Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

DISCOURSE FEATURES OF BLOGS IN SUBCORPUS OF RUSSIAN RU-RSTREEBANK[1]

Toldova S. (toldova@yandex.ru)
NRU Higher School of Economics, Moscow, Russia;

Davydova T. (tdadidik@gmail.com),
Kobozeva M. (kobozeva@isa.ru),
Pisarevskaya D. (dinabpr@gmail.com)
FRC CSC RAS, Moscow, Russia

The paper presents a corpus study of the discourse features in the corpus of blogs. It is based on the data of Ru-RSTreebank annotated within the framework of the Rhetorical Structure theory [Mann, Thompson 1988]. The Ru-RSTreebank represents genres of news and popular science, scientific papers, and blogs texts. Blog subcorpus contains such topics as travelling, cosmetics, sports and health, psychology, IT and tech and some others. Blogs texts constitute a specific genre as they combine properties of written and spoken discourse. The purpose of the paper is to investigate discourse features of blogs in comparison with other genres. We analyze the variation in rhetoric relations distribution among genres, and single out the differences in discourse connectives usage. Furthermore, we check the distribution of other discourse features reported in different studies for spoken discourse and for social media in the Ru-RSTreebank blogs subcorpus. The general frequency analysis and the experiments on RandomForest classifier application to genre recognition have shown that the most important rhetoric relations specific to blogs are Evaluation and Contrast, that there is a tendency to use shorter discourse units and not to express the discourse relations overtly via subordinative conjunctions.

Keywords: discourse analysis, rhetorical structure theory, blogs annotation, corpus linguistics, corpus annotation

DOI: 10.28995/2075-7182-2020-19-747-761

[1] The study was funded by Russian Foundation for Basic Research according to the research project № 17-29-07033.

ДИСКУРСИВНЫЕ ОСОБЕННОСТИ БЛОГОВ НА МАТЕРИАЛЕ РУССКОЯЗЫЧНОГО ДИСКУРСИВНОГО КОРПУСА RU-RSTREEBANK

1. Introduction

The research on discourse coherence and on how this coherence is achieved has experienced a revival in the last few years. The core questions are how the discourse relations are established and what are signals of these relations. In order to answer these questions, corpora supplied with discourse annotation have been constructed for English and many other languages [Matthiessen, Teruya 2015]; [da Cunha 2016]; [Iruskieta et al. 2015]; [Zeldes 2017]. Besides serving as a source for theoretical studies, these corpora are used as datasets for building discourse parsers.

The structure of texts from different registers (e.g. written vs. spoken texts) and genres varies in a number of parameters, such as typical discourse relations, sentence length, verb forms, conjunctions usage etc. [Chafe 1982]. This variation should be considered in discourse research and in building the applications. In recent decades, new genres have emerged, pertaining to online social media and blogs communication. Its discourse characteristics are widely discussed (e.g. [Simaki et al. 2018]; [Berger, Hennig 2015]; [Germasheva 2010] etc.). It is generally assumed that it combines properties of written and spoken modes and, besides, manifests its own features. Thus, the investigation of the discourse parameters of social media as compared to other genres and registers is of great importance.

The present work deals with the analysis of discourse features of blogs. The data is taken from Ru-RSTreebank corpora [Pisarevskaya et al. 2017]; [Chistova et al. 2019] annotated within the RST framework (Rhetorical Structure theory) [Mann,Thompson 1988]. The new release of this corpus in 2019 includes 104 texts of blogs. The main research questions are: (a) whether there is a significant difference in rhetoric relations distribution among genres; (b) what other characteristics of discourse units distinguish blogs from other genres; (c) whether there is a difference in discourse connectives usage among the three genres represented in Ru-RSTreebank; (d) what other discourse-oriented devices let blogs differ from written texts. Our aim is to check the claim, articulated in the literature (e.g. [Simaki et al. 2018]), that blogs carry some of the features of spoken discourse. One can find the correlates of specific types of features pertaining to spoken discourse, such as short subclausal discourse units (elaborations, parcellations etc.), markers of interaction or regulatory markers and some others. We enhanced the corpus with additional annotation for these features. In this work, we provide the corpus-based analysis of blogs with respect to spoken discourse features.

The paper is structured as follows. We start with a brief description of theoretical assumptions (2.1) and short overview of discussions, devoted to the written vs. spoken discourse opposition (2.2), and blogs discourse (2.3). Next, we describe the

corpus data and its preparation for the analysis (3.1–3.3). After that, we turn to quantitative (4.1–4.3) and qualitative analysis of various discourse features (5.1–5.2).

2. Background

2.1. Discourse structure

Our study is based on the corpus analysis of Ru-RSTreebank (https://rstreebank.ru/dataset), the first discourse-annotated corpus for Russian [Pisarevskaya et al. 2017]. According to RST, the discourse is a hierarchical structure. A text can be successively split into spans (discourse units, DUs, up to elementary discourse units, EDUs) based on the rhetorical relations between them (e.g. Cause-Effect, Concession, Joint etc.). The relations between DUs are somehow parallel to the relations between clauses in complex sentences. They can be asymmetrical (cf. subordinate vs. main clause) or symmetrical ones. The 'canonical' EDUs usually describe events or states and, hence, syntactically, the typical EDUs are simple clauses [Kibrik, Podlesskaya eds. 2009]. However, there are different types of EDUs smaller than a clause (subclausal) in spoken discourse (ibid.).

2.2. Difference in register for spoken and written discourse

There is a claim in the literature that microblogs and blogs are similar to spoken conversations (e.g. [Scheffler et al. 2019], [Akhapkina 2014] etc.). According to [Chafe 1982] (see also [Kibrik, Podlesskaya eds. 2009]), spoken vs. written discourse are opposed with respect to two basic dimensions. The first one is **'integration/fragmentation'**. Written texts contain more complex sentences: nonfinite clauses (nominalizations, relative clauses, infinitives). In speaking, units are shorter; the relations between them are often expressed covertly without special conjunctions. Consequently, sentences and clauses are longer in a written discourse. The second opposition is **'detachment/involvement'**. Passive voice is more common for written texts. Spoken discourse contains frequent reference to the speaker, more indexicals (*you, me, here, now*) and more particles.

According to [Castellà 2004], spoken conversation has a verbal style (more verbs and verbal complements); more reduced phrases and clauses, flexible structures and more repetition as well as profuse use of discourse markers. Spoken genres use a reduced variety of connectives, they are polyfunctional.

Other important features are **fragmentation and special discourse particles**. Thus, speech is discontinuous, its production is a sequence of segments, and the standard concept of a sentence is inapplicable to it (cf. [Lapteva 1976]).

In [Wang et al. 2019: 156] additional relations for RST-style annotation of spontaneous speech are suggested, i.e. unfinished utterance relations, discourse particle relations (*as you know,* or *right, Okey*, which are satellites of adjacent spans). There are spoken discourse corpora annotated according to PDTB (Penn Discourse Treebank), [Prasad et al. 2008] standards, the annotation was discussed in [Tonelli et al. 2010]; [Rehbein et al. 2016]. According to the study [Crible, Cuenca 2017], based on this type

of annotation, the structures containing discourse markers in unplanned speech are often truncated (the second argument is missing/not complete). The clusters of discourse markers are also quite frequent in spoken discourse (cf. *and so, because if*).

To sum up, according to the literature, clauses in spoken discourse are simpler and shorter as compared to the written one. Moreover, there are a number of incomplete subclausal units in the former. The spoken discourse is characterized by high frequency of particles. A high percentage of complex sentences with non-finite forms is a written text feature.

2.3. Spoken discourse features in blogs

There is a considerable number of blog corpora ([Macdonald, Ounis 2006], [Burton et al. 2009], [Mishne et al. 2005], [Quan, Ren 2009], [Santos et al. 2018] etc.), primarily designed to model topic classification or opinion mining tasks. The annotation of blog subcorpus of Ru-RSTreebank focuses on discourse structure.

Many studies deal with the features that blogs share with spoken conversations ([Simaki et al. 2017: 14], [Scheffler et al. 2019] etc.). Fragmented or incomplete clauses, dialogic interaction expressions are among them. According to [Berger, Hennig 2015], blog texts look more personal and diary-like than regular news. There are in-depth studies of blogs in Russian (e.g. [Germasheva 2010], [Kuznetsova 2008], [Novikova 2005]). The researches point out the high frequency of 'dialogical' features in blogs, including interaction markers, questioning particles (*da?* 'yes'*, a?, pravilno?* 'right?'), the high frequency of second person verb forms, imperatives, politeness formulas, questions, and others. Another bulk of features mentioned in studies deals with intentional simulating spoken discourse, its spontaneity and disfluency: ellipsis, *a* 'but' for a topic shift and regulatory particle *nu* for summarization; attitude expressions; ellipses marks that simulate slowed-down speech under the effect of emotional states etc. Other spoken discourse features are nouns in nominative case, incomplete phrases, self-corrections, high frequency of simple sentences, parcellations, colloquialisms and usage of highly expressive lexis.

Therefore, many features differentiating the spoken and written discourse are relevant to blogs characterization. We focus on the ones that are related to discourse structure and discourse properties. These are segmentation into DUs and their structural properties (e.g. part of speech proportions), distribution of rhetoric relations and some properties of discourse markers (e.g. interaction and regulatory markers).

3. Data

3.1. The corpus

The data for our analysis consists of three subcorpora of Ru-RSTreebank:
(1) news and popular science texts (129 texts);
(2) scientific papers (100 texts);
(3) blog subcorpus.

(1) and (2) have 279,426 tokens in total. (3) contains 104 blog texts, 128,917 tokens. Their topics are traveling, cosmetics, sports and health, everyday life, psychology, IT and tech, politics, social aspects (13 texts per a topic). Three main types of blogs—news, commentary, journal (diaries) are presented.

3.2. Blogs segmentation

Though typical elementary discourse units (EDUs) are clauses, according to Ru-RSTreebank annotation rules, certain types of subclausal EDUs are possible in all genres and registers. These are prepositional phrases for cause, purpose etc.: e.g. *из-за переездов* 'due to relocation'.

Besides, blogs authors sometimes separate text fragments lacking overt predicates by sentence punctuation marks (e.g. *Зря.* 'In vain' or *Конечно.* 'Of course'). Some of these fragments are similar to subclausual EDUs, used in spoken discourse, e.g. parcellation or increments. These segments are considered as EDUs according to annotation rules. Moreover, there are many borderline cases when an EDU consists of a noun phrase in Nominative. These EDUs are treated as full clausal units with a zero copula: *Остановка автобуса. 5 утра. Ни одного человека на улице.* 'A bus stop. 5 o'clock in the morning. Nobody on the street.'

For experiments, we excluded headlines, bibliography and other metainformation from texts, as well as vocatives and politeness formulas (e.g. 'Thank you all! Good night', 'And how are you doing?'). We also excluded markers for images (IMG) in blogs. This is a controversial issue since some pictures in blogs can be considered as EDUs participating in meaningful discourse relations (e.g. Evidence). IMG can be a part of an EDU, where gestures or deictic indication would be used in speaking, so eliminating IMGs in this case would result in the incoherent text. We keep this type of references to images as part of EDUs.

3.3. Data preprocessing

For our research, we enhance the annotation of Ru-RSTreebank by adding syntactic and morphological layers (Universal dependencies standards). We used DeepPavlov library (http://deeppavlov.ai/) for this task (a pretrained model for Russian—ru_syntagrus_joint_parsing). General statistics, involving morphological and syntactic properties of discourse units, is based on the automatic annotation (accuracy reported for DeepPavlov morphological tagging 96.23). We assume that parsing errors do not affect the differences in relative frequencies of parameters among genres.

Besides, we added four other discourse-related layers:

1) **morphological type of EDU** (finite or non-finite EDUs dependent on grammatical form of the predicate);
2) **syntactic type of EDU**: prepositional phrase, subordinate clause, coordinate clause, etc.;
3) **the type of subclausal EDUs** (parcellation, external topics, splits etc. according to [Kibrik, Podlesskaya eds. 2009] classification;
4) parenthetical and other **types of intervening constructions**.

4. EDU structural properties

4.1. Rhetoric relations in blogs as compared to other genres

We examine the frequency distribution of different rhetoric relations among genres, taking into account the relations between all level discourse units (not only EDUs). The Elaboration relation is the most frequent one for all three genres—scientific articles, news and blogs (more than 30% of all the segments). General statistics for relations is given in Fig. 1.

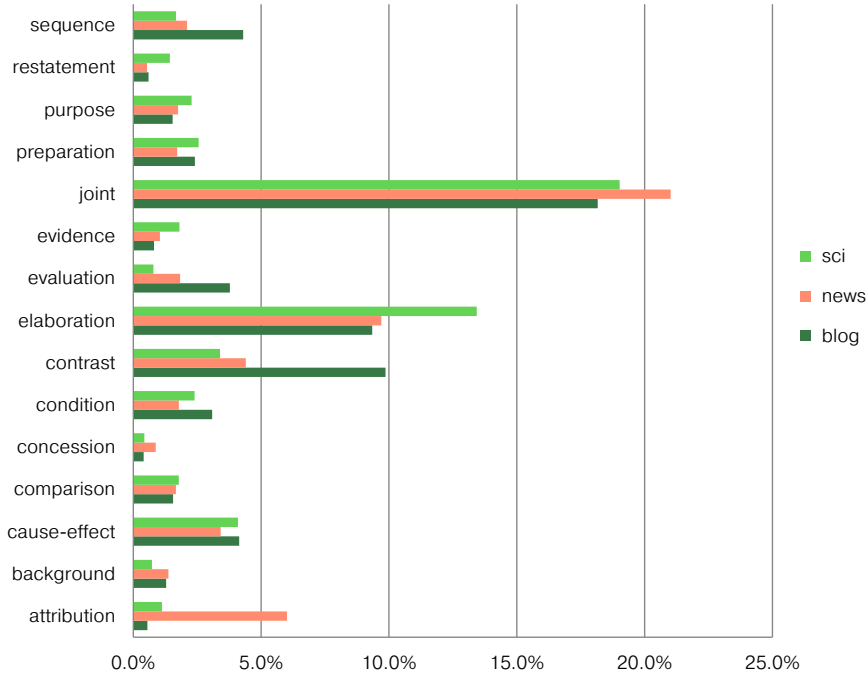

Fig. 1. Rhetoric relations relative frequency for different genres

Fig. 1 shows that for all genres the Joint and Elaboration relations are the most frequent ones. While Attribution is a specific relation for news, and Elaboration is more frequent in scientific texts, blogs are characterized by higher proportion of Sequence, Evaluation and Contrast as compared to other genres. In order to find out the impact of particular relations for genre differentiation, we built a RandomForest classifier with rhetorical relations as features (f1—82%). The relations with the highest impact are Contrast (0.15), Attribution (0.14), Evaluation (0.12) and Elaboration (0.8).

The high proportion of relations with no overt marking (Contrast and Sequence) reflects the tendency for fragmentation. The large number of Evaluation is an evidence of high degree of 'involvement'. Thus, both tendencies for spoken discourse, named by W. Chafe ([Chafe 1982], see also **2.3**) are supported by the rhetoric relations distribution in blogs.

4.2. EDU length

Taking into consideration the tendency for fragmentation in spoken discourse, one would expect a smaller number of complex sentences in blogs, fewer EDUs per sentence. However, the data in Table 1 shows that there is no statistically significant difference in the number of EDUs per sentence among genres.

Table 1. The average number of EDUs per sentence

Text Genre	Average of sentence segments number	StdDev of sentence segments number
blog	2.12	0.47
news	2.19	0.47
science	2.18	0.44
Grand Total	**2.16**	**0.46**

The difference lies in the length of the EDUs (Table 2):

Table 2. The average number of tokens per EDU

Text Genre	Average of tokens/segment	StdDev of tokens/segment
blog	8.59	1.50
news	11.08	1.93
science	14.75	2.77
Grand Total	**11.02**	**3.03**

While the longest EDUs are in scientific texts, the EDUs in blogs are significantly shorter.

4.3. Morphological and syntactic features of EDUs

Other features referring to the 'fragmentation/integration' opposition are sentence complexity, POS (part-of-speech) distribution and subordinate vs. coordinate conjunctions usage.

4.3.1. Verb and noun forms distribution

Table 3 presents the distribution of finite/non-finite verb forms in EDUs among genres (verbless EDUs are not included into the table):

Table 3. The proportion of different verb forms per EDU

Text Genre	Finite	Participle	Infinitive	Converb
blog	63%	4%	22%	2%
news	67%	7%	22%	2%
science	58%	12%	19%	2%

In general, the distribution of verb forms looks similar. For all text genres, the most widely used grammatical type is a finite verb EDU, while converbs are rarely used in all genres. There is a slight difference in EDUs headed by participles. They are more frequent in scientific texts.

As for nouns, the distribution of nouns per EDU is shown in Table 4:

Table 4. Nouns frequency

Text Genre	Nouns average per EDU
blog	1.92
news	2.90
science	4.13

This data agrees with the expectations that written discourse is characterized by a high proportion of nouns. The rate of nouns in blogs is lower than in news or scientific texts (cf. [Chafe 1982] concerning the opposition of written vs. spoken discourse).

4.3.2. Verbless EDUs

According to 4.1, there is no big difference in verb forms distribution across genres. The main difference is in EDUs without verbs, see Table 5 (we do not include sentences, erroneously parsed as rootless, though many of them are also clauses with no overt copula):

Table 5. The average number of root sentences without verbs per EDU

Text Genre	Average of root sentences without verbs per EDU
blog	18.01%
news	8.96%
science	10.98%
Grand Total	**12.55%**

There are verbless EDUs in news and science subcorpora. These EDUs are often prepositional phrases signaling purpose or cause relations. Another type of verbless EDUs are clauses with no overt copula. In scientific texts, these are primarily definitions or characterizations:

(1) *Аргументация это универсальный феномен.*
Argumentation is a universal phenomenon.

However, the highest proportion of verbless sentences is in blogs. Indeed, there are a lot of sentences with zero copula (see also 5.2):

(2) *Кухня одна на всех.*
lit. Kitchen one for all.

(3) *Медынь, автостанция.*
lit. Medyn, a bus-station.

The high proportion of EDUs consisting of noun phrases only is also an evidence for high degree of fragmentation in blogs.

4.3.3. Subjectless EDUs

Based on general assumption that scientific writers often try to avoid the use of personal expressions and to demote human subjects, we expected the high proportion of subjectless clauses in the scientific subcorpus. However, our research has shown that in the blogs subcorpus the proportion of subjectless clauses is the highest one (Table 6):

Table 6. Subjectless main clauses

Text Genre	No subject + root
blog	17.12%
news	7.27%
science	9.92%
Grand Total	**4341**

While the subjectless clauses in scientific texts or news are often impersonal clauses (e.g. *было показано* '(it) was shown'), in blogs the majority of the clauses of this type are clauses with an anaphoric or a personal zero pronoun (pro) as Subject (or with pronominal ellipsis), as in *Встала рано. Вышла ровно в 6* '(I/she) got up early in the morning. (I/she) went out at 6 a.m.'

4.3.4. Impact of POS and dependency relations distribution on genre differentiation

In order to check features impact on genre differentiation, we built a RandomForest classifier based only on POS and some syntactic relations features (Accuracy: 0.84 (± 0.02)). The top 11 features (out of 25) are presented in Table 7.

Table 7. POS and syntactic relation features impact

Feature	Impact
number of nouns per EDU	0.14
number of tokens per EDU	0.10
relative frequency of adverbs	0.07
number of parataxis relations per EDU	0.07
relative frequency of pronouns	0.07
relative frequency of verbless EDU	0.05
relative frequency of particles	0.05
relative frequency of EDUs started with a coordinative conjunction	0.05
relative frequency of prepositional phrases (CASE syntactic relation)	0.04
relative frequency of subjectless clauses	0.04
relative frequency of nouns	0.04

Table 7 shows that ranking of features goes hand in hand with the analysis of features suggested above. The most important features are number of tokens per EDU, number of nouns per EDU (these are highly correlated features), the proportion of verbless and subjectless per text.

Another significant difference between blogs as opposed to news and scientific texts is the distribution of subordinate vs. coordinate conjunctions. While there is no significant difference in subordinate conjunctions distribution, the number of EDUs started with coordinate conjunctions is a quite important feature (20% EDUs in blogs, 12% in news and 8% in scientific texts), as well as number of particles per token (4% of tokens in blogs, 2% in news, and only 1% in scientific texts). The high proportion of coordinative conjunctions and particles in blogs agrees with the high number of Contrast relations in blogs.

5. Parallels between blog and spoken EDUs features

5.1. Non-canonical EDUs in blogs

5.1.1. Speech disfluency and prosodic features

Some of the prosodic features of spoken discourse can be mapped into special cases of punctuation in blogs. While incomplete EDU boundaries in speech are usually detected via pauses and prosodic contour, the speaker in blogs often uses punctuation marks for separating the parts of a clause (a phenomenon similar to parcellation or incrementation):

(4) *Вот, встретили по пути борзую.* ***В махровых тапочках.***
Here, we met a greyhound along the way. **In terry slippers**.

Besides, speakers use ellipses marks in blogs as an alternative of hesitation markers.

5.1.2. "Quasi-echo" EDUs

One of the constructions imitating spoken phenomena in blogs is a sequence of incomplete clauses (with an ellipses mark) with the repetition of the same idea:

(5) a) ***Знаете, такое ощущение…***
b) ***вот мне почему-то кажется, что это…***
c) ***нет, ну я могу ошибаться…*** *А что если это связано с новым президентом?*
a) You know, **I have such a feeling…**
b) for some reason, **it seems to me that…**
c) no, **well, I could be wrong…** What if this could be related to the new president?

This construction can imitate false start (cf. (a) and (b) in (5)).

The EDUs in this construction can contain regulatory markers and markers of interaction (e.g. *вот* (interaction marker), *ну* 'well', *просто* 'just', *знаете* 'you know', etc.). Their function is to express uncertainty or to focus the attention on a particular DU.

5.1.3. Focus/topic extraction

An isolated noun phrase can precede a clause where the same noun phrase is an argument (10 examples in blog subcorpus). It can be either in focus or in topic in this clause. Author uses topic repetition to highlight the main topic.

(6) ***Факты.*** *Всегда смотрите на факты.*
Facts. Always look at the facts.

Another case is topic extraction with corresponding noun phrase ellipsis:

(7) ***Внутренние монголы.*** *Путешествуют всегда организованными группами.*
Inner Mongols. Travel always in organized groups.

5.1.4. Retrospective subclausal EDUs

Blog-writers use retrospective subclausal EDUs mostly as "adjuncts or attributes that semantically belong to a clause but constitute a separate short EDU" [Kibrik 2015: 229]. There are 101 examples of retrospective EDUs in blog subcorpus (cf. 12 cases in news, five are in reported speech):

(8) *И каждый менеджер уникален.* ***Как снежинки.*** *Уникальные люди-снежинки.*
And every manager is unique. **Like snowflakes.** Unique snowflake people.

5.1.5. Noun phrase chaining

Another construction, simulating spoken discourse features, is nominal clause chaining or chaining of noun phrases with no overt copula (their exact syntactic status is often unclear):

(9) *Отель-чемодан-метро. Тайм-чек.*
Hotel-suitcase-metro. Time check.

(10) *Моя кожа: 29 лет, комбинированный тип, чувствительная, акне в ремиссии, пост-акне, быстро забиваются поры.*
lit. My skin: 29 years old, mixed type, sensitive, acne in remission, post-acne, pores are quickly logged.

5.1.6. Mixture of predicate types in coordinate constructions

One more construction of special interest is a multinuclear relation where syntactically heterogeneous phrases (noun phrases, finite clauses, etc.) form a coordinating construction.

(11) ***Электричка*** *"Стандарт-плюс".* ***Цивильная*** *и комфортная, с мягким ходом,* ***есть*** *вай-фай. Два часа блаженства и резкий контраст с тем, что будет дальше.*
Electric train Standard plus. Civil and comfortable, with a smooth ride, there is Wi-Fi. Two hours of bliss and a sharp contrast with what will happen next.

5.1.7. Splits

There are special types of splits in blogs that are marked with three dots or exclamation marks, interjections or emoticons (77 cases)

(12) ***А ты мама…*** *вот тогда злилась на меня, когда мы опаздывали на электричку!*
And you mom... then got mad at me when we were late for the train!

(13) ***Ой:)*** *сейчас же в моде такие словечки…*
Oh:) these words are now in fashion ....

5.2. Embedded parataxis constructions

The parenthetical phrases intervening into EDUs occur in all three genres (blogs—528 (3.7% of EDUs), news 482 (3.9%), scientific texts—592 (6.8%)). Table 8 presents the distribution of factual vs. evaluative EDUs occurring in parenthesis:

Table 8. The distribution of different types of parenthetical constructions among genres

Parameter / text genre	blogs	news	scientific texts
factual information: elaboration / disclosure	266	513	580
evaluation / interpretation	189	43	12
Total	455	546	592

In news, the function of phrases in parenthesis is to enforce author's interpretation of facts. Among the three genres, the parenthetical constructions most often occur in scientific texts. Interactive communication with the audience in scientific texts is represented only by reduced scientific abbreviation-clichés: *ср*. 'cf.', *см*. 'see', etc.

In blogs, parenthetical constructions are used as blog-writer "protocols". They often include evaluative expressions:

(14) *Сладости я оставила только без сахара и органический шоколад в очень умеренных дозах* ***(чтоб не так хотелось убивать людей)))***.
I kept only sweets without sugar and organic chocolate in very moderate doses **(not to want to kill people)))**.

Many of them contain irony:

(15) *Газовщица заполнила договор* ***(один экземпляр)****, теща подписала и спрашивает* ***(все-таки бывший директор школы)****, а ей экземпляр?*
The gas worker filled out the contract **(one copy)**, the mother-in-law signed and asked **(after all, the former principal)**, but what about her copy?

Moreover, sometimes they are markers of interaction:

(16) *Сегодня будем размышлять над следующим изречением* ***(тем, кто не знает иностранных языков — Google в помощь)***.
Today we will reflect on the following saying **(for those who do not know foreign languages—Google can help (you))**.

6. Conclusion

To sum up, our research shows the difference in discourse properties of the three subcorpora in Ru-RSTreebank. The distribution of rhetoric relations differs among genres.

Our corpus study provides an additional evidence for the claim that blogs have spoken discourse features. The rhetoric relations distribution, the occurrence of subclausual EDUs of certain types, the part of speech distribution differ blogs from the other two genres, science articles and news. There are special devices that are used by blog authors to imitate the spontaneous speech.

References

Examples of sources for blog texts:

- http://www.livejournal.com
- https://medium.com
- http://www.blogspot.com
- https://kosmetista.ru
- https://habr.com
- https://airfit.ru/blog

1. *Akhapkina Ya. E.* (2014), About the grammar of oral-written speech [O grammatike ustno-pis'mennogo viskazivaniya],Contemporary Russian language in Internet, pp. 181–194.
2. *Berger P., Hennig P., Schoenberg V., and Meinel C.* (2015), Blog, forum or newspaper? Web genre detection using SVMs, 2015 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 3, pp. 64–68.
3. *Burton K., Java A., and Soboroff I.* (2009), The icwsm2009 spinn3r dataset, Third Annual Conference on Weblogs and Social Media (ICWSM 2009).
4. *Castellà J. M. Oralitat i Escriptura.* (2004), Dues Cares de la Complexitat del Llenguatge, Publicacions de l'Abadia de Montserrat, Barcelona.
5. *Chafe, Wallace* (1982), Integration and involvement in speaking, writing, and oral literature, Spoken and written language:Exploring orality and literacy, ed. D. Tannen, Norwood: Ablex, pp. 35–54.
6. *Chistova, E., Shelmanov, A., Kobozeva, M., Pisarevskaya, D., Smirnov, I., and Toldova, S.* (2019), Classification models for RST discourse parsing of texts in Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019", pp. 163–176.
7. *Crible L., Cuenca M. J.* (2017), Discourse markers in speech: characteristics and challenges for corpus annotation, Dialogue and Discourse, Vol. 8(2), pp. 149–166.
8. *Da Cunha, I.* (2016), Towards discourse parsing in Spanish, TextLink–Structuring Discourse, Multilingual Europe Second Action Conference Károli Gáspár University of the Reformed Church in Hungary, Budapest, 11–14 April, p. 40.
9. *Germasheva T. M.* (2010), Studies of linguistic and paralinguistic features of blogs discourse [Issledovanie lingvisticheskih i paralingvisticheskih kharakteristik blog-discursa], Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A. I. Gertsena, 126, pp. 150–155.

10. *Iruskieta, M., Da Cunha, I., and Taboada, M.* (2015), A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora, Language resources and evaluation, 49(2), pp. 263–309.
11. *Kibrik A. A.* (2009), Modus, genre and other parameters of discourses classification [Modus, ganr i drugie parametri klassifikatsii discursov], Voprosi yasikoznaniya, 2 (3).
12. *Kibrik A. A.* (2015), The problem of non-discreteness and spoken discourse structure, Computational linguistics and intellectual technologies, 14 (21), vol. 1, pp. 225–233.
13. *Lapteva O. A.* (1976), Russian spoken syntax [Russkiy razgovorniy sintaksis], Moscow, Nauka.
14. *Macdonald C., Ounis I.* (2006), The trec blogs06 collection: Creating and analysing a blog test collection. Department of Computer Science, University of GlasgowTech Report TR-2006–224, 1:3–1.
15. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, pp. 243–281.
16. *Matthiessen C. M., Teruya K.* (2015), Grammatical realizations of rhetorical relations in different registers, Word, 61(3), pp. 232–281.
17. *Mishne G. et al.* (2005), Experiments with mood classifi-cation in blog posts, Proceedings of ACM SIGIR 2005workshop on stylistic analysis of text for information access, Vol. 19, pp. 321–327.
18. *Novikova E. G.* (2005), Language features of text structuring in classic and online diares [Yazikovie osobennosti organizacii tekstov klassicheskogo i setevogo dnevnikov], AutoAbstract of Ph. D. Thesis, Stavropol.
19. *Pisarevskaya D. et al.* (2017), Towards building a discourse-annotated corpus of Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017", pp. 194–204.
20. *Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., and Webber, B.* (2008), The penn discourse treebank 2.0, Proceedings of the 6 International Conference on Language Resources and Evaluation, LREC'08, pp. 2961–2968.
21. *Quan C., Ren F.* (2009), Construction of a blog emotion corpus for chinese emotional expression analysis, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol. 3, pp. 1446–1454.
22. *Rehbein I., Scholman M., and Demberg V.* (2016), Annotating Discourse Relations in Spoken Language: A Comparison of thePDTB and CCR Frameworks, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1039–1046.
23. *Santos H., Woloszyn V., Vieira R.* (2018), BlogSet-BR: A Brazilian Portuguese Blog Corpus, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 661–664.
24. *Scheffler T., Aktaş B., Das D., and Stede M.* (2019), Annotating Shallow Discourse Relations in Twitter Conversations, Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pp. 50–55.

25. *Simaki V., Skeppstedt M., Paradis C., Kerren A., and Sahlgren M.* (2017), Annotating Speaker Stance in Discourse: The Brexit Blog Corpus, Corpus Linguistics and Linguistic Theory, 34 pages. Available at https://portal.research.lu.se/ws/files/34256897/_Corpus_Linguistics_and_Linguistic_Theory_Annotating_Speaker_Stance_in_Discourse_The_Brexit_Blog_Corpus.pdf.
26. *Simaki V., Paradis C., and Kerren A.* (2018), Evaluating stance-annotated sentences from the Brexit Blog Corpus: A quantitative linguistic analysis, ICAME Journal, Vol. 42, pp. 133–165.
27. *Night dream stories:* Russian Corpus Study of Oral Discourse (2009) [Rasskazy o snovideniyah: korpusnoe issledovanie ustnogo russkogo diskursa], Ed. Kibrik A. A., Podlesskaya V. I.
28. *Tonelli S., Riccardi G., Prasad R., and Joshi A.* (2010), Annotation of discourse relations for conversational spoken dialogs, Proceedings of the 7 International Conference on Language Resources and Evaluation, LREC'10, pp. 2084–2090.
29. *Wang X., Gyawali B., Bruno J. V., Molloy H. R., Evanini K., and Zechner K.* (2019), Using Rhetorical Structure Theory to Assess Discourse Coherence for Non-native Spontaneous Speech, Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pp. 153–162.
30. *Zeldes A.* (2017), The GUM corpus: Creating multilayer resources in the classroom, Language Resources and Evaluation, 51(3), pp. 581–612.