ISSN 2039-2117 (online)
ISSN 2039-9340 (print)

Mediterranean Journal of Social Sciences
MCSER Publishing, Rome-Italy

Vol 6 No 2 S4
April 2015

Genre Aspect as Metamarker of Tomsk's Dialect Corpus Studies of the Middle-Ob Dialects

Anna Alekseevna Dolganina (Plotnikova)

National Research Tomsk State University, PhD in Philology
E-mail: plotnikowa.anna@gmail.com

Doi:10.5901/mjss.2015.v6n2s4p99

Absrtact

Genre marker parameter of dialectal speech is discussed based upon the comparative assessment with practice of genre marker linguistic corpuses, and dialect subcorpus of the National Russian language corpus, Saratov Dialect Corpus and Tomsk Dialect Corpus. The problems of representativeness of dialect genres in the corpus and principles of determination of this parameter in texts in Tomsk Dialect Corpus are separated. It is proved that determination of genre limits in the dialect corpus is specific in comparison to the literary language.

Keywords: *genres; dialect; dialectal corpus; metamarker.*

1. Introduction

Corpus linguistics of the end of XX century became indispensable part of the linguistic studies and formed as the independent scientific direction, represented by the range of large projects such as British corpus (BC), Czech national corpus (CNC), National corpus of Russian language (NCRL) and studies (Kilgarriff; Catherine, Fillmore, Atkins; Martin; Oakes; Sinclair). The corpus approach to language allows not only analyzing the frequency of word usage and conduct statistical analysis, but also to check the range of philological hypotheses on the big corpus of materials, optimize the search of data, preserve and present for wide range of users earlier inaccessible archives of data (for example, materials of dialectological expeditions).

The development of principles of corpus construction is one of the most actual and complex problems of corpus linguistics. The subject of investigation in this study is the problem of genre linguistic corpus marking, in particular dialect one (based upon the material of Russian dialects of Middle-Ob dialects. The genre is one of the parameters of metatextual marking of majority of famous corpuses (for example, BC, CNC, NCRL); this is the sign, which characterizes the texts on the whole. However, the single methodological principles in the text marking in this aspect don't exist presently.

Primarily, the genre aspect of corpus metamarking has been formed spontaneously. So, at first we'll give the brief review of how the text genre belonging in the modern corpus is reflected. Then this problem will be considered in the applied aspect: formation of metamarking principles of Tomsk dialect corpus as a linguistic problem.

Presently, the lexical marking (abstract) that is automated in many respects is the most developed. However it is impossible to restrict only with the area of word usage: the results of study can't be representative, if not take into account different types of speaking and writing situations, then there is important metatextual information, including the genre belonging.

2. Discussion

Genre is one of the parameters of metatextual marking of the majority of famous corpuses (for example, BC, CNC, NCRL), this is the sign that characterizes texts on the whole along with such signs as area of functioning, text type, heading, volume, theme, author and data of creation. However, the principles of genre marking are ambiguous and cause the range of theoretical and practical problem issues, the attempt to designate which is taken in the present article.

For the representative corpuses the requirement of representativeness and balance is set up. Thus, in the genre relation in ideal the corpus should represent the system, reflecting the diversity of genre content of language and at that the different genres must be represented proportionally to each other. This principle is not always followed, as an example we can state the corpus of modern American English language (CCAE), in which the genre strategy is one of

the principles of quantitative selection of the material and structure of corpus on the whole, that's why at the request "the quantity of word usages in texts divided into 5 groups, is approximately similar – 70-80 millions. This allows user automatically seeing the frequency according to genre at request issue. The obviousness and simplicity of such analysis impress" [Mordovin, 2009].

As A. Y. Mordovin showed the majority of representative corpuses are characterized by "absence of the strict genre programme" [Mordovin, p. 48], especially the quantitative disproportion of different genre texts is typical for corpuses, fixing the spoken language. However, the genre content of real communication has not strict structural system and is done not by one, but by the complex of bases. So any attempts to lead it to the simplified form are always artificial in certain extent and should be directed for solution of the certain tasks. The list of speech genres aspires to endlessness that is not positive phenomenon for the corpus, besides, there is essential quantitative inequality of different type genres that conditioned the problems of imbalanced content.

The other essential problem of corpus genre marking is the problem of genre classification and basis selection for their separation. "It is important, that the system of genres, offered in the corpus, should correspond to representations about genres of those linguists, which are the target corpus audience. In case if the gap between classification corpus scheme and traditional conceptions is too big, there is some risk, that many linguists would prefer ignoring the genre diversity" (Piperski, 2013).

For example, in Czech national corpus (CNC) the mixture of thematic and genre characteristics are also observed. The "type of genre" properly includes only 60 types, for example, "drama", "novel", "philosophy", "industry", "sport" at that there is a separate parameter of "subgenre": "text-book", "critical article", "encyclopedia", but genres and subgenres are not correlated hierarchically (*Český* Národní Korpus). Thus, the statistical comparison of different type texts becomes complicated and ineffective.

Based upon the British national corpus BNC D. Lee made the special basis (The BNC Index), in which he separated 24 spoken language genres, 46 – for written one, but the bases of their separation are unambiguous in many respects. For example, along with such genres as "biography", "school compositions", "personal letters", and "business letters" macrogenre formations as "academic prose": "natural sciences", "fiction literature": "drama", "central newspapers": art/culture, "central newspapers: different" were separated (Lee, 2001).

In the Russian practice of text metamarking in corpus linguistics the support on the communicative scheme of J. Sinclair (Thomspon), using the traditional nomenclature of genres (American tradition of abstracting) is typical, on the other part, on the categories traditional for the Russian linguistic: "the area of functioning" and "speech genre" as a text type, characterizing the unity of thematic content, compositional and language filling (Bakhtin, 1979). In this understanding the speech genre includes the complex of signs, which are taken into account in text metamarking. Usually, the researchers refer to the parameters, offered by M.M. Bakhtin and T.V. Shmeleva: communicative aim (informative/imperative/etiquette or social/evaluative), author's image, addressee's image, dictum (text subject, event-trigger basis) (Shmeleva, 1997).

In NCRL for marking the genre belonging of the text 2 terms are used: "the fiction literature genre" and "text type", which are understood as the speech genres. S.O. Savchuk marks that "…the shortage of term *genre* is its polysemy as along with the above-mentioned linguistic understanding of the term there exist the literary tradition of separation and description of genres of fiction literature" (Savchuk). About 100 speech genres or text types properly are distinguished in NCRL (for example, riddle, notes, legend, libretto, miniature: anecdote, miniature: joke, parable, story, novel, fairy-tale and etc.) and their list is open.

The list of fiction literature "genres" in NCRL is limited as only 11 of them are separated: "non-genre prose", "detective, thriller", "children's", "historical prose", "adventures", "fiction", "love story", "humour and satire", "documentary prose", "dramaturgy", "translation". Marking these types of fiction texts is based upon the following determination of genre: "The partial display of type [fiction text], determined by the story theme, is called the genre of fiction literature" (Gorshkov, 2001). On the background of complex and many-aspect theory of genre in literature this typology, on one hand, is good because it finds the single (subject-thematic) basis for distinguishing the big volume of text, on the other hand, it is insufficient for own genre studies, besides from the positions of theoretical literature it makes the issue about limits and essence of "genre" notion.

Recently the methods, which execute automated classification of genres, appear, but no one of them became generally accepted presently. Not very traditional approach to genre marking is taken by the General Internet-corpus of Russian language: "the genre categories are separated a posteriori on the basis of similarity of texts, which enter into the corpus, between each other" (Piperski, 2013) and marking is executed in the automated type through the machine education according to 15 issues-parameters of S.A. Sharov, which have four-frequency scale [see the same]. However, presently this study is on stage of development. Unfortunately, this method is not applied on the present stage to dialect

corpus because of character of the available dialect material, the complexity of composition of issues-parameters, which suit for the available material.

V.P. Zakharov and S.Y. Bogdanova separate several types of the specialized corpuses according to the genre sign: "literary", "folklore", "dramaturgic", "publicist" (Zakharov, Bogdanova, 2011). Moreover, the corpus can be made of texts of one genre, for example, the famous project "The story about dreams" (Kibrik, 2009).

The separate consideration deserves the genre metamarking specialized textual corpuses, within the framework of present study – dialect corpus. The genre dialect system due to the peculiarities of communication character in the national-speech culture is distinguished not only by the quantity of genres, but by the quantitative characteristics and at abstracting the corpus of dialect texts it is necessary to take into account this specificity.

The nature of speech genre of dialect discourse is distinguished from the literary language. In the theory of speech genres on the material of literary language it is stated that the "genre is correlated with situation, event, text, and has the quantitatively and qualitatively more complex nature, than the speech act, even if it represents subgenre, which volume is equal to single-act statement, as for example, in different signboards" (Komleva, 2011). However, in the dialect discourse the text is almost always polythematic, however, one speech situation contains a lot of propositions, referring to other situations, at that the transition from one to another can be both sharp and implicit. So, almost any text requires having several genre determinants.

The dialect subcorpus on NCRL almost doesn't separate the speech genres, as there only subdivision into 4 large categories is supposed: 1) spoken non-folklore texts, 2) written non-folklore texts, 3) spoken folklore, 4) written folklore texts (Savchuk). At that the basic attention is paid to the thematic marking. The dialect subcorpus of NCRL and Saratov dialect corpus are distinguished according to the character of thematic and genre marking of corpuses: "In DC of NCRL the separation of text according to the thematic principle gives the excessive thematic marking of each separate text. The inconsiderable volume of thematically integrate texts in the Dialect subcorpus of NCRL conditions also inappropriateness of their genre marking: each text is usually multi-genre ("daily sphere")" (Kryuchkova, Golding, Retrieved from http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/36.pdf4). However, it should be noted that the materials of Tomsk dialect expeditions contain small quantity of such type notes.

In Saratov and Tomsk dialect corpuses the genre marking is provided as the part of metamarking structure. In the principles of genre system separation there is certain succession, but differences too.

The elements of genre parameterization of Saratovsk corpus are, for example, "narration story", "argumentation", "description", "fairy-tale", "song", "proverbs, and sayings".

In Tomsk dialect corpus, based upon the results of material studies of numerous expeditions and concrete language personality, it is stated that such genres as "biographical story", "interview", "description" and "argumentation" are marked [Kazakova; Voloshina]. First of all, the corpus on the materials of expeditions reflect and confirm the fact, that "one of most important functions of dialect as the spoken form of communication is the transmission of culturally important information from one generation to another, embodied into the system of genre forms. The speech genres of remembrance and autobiographical story are core ones in this relation" (Voloshina, Demeshkina, 2012).

In text metamarking not only the type of genre is designated, but more exact subgenre that is characterized by the connotative direction. Subgenre in Tomsk dialect corpus is expressed in the form of closed system and represented by 4 groups:

- Informative: event message, intent message, opinion message, citation message, supposition, explanation, complaint, warning.
- Imperative: request, disposal, instruction, order, offer, advice.
- Ritual: welcoming, parting, apology, gratitude, invitation, treating, wish.
- Evaluation: praise, blame, self-appraisal, evaluation (Yurina, 2011).

The source of the materials of Tomsk dialect corpus are materials of expeditions, conducted on the basis of Tomsk dialect school, beginning from 1946 up to present days. From the point of view of genre balance the materials of expedition are vulnerable, and in many respects fragmentary reflect the genre dialect system, as the majority of texts were received in the result of conversations with dialectologist, when in expeditions of last decades frequently there was no possibility to fix the spontaneous and bounded dialect texts. Thus, the problem of unbalanced genre system in dialect corpus is inevitable. However, in spite of the above-mentioned facts, the genre corpus marking, oriented "from the material", allow fixing the genre diversity and material peculiarity with the range of slips of tongue.

One of the aims of Tomsk dialect corpus creation is possibility to model fragment of dialect discourse with its assistance. Namely this dialect corpus should be textual, instead of being limited with only country dance construction. In the natural communication even the strictly formed folklore genre are not isolated. So possibility to see not only the

context of 1 word, but the expanded context for the text on the whole that will allow seeing both the other genre introduction function, and the ways of its entering into the speech flow, is important.

It is appropriate for the dialect to separate 2 types of genres: folklore and speech ones. The folklore genres, developed by the long tradition of national culture, as a rule, are well separated in the speech flow. The folklore genres are separated in the separate group, in connection to their specificity and isolation from other speech types of text: plot story, fairy-tale, song, chastushka and proverb.

In the practice of replenishment of linguistic corpus there is a problem of definition of text limit, especially in those cases, when there is interaction of some genre forms in the speech. The issue about presence of 1 or 2 texts is disputable, when any folklore genres are used by the speaker in the frameworks of single narration or description of anything, for example, as an illustration or confirmation of the thoughts. On the one part, in this case there is 1) commonality of speech situation, 2) theme unanimity, 3) auxiliary and reproducible character of the used folklore text, 4) correspondence of the folklore text of general statement aim. On the other part, the folklore text has 1) compositional completeness, 2) presence of distinguishable language genre signs, and 3) reproduction ability in the ready form in the similar situations.

For example, the limits of chastushka text are easily separated in the speech flow, while the statements framing it are not texts, which own the signs of integrity and coherence.

The additional complexity is represented by the processing of dialect notes of different time periods. If the modern notes of dialects aspire to maximal reflection of own communication, then notes made from 50th years of 20th century were kept on the assumption of other methodological principles. At that in the materials of 70-80th years the researches often takes role of communication moderator, but unfortunately requesting remarks are not frequently preserved in the notes, however the notes of this period represent the peculiar value for the scientific comprehension within the framework of replenishment of the dialect corpus. In particular, these notes also represent certain interest from the point of genre variety. Answering to the researcher's request, dialect speaker can remember and tell fairy-tale, parable, perform song or chastushka, and list specific signs. At that the dialect speaker distinctly recognizes the difference between chastushka and song, peculiar sign and saying, and also represents the set of situations, in which these genres will be appropriate. However, it is important to mark that in the materials of dialect expeditions these genres are most frequently provoked.

Presently, the genre marking of text in corpus linguistics is done by hand. Let's consider the task of genre marking on the example of small fragment of the decoded note of the sounding speech made in 1979 in Belyi Yar village of Verkhneketskii district of Tomsk region. The issues of informant are not represented in the notes.

> *Here is meatea'ter, I don't know when, well, perhaps, he is usually in spring, well, meatea'ter, and they eat meat, all days eat meats. Well, now meatea'ter, there is meat in the store – well meatea'ter, it there is no it – then advent.*
> *Matan'ia, we used to speak so, well it is, matania, now, Gods forgive me, there are frayer', while we speak motan'ia. Motan'ia, mota'nechka – the both are tender. Well, this is tender, motania, so this is good.*
> *Well, there are a little of "flies, matan'ia is still there, and we composed songs about matan'ia.*
> *Well, now it is happened with us too. Being drunk the militia men drove his car and knocked the boy and girl on the motorcycle. He should be taken in the army, today seeing-off might be, and today he was knocked.*
> *I'll feed you with my soup. Perhaps, you would like it, but all the same)*
> *(Вот мясое'д, я не знаю когда, вот, наверно, весной он бывает, мясоед-то, мясо едят, всё скоро'мно. Да, щас мясое'д-то, мясо в магазине есть – вот мясое'д, а нет – так пост.*
> *Мата'ня, у нас всегда так говорили, ну а как же, матаня, щас как-то, прости меня господи, фраера', а у нас мота'ня. Мота'ня, мота'нечка – одно и то же ласково. А чё, это ласково, мотаня, хорошо.*
> *А «залётка» как-то мало, у нас всё мата'ня, у нас и песни складывали про мата'ню.*
> *А но'нче как у нас получилось тоже. Мильцанер ехал пьяный на своей машине и парня и девушку на мотоцикле сбил. Его в армию, сёдня проводи'ны и сёдня его сбили.*
> *Я своим супом покормлю вас. Может, не поглянется вам, но всё равно).*

(Copy-book 961, The place of note is Belyi Yar, Verkhneketskii district, Tomsk region, 1979, informant Talaeva (Drozdova) Praskov'ia Dmitrievna, 54-years-old).

Most of all, the fragment represents the part of linguistic interview with dialectologist, in which the range of metalanguage argumentation of dialect speaker is distinctly separated:

> *"meateater, and they eat meat, all days eat meats"; Matan'ia, we used to speak so, well it is, matania, now, Gods forgive me, there are frayer', while we speak motan'ia. Motan'ia, mota'nechka – the both are tender"; "well, there are a little, still there…", "and we composed songs" (мясоед-то, мясо едят, всё скоро'мно.»; «Мата'ня, у нас всегда так говорили, ну а как же, матаня, щас как-то, прости меня господи, фраера', а у нас мота'ня. Мота'ня,*

мота'нечка – одно и то же ласково»; «как-то мало, у нас всё...», «у нас и песни складывали»).

At that we should mark 2 typical genre features for dialect communication: 1. semantization through motivation (*meateater = eat meat)* (*мясоед = мясо едят*) and 2) expressed opposition "own" – "alien" (*we have* (*у нас*)). At that the markers of linguistic interview are "speak" and "compose" <songs/chastushkas/verses>.

In the fragment the speaker twice changed the theme sharply: at first the genre "event message" is introduced, that is marked by the typical beginning and event informing phrase:

> *Well, now it is happened with us too. Being drunk the militia men drove his car and knocked the boy and girl on the motorcycle. He should be taken in the army, today seeing-off might be, and today he was knocked (А но'нче как у нас получилось тоже. Мильцанер ехал пьяный на своей машине и парня и девушку на мотоцикле сбил. Его в армию, сёдня проводи'ны и сёдня его сбили); and treating: "I'll feed you with my soup. Perhaps, you would like it, but all the same" ("Я своим супом покормлю вас. Может, не поглянется вам, но всё равно").*

The belonging to subgenre of treating is conditioned by the statement dictum and communicate future put in it – then the refusal will follow, or the supposed action will be realized *("I'll feed you with my soup" ("супом угощу")).* Moreover, in the stated discourse fragment there is also evaluation: *"Motan'ia, mota'nechka – the both are tender. Well, this is tender, motania, so this is good." ("Мота'ня, мота'нечка – одно и то же ласково. А чё, это ласково, мотаня, хорошо").*

In the example stated the following marking variant is offered:

#genre: interview, narration;

#subgenres: explanation, event message, treating.

From the point of different speech genre theory, the genre belonging as the evaluation statement both event message and treating is disputable and depends on the genre comprehension, in particular, from the criteria of genre differentiation in the speech flow. Thus, there exists opinion that the genre limits must coincide with the limits of statement text, which are determined by the change of speech subjects [Bakhtin 1986: 255].

It is obviously that not any evaluative statement is genre as, for example, in the studied text the evaluation doesn't belong to the described object, but to the language nomination. Moreover, the evaluation of one or another dialect word in comparison to the literary or low colloquial is one of the typical sings for metalanguage interview.

3. Concluding Remarks

As a conclusion it should be noted that at by-hand marking the volume of material becomes complex practical task and requires optimization of temporary expenditures. On the one part, the proper text analysis is required, and on the other part, it is necessary to take into account the factor of target purpose of corpus and comfortable usage, and finally, the genre marking must if possible completely reflect the originality of genres of dialect speech. In connection to the mentioned specificity stage-by-stage abstracting is used at the genre marking: on the first stage the genres, which are most typical for dialect and texts and having the obvious multigenre nature, are marked (especially the group of folklore texts); on this stage many texts are outside the genre definitions. On the second stage subgenres are marked, and then (along with corpus replenishment) the correction and specification of genre definition take place in accordance with researchers' requirements.

References

Ball Catherine N. Tutorial: Concordances and Corpora. Retrieved from http://www.georgetown.edu/cball/corpora/tutorial.html.

Bakhtin, M. (1986). *The Problem of Speech Genres. Aesthetics of Verbal Creativity*. M.: Iskusstvo. 250-297.

BNC: The BNC Users Reference Guide (2000). Retrieved from http://www.natcorp.ox.ac.uk/World/HTML/.

Fillmore, C. (1994). Atkins B.T.S. Starting Where the Dictionaries Stop: the Challenge of Corpus Lexicography. *Computational Approaches to the Lexicon*. 151–159.

Gorshkov, A. (2001). *Russian Stylistics*. Moscow.

Gellerstam M. (1992). Modern Swedish Text Corpora. Directions in Corpus Linguistics. Berlin.

Kazakova O. (2007). *Dialect Language Personality in Genre Aspect.* Tomsk: Publ. house Tomsk Polytechnic. University. 196.

Kilgarriff, A. Web as Corpus. Retrieved from http://www.itri.bton.ac.uk/Adam.Kilgarriff/wac_cfp.html

Kibrik, A., Podleskaia, V. (ed.) (2009). *Dream Stories: Corpus Study of Spoken Russian Discourse.* M: IASK.

Kocek, J., Kopřivová, M., Kučera, K. (eds.) (2000). *Český Národní Korpus – Úvod a Příručka Uživatele*. Praha.

Komleva, E.(2011). Correlation of Notion "Speech Genre" and "Speech Act" (on German Appellate Text Material). *Theory and Practice of*

Public Development. 5. pp. 296−301.
Kryuchkova, A., Golding, V. Corpus of Russian Dialect Speech: Evaluation Conception and Parameters. Retrieved from [Electronic resource] http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/36.pdf4.
Lee, David Y. (2001). Genres, Registers, Text Types, Domain, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, Vol. 5. Nr. 3. pp. 37-72. Sep.
Mordovin A. (2009). Completeness of Genre Composition of Modern Non-specialized Text Corpora Revisited. *Vestnik of IGLU: Language, Culture, Communication*. Irkutsk. 48-52.
Oakes, M. (1998). Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh.
Piperski, A. (2013). Genre Classification in the General Internet Corpus of Russian Language. *Modern Problems of Science and Education*. 4. Retrieved from www.science-education.ru/110-9762 (request date: 01.09.2014).
Proceedings of the LREC (Language Resource Evaluating Conference). (2002, 2003, 2004, 2005).
Sinclair, J. (1991). Corpus, Concordance, Collocation, Oxford University Press.
Shmeleva, T. (1977). Model of Speech Genre. *Speech Genres: Coll. of art. Iss. 1.* Saratov: College. pp. 88-99.
Voloshina, S. (2010). Speech Genre of Autobiographical Story (on Dialect Material).*Vestnik of Tomsk State University. Philology.* 2010. 2 (10). pp. 5-10.
Voloshina, S., Demeshkina, T.(2012). World Modelling Potential of Speech Genre (on Dialect Material). *Vestnik of Tomsk State University. Philology.* 3 (19). pp. 14-20.
Yurina, E. (2011). Tomsk Dialect Corpus: Beginning of Way. *Vestnik of Tomsk State University. Philology.* Tomsk.. 2(14). pp. 58 – 63.
Zakharov V., Bogdanova S.(2011). *Corpus Linguistics: Text-book for Students of Humanitarian HEIs.* Irkutsk: IGLU. 161.