Authomatic direct speech tagging in Russian prose: markup and parser

Irina Nikishina, Irina Sokolova and Daniil Tikhomirov

Higher School of Economics, Moscow, Russia
irina.nikishina@mail.ru, irina.sokolovalxxxix@gmail.com,
dan.tijomirov@gmail.com

Abstract. Identifying speech in literary texts is one of the crucial tasks not only for Digital Humanities, but also for Natural Language Processing, as it might be useful for various literary studies as well as for Coreference resolution and even Dialogue systems. While parsers and corpora for automatic speech tagging already exist for some European languages, there are no such instruments for Russian. In the current article we present a gold-standard corpus with manually annotated sentences containing direct speech, new approaches to annotating speech with TEI markup, and a promising automatic parser that could help annotators to speed up annotation process.

Keywords: Direct speech tagging · Russian prose analysis · TEI · Automatic annotation.

1 Introduction

This article describes the results of a pioneering project in automatic direct speech tagging in Russian literary prose. While automatic and semi-automatic speech tagging tools already exist for the English language, they are not well-suited to Russian texts. Tagging direct speech is a language-specific task, mainly because graphic conventions and punctuation can vary greatly in different languages, and Russian punctuation is significantly different from English. Thus, rule-based approaches that work well for English texts are not directly applicable to Russian literature.

In this article we introduce a parser for automatic direct speech tagging in Russian literary texts that recognizes direct speech and accompanying author comments, identifies speech verbs and their properties, and attributes every instance of speech to a character. Our second goal is to create an annotated corpus with direct speech marked up using TEI standards, to be used for further research. The corpus contains a wide variety of types of direct, indirect, and free indirect speech, including complicated and ambiguous cases; it has been expert-annotated with high inter-annotator agreement; and it is openly available.

We use a sizeable (492 works) collection of classical Russian texts written in the 19th, 20th, and 21st century to study the many ways direct speech is marked in Russian prose; a shorter collection of 365 excerpts from 52 of these

texts, containing a wide variety of types of direct, indirect, free indirect speech as well as challenging or ambiguous cases, has been compiled for manual annotation and testing.

Using insights gained from manual annotation, we have created a pipeline that pre-processes plain text to replace irregular quotation marks with proper smart quotes; identifies speech in the text; distinguishes the character's utterance from the author's comment (if present) within the direct speech construction; finds the speech verb or its equivalent in the comment and recognizes its semantic and emotional characteristics; finally, it determines speech characteristics, and attributes each utterance to a character.

We then use the parser on the large collection for evaluation purposes. The parser accepts plain text as input and returns the same text with XML markup.

The authors of [Brooke 2015] [2] emphasize that "using NLP for literary analysis will require building literature-specific modules, even for tasks that are otherwise well-addressed in the field", as literary texts are "too different from the newswire and web texts that have been the subject of the vast majority of work in the field". As far as we are aware, no work has been done on tagging direct speech in Russian literary prose, as opposed to news texts. Thus, our work serves as the first step to automatic speech annotation in Russian prose and our parser is a contribution to quantitative analysis of Russian literature, and has a large number of applications in the field of literary analysis and digital humanities; it can be used to annotate large corpora quickly and in detail, and obtain valuable insights into Russian literature. The gold standard corpus can be used for further research into speech in novels, including research of literary history, gender studies, or stylometry; it is a valuable resource for machine learning as training data.

This paper is organized in the following way: first, in Section 2, we discuss previous works, relevant to our project: the output format, qualitative and quantitative research on speech in novels, speech detection in Anglophone literary texts, and character recognition in novels. Next, we describe our task (Section 3) and our pipeline (Section 4). In Section 5 we evaluate the obtained results and in Section 6 we discuss the outcome, the difficulties our parser faces at the moment and possible solutions, as well as ways to develop and broaden our work, and new features that can be introduced to increase the project's usefulness.

2 Related Work

In this section we discuss quantitative and qualitative studies on speech tagging and speech markers in text and provide an overview of digital research that is closely related to ours.

2.1 Qualitative studies

Understanding how speech is represented in literary texts is vital to identifying it automatically. Speech in literature is well-researched in academic papers (see,

for example, Lomov 2012 [14], Choi 2000 [17]), and free indirect discourse, as well as other non-classical forms of speech that appear in modern literature, have been the subject of special attention among researchers (see Pokrovskaya 2005 [15], Verdesh 2015 [12], Voronovskaya 2011 [13], Brooke et al. 2016 [3]). While direct speech can be found quite reliably using punctuation, free indirect speech is not marked in text in the same way, therefore further research is needed to understand its features and to detect it automatically, should we decide to add free indirect speech tagging to our parser.

In the late 20th-century and 21st-century Russian literature, not only free indirect speech, but also plain direct and indirect speech tagging becomes more problematic. Borders between the author's speech and characters' speech become increasingly blurred, and graphical and punctuation conventions are frequently defied: some authors use brackets instead of dashes and quotation marks, or do not mark a character's speech in the text in any way at all, which makes it visually inseparable from other characters' speech or the author's words (see Verdesh, Pokrovskaya, Arzyamova [12,15,11]).

Where punctuation alone is not a reliable marker for tagging, grammatical and lexical features can be used instead. Philological studies find that characters' speech is usually contrasted with the author's speech in terms of syntactic structure and lexical choices (see Shatalova 2011 [18], Voronovskaya 2011 [13]); it typically consists of shorter and simpler sentences, and includes more interjections and emotional epithets.

2.2 Quantitative & digital studies

The closest project to ours is GutenTag [2], a software tool that facilitates analyzing texts from the Project Gutenberg corpus. The tool provides detailed metainformation about the texts, allows the user to build subcorpora based on a large number of criteria, such as publication date, author gender or genre, and analyze the selected texts. Direct speech tagging and speaker identification are available, which is useful for literary analysis.

GutenTag uses the rule-based approach, relies on quotation marks to find direct speech, and outputs a set of tags in TEI format. Its authors suggest that new tags might be added to their parser in order to identify "some of the classic poetic elements such as rhyme scheme, meter, anaphora, alliteration, onomatopoeia, and the use of foreign languages", or tag "entire scenes with a physical location, a time of day, and a list of participants" and find elements of the plot structure. The parser we present in this paper has similar functions and can have similar applications.

An important paper by Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky introduces a new metric, dialogism, to better understand how spoken language is represented in novels [7]. The authors use abstract grammatical features, such as parts of speech or the structure of sentences, to measure to what extent a certain span of text is dialogic, i.e. close to natural spoken dialogue.

This paper is relevant to our own work in two ways: first, the model described in it relies on being able to extract speech from novels, which is what we attempt

to do with our parser. Second, the conclusions about dialogism in novels and ways of measuring it could be useful for correcting and fine-tuning automatic speech detection algorithms when the rule-based approach fails.

The authors have successfully used regular-expression-based algorithms to extract dialogue; however, such algorithms run into problems with texts which are messy due to automatic character recognition errors or changing publication standards. Typically, difficulties arise when an utterance is stretched over more than one paragraph or quotes are in the wrong direction; additionally, apostrophes often confuse the algorithm.

In Russian texts, the difficulties are different. Apostrophes are rarely seen, mostly when there is text in a foreign language, such as French or English, inside the Russian text. If an utterance is more than one paragraph long, it is not enclosed in quotation marks at all and begins with a dash instead, so there is no way of knowing where an utterance ends, even for a human reader.

The authors of [Muzny et al.] have developed a quote extraction system that they call QuoteAnnotator, which considers each quotation mark in context and decides if it is an opening or a closing quote, following a set of rules. The system also uses "fail safe measures like limiting the length of extracted quotations, allowing it to recover from garden paths caused by bad quotations". It is also able to apply different rules and then choose whichever one returns the best result.

Having annotated a large collection of English-language literature with QuoteAnnotator, the authors identify factors (parts of speech and grammatical forms) that reliably distinguish dialogue from narration. Their findings are quite intuitive, e.g. they conclude that modal verbs, "used to indicate desire, obligation, and ability", are "strongly associated" with speech, while third-person pronouns are used to talk about other people and thus are associated with narration.

The speech-extracting algorithm described in this paper is fitted to the English language and cannot be used directly on Russian texts. However, certain insights and heuristics are useful, as are the conclusions about the levels of dialogism in speech, which could help identify it in Russian texts.

2.3 Speaker identification

An important work on speaker recognition is *Identification of Speakers in Novels* by Hua He, Denilson Barbosa and Grzegorz Konrak [5], where the authors have chosen a supervised machine learning approach to attributing utterances to speakers. The approach is based on the assumption that all utterances within one paragraph can be attributed to a single speaker and each utterance is contained within a single paragraph (exceptions to this are rather easily identified by detecting quotation marks).

The authors' model exploits the fact that there normally is a pattern in conversations: when two characters talk, both are usually explicitly named in the beginning of the conversation, and then they take turns to speak. If the pattern is broken, there is a clue in the text that lets the reader know about it. Thus, consecutive utterances are usually attributed to different speakers, and

the speaker of the nth utterance is likely to be the same as the speaker of the $(n-2)$-th utterance. The authors have compared the results of several models, and the model that took into account the speaker alternation pattern showed superior performance.

A problem associated with speaker identification is speaker disambiguation, and it is addressed in a paper called *Mr Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized* [9]. Here character extraction is carried out by a rule-based pipeline, which builds a graph of character names connected by edges if the names denote the same character. This method allows to construct character lists automatically, which facilitates attributing utterances to speakers.

3 Task description

Our main goal, as briefly outlined above, can be summarized as follows: we aspire to create an automatic parser that identifies and annotates direct speech and identifies components thereof, such as the speakers' identities. Such an annotator may prove useful in any future literary research, including, but not limited to, social network extraction and comparative studies, as showed by studies conducted using English-based counterparts. The present paper aims to describe our progress towards developing such a parser and the results it yields in comparison to human annotators

For the purpose of speech annotation we have developed a series of our own task-specific TEI elements, described in full in Section 3.2. The outline of our method is provided in Section 3.3.

3.1 Markup

For the purpose of our task we have decided to build upon the existing markup language, The Text Encoding Initiative (TEI). TEI is a consortium that defines and maintains standards for the representation of texts in digital form. The TEI *Guidelines for Electronic Text Encoding and Interchange* document a markup language for representing structural and other features of texts. The Guidelines are expressed as an extensible XML schema. TEI markup language is by far the most widespread, and can be easily extended to fit particular research needs. On the other hand, the Guidelines are fairly strict and rigorously maintained, with all possible extensions conforming to a predetermined format.

Original TEI Guidelines do not offer much in terms of direct speech markup. Originally, only one TEI element, <said>, which "indicates passages thought or spoken aloud" [4], fits the task at hand without significant alternations. This element has the following attributes which are of interest for the task at hand:

@aloud whether the speech section represents thought or actual spoken words;
@direct whether speech is direct or indirect;
@who which named entity is considered the in-text speaker.

However, the task of creating an automatic direct speech annotator calls for a more nuanced identification of relevant sections of the written text, so the existing markup was extended to better reflect the inner structure of such passages.

As detailed above in Section 2.2, most studies that deal with direct speech in Russian literature and its punctuation (most notably, *D.E. Rosenthal's Russian Language. Orthography and Punctuation* handbook [16] which relates all typographical conventions concerning direct speech which are adhered in all contemporary editions) divide all instances of speech in two parts. The first one is direct speech itself, i.e. what is being said by a particular character related verbatim or represented in narrator's own words. The second part is a narration which introduces the character's utterance, providing context-specific details about the circumstances of this utterance, its emotional content and details on whom this utterance can be attributed to.

For the purposes of our project the first part is designated with a tag <**said**>, as it neatly corresponds to the similarly named element of original TEI markup, and the second part is marked by our own element <**author_comment**>. An umbrella tag <**speech**> is used to mark those sections of the text which include at least one instance of <said> with corresponding <author_comment> tags.

Another crucial addition to the existing TEI markup is a <**speech_verb**> element, which designates a verb related to a certain utterance which carries a lot of semantic information. This verb extraction is carried out using a pre-made vocabulary of lexical items that carry the meaning of saying or thinking something, with assigned **@emotion** and **@semantic** attributes. These attributes are subject to a later revision, and in their current state **@semantic** represents those elements of verb's lexical meaning that characterize the speech act itself (whether the character interrupts someone, speaks above or below the neutral sound volume, laughs or cries, etc.), while the **@emotion** attribute characterizes the emotional state of the character which produces the utterance, such as anger, sadness or happiness.

Finally, an important addition to the existing TEI element <said> comes in the form of attribute **@corresp**, which mirrors the already existing attribute **@who**, designating the character to whom the utterance is addressed.

All in all, our markup with all extensions can be summarized as follows:

<**speech**> an umbrella tag which includes both <said> and <author_comment> that can be ascribed to a single speech act by a single character;
<**said**> an utterance by a certain character, with the following attributes:
  @who the character who produced the utterance;
  @corresp the character to whom the utterance is addressed;
  @aloud whether the instance of speech relates an actual spoken utterance or a thought;
  @type designates direct or indirect speech.
<**author_comment**> a part of narration that accompanies the utterance;
<**speech_verb**> a verb included into an <author_comment> tag that carries semantic information about the utterance, with the following attributes:

@semantic characterizes the speech act as a whole;
@emotion characterizes the emotional state of the character.

Here we provide an example of a fully-parsed sentence:

<**speech**>
<**said who**="**Фома**" **corresp**="**None**" **aloud**="**true**" **type** = "**direct**">
– Позвольте вам заметить, что я жду</**said**><**author_comment**>
<**speech_verb, semantic**=**speech, emotion**=**neutral**>замечает</**speech_verb**>
Фома обидчивым голосом.</**author_comment**>
</**speech**>

As can be seen here, the <**speech**> tag captures a sentence, which occupies a whole paragraph, with correctly identified speaker and an unidentified addressee. The speech verb is characterized as semantically and emotionally neutral, as it does not explicate any specific characteristics of the speech act or the character's mental state.

3.2 Method

Our approach is rule-based, borrowing heavily from punctuation handbooks such as Rosenthal 2011 [16]. There are many ways speech can be punctuated in Russian prose: as a dialogue with each utterance occupying its own line, or with the whole conversation taking place inside a single paragraph with each utterance marked with quotation marks. However, all combinations of dashes, quotation marks and line breaks constituting an instance of direct speech confirm to a limited range of possible patterns – the fact that was thoroughly utilized during the development of rules for our annotator. It should be noted right away that the scope of our parser is, for the time being, limited exclusively to direct speech: it amounts to a sizeable portion of all non-narrative text in literature, and existence of clear-cut typographic conventions, on the one hand, makes a rule-based approach viable, and on the other, deviations from said conventions present a fairly challenging problem.

Our parser employs a five-step pipeline, with each stage dedicated to either bringing the text to a proper formatting or to extracting parts of text corresponding to certain elements in our markup. The output of each step becomes an input for the subsequent step, where, with the help of task-specific regular expressions, based on already present punctuation and speech boundaries provided on previous stages, the annotation is enriched and built upon. Method for each step of our pipeline is described in more detail in the next section.

4 Pipeline

Our pipeline comprises 5 steps: "Quotation marks replacement", "Speech tagging", "Author comment and Said tagging", "Speech verb tagging" and "Speaker identification". Each step except for the first one is named after the corresponding tag.

4.1 Quotation marks replacement

The very first, preliminary step in our pipeline is devoted to the preparation of the document for further parsing. During the development of our parser and upon deciding on the rule-based approach we have encountered the same problem as mentioned in Section 2.1 of Muzny et al. 2017 [7]: due to the nature of our corpus, namely, the fact that contributions to Moshkov's Library employ a wide range of digitization techniques (ranging from scans and optical recognition to simple re-typing of physical books), the punctuation in its items is far from being homogenous. As regular expressions used to extract direct speech are based mostly on punctuation, and the fact that whether the quotes are opening or closing determine the limits of speech, a special step included in our pipeline changes all punctuation marks to guillemets (« »). The use of guillemets instead of quotation marks has the following advantages:

1. They allow to easily discern the beginning and the end of a quote;
2. It allows to bypass a number of problems associated with errors made during Optical Character Recognition and changing typographical conventions;
3. Enables us to completely ignore tricky punctuation cases, such as apostrophes inside an utterance.

This step is also rule-based, and is comprised of a number of regular expressions detailing every combination of opening and closing quotation marks we have encountered in our corpus. The introduction of this step allowed to drastically reduce the number of specific rules written for later steps in the pipeline, and can be considered one of the key features of our annotator.

4.2 Speech (boundary) detection

The first step of the tagging stage of our pipeline receives text with all quotation marks changed to guillemets, and produces as its output a text with <**speech**> tags identifying sentences which contain at least one instance of direct speech, and, therefore, subject to further parsing. Rules used at this stage rely heavily on punctuation conventions, such as use of line breaks in dialogues or separation of author comment from direct speech with commas. As items from our corpus are plagued by severe typographic inconsistencies, each rule took into account the differing indents, dashes/hyphens and spacing which can be found in the same pattern.

In terms of its scope <**speech**> usually corresponds to a full sentence which includes at least one <**said**> and optional <**author_comment**> elements; however, since in some instances an utterance and the accompanying comment are embedded into a larger sentence with parts of narrative that have no relation to the utterance whatsoever, in certain <**said**>-<**author_comment**> patterns it seemed more appropriate to limit the scope of <**speech**> tag to the nearest comma after the last instance of <**said**> or <**author_comment**>. Below we provide one such case:

– Этот у меня будет светский молодой человек, – сказал папа, указывая на Володю,– а этот поэт, – прибавил он, в то время как я, целуя маленькую, сухую ручку княгини, с чрезвычайной ясностью воображал в этой руке розгу, под розгой – скамейку, и т. д.

The perfect algorithm would identify the scope of the tag <**speech**> as ending with the first comma after the direct speech, as the further narrative has little to do with speech. Therefore, the perfectly parsed sentence would look like this:

<**speech**> <**said who="папа" corrsep="я" aloud="true" type="direct"**>– Этот у меня будет светский молодой человек, –</**said**> <**author_comment**> <**speech_verb, semantic=speech, emotion=neutral**>сказал</**speech_verb**> папа, указывая на Володю,</**author_comment**> <**said**>– а этот поэт, –</**said**> <**author_comment**> <**speech_verb, semantic=speech, emotion=neutral**>прибавил</**speech_verb**> он</**author_comment**> </**speech**>, в то время как я, целуя маленькую, сухую ручку княгини, с чрезвычайной ясностью воображал в этой руке розгу, под розгой – скамейку, и т. д.

However, the line between what is relevant for the speech and what is not is a very fine one, and very rarely can be formalized in a rule-based approach like we employ here. This makes the "one speech – one sentence" approach much more feasible, even at the expense of semantic accuracy, as it is not always clear whether the narration can be better described as an author comment or as something else entirely. It should also be noted that the step of sentence splitting is absent from our pipeline: even though all instances of speech are punctuated as sentences, inclusion of straightforward sentence tokenization would make it impossible to use line breaks as reference points for determining speech boundaries and would yield poor results when faced with long instances of direct speech that includes many embedded sentences.

When an instance of direct speech is embedded into another instance of direct speech, the sentence including this embedded utterance is treated as an instance of <**speech**> in its own right. This is done partially because the subsequent steps of our pipeline rely heavily on the <**speech**> tags created during this step.

4.3 "Author comment" and "Said" detection

This stage has as its input a document with <**speech**> tags already in place. Because of that, rules used to extract these elements are less based on punctuation, and, therefore, less susceptible to typographic errors. As such, the rules are more dependent on the common patterns found inside sections that include speech.

This stage starts with identifying combinations of punctuation symbols that separate utterances from narration. If no such patterns could be found, the whole section is marked as <**said**>; if they could be found, the very first part is identified as an utterance or as a comment (based on punctuation), and the

next is marked as the opposite. Our parser exploits the fact that in most cases where a single sentence includes several utterances and comments they tend to alternate in a very predictable pattern.

4.4 Speech verb detection

This step receives as an input the marked-up text with <**said**> and <**author_comment**> boundaries already in place. As such, the search for speech verbs is limited to those that are directly included inside an <**author_comment**> tag. Each word inside this tag is lemmatized and checked against a premade dictionary of verbs. Each verb and its corresponding attributes (**@emotion** and **@semantic**) are included into the dictionary as lexical characteristics. Because of that each verb can have only one attribute, and our verb tagger is deaf to all additional semantic information which can be extracted from the context.

4.5 Speaker detection

It is quite important to mention that our Speaker detection is confined to <**author_comment**> boundaries. Our goal is to correctly identify the speaker if they are somehow mentioned inside the tag.

For detecting speakers we use an external framework Texterra [8] via API. The system performs deep linguistic analysis including syntax parsing, named-entity recognition that we implement for identifying speakers. We use this framework for our analysis in order to have similar framework output format (entities provided with borders), whereas separate frameworks may provide different output formats that would be rather difficult to combine. It is clear that in the future we are likely to implement and compare other frameworks for the current step, but for the present Texterra is considered to be our baseline solution. In the scope of <**author_comment**> we first of all identify the subject with the usage of Texterra's syntax parser. Then we find all possible Named entities in the same borders.

As a result, Texterra gives us a possible subject (if present) and a list of Named entities. Then we apply the following algorithm to identify the speaker:

1. if the subject is defined and the Named-entity list is empty, consider subject as speaker
2. if the subject is defined and the Named-entity list is not empty, compare each Named-entity to subject:
   2.1 if Named-entity string contains string of subject, consider the current Named-entity as speaker (this step permits to get full name for speaker)
   2.2 if the subject is not presented in the Named-entity list, consider subject as speaker
3. if subject is not defined and Named-entity list is not empty, then consider the first item of Named-entity list as speaker
4. if subject is not defined and Named-entity list is empty, then assign "None" value to speaker

5 Evaluation

The current section describes data used for training and testing. Moreover, we present two types of evaluation for the results obtained using Pipeline we described above.

5.1 Data

For building our gold corpus we have manually selected 365 excerpts from literary texts written in the XIX and XX century from Maksim Moshkov's Library (lib.ru).

While selecting excerpts to include in our corpus, our main criterion was a significant presence of conventionally punctuated direct, indirect and free-indirect speech. Moreover, we also tried to pay attention to the presence of various types of punctuation and some extra linguistic features: authorship and periodization. As a result, we used 52 works for building our gold-standard corpus that are listed in 6. Is commonly available on github[1].

We have not used the help of third-party experts, so we annotate the corpus ourselves, as native speakers getting a master's degree in linguistics. In other words, the assessors of the current corpus might be considered experts in the field. In order to evaluate confidence level of our corpus we measured inter-annotator agreement. The results are described in table 1 (we use Fleiss' kappa for the task).

Table 1. Annotator agreement (Fleiss' kappa)

	A1	***A2***
A2	Speech Kappa = 0.983 Said Kappa Kappa = 0.99 Author_comment Kappa = 0.957 Speech verb Kappa = 0.921 Mean = 0,96275	
A3	Speech Kappa = 0.836 Said Kappa = 0.776 Author_comment Kappa = 0.782 Speech verb Kappa = 0.796 Mean = 0.7975	Speech Kappa = 0.843 Said Kappa Kappa = 0.769 Author_comment Kappa = 0.81 Speech verb Kappa = 0.692 Mean = 0.7785
overall	Speech Kappa = 0.885 Said Kappa = 0.843 Author_comment Kappa = 0.85 Speech verb Kappa = 0.811 Mean = 0.84725	

[1] http://github.com/DanilSko/speech/gold_corpus

5.2 Evaluation Results

Evaluation of our study comprises three stages: first of all, we measure inter-annotator agreement in order to estimate corpus annotation consistency and its confidence level,

In order to conduct the next step, known as "manual evaluation" we parse annotated data into a table that contains the following columns:

- speech
- said
- who
- author comment
- speech verb
- semantics of speech verb
- emotion characteristics of speech verb

An example of the output can be seen in table 2

Table 2. Data output example for the first stage of evaluation

Кичкене с негою во взоре обняла старушку: Искендер-беку послышалось, что она даже вздохнула; я не слыхал, я не уверю в этом.
– Дядюшка Фетхали говорит, **что я еще слишком молода,** **– примолвила она почти грустно.**
– А что говорит твое сердечко, малютка моя? – возразила смеючись Аджа-Ханум. Кичкене резко схватила бубен, висевший на стене, и, колебля его звонки между расцвеченными хной пальчиками, вместо ответа пропела известную песню.

Said	***Who***	***Author comment***	***Speech verb***	***Semantics***	***Emotion***
– Дядюшка Фетхали говорит, что я еще слишком молода	она	примолвила она почти грустно.	примолвила	speech	neutral

It can be seen from the table that each syntactic structure that contains speech is embedded in its context. Moreover, each type of annotation is assigned to the appropriate column.

For the first part of evaluation we define a very specific system that is described in the table 3

The results for each instance of speech are provided below (see table 4). The scores might be interpreted as similar to precision. It should be noted that during this step we do not count the recall measure, as our goal is to detect drawbacks and limitations of our approach which are also described in table6 in 6.

From table 4 we can see, that according to our manual evaluation <**said**>, **@who** and <**author_comment**> show rather good results. At the same time,

Table 3. Evaluation I criteria

Score	***Description***
0	speech instance is not defined in the scope
1	speech instance and its borders are correctly defined in the scope
-1	speech instance is defined in the scope, but its borders are incorrect

Table 4. Precision for speech instances annotation

	Said	***Who***	***Author comment***	***Speech verb***	***Semantic***	***Emotion***	***Overall***
precision	0.9054	0.7608	0.8164	0.7837	0.7706	0.7662	**0.8021**

lower scores of **speech_verb** annotation (including parameters) in comparison with the aforementioned tags are quite confusing. However, this phenomenon might be explained by the fact that **speech_verb** obtains zero points if it is not identified in <**author_comment**> as well as <**author_comment**> is not identified, that can be considered a limitation of the system (or, to some extent, of our parser, as it relies on <**author_comment**> and not <**speech**>).

The analysis of the drawbacks is discussed in Section 6.

The second part of our evaluation (hereinafter Evaluation II) is done automatically, as we already possess annotated test set we consider to be gold standard. We chose Constrained Entity-Alignment F-Measure (CEAF, [6]) as our evaluation algorithm.

The authors of CEAF postulate that their algorithm is both simple and effective, as it begins with aligning reference and system entities, and then proceeds to count precision, recall and F-measure. They also prove that the methods proposed by [1] (known as B-cubed metric) and by [10] (MUC F-measure), are less representative then CEAF. In fact, B3 computes individual precision, recall and F-measure for each entity and then takes the weighted sum of these individual results as the final metric. MUC F-measure is computed by "first counting the number of common links between the reference and the system output". However, their common limitation (in comparison to CEAF) is the process of "intersecting" the reference and system entities, which allows an entity to be used more than once.

Despite the fact that these measures are mostly used for Coreference resolution task, this approach is perfectly suitable for our task as well. Thus, we provide the results of evaluation performed by CEAF in table 5. From the numbers above we conclude that speech verb (and its parameters) detection is quite successful, while correct annotation of "said" instance and "author comment" leaves much to be desired. The obtained results are also described in Section 6.

Table 5. Automatic evaluation using CEAF

	Speech	***Said***	***Author comment***	***Speech verb***	***Semantic***	***Emotion***	***Overall***
precision	0.6620	0.6494	0.6231	**0.6941**	**0.6941**	**0.6941**	0.6694
recall	0.6426	0.6098	0.5163	**0.8268**	**0.8268**	**0.8268**	0.7081
F1	0.6520	0.6289	0.5647	**0.7547**	**0.7547**	**0.7547**	0.6850

6 Discussion

As has been already mentioned, the aim of our study is to create a gold standard corpus with annotated speech, to develop new TEI structure for speech annotation and to create an automatic parser for speech detection. In sum, the obtained results mostly correspond to our objectives.

Even though the total volume of manually annotated excerpts is only 37,58 percent of the whole corpus of XIX and XX century literature and comprises 155632 tokens, this corpus might be used for further training. Moreover, it is the first commonly available dataset for speech detection in literary texts in Russian language. This corpus may be expanded by texts annotated automatically via our parser with further manual correction.

Considering the performance of the developed parser, we might say that it produces quite adequate annotation for those sentences which are not overburdened with complex grammatical constructions. As already mentioned above, we received a fairly good performance for detecting speech verbs and its parameters. However, the following improvements should be taken into consideration for our future work:

- add more rules for detecting speech in raw text
- do not remove punctuation during annotation process
- make rules more independent of line start
- make rules more resistant to syntax structure
- apply syntax parser to the speech verb detection pipeline step
- apply coreference resolution in order to identify character and not the speaker
- develop pipeline step for detecting the intended recipient of speech
- implement sentiment analysis algorithm for detecting author comment characteristics

Moreover, it should be also said that in addition to the current state of our parser we also plan to add a classifier (or, to be precise, a sequence labeler) that would label a sentence as sentence that contains speech or not, For this task we should create a list of useful features for labeling the data. Moreover, we would like to compare performance of machine learning (e.g. SVM or CRF) and neural networks for the task.

Another direction for further development might be detection of indirect and free-indirect speech. It could be also interesting to build sentence classifier that

detects speech type using specific features. Furthermore, given neatly annotated corpus any king of literary analysis may be performed. For instance, from corpus we may gather some information about most common speech verbs, its semantic, emotional characteristics, common speakers etc. In other words, the utility and value of "speech" corpus are fairly significant.

References

1. Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada, 1998.
2. Julian Brooke, Adam Hammond, and Graeme Hirst. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47, 2015.
3. Julian Brooke, Adam Hammond, and Graeme Hirst. Using models of lexical style to quantify free indirect discourse in modernist fiction. *Digital Scholarship in the Humanities*, 32(2):234–250, 2016.
4. TEI Consortium. TEI Guidelines "3.3.3 quotation" tei p5: Guidelines for electronic text encoding and interchange, 2018.
5. Hua He, Denilson Barbosa, and Grzegorz Kondrak. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1312–1320, 2013.
6. Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics, 2005.
7. Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32(suppl_2):ii31–ii52, 2017.
8. D Yu Turdakov, NA Astrakhantsev, Ya R Nedumov, AA Sysoev, IA Andrianov, VD Mayorov, Denis G Fedorenko, AV Korshunov, and Sergei D Kuznetsov. Texterra: A framework for text analysis. *Programming and Computer Software*, 40(5):288–295, 2014.
9. Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, 2015.
10. Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics, 1995.
11. Ольга Витальевна АРЗЯМОВА. АВТОРСКОЕ ОФОРМЛЕНИЕ ПРЯМОЙ РЕЧИ КАК СРЕДСТВО ВЫРАЖЕНИЯ ИДИОСТИЛЕМЫ В НОВЕЙШЕМ РУССКОМ ХУДОЖЕСТВЕННОМ ДИСКУРСЕ_. *Известия Воронежского государственного педагогического университета*, (3):147–152, 2016.
12. АА Вердеш. Прямая речь в художественных текстах неклассической парадигмы: новые явления. *Мир науки, культуры, образования*, (5):319–321, 2015.

13. Ирина Андреевна Вороновская. Порядок слов в авторской речи Б. Акунина. *Известия Саратовского университета. Новая серия. Серия Филология. Журналистика*, 11(4), 2011.
14. Анатолий Михайлович Ломов. Чужая речь в письменном тексте. *Вестник Балтийского федерального университета им. И. Канта. Серия: Филология, педагогика, психология*, (8), 2012.
15. Елена Александровна Покровская. Чужая речь и диалог в потоке сознания (на материале русской литературы XX в). *Политическая лингвистика*, (16), 2005.
16. Дитмар Эльяшевич Розенталь. *Русский язык. Орфография и пунктуация*. М.:Эксмо, 2011.
17. Чжи Ен Чой. Способы передачи чужой речи и тип художественного повествования (на материале рассказа АП Чехова «Скрипка Ротшильда»). In *ЯЗЫК, СОЗНАНИЕ, КОММУНИКАЦИЯ*, pages 89–98. 2000.
18. Ольга Васильевна Шаталова. Автор и персонаж: синтаксическая репрезентация в несобственно-прямой речи. *Вестник Костромского государственного университета*, 17(2), 2011.

Appendix A

Table 6. Errors and limitations

Speech instance type	*Error description*	*Example*
-	punctuation is deleted	<speech><said who="None" corresp="None" aloud="true" type = "direct"> – Прощайте, monsieur Irteneff</said> <author_comment> <speech_verb, semantic=speech, emotion=neutral>сказала</speech_verb> мне Ивина</author_comment></speech>
speech	borders are too wide (after "." in author comment)	<speech><said who="Пьер" corresp="None" aloud="true" type = "direct"> – Что же</said> <author_comment><speech_verb, semantic=speech, emotion=neutral>сказал</speech_verb> Пьер, всё так же улыбаясь. Становилось страшно.</author_comment></speech>
speech, said author comment	long sentences with more the 1 author comment are wrongly parsed after the first "said" instance	<speech><said who="Николай" corresp="None" aloud="true" type = "direct"> – Ну, этого ты никак не знаешь</said><author_comment> <speech_verb, semantic=speech, emotion=neutral>сказал</speech_verb> Николай; – но мне надо поговорить с ней. Что за прелесть, эта Соня</author_comment> <said who="Николай" corresp="None" aloud="true" type = "direct"> прибавил он улыбаясь. </said></speech>
speech, said, author comment	long sentences with different speakers: speakers and instances are wrongly identified	– Знаете ли что, Африкан Семеныч начала Дарья Михайловна вы недаром так озлоблены на женщин. Какая-нибудь, должно быть, вас.. Обидела, вы хотите сказать перебил ее Пигасов.
speech verb	speech verb not in list	усомнился
speech verb	wrong speech verb identified	– А! вот и Николенька Наташа подбежала к нему.
author comment	author comment is not defined	"Замолчи, я тебя убью!" – закричал в испуге Анлрей
author comment	author comment is inside "said"	Еще другой куплет, –говорила она, не замечая Николая.
said, *author comment,* *speech verb*	speech is not detected	Доктор посмотрел на меня и сказал торжественно, положив мне руку на сердце: – Она вам знакома!..
said, *author comment,* *speech*	speech is inside speech	Раз и зовет меня старая пани: «Харько, я тебя женить хочу!» – «Воля ваша, говорю, пани»...

Appendix B

1. **Ahsharumov, N. D.** «Kontsy v vodu» *(Ахшарумов, Н. Д. «Концы в воду»)*
2. **Annenskaya, A. N.** «Brat i sestra» *(Анненская, А. Н. «Брат и сестра»)*
3. **Astafyev, V. P.** «Veselyj soldat» *(Астафьев, В. П. «Веселый солдат»)*
4. **Babel, I.** «Konarmija» *(Бабель, И. «Конармия»)*
5. **Bestuzhev-Marlinsky, N. N.** «Mulla-Nur» *(Бестужев-Марлинский, Н. Н. «Мулла-Нур»)*
6. **Boborykin, P. D.** «U plity» *(Боборыкин, П. Д. «У плиты»)*
7. **Chekhov, A. P.** «Chelovek v futljare» *(Чехов, А. П. «Человек в футляре»)*
8. **Chekhov, A. P.** «Dom s mezoninom» *(Чехов, А. П. «Дом с мезонином»)*
9. **Chekhov, A. P.** «Palata» *(№6 Чехов, А. П. «Палата №6»)*
10. **Danilevsky, G. P.** «Beglye v Novorossiii» *(Данилевский, Г. П. «Беглые в Новороссии»)*

11. **Dostoevsky, F. M.** «Belye nochi» *(Достоевский, Ф. М. «Белые ночи»)*
12. **Dostoevsky, F. M.** «Igrok» *(Достоевский, Ф. М. «Игрок»)*
13. **Dostoevsky, F. M.** «Selo Stepanchikovo i ego obitateli» *(Достоевский, Ф. М. «Село Степанчиково и его обитатели»)*
14. **Dovlatov, C. D.** «Kompromiss C.» *(С.Д. Довлатов, «Компромисс»)*
15. **Durova, N. A.** «Ugol» *(Дурова, А. Н. «Уголь»)*
16. **Furman, P. R.** «Sardaamskij plotnik» *(Фурман, П. Р. «Сардаамский плотник»)*
17. **Gogol, N. V.** «Mertvye dushi» *(Гоголь, Н. В. «Мёртвые души»)*
18. **Gogol, N. V.** «Portret» *(Гоголь, Н. В. «Портрет»)*
19. **Gogol, N. V.** «Povest' o tom, kak possorilsja Ivan Ivanovich s Ivanom Nikiforovichem» *(Гоголь, Н. В. «Повесть о том, как поссорился Иван Иванович с Иваном Никифоровичем»)*
20. **Gogol, N. V.** «Taras Bul'ba» *(Гоголь, Н. В. «Тарас Бульба»)*
21. **Goncharov, I. A.** «Oblomov» *(Гончаров, И. А. «Обломов»)*
22. **Gorky, M.** «Foma Gordeev» *(Горький, М. «Фома Гордеев»)*
23. **Gorky, M.** «Staruha Izergil'» *(Горький, М. «Старуха Изергиль»)*
24. **Grigorovich, D. V.** «Koshka i myshka» *(Григорович, Д. В. «Кошка и мышка»)*
25. **Ilf, I. and Petrov, E.** «Zolotoj telenok» *(Ильф, И. и Петров, Е. «Золотой теленок»)*
26. **Kokorev, I. T.** «Sibirka» *(Кокорев, И. Т. «Сибирка»)*
27. **Kryukov, F. D.** «Gulebschiki» *(Крюков, Ф. Д. «Гулебщики»)*
28. **Kugushev, K. V.** «Kornet Otletaev» *(Кугушев, К. В. «Корнет Отлетаев»)*
29. **Kuprin, A. I.** «Olesja» *(Куприн, А. И. «Олеся»)*
30. **Lermontov, M. J.** «Geroj nashego vremeni» *(Лермонтов, М. Ю. «Герой нашего времени»)*
31. **Leskov, N.** «Levsha» *(Лесков, Н. С. «Левша»)*
32. **Leskov, N. S.** «Zahudalyj rod» *(Лесков, Н. С. «Захудалый род»)*
33. **Leskov., N.** «Obojdennye» *(Лесков, Н. С. «Обойденные»)*
34. **Mamin-Sibiryak, D. N** «Ohoniny brovi» *(Мамин-Сибиряк, Д. Н. «Охонины брови»)*
35. **Nabokov, V. V.** «Priglashenie na kazn'» *(Набоков, В. В. «Приглашение на казнь»)*
36. **Panaev, I. I.** «Belaja gorjachka» *(Панаев, И. И. «Белая горячка»)*
37. **Pushkin, A. S.** «Dubrovskij» *(Пушкин, А. С. «Дубровский»)*
38. **Pushkin, A. S.** «Kapitanskaja dochka» *(Пушкин, А. С. «Капитанская дочка»)*
39. **Pushkin, A. S.** «Stantsionnyj smotritel'» *(Пушкин, А. С. «Станционный смотритель»)*
40. **Saltykov-Schedrin, M. E.** «Gospoda Golovlevy» *(Салтыков-Щедрин, М. Е. «Господа Головлевы»)*
41. **Saltykov-Schedrin, M. E.** «Liberal» *(Салтыков-Щедрин, М. Е. «Либерал»)*
42. **Saltykov-Schedrin, M. E.** «Novyj Nartsiss» *(Салтыков-Щедрин, М. Е. «Новый Нарцисс»)*

43. **Severin, N.** «Poslednij iz Vorotyntsevyh» *(Северин, Н. «Последний из Воротынцевых»)*
44. **Sologub, F.** «Ulybka» *(Сологуб, Ф. «Улыбка»)*
45. **Tolstoj, L. N.** «Detstvo» *(Толстой, Л. Н. «Детство»)*
46. **Tolstoj, L. N.** «Dva Gusara» *(Толстой, Л. Н. «Два Гусара»)*
47. **Tolstoj, L. N.** «Junost'» *(Толстой, Л. Н. «Юность»)*
48. **Tolstoj, A. N.** «Kazatskij shtos» *(Толстой, А. Н. «Казацкий штос»)*
49. **Tolstoj, L. N.** «Vojna i mir» *(Толстой, Л. Н. «Война и мир»)*
50. **Turgenev, I. S.** «Rudin» *(Тургенев, И. С. «Рудин»)*
51. **Veresaev, V.** «Zapiski vracha» *(Вересаев, В. «Записки врача»)*