COGNITIVE STUDIES | ÉTUDES COGNITIVES, 13: 13–23
SOW Publishing House, Warsaw 2013

DOI: 10.11649/cs.2013.001

**MARK KIT**[1,A] & **VIOLETTA KOSESKA-TOSZEWA**[2,B]
[1]Language Interface Inc., New York
[2]Institute of Slavic Studies, Polish Academy of Science, Warsaw
[A]mark.kit@langint.com ; [B]amaz@inetia.pl

# DIALOG BETWEEN A LEXICOGRAPHER AND A TRANSLATOR

**Abstract**

The discussion between the authors of the paper concerns the most pressing issues encountered in natural language semantics, as well as in corpus linguistics and computational linguistics. A broad range of knowledge, allowing linguists and information scientists to work together, is required in these areas. The paper describes some primary problems of human and machine translation caused by gaps between different fields of knowledge. The authors suggest that interdisciplinary approach is required when it comes to contrastive  studies in linguistics.

**Keywords:** reversibility of dictionaries, lexical structures, multilingual dictionaries, perfective aspect, semantic tags, parallel corpora, corpus linguistics, computational linguistics, contrastive linguistics.

**Violetta Koseska**: To a large degree, the development of science is driven by cross-disciplinary research. It requires researchers who work in different fields to deeply understand new issues they face and, of course, respect each other.

My partner in this discussion is a translator, engineer and researcher studying problems of translation quality and efficiency assurance. Mark Kit has a great deal of hands-on experience in electronics, automated control and radio communications. He is certified as a NASA Mission Control Center Interpreter and in that capacity has supported 11 space missions at the flight controller's console. He also worked as an interpreter at numerous conferences and technical meetings. Mr. Kit is certified by the American Translators' Association, by the Department of Social and Human Services of the State of Washington and other institutions. He is a founder and CEO of Language Interface Inc. (USA), a company that does high-quality translations for international projects.

**V.K.**: Very often we happen to hear about advantages of electronic dictionaries. However, many of us prefer to use traditional printed dictionaries. What do you think about it?

**Mark Kit**: I do not claim that an electronic dictionary is better than a printed one. I am only saying that these are different products that are intended for different purposes and, therefore, it does no good to compare them against each other. Instead one should determine the most advantageous application of each of them or, even better, try to use the strongest features of each of them in order to create an optimal tool.

A printed dictionary is a collection of lexical units in the form of a snapshot at a certain point in time (i.e. the date when it was printed) and in that capacity it can (and often does) serve as a source of linguistic norm, which is critically important, for instance, for linguistic expert evaluation or in selection of an appropriate use of language..

But the most important is the fact that the printed dictionary is a primary lexical source for creating the content of electronic dictionaries and the quality of the latter is manly driven by the quality of that source.

In contrast, the content of electronic dictionaries, under which we mainly mean online dictionaries, can be changed any minute, especially if general public has access to editing or adding lexical entries in it (which is often the case since the owners of such dictionaries seek to expand them as much as possible). For that reason the correctness or legitimacy of lexical entries in those dictionaries is quite questionable. I am not aware of any online dictionary that could be recommended as a source of lexical norms.

On the other hand, the capacity of online dictionaries to be updated any time opens and opportunity to keep them up-to-date, capable to reflect the lexical content of the today's language. In case of cross-language dictionaries it means that this tool can offer the most current views on translation of lexical units.

Translation activity today calls for high performance and for that purpose online dictionaries have a great advantage, offering fast search for potential translations of the lexical units (words or combinations of words) entered by the translator. According to our experiments, translators access the online dictionary 45 times per hour in average. At this frequency searches in printed dictionaries result in tremendous loss of performance (up to 74%) (Kit, M. 2010).

Besides, when working with printed dictionaries translators have to use separate books for translations in both directions (e.g. one book for translations from Russian to English and another one for translations from English to Russian), while in the online dictionary it can be done without the need to leave the website.

**V.K.**: You are saying that the traditional dictionaries are irreversible. What do you mean by that?

**M.K.**: From the translator's standpoint, a reversible dictionary can be used for translations to both directions, i.e. a reversible Polish-Bulgarian dictionary is equally good for translations from Polish into Bulgarian and from Bulgarian to Polish. Printed dictionaries do not have this feature and that is why Russian-English dictionary, for instance, is a separate publication than an English-Russian dictionary. Emergence of electronic dictionaries made it possible to find translations in any direction within the same program or website.

However, this is not the same as full reversibility, which is a capacity to provide round-trip translations. The round-trip translation is a subsequent events of translation from one language to another and then backwards without distortion of the meaning. For example, it would be translation from Russian into English, and then the resulting text is translated back into Russian and so on. This method is sometimes used to test the results of machine translation — and always unsuccessfully. For example, a translation is looking for a Russian equivalent of the word *craft* (meaning *ремесло (artificer's skill*, *искусство (art)*) in the automatic translation system Google Translate. Here are the results:

*Craft — мастерство — skill — развивать — develop* and so on, so each new translation attempt is exposed to a great deal of risk to obtain a new meaning that is different from that of the source. It is even worse when trying to run a round-trip translation of combination of words or phrases:

*Springs were running all over the hill — Пружины бегали по всему холму — The springs were running around the hill — Источники бегали в гору — The sources ran up the hill — Источники побежал вверх по склону.*[1]

Structures of the existing dictionaries do not allow for full reversibility of translations they provide.

**V.K.**: But why would translators need round-trip translations and full reversibility?

**M.K.**: This is often necessary when working on a translation. A translator, even an experienced one, does not always feel subtle flavors of the foreign language. But he either has to understand the exact meaning of the unit in question (when translating from a foreign language) or find an exact equivalent of the unit (when translating into a foreign language). And if the translator does not have excellent skills (which is often the case), he is challenged by the need to select the right translation from the list offered by the dictionary. It is often a good technique to use reverse translation to check the choice of translation.

For example, when translating the word *тонуть (sink, drown)* into English one has to make a choice between potentially correct options *sink* and *drown*. What choice shall be made in the phrase "*Я тону в море Вашего милосердия*[2]" — *drown* or *sink* (the options offered by the dictionary)? The speaker of foreign language may not feel it. Or in the phrase: *"Она утонула в своем горе"*[3]? In order to make the choice one has to find examples of usage, or the exact description of the meaning, or, best of all — both. In these situations reverse translation helps, as well corpora, especially the parallel ones.

Another application of reversible lexical systems would be automatic verification of translations. Such system would allow for checking automatic or human-produced translations by converting the translated texts to the original language and comparing the resulting text with the original source. At the present time this

[1] This experiment was run on the dictionary/machine translation system `http://translate.google.com/`

[2] *I am sinking in the sea of your grace.*

[3] *She drowned in her grief.*

is performed by humans, which costs large amounts of time and money spent for producing reverse translations. But the main shortcoming of this method is the fact that no two translations done by humans are the same, since they bear specifics of unique minds of the translators. Because of that reverse translations done by humans almost never match the source text on the word level.

**V.K.**: What do you see as primary roadblocks on the way to reverse dictionary creation? And are there such roadblocks at all?

**M.K.**: Irreversibility is inevitable under the traditional way to view lexical data as a tree-like structure with the symbolic representation of the unit in its root (symbolic form).

test

1. *n*

**1)** испыт**а**ние;

to put to test подверг**а**ть испыт**а**нию;

to bear the test в**ы**держать испыт**а**ние;

to stand the test of time в**ы**держать испыт**а**ние вр**е**менем

**2)** мер**и**ло; крит**е**рий

**3)** пров**е**рочная, контр**о**льная раб**о**та;

a test in English контр**о**льная раб**о**та по англ**и**йскому язык**у**

**4)** *психол.* тест

**5)** *хим.* иссл**е**дование, ан**а**лиз; пров**е**рка;

a test for the amount of butter in milk определ**е**ние ж**и**рности молок**а**

**6)** *хим.* реакт**и**в

**7)** *attr.* испыт**а**тельный, пр**о**бный; контр**о**льный, пров**е**рочный;

test station контр**о**льная ст**а**нция

2. *v*

**1)** подверг**а**ть испыт**а**нию, пров**е**рке; исп**ы**тывать

**2)** *хим.* подверг**а**ть д**е**йствию реакт**и**ва

**3)** производ**и**ть **о**пыты

Data structure in this entry looks like this:

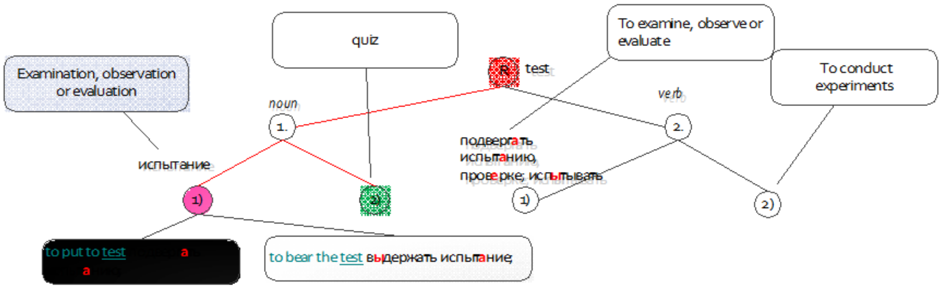
quiz
To examine, observe or evaluate
Examination, observation or evaluation
R
test
To conduct experiments
noun
verb
1.
2.
подвергать испытанию, проверке; испытывать
испытание
1)
2)
1)
2)
to put to test
a
a
to bear the test выдержать испытание;

**Figure 1**. Traditional dictionary data structure.

Apparently, each meaning (shown in the diagram as a leaf of the tree) can be reached only from the root, while each root (the symbol) is linked to multiple meanings and the translator has to decide which of them is the correct meaning that suits a specific textual situation.

When building multilingual dictionaries the situation aggravates. Here is an example of a lexical entry taken from a Polish-Bulgarian-Russian dictionary described in work (Koseska-Toszewa, Satoła-Staśkowiak, Duszkin 2012).

| **Russian** | **Bulgarian** | **Polish** |
|---|---|---|
| **Покупать, -ю, -ешь** vi, state, transitiv<br>1. ‘purchase something at a certain price, for a certain amount of money’ покупать клубнику<br>2. met. (кого-либо) ‘secure somebody’s support, favour for oneself by giving them money, gifts, a bribe’ покупать чиновника | **купувам, -аш, -а** vi., state, transitiv<br>1. ‘purchase something at a certain price, for a certain amount of money’ да купувам ягоди<br>2. met. ‘secure somebody’s support, favour for oneself by giving them money, gifts, a bribe’ да купувам чиновника | **kupować, -uję, -ujesz** vi, state, transitiv<br>1. ‘purchase something at a certain price, for a certain amount of money’ kupować truskawki<br>2. ---- no meaning (see przekupywać) |

It can be seen in this table that 2 meanings of the symbolic form of the Russian word “покупать” (to buy, to purchase) correspond to 2 meanings in the Bulgarian language and 1 meaning in the Polish language. The diagram below shows that the first meaning, shown in the dictionary, is connected to 3 symbolic forms (Ru, Bu, Pl), while the second meaning is connected only to 2 forms (Ru-Bu). In Polish the third meaning is represented by another symbol, so it is not missing on the language, however the one-to-one relationship is lost.

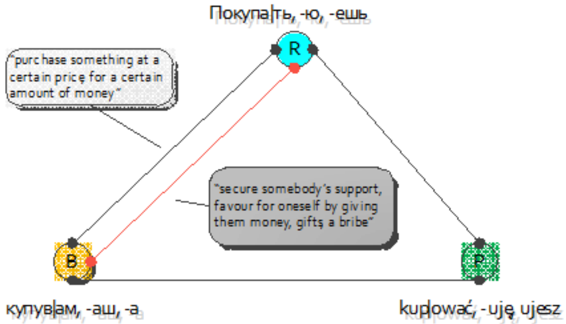
Покупа|ть, -ю, -ешь
R
"purchase something at a certain price, for a certain amount of money"
"secure somebody's support, favour for oneself by giving them money, gifts, a bribe"
B
P
купув|ам, -аш, -а
kup|ować, -uję ujesz

**Figure 2**. Linkage between lexical entries in a multiple-language dictionary.

This irreversibility is due to the fact that the symbolic representation of the word can be linked to several meanings due to homonymy or polysemy. When a translator looks for the translation, the dictionary returns a list of all possible translations or definitions of this form. Selection of any form out of this list shifts the user to a new set of meanings that this form is linked to.

This irreversibility can be overcome by moving the meaning to the root of the tree data structure. This structure "makes semantically equal" all forms that represent the meaning $M$. Now, in order to translate any form that represents the meaning $M$ is to climb up to the root of the tree and then move down to the branch that represents the desired language.

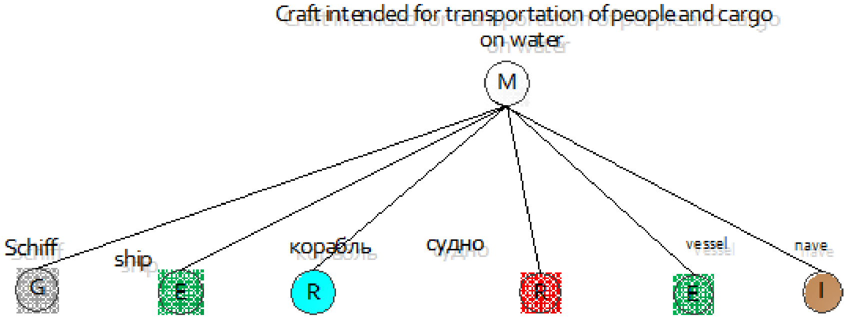
Craft intended for transportation of people and cargo on water
M
Schiff
ship
корабль
судно
vessel
nave
G
E
R
R
E
I

**Figure 3**. Meaning-based dictionary structure.

**V.K.**: It seems that you support my concept stating that dictionaries shall be based on meanings and not forms.

**M.K.**: By all means! My experience in translation makes me confident that a real translator almost always is looking for meaning and only then expresses that meaning in a suit of forms (words or expressions). Search for an exact form of a meaning is also required, e.g. when looking for a synonym that would convey that meaning, but it is always proceeded by clear understanding of the meaning. No search will help if the translator does not grasp the meaning.

**V.K.**: I would like to get back to the problem of distinguishing between forms and their meanings in dictionaries and corpora. This is an open issue while, no doubt, without solving it we cannot obtain high quality translations, especially in the machine translation area.

If, for instance, we are looking for forms of imperfective and perfective verbs in a parallel corpus where one of the languages is English, we won't find any, because in the English language verbs do not have the perfective aspect.

Modern corpus and computational linguistics are based mostly on forms existing in the language. In some researches computational linguists do not distinguish between forms and meanings at all, in others they think that they do, but is it really so? I have come across works on computational linguistics where the researchers put imperfective and perfective verbs as forms in one place and indicate the same in another place as the meaning. And the problem is not that in such occasions we have the same term for both the form and the meaning. The thing is that there is no meaning in the "spot" assigned for the meaning!

There are monographs by linguists studying semantics of the perfective aspect *and* the tense, however, the outcome of those works unfortunately have not been included in the conventional grammars while these grammars are exactly what the computational linguists use. This explains why there is no meaning of the aspect and tense the in the place designated for meanings. That is why cross-disciplinary studies are needed — to remove issues of this kind.

In the Bulgarian-Polish comparative grammar, based on the math model of the process theory (Petri net theory), we believe to have exactly identified aspect-temporal meaning of imperfective/perfective forms in the Slavic languages (Koseska-Toszewa, Penčev (Eds.) 1988–2008), (Koseska 1982), (Koseska-Toszewa 2011), (Koseska, Mazurkiewicz 2010), (Koseska, Dimitrova, Roszko, (Eds.) 2009), (Dimitrova, Koseska, Garabík, Erjavec, Iomdin, Shyrokov, 2010, Mondilex).

We have proved that the primary meaning of an imperfective symbolic form is either a state (state1) or a sequence of states and events resulting in a state (state2). While the primary meaning of an perfective symbolic form is an event (event1) or a sequence of states and events resulting in an event (event2) (Mazurkiewicz 1986), (Koseska, Gargov, 1990), (Koseska, Mazurkiewicz 2010).

That is why we suggest incorporating in both dictionaries and corpora the new semantic classifiers state1, state2, event1, event2 wherever information scientists enter imperfective or perfective form in place of the meaning.

Another example is from the semantics of linguistic category of definiteness/indefiniteness.

Traditional grammars assume that the Polish and the Russian languages do not have this category because they are synthetic languages and have no morphological means to express definiteness or indefiniteness. Contractive semantics, however, shows that the languages although using different ways to express this category, do have any means for that.

In the Polish and the Russian languages this category is represented mainly through lexical means, while the Bulgarian language uses morphological means for that, the same way as in the English language where in noun phrases the definite and indefinite articles are used.

There is a variety of means to express this category. But is there a common understanding of what the notion of definiteness/indefiniteness is? My studies show that it is not so. To resolve this issue I had to start working on a description of the semantic category of definiteness/indefiniteness.

In my works I assumed that definiteness is a uniqueness of an object and a set while indefiniteness is an existentiality and generality. If we do not know what the uniqueness of an object/set is, what their existentiality/generality is, we are unable to find adequate means to express these types of quantification in terms of comparison for two or more languages.

The tradition assigned this category to noun phrases since the definite and indefinite articles are linked to the noun. As back as in 1982 I wrote that definiteness/indefiniteness is a phrase category and not a noun phrase category as it was assumed in the literature. Object/set uniqueness, existentiality and generality also are applicable to a verb phrase where these meanings are expressed through lexical and grammatical means. The meaning of the definite article in the Bulgarian language can be found only on the phrase level, not the noun phrase level (Koseska, 1982).

Compare: *Човекът, който те търсеше, излезе преди 5 минути.* "The man who was looking for you left 5 minutes ago" (quantified uniqueness in the noun phrase).

*Човекыт е смъртен* (People are mortal) (Each person is mortal) (quantified generality in the noun phrase).

*Само човекът е действително мислещо същество.* "Only a human is a truly thinking entity" (quantified uniqueness in the noun phrase).

**M.K.**: The English word *fish*, for example, can mean "*some fish*": "*I have a fish so I will make us dinner*", but it can also mean "*fish in general*": "*In the process of evolution fish came ashore*", that is this form means both uniqueness and generality.

**V.K.**: Indeed. On the level of a verb phrase adverbs as forms express uniqueness, existentiality and generality. For example, compare:

*Он гуляет по вечерам (He takes walks at night).* This phrase does not carry information on definiteness or indefiniteness. We do not know whether he always walks at night (generality) or he sometimes takes a walk at night.

On this issue you can see (Косеска, Гаргов 1990). In this case we can determine what quantified meaning in the verb phrase is by adding linguistic quantifiers of generality or existentiality: *always, sometimes*:

He *always* takes walks at night (general quantified meaning of the verb phrase).

He *sometimes* takes walks at night (existential quantified meaning of the verb phrase).

We call the meanings discussed here *quantified* because they are ordered in the logical quantified model that enables us to write these exact meanings, distinguishing them from symbolic forms that express quantified uniqueness, generality and existentiality (see Koseska 1982, Koseska, Gargov 1999).

The quantified model enables us to determine the meaning of linguistic means related to quantification in each natural language.

Our studies will help us to incorporate quantification tags into corpora and dictionaries. And, for example, lexical units like: *always, sometimes, every now and then,* etc. will be needed in the place assigned for forms as adverbs, while in the place assigned for the meaning we can enter the information on their area of quantification. As soon as we can separate the form from the meaning, it will become possible to search dictionaries and corpora for meanings and not only forms.

**V.K.**: All efforts of information scientists are focused mostly on formal tagging of corpora. Very few solutions have been proposed for semantic tagging of parallel corpora and this impedes the improvement of machine translation, so demandable in the modern world. This issue has been discussed (Dimitrova, Koseska-Toszewa

2012), (Dimitrova, Koseska-Toszewa, D. Roszko and R. Roszko 2009), (Dimitrova, Koseska-Toszewa, Roszko, D. and Roszko, R. 2011).

**M.K.**: Parallel corpora is a powerful tool for translators. Let's take a look at the application of Computer-aided Translation (CAT) tools. These systems search for previously translated phrases or segments of the same and suggest the translator to use them. These segments are stored in databases called translation memories (TM). Parallel corpora are a valuable source for creation of these databases. Studies show that when using CAT tools translators can save more than 20% of the text volume to be translated (Kit 2011).

Another application of parallel corpora is searches for examples of usage of words, terms or expressions. Availability of such examples is an very important for producing high-quality translations. One can find many examples in the Internet, but there are no guarantees of their quality, while a corpus provides standards of usage.

Additionally, being a normative source the corpora, as well as the dictionary, can affect solutions made in the course of linguistic expert reviews, including forensic studies. And, certainly, it is hard to overestimate the scientific value of parallel corpora. There is hardly another tool of this magnitude of capabilities for studies in the field of comparative linguistics.

**V.K.**: What is a connection between parallel corpora and translation or bilingual/trilingual dictionaries?

**M.K.**: Unfortunately I have not knowledge on any efforts to arrange such a connection, which could open new opportunities in lexicography. For example, it could help to build automatic generation of bilingual and multilingual dictionaries through processing of parallel corpora. Automatic translation systems and search engines could use links between components of parallel corpora and dictionaries.

For translators the integration of corpora with the dictionaries would be extremely useful. As a rule, dictionaries have very few examples of usage of lexical units. The integrated corpus-dictionary system would enable users to find not only the required translations of lexical units, but also examples of their normative usage. I think this is a promising field.

**V.K.**: What would you recommend to the developers of dictionaries to facilitate transfer of their entries to electronic format?

**M.K.**: This is a very important and pressing issue. The thing is that many dictionaries are built in such a way that their conversion into electronic format (using the concepts I just talked about, i.e. through creation of source-target semantic pairs) is very time consuming. Here is an example of a typical entry in a dictionary:

> **адрес** *м* **1**. *(пощенски)* адрес; **пиши на този ˜** пиши по адресу; **сбъркал съм ˜а** я ошибся адресом; не тот адрес; не туда попал **2.** *(писмено поздравление)* адрес ◊ **говоря по ˜ на *н-го*** говорить в адрес *(Р)*; *прен* злословить.

In this record it is very difficult to separate attributes from lexical units, since the entries are not uniform and not only in terms of their content, but also their structure is different, so a computer cannot exactly determine what categories the attributes and lexical units should be assigned to.

If a certain place is assigned to every meaningful fragment of the entry or it would be marked with some identifiers, it would be easy to automatically transfer the content in an electronic database. For example, in the standard XML format the entry could look like this:

<lex_unit><lex-BU>Адрес</lex-BU><tag1>пощенски</tag1><example>пиши на този</example></lex_unit>.

Here the boundaries of the entry are marked with the tags <lex_unit> and </lex_unit>, other fragments are also tagged. For a linguist this way of making a record may seem to be sophisticated, but for them a simple graphical interface can be made, offering them a convenient tool while taking care of the tagging on the background. But even without such an interface a lexicographer can use such well-known systems as Excel spreadsheets, entering lexical units and their attributes in the appropriate cells of the table and then automatically convert the table into a database content. Even such a simple method allows to avoid tremendous efforts that otherwise should be made to manually populate the database with lexical entries.

**V.K.**: Your comments are very interesting and useful for contemporary lexicographers and linguists. Thank you very much!

## References

Dimitrova, L., Koseska-Toszewa, V. (2012). Bulgarian-Polish Parallel Digital Corpus and Quantification of Time, *Cognitive Studies / Études Cognitives, 12*: p. 199–207.

Dimitrova, L., Koseska-Toszewa, V., Garabík, R., Erjavec, T., Iomdin, L., Shyrokov, V. (2010). *Conceptual Scheme for a Research Infrastructure Supporting Resources in Slavic Lexicography.* Sofia, Demetra Ltd. Publisher.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D. & Roszko, R. (2009). Bulgarian-Polish-Lithuanian Corpus — Problems of Development and Annotation. In: Erjavec, T. (Ed.), *Research Infrastructure for Digital Lexicography. MONDILEX Fifth Open Workshop. Ljubljana, Slovenia, October 14–15, 2009, Proceedings of the 12th International Multiconference Information Society 2009*, Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, p. 72–86.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D. & Roszko, R. (2011). Bulgarian-Polish-Lithuanian Corpus — Recent Progress and Application. In: Majchráková, D., Garabík, R. (Eds.) Natural Language Processing, Multilinguality. ISBN: 978-80-263-0049-6, Slovenská akadémia vied Jazykovedný ústav Ludovíta Štúra, p. 30–43.

Koseska, V., (1982). *Semantyczne aspekty kategorii określoności/nieokreśloności (na materiale z języka bułgarskiego, polskiego i rosyjskiego)*, Wrocław.

Koseska-Toszewa, V., Dimitrova, L. & Roszko, R. (Eds.) (2009). *Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open Workshop, 29 June – 1 July 2009, Warsaw.* SOW, Warsaw.

Koseska-Toszewa, V., Gargov, G. (1990). *Bylgarsko-polska sypostavitelna gramatika*, vol. II. *Semanticznata kategorija opredelenost/neopredelenost*, BAN, Sofia.

Koseska-Toszewa, V., Mazurkiewicz, A., (2010). *Time Flow and Tenses*, SOW, Warsaw.

Koseska-Toszewa, V., Penčev, J. (Eds.) (1988–2008). *Gramatyka konfrontatywna bułgarsko-polska*, vol. 1–8., Sofia–Warszawa.

Koseska-Toszewa, V., Satoła-Staśkowiak, J., Duszkin, M. (2012). Polish-Bulgarian-Russian, Bulgarian-Polish-Russian or Russian-Bulgarian-Polish Dictionary?): *Cognitive Studies / Études Cognitives, 12*: p. 51–56.

Mazurkiewicz, A. (1986) Zdarzenia i stany: elementy temporalności, In: *Studia gramatyczne bułgarsko-polskie*, vol. I, *Temporalność*, Wrocław, p. 7–21.

Kit M. (2010). Кит, М., О стратегии построения высокоэффективных сетевых словарей. На примере разработки словаря LexSite. *Вестник РГГУ № 9*; Сер. «Языкознание», М.: РГГУ, 2010. pp. 151.

Kit, M. (2011). Кит, М., Об использовании повторяемости сегментов текста для повышения эффективности процесса перевода. *Вестник РГГУ № 11*; Сер. «Языкознание», М.: РГГУ, p. 196–208.