Jakub Banasiak
(Instytut Slawistyki PAN, Warszawa)

Synonymy and search synonymy in an IR system (on the basis of linguistic terminology and the iSybislaw system)

iSybislaw is an online IR (information retrieval) system presenting bibliographic information on the works in the field of Slavic linguistics, Slavic-non--Slavic contrastive studies and (to some extent) general linguistics. A keyword language is used as the main IR tool, there is, however, also a classification system implemented. The classification language is traditional and similar to the one used in the printed predecessor of the database and will not be subject to further deliberation in this paper.

Linguistic terminology, which is the core of the vocabulary reflected in the iSybislaw's keywords, is in its primal metalinguistic function subject to the same regularities and changes that occur in general vocabulary. From the viewpoint of application in IR systems this presents a serious inconvenience to both the user and the database indexer. The language is subject to numerous processes that result in the emergence and disappearance of phenomena such as variantivity, homonymy/homography and polysemy. A sharp

distinction between synchronic and diachronic phenomena, currently considered standard in linguistic studies, is difficult to apply in the case of vast data banks in which older works coexist with new ones. Practical application of consistent and current terminology in the description of all of the indexed information seems almost impossible because of the diversity of research methods and methodological trends. Such standardization of terminological system, along with the elimination of contradictions and ambiguities would be a great help in the process of creating an IR system. It should be noted, however, that it would be a major simplification of the image of the scientific field that emerges from the database. A significant problem lies thus in the ambiguity of linguistic signs as such. The relationship between a linguistic exponent and the concept (i.e. the semantic component of a linguistic unit) is rarely unambiguous. One concept may be expressed by multiple strings of phonemes/graphemes (synonymy) and one string of phonemes/graphemes may express different concepts (ambiguity). These phenomena, non-relevant from the perspective of everyday communication (because of context etc.), turn out to be crucial in the process of optimization of information retrieval both in closed and open collections.

There are two distinctive levels considered in this paper. The first one is primarily metalinguistic resulting from the character of linguistics itself and it being the subject presented in iSybislaw, the second is meta-informative and is a result of the character of iSybislaw (it being an IR system).

Before I can proceed any further in the deliberation of the impact that synonymy and similar phenomena have on IR, I must note that the elimination of ambiguity is a necessary preliminary condition for such an analysis. Due to the binary character of the study we should first establish the notions of synonymy in natural language (including metalanguage) and synonymy in IR languages, such as the keyword language implemented in iSybislaw. One has to note that whilst synonymy in natural language is not a problem per se (it may however be subject to study), synonymy in IR systems is not only an interesting phenomena, but mainly a problem of practical nature (limiting the effectiveness of a search in terms of its completeness). On the basis of *Encyklopedia językoznawstwa ogólnego* we can give the following definition of synonymy: expressing the same content using two or more different linguistic forms (cf. Polański, 1999). Although owing to language economy, also typical of specialized languages, diachronically synonymous terms may differentiate their meaning. In the case of IR tools it is necessary

to combine synonymous expressions or remove those of them that the creator of the system would consider (for various reasons) redundant or non-preferable. The second of these solutions, however, requires the user of an IR system to be accurately acquainted with the conceptual apparatus used in indexing, and thus it makes information retrieval problematic. Of course, the creators of iSybislaw are aware of the complexity of the phenomena and changes characteristic for the terminological subsystem and to some extent take them into account in the database. In any synonymous string a single word is highlighted as a key descriptor, based on its usage, frequency, linguistic correctness and clarity, see the entry: *termin preferowany* (Eng. *preferred term*) in *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych* (Bojar, 2002).

A linguistic sign is considered to consist of its form (phonemic or graphemic), connotation and denotation. The denotation of a sign is widely believed to be dependent on its connotation. This matter is more complicated in IR languages (even those para-natural) because of the meta-informative function of IR in general, resulting in keywords having both direct and indirect connotation and denotation (cf. Bojar, 2002). Therefore the relation of search synonymy requires two or more expressions in an IR language to have identical direct and indirect denotation and connotation (cf. Bojar, 2002). The indirect connotation and denotation of keywords can obviously be derived from the para-natural character of the keyword language. The direct denotation of a keyword (being a set of documents on the subject) must be created during indexation by ascribing the given keyword to bibliographic records (or other [meta]data depending on the system in question). We may therefore conclude that keywords may be indirectly synonymous (i.e. have identical indirect connotation and denotation) as a result of the para-natural character of the used IR language. Their direct synonymy can only be achieved through the optimization of the used IR language and only then can we speak of search synonymy. In iSybislaw this can be accomplished simply by linking synonymous keywords. This is a great advantage over some popular software packages used for creating open source repositories such as DSpace, in which synonymous keywords cannot technically be linked to one another. Having to add every synonymous keyword separately in every record in DSpace makes search synonymy virtually impossible and may lead to information overload.

The core of linguistic terms functions as metalinguistic. Used in information retrieval system, their equivalents function providing data on metalinguistic

and meta-scientific information contained in the described works. Within the framework of scientific information the need for such a choice of keywords that they be as informative as possible and thus have their scope defined in the most unambiguous manner possible is often highlighted. There is no doubt that strict definitions are an extremely important component of good scientific workshop. Obviously, even within a single language the same denotation can be assigned to different names, defined and understood in slightly different ways. This phenomenon in general language is described as the so-called profiling. In the case of terminology, however, the problem is often not limited to random semantic features (different associations of a given expression) and considers qualities essential to the definition (i.e. its differentia specifica). Used in IR terms refer indirectly to themselves (the concepts they name) and directly to documentary reality (the set of documents on the subject). The users information needs seem a good standpoint for further deliberation. Since iSybislaw is mainly used by linguists we can assume that they seek primarily metalinguistic content (information on the phenomena of linguistic reality). Therefore denoting the same set of linguistic elements seems more relevant than the means by which they are defined. The division between purely metalinguistic and meta-scientific terms was mentioned above, in reality there is a large group of mixed terms:

Table 1. Types of linguistic terms

Metalinguistic terms	Meta-scientific terms	Mixed terms
Pol. *rzeczownik* (Eng. *noun*)	Pol. *lingwistyka antropologiczna* (Eng. *anthropological linguistics*)	Pol. *referencja* (Eng. *reference*)
Pol. *podmiot* (Eng. *subject*)	Pol. *lingwistyka korpusowa* (Eng. *corpus linguistics*)	Pol. *kwantyfikacja logiczna* (Eng. *logical quantification*)
Pol. *zdanie* (Eng. *sentence*)	Pol. *lingwistyka kognitywna* (Eng. *cognitive linguistics*)	Pol. *intensjonalna teoria rodzajnika* (Eng. *intensional theory of the article*)

Such terms present additional difficulty in the process of indexing. Adding a methodologically more neutral keyword is one of the possible solutions. For the

mixed terms presented in table 1. adding Pol. *określoność/nieokreśloność* (Eng. *definiteness/indefiniteness*) seems like a plausible solution.

For example, there is no doubt that in all the Polish works in the field of Slavic studies the following terms for imperceptive mood: *tryb nieświadka*, *narrativus/narratyw*, *imperceptivus* and *tryb imperceptywny* all refer to the same set of verb forms in Bulgarian and/or Macedonian, but they do it in a different way. The diversity of meanings of linguistic terms with this denotation in Polish, Russian and Bulgarian is presented in the table below. The confusion is such that it results even in abandoning domestic terminology. For instance M. Ledzion--Jelen chooses to use the Macedonian term *прекажаност* (Eng. *re-narrativeness*) (cf. Ledzion-Jelen, 2009, p. 130).

Table 2. Eng. *imperceptive mood*

meaning	Polish	Bulgarian	Russian
re-narration	*narrativus/ narratyw*	*преизказно наклонение*	*пересказыватель- ное наклонение*
not witnessing	*tryb nieświadka*	*несвидетелско наклонение*	*несвидетельское наклонение*
lack of perception	*imperceptivus/tryb imperceptywny*	*заглазное накло- нение*	*имперцептив*

All terms in the above table can be defined in such a way that their scope is strict and the only loss of information occurs because of some connotational differences. Such terms are combined into sets of synonyms in one language and sets of equivalents on multilingual level enabling cross-lingual IR in iSybislaw.

In the database we consistently distinguish between two levels of linguistic reality – the formal and the content plane. This results in the separate treatment of semantic units such as Polish *imperceptywność* (Eng. *imperceptivity*) and the means of expressing a given notion/semantic category etc. (both grammatical and lexical) such as Polish *tryb imperceptywny* (Eng. *imperceptive mood*). This division is sometimes troublesome because such an approach is not yet prevalent in all linguistic frameworks. It is worth noting that the picture emerging in this regard from particular languages is largely due to the usage and tradition. Both in Polish, Bulgarian, and Russian the term for predicate

acts both as the name of a semantic and syntactic (i.e. formal) component. To maintain consistency we found it necessary to add a subscript to the second/ secondary (formal) meaning of the term. The table below presents synonymous strings for the term in Polish, Russian, and Bulgarian.

Table 3. Eng. *predicator*

Polish	Bulgarian	Russian
wyrażenie predykatywne	*предикативен израз*	*предикативное выражение*
predykat 2	*предикат 2*	*предикат 2*
predykator	*предикатор* (seldom)	*предикатор* (seldom)
predykat składniowy	*синтактичен предикат*	*синтаксический предикат*

There is no doubt that the interchangeable use of all of the specified terms in one scientific work (or even more broadly – one terminological idiolect) would lead to inconsistencies. It turns out that authors' preferences in this area vary and have different motivations. For instance Z. Topolińska uses the term Pol. *wyrażenie predykatywne* (Eng. lit. *predicative expression*) very consistently (cf. Topolińska, 1999). As we can see in the table above the presented terms can even be grouped in such a way that they correspond not only by meaning but also by form. Such is not always the case as can be seen in table 4. presenting the Polish equivalents of the Russian term *предикатив* (Eng. *non-inflectional verb*) (with probably stabilized meaning in Russian) (cf. Ахманова, 1966; Немченко, 2008) and its synonyms. The use of Polish terms such as *przysłówek predykatywny* (Eng. *predicative adverb*) is very rare and may be viewed as a result of Russian influence. And thus arises the question (relevant in translation) which of the non-corresponding terms should be viewed as the most strict equivalents. For example, the distinction between verbs and adverbs seems well documented in linguistics and yet Polish and Russian differ slightly in the manner they treat non-inflectional verbs (cf. the use of Rus. *наречие* [Eng. *adverb*] in two-word terms in Russian as opposed to the use of *czasownik* [Eng. *verb*] in Polish). Therefore, one can concur that in Russian terminology the phenomenon is viewed as a certain kind of an adverb. In Polish terminology, however, the view that it is a special kind of verb seems prevalent. Of course these are only preliminary observations and it seems that a deepened research should take into account the text frequency of the considered terms.

Table 4. Eng. *non-inflectional verb*

Russian	Polish
категория состояния	*kategoria stanu*
безлично-предикативное слово	--
предикативное слово	*leksem predykatywny* (very seldom)
предикативное наречие	--
безличное наречие	--
предикатив	*predykatyw*
бессубъектное прилагательное	*czasownik niewłaściwy*
--	*czasownik niefleksyjny*
--	*czasownik nieosobowy*

It should be noted that the classification of parts of speech is rarely strict enough to create separate sets of units without any ambiguity. For example, in Polish terminology it is possible to use the name *predykatyw* (Eng. lit. *predicative*) in a broad sense, synonymous with widely understood *czasownik* (Eng. *verb*) (and therefore *predykatyw* 2 [i.e. *predykatyw* in the above sense] would determine a set of linguistic units, such that *predykatyw* in its primary meaning would be a part of) (cf. Kubiszyn-Mędrala, 2000). One should also note that both in Polish and Russian the respective terms are also used as a case name (cf. Topolińska, 1999; Жеребило, 2010).

Complex semantic relations occurring between terms and varying terminological conventions do not alter the fact that the lexical subsystem is characterized by the pursuit of systematic organization. Terms that become ambiguous sometimes "wear out" and gradually become obsolete, see e.g. the abandonment of the Polish term *agens* (Eng. *agent*) in the works of M. Korytkowska (cf. Korytkowska, 1992). Potential units often remain only potential in the absence of clear nominative need. The observation of this state of affairs leads to the trivial conclusion that linguists are expected to be competent in the field of linguistic terminology. Languages may differ greatly and conclusions based on monolingual material are often not representative for multilingual purposes. Even closely related languages are characterized by lexical asymmetry. The traditional approach of source and target language may not result in a complete picture of the target language. An important novelty in the works on iSybislaw is the rejection of such an approach (i.e. projecting one language onto another). This results in the parallel research of confronted

languages. The following table shows the relation of synonymy for three different languages. In these sequences one should also distinguish certain pairs of terms being combinatorial variants. The table also includes potential units (crossed out expressions).

Table 5. Eng. *nominal phrase*

Polish	Bulgarian	Russian
grupa nominalna	*номинална група*	~~*номинальная группа*~~
grupa imienna	*именна група*	*именная группа*
fraza nominalna	*номинална фраза*	*номинальная фраза*
fraza imienna	*именна фраза*	*именная фраза*
syntagma nominalna	*номинална синтагма*	*номинальная синтагма*
~~*syntagma imienna*~~	*именна синтагма*	*именная синтагма*

In IR the distinction between synonymy and variantivity seems irrelevant. In both cases, different language forms express the same content and search engine optimization requires combining them in one equivalence class. There are various views on variantivity on the level of morphemes and word formation, which forces us to ask the question about the nature of the relationship between complex terms in which one of the elements is interchangeable with a functionally identical element (see above). The systematic character of such phenomena allows to predict the so-called potential units. Synonymy (being a lexical phenomenon) is more irregular. A separate problem is the possibility that variants of the same term in different languages differ in their nature (e.g. phonetic vs. inflectional), cf. Russian *алломорф/ алломорфа* (Eng. *allomorph*) and Polish *allomorf/alomorf*. True variantivity is a rarity in the terminological subsystem, however.

A separate problem is also a kind of ambiguity of terms resulting from their different definitions and the application of various research methods. Such terms as *określoność* (Eng. *definiteness*) in S. Karolak's works (cf. Karolak, 2001) have a different meaning and scope than in the works of V. Koseska (cf. Koseska-Toszewa, Korytkowska, & Roszko, 2007). In the case of S. Karolak it can be considered synonymous with Pol. *intensjonalna zupełność* (Eng. *intensional completeness*), in other terms with *uniqueness* and *generality*.

V. Koseska does not use *intensional completeness* as a term not due to idiolectal preferences discussed above. The absence of the term is motivated by a different research method implemented in her works in which Pol. *określoność* (Eng. *definiteness*) is understood more narrowly and does not cover *ogólność* generality (generality is considered indefinite in works based on the quantificational model sic!). Distinguishing two meanings for each of the two following terms Pol. *określoność* (Eng. *definiteness*) and Pol. *nieokreśloność* (Eng. *indefiniteness*) in the case of an IR system such as iSybislaw seems a bit far stretched, however.

It seems that true synonymy in terminology is problematic because definitions vary in different works (even of the same author) and establishing it requires a depend research. In IR, when creating synonymy/equivalence classes (multilingual and/or including variants), the depth of analysis should be restricted to a more moderate level. It is preferable for the user to receive a complete set of information even at the cost of obtaining some redundant (from his point of view) data. The optimization of IR requires some compromises, but (unfortunately) there are no shortcuts and every case should be analyzed separately.

Bibliography

Bojar, B. (2002). *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich. (Nauka, Dydaktyka, Praktyka; 56).

Karolak, S. (2001). *Od semantyki do gramatyki*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Kiklewicz, A. (2003). Podmiot i orzeczenie jako kategorie gramatyki funkcjonalnej. *Prace Językoznawcze*, (5), 117–139.

Korytkowska, M. (1992). *Gramatyka konfrontatywna bułgarsko-polska* (Vol. 5/1: *Typy pozycji predykatowo-argumentowych*). Warszawa: Slawistyczny Ośrodek Wydawniczy.

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Warszawa: Dialog.

Kubiszyn-Mędrala, Z. (2000). Czasowniki nieosobowe (niewłaściwe), ich miejsce w polskim systemie morfologiczno-syntaktycznym. *Biuletyn Polskiego Towarzystwa Językoznawczego*, *58*, 95–104.

Ledzion-Jelen, M. (2009). Sposoby oddawania macedońskiej kategorii прекажаност w języku polskim i niemieckim. In M. Cichońska (Ed.), *Kategorie w języku, język w kategoriach*

(pp. 130–144). Katowice: Wydawnictwo Uniwersytetu Śląskiego. (Prace Naukowe Uniwersytetu Śląskiego w Katowicach; 2631).

Polański, K. (Ed.). (1999). *Encyklopedia językoznawstwa ogólnego*. Wrocław (etc.): Zakład Narodowy imienia Ossolińskich.

Rudnik-Karwatowa, Z. (1995). Wczoraj, dziś i jutro informacji dokumentacyjnej językoznawstwa slawistycznego. *Zagadnienia Informacji Naukowej*, (1–2), 63–67.

Rudnik-Karwatowa, Z. (2002). Język informacyjno-wyszukiwawczy dokumentacyjnego systemu językoznawstwa slawistycznego. Doświadczenia z realizacji projektu. In Z. Rudnik-Karwatowa (Ed.), *Językoznawstwo. Prace na XIII Międzynarodowy Kongres Slawistów w Lublanie 2003* (pp. 207–212). Warszawa: Komitet Słowianoznawstwa PAN. (Z Polskich Studiów Slawistycznych; seria 10).

Rudnik-Karwatowa, Z., Mikos, Z., & Bojar, J. (2007). Nowoczesny system informacji slawistycznej. Zadania, dotychczasowe wyniki i perspektywy. *Zagadnienia Informacji Naukowej*, (2), 19–40.

Topolińska, Z. (1999). *Język, człowiek, przestrzeń*. Warszawa; Kraków: Towarzystwo Naukowe Warszawskie.

Аврамова, В. (2007). Строение человека в болгарской языковой картине мира (на фоне русской языковой картины мира). In Т. Иванова (Ed.), *Проблемы когнитивного и функционально-коммуникативного описания русского и болгарского языков* (Vol. 5, pp. 153–169). Шумен: Университетско издателство Епископ Константин Преславски.

Ахманова, О. С. (1966). *Словарь лингвистических терминов*. Москва: Советская энциклопедия.

Жеребило, Т. В. (2010). *Словарь лингвистических терминов*. Назрань: Пилигрим.

Косеска-Тошева, В. (2001). Функции и формите на перфекта в българските диалекти. In В. Радева (Ed.), *Българският език през XX век* (pp. 130–138). София: Академично издателство Проф. Марин Дринов.

Немченко, В. Н. (2008). *Введение в языкознание*. Москва: Дрофа.

Bibliography (transliteration)

Akhmanova, O. S. (1966). *Slovar' lingvisticheskikh terminov*. Moskva: Sovetskaia entsiklopediia.

Avramova, V. (2007). Stroenie cheloveka v bolgarskoĭ iazykovoĭ kartine mira (na fone russkoĭ iazykovoĭ kartiny mira). In T. Ivanova (Ed.), *Problemy kognitivnogo i funktsional'no-kommunikativnogo opisaniia russkogo i bolgarskogo iazykov* (Vol. 5, pp. 153–169). Shumen: Universitetsko izdatelstvo Episkop Konstantin Preslavski.

Bojar, B. (2002). *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich. (Nauka, Dydaktyka, Praktyka; 56).

Karolak, S. (2001). *Od semantyki do gramatyki*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Kiklewicz, A. (2003). Podmiot i orzeczenie jako kategorie gramatyki funkcjonalnej. *Prace Językoznawcze*, (5), 117–139.

Korytkowska, M. (1992). *Gramatyka konfrontatywna bułgarsko-polska* (Vol. 5/1: *Typy pozycji predykatowo-argumentowych*). Warszawa: Slawistyczny Ośrodek Wydawniczy.

Koseska-Tosheva, V. (2001). Funktsii i formite na perfekta v bŭlgarskite dialekti. In V. Radeva (Ed.), *Bŭlgarskiiat ezik prez XX vek* (pp. 130–138). Sofiia: Akademichno izdatelstvo Prof. Marin Drinov.

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Warszawa: Dialog.

Kubiszyn-Mędrala, Z. (2000). Czasowniki nieosobowe (niewłaściwe), ich miejsce w polskim systemie morfologiczno-syntaktycznym. *Biuletyn Polskiego Towarzystwa Językoznawczego*, *58*, 95–104.

Ledzion-Jelen, M. (2009). Sposoby oddawania macedońskiej kategorii prekažanost w języku polskim i niemieckim. In M. Cichońska (Ed.), *Kategorie w języku, język w kategoriach* (pp. 130–144). Katowice: Wydawnictwo Uniwersytetu Śląskiego. (Prace Naukowe Uniwersytetu Śląskiego w Katowicach; 2631).

Nemchenko, V. N. (2008). *Vvedenie v iazykoznanie*. Moskva: Drofa.

Polański, K. (Ed.). (1999). *Encyklopedia językoznawstwa ogólnego*. Wrocław (etc.): Zakład Narodowy imienia Ossolińskich.

Rudnik-Karwatowa, Z. (1995). Wczoraj, dziś i jutro informacji dokumentacyjnej językoznawstwa slawistycznego. *Zagadnienia Informacji Naukowej*, (1–2), 63–67.

Rudnik-Karwatowa, Z. (2002). Język informacyjno-wyszukiwawczy dokumentacyjnego systemu językoznawstwa slawistycznego. Doświadczenia z realizacji projektu. In Z. Rudnik-Karwatowa (Ed.), *Językoznawstwo. Prace na XIII Międzynarodowy Kongres Slawistów w Lublanie 2003* (pp. 207–212). Warszawa: Komitet Słowianoznawstwa PAN. (Z Polskich Studiów Slawistycznych; seria 10).

Rudnik-Karwatowa, Z., Mikos, Z., & Bojar, J. (2007). Nowoczesny system informacji slawistycznej. Zadania, dotychczasowe wyniki i perspektywy. *Zagadnienia Informacji Naukowej*, (2), 19–40.

Topolińska, Z. (1999). *Język, człowiek, przestrzeń*. Warszawa; Kraków: Towarzystwo Naukowe Warszawskie.

Zherebilo, T. V. (2010). *Slovar' lingvisticheskikh terminov*. Nazran': Piligrim.

Synonymy and search synonymy in an IR system (on the basis of linguistic terminology and the iSybislaw system)

Summary

The paper focuses on some problems of synonymy in the linguistic terminology and solutions for the optimal representation of information in the structure of an IR language. Linguistic terms in addition to metalinguistic meaning also carry some meta-scientific information (e.g. on the methodological school). It is thus possible that two different terms refer to the same linguistic phenomenon within various research trends. The issue of usage is also addressed here (including idiolectal preferences). The above phenomena on the one hand and various user information needs on the other result in some significant difficulties in the work on the optimization of IR in iSybislaw.

Keywords: information retrieval system; iSybislaw; linguistic terminology; search synonymy; Slavic languages; synonymy

Słowa kluczowe: języki słowiańskie; syonimia; synonimia wyszukiwawcza; system informacyjno--wyszukiwawczy; terminologia