Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 154 (2014) 175 – 178

Procedia
Social and Behavioral Sciences

THE XXV ANNUAL INTERNATIONAL ACADEMIC CONFERENCE, LANGUAGE AND CULTURE, 20-22 October 2014

The Concept of the Tomsk Regional Corpus: Balance and Representativeness

Olga Sologub[a], Zoya I. Rezanova[b]*, Irina G. Temnikova[c]

[a] National Chengchi University, NO.64,Sec.2, ZhiNan Rd.,Wenshan District, Taipei City 11605, Taiwan (R.O.C)
[bc] National Research Tomsk State University, 36, Lenin Ave., Tomsk, 634050, Russia;
[b] National Research Tomsk Polytechnic University, 30, Lenin Ave., Tomsk, 634034,Russia

Abstract

The paper discusses two issues of the conception of the Tomsk Regional Corpus developed at Tomsk State University, to obtain a more exact balance and representativeness of the Corpus. These parameters of the Tomsk Regional Corpus are viewed in comparison with the Russian National Corpus (basic subcorpus), its dialectal subcorpus and the Saratov Dialectal Corpus.

Keywords: The Russian National Corpus; the Tomsk Regional Corpus; balance; representativeness

1. Introduction

Corpus Linguistics is applied linguistics with a long history that, as it is known, began with the creation of the Brown Corpus in the USA, 1963, then the Lancaster-Oslo-Bergen Corpus in the UK (Zakharov, 2005, p. 4). Currently, there are a lot of global national projects such as the British National Corpus (http://www.natcorp.ox.ac.uk), the International Corpus of English, the Bank of English (http://www.collins.co.uk/Corpus /CorpusSearch.aspx), the National Corpus of Polish (http://nkjp.pl), the Russian National Corpus (http://ruscorpora.ru), the Slovak National Corpus (http: //korpus.juls.savba.ski), the Czech National Corpus (http://ucnk.ff.cuni.cz), the Corpus of the Slovenian language (http://www.korp.fdv.uni-lj.si), etc. We can also mention some corpora of a huge variety of specialized ones differing in the composition of texts and types of

*Zoya Rezanova Tel.: +7 903 952 23 47; fax: 83822 534 077.
E-mail address: rezanovazi@mail.ru

doi:10.1016/j.sbspro.2014.10.131

linguistically relevant markup: the Polish and English Language Corpora for Research and Applications (http: //korpus.ia.uni.lodz.pl), the St. Petersburg Corpus of hagiographic texts of the XV-XVII centuries (http://project.phil.pu.ru/skat), etc.

Creating a corpus stimulates the development of linguistic theory; large databases of linguistically marked texts have an immediate applied relevance and become an empirical basis for further linguistic research. The number of representative texts and their configuration (representativeness and balance of corpus) are of fundamental importance. For the time being the corpora the texts of which exceed 100 million word usages are considered to be representative. A balanced and representative linguistic corpus allows the meeting of research challenges at a new level. Thus, the use of data of ruscorpora allows researchers to make significant adjustments to the existing theories concerning the so-called "linguistic reality" including the compatibility or management of particular lexical items, lexical and grammatical meanings, syntax features, etc. (Pertsov, 2006a, p. 318 - 331; Pertsov, 2006b, p. 227 - 245). Generalization of half a century's experience of creating linguistic corpora provides a possibility of their classification based on a set of relevant features, for example, in Zakharov (2005, p. 13). Currently multidirectional development of corpora displays two pronounced trends - creating global national corpora and special ones of various types. These trends are not contradictory, but complementary. This is evidenced by the development of the Russian National Corpus, which now replenishes its basic stock and develops due to specialized sub corpora: multimedia, dialectal, poetry, accentual, and oral speech ones. The opposition universal vs. specialized corpora is directly correlated with their representativeness and balance, as well as with the principles of markup and metamarkup of corpora.

2. Problem statement

In this paper we discuss one of the problems of the structure of the Tomsk Regional Corpus (TRC), which is developed by a team from Tomsk State University. As its name implies, the basic differential feature of the corpus is the regional limitation of texts. The language of the region comprises a complexly organized unity of different forms of the national language: a standardized literary language, dialects, sociolects, urban and rural vernacular speech. The authors of the project aim to present a representative and balanced regional version of the Russian language in this corpus.

We believe that creating a regional corpus of the Russian language is a very urgent problem the solution of which is due to both further development of corpus linguistics and activation of the problems of regional variation of the literary language in Russian Studies. Russian linguistics interprets regional variation of the national language primarily within the framework of dialectology. Regional variation of the Russian national language and its literary form in Russian studies is just beginning to be investigated and the creation of the regional corpus of the Russian language, which represents extensive linguistically marked text massifs, will be a strong foundation for statistically valid studies of the nature and directions of regional language variations as a complete system.

For the moment Russian linguistics has only dialectal corpora as the implementation of the idea of regionalism: Saratov dialectological corpus and dialectal subcorpus of ruscorpora. It is known that these projects were developed on different theoretical grounds, and they are focused on different tasks (Letuchy, 2005, p. 215 – 232; Letuchy, 2009, p. 114 – 128) (Kryuchkova and Goldin, 2011). In what follows we are going to discuss the theoretical objectives of the authors concerning the Tomsk Regional Corpus in their correlation with the main corpus of the RNC, its dialectal Subcorpus and the Saratov Dialectological Corpus.

3. The representativeness and balance of the Tomsk Regional Corpus as reflection of communication structure

If we compare the TRC with the corpora of national languages, it can be characterized as specialized, since the authors of the TRC do not set the task of textual representation of all the spheres of communication in Russian. The TRC is aimed at representing regional language texts with representation of local variation to be of primary importance. Such an aim of the corpus under development correlates directly with the "ideological preferences of the RNC authors": "Attention to the synchronous variation of language" and the conception of "non-literature centrism" (Plungyan, 2008), that is, when a text is included into the corpus its location should reflect the role of this type of texts in the structure of regional communicative practices combining texts of special business, scientific, everyday and other communication types. This principle is related to the balance of the Corpus. In a balanced corpus the principle of "literature is not in the center" is being implemented by increase in the share of publicism; literary texts at the same time are in the second place in the total corpus, the third position is occupied by the so-called specialized texts. For example, the balanced Slovak National Corpus includes texts in the following proportion: publicism (60.6%), fiction (17.5%), specialized texts (11.6%), and other (10.3%). Two variants of the Slovenian National Corpus FIDA and FidaPLUS contain the texts in the following proportions: literary texts (6 vs. 3.47%), research (18.5 vs. 10%), others (75.5 vs. 86.34%) ; books (22.7 vs. 8.74%), newspapers (46.6 vs. 65.26%), magazines (23.9 vs. 23.26%), the texts from the Internet (electronic texts) (0.02 vs. 1.24%), other (including a small proportion of oral speech - Transcripts of parliamentary hearings) (6.78 vs. 1.5%).

RNC authors include texts representing modern Russian literary (written) language; they stress that "the texts are presented in a certain proportion, reflecting their share in the total array of modern texts. Thus, the share of literary texts (including drama and memoirs) is not more than 40%." and that "all of these texts are part of the body as much as possible in proportion to their share in the language of the relevant period" (Russian National Corpus // URL: http: //ruscorpora.ru). Thus, modern balanced corpora generally reflect the structure of written communication, often of written institutional communication, downplaying the share of everyday personal communication, texts generated in direct informal communication.

The principles of selecting material for the Tomsk Regional Corpus - representativeness and balance - correlate with the principles of ruscorpora; nevertheless, there are significant differences in the issue of balance in the TRC. We should note significant differences in the interpretation of balance in TRC. The authors of the project consider reflection of the structure of communication in the region in the structure of the corpus to be the basic principle. As a consequence, there are significant differences between the principle of balance of the TRC and ruscorpora.

1. The TRC includes texts of written and oral communication. Note that the proportion of ordinary oral communication should be provided in accordance with its role in the communicative existence of modern man. Because of this, the authors of the TRC must solve the problem of creating a database of oral texts and their transcripts. Oral texts should represent the maximum variety of genres of everyday communication. In the TRC balance can also be achieved by increasing the proportion of ordinary personal written communication, including genres that are in zones of overlap of institutional and personal communication (e.g., application letters to the official authorities, meeting minutes, etc.), various genres of modern computer communication.

2. The second fundamental difference between the structure of the TRC and ruscorpora, which is balanced in relation to written texts of the literary language, is that TRC authors aim to reflect the structure of communication in the region as a complex unity of codified and uncodified forms of the Russian language. As noted, the language of the region includes not only standardized literary language, but also the rural and urban vernacular, dialects, sociolects, jargons. Thus, the balance in the TRC is achieved by proportional representation of discourses and genres of uncodified language forms that are specific for the Tomsk region.

3. The third feature of the balance in the TRC is detected in comparison with the existing regional corpora. The Dialect Subcorpus of ruscorpora is a corpus of a differential type, it dialectal material is interpreted against the background of the literary norm and the deviations from it are marked. In the Saratov Dialectological Corpus the dialect is presented and interpreted as a standalone system. The TRC, in the same way as the Saratov Dialectological Corpus, is a collection of texts of non-differential type, focused on the system of regional representation of linguistic unity, in which the boundaries between literary and vernacular norm can be weakly manifested. The TRC includes codified, uncodified texts and the texts of transient types.

4. The fourth feature of balance in the TRC is that, according to the authors of the conception, it is supposed to include proportionally texts representing the Russian-speaking bilingual communicative practice with different types of bilingualism. Since the authors of the corpus aspire to represent the language of the region as a conglomerate of different forms of the national language, they should take into account the effects of its contamination with the contact languages. Tomsk Region is predominately populated by Russians (according to the census in 2002, 90.84%), and at the same time, the other 10% are representatives of more than two dozen nationalities - Tatars 1.93%, 1.60%, Ukrainians, Germans, 1.29%, and the rest refer to a range from 0.56 to 0.06% (in descending order) - Chuvash, Belarusians, Armenians, Azerbaijanis, Bashkirs, Mordvinians, Selkups, Uzbeks, Umurty, Moldavians,

Poles, Kazakhs, Koreans, Khanty, Mari, Jews Estonians, Latvians, Chechens, Georgians. In the result of ethnic interactions in the region divergent bi- and polylingual language situations have been formed. While the languages of the indigenous peoples have been studied extensively, the variants of the Russian language functioning in bilingual conditions have hardly been studied. TRC shall provide representative material to fill this research lacunae.

Provision of this option of the balance is achieved by taking into account intersecting parameters: a balanced representation of different genres, forms of the national language, within the latter - a representative view of texts, monolinguals and bilinguals.

4. Conclusion

Thus, the Tomsk Regional Corpus is based on the following typologically important parameters selected by V. P. Zakharov: 1) the language of texts – Russian; 2) data type - mixed, that is, including written and transcribed oral texts; 3) monolinguality VS. “literariness” - mixed, that is, the material is both proper literary proper and belonging to other forms of the national language, including the variants that will arise in the bilingual environment the speakers live in; 4) genres - representation of genre diversity of different functional-stylistic registers of speech.

Acknowledgement

The paper is developed as a part of project “Linguistic corpus “Tomsk local text”: conception and structure” №14-14-70010 supported by Russian Foundation for Humanities.

References

Bank of English (2012). Retrieved from: http://www.collins.co.uk/Corpus /CorpusSearch.aspx.

British National Corpus (2010). Retrieved from: http://www.natcorp.ox.ac.uk.

Corpus of the Slovenian language (2006). Retrieved from: URL: http://www.korp.fdv.uni-lj.si.

Czech National Corpus (2012). Retrieved from: http://ucnk.ff.cuni.cz.

Goldin V. E., Kryuchkova O. Yu. (2007). Electronic corpus of Russian dialect speech and principles of its markup. *Journal of Saratov State University. Philology. Journalism*. Vol. 7. Issue 1. Saratov.

International Corpus of English (2014). Retrieved from: http://ice-corpora.net/ICE/INDEX.HTM.

Kryuchkova O. Yu., Goldin V. E. (2011). Corpus of Russian dialectal speech: the concept and parameters of evaluation. *Computer-based linguistics and intellectual technologies. Proceedings of International Conference «Dialogue»* (Bekasovo, May 25 – 29, 2011 г.), issue 10 (17), Moscow, 359 – 367.

Letuchy A. B. (2005). Corpus of dialectal texts: objectives and challenges. *National Corpus of the Russian Language*: 2003—2005. Moscow: Indrik, 215– 232. Retrieved from: http://ruscorpora.ru/sbornik2005/13letuchy.pdf.

Letuchy A. B. (2009). Dialectal corpus: composition and markup peculiarities. *National Corpus of the Russian Language*: 2006—2008. New results and perspectives. SPb: Nestor-Istoriya, 114 – 128.

National Corpus of Polish (2007). Retrieved from: http://nkjp.pl.

Pertsov N. V. (2006a). About the role of corpora in language studies. *Proceedings of International Conference «Corpus Linguistics-2006»*. SPb., 318 – 331.

Pertsov N. V. (2006б). About judgments concerning the Russian language facts in view of corpus data. *The Russian language from scientific position, 1(11),* 227 – 245.

Plungyan V. A. (2008). Corpus as a tool and ideology: some lessons of contemporary corpus linguistics. *The Russian language from scientific position*, *16 (2),* 7–20.

Polish and English Language Corpora for Research and Applications. (2014). Retrieved from: http: //korpus.ia.uni.lodz.pl.

Russian National Corpus. (2003-2014). Retrieved from: http://ruscorpora.ru.

Slovak National Corpus. (2013). Retrieved from: http: //korpus.juls.savba.ski.

St. Petersburg Corpora of Hagiographic Texts of the XV-XVII centuries (2008 - 2014). Retrieved from: http://project.phil.pu.ru/skat.