Automatic direct speech tagging in Russian prose: markup and parser

Nikishina I. A.[†]* (irina.nikishina@mail.ru)
Sokolova I. S.[†] (irina.sokolovalxxxix@gmail.com)
Tikhomirov D. O.[†] (dan.tijomirov@gmail.com)
Bonch-Osmolovskaya A.[†] (abonch@gmail.com)

[†]Higher School of Economics, Moscow, Russia
*Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

Identifying speech in literary texts is one of the crucial tasks not only for various literary studies, but also for Natural Language Processing tasks, such as for Co-reference resolution, Information Extraction and also when building Dialogue systems. While parsers and corpora for automatic speech tagging already exist for some European languages, there are no such instruments for Russian. Thus, in this paper we present a gold-standard corpus with manually annotated sentences containing direct speech. We also describe new approaches to annotating speech with TEI markup, and a promising parser that could help speed up annotation process.

Key words: direct speech tagging, Russian prose analysis, TEI, automatic annotation

Прямая речь в русской прозе: разметка и автоматический аннотатор

Никишина И. А. А.[†]* (irina.nikishina@mail.ru)
Соколова И. С.[†] (irina.sokolovalxxxix@gmail.com)
Тихомиров Д. О.[†] (dan.tijomirov@gmail.com)
Бонч-Осмоловская А. А.[†] (abonch@gmail.com)

[†]Высшая школа экономики, Москва, Россия
*Институт системного программирования им. В.П. Иванникова РАН, Москва, Россия

Определение прямой речи в тексте – одна из важнейших задач автоматической обработки естественного языка, которой в настоящее время, к сожалению, не уделяется должного внимания. В то время как парсеры и корпуса прямой речи существуют для многих европейских языков, для русского языка исследования в данной области практически не проводились. Знание границ прямой речи позволило бы значительно улучшить качество автоматических аннотаторов в таких задачах, как разрешение кореферентности, а также при создании диалоговых систем. Таким образом, основной целью данной статьи является привлечение внимания российского NLP-сообщества к задаче выделения прямой речи в тексте, путем создания эталонного корпуса и автоматического аннотатора для оценки и сравнения результатов.

Ключевые слова: прямая речь, анализ русской прозы, TEI-разметка, автоматическое аннотирование

1 Introduction

This article describes the results of a pioneering project in automatic direct speech tagging in Russian literary prose. Identifying direct speech in news texts is already possible, however the task of finding direct speech in literary texts is a different kind of challenge, which has

not yet received much attention. Creating tools for tagging direct speech in literary texts is a necessity for the field of natural language processing and the digital humanities.

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 an annotated corpus with direct speech marked up with TEI standards, to be used for further research. The corpus contains a wide variety of direct, indirect, and free indirect speech, including complicated and ambiguous cases; the markup includes the speaker and addressee of every instance of speech. The corpus has been expert-annotated with high inter-annotator agreement, and it is openly available. We also introduce a parser tailored to direct speech tagging in Russian literary texts: it recognizes direct speech and accompanying author comments, and identifies speech verbs and their properties.

Thus, our work serves as the first step to automatic speech annotation in Russian prose and our parser has a large number of applications in the field of quantitative literary analysis and digital humanities. The parser 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 as training data for further research into speech in novels, including research of literary history, gender studies, or stylometry.

This paper is organized in the following way: first, in Section 2, we discuss previous works, relevant to our project: qualitative and quantitative research on speech in novels, speech tagging in Anglophone literary texts, and identification of speakers 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 make the project more useful.

2 Related Work

In this section we briefly review quantitative and qualitative studies on speech in novels, speech tagging, and speaker identification that are relevant to our task. Although our parser does not as yet identify speakers, they are identified in our annotated corpus.

2.1 Qualitative studies

Understanding how speech is represented in novels is vital to identifying it automatically. While direct speech can be found quite reliably by punctuation, free indirect speech is not marked in the same way, and since the late 20th century, ordinary direct and indirect speech in novels become trickier to identify as well. Borders between the author's and characters' speech are increasingly blurred, and punctuation conventions are often defied: instead of dashes and quotation marks, authors use brackets or no punctuation at all, thus making characters' lines visually inseparable from other characters' speech and narration (see Verdesh, Pokrovskaya, Arzyamova [11, 6, 1]).

Where punctuation cannot be relied on, grammatical and lexical features help identify dialogue. Philological studies find that characters' speech is often contrasted with narration in terms of syntax and lexical choices (see Shatalova 2011 [8], Voronovskaya 2011 [12]): dialogue uses shorter and simpler sentences and has more interjections and emotional epithets.

2.2 Quantitative & digital studies

Project GutenTag [2], a rule-based tool for analyzing the Project Gutenberg corpus, is close to what we set out to do. GutenTag relies on quotation marks to find direct speech, and outputs a set of tags in TEI format.

Muzny et al. 2017 [5] measure to what extent a span of text is dialogic, i.e. close to natural spoken dialogue. The authors successfully use regular-expression-based algorithms to extract dialogue, then identify factors (parts of speech and grammatical forms) that reliably distinguish dialogue from narration: for example, modal verbs are "strongly associated" with speech, while adjectives are more frequent in narration. The conclusions, some of which contradict those of the philological studies cited above, can be used for correcting and fine-tuning speech tagging tools when the rule-based approach fails.

2.3 Speaker identification

The speaker identification model described in He et al. 2013 [4] exploits the typical pattern in conversations: when two characters talk, both are explicitly named in the beginning of the conversation, and then they take turns to speak. Thus, consecutive utterances are spoken by different speakers, and the speaker of the nth utterance is likely to be the same as that of the $(n-2)$-th utterance. The article compares several models, and the one that takes into account this pattern performs better.

Speaker disambiguation is a related problem, addressed in Vala et al. 2015 [10]: a rule-based pipeline extracts characters from text and builds a graph of character names connected by edges if the names refer to the same character. This 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 in the following way: we aspire to create a gold standard corpus annotated with tags for direct, indirect and free-indirect speech. We also develop a rule-based parser that focuses on direct speech and some of its components. Both corpus and parser may prove useful in any future literary research, including, but not limited to, social network extraction and comparative studies, as demonstrated by studies which employ English language counterparts. The present paper aims to describe our progress towards developing such a parser and the results it yields compared to human annotators. We assume that clear and properly defined rules for usage and punctuation of direct speech in Russian texts could be easily transformed into a rule-based parser that provides adequate results.

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[1]). 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

[1] http://www.tei-c.org/index.xml

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" [3], 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.

Most studies that deal with direct speech in Russian literature and the relevant punctuation (notably, *D.E. Rosenthal's Russian Language. Orthography and Punctuation* handbook [7] which lists the typographical conventions followed in all contemporary publications) 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 an 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 in the following way:

<**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 whether speech is 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 hint towards 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 [7]. There are many ways speech can be punctuated in Russian prose: as a dialogue with each utterance starting on a new line, or with the whole conversation taking place inside a single paragraph with each utterance enclosed in 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 here that the scope of our parser is, for the time being, limited exclusively to direct speech. Direct speech constitutes a sizeable portion of all non-narrative text in literature and can be identified by rules quite well due to clear-cut typographic conventions. Deviations from said conventions, however, present a serious challenge.

Our parser employs a four-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 4 steps: "Quotation marks replacement", "Speech tagging", "Author comment and Said tagging", "Speech verb tagging". 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 that is mentioned in Section 2.1 of Muzny et al. 2017 [5]: due to the nature of our corpus (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 the texts is far from homogeneous. To combat that, we have included a special step in our pipeline, which converts different kinds of quotation marks into opening and closing guillemets (« »). This way, we can safely use regular expressions at later steps to detect the beginnings and ends of utterances. 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 scope, <**speech**> usually corresponds to a full sentence which includes at least one <**said**> and may include optional <**author_comment**> elements; however, since in some instances an utterance and the accompanying comment can be embedded into a larger sentence with parts of narrative that have no relation to the utterance, in certain <**said**>-<**author_comment**> patterns it seems 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" **corresp**="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 the one 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 sentence splitting step 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 such an 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 preexisting 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 only have one attribute, and our verb tagger ignores any additional semantic information which can be extracted from the context.

5 Gold corpus

We use a sizeable (492 works) collection of classical Russian texts written in the 19th, 20th, and 21st century from Maksim Moshkov's Library[2] to provide a sufficient number of examples for each of the many ways direct speech is marked in Russian prose. Furthermore, we have also compiled a shorter collection, containing a wide variety of types of direct, indirect, free indirect speech as well as specially selected challenging or ambiguous cases.

During selection of excerpts to include in our corpus, our main criterion was a significant presence of conventionally punctuated direct, indirect and free indirect speech. We pay attention to position and number of <**said**> and <**author_comment**> elements, take into consideration different values for **@aloud**,**@semantic** and **@emotion** attributes. Here we provide some relevant examples from our corpus:

(1) *Помнится, когда я ехал по берегу Байкала, мне встретилась девушка бурятка, в рубахе и в штанах из синей дабы, верхом на лошади;* ***<speech> <author_comment>****я* ***<speech_verb semantic="speech" emotion="question">****спросил****</speech_verb>*** *у неё,* ***</author_comment> <said type="indirect" who="Narrator" corresp="Buryat" aloud="true">*** *не продаст ли она мне свою трубку,* ***</said> </speech>****и, пока мы говорили, она с презрением смотрела на моё европейское лицо и на мою шляпу, и в одну минуту ей надоело говорить со мной, она гикнула и поскакала прочь.*

(2) *Но между тем странное чувство отравляло мою радость: мысль о злодее, обрызганном кровию стольких невинных жертв, и о казни, его ожидающей, тревожила меня поневоле:* ***<speech> <said who="Author" corresp="None" type="direct" aloud="false">****"Емеля, Емеля!* ***</said>*** *–* ***<author_comment><speech_verb semantic="thought" emotion="neutral">****думал****</speech_verb>*** *я с досадою;–* ***</author_comment> <said who="Author" corresp="None" type="direct" aloud="false">*** *зачем не наткнулся ты на штык или не подвернулся под картечь? Лучше ничего не мог бы ты придумать".* ***</said> </speech>***

(3) ***<speech> <said type="direct" who="Mikhail" corresp="Andrey" aloud="true">****— А вот и я!* ***</said> <author_comment>*** *—* ***<speech_verb semantic="speech" emotion="neutral">****говорит****</speech_verb>*** *он, входя к Андрею Ефимычу.* ***</author_comment> <said type="direct" who="Mikhail" corresp="Andrey" aloud="true">****— Здравствуйте, мой дорогой! Небось я уже надоел вам, а?* ***</said> </speech>***

We also tried to pay attention to the presence of various types of punctuation and some extralinguistic features: authorship and creation date.

Altogether, we used 52 works for building our gold-standard corpus of 365 excerpts that are commonly available on github[3].

In the table below we provide some statistics about our corpus:

We have not used the help of third-party experts, so we annotate the corpus ourselves, as native speakers with the sufficient linguistic background. In order to estimate corpus annotation consistency and its confidence level we measure inter-annotator agreement. Therefore, Fleiss' kappa test results are provide in the table 1.

It can be seen that the Fleiss' Kappa is fairly high, especially for annotators A1 and A2. Thus, we assume our corpus to be accurately and thouroughly annotated.

[2] http://lib.ru/

[3] anonymized

Total number of tokens:	10161
total number of labels	**value**
speech	4417
author_comment	4919
said	6823
speech_verb	4141
total number of mentions	**value**
speech	152
author_comment	126
said	206
speech_verb	111

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	

6 Evaluation Results

Evaluation of our parser comprises four parts corresponding to the number of instance of speech implemented: <**speech**>, <**said**>, <**author_comment**> and <**speech_verb**>.

For the evaluation process, we converted both gold data and automatically tagged gold texts into conll format and applied Perl evaluation script from CoNLL-2000 shared-task [9]. The results for each instance of speech are provided below (see tables 2 and 3 and 4).

From table 2 we can see, that our parser detects almost 60% of speech constructions in text. At the same time, lower scores of **said** and **author_comment** annotation are quite evident, as we do not implement any semantic analysis of **speech** content and rely on punctuation only. Moreover, the phenomenon of a low score for **speech_verb** tag might be explained by the fact that it could not be defined at all if its <**author_comment**> umbrella tag is not identified. Actually, that can be considered a limitation of our parser, as it relies on <**author_comment**> and not <**speech**>).

The obtained results as well as the analysis of the drawbacks are presented in Section 7.

Table 2: Accuracy

tag	*accuracy*
speech	58.98%
said	52.68%
author_comment	55.23%
speech_verb	60.14%

Table 3: Raw results

tag	*precision*	*recall*	*F1*
speech	52.03%	72.11%	60.45%
said	88.79%	33.80%	48.96%
author_comment	62.33%	19.01%	29.13%
speech_verb	91.74%	2.41%	4.71%

7 Discussion

As we mentioned above, 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.

The total volume of manually annotated excerpts is 37,58 percent of the whole corpus of XIX and XX century literature and it comprises 155632 tokens that might be used for further training, as it contains both typical and more complex cases. It is also the first commonly available dataset for speech detection in Russian. This corpus could be also expanded by texts annotated automatically with help of our parser with further manual correction.

Evaluation (see above) shows that our parser performs well on sentences which are not overburdened with complex grammatical constructions. In addition, we achieved a rather good performance for detecting <**speech**> as the largest direct speech container. However, the following improvements are still required:

- Adding more rules for detecting speech in raw text;
- Making rules more independent of line breaks;
- Making rules more flexible to cover complex syntax structure;
- Applying a syntax parser to the speech verb detection pipeline step;
- Applying co-reference resolution in order to identify all participating characters, like an intended recipient of speech, and not just the speaker;
- Implementing a sentiment analysis step for detecting semantic characteristics of author comments.

Another direction for further development might be detection of indirect and free-indirect speech. It would be also interesting to build sentence classifier that detects speech type using specific features. Furthermore, with a well-annotated corpus, any kind of literary analysis may be performed. For instance, from such corpus we may gather information about most common speech verbs and their semantic characteristics. The utility and value of such an annotated corpus of direct speech are hard to overstate.

Table 4: Mention results

tag	*precision*	*recall*	*F1*
speech	16.88%	30.95%	21.85%
said	6.25%	5.78%	6.00%
author_comment	7.35%	7.46%	7.41%
speech_verb	52.78%	45.24%	48.72%

8 "Speech" web service

Another feature of our project is the web service we provide to ease the user's interaction within the corpus and parser. We allow users to download corpus directly from the website[4] and to process their own text files online. Moreover, advanced users might connect directly to our parsers by means of an API or even install it as a python library via setup. The interface of our web service is presented in the images 1, 2 and 3.

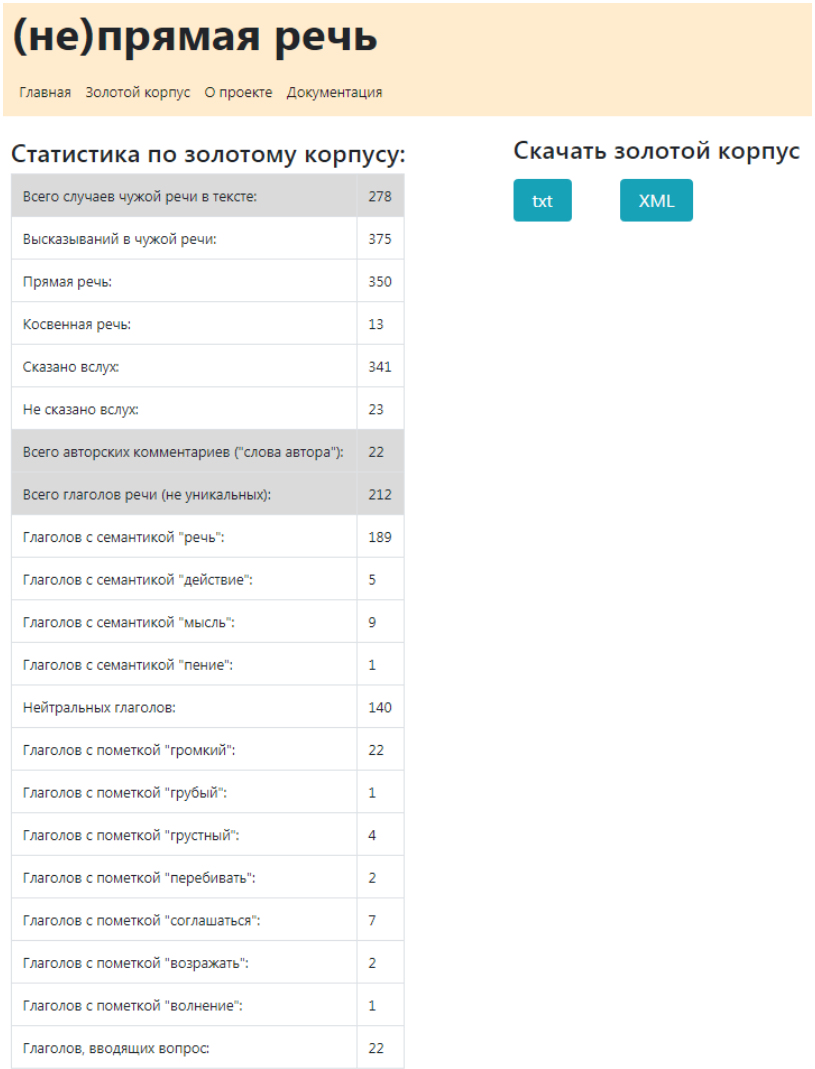

Figure 1: Page for downloading gold corpus

[4] https://direct-speech.linghub.net/web/

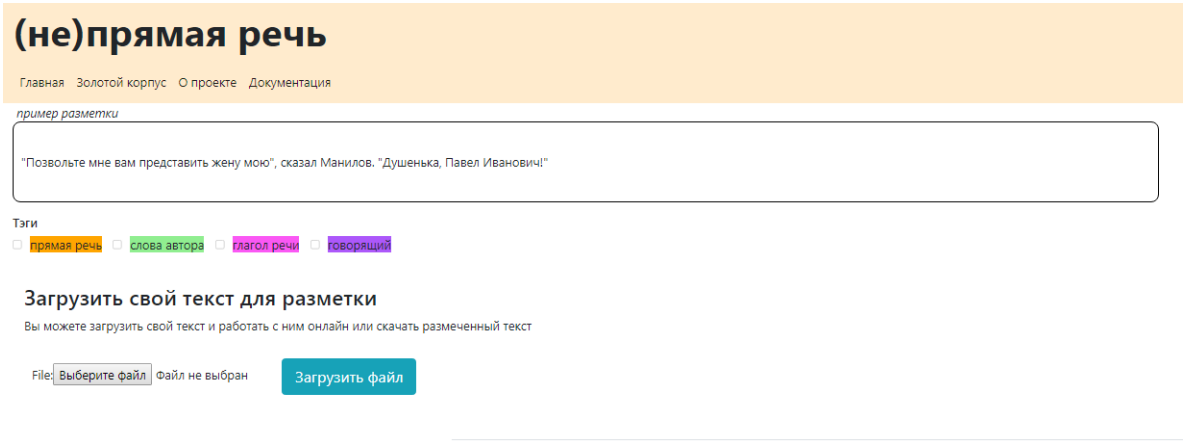

Figure 2: Main page for document processing

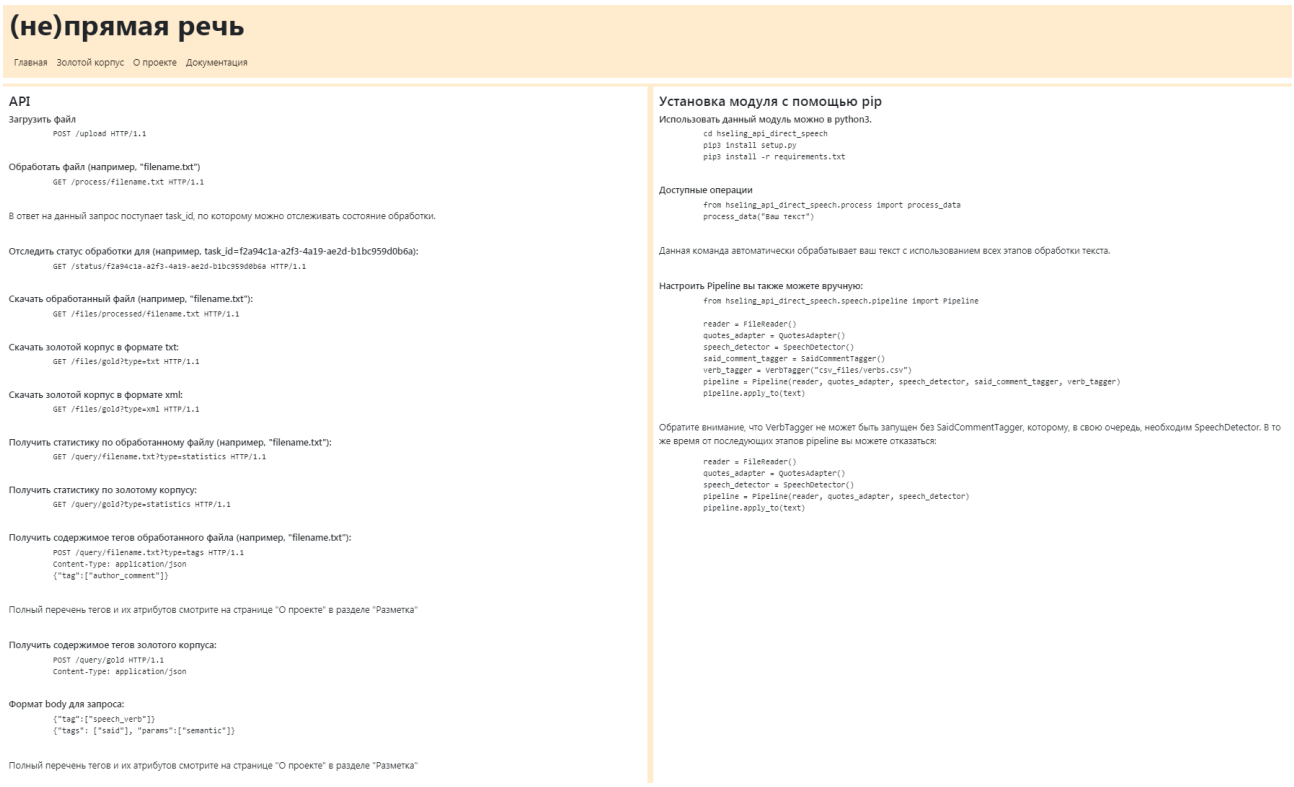

Figure 3: Documentation page

References

[1] Olga Arzyamova. Avtorskoe oformlenie pryamoj rechi kak sredstvo vyrazheniya idiostilemy v novejshem russkom hudozhestvennom diskurse_. *Izvestiya Voronezhskogo gosudarstvennogo pedagogicheskogo universiteta*, (3):147–152, 2016.

[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] TEI Consortium. TEI Guidelines "3.3.3 quotation" tei p5: Guidelines for electronic text encoding and interchange, 2018.

[4] 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.

[5] 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.

[6] Elena Pokrovskaya. Chuzhaya rech i dialog v potoke soznaniya (based on materials of russian literature of XX century). *Political linguistics*, (16), 2005.

[7] Dietmar Rozental. Russian language referencebook: orthography and punctuation, 1994.

[8] Olga Shatalova. Avtor i personazh: sintaksicheskaya reprezentaciya v nesobstvennopryamoj rechi. *Vestnik Kostromskogo gosudarstvennogo universiteta*, 17(2), 2011.

[9] Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics, 2000.

[10] 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.

[11] Anush Verdesh. Pryamaya rech v hudozhestvennyh tekstah neklassicheskoj paradigmy novye yavleniya. *The World of Science, Culture, Education*, (5):319–321, 2015.

[12] Irina Voronovskaya. Poryadok slov v avtorskoj rechi b akunina. *Izvestiya Saratovskogo universiteta. New edition. Philology edition. Journalism*, 11(4), 2011.