FROM THE PROBLEMS OF DICTIONARIES
AND MULTI-LINGUAL CORPORA

Abstract

The article describes the work on a number of dictionaries being developed by the Corpus Linguistics and Semantics Group of the Institute of Slavic PAS. They include “Contemporary Bulgarian-Polish Dictionary”, “Bulgarian-Polish Online Dictionary” and “Russian-Bulgarian-Polish Dictionary”. The dictionaries differ in the numbers of entries, as well as in the different degrees of their connection with parallel corpora being elaborated under the “Clarin” project. All the discussed dictionaries are similar with respect to their use of traditional, syntactic classifiers and of semantic classifiers, introduced for the first time in the existing lexicographical practice. Thanks to the “Polish-Bulgarian-Russian Corpus”, the Group has managed to verify the results of contrasting Polish and Bulgarian in the light of scope-based logical quantification. Thanks to the Russian material added to the trilingual corpus, the researchers have managed to confirm the fact that from the viewpoint of “incomplete quantification” Russian and Polish (synthetic languages) behave similarly, and are opposed to the analytic Bulgarian.

Keywords: Bilingual dictionary, trilingual dictionary, online dictionary, traditional classifier, syntactic classifier, semantic classifier, bilingual parallel corpora, trilingual parallel corpora, linguistic quantification, “incomplete quantification”.

The idea of creating a “Contemporary Bulgarian-Polish Dictionary” is closely related to the collective work of the Bulgarian linguists L. Andrejczin, L. Georgiev, S. Ilcev, N. Kostov, S. Stojkov, S. Todorov (a dictionary of 1973), and especially I. Lekov (1945) and I. Lekov, Fr. Slavski (Eds.) (1961), as well as the greatest Bulgarian-Polish dictionary by Fr. Slavski of 1987. Those outstanding linguistic studies gave rise to the idea of creating a separate work containing about 40,000 entries and introducing the most important, in the authors' opinion, grammatical and semantic information about Polish and Bulgarian verbs.
The book form of this dictionary will consist of 4 fascicles. The first of them should appear in 2013. As far as possible, the language material will be connected with the “Bulgarian-Polish Parallel Corpus” (authors: L. Dimitrova, V. Koseska-Toszewa) and with the “Polish-Bulgarian-Russian Parallel Corpus” (authors: V. Koseska-Toszewa, J. Satola-Staśkowiak, W. Sosnowski, A. Kisiel) (comments on that subject can be found in publications dating from 2009, 2011 and 2012, see the references. The authors do not take into consideration archaic and dialectal lexis, and try to introduce the maximum possible number of Polish, Bulgarian and Russian neologisms that are already commonly used in the selected languages. Beside traditional classifiers, the authors introduce into the entries new semantic classifiers, which help distinguish between language forms and their meanings. They concern first of all verb forms and adverbia. Such an approach to elaborating those parts of speech is a novum in the existing lexicographic practice.

The task of Joanna Satola-Staśkowiak is to prepare a thesaurus of neologisms, and to introduce the newest lexis into the dictionary. She is responsible for translating the meanings of Bulgarian entries to Polish and, for the correspondence between the Polish and Bulgarian meanings (Dimitrova, Koseska-Toszewa, Satola-Staśkowiak, 2009, 2012; Koseska-Toszewa, Satola-Staśkowiak, Duszkin, 2012).

V. Koseska-Toszewa verifies Bulgarian entries, checks if their translation to Polish is correct, defines semantic classifiers, and is responsible for their introduction into verb entries (Dimitrova, Koseska-Toszewa, 2009; Dimitrova, Koseska-Toszewa, Satola-Staśkowiak, 2009, 2012; Koseska-Toszewa, Satola-Staśkowiak, Duszkin, 2012; Dimitrova, Koseska-Toszewa, 2012), see the article by V. Koseska-Toszewa in this volume, see also the article by J. Satola-Staśkowiak in this volume.

L. Dimitrova is responsible for verifying Bulgarian entries and checking the usage frequency of Bulgarian lexems based on the “Bulgarian-Polish Corpus” (Dimitrova, Koseska-Toszewa, 2009; Dimitrova, Koseska-Toszewa, Satola-Staśkowiak, 2009, 2012; Dimitrova, Koseska-Toszewa, 2012).

The “Bulgarian-Polish Online Dictionary” is an experimental bilingual dictionary being developed by the authors of the “Contemporary Bulgarian-Polish Dictionary”. The dictionary numbers just about 5,000 entries. The task of preparing a proprietary computer program for handling the electronic Bulgarian-Polish dictionary has been undertaken by L. Panova and R. Dutsova. Besides adding traditional and semantic classifiers, the priority task of the authors of the “Contemporary Bulgarian-Polish Dictionary” in an online version is to include the newest Bulgarian lexis (being developed by J. Satola-Staśkowiak). The most important issue has become the way of examining an individual lexical unit, which consists in establishing whether it is fully accepted by the language system (the determinants in Bulgarian were, for example: the possibility of using the neologism with the article, the fact of possessing a plural form, and others). An appropriate verification of the newest lexems will protect the authors against the erroneous inclusion in the dictionary of, e.g., occasional expressions, whose life in the system of each language is limited in time.

The plan for developing a Russian-Bulgarian-Polish dictionary was set up two years ago by the Corpus Linguistics and Semantics Group at the Institute of Slavic of the Polish Academy of Sciences. The dictionary is a kind of lexicon containing
short definitions and a large number of meanings in all the three languages. According to the plan, the dictionary will contain about 10,000 entries for each of the three languages, and will number about 30,000 entries altogether. After long discussions, the final order of the described languages was selected. The dictionary starts with Russian due to the largest number of recipients knowing that language; Bulgarian is presented next, and then Polish. The above dictionary concerns languages representing the southern, eastern and western groups of Slavic languages, and will be of a large importance not only for slavists. According to the original plan, its authors were to be Maksim Duszkin, Violetta Koseska-Toszewa, Joanna Satola-Stackowiak (on the Polish side) and Anastasja Tzoneva (on the Bulgarian side). For reasons beyond our control, the composition of authors of the trilingual dictionary has been changed. At present, work on the dictionary is carried out by W. Sosnowski, V. Koseska-Toszewa and A. Kisiel. The first issue of the dictionary is to be published in 2014/2015. Examples:

<table>
<thead>
<tr>
<th>Русский</th>
<th>Bulgarian</th>
<th>Polish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>безимённый</strong> adj.</td>
<td><em>безимён</em></td>
<td><em>bezimienia</em></td>
</tr>
<tr>
<td>безимённая могила</td>
<td><em>безимённая могила</em></td>
<td><em>bezimienна могила</em></td>
</tr>
<tr>
<td>безимённый палец</td>
<td><em>безименній палец</em></td>
<td><em>bezimienны палец</em></td>
</tr>
<tr>
<td><strong>безысходный</strong> adj.</td>
<td><em>безъход</em></td>
<td><em>bezъход</em></td>
</tr>
<tr>
<td>безысходное положение</td>
<td><em>безъходното положение</em></td>
<td><em>bezъходно положение</em></td>
</tr>
<tr>
<td>see <strong>безъходность</strong></td>
<td>see <em>bezъходност</em></td>
<td></td>
</tr>
<tr>
<td><strong>бела</strong></td>
<td><em>побеля</em></td>
<td><em>pobilić</em></td>
</tr>
<tr>
<td>-ть, -я, -ешь</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Стена</em></td>
<td>see <em>niebo bieleje</em></td>
<td></td>
</tr>
<tr>
<td>see <strong>побелить</strong></td>
<td>see <em>zbieleć</em></td>
<td></td>
</tr>
<tr>
<td><strong>близи</strong> n. f. (Sg. Tantum)</td>
<td><em>близка</em></td>
<td><em>bliżka</em></td>
</tr>
<tr>
<td><em>снега</em></td>
<td>see <em>śniegu</em></td>
<td></td>
</tr>
<tr>
<td>see <strong>белый</strong></td>
<td>see <em>śnieżny</em></td>
<td></td>
</tr>
<tr>
<td><strong>близить</strong> -я, -ишь</td>
<td><em>ближить</em></td>
<td><em>pobilić</em></td>
</tr>
<tr>
<td><em>стену</em></td>
<td>see <em>śnieżną ścianę</em></td>
<td></td>
</tr>
<tr>
<td>see <strong>побелить</strong></td>
<td>see <em>pobilić</em></td>
<td></td>
</tr>
</tbody>
</table>

Examples:

- **безимённый** | *безимён* | *bezimienia* | 'having no name' |
- **безысходный** | *безъход* | *bezъход* | 'giving no hope for a positive result' |
- **бела** | *побеля* | *pobilić* | vi. (state, intransitive) | 'become white' |
- **близи** | *близка* | *bliżka* | 'blinding white colour of sth.' |
- **близить** | *ближить* | *pobilić* | vi. (state, transitive) | 'robić coś białym' |
As we have already mentioned, our dictionaries will employ syntactic classifiers with the abbreviations: transitive and intransitive, which indicate transitivity and intransitivity of a verbum. By transitivity we will mean the possibility of a direct object appearing in the sentence after the verb. In Polish and Russian sentences, a transitive verb is followed by a(n) in the accusative case. Intransitivity excludes the above possibility. Intransitive verbs cannot be followed by a direct object, and in Polish and Russian the verbum form is followed by all case forms except the accusative case. As our dictionary is not a valence dictionary, the above definition should be satisfactory for the reader.

Yet another group of classifiers will be semantic classifiers, which were not elaborated earlier, and were distinguished thanks to the many years of work on the first semantic confrontation in the word, carried out in the 12-volume monograph *Contrastive Polish and Bulgarian Grammar* with a semantic interlanguage (Koseska-Toszewa, 2006). The authors are aware that distinguishing semantic classifiers is not an easy task, and in order to achieve such a goal, one should consistently distinguish between the form and the meaning, e.g. decide whether the imperfect or perfect aspect used as a classifier in the commonly known dictionaries is a form of the verb or its meaning? In the authors’ opinion, this a form, and its meanings are either 1) states or sequences of states and events ending with a state, or 2) events or sequences of states and events ending with an event.

The notions of events, states and their configurations are understood here as in the network-based description of time and aspect, i.e. so that an event does not last in time (it begins, ends or interrupts states, while a state lasts and is begun or ended by an event, see in more detail the use of Petri net theory and its application in a natural language (Mazurkiewicz, 1986; Koseska-Toszewa, A. Mazurkiewicz, 1988, 2010; Koseska-Toszewa, 2006). Assuming that a verb form marked as “imperfective” is a form, we will present its meanings as either 1. state
From the Problems of Dictionaries and Multi-lingual Corpora

— a state or 2. — a sequence of events and states ended with a state, while a verb form marked as “perfective” will have a semantic classifier either 1. event, i.e. an event, or 2. a sequence of states and events ended with an event. Meanings 1 and 2 for an event and 1 and 2 for a state can be clearly illustrated on the example of an aspectually-temporal relation, i.e. when the verb form expresses a specific tense, due to which the infinitive form, which is “timeless”, can only be accompanied by the state and event abbreviations. Examples:

(Rus.) писать vi. (state, transitive) (Bulg.) писа vi. (state, transitive) (Pl.) писа vi. (state, transitive)

(Rus.) написать vp. (event, transitive) (Bulg.) напиша vp. (event, transitive) (Pl.) писа vp. (event, transitive)

(Rus.) жалеть vi. (state, transitive) (Bulg.) съжалявам vi. (state, transitive) (Pl.) żałować vi. (state, transitive)

Semantic classifiers concerning quantification of nomen and predicates are initially introduced in the “Polish-Bulgarian-Russian Parallel Corpus” (the corpus, being developed by the authors, is created within the European Project Clarin\footnote{Common Language Resources and Technology Infrastructure} — a scientific project, which in February 2012 obtained the European Commission legal status of ERIC (European Research Infrastructure Consortium). The founders of Clarin ERIC are Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Germany, the Netherlands and Poland. CLARIN is a project from the so-called ESFRI road map (European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures). The main goal of the project is to combine language resources and tools for European languages into a single, common and uniform network, which is to become an important working tool for scientists from the humanities in the broad sense of the term.

The discussed three dictionaries differ in the numbers of entries. The Bulgarian-Polish Online Dictionary contains a much smaller number of entries than the Contemporary Bulgarian-Polish Dictionary or the Russian-Bulgarian-Polish Dictionary, which follows from the fact that it is treated as a scientific experiment by the authors. However, there is a possibility of extending the language resources of the dictionary. The authors are planning to make use of this option, and introduce supplements in the subsequent years. The second largest dictionary is the Russian-Bulgarian-Polish Dictionary, which includes about 30,000 entries. The largest dictionary in the cycle is the Contemporary Bulgarian-Polish Dictionary, having about 40,000 entries.

The dictionaries differ also in the number of semantic classifiers. The Contemporary Bulgarian-Polish Dictionary is equipped in semantic classifiers concerning solely verbum. This is also the case with the Bulgarian-Polish Online Dictionary, where only verbum is characterized with help of semantic classifiers. We are planning that in the Russian-Bulgarian-Polish-Dictionary semantic classifiers will only concern verbum, but a classifier with quantification of adverbia will also be introduced in all three languages.

All the three dictionaries described in the article are connected with parallel corpora, and can be called “contemporary” due to the emphasis placed on the newest
and contemporary lexis in the language material under preparation, as well as due
to the analysis of the “usage frequency” verified by the parallel corpora mentioned
here. In lexicographic work, the most problems are posed by the trilingual Russian-
Polish-Bulgarian dictionary due to the methodological assumption concerning the
entries elaborated in all languages, saying that the material should be described in
a way enabling unproblematic establishment of a database for future dictionaries,
also electronic ones. This means, among others, that the entries should contain
meanings and classifiers in each of the languages rather than as is the case in
traditional bi- and trilingual dictionaries, where the meanings and classifiers concern
the initial language only.

By way of example, the semantic annotation applied results from a very com-
plicated process, see the article by V. Koseska-Toszewa in this volume. For this
reason, we could limit it to marking just the quantification of nomen at the sub-
ject position, and to specifying events — event 1, event 2 (sequence of events and
states, finally ended with an event) and states — stat1 and stat2 (sequence of states and events, finally ended with a state). It is also worthwhile to distin-
guish between adverbial forms and their quantifying meanings, see the table with
„zawsze” — adverbium with the universal quantificational meaning — given below.

By existentiality we mean here expressions of the form (\exists x)P(x) preceding
the predicate, i.e. a sentential function P (from now on, P), in the semantically-
logical structure of the sentence, read using the phrases “there is an x, such that”,
“for some x”. By universality we mean expressions of the form (\forall x)P(x) preced-
ing the predicate P in the semantically-logical structure of the sentence. Finally,
by uniqueness we mean an expression of the form (ix)P(x) which assumes that
the given sentential function (P) is satisfied either by exactly one element of the
considered universe, or by one and only one set of elements. As generally accepted
in logical literature by now, we treat the iota operator as the unique quantifier.
Quantificational expressions are not unambiguous (Koseska, Gargov, 1990). For
example, “each such that P” may be understood as “all elements satisfying P”,
which we can write in the form (\forall x)P(x). See Днес всяко момче кара ски.
(Dzisiaj każdy chłopiec jeździ na nartach.) [Today each boy is skiing]. In a context
different from the above, the expression written as (\forall x)P(x) can be understood
as “the set (of generally many elements) satisfying P as the only one”. Then this
quantificational meaning should be written as (Ix)P(X), and it would represent an
expression with a unique meaning. When encountering “incomplete quantification”,
we precede the quantification symbols with the question mark “?” — e.g. ?(Ix)Px)
—incomplete unique (iota-operator) quantification, see below the quantificational
meanings of adverbs referring to time and aspect of verbs. This is why we have to
do with quantification of the predicate rather than nomen here.

<table>
<thead>
<tr>
<th>Всегдa</th>
<th>Винаги</th>
<th>Zawsze</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv. (\forall X) P(X), univ.</td>
<td>adv. (\forall X) P(X), univ.</td>
<td>adv., (\forall X)P(X), univ.</td>
</tr>
</tbody>
</table>

See also (Tab. 1) semantic annotation taking into consideration quantification
of names at the subject position, and the aspectually-temporal meaning of the
Polish, Bulgarian and Russian verbs in the Bulgarian fable translated to Polish
and Russian Цар Безсънко from the Polish-Bulgarian-Russian Parallel Corpus.
Table 1.

<table>
<thead>
<tr>
<th>Polish</th>
<th>Bulgarian</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Król Bezsenek</strong></td>
<td><strong>Цар Безъяко</strong></td>
<td><strong>Царь Бессон</strong></td>
</tr>
<tr>
<td>Był sobie ((\exists X)P(X)) krasnoludek ((\exists x)P(x)) z długą białą brodą w czerwonej czapce.</td>
<td>Имаше някога ((\exists X)P(X)) едно джудже ((\exists x)P(x)) с дълга бяла брадица и алена папчица.</td>
<td>Жил некогда ((\exists X)P(X)) на свете гном ((\exists x)P(x)) с длинной белий бородой.</td>
</tr>
<tr>
<td>((\exists X)P(X)) (event) Pewnego wiosennego poranka ((ix)P(x)) krasnoludek jak zwykle przed swoim domkiem i począł liczyć ziarenka w swoim różańcu.</td>
<td>((\exists X)P(X)) Една про- летна утрин ((ix)P(x)) нашето джудже излезе както винаги пред вратата на своята къщица и започна да върти броеничката си от просене зърнца.</td>
<td>((\exists X)P(X)) (state) Както весенним утром ((ix)P(x)) гном по обыкновению сидал перед своей избушкой и перебирал четки.</td>
</tr>
<tr>
<td>((ix)P(x)) Myszka ((iX)P(X)) przyglądała ((state)) mu się bacznie błyszczącymi oczyma, oddychając ciężko.</td>
<td>((ix)P(x)) Мышка го гледаше ((state)) с малките си лъскави очички, без да мига, и дишаше тежко.</td>
<td>((ix)P(x)) Мышка ((iX)P(X)) (state) смотрела на него маленькими блестящими глазами и тяжело дышала.</td>
</tr>
<tr>
<td>((ix)P(x)) Krasnoludek ((iX)P(X)) (event) przerwał śpiew i wstał, aby przywitać nieznajomego gościa.</td>
<td>((ix)P(x)) Джуджето ((iX)P(X)) (event) преста- на да пее и се изправи на нозете си.</td>
<td>((ix)P(x)) Гном ((iX)P(X)) (event) перестал петь и встал.</td>
</tr>
</tbody>
</table>

Attention! The symbol "?" preceding logical symbols related to quantification represents incomplete quantification.

The material from our three-lingual parallel corpus reveals the fact, known from the „Bulgarian-Polish Contrastive Grammar”, that synthetic languages differ from analytic ones in having language means expressing especially unique quantification omitted in them, see (Koseska, Gargov, 1990; Koseska, 2006; Koseska, Korytkowska & Roszko, 2007). Let us remind here that, following Ajdukiewicz, such a phenomenon is referred to as “incomplete quantification” (Ajdukiewicz, 1974). As he wrote in his “Logika pragmatyczna”: “...an expression is called, for its certain meaning, a statement (or a sentence in the logical sense) if the expression, with just this meaning, utters some judgment, i.e. some thought referring in a reporting way to a certain state of things”. Ajdukiewicz called statements sentences in the logical sense. In his opinion, they had a certain common property distinguishing them from all other types of sentences. “Namely, all statements (and only them!)
are either true or false.” However, in a natural language we often encounter statements without either the true or the false value, such as, e.g., the sentence from “Logika pragmatyczna” by Ajdukiewicz: *Anglicy są flegmatyczni* [The English are phlegmatic]. Ajdukiewicz draws the reader’s attention to the fact that the above sentence represents incomplete quantification, which is why it has neither the true or the false value. Indeed, we do not know whether it refers to all the English, to their majority, or only to some of the English. Roughly speaking, the sender of the above information does not know whether the predicate “x are phlegmatic” is satisfied for the name “the English”. In this case, we need additional information, which can give the true value to this predicate. Namely, the information about scope-based quantification of the name “the English” is missing here. Ajdukiewicz assessed this phenomenon as error in articulation. Omitting of some component in a sentence prevents judging it as either true or false, and leads to misunderstanding. Incomplete quantification or relativization was deemed to lead to misunderstanding the speaker’s intention or to corrupting it. Clearly, a natural language has an effective defence against such a situation. Whoever does not understand the information sender’s intention, asks them a supplementary questions, and in the end clarifies any misunderstanding. However, there is still the question if “incomplete quantification”, which is the subject of our observations, is indeed a marginal problem — an ordinary error? Observations of research on the semantic category of definiteness / indefiniteness in Polish and Bulgarian show that incomplete quantification cannot be an ordinary language mistake, since languages differ among themselves in this respect. In Polish we encounter such phenomenon more often than in Bulgarian. Ajdukiewicz spoke of incomplete quantification having in mind the nominal phrase only, while we speak of incomplete logical scope-based quantification having in mind also the verbal phase in sentences of both the contrasted languages.

The languages studied earlier — Polish and Bulgarian — differ from each other with respect to incomplete quantification. This fact can be explained by differences in the morphological systems of both languages. In Polish this is a common phenomenon. Incomplete unique quantification is limited in both languages, but in Bulgarian it is not encountered on the nominal phrase level first of all because Bulgarian is an article language, see (Koseska, Gargov, 1990, p. 138–139), e.g.:

Pl. *Odwiedzilem staruszkę* (This only one, some, certain?) — incomplete quantification of the nominal phrase.

Bulg. *Посетих старица*, but it is impossible to build the sentence *Посетих една старица* with the same meaning. / *Odwiedzilem tę jedyną staruszkę* [I visited this unique old woman] (unique quantification).

*Bulg. Посетих една старица* / *Odwiedzilem pewną staruszkę* (existential quantification of the nominal phrase).

The examples from our corpus show that Russian behaves from the viewpoint of incomplete quantification just like Polish, and that synthetic languages (Polish and Russian) are in opposition to the analytic Bulgarian language with respect to „incomplete quantification”. See below examples isolated in this respect from
the “Polish-Bulgarian Russian Parallel Corpus”, which is being developed by the authors of this article under the European „Clarin” grant.

Authors of the article are aware that the plans discussed here require many years of work of the small Corpus Linguistics and Semantics Group of IS PAN.

References


