Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2016)

Выпуск 15

Computational Linguistics and Intellectual Technologies

Proceedings of the Annual International Conference “Dialogue” (2016)

Issue 15

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает искренню благодарность Российскому фонду фундаментальных исследований за финансовую поддержку

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). — М.: Изд-во РГГУ, 2016.

Сборник включает 68 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2016», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

COREFERENCE IN RUSSIAN ORAL MOVIE RETELLINGS (THE EXPERIENCE OF COREFERENCE RELATIONS ANNOTATION IN "RUSSIAN CLIPS" CORPUS)[1]

Toldova S. Yu. (stoldova@hse.ru),
Bergelson M. B. (mbergelson@hse.ru),
Khudyakova M. V. (mkhudyakova@hse.ru)

National Research University "Higher School of Economics",
Moscow, Russia

The work deals with adapting the Russian coreference corpus RuCor annotation system (used for written Russian) to the corpus of Russian oral narratives from the Russian Clinical Pear Stories Corpus (Russian CliPS) (Khudyakova et al., 2016). Russian CLiPS is a corpus of Russian "Pear stories" movie (Chafe, 1980) retellings in clinical populations as compared to neurologically healthy people. The analysis deals with 11 texts by healthy people and 9 texts by people with various types of aphasia. The focus is on the specificity of reference choice in oral retellings and the parameters to be used for the annotation procedure to register deviations in referential choice in spoken discourse as compared to the written one. The specific features for annotation of referential choice in clinical populations are also under discussion. The main claims are as follows. Certain types of speech disfluencies should be integrated into the coreference annotation scheme. These are noun phrases, which are repetitions of a previous referent mention, referent renaming, or name correction. Such occurrences can influence the referent activation; on the other hand, they could shed some light on the process of the referential expression choice. The NP morphosyntactic structure and zero-anaphora should have more granulated set of features for coreference devices, as they are more diverse in spoken discourse. Moreover, certain structures, such as adjectives postposition etc. and some types of zeros are characteristic of referential expressions in spoken discourse.

Keywords: coreference, oral retellings, coreference corpus annotation

[1] The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016 (grant №16-05-0024) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

КОРЕФЕРЕНТНЫЕ ОТНОШЕНИЯ В РУССКИХ УСТНЫХ ПЕРЕСКАЗАХ (ИЗ ОПЫТА РАЗМЕТКИ КОРЕФЕРЕНТНЫХ ОТНОШЕНИЙ В КОРПУСЕ «RUSSIAN CLIPS»)

Толдова С. Ю. (stoldova@hse.ru),
Бергельсон М. Б. (mbergelson@hse.ru),
Худякова М. В. (mkhudyakova@hse.ru)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Статья посвящена опыту разметки кореферентных связей в корпусе устных пересказов Russian CliPS (Khudyakova et al., 2016). Корпус представляет собой пересказ фильма о грушах (Chafe, 1980). В статье представлен анализ параметров, которые необходимо учитывать при разметке такого рода текстов. В результате анализа данных, мы предлагаем подходить к разметке кореферентных связей в устных текстах с позиции взаимодействия разных систем: собственно кореферентных цепочек в нарративе, элементов речевых сбоев (например, случаев переименования референта и др.), а также элементов интеракции (например, оценка говорящим степени уверенности в выбранной номинации).

Ключевые слова: кореферентные отношения, устные пересказы, кореферентная аннотация корпуса

1. Introduction

When producing a coherent text, a speaker can use different linguistic devices (NPs) to name an entity (referent): full NPs (*a man, the man with the goat, that man*), anaphoric pronouns (*he, his*), and zero pronouns. When comprehending the text, the listener must make a decision regarding whether a certain NP introduces a new referent in the discourse or relates to a previously mentioned referent. Establishing coreference relations in discourse is a complex process, which depends on various cognitive, discourse and grammatical factors. To study these factors in their interaction corpora with coreference annotation are needed. Recently, the task of creating such corpora not only for written texts, but also for different genres of spoken discourse has become topical.

One of the aims for coreference annotation in spoken discourse within the NLP paradigm is the frequency distribution of basic types of referring expressions (full NPs vs. underspecified expressions such as pronouns and zeros) and the distribution of different features used by coreference resolution systems in spoken discourse in comparison to written texts. For this purpose, spontaneous speech undergoes a kind of normalization when different types of disfluencies are removed from the texts submitted to annotation process (for speech corpus normalization see (Fitzgerald and Jelinek, 2008; Hajič et al., 2008). Various kinds of disfluencies are annotated and studied separately at a separate level of annotation (Heeman et al., 2006).

Our study is based on the material of Pear story film retellings (Chafe, 1980) by healthy speakers of Russian and people with aphasia (PWA)—a language pathology resulting from damage to the language-dominant hemisphere of the brain. Each aspect of this topic has been well researched before. In the area of automatic text processing the task of developing corpora that include coreference annotation has been important for several decades (see for example the manual for coreference annotation (Chinchor and Robinson, 1997; Hirschman et al., 1997). A significant number of studies focus on different parameters and mechanisms involved in referential choice (see Fedorova, 2014; Kibrik, 2011 inter alia). The problem of Russian spoken discourse transcription, annotation, and analysis is covered by Kibrik and Podlesskaya (2009). The pear film has been used for four decades as elicitation stimulus for collection and analysis of narratives in a number of typologically different languages (Chafe, 1980; Erbaugh, 1990; Fedorova, 2014). Pear film retellings by English speakers are a part of a corpus with coreference annotation—ARRAU (Poesio and Artstein, 2008).

Unlike written discourse, spoken discourse has certain distinct features (Biber et al., 1999; Kibrik, 2009) such as hesitation pauses, self-corrections, discourse markers, and markers of word-finding difficulties (Bergelson et al., 2015; Podlesskaya and Kibrik, 2007; Shriberg, 1994). These disfluencies can affect the process of referent naming or the assessment of its prominence (extra referent mentioning attracts more attention to it and thus influences the referent prominence assessment). In clinical linguistics domain, analysis of reference in speech pathologies is not a common topic for research. The studies focus on finding differences between brain-damaged and healthy groups in the frequencies of basic classes of referential devices (full NP/anaphoric pronoun/zero pronoun) (Peng, 1992; Romanova, 2010), or on pronouns as means of establishing cohesion and coherence (Davis and Coelho, 2004). There are even more cases of disfluencies in the speech of the brain-damaged populations. The focus of this study as compared to the above mentioned is on the parameters that must be accounted for in coreferential chains annotation under the following condition: our data is comprised of the text retellings, not spontaneous production, including retellings by the brain-damaged individuals. We suggest some features to be employed for registering differences in written, spoken and clinical discourse.

The aim of this work is to describe issues that arise when the annotation scheme designed for written texts is adapted for spoken discourse analysis. In particular, we analyze the specific features of referring expressions in spoken discourse, including possible disfluencies and errors related to the referential choice.

2. Method and material

2.1. Participants and procedure

As mentioned above, our study is conducted on narratives from Russian CliPS (Clinical Pear Stories) corpus which contains Pear film (Chafe, 1980) retellings by people with aphasia (PWA) and neurologically healthy adults. The recorded narratives were transcribed and annotated with attention to speech failures and disfluencies in ELAN[2]. The narrative recording procedure, information about the speakers, and annotation scheme is described in (Khudyakova et al., 2016).

2.2. Coreference subcorpus

For the current study 11 texts by healthy speakers (norm) were chosen for developing annotation principles, and those principles were then applied to 9 texts by PWA. We have chosen texts by people with acoustic-mnestic and efferent motor aphasia (for description of aphasia types see, for example Akhutina, 2015; Luria and Hutton, 1977). Although PWA have deficits on micro-linguistic level, the narrative structure and coreference relations can be established (Marini, 2012). Lexical transcripts (with no annotation for pauses) of the texts were run through automatic lemmatizer and morphological analyzer[3]. The general statistics is shown in Table 1.

Table 1. The general statistics for the experimental corpus

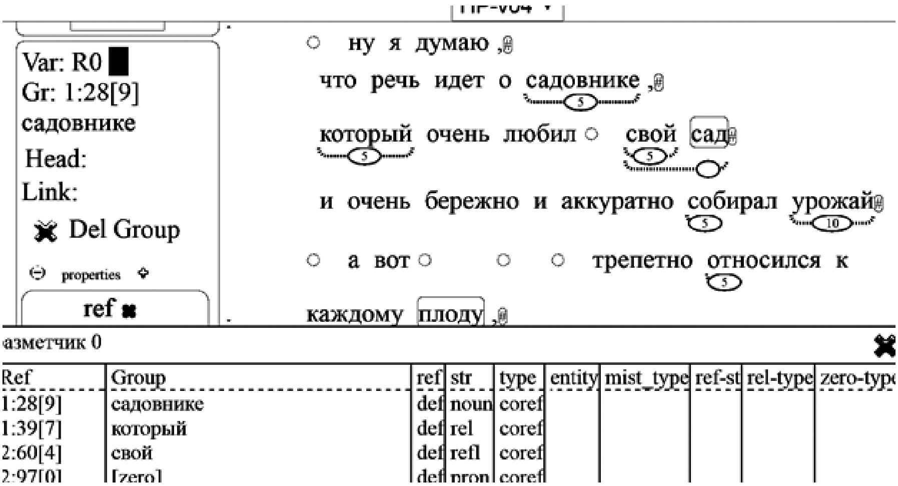
	Healthy speakers	**PWA**
Number of texts	11	9
Min length in tokens	106	233
Max length in tokens	391	419
range	285	186
median	299	302
total	3,324	2,934

2.3. Annotation tool

As a starting point we have chosen to use the annotation scheme and annotation tool of RuCor corpus that was created for RU-EVAL forum on automated anaphora and coreference resolution (Toldova et al., 2014; http://ant0.maimbava.net/). Figure 1 demonstrates a fragment of the coreference annotation tool.

[2] https://tla.mpi.nl/tools/tla-tools/elan/

[3] We used Treetagger and lemmatizer for Russian http://corpus.leeds.ac.uk/mocky/



○ ну я думаю ,
что речь идет о садовнике ,
который очень любил ○ свой сад
и очень бережно и аккуратно собирал урожай
○ а вот ○ ○ ○ трепетно относился к
каждому плоду ,

Ref	Group	ref	str	type	entity	mist_type	ref-st	rel-type	zero-typ
1:28[9]	садовнике	def	noun	coref					
1:39[7]	который	def	rel	coref					
2:60[4]	свой	def	refl	coref					
2:97[0]	[zero]	def	pron	coref					

Figure 1. Annotation of coreference relations in the coreference annotation tool

When a coreference relation is established between two NPs, the same coreference chain ID is assigned to both NPs (see NP ***садовник*** ‘*a gardener*’, ***свой*** ‘*his own*’ and zero pronoun with the verb ***берет собирал*** ‘collected’; they all have index 5 on the arc). The tool allows manual annotation of NP’s head and embedded NPs, as well as assigning values for different NP features (the assigned values can be seen in the table on Figure 1). In case of ambiguity, it is possible to link NP to several coreference chains (cf. annotation scheme for “Pear stories” in Poesio and Artstein, 2008).

3. Annotation principles adaptation

To our knowledge none of the papers on coreference chain annotation in spoken discourse (see for example Poesio and Artstein, 2008) discussed the problem of annotating self-corrections, false starts etc., despite the fact that experimental research on reference shows that speech failures affect production and comprehension processes. For example, filled hesitation pauses in case of referential conflict facilitate its resolution and the choice of a newer referent as antecedent (Arnold et al., 2003).

In our coreference in spoken discourse annotation scheme we tried to pay attention to these phenomena and annotate not only specific features of coreference chain annotation, but also speech failures related to production and choice of an appropriate referential expression.

3.1. Markable boundaries

Many international standards for annotation of written texts define a markable as the maximally full NP up to the nearest comma to the right (see for example

Krasavina and Chiarcos, 2007). However, spoken discourse has certain features that do not allow using this criterion.

Markable borders in case of sequential nomination of a referent (‘renaming constructions’)

There are cases where two NPs denoting the same entity participate in so-called “renaming” constructions, e.g. an alternative construction, as in (1). Such constructions are means to express overtly the mental process of referent naming, that is—seeking a proper referring expression, correcting or refining the chosen one, as in (2) (see Bergelson et al., 2015 for more detailed discussion).

(1) [*мальчик*] [*или парень*]
‘a boy [*or a guy*]*’*

(2) [*груши*], [*или, как там их, грушины*]
‘pears, [*or how do you call them, pear-things*]*’.*

Such cases pose a problem for classical approach to coreference annotation. They are not coreferential expressions in the proper sense of the word, but they also can influence referential choice by being a factor of additional referent activation (see Givon’s notion of topicality in Givón, 1983).

We decided to define as separate markables all NPs that name one referent in one point of discourse, and to establish a special type of relation between them rather than include them in the chain as coreferent NPs. The connectives, discourse markers and parenthetical words are also included into markable.

Markable boundaries under non-standard syntactic environment

In retellings various discourse markers can be embedded in or adjacent to NPs though their standard syntactic position is the sentential modifier. These markers can be attributed to the degree of confidence of referential choice or interaction components of the discourse:

(3) *поглощающие по внешнему виду его груши* (HP-v02)
‘Consuming as it looks his pears’

Markables in case of postpositional adjectival modifiers: apposition vs. an entire NP

Adjectival phrases in postposition to the head in a referring expression pose a separate problem. When annotating written texts, one would normally use a principle of the ‘left’ punctuation border. In retellings speakers not only verbalize the procedure of choosing the most appropriate referential device, but also the procedure of ‘attributive description choice’ for the referent. That is why adjectival phrases in postposition are quite common in oral narratives:

(4) *… и прошли как раз мимо* [*хозяина груши этой большой*]
Lit. ‘And passed by the owner [*of the pear tree this big*]*’* (c.f. this big pear tree)

In spoken discourse annotation the "left punctuation border" criterion cannot be applied. Postpositive adjectival phrases (such as (4) can be interpreted as parcellation, however, when descriptions in postposition appear in written texts, they are usually characterized by a certain syntactic structure (see Ljutikova, 2015). Our decision was to place such descriptions into the same markable unless any specific signs of border (e.g. declining intonation and long pause) are present. However, postposition of the adjectival phrase is reflected in the NP morphosyntactic structure parameter.

Split NPs

In spoken discourse split NPs are more prevalent, as well as non-projectiveness and non-canonical word order, as in (5):

(5) *А куда* [*корзина-то*]$_{gr1}$ *делась* [*одна*]$_{gr1}$ (HP-v03)
And somewhere [*the basket*]$_{gr1}$ *has gone4* [*one*]$_{gr1}$

In (5) the numeral *одна* '*one*' is in postposition to its head (cf. *одна корзина* '*one basket*') and is separated from it by the verb.

3.2. Entity types

The nature of texts in the corpus, which are retellings of the same film makes it somewhat easier to analyze coreference, because the characters of the story are the same for all narratives. The speaker can pick an inadequate referential device, use a deictic device *вот этот* '*this* one' or anaphoric pronoun *он* '*he*' without introducing the referent, but in this case the annotator would still be able to refer it to the appropriate NP based on the context and the annotator's knowledge of the story. Besides, in the film a number of characters and objects belong to the same ontology class and can be referred to by the same name, which allows for the referential conflict to appear not only in case of anaphoric pronouns, but also full NPs.

To differentiate entities in the narrations we included the special labels such as 'man', 'baskets_man', 'boy', 'three_boys' etc. This type of annotation allows us to compare the referential expressions used by different speakers for the same entities, and different expressions used for the entities from the same ontological class.

3.3. Morphosyntactic types of NPs

Our investigation of the NP morphosyntactic properties for markables in the corpus has shown that there are some special cases that are characteristic of oral narration both by PWA as well as healthy speakers.

Occurrences of NPs with demonstratives as the introductory NPs, e.g. *И вот вдруг приехал этот мальчик* '*And suddenly this boy came*') are not common in written texts. Moreover, the NPs with demonstratives are rare in Russian news texts (Nedoluzhko et al., 2015).

Another peculiarity that was revealed for referential devices in spoken discourse is that a parenthetical word can serve as a prenominal modifier, as in (6) and (7):

(6) *залез на, видимо, груши* (HP-v01)
'Got onto, obviously, pear'

(7) *ну подъехал его, наверное, его сын младший*
'And came his, maybe, his son younger'

As it was mentioned earlier, an important parameter is the place of the modifier relative to the NP head *груши спелые 'pears ripe'* vs. *спелые груши 'ripe pears'*.

(8) *… в* [*такой шапке летней*]
'… in [*such a hat summer-y*]

The following morphosyntactic types taken from RuCor annotation scheme are retained: NPs with demonstratives, NPs with other modifiers (adjectives, numerals, indefinite pronouns); bare nouns, anaphoric pronouns (both 3rd person pronouns and reflexives), relative pronouns and zero pronouns (pro). In order to check some specific features of NPs used in oral retellings we use the more detailed classification of NPs: the type of pronoun or numeral is taken into consideration, the type of modifiers, as well as the type of word order.

3.4. Types of zeroes

Russian is a so called pro-drop language, that is finite clauses with no overt subject are possible as well as 'omitted' anaphoric pronouns in some other positions. While zero subjects are very rare in news texts, zero subjects chaining is a standard strategy for spoken discourse. Number of zero pronouns can be an important parameter in the analysis of pathological discourse compared to healthy discourse. Each predicate belongs to an elementary discourse unit (EDU), and, thus, we restore zero subjects for all the verb forms with no overt subjects.

We added distinction of different types of zero pronouns into the annotation scheme: syntactically motivated zeroes, conjunction zeroes, subject zeroes in separate clauses.

3.5. Link types in chains: annotation of non-coreference relations between NPs

Naming relations: renaming and speech disfluences

There are cases in corpus when NPs do not refer directly to an entity rather they denote the process of referent's naming or they represent speech disfluencies that can affect the degree of entity activation (for the factors influencing referemt's a\ctivation see Kibrik A. A. 2011): false-starts, repetitions, self-corrections (10), (11); name elaboration (12) and alternative naming (9):

(9) *Плоды* [*груши*] [***или*** *авокадо*] (HP-v02)
‘Fruit of pears or avocadoes’

(10) *Яблоки,* [***точнее****, груши*]
‘Apples, no, pears’

(11) *И поставили корзину*$_i$ *на витрину*$_j$*, ведро*$_i$ *на велосипед*$_j$ (AP-v01)
‘And put the basket on the window bucket on the bike’

(12) *Ребята* ***ну*** *друзья его из деревни* (AP-v07)
‘Guys well friends his from the village’

For some occurrences of referential expressions, NP elaboration is hard to distinguish from NPs in postposition (12).

Self-corrections may apply to a NP due to the wrong choice of its referential status:

(13) *Эту грушу каждую грушу вытирал*
‘This pear (wrong referential expression) *every pear*

Special cases of apposition links

A special type of apposition is used in spoken discourse (it does not occur in written texts). That is an anaphoric pronoun followed by full NP (14). In spoken discourse they reflect monitoring the process of hearer’s referential expression interpretation by the speaker.

(14) *И забрал* [*ее*]*,* [*корзину*]
And he took [*it*]*,* [*the basket*]

The following basic annotation features are used for written texts: link type (coreference, apposition, predicative (Toldova et al. 2014)). Besides these link types, we have introduced additional values of the feature ‘Link Type’, namely ‘repetition’, ‘self-correction’, false start’, ‘alternative nomination’ and some others.

3.6. Error types

Spoken discourse demonstrates various errors in naming and in choice of NPs, that’s why we introduced a specific parameter to capture these errors. A detailed typology of these errors requires additional research. As mentioned in (Bergelson et al. 2015) referential errors are caused by different mechanisms of speech production. In our initial annotation we pay attention to only basic type of errors:

(a) morphological errors—the speaker chose the wrong number or person agreement marker, or a wrong case marker (15):

(15) *То ли не все положили, остались у него* (у них) *груши в руках*
'As if not all were placed, pears remained in his (in their) *hands'*

(b) wrong lexical choice (semantic paraphasia) (*мешок* '*bag*' instead of *корзина* '*basket*')
(c) wrong choice of referential expression, like *эти мальчики* '*these boys*' instead of *три мальчика* '*three boys*' when introducing the referent.

(16) *Эту грушу каждую грушу вытирал*
'He wiped this pear, every pear'

3.7. Summary statistics

Deviations of the coreference annotation scheme for spoken discourse from that for the written texts reflect specific features of coreference in the former—both the speech of healthy people as well as the aphasic speech.

The comparison of basic morphosyntactic types distribution for written vs. spoken discourse is given in Table 2. The figures for written texts are taken from (Nedoluzhko et al., 2015) where the written text corpus consists of 16 short news texts on political and economic topics (the average length is 30 sentences). The figures are given for the markables including appositions and excluding various kinds of renaming and disfluencies (e.g. repetitions and false starts).

Table 2. The distribution of morphosyntactic types of referential devices in written texts, retellings by neurologically healthy people and by PWA

NP morphosyntactic type		Written texts		Pear Stories			
				Healthy speakers		PWA	
anaphoric and reflexive pronouns	subject position	39	3.8%	148	15%	138	14%
	non-subject position	95	9.3%	145	14%	112	11%
relative		42	4.1%	19	2%	21	2%
zero (pro)		13	1.3%	199	20%	196	20%
bare noun		164	16.0%	338	33%	338	33.1%
NP with a demonstrative		20	1.9%	49	5%	38	4%
Other NPs		652	63.6%	163	18%	131	13%
TOTAL		**1,025**	**100%**	**1,061**	**100%**	**974**	**100%**

As demonstrated in table 2. the reduced referential expressions (pronouns and zero pronouns) are much more frequent in retellings than in written news texts. There is a great difference in the anaphoric pronouns distribution for news texts vs. retellings. However, the difference on pronoun frequency between healthy speakers' texts and PWA texts is not so substantial. The more striking contrast is in zero anaphora distribution (the frequency is 15 times greater in in retellings than in written texts). It is also worth mentioning that the distribution of demonstratives is significantly lower in written texts as compared to spoken discourse.

As for different types of disfluencies, they make approximately 8% of all markables for the healthy speakers and 10% for the PWA.

4. Conclusions

Disfluencies represent one of the most eye-catching features of spoken discourse. They mark the process of speech production directly in the resulting text. Often when performing coreference chains annotation for spoken discourse the text is 'purified' from disfluencies and interaction markers. It means that two objects—a 'normalized' text and various disfluencies are studied as separate systems. At the same time presence of disfluencies in the text has impact on the interpretation of other text elements and also on the speaker's verbalization choices. While adapting the initial coreference annotation scheme we came to a conclusion that besides the referential ambiguity, which is normally taken into account in spoken discourse analysis, and basic taxonomy of the referential devices (full NP vs. anaphoric pronoun vs. anaphoric zero) we need to include there both disfluencies and interactional markers.

Thus, we suggest an approach to the coreference relations annotation of spoken discourse that integrates various phenomena. Those are coreferential narrative chains, disfluencies (like changing the name of the referent) and interactional elements (for instance, speakers' assessment of the correctness of their choice of nomination).

References

1. *Akhutina, T.* (2015). Luria's classification of aphasias and its theoretical basis. Aphasiology, 1–20. doi:10.1080/02687038.2015.1070950.
2. *Arnold, J. E., Fagnano, M., and Tanenhaus, M. K.* (2003). Disfluencies signal theee, um, new information. in Journal of Psycholinguistic Research, 25–36.
3. *Bergelson, M. B., Akinina, Y. S., Dragoy, O. V., Iskra, E. V., and Khudyakova, M. V.* (2015). Markers of word production difficulties in normal and clinical discourse production: continuity of norm in language and discourse [Zatrudnenija pri poroždenii slov v diskurse i ix formal'ные markery: norma i patologija, ili o nediskretnosti normy v jazyke. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue 14, 41–51.
4. *Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E.* (1999). The Longman grmmar of spoken and written English.

5. *Chafe, W.* (1980). The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. , ed. W. Chafe Norwood, New Jersey: Ablex.
6. *Chinchor, N., and Robinson, P.* (1997). MUC-7 named entity task definition. in Proceedings of the 7th Conference on Message Understanding, 29.
7. *Davis, G. A., and Coelho, C. A.* (2004). Referential cohesion and logical coherence of narration after closed head injury. Brain and Language 89, 508–523. doi:10.1016/j.bandl.2004.01.003.
8. *Erbaugh, M. S.* (1990). Mandarin Oral Narratives Compared with English: The Pear/Guava Stories. Journal of the Chinese Language Teachers Association 25, 21–42.
9. *Fedorova, O. V* (2014). Experimental discourse analysis [Eksperimental'nyj analiz diskursa]. Languages of Slavonic Culture.
10. *Fitzgerald, E., and Jelinek, F.* (2008). Linguistic resources for reconstructing spontaneous speech text. in LREC Proceedings (Marrakech, Morocco), 1–8.
11. *Givón, T.* (1983). "Topic Continuity in Discourse: An Introduction," in Topic Continuity in Discourse: A Quantitative Cross-language Study, ed. T. Givón (John Benjamins), 3–41.
12. *Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptáček, J., Toman, J., et al.* (2008). An annotated resource for speech reconstruction. in 2008 IEEE Workshop on Spoken Language Technology, SLT 2008—Proceedings, 93–96.
13. *Heeman, P. a, McMillin, A., and Yaruss, J. S.* (2006). An annotation scheme for complex disfluencies. INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing 3, 1081–1084.
14. *Hirschman, L., Robinson, P., Burger, J., and Vilain, M.* (1997). Automating Coreference : The Role of Annotated Training Data. in Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloqium.
15. *Khudyakova, M. V., Bergelson, M. B., Akinina, Y. S., Iskra, E. V., Toldova, S., and Dragoy, O. V.* (2016). Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals. in LREC Proceedings (Portoroz, Slovenia).
16. *Kibrik, A. A.* (2009). Modus, genre and other parameters of discourse classification [Modus, zhanr i drugie parametry klasifikatsii diskursov]. Topics in the study of language [Voprosy Jazykoznanija] 2, 3–21.
17. *Kibrik, A. A.* (2011). Reference in discourse. Oxford University Press.
18. *Kibrik, A. A., and Podlesskaya, V. I. eds.* (2009). Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijah: korpusnoe issledovanie ustnogo russkogo diskursa]. Moscow: Languages of Slavonic Culture.
19. *Krasavina, O. N., and Chiarcos, C.* (2007). PoCoS: Potsdam coreference scheme. in Proceedings of the Linguistic Annotation Workshop (Association for Computational Linguistics), 156–163.
20. *Ljutikova, E.* (2015). Agreement, features and structure of NP in Russian [Soglasovanie, priznaki i struktura imennoj gruppy v russkom jazyke]. Russian Language and Linguistic Theory [Russkij yazyk v nauchnom osveshchenii] 2.
21. *Luria, A. R., and Hutton, J. T.* (1977). A modern assessment of the basic forms of aphasia. Brain and Language 4, 129–151.

22. *Marini, A.* (2012). Characteristics of narrative discourse processing after damage to the right hemisphere. Seminars in Speech and Language 33, 68–78. doi:10.1055/s-0031-1301164.
23. *Nedoluzhko, A., Toldova, S., and Novák, V.* (2015). Coreference Chains in Czech, English and Russian: Preliminary Findings. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” 14, 456–469.
24. *Peng, V. M.* (1992). The usage of reference items in aphasic and normal conversations. Journal of Neurolinguistics 7, 295–307. doi:10.1016/0911-6044(92)90020-W.
25. *Podlesskaya, V. I., and Kibrik, A. A.* (2007). Self-corrections of the speaker and other types of speech failures as object of annotation in spoken corpora [Samoispravlenija govorjaščego i drugie tipы rečevых sboev kak ob”ekt annotirovanija v korpusax ustnoj reči]. Science-Technical Information [Naučno-texničeskaja informacija] 2, 2–23.
26. *Poesio, M., and Artstein, R.* (2008). Anaphoric annotation in the ARRAU corpus. in Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08) (Marrakech, Morocco).
27. *Romanova, A.* (2010). Referential Choice: Distribution of Subject Types in Russian Aphasic Speech. The realization of L*+ H pitch accent in Greek, 139–147.
28. *Shriberg, E. E.* (1994). Preliminaries to a theory of speech disfluencies.
29. *Toldova, S., Roytberg, A., Ladygina, A. A., Vasilyeva, M. D., Azerkovich, I. L., Kurzukov, M., et al.* (2014). RU-EVAL-2014 : Evaluating Anaphora and Coreference Resolution for Russian. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue 14, 1–14.