AN ANAPHORA RESOLUTION SYSTEM FOR RUSSIAN BASED ON ETAP-4 LINGUISTIC PROCESSOR[1]

Inshakova E. S. (e.s.inshakova@gmail.com)

Laboratory of Computational Linguistics,
A. A. Kharkevich Institute for Information Transmission
Problems, Moscow, Russia

The paper presents a rule-based system of automated anaphora resolution for Russian. The system is based on the resources of ETAP-4 linguistic processor: the Russian combinatorial dictionary (RCD), the ETAP parser, and the ontology OntoEtap. In this paper, I describe the ordered algorithms for resolution of different pronouns and provide the results of their evaluation.

Keywords: anaphora resolution, ETAP-4 linguistic processor

СИСТЕМА РАЗРЕШЕНИЯ АНАФОРЫ ДЛЯ РУССКОГО ЯЗЫКА НА БАЗЕ ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ЭТАП-4

Иншакова Е. С. (e.s.inshakova@gmail.com)

Лаборатория компьютерной лингвистики,
Институт проблем передачи информации
им. А. А. Харкевича РАН, Москва, Россия

---

1.   Introduction

In this paper, I present a system of automated anaphora resolution, which is a module of the ETAP linguistic processor [Boguslavsky et al., 2008][2]. This system is basically a system of deterministic rules written in the FORET formal language [Cinman 1995], with access to the Russian combinatorial dictionary (RCD) and the OntoEtap ontology [Boguslavsky et al. 2013]. It is domain and genre unspecific, meant to provide an antecedent for every pronoun in texts of any type; thus, it makes an important part of SemEtap—a system for deep semantic analysis and question answering being currently developed in IITP.

The paper has the following structure. In **section 2**, I describe the anaphora resolution module of ETAP: in **subsection 2.1**, I give its general outline, in **subsections 2.2**, **2.3**, **2.4** and **2.5** I discuss in more detail the resolution of reflexives, relative pronouns, the *tot* pronoun and 3rd person pronouns respectively. The subject of **section 3** is evaluation of the system's performance. **Section 4** contains conclusions and directions for further work.

2.   Describing the system

2.1. A general overview

ETAP-4 is a linguistic processor whose main options are machine translation between Russian and English and semantic analysis of Russian texts; ETAP was also used to create SynTagRus[3], a Russian dependency treebank [Boguslavsky et al. 2014].

The stages of text processing in ETAP are as follows:

0.   Splitting a text into sentences.
1.   Morphological analysis.
2.   Syntactic analysis.
3.   Anaphora resolution.
4.   Different options of translation from Russian (into semantic structures / into English).

The anaphora resolution module comes into action after the input text is fully parsed (its input can be a separate sentence / group of sentences or a corpus of parsed sentences in the .tgt format). It augments the parse trees with two types of non-tree links: 1) 'quasi-syntactic' links for zero anaphora that connect the syntactic host and the controller of zero pronouns; 2) coreference links between pronouns and the heads of their linear closest non-pronominal antecedents. All the mentions that form a coreferential chain are merged into coreference groups. Each coreference group automatically gets a unique number, which serves as a coreferential index.

---

[2]   ETAP-4 is downloadable at http://proling.iitp.ru/etap4download.

[3]   http://ruscorpora.ru/search-syntax.html.

Our algorithm is an ordered set of rules (each rule uses the results of earlier ones' work—coreference links or absence thereof). The (groups of) rules are listed here in the order they are applied:

1. A rule for various types of zero anaphora[4].
2. Rules for reflexive pronoun *sebja* '-self', reflexive possessive *svoj* '-self-s' and reciprocal *drug druga* 'each other'.
3. Rules for relative pronouns *kotoryj* 'which', *kto* 'who', *chto* 'that' and *chej* 'whose'.
4. A rule for the switch-reference pronoun *tot*.
5. Rules for 3rd person anaphora.

Like in many other rule-based or hybrid anaphora resolution systems (e.g. [Mitkov 2002], [Haghighi and Klein 2010], [O'Connor and Heilman 2013], [Lee et al. 2011]), the process of resolving 3rd person pronouns in ETAP falls into three stages:

a) filtering out the nouns that cannot control 3rd person pronouns at all;
b) generating a set of potential antecedents for every pronoun;
c) passing them through the ordered sieves of restrictional rules to select the right candidate:
    i.   the ontological sieve;
    ii. the syntactic sieve that favors anaphoric pairs in certain syntactic configurations and filters out the other ones;
    iii. the discourse sieve that sorts out candidate antecedents on the basis of their relative discourse prominence.

2.2. Reflexive and reciprocal pronouns

The algorithm for reflexive (*sebja* 'oneself', *svoj* '-self's') and reciprocal (*drug druga* 'each other') pronouns includes the following stages:

1. Searching for a predicate Z—the head of the local finite or non-finite clause (including nominalizations) that contains the pronoun. That means finding a predicate (a verb, event noun, adjective or adverb) that governs the pronoun directly or via a sequential chain of words that are not finite verbs and don't have their own subjects.

2. Finding the subject of this clause (the 'short-distance' antecedent). It is the noun that depends on the predicate Z via PREDIC, or DAT-SUBJECT, or QUASI-AGENT / AGENT[5] (if Z is a noun) relation. For the reciprocal pronoun *drug druga*, the antecedent can be only short-distanced (within the same minimal clause).

---

[4]   I won't discuss here the process and examples of resolution of this type of anaphora for space reasons and because zero anaphora annotation standard in ETAP has not yet been developed, so the quality of its resolution cannot be evaluated.

[5]   See a comprehensive list of Russian SSRs in Mel'čuk 1974 or on the SynTagRus website, and of English SSRs in [Mel'čuk 1988].

Our rules handle reflexive cataphora: *Svoj*$_i$ *put' v zhurnalistiku on*$_i$ *nachal barabanschikom v Nju-Jorke*—lit. 'His$_i$ road to journalism, he$_i$ began as a drummer in New York'). They contain subrules for detecting oblique antecedents [Paducheva 1983], e.g. *Mne*$_i$ *ne hvataet vremeni dlja sebja*$_i$ 'I$_{DATi}$ am short of time for myself$_i$', *Kazhdomu avtoru*$_{DATi}$ *svoj*$_i$ *podhod kazhetsja bolee jestestvennym*—lit. 'To each author$_i$, their$_i$ own approach seems more commonsensical'.

3. If the predicate Z is a nominalization or an infinitive governed by a predicate with the feature OB-INF (and the pronoun is reflexive), the algorithm goes up the parse tree to the predicate that governs Z and selects its subject as the 'distant' antecedent (in Russian, the binding domain for reflexive pronouns is a finite clause: see Rappaport 1986). In most cases, the next rule selects the most recent antecedent, and only in the contexts like *zastavljat' sebja* + INF—the 'distant' one (...*ona predlozhila sozdat' jedinuju platformu*$_i$, *kotoraja*$_i$... *zastavit vlast' sebja*$_i$ *slushat'* '...she suggested creating a united platform$_i$ that$_i$... would make the government listen to it$_i$').

4. Ruling out idiomatic expressions where reflexive pronouns are used non-referentially: *vesti sebja* 'to behave', *ne po sebe* '[to feel] uneasy', *svoego roda* 'a sort of' and many others[6].

2.3. Relative pronouns

The rule for this type of anaphora first finds a noun (Z) that is the syntactic host of a relative clause, i.e. governs its main predicate (Q) via the RELAT relation. Then it proceeds to the predicate Q and goes down the tree to the relative pronoun (it finds pronouns that depend on Q immediately or via a preposition, a noun, a noun and a preposition). If the pronoun fills a valency slot of some predicate, the rule checks whether the noun Z fits its selectional restrictions.

(1) *protestantskaja etika*$_i$ (Z), *znachenie kotoroj*$_i$ *raskryl* (Q) *Maks Veber...*
    'protestant ethics$_i$ (Z), the meaning of which$_i$ was (Q) revealed by Max Weber...'

If the noun Z depends on another noun (but not via APPOS, COORD or EXPLIC), which can also depend on a noun, the rule the rule checks the selectional restrictions for all those nouns to decide which of them is the antecedent.

However, many cases of such syntactic ambiguity require world knowledge, for which the state-of the-art ETAP does not yet provide any ready resources.

2.4. The pronoun "tot"

The rule for this pronoun is mainly based on the paper [Kreydlin, Chekhov 1988]. It also resolves the pronoun *tot* in '...*da i tot*'-constructions, which were not analysed in that paper. The rule deals with 3 types of contexts:

---

6   See a full list of such expressions in [Inshakova 2016].

1. The antecedent of *tot* is an argument (but not a subject) in the clause that precedes the clause containing the pronoun *tot*. The predicate that governs its antecedent also has a topically more prominent argument (it is usually a subject):

(2) *Bauer$_j$ pytalsja privlech k sudu i avtora «Njurnbergskix zakonov» Globke$_i$, no tot$_i$ okazalsja jemu$_j$ ne po zubam...*
'Bauer$_j$ tried to bring Globke$_i$, the author of the 'Nuremberg laws', to trial, but he$_i$ turned out to be a hard nut to crack for him$_j$.'

2. The antecedent of *tot* is a nominal subject, and there is another nominal subject between the pronoun *tot* and its antecedent.

The rule checks number and gender agreement between the pronoun and its antecedent as well as selectional restrictions of the predicate that governs the pronoun. It prefers candidate antecedents from the class 'PhysicalObject' to other nouns.

2.5. 3rd person anaphora

2.5.1. Filtering out 'non-antecedents'

At the preliminary stage of 3rd person pronoun resolving, the system discards nouns that cannot be antecedents of such pronouns at all (by means of assigning them the NON-ANTEC feature). The list of 'non-antecedents' includes many groups of expressions; here I will give only four examples (others are listed in [Inshakova 2016]):

1. **Nominal predicates**:

(3) *On skazal, chto Efimova chestnyj, kvalifitsirovannyj rabotnik.*
'He said that Efimova was an honest, competent <u>employee</u>.'

(4) *Glavnyj <u>element</u>$_j$ konstruktsii vertoleta—rama$_j$.*
'The main <u>element</u>$_j$ of the helicopter structure is a frame$_i$.'

2. **Nouns** from the class **GradableParameter** or with the feature **CHARAC-TROD** 'genitive of characterization', depending by **ATTRIB**: *problema pervostepennoj <u>vazhnosti</u>* 'an issue of primary <u>importance</u>'; *kategorija eksponatov povyshennogo <u>riska</u>* 'the category of higher <u>risk</u> exhibit items'.

3. Nouns that depend by the **COMPL-APPOS** relation (directly or via a preposition): *dom <u>nomer</u> vosem'* 'house <u>number</u> eight'; *ves v sto <u>tonn</u>* 'a hundred-ton weight'; by **LOCUT**: *<u>drug</u> druga* '<u>each</u> other'; *kuram na <u>smex</u>* ≈'enough to make a <u>cat</u> laugh'.

4. **Nouns in idiomatic expressions**, e.g. *slava <u>bogu</u>* 'thanks <u>God</u>', *chto <u>tolku</u>...* 'what's the <u>use</u> of...', *ne mozhet byt' i <u>rechi</u> o...*'[It's] out of the <u>question</u>', *<u>delo</u> ne v etom* 'it is not the <u>point</u>'.

2.5.2. Creating the set of candidate antecedents

The next rule picks out possible antecedents (within the search scope of 3 sentences—the current one plus two preceding sentences). It sets the following restrictions:

1. The antecedent's **location and POS feature**: to the left of the pronoun (our system does not resolve 3rd person cataphora yet). It can be a noun, an adjective or a participle (*mitingujuschie* 'protesters').

2. **Gender, number and person features:** agreement or disagreement, e.g.— *Ja$_i$ zhit' hochu!—zakrichal on$_i$* '— I$_i$ want to live!—he$_i$ cried'; *redaktor Znamenskaja* 'editor-M Znamenskaja-F'; *gosudarstvo$_i$*-N *v litse prinadlezhaschix jemu$_i$*-M *monopolij* 'the state$_i$ in the person of monopolies that belong to it$_i$'. For unknown words with missing morphological features, the rule checks the gender and number of their predicates or adjectival modifiers (if there are any).

3. **Selectional restrictions** for the candidate antecedents (if the pronoun fills a valency of some word). Candidates that don't belong to any of the ontological concepts listed in the corresponding column of the predicate word's GP (if there are any) are filtered out.

4. **Syntactic restrictions**. For 3rd person pronouns, the best known ones are **binding principles B and C** (Chomsky 1981) or **co-dependency** for dependency trees [Paducheva 1977]. The rule forbids contexts like *Masha$_i$ znala bol'she nejo$_{*i}$* 'Masha$_i$ knew more than her$_{*i}$'; *Petja$_i$ vidit jego$_{*i}$ dom* 'Petja$_i$ sees his$_{*i}$ house'; *Kak najti Petju$_i$, on$_{*i}$ ne znaet*—lit.'How to find Petja$_i$, he$_{*i}$ does not know'. There is also a counterpart of the i-within-i restriction, (*problemy$_i$ ix$_{*i}$ povtornoj restavratsii* 'issues$_i$ of their$_{*i}$ repeated restoring').

5. The rule forbids anaphoric links to conjuncts, except for cases like *arxitektor i sotsial'nyj reformator Vjacheslav Glazychev* 'architect and social reformer Vjacheslav Glazychev'.

6. The rule also forbids anaphoric links from subjects or addressees of speech verbs to nouns within the direct speech governed by these verbs:—*Skoro pridet Petja$_i$,—skazal on$_{*i}$ jemu$_{*i}$* '— Petja$_i$ will come soon,—he$_{*i}$ told him$_{*i}$'.

2.5.3. Sorting out incorrect candidates

The architecture of ETAP processor does not allow the system to rank a set of candidate antecedents, which can consist of >15 words for each pronoun. Instead, the anaphoric rules are 'eliminating': they either discard the candidate antecedents that don't meet certain ontological or syntactic constraints, or take every two nouns/adjectives Q and Z from the set of possible antecedents, compare the relative prominence of Q and Z and delete the coreferential link with the less prominent candidate.

The **ontological sieve** deals with such selectional restrictions that are not specified in the government patterns in RCD, but can be extracted from the dictionary zone SEM-ONTO of the RCD lexemes, which contains their ontological correlates from the ontology OntoEtap.

Table 1. Ontological restrictions stipulated in the ontology-based rule

Ontological correlate of the word W that governs the pronoun	Ontological correlate the antecedent should have	Syntactic relation between W and the pronoun	Examples
1. IntentionalProcess, SocialRole, SocialRelation, BiologicalAttribute, IntentionalRelation, NormalBiologicalEvent, PathologicProcess, PhysiologicProcess, EmotionalBehavioralProcess, StateOfMind, Human, Proposition, Model, PropositionalAttitude, TraitAttribute, Intelligence	Agent	PREDIC / QUASI-AGENT / AGENT / DAT-SUBJECT	*On vidit* 'he/it sees'; *oni studenty* 'they are students'; *jego teorija* 'his theory'
2. Human&SocialRelation	Human	1-COMPL	*Jejo drug* 'her friend'
3. Creation, Manufacturing	Artifact	1-COMPL	*Izgotavlivat' ix* 'to produce them'
4. SocialInteraction	Human, Organization	2-COMPL + preposition *s* 'with'	*Videt'sja s nim* 'to meet him'
5. AnimalAnatomicalStructure	Animal	QUASI-AGENT	*Jego plecho* 'his shoulder'
6. ChemicalProcess, NaturalProcess, PhysicalAttribute, StateChange, SurfaceChange	PhysicalObject	PREDIC / QUASI-AGENT / AGENT	*On plavitsja* 'it melts'
7. Motion, ShapeChange, StateChange, SurfaceChange	PhysicalObject	1-COMPL	*Chistit' jejo* 'to clean it/her'
8. Artifact, GeopoliticalArea, Human, Animal	Human	ATTRIB	*Ix gorod / sobaka / mashina* 'their town / dog / car'
9. Event	Event	PREDIC	*On byl obrjadom* 'it was a ceremony'
10. Location, BodyPosition + certain locative prepositions	PhysicalObject	1-COMPL / ADVERB	*Ivan zhil cherez tri doma ot nejo* 'Ivan lived three houses away from her/it'

Candidates that don't belong to the needed ontological classes are filtered out, as the bolded noun in (5):

(5) *Nel'zja cheloveka*$_i$ *zastavit' idti na* **miting**$_{*i}$. *Jemu*$_i$ *mozhno rekomendovat', jego*$_i$ *mozhno prizvat'.*
'One cannot force people$_i$ to go to **rallies**$_{*i}$. They$_i$ can be recommended to, they$_i$ can be encouraged.'

The **lexical functional rule** is applied to pronouns that depend on verbs/nouns that are values of lexical functions, if some of the candidate antecedents belong to the set of arguments of the given LF, and some do not:

(6) *...vpolne jestestvennymi byli dva proekta, kotorye Xruschev predlozhil* **nomenklature**$_{*i}$ *... Vo-pervyx, snjat' s nejo otvetstvennost'$_i$ za terror, perelozhiv jejo$_i$ na Stalina...*

'...two projects Khrushchev offered to the **Nomenklatura**$_{*i}$ were quite natural... Firstly, to exonerate it from responsibility$_i$ for the terror and to shift it$_i$ on Stalin...'

The **syntactic sieve** up to date deals with the following types of constructions:

1. Coordinate chains where the antecedent is an nth conjunct noun and the pronoun directly or indirectly depends on the $n+1^{th}$ conjunct noun:

(7) *Vo vremja Olimpiady politsejskim$_{*i}$ dano ukazanie: vystupat' protiv bezdomnyx$_i$ i zaschischajuschix ix$_i$ aktivistov...*

'During the Olympics, policemen$_{*i}$ were instructed to force against the homeless$_i$ and the activists who defended them$_i$ ...'

2. Specifying constructions where the minimal clause that contains the pronoun depends on its antecedent (or its syntactic host) via EXPLIC or JUXTAPOSE relation:

(8) *Ljudi$_{*i}$ zdes' xodjat s <u>telezhkami</u>$_i$ iz supermarketov: v <u>nix</u>$_i$ udobno skladyvat' banki i butylki.*

'People$_{*i}$ here walk with shopping <u>carts</u>$_i$: <u>they</u>$_i$ are convenient to collect cans and bottles.'

3. Some types of syntactic parallelism. These are present in the following constructions:

1) In a coordinate chain, the antecedent depends on the nth conjunct via a syntactic relation R, and the pronoun depends on the $n+1^{th}$ conjunct via the same relation R:

(9) *...ljudej$_i$, <u>ix</u>$_i$ otnoshenija$_{*i}$ s Bogom, <u>ix</u>$_i$ put', prisutstvie Xrista v <u>nix</u>$_i$.*

'...people$_i$, <u>their</u>$_i$ relationship with God, <u>their</u>$_i$ [spiritual] path, presence of Christ in <u>them</u>$_i$.'

2) The antecedent depends on a word that belongs to a lexeme L via a syntactic relation R, and the pronoun depends on another word that is also a form of the lexeme L, via the same relation R:

(10) *...po isku$_{*i}$ ...on priznal <u>ispol'zovanie logotipa</u>$_i$ nezakonnym i nalozhil zapret na <u>jego</u>$_i$ <u>ispol'zovanie</u>.*

'...on the suit$_{*i}$ ... he declared the <u>use of the logo</u>$_i$ to be illegal and imposed a ban on <u>its use</u>$_i$.'

Discourse rules compare candidates by pairs. They deselect a candidate antecedent:

1) if it is an adjunct (i.e. depends via ADVERB, or ATTRIB, or 4/5-COMPL relation) located 1 or 2 sentences away from the pronoun or it is inside an adjunct clause, and the other candidate holds an argument position in the current sentence.

2) if it is a possessor of another candidate noun, which belongs to the ontological class PhysicalObject.

3) if it is not not the previous finite clause's 'forward-looking center' in terms of the Centering theory [Grosz et al., 1995]:

(11) *Kogda krestjanin $_i$ zhenil starshego* **syna**$_{*i}$, *on$_i$ schital sebja$_i$ «starikom»*
i otdeljalsja so «staruxoj» v otdel'noe pomeschenie.
'When a peasant $_i$ married off [his] eldest **son**$_{*i}$, he $_i$ considered himself$_i$ an "old man" and moved with his $_i$ "old-wife" to a separate room.'

If, after all those sieves are applied, there still remain several possible antecedents for a pronoun, the last rule selects the most recent candidate.

3. The testing corpus, evaluation and error analysis

Our system participated in the AnCor evaluation campaign (2019), where only 3rd person pronouns were considered. It was evaluated on the test corpus of texts collected from the OpenCorpora online corpus (opencorpora.org), untagged but split into sentences. Its performance on this corpus and for this class of pronouns turned out to higher in precision but poorer in recall (micro-averaged) than the results shown on ETAP team's own test set. In this paper, I will also present the results of evaluation on our own testing corpus[7].

This corpus is an anaphorically annotated subset of SynTagRus in the .tgt format, collected in 2017–2019. It comprises 43 texts (mainly newspaper articles and fiction, a small amount of news texts and interviews), and in total is 6,315 sentences long. It contains 3,621 pronoun—antecedent pair.

When annotating the testing corpus I kept to the 'soft' criterion of annotation, i.e. agreed with the tagger's choice of non-closest and/or pronominal antecedents, if the resulting coreference groups were correct. Anaphora to coordination chains, comitative constructions and split antecedents was annotated as several anaphoric links from a pronoun to each element of its disjoint antecedent.

The SynTagRus-based corpus comes in two versions: 1) with 'gold', manually corrected syntactic structures and morphological tags, and 2) with 'raw', uncorrected structures and morphology. In this version, unknown words (= absent in RCD) were often unidentified by ETAP parser and lack morphological and POS tags. In this way, the impact of syntactic and morphological correctness on the system's performance can be estimated (Table 2).

---

7   Texts from the RuCor corpus ([Toldova et al., 2014]; downloadable at http://rucoref.maimbava.net), parsed by ETAP, comprised the development corpus.

Table 2. The results of the system's evaluation

Corpus	Pronouns	Syntactic structures	Criteria	Precision	Recall	Pairwise F1
AnCor	3rd person	'Raw'	macro, soft	69.90	53.80	59.30
			micro, soft	78.70	52.40	62.90
			macro, strong	58.10	45.00	49.40
			micro, strong	58.70	39.10	46.90
SynTagRus	3rd person	'Raw'	macro, soft	68.00	63.78	65.82
			micro, soft	66.81	62.46	64.56
		Gold	macro, soft	72.51	68.72	70.56
			micro, soft	75.40	71.05	73.16
	Reflexive and reciprocal	'Raw'	macro, soft	88.31	84.14	86.17
			micro, soft	84.76	76.25	80.28
		Gold	macro, soft	91.78	89.75	90.75
			micro, soft	90.94	87.26	89.06
	All	'Raw'	macro, soft	76.96	71.23	73.98
			micro, soft	73.78	68.04	70.79
		Gold	macro, soft	82.33	78.76	80.50
			micro, soft	81.99	78.12	80.01

Because the ETAP-based system extensively employs syntactic information, it is quite predictable that its performance depends on the quality of parsing to some extent. However, this dependence is not as strong as might be expected. The difference in performance on SynTagRus-based corpora with 'gold' and 'raw' syntactic structures for 3rd person pronouns is 4.74 F1 points / 6.72% (macro), or 8.60 F1 points / 11.76% (micro). For reflexive and reciprocal pronouns the difference is 4.58 F1 points / 5.05% (macro), or 8.78 F1 points / 9.86% (micro). For all types of pronouns it is 6.52 F1 points / 8.10% (macro), or 9.22 F1 points / 11.52% (micro). The fact that correct parse trees turn out to have a lower impact on resolution of reflexives and reciprocals than on resolution of pronominals is also somehow against expectations (because the correctness of structure seems to be crucial for syntactic anaphora resolution).

Error analysis. The comparison of scores for pronominals vs. reflexives and reciprocals supports the conclusion from [Toldova et al. 2016] that systems that handle syntactic anaphora quite well tend to have more mistakes in discourse anaphora resolution. Table 3 shows the actual role of salience-related errors (in the AnCor test set, the ETAP-based system made 250 errors in total).

In fact, the top three causes of errors have something to do with discourse issues. Such factors as spelling and punctuation errors (present in the AnCor corpus but not in the SynTagRus-based corpora) and unknown words, mostly proper names, on the contrary, don't decrease the score as dramatically as discourse rules do, although ETAP was designed to parse error-free texts and fails to process such words and sentence structures.

Table 3. Distribution of errors

Error type	Percentage
1. Discourse salience rules discard the correct candidate	19.2
2. Parsing errors and incorrect morphological tags that result in impossible syntactic configurations for pronouns and their antecedents (e.g. i-within-i, incorrect binding, attaching pronouns to argument positions with such selectional restrictions that disallow the antecedent, etc.). Some of the discourse-related errors may also be caused by incorrect parsing	15.2
3. Wrong choice of the linear closest candidate (i.e. the rules' failure to choose the most prominent candidate)	12.8
4. Antecedent beyond the search scope of 2 sentences	10.0
5. Too restrictive ontological rule that can deselect the correct candidates because they 'don't satisfy' the ontological constraints of words that govern the pronouns	8.0
6. Too restrictive rule that creates anaphorical links	6.8
7. The antecedent is an unknown lexeme	4.8
8. Pronoun—antecedent number disagreement (e.g. *Zimbabwe—they*)	4.8
9. Imperfections in RCD entries (e.g. lack of selectional restrictions or certain lexical functions)	3.6
10. The parser selects a wrong homonym (e.g. *predpolagat'1* 'suppose' instead of *predpolagat'3* 'presuppose')	2.4
11. Misprints and punctuation errors	2.4
12. Incorrectly assigned or lacking NON-ANTEC feature (for non-referential expressions)	2.4
13. Incorrect choice of a non-closest candidate in the case of JUXTAPOSE construction	1.6
14. Malfunction of the LF-based rule	1.6
15. Other types of errors (a single error for each type)	4.0

4.   Conclusion and further work

The first item on our laboratory's agenda now is developing non-pronominal coreference resolution algorithms. Secondly, the anaphora resolution system has to access sources of world/encyclopaedic knowledge and to fully use the potential of the ontology OntoEtap (which is still rather limitedly used in anaphora resolution) and its inference engine. This will enable it to resolve more complicated cases of anaphora, like those in the Winograd Schema Challenge, at the stage of semantic analysis (see an account of the initial stage of work on this issue in [Boguslavsky et al. 2019]). Thirdly, the anaphorically annotated part of the SynTagRus corpus has to be made publically available.

References

1. *Apresyan Yu. D. et al.* (2007) Lexical Functions in Actual NLP-Applications. In: Wanner L. (ed.): Selected Lexical and Grammatical Issues in the Meaning–Text Theory. In honour of Igor Mel'čuk. Studies in Language Companion. Series 84. Amsterdam: Benjamins Academic Publishers. P. 199–230.
2. *Boguslavsky I. M.* (2014) SynTagRus—a deeply annotated corpus of Russian. In: Blumenthal P., Novakova I., Siepmann D. (eds): Nouvelles perspectives en sémantique lexicale et en organisation du discours. Frankfurt am Main. P. 367–379.
3. *Boguslavsky I. M. et al.* (2008) Parser of the ETAP system and its evaluation with the aid of a deeply annotated corpus of Russian texts [Sintaksicheskij analizator sistemy ETAP i jego otsenka s pomoschju gluboko razmechennogo korpusa russkix tekstov]. In: Proceedings of the international conference 'Corpus linguistics—2008' (St. Petersburg). P. 56–74.
4. *Boguslavsky I. M. et al.* (2013). Semantic representation for NL understanding. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2013), p. 132–144.
5. *Boguslavsky I. M. et al.* (2019) Knowledge-based approach to Winograd Schema Challenge. In: current volume.
6. *Chomsky N.* (1981) Lectures on Government and Binding. Foris.
7. *Grosz, B., Aravind J. and Weinstein, S.* (1995). Centering: a framework for modelling the local coherence of discourse. Computational Linguistics, 21 (2), 203–225.
8. *Haghighi A. and Klein D.* (2010) Coreference resolution in a modular, entity-centered model. In Proc. of HLT-NAACL. P. 385–393.
9. *Inshakova E. S.* (2016) Resolution of syntactic pronominal anaphora in the ETAP-3 system [Razreshenie sintaksicheskoj mestoimennoj anafory v sisteme ETAP-3]. In: Information Technology and Systems (ITaS'16). Proceedings of the 40th Interdisciplinary Conference and School (St. Petersburg, 2016).
10. *Kreydlin G. E., Chekhov A. S.* (1988) Interrelation of semantics, information structure and pragmatics in lexicographic description of anaphoric pronouns (the case of pronouns of the TOT group) [Sootnoshenie semantiki, aktual'nogo chlenenija i pragmatiki v leksikograficheskom opisanii anaforicheskix mestoimenij (na materiale mestoimenija gruppy TOT)] Institute of Russian Language of AS USSR. The experimental and applied linguistics task group [IRYa AN SSSR. Problemnaja gruppa po eksperimental'noj i prikladnoj lingvistike]. Preprints. Iss. 178. Moscow.
11. *Lee H. et al.* (2011) Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In Proceedings of the CoNLL-2011 Shared Task.
12. *Mel'čuk I.* (1974) Toward a theory of Meaning ⇔ Text linguistic models [Opyt teorii lingvisticheskix modelej "Smysl ⇔ Text"]. Moscow.
13. *Mel'čuk I.* (1988) Dependency syntax: Theory and practice. Albany, NY.
14. *Mitkov R.* (2002) Anaphora resolution. Longman.
15. *O'Connor B., Heilman M.* (2013) Arkref: A rule-based coreference resolution system. arXiv preprintarXiv:13101975.

16. *Paducheva E. V.* (1977) On the semantics of syntax. Materials toward the transformational grammar of Russian [O semantike sintaksisa. Materialy k transformatsionnoj grammatike russkogo jazyka]. Moscow.

17. *Paducheva E. V.* (1983) Reflexive pronoun with an oblique antecedent and the semantics of reflexivity [Vozvratnoe mestoimenie s kosvennym antetsedentom i semantika refleksivnosti]. Semiotics and Informatics [Semiotika i informatika]. Iss. 21. P. 3–33.

18. *Rappaport G. C.* (1986) On anaphor binding in Russian. In: Natural Language & Linguistic Theory, 4(1):97–120.

19. *Toldova S. Ju., Roytberg A., Nedoluzhko A., Kurzukov M., Ladygina A., Vasilyeva M., Azerkovich I., Grishina Y., Sim G., Ivanova A., Gorshkov D.* (2014) Evaluating Anaphora and Coreference Resolution for Russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue. 13(20). P. 681–695.

20. *Toldova S., Roytberg A., Ladygina A., Azerkovich I., Vasilyeva M. D.* (2016) Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language, in: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL 2016, San Diego, California, June 16, 2016. Stroudsburg, PA: Association for Computational Linguistics. P. 74–83.