waw.pl/TermoPL). It is the future of the digital dictionary of keywords to integrate these tools and optimize the work of lexicographers compiling a dictionary.

To sum up the reflection about the Polish printed dictionary of keywords and digital multilingual dictionary of keywords integrated with the bibliographic database of Slavic linguistic documents we can argue with a well-known statement about electronic lexicography that in electronic (digital) dictionary you only get what you search for. Instead of this, in the digital dictionary of keywords due to equivalent classes by querying you have got an opportunity to travel around the lands of varieties of multilingual words; at the beginning, you never know where you will end. This is where the printed dictionaries and electronic dictionaries meet.

References

Banasiak, Jakub. 2014. Synonymy and search synonymy in an IR system: on the basis of linguistic terminology and the iSybislaw system. *Studia z Filologii Polskiej i Słowiańskiej* 49. 176–187. dx.doi.org/10.11649/sfps.2014.017.

Bekisz, Wiktor; Fontański, Henryk. 1997. *Białorusko-polsko-rosyjski słownik terminów lingwistycznych i leksyki specjalnej.* Wydawnictwo Uniwersytetu Śląskiego. Katowice.

Bergenholtz, Henning; Gouws, Rufus H. 2012. What is Lexicography?. *Lexikos* 22. 31–42. doi.org/10.5788/22-1-996.

Bojar, Bożenna. 2002. *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Wydawnictwo SBP. Warszawa.

Fuertes-Olivera, Pedro A. 2010. *Specialised Dictionaries for Learners*. De Gruyter. Berlin – New York.

Fuertes-Olivera, Pedro A. 2017. *The Routledge Handbook of Lexicography.* Routledge. London.

Granger, Sylviane; Paquot Magali. 2012. Electronic lexicography: From challenge to opportunity. *Electronic Lexicography.* Eds. Granger, Sylviane; Paquot, Magali. Oxford University Press. Oxford.

Łuczków, Iwona. 2014. Uwagi o ekwiwalencji słów kluczowych w systemie informacyjnym iSybislaw: na przykładzie wybranych terminów gramatycznych. *Studia z Filologii Polskiej i Słowiańskiej* 49. 203–218. dx.doi.org/10.11649/sfps.2014.019.

Karpilovska, Evgeniia A. 2014. Rol terminiv-nepriamych nominacij u tezaurusi informacijno-poshukovoï sistemi slavistichnogo movoznavstva. *Studia z Filologii Polskiej i Słowiańskiej* 49. 150–163. dx.doi.org/10.11649/sfps.2014.015.

Kowalski, Paweł; Banasiak, Jakub. 2017. Ključovi slova ta klasi ekvivalentnosti v sistemi iSybislaw jak znarjaddja ta predmet doslidžen. *Ukrainska mova* 4/64. 14–26.

Kulpina, Valentina G.; Tatarinov, Viktor A. 2014. Sinonimija ključovih slov v sisteme slavističeskoj bibliografii kak epistemologičeskije refleksy razvitija lingvističeskoj terminologii. *Studia z Filologii Polskiej i Słowiańskiej* 49. 188–202. dx.doi.org/10.11649/sfps.2014.018.

Rudnik-Karwatowa, Zofia. 2002. Słowa kluczowe – elementarne jednostki leksykalne języka informacyjno-wyszukiwawczego. *Słowo z perspektywy językoznawcy i tłumacza*. Eds. Pstyga, Alicja; Szcześniak, Krystyna. Wydawnictwo Uniwersytetu Gdańskiego. Gdańsk.

Rudnik-Karwatowa, Zofia; Karpińska, Hanna. 1999. *Słownik słów kluczowych językoznawstwa slawistycznego*. Slawistyczny Ośrodek Wydawniczy. Warszawa.

Siatkowska, Ewa. 1964. Syntetyczne i analityczne nazwy w języku czeskim i polskim. *Prace Filologiczne* 18/2. 219–231.

Stanojević, Marek; Kryżan-Stanojević, Barbara. 2014. Słowa kluczowe podawane przez autora publikacji jako podstawa opisu bibliograficznego w iSybislawie. *Studia z Filologii Polskiej i Słowiańskiej* 49. 219–231. dx.doi.org/10.11649/sfps.2014.020.

Toporišič, Jože. 1992. *Enciklopedija Slovenskega Jezika*. Cankarjeva založba. Ljubljana.

Višejezični rječnik ključnih riječi kao alat za digitalnu bibliografsku bazu svjetske slavistike

Sažetak

Rad prikazuje strukturu višejezičnoga rječnika ključnih riječi koji je sastavni dio bibliografske baze slavenske lingvistike iSybislaw (www.isybislaw.ispan.waw.pl). Ključne su riječi u bazi prije svega jezikoslovni nazivi. Označavaju skupove dokumenata i mapiraju semantičko polje posebne discipline, u ovome slučaju slavenskoga jezikoslovija. Rječnik je, dakle, internetski specijalni rječnik utemeljen na jezikoslovnoj domeni koji služi kao alat za korisnike bibliografske baze. Rječnik je namijenjen svima koji se bave jezikoslovljem i jezikoslovnim nazivljem, prije svega znanstvenicima jezikoslovcima, studentima filologija, doktorandima te prevoditeljima znanstvenih radova iz područja jezikoslovlja.

Keywords: electronic lexicography, terminological lexicography, linguistic terminology, keywords, information retrieval system

Ključne riječi: elektronička leksikografija, terminološka leksikografija, lingvistička terminologija, ključne riječi, informacijski sustav pretraživanja