L. V. RYCHKOVA, A. YU. STANKEVICH

SPECIALIZED LANGUAGE DATA BASE INCORPORATING THE TEXTS OF THE REGIONAL BELARUSIAN NEWSPAPERS

Main principles of creation of the full-text database incorporating the texts of regional Belarusian mass-media being elaborated as an experimental basis within the framework of the research project «Functioning of the Belarusian language in bilingual regional mass media» are described. The choice of the bilingual regional newspapers texts as the authentic language material is justified and the annotation system is explained.

Keywords: linguistic resources, Belarusian language, interaction of languages, regional newspapers, full-text database, meta-annotation system, mixed-type corpus.

The creation of representative linguistic resources on an authentic Belarusian language material is a very urgent task for language planning in the Republic of Belarus. Firstly, de facto, the communicative matrix of the Belarusian language is not complete in terms of its real use, as Belarus provides an example of predominantly Russian-speaking region not belonging to Russia itself. Secondly, the issue of delimitation of authentic primary Belarusian-language texts from secondary texts of translations into Belarusian becomes rather difficult in practice, since as a rule only the texts already translated into Belarusian become apparent while the original texts in Russian remain behind the scene, since they serve as an intermediate language material in the creation of the final Belarusian texts. Thirdly, the real distribution of languages in various discourses in the complex context, which is comprised by the official bilingualism and the interaction of two closely related by their origin East-European languages as well as by the interaction of the appropriate linguocultures, requires investigation.

Within the framework of the research project «Functioning of the Belarusian language in bilingual regional mass media» being fulfilled in the frames of the State Programme of Scientific Researches of the Republic of Belarus, agreement A70-16, an original electronic language resource — a full-text database of regional Belarusian mass-media (RBM) is created, representing meta- and thematically annotated regionally distributed corpus of texts, which is in its final version should be regionally volume-balanced. As the main aim of the project is to study the peculiarities of the Belarusian language real usage in the regional newspapers of Belarus, RBM is planned to include the full texts of the newspapers, reflecting the interaction of the two state languages in the context of the Belarusian-Russian linguocultural community. Considering the purpose of revealing the regional peculiarities as well, namely those, which result from the influence of bordering languages and cultures, the mass media are chosen from all the regions of the Republic of Belarus, which border other countries, so the newspapers from Minsk region are not included.

It is known that the functioning of the language in a permanent contact with other language(s) has its own characteristics, which under certain conditions can be fixed in the usual norm, adapting to the functioning in different conditions, especially in the oral and written speech of bilinguals in the conditions of interaction of cultures (see, for instance,[Smułkowa, 2000]) and other numerous publications on the problem of language contact). From this point of view, the Belarusian language was not investigated, and the question of regional variation of its literary form was not raised at all. Nevertheless, this task is very relevant for the language building in the Republic of Belarus, and its solution is possible only by creation and use of the targeted representative language resources on the material of authentic texts as an experimental evidence base for such kind of the research.

We should mention here that in traditional academic lexicography the texts of newspapers publications have been considered as «sources of an unconventional type» [Korovanenko, 1995, p. 35] (here and further the citations were translated by L. Rychkova). However, in the 90s of the twentieth century, the attitude of lexicographers to the print media has changed dramatically. The inclusion of media in the range of the important language material for the creation of language resources is due to a number of factors, among which the following should be noted as the most significant: 1) reassessment of the «language standard»; 2) considering the newspapers as «inexhaustible donors of neology»; 3) the reflection of «standardized phrases ready for lexicographical use as words» in the texts of media; 4) a wide range of «expressive-stylistic means» presented in the texts of newspapers; 5) their function of the «conductor of terminology»; 6) the trend of «resurrection of obsolete vocabulary» explicit in the newspapers; 7) the function of the newspaper as the «supplier of regional geographic and international vocabulary» [Korovanenko, 1995, p. 35–37]. If we consider the Belarusian regional newspapers, one more feature is of utmost importance — that is the character of their bilingualism, when any author can chose Belarusian and / or Russian as a language of his / her publication on his / her own. This feature makes it unnecessary to translate any text and proves authenticity of the texts in Belarusian functioning in such newspapers. Besides, the texts, which reflect the interaction of languages in the same newspaper publication, represent «mixed language material» first differentiated during the procedure of the first mixed-type corpus creation, namely the illustrative linguistic corpus of Grodno region mass media [Rychkova, Stankevich, 2017]. This corpus based on the sources that included both authentic texts in Belarusian and Russian allowed to show the peculiarities of the Russian language in the bilingual (Russian-Belarusian) regional media of Belarus and designate the research direction for studying, besides Belarusian, regional features of the Russian language in Belarus as well. The illustrative linguistic corpus of Grodno region mass media is an open electronic (digital) resource and is freely accessible through the regional module (Corpus of regional and foreign press) of the media corpus in the frames of the Russian National Corpus. In order to be able to work only with the illustrative linguistic corpus of Grodno region mass media one must go to the search window of the regional module at http://ruscorpora.ru/search-regional.html and create the appropriate subcorpus using meta-characteristics of the country and region.

In the modern world, «the creation and effective use of information resources» is considered «as the most important factor in the social and economic development of mankind» [Antopolski, 2004, p. 8]. As a rule, information resources are understood as «separate documents and separate arrays of documents, documents and arrays of documents in information systems, data banks, other information systems» [Antopolski, 2004, p. 12]. Citing this definition, A. Antopolsky rightly notes that not always the data circulating in information systems can be identified only as documents, i.e. texts and data with a different structure, supplied with unique details for their search and distinguishing them among a number of other similar documents. Nevertheless, this is exactly the document approach that was taken as the basis for the development of meta-annotation for full-text electronic language resources. Combinations of tags, attributed to texts as information objects, allow identifying each specific text and, accordingly, select only such sets of texts (= information arrays) that possess a set of specific meta-text characteristics by tag values.

The infological model of RBM takes into account its complex nature as a type of full-text annotated electronic language resource, for which, taking into account the goal of the performed research, an important characteristic of the region, which, when implemented, creates the possibility of working with both the resource as a whole and with data sets representing the newspapers from specific region(s) has been provided.

The system of RBM meta-annotation represents the infological model of the resource and includes the following characteristics of the texts as search objects:

A. Structural markup (limited only to paragraphs);

B. Meta-markup, which includes the following set of characteristics (meta-parameters):

1. Author (an open list of values);

2. Title (an open list of values);

3. Date (an open list of values, though they should be given in the format DD.MM.YYYY);

4. Topic (a closed list of values provided – is given below);

5. The name of the newspaper (an open list of values);

6. Region (a closed list of values, which includes the following items: «Grodno region», «Gomel region», «Brest region», «Vitebsk region», «Mogilev region»);

7. The quantity of tokens (an open list of values);

8. The Internet address of the source to which a certain article belongs (an open list of values);

9. Language of the text of the article (a closed list of values, which includes the following options: «ru» (‘Russian’), «be» (‘Belarusian’), «mfr» (‘Russian with Belarusian inclusions’));

10. Language of the headline of the article (with the same list of options as for the meta-parameter 9 — «Language of the text of the article»).

The list of values for the meta-parameter 4 — «Topic» is divided into 2 sets: 1) general values (these values are used for the general, not specialized content); 2) specialized values (those values are directly related with special fields of knowledge).

The general values of the meta-parameter «Topic» are as follows: 1) «administration and management»; 2) «the army and armed conflicts»; 3) «astrology, parapsychology, esoterics»; 4) «business, commerce, economics, finance»; 5) «home and household»; 6) «health and medicine»; 7) «leisure, sights and entertainment» 8) «art and culture»; 9) «crime»; 10) +«forestry»; 11) «science and technology»; 12) + «education»; 13) «politics and public life»; 14) «law»; 15) + «nature»; 16) + «industry»; 17) «accidents»; 18) «travelling»; 19) «religion»; 20) + «agriculture»; 21) «sports»; 22) + «construction and architecture»; 23) + «engineering»; 24) + «transport»; 25) «private life».

The value «science and technology» can be used isolated in annotation of the texts or it can be used in combination with any general value marked above by «+», as well as in combination with a certain value from the list of the specialized values (see below).

The closed list of the specialized values of the meta-parameter «Topic» is divided into 3 subsets: a) natural sciences; b) applied sciences; c) humanities. All the specialized values can be used only in combination with the general value «science and technology».

The values belonging to the subset of natural sciences are as follows: 26) «astronomy»; 27) «biology»; 28) «geography»; 29) «geology»; 30) «computer science»; 31) «mathematics»; 32) «statistics»; 33) «physics»; 34) «chemistry».

The subset of values belonging to applied sciences includes the following items: 35) «military affairs»; 36) «medicine»; 37) «energy».

Finally, the humanities subset comprises 10 more values: 38) «art history»; 39) «history»; 40) «cultural studies»; 41) «political science»; 42) «psychology»; 43) «religious studies»; 44) «sociology»; 45) «philology»; 46) «philosophy»; 47) «economy».

Comparison of the lists of general and specialized values shows the explicit correlations between certain items:

2) «the army and armed conflicts» — 35) «military affairs»;

4) «business, commerce, economics» — 47) «economy»;

6) «health and medicine» — 36) «medicine»;

8) «art and culture» — 38) «art history»/ 40) «cultural studies»;

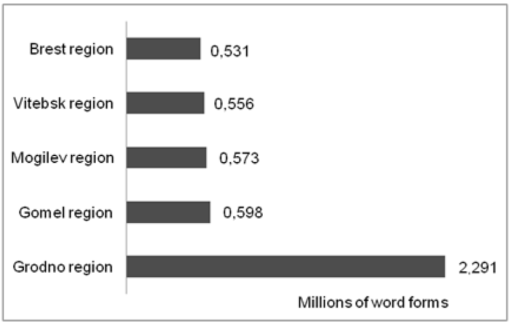
13) «politics and public life» — 41) «political science»;

19) «religion» — 43) «religious studies».

Such a correlation is a very valuable tool for the study of consubstantial lexis as it provides an option to find thematically correlated texts, which belong to general and / or scientific discourse. The analysis of the variety of texts in contemporary newspapers shows that specialized texts are no longer represent exceptions from the rule [Rychkova 2014, 2015].

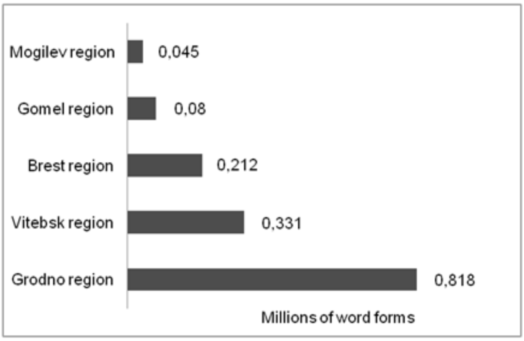
Like in the illustrative linguistic corpus of Grodno region mass media, the system of meta-annotation used for infological modeling of RBM is based on the variety of parameters of the meta-annotation elaborated for the Corpus of regional and foreign press in the frames of the Russian National Corpus. However, it is replenished with a unique parameter (10 — Language of the headline of the article). This meta-parameter is very important for RBM as it gives possibility to find not so rare cases of texts written in Belarusian but having headlines written in Russian as well as quite rare occasions of the Russian texts with Belarusian headlines. Therefore each headline in RBM has a parallel representation in the form of a set of lemmas, that can be considered as one more distinguishing feature of the infological model of RBM.

The current state of RBM is far from being regionally balanced in volumes of the language material (see the diagram in the picture 1 below) and especially it concerns the texts published in Belarusian (see the diagram in the picture 2). The list of all the newspapers processed in the RBM distributed by region is given in the attachment to the article.



Picture 1. Distribution of the whole volume of language material in RBM by the regions of Belarus

This means that the main attention further should be paid to the replenishment of the RBM in order to ensure the regional balance of the language material and enlarge the volume of the Belarusian language material proper.



Picture 2. Distribution of Belarusian language material in RBM by the regions of Belarus

When completed, RBM can be used in an independent manner for different scientific and educational purposes, and its infological model described above makes it obvious that RBM can be also easily included into the Corpus of regional and foreign press in the frames of the Russian National Corpus for the open access through Internet.

REFERENCES

Antopolski, 2004	—	*Антопольский А. Б.* Информационные ресурсы России. — М., 2004. — 423 с.
Korovanenko, 1995	—	*Корованенко Т. А.* Источники нового академического словаря // Очередные задачи русской академической лексикографии. — СПб., 1995. — С. 31–43.
Rychkova, Stankevich, 2017	—	*Рычкова Л. В.* Лингвистический корпус СМИ Гродненщины: технология создания,

направления использования / Под науч. ред. Л. В. Рычковой. — Гродно, 2017. — 115 с.

Rychkova, 2015 — *Рычкова Л. В.* Лексіка рэлігійнай тэматыкі ў беларускамоўных тэкстах СМІ Гродзеншчыны і яе лексікаграфічнае адлюстраванне // Беларуская мова і літаратура ў славянскім этнакультурным кантэксце : Матэрыялы ІІ Рэсп. навук.-практ. канф. (Віцебск, 19–20 лістапада 2015 года) / Рэдкал.: Г. А. Арцямёнак (адк. рэд.) і інш. — Віцебск, 2015. — С. 134–138.

Rychkova, 2014 — *Рычкова Л. В.* Специальная лексика в языке региональных СМИ // Терминология и знание: Материалы IV Междунар. симпозиума (Москва, 6–8 июня 2014 г.) / Отв. ред. С. Д. Шелов. — М., 2014. — Вып. 4. — С. 157–172.

Smułkowa, 2000 — *Smułkowa E.* Dwujęzyczność po białorusku: bilingwizm, dygłosja, czy coś innego? // Język i tożsamość na pograniczu kultur. — Białystok, 2000. — S. 90–100.

ATTACHMENT

List of newspapers in the RBM distributed by region

Regions of Belarus in alphabetical order	Original newspapers' titles	Transliterated newspapers' titles
Brest region	Заря над Бугом	Zarja nad Bugom
Brest region	Навіны Камянеччыны	Naviny Kamjanjechchyny
Gomel region	Дняпровец	Dnjaprovjets
Gomel region	Светлагорскія навіны	Svjetlagorskija naviny
Gomel region	Светлае жыццё	Svjetlaje zhytstsjo
Gomel region	Хойніцкія навіны	Hojnitskija naviny
Grodno region	Астравецкая праўда	Astravetskaja prawda
Grodno region	Бераставіцкая газета	Bjerastavitskaja gazjeta
Grodno region	Вечерний Гродно	Vjechjernij Grodno
Grodno region	Воранаўская газета	Voranawskaja gazjeta
Grodno region	Іўеўскі край	Iwjewski kraj
Grodno region	Перспектива	Pjerspjektiva
Grodno region	Праца	Pratsa
Grodno region	Свіслацкая газета	Svislatskaja gazjeta
Mogilev region	Прыдняпроўская ніва	Prydnjaprowskaja niva
Mogilev region	Родная ніва	Rodnaja niva
Vitebsk region	Герой працы	Gjeroj pratsy
Vitebsk region	Жыццё Прыдзвіння	Zhytstsjo Prydzvinnja

(Grodno, Belarus)

РИЧКОВА Л.В., СТАНКЕВИЧ А. Ю.

СПЕЦІАЛІЗОВАНА МОВНА БАЗА ДАНИХ З ВИКОРИСТАННЯМ ТЕКСТІВ БІЛОРУСЬКИХ РЕГІОНАЛЬНИХ ГАЗЕТ

У статті описано головні засади створення повнотекстової бази даних, що містить тексти регіональних білоруських ЗМІ і була розроблена як експериментальна основа в межах дослідницького проекту «Функціонування білоруської мови в двомовних регіональних засобах масової інформації». Обґрунтовується вибір двомовних регіональних газет як джерел автентичного мовного матеріалу і висвітлюється система анотування.

Ключові слова: мовні ресурси, білоруська мова, міжмовна взаємодія, регіональні газети, повнотекстова база даних, система мета-анотування, корпус змішаного типу.