NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Svetlana Yu. Toldova, Elizaveta I. Ivtushok,
Kira M. Shulgina, Mira B. Bergelson,
Mariya V. Khudyakova

COREFERENCE ANNOTATION IN THE RUSSIAN CLINICAL PEAR STORIES CORPUS: ANNOTATION FEATURES AND PRELIMINARY RESULTS

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 50/LNG/2016

Svetlana Yu. Toldova[1], *Elizaveta I. Ivtushok*[2], *Kira M. Shulgina*[3], *Mira B. Bergelson*[4], *Mariya V. Khudyakova*[5]

COREFERENCE ANNOTATION IN THE RUSSIAN CLINICAL PEAR STORIES CORPUS: ANNOTATION FEATURES AND PRELIMINARY RESULTS[6]

This work is devoted to the distribution of different referential devices in spoken discourse produced by healthy speakers and people with aphasia and its comparison to written discourse. We discuss some special annotation issues for the corpus of Pear film retellings (Russian CliPS) by people with aphasia (PWA), right hemisphere damage (RHD), and healthy speakers (HP for healthy people) of Russian. The study summarizes the comprehensive annotation schema developed for this task and the preliminary research of the referential choice features based on the corpus. Comparing retellings and written texts, we found a significant difference in the use of basic coreferential expressions between the two. Firstly, there is a significant difference in the distribution of basic NP types. Speakers use reduced devices such as zero anaphora or bare nouns in retellings more frequently than in written texts. There are also differences in the distribution of more granulated features such as the word order within an NP, the use of anaphoric and reduced expressions (demonstratives or zero NPs) for the first mention of an entity, and the inclusion of epistemic markers into NPs. We also found that the retellings produced by PWA and HP do not differ much in terms of the distribution of basic NP types. However, a detailed analysis within different NP types and taking into consideration various disfluencies reveals some prominent differences between the two populations. These include a difference in zero subject distribution, the frequency of non-referential NP links, the frequency of co-reference errors. While adapting the initial coreference annotation scheme we concluded that besides 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), other features need to be considered.

Key words: coreference annotation, retellings corpus, discourse, brain damage, aphasia.

JEL Classification: Z

[1] National Research University Higher School of Economics, Moscow, Russia. toldova@yandex.ru

[2] National Research University Higher School of Economics, Moscow, Russia. e.ivtushok@gmail.com

[3] National Research University Higher School of Economics, Moscow, Russia. track_5@mail.ru

[4] National Research University Higher School of Economics, Moscow, Russia. mirabergelson@gmail.com

[5] National Research University Higher School of Economics, Moscow, Russia. mariya.kh@gmail.com

[6] 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 №16-05-024) and by the Russian Academic Excellence Project «5-100».

Introduction

A speaker can use different natural language expressions such as full noun phrases (*a boy, the boy with the bicycle*), demonstratives (*this, those*), or pronouns (*it, he*) to refer to an entity in discourse. Two expressions referring to the same object are said to be coreferent. Establishing coreferential relations in discourse is a complex process depending on various cognitive, discourse and grammatical factors. This phenomenon is the topic of multidisciplinary research, unifying scientists in various fields of linguistics, such as syntax, computational linguistics, and psycholinguistics [Gordon, Hendrick 1998].

One of the methods of investigating the coreferential phenomenon is the analysis of corpus data with deep linguistic annotation including coreferential chain annotation.

The majority of such corpora are based on written texts. However, nowadays there is greater investigation of spoken data as well. Spoken and written texts have different distribution of various linguistic features, they are produced and processed differently [Biber, 1988, Biber et al., 1999; Kibrik, 2009, Cuenca, 2015]. The essential properties of spoken discourse are that it has a high degree of interactivity and is produced in real time. These features can greatly influence the distribution of coherence devices and the rate of disfluent structures, such as hesitation, self-corrections, difficulties in the naming process and markers of word-finding [Bergelson et al., 2015; Podlesskaya and Kibrik, 2007; Shriberg, 1994]. These features can also influence the referential choice (the choice of a particular type of a noun phrase for referent maintenance) as well.

We discuss the annotation of coreferential relations in spoken discourse. The set of parameters used in our annotation scheme is described in [Toldova et al., 2016]. The present work is devoted to a more detailed analysis of these parameters. We try to find whether there are some specific features of referential choice in spoken discourse compared to coreference in written texts.

Our analysis is based on the data of a subcorpus of the Russian Clinical Pear Stories corpus (Russian CliPS) [Bergelson et al., 2015] which contains retellings of the Pear film (Chafe 1980). The film was created in order to get comparable speech samples with a clear story line and to study the flow of discourse. The Russian CliPS is a multimedia corpus of narratives produced by people with aphasia (PWA) and right hemisphere damage, and neurologically healthy speakers (HP for healthy people) of Russian [Khudyakova et al., 2016]. The annotation scheme should reflect the specificity of referential choice of speakers with different types of aphasia.

The focus of this study is on the parameters that should be taken into account in coreferential chains annotation of text retellings, including retellings by brain-damaged individuals. We analyse the specific features of referring expressions in spoken discourse, including possible disfluencies and errors related to the referential choice. The distribution of these features in texts of neurologically healthy people as compared to clinical discourse is discussed, as well as its comparison to the written text coreference analysis.

Related work

Coreference corpora: annotation principles

Coreference annotation has a relatively long history [cf. MUC-6 corpus, 1996, Bagga, Baldwin, 1998]. However, researchers still have interest in aspects concerning different types of texts. This interest has led to the creation of various corpora of coreference annotation. Some of them were created to evaluate automatic anaphora and coreference resolution systems (see for example the manual for coreference annotation [Chinchor, Robinson, 1997; Hirschman et al., 1997]), such corpora are also used in theoretical research focused on different features of coreferential choice.

Depending on the purpose of the corpus, the annotation principles may vary. For example, the corpora of MUC-6 conference[7] are focused on annotating only specific noun phrases referring to entities from the real world. In others, like ARRAU [Poesio, Artstein 2008], the annotation of generic noun phrases is also presented. The majority of coreference corpora of written texts have clear-cut and well described annotation schemes.

There are some coreference corpora containing spoken texts (cf. a subpart of ARRAU corpus). However, there are very few discussions concerning the annotation scheme for these texts.

For Russian, the first open coreference corpus of Russian language (RuCor[8]) was created in 2014. Its annotation was based on the MUC-6 annotated scheme. The RuCor scheme is the starting point in our annotation scheme. According to RuCor, the expressions that have to be annotated are the maximal NPs without left modifiers, separated by the comma from the head noun. The two NPs are in a coreferential relation only if the entities they refer to are identical. For an NP, its basic morphosyntactic type is annotated, such as, whether it is a certain type of pronoun or a full noun with a demonstrative (the detailed scheme is discussed in the following chapters).

[7] http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html
[8] http://rucoref.maimbava.net

Pear stories

The Pear film was produced in 1975 and was originally designed to investigate relations between language and human experience. The film lasts approximately 6 minutes and introduces viewers to a short silent story of a gardener who harvests pears, a boy who steals them, and other characters. For the last four decades, data from retellings of the Pear film have been used in discourse studies in different languages, including English, Finnish, Japanese and Russian [e.g. Fedorova 2015].

Spoken discourse annotation

As mentioned above, spoken discourse has some special features, such as unfinished utterances, various disfluencies, special interaction markers. As the annotation procedure for this register is more complicated than the written one, it needs further specification.

On the one hand, there are a lot of written corpora with deep linguistic annotation including the annotation of discourse features. Such annotation procedures are usually well documented; they are based on special instructions and standards [Gordon, Hendrick 1998]. There is a tendency to adapt the schemes used for the annotation of written texts to the spoken register. However, the features of spoken discourse mean the existing instructions cannot be used as they are [Kríž et al. 2015]. One of the possibilities to override this problem is to have two separate layers: one for the transcribed discourse as is and the other for the "normalized" or "reconstructed" text [ibid.]. This would allow the study of various disfluencies separately. This strategy has resulted in the majority of spoken corpora having coreference annotation only for the "normalized" level (e.g. ARRAU, RuCor).

On the other hand, there are numerous works on the annotation of various specific spoken discourse features. However, their focus is primarily the issue of discourse segmentation and different spoken discourse phenomena, such as hesitation pauses, self-corrections, discourse markers, markers of word-finding difficulties, and repetitions [Bergelson et al., 2015; Podlesskaya, Kibrik, 2007; Shriberg, 1994].

In our opinion, disfluencies can affect various discourse mechanisms such as information flow manipulation, reference tracking. They can affect the choice of referential device or the assessment of referent's discourse status (extra referent mentioning attracts more attention to it and thus influences the referent prominence assessment).

All in all, we consider that deviations from "normalized" noun phrases and some types of disfluencies related to the naming procedure should be taken into consideration in the coreference annotation scheme.

Aphasic speech

Coreference analysis is even more complicated taking into account that the corpus contains the oral retellings by PWA, a partial (or full) loss of language abilities caused by brain damage. The majority of works concerning referential choice in this type of discourse discuss the frequency distribution of the basic NP types (full NP/anaphoric pronoun/zero pronoun) compared to the texts of HP [Peng, 1992, Romanova, 2010].

Furthermore, in our research we compare the coreference phenomenon under these two conditions in detail. We single out specific features of referential choice in retellings, along with different types of deviations.

Coreference relations in Pear stories retellings: the parameters and their distribution

Data and data pre-processing

This study is based on the subcorpus of Russian CliPS, a multimedia corpus of Pear film retellings by PWA, right hemisphere damage (RHD), and healthy speakers of Russian [Khudyakova et al., 2016]. All speakers were asked to watch the film and then retell it in detail to a person who had not seen it before.

The annotation of the corpus was performed in ELAN[9] [Wittenburg et al., 2006]. The annotation scheme 1.0 includes the following tiers: quasi-phonetic, lexical, lemma, POS, grammatical properties, errors, laughter, segmentation into clauses and utterances. The quasi-phonetic tier (Transcript) is aligned with the media files, and contains an orthographic transcript of the speech recorded. Texts were also divided into elementary discourse units (EDU) to mark the borders of clauses in the spoken discourse.

Our subcorpus contains nine retellings produced by PWA and twelve narratives from HP. For the pilot study, we concentrated on the texts from people with two types of aphasia: acoustic-mnestic and efferent motor aphasia (for details concerning different types of aphasia see, for example [Akhutina, 2015; Luria, Hutton, 1977]). Although these groups of PWA have deficits on the micro-linguistic level of discourse, the narrative structure of their speech and coreference relations can be established [Marini, 2012].

We consider three text forms. The first is the full transcript version (see Fig.1, the examples of pauses and filled pauses are enclosed in the grey boxes). The second is the lexical transcript (with no annotation for pauses, laughter and so on). Lexical transcripts were run

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

through an automatic lemmatizer and morphological analyser. The third was without fillers, interaction markers, some of the speaker's remarks (e.g. "I think that it is nice…" and fillers such as *vot* 'so', *eto* 'well', *nu* 'well' etc., cf. Fig. 1, where filler is in the black circle).

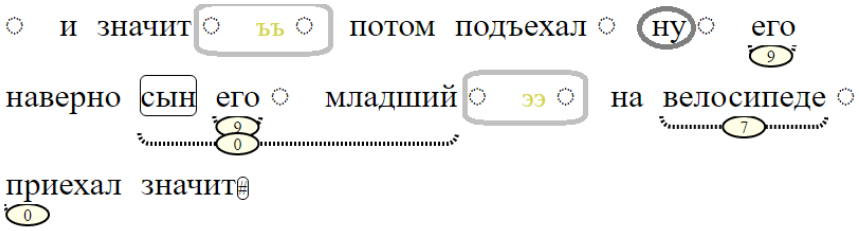

Fig. 1. The coreferential links annotation tool.

Table 1 shows the general condition of the two parts of subcorpus under consideration.

	PWA	HP
Number of texts	9	12
Min length in tokens	233	106
Max length in tokens	419	391
Range	186	285
Median	302	299
Average text length	326	277
Total	2934	3324

Table 1. The general statistics of the PWA and HP subcorpora.

Figure 2 illustrates the general text statistics for the overall number of tokens in the two groups of texts.

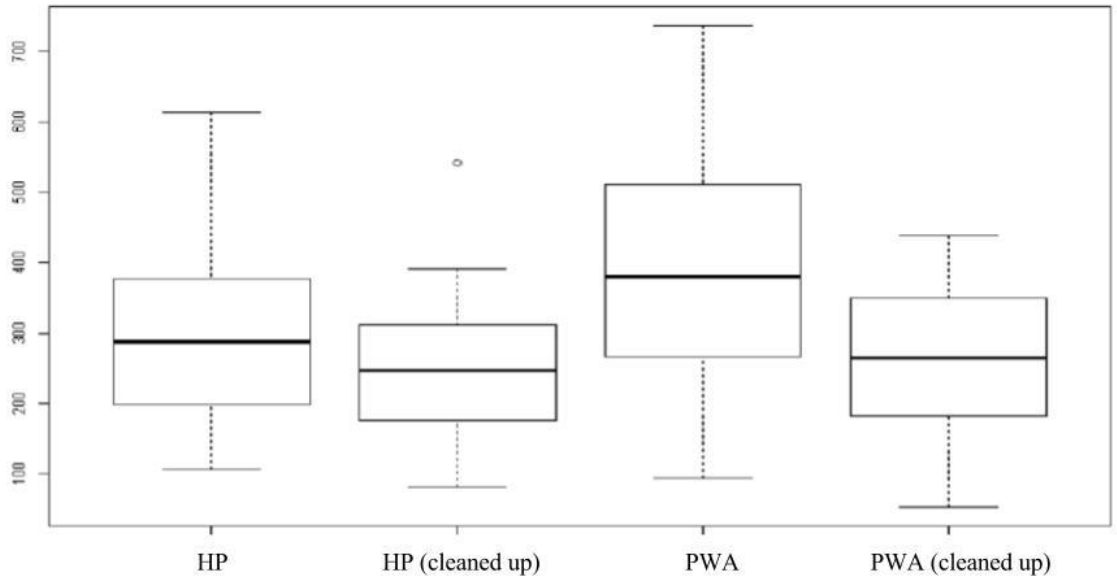

Fig. 2. The general statistics (number of tokens) box plot for PWA and HP texts before and after they has been cleaned up.

The figure shows that there is no significant difference in the median token number per text between two groups of people in texts without fillers (cf. (2) for HP and (4) for PWA respectively). However, the PWA manifests the higher variability. Moreover, the average token number including fillers is higher in PWA texts (cf. (1) for HP vs. (3) for PWA).

Coreference annotation procedure and tools

Considering the purposes of the study and the main features of the retellings in the corpora, the annotation consists of several important components. First, one needs to single out the referential expressions (NPs that refer to a real-life entity), or markables. For the film retellings, there is a set of entities, which appear in the film. These are people (such as the gardener, and the boy) and different items such as bicycles, pears, a hat [Fedorova 2015]. As this study is dealing with retellings of the same film, the number of entities is limited. The procedure of defining the markables is easier than in other texts types such as news. Accordingly, we single out NPs referring to entities from the predefined set (e.g. the horizontal braces below the NPs on Fig. 1). We take into consideration all NP types such as bare nouns, modified nouns (full noun phrases), anaphoric pronouns, syntactic and non-syntactic zeroes. Second, we also take into account non-referential mentions, the NPs that are repetitions of referential NPs in false-start constructions, renaming constructions etc. Third, we annotate all the links between markables referring to the same entity. (cf. the indices in bubbles on Fig. 1).

The coreference annotation is the process that needs a special instrument. The instrument should allow marking zero pronouns (we place the zeroes in front of the corresponding verb as in *Mal'chik padaet. Ø Ronyaet korzinu* 'lit. The boy falls down. *Ø* Drops the basket'). We use a special coreference annotation tool that was designed for RuCor (Russian Coreference Corpus). It is based on MySQL database engine and has a convenient web interface that allows parallel annotation by two or more people and online tracking of discrepancies between annotators. It also supports an extension of the feature set associated with a markable, and the linking of coreferent markables.

During the annotation process, we considered several parameters to be important when dealing with the spoken retelling. In order to specify such parameters, we conducted the experiment described in Toldova et al. [2016]. We took the ready-made scheme used for corpora annotation in NLP tasks. Our analysis of the data has shown that in order to capture various phenomena specific to the retellings some basic concepts, annotation parameters and their values needed refinement. The notion of markable needs adjustment. Some additional features might be highly informative, such as the entity type, the fine-grained taxonomy of NP structure, word order within an NP and zero anaphora types. Moreover, the annotation of various "non-coreferential" mentions and annotations of different types of errors or discourse disfluencies are also of great importance for our task.

We discuss the realization of these parameters in our data providing the comparison of retellings and the written texts as well as comparison of HP vs. PWA speech in the next section.

Analysis of parameters specific for coreference annotation in retellings: HP vs. PWA populations comparison

Markables

Many international standards for the annotation of written texts define a markable as a full noun phrase down to the nearest comma to the right [Krasavina, Chiarcos, 2007]. However, given the specific nature of spoken discourse, there can be possible limitations when dealing with the markable borders, making the standard principle of the right-comma-boundary inapplicable.

The first problem we are dealing with is various types of renaming, a construction when a false-name is followed by self-correction with a conjunction is quite frequent:

[1] *malchik [ili paren']*

'a boy [or a guy]'

The question is whether there are two markables in this case or one. Taking into consideration that both parts of such constructions are NPs with their own structure and lexical choice, these two NPs should be treated separately.

In the current study we take in consideration additional types of markables (a more detailed analysis is presented below). Two NPs with a conjunction between (e.g. clauses like 'X or Y') are considered to be two separate markables.

Secondly, we treat false starts as separate markables as well. By a false start we mean chunks of discourse where a participant begins to name an entity but stops or hesitates. Those are also annotated as separate NPs.

Another feature of spoken discourse is that an NP modifier (e.g. adjective or an apposition) can be postpositional:

[2] *... i proshli kak raz mimo khoz'aina [grushy etoy bolshoi]*

'lit. …and passed by the owner [of the pear tree this big]' (c.f. this big pear tree)'

In written discourse, such postpositive adjective phrases are interpreted as parcellation or a detached phrase. There is no punctuation marks in spoken discourse, so there are certain difficulties in drawing clear-cut boundaries between phrases. We treat the postpositional adjective in spoken discourse as a part of the markable that includes the preceding head noun the former modifies.

The third issue to be clarified is the referential properties of the markables. Though there are real-life entities appearing in the film such as trees in the garden or pears on these trees, the referential properties of corresponding NPs are not so clear in the retellings. A speaker can use an expression such as *sobiraet grushi* 'collects pears' in a generic sense. These entities are not "individualized" in the conceptualization of the corresponding scene by the speaker. He can use it as a generic expression. We annotate this type of expression as well.

Entity type

An analysis of an oral retelling of the same source such as a film or a set of pictures is easier and more productive, as the narrator and the researcher are limited in the number of known entities. As mentioned earlier, we annotate the entity type using a predefined set of entities. The set includes the main characters such as the boy and the gardener; the main items, which they manipulate with (baskets, bikes) and subsidiary items (a hat, a stone). The cases of erroneous naming (e.g. 'an apple' instead of 'a pear') are also tagged for the entity type.

One of the properties that can influence the referential choice and the type of NP is the discourse role of the mentioned entity, whether it is in the main focus of narration [Grozs, 1983]. Also, the frequency distribution of main vs. extra entities is a reflection of speaker's ability to

focus on the salient information in discourse. These parameters can differ for the HP and PWA populations. The distribution of entity types is illustrated in the table below.

	PWA			HP		
Main characters and entities						
Boy	276	27%	9/9	296	26%	12/12
Man	166	16%	9/9	212	18%	12/12
Three boys	136	13%	9/9	164	14%	12/12
Basket the boy took	50	5%	9/9	53	5%	12/12
Pears (generic)	43	4%	9/9	53	5%	12/12
Extra characters and minor entities						
One boy (of three)	37	4%	6/9	13	3%	5/12
Pears (taken by the boy)	35	4%	8/9	35	26%	10/12
Girl	25	2%	9/9	34	3%	12/12
Hat	33	3%	9/9	32	3%	11/12
Tree	17	2%	7/9	30	3%	10/12
Bike	31	3%	9/9	28	3%	10/12

Table 2. The distribution of basic entities in PWA and HP retellings.

The frequency mentions of the main characters and the key inanimate entities is approximately equivalent in both the HP and PWA population. As far as the minor items are concerned, they are more frequent for PWA population. According to these data a hypothesis can be formulated, that there is a tendency in PWA populations (namely people with acoustic-mnestic and efferent motor aphasia, whose narrations are analysed in the present research) to pay more attention to irrelevant details. A detailed discussion of entity tracking through discourse and their salience will be considered in future work.

NP morpho-syntactic structure

One of the main characteristics for a coreference phenomenon in a certain type of discourse (or a language) is the distribution of basic NP types. The distribution of these morpho-syntactic types

in coreferential choice is presented in the following table. The data on written text is based on the corpus of written news [Nedoluzchko et al., 2015]:

NP morphosyntactic type		Written texts		Pear Stories			
				PWA		HP	
anaphoric and reflexive pronouns	subject position	39	3.8%	138	14%	148	15%
	non-subject position	95	9.3%	112	11%	145	14%
Relative		42	4.1%	21	2%	19	2%
zero (pro)		13	1.3%	196	20%	199	20%
bare noun		164	16.0%	338	33%	338	33.1%
NP with a demonstrative		20	1.9%	38	4%	49	5%
Other NPs		652	63.60%	131	13%	163	18%
TOTAL		1025		974		1061	

Table 3. The distribution of basic morpho-syntactic types in written texts and PWA and HP retellings.

The basic referential expressions of the considered morpho-syntactic types do not differ much among retellings produced by HP and PWA. A chi-squared test confirms the absence of any statistically significant difference between these two groups ($\chi 2$= 0.83, df = 6, p = .9911). However, there are some differences, for instance, HP tend to use anaphoric and reflexive pronouns in non-subject position more often.

As for the written text, distribution patterns are considerably different in comparison with spoken discourse ($\chi 2$= 57.36, df = 6, p < .001).

To sum up, there is a significant difference in coreference devices in written texts and spoken retellings. As we can see, the more reduced devices (zeroes, bare nouns) are more likely in spoken discourse. In HP vs. PWA discourse, the general NP type statistics do not reveal any difference. However, a more precise analysis of referential expressions structure can reveal possible differences in the reference tracking models for these two populations.

NP morphosyntactic structure

The majority of annotation schemes for coreference reflect only basic types of noun phrases. However, more granulated NP morpho-syntactic features can matter.

In addition to NP basic types taken from the RuCor annotation scheme, we used a more detailed classification of NPs with modifiers, including:

- special types of modifiers such as indefinite pronouns, numerals, quantifiers and alternators (e.g. *drugoj* 'another', *takoj*, *pokhozhij* 'similar');
- the modifier morphological type (e.g. Genitive NPs, Comitative construction);
- the word order.

We analysed this parameter in interaction with the parameter of first / non-first entity mention and we found some specific spoken discourse features in using anaphoric devices such as zeroes or demonstratives in an introductory NP:

[3] *... potomu chto etot chelovek, sobiraiushchii grushi, on naverniaka vsio-taki zvuk slyshit khorosho*

'because this man, who collects pears, he likely still hears sound well'

The use of the demonstrative *etot* for the first entity mention, while possible in speech (9 times in HP texts vs. 6 times in PWA texts) would be considered an error in written discourse.

As the use of various anaphoric and zero-type links in spoken discourse is very common, the speakers may, although they do not very often, use them as antecedents, i.e. to introduce an entity. Table 4 shows the number of introductory mentions via zeroes, anaphoric pronouns, and NPs including demonstratives.

First-mention type	PWA	HP
Zero	1	7
anaphoric pronouns	1	9
Dem+N	7	11

Table 4. The distribution of non-standard first mention (introductory) expressions in PWA and HP.

Table 4 illustrates the fact that though a violation of the rules (of the standard usage of anaphoric devices) in introductory NPs is rare in our corpus, this type of deviation from the norm does occur in spoken discourse both in HP and in PWA speech. It is more frequent in HP speech. However, the frequencies are too low to check whether the difference is statistically significant.

One more interesting feature of spoken discourse is the inverse order of elements in a noun phrase:

[4] *uzhe gotovye v korzine sobrannye grushi*

'already ready in a/the basket collected pears'

Zero anaphora

Another important parameter in coreference annotation in both spoken and, less commonly, written discourse is the presence of syntactic zeroes: the absence of an overt noun phrase in a clause (e.g. 'He took the basket and Ø drove away'). Russian is a pro-drop language, meaning that a finite clause may have no overt subject. There may be no overt anaphoric pronoun in some other positions. Even given the fact that so-called zero pronouns can occur in written discourse, the nature of their usage in speech is different. When using zero-type units in written text, one has to make sure the meaning of it is recoverable from the context. A more common strategy is using overt NPs and leaving zero pronouns for syntactically determined positions. In Nedoluzhko et al. [2015] it is reported that zero pronouns in Russian newswire texts appear only in syntactically motivated positions. However, this 'syntactical motivation' is not that crucial for spoken discourse: full NPs are mostly used when introducing the entity or removing it from the narration, while in between it is preferable to use pronouns and zeroes [Fox, 1987, Kibrik, 1997].

Considering this, the 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 we restore zero subjects for all the verb forms with no overt subjects.

In addition, there are also syntactic zeroes in Russian. The overt subject is impossible in an infinitival clause, PRO (a pronominal determiner phrase that denotes an empty category, is used in non-finite clauses, and has strictly syntactic functions) is postulated in this case (e.g. *I oni pomogli mal'chiku Ø sobirat' grushy s zemli.* – 'And they helped the boy Ø to gather the pears from the ground'). We take these cases in consideration in our annotation scheme. Another syntactically motivated case is pro (an omitted pronoun) in verb coordination construction (e.g. *…on vylozhil vse grushi i Ø polez opiat' na grushu* – '… he laid out all pears and climbed again the pear [tree]')

The use of syntactic and non-syntactic zeroes is a common strategy in coreferential choice in spoken discourse [Grenoble 2001]. The number of zeroes used in retellings may also be a crucial parameter when comparing discourse produced by PWA and HP, therefore it is considered to be one of the most important parameters of the annotation in the present study.

The distribution of all zero NPs in the subject position as compared to non-zero subjects is presented in the Table 5.

	PWA		HP	
pro	93	17%	198	33%

PRO	48	9%	36	8%
non-zero	320	74%	373	61%
Total	541		607	

Table 5. The distribution of zero-type anaphoric tools in PWA and HP.

The proportion of zero non-PRO subject NPs in HP texts is nearly twice as high as in PWA texts and we conclude that the referential choice in HP discourse conforms to the accessibility hierarchy-based principle [Ariel 1990] to a greater degree than PWA discourse. Healthy people freely chose the most reduced device for salient referents (that is zero pronoun in Subject position for spoken discourse in Russian) to maintain reference in discourse.

Non-coreferential links

Among other features of spoken discourse, we also distinguish several types of non-coreferential links, such as renaming, self-correction, false starts, alternative naming.

Those link types almost never appear in written discourse. This parameter, however, is crucial for spoken discourse because it reflects the naming (or reference choice) procedure performed by a speaker as such. This parameter is defined as a *link type* in our annotation scheme. The detailed analysis of these cases is given in [Bergelson et al. 2014]. They are also discussed in [Toldova et al. 2015]. Consider the following examples:

[5] *…eti samyie briuchki … shtanishki* – renaming

‘these trousers … pants’

[6] *navervo on= mysl’ byla, shto on ili real’no durak ili…* – false start

‘maybe he= there was a thought that he is either a fool or…’

[7] *a mimo tri khuligana, mozhet i nie khuligana* – auto-correction

‘beside three bullies, maybe not bullies’

[5] illustrates the renaming procedure where a speaker suggests a more precise common name for an entity. In [6] a speaker makes a false start and then restarts an utterance changing its structure. In [7] he/she repeats the same NP. We consider these NPs in our annotation scheme for they may influence the increase of a referent’s salience.

The general statistics for different link types is given in the Table 6.

	PWA	HP
renaming	23	21
false starts	47	25
repetition	44	11
% of total markables	12%	6%

Table 6. The distribution of the most frequent non-coreferential links in PWA and HP.

The frequency of non-coreferential links use in PWA tests is considerably more frequent (F-test, $p < .05$).

Error annotation

We introduced a special parameter for marking common errors when choosing a noun phrase. However, these errors require an additional, more thorough analysis. We attribute the errors to three basic classes: (a) morphological errors; (b) lexical choice; (c) referential choice. The errors are illustrated in [8], [9] and [10] respectively.

[8] *pokazyvaiut sadovnika … muzhchina kotoryi im … na etogo malchika kotoryi*

'show a gardener … a man who them … on this boy who' (number agreement in pronoun)

[9] *paket nu…*

'bag well… (instead of basket)'

[10] *on otblagodaril tremia grushami… eshcho chego… i vsio. on spuskaetsia vniz…*

'he (the boy) thanked with pears… what else… and that's it. he (man) came down…' (ambiguity)

There are morphological errors when an anaphoric pronoun chosen by the speaker disagrees in number or gender with its antecedent. In [8] the speaker has chosen a wrong lexeme for naming the referent: *them* instead of *him*. In [3] the speaker has chosen a wrong referential device, namely, he used the demonstrative for the first mention of a referent. Other cases of erroneous referential choice are cases, when speaker's choice is ambiguous, an NP choice leads to a referential conflict (it can refer to more than one entity) or it can be mistakenly interpreted as referring to another entity.

These error types pertain to the referential choice in spoken discourse and do not occur in written discourse.

Conclusion

Spoken discourse has specific coreference features compared to written discourse. As we have shown, these features are essential for spoken discourse coreference characterization. First, there is a significant difference in the distribution of basic NP types. The main difference is significantly more frequent use of reduced devices such as zero anaphora or bare nouns. Other differences are revealed only through a more granulated description of referential expressions. These are more frequent use of anaphoric elements for first mention, inverse order in NPs, the inclusion of epistemic modality expressions into NP, and non-coreferential links between NPs naming the same entity.

We found that the retellings produced by PWA and healthy speakers of Russian do not differ significantly in terms of the distribution of basic NP types. However, a detailed analysis within different NP types and taking into consideration various cases of disfluencies reveals some prominent differences within the two populations such as the difference in zero subject distribution, the frequency of non-referential NP links, the frequency of coreference mistakes.

Often when performing coreference chain annotations for spoken discourse the text is 'purified' from disfluencies and interaction markers. It means that two objects—the 'normalized' text and various disfluencies—are studied as separate systems. However, these disfluencies in the text have impact on the interpretation of other text elements and 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 needed to include more granulated features. We need to consider various stipulations from the norm and disfluencies, as far as referential choice is concerned.

References

1. Akhutina, T. (2015). Luria's classification of aphasias and its theoretical basis. *Aphasiology*, *30*(8), 878-897.
2. Ariel, M. (1991, November 30). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, *16*(5), 443-463.
3. Biber, D., Finegan E. (1988, January 1). Adverbial stance types in English. *Discourse processes*, *11*(1), 1-34.

4. Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. (1999). *Longman grammar of spoken and written English*. (1204 pp). Harlow: Pearson Education Limited.
5. Cuenca, M. J. (2015). La connexió textual: l'adversitat en el nivell textual. *Caplletra. Revista Internacional de Filologia, 7,* 93-116.
6. Chafe, W. L. (ed.). (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production.* Advances in Discourse Processes, 3. Norwood, N.J.: Ablex.
7. Chafe, W. L. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* University of Chicago Press.
8. Chinchor, N., Robinson, P. (1997). MUC-7 named entity task definition. *Proceedings of the 7th Conference on Message Understanding.* p. 29.
9. Fedorova, O. (2015). *Eksperimental'nyj analiz diskursa*. Jazyki Slav'anskoj Kul'tury.
10. Gordon, P. C., Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cognitive science, 22*(4), 389-424.
11. Grenoble, L. A. (2001). Conceptual reference points, pronouns and conversational structure in Russian. *Glossos, 1*(1).
12. Hirschman, L., Thompson, H. S. (1977) Overview of evaluation in speech and natural language processing. *Survey of state of the art of human language.* 409-415. Cambridge: Cambridge University Press.
13. Fox, Barbara. (1987). *Discourse Structure and Anaphora.* Cambridge: Cambridge University Press.
14. Grosz, B. J., Joshi, A. K., Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, 44-50. Association for Computational Linguistics.
15. Khudyakova, M., Bergelson, M., Akinina, Y., Iska, E. (2016, May 23). Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals. *Proceedings of LREC 2016 Workshop.* Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016). Linköping University Electronic Press.
16. Kibrik, A. A. (1997). Modelirovanie mnogofaktornogo processa: vybor referencialnogo sredstva v russkom diskurse. *Vestnik Moskovskogo universiteta: Filologija, 53*(9), 94–105.
17. Kibrik, A. A., Podlesskaja, V. I. (2009). *Rasskazy o snovidenijah: korpusnoe issledovanie ustnogo russkogo diskursa.* Moscow, Jazyki Slav'anskoj Kul'tury.
18. Kríž, V., Hladká, B., Urešová, Z. (2015) *Czech Legal Text Treebank.*

19. Krasavina, O., Chiarcos, C. (2007). PoCoS: Potsdam coreference scheme. *Proceedings of the Linguistic Annotation Workshop.* Association for Computational Linguistics, 156-163.
20. Luria, A. R., Hutton, J. T. (1977). A modern assessment of the basic forms of aphasia. *Brain and Language*, *4*(2), 129-151.
21. Marini, A. (2012). Characteristics of narrative discourse processing after damage to the right hemisphere. *Seminars in speech and language*, *33*(1), 68-78. Thieme Medical Publishers.
22. Nedoluzhko, A., Toldova. S., Novák. M. (2015). Coreference Chains in Czech, English and Russian: Preliminary Findings. *Computational Linguistics and Intellectual Technologies, 14*, 456-469.
23. Bagga, A., & Baldwin, B. (May 1998). Algorithms for scoring coreference chains. *The first international conference on language resources and evaluation workshop on linguistics coreference, 1*, 563-566.
24. Bergelson, M., Akinina, Yu., Dragoy, O., Iskra, E., Khudyakova, M. (2015). Zatrudneniya pri porozhdenii slov v diskurse i ih formalnye markery: norma i patologiya ili o nediskretnosti normy v yazyke i rechi. *Proceedings of the International Conference "Dialogue 2015",* 41-45.
25. Peng, V. M. (1992). The usage of reference items in aphasic and normal conversations. *Journal of neurolinguistics*, *7*(4), 295-307.
26. Podlesskaya, V., Kibrik, A. (2007). Samoispravlenia govorjaš'ego i drugije tipy reč'evyh sbojev kak objekt annotirovanija v korpusah ustnoj reč'i. *Naučno-tehničeskaja informacija*, *2*, 2-23.
27. Poesio, M., Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech.
28. Shriberg, L. D., Kwiatkowski, J. (1994). Developmental phonological disorders: a clinical profile. *Journal of Speech, Language, and Hearing Research*, *37*(5), 1100-1126.
29. Toldova, S. Y., Khudyakova, M. V., Bergelson, M. B. (2016). Coreference in Russian oral movie retellings (the experience of coreference relations annotation in "Russian CliPS" Corpus). *Proceedings of the International Conference "Dialogue 2016".*
30. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559)

Contact details and disclaimer:

Svetlana Yu. Toldova
National Research University Higher School of Economics (Moscow, Russia). School of linguistics, associate professor.
E-mail: stoldova@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.