ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Максимова Дарья Максимовна

КЛАССИФИКАЦИЯ КОРОТКИХ ТЕКСТОВ НА ПРИМЕРЕ АВТОРСКИХ РЕМАРОК В РУССКОЙ ДРАМЕ XVIII-XX ВВ. (SHORT TEXT CLASSIFICATION: A CASE OF STAGE DIRECTIONS IN RUSSIAN DRAMA FROM THE 18TH TO THE 20TH CENTURY)

Выпускная квалификационная работа студентки 4 курса бакалавриата группы БКЛ151

Академический руководитель образовательной программы	Научный руководитель
канд. фил. н., доц.	Ph.D., доц.
Ю.А. Ландер	Ф. Фишер
______________________	______________________
«____» ____________ 2019 г.	

Москва, 2019 г.

Contents

1 Introduction **1**

2 Background **5**

2.1 Literature review 5

2.1.1 Digital drama studies and stage directions 5

2.1.2 Short text classification 8

2.2 Russian Drama Corpus and its analogues 10

2.3 TEI element <stage> and its documentation 12

3 Working with corpus and its data **14**

3.1 Play selection 14

3.2 Annotation 15

3.3 Data analysis annotation-wise 16

3.3.1 Ambiguous cases 16

3.3.2 Type distribution 18

4 Classification **21**

4.1 Preprocessing and feature extraction 21

4.1.1 Workflow 21

4.1.2 Final dataset description 25

4.2 Problem statement 26

4.3 Possible approaches 26

4.3.1 Rule-based 26

4.3.2 Logistic regression 27

4.3.3 Other algorithms 28

4.3.4 Neural networks and their perspectives 29

5 Results **31**

5.1 Metrics 31

5.2 Best models and their performance 32

6 Conclusions **35**

6.1 Classification improvement 35

6.2 TEI standard and Russian stage directions 36

Appendix 1: Links **41**

Appendix 2: Annotation guide **42**

Appendix 3: Model performance on direction types **45**

«Ô didascalie, dis-moi ce que tu lis, je te dirai qui tu es. . . »

— (Issacharoff 1981)

1 Introduction

Drama is one of the three main genres that originated in Ancient Greece, along with poetry and prose. With this in the background, it had always received much attention from the researchers. Drama analysis, consequently, had always been present in the field of literary studies.

There already is numerous research on the subject, as drama is a unique genre which combines written and oral text (i.e., both the text to be pronounced by the actors and the text to be read and interpreted by the cast and/or the readers are present). Such research dealt both with structure and content; however, the majority of this research had been conducted in a "philological" way.

With the advent of digital literary studies, research on drama has gained new impulses — it was studied in depth due to the fact that dramatic texts are relatively easy to parse and analyze with statistical and computational methods. This structure uncovers when one has read a number of plays. For example, character names are typed in all capitals; the list of characters is placed between the title of the play and the beginning of the first act; the structural parts of the play (that is, the scenes, the acts, etc.) can be identified with simple morphological analysis — if a line of the drama has the word "act" or "stage" inside it, this line is very likely to be a structural part identifying the beginning of a scene/act.

Having identified and parsed literary texts — including, but not limited to scenic ones — researchers were provided with an opportunity to study the same subject from a new angle. One of the aspects of digital literary studies is to support numbers to theoretical statements which were made during many years of literary studies; therefore, it is logical to call this type of research *proof-providing*.

Good examples of such research are stylometry and authorship attribution studies. The first research in the field was conducted in the mid-15th century — Lorenzo Vella published "De falso credita et ementita Constantini Donatione declamatio" (*Discourse on the Forgery of the Alleged Donation of Constantine*) in 1440 and proved that Declamation of Constantine is a forgery. Numerous works were written after that; even though stylometry has a long history, it is still popular with some researchers and non-fiction writers. A recent example is the case of Joanne K. Rowling who published a detective under the nickname of Robert Galbraith but was revealed with the help of stylometry.

The other type is rather *re-discovering* the well-known material. After digitizing the material, some quantitative research became possible. This type can be illustrated with topic modelling. The general task is based on the idea that all documents are composed of several topics which respectively have a vocabulary of their own. With the help of mathematical statistics and probability theory, it is possible to create such vocabularies and calculate their deal in the document under question. This task is also a part of research inspired by "distant reading" philosophy. Introduced by Franco Moretti, it became popular with digital humanists. Distant reading also made an impact on studies on social networks of characters, their speech, and other aspects.

Digital drama studies were obviously influenced by some of the methods described earlier. However, drama researchers could choose the material they would work with; one of the questions they had to answer was whether or not include stage directions. Some studies concerned them a merely functional part of the text that, as a result, was left out of the analysis. These researchers claimed that there is no need to pay any attention to the directions (otherwise called *didascalia*). On the other hand, other researchers included stage directions in their analyses or even focused on them. Existing research covers some European dramas (for example, German, French, and English), yet there is just one quantitative study on the material of Russian drama — (Sperantov 1998) — which covered only tragedies from the 18th century. This work, on the other side, covers this and other drama types over a bigger time span.

As already mentioned, stage directions are functional, which means they are all present in the plays intentionally and serve for a definite reason. In this paper, I try to classify these directions according to their function in the text. Such a research goal does not cover the intentions with which the authors used the directions or wrote them in a certain manner; neither it shows authors' motivations for including certain direction types — both of these questions are out of the scope of this work, as well as some other epistemological concerns.

This work is based on Russian Drama Corpus which uses TEI-P5 schema for its annotation. The directions are assigned a special tag which has an optional argument for the direction type. Stage directions functions I will be primarily concerned with are those listed in the latest version of TEI documentation[1] as variants (or values) for a stage direction tag.

To make the classification, the following steps are taken. First, the required data is extracted from the corpus. After that, this data undergoes automated linguistic analysis. Finally, computational instruments and machine learning is used to create a final classification tool.

The present study is, in a way, an answer to several questions. The first is, "Is it possible to get an impression of a genre without really reading many examples of this genre" This is rather a more general distant reading-inspired question than a question for a practice-oriented study. Nonetheless, the results of data analysis performed as part of the preprocessing operations allow drawing several conclusions on the general state of the corpus and its special aspects.

Apart from that, this work is an answer to another question, "Is it possible to assign a Russian stage direction its function automatically, without using any human resource?". This question poses a simply formulated, yet challenging problem that arose from corpus usage and the general wish to enrich the corpus annotation. Annotating direction types manually takes time and can be inconsistent due to the fact that different annotators may have different views on the subject, even if they are given the annotation guide.

The main objective of the present work is to develop a tool which would allow to classify stage directions from a given play, therefore to enrich annotation of Russian Drama Corpus. An additional goal is to analyze Russian stage directions and develop a system of the types present in the plays.

[1] As of May 2019, edition 3.5.0, revised on January 29, 2019.

This work is organized as follows: in the second chapter, I will provide background information: literature review, corpus introduction, and TEI documentation on the matter. In the third chapter, I will cover the annotation process from play selection to statistical analysis of ready annotations. In the fourth chapter, I will describe the classification process in detail: problem statement, algorithms which were used or considered, approaches to different direction types. The fifth chapter is dedicated to the results — metric choice, models quality, and result analysis. Finally, the sixth chapter draws the conclusions and states new research questions.

2 Background

2.1 Literature review

2.1.1 Digital drama studies and stage directions

Digital drama studies are generally based on, or somehow related to, the concept of "distant reading", introduced in (Moretti 2000). Its main idea is to abstract from focusing on a single item[2] to observing and processing many items at once. Having used "distant reading" techniques on Shakespeare's *Hamlet* in (Moretti 2011), Moretti created graphs representing character relationships: the vertices represented play characters, edges showed whether the two given characters had interacted. A more complex version of such graph included edge weights that showed the amount of interaction: the more the interaction was, the more weight an edge had. That allowed to represent the relationships throughout the play: the two characters may only have interacted once, but the edge connecting them always stayed in the graph. With the help of these graphs Moretti came to a conclusion that was not seen when close-reading the play: Horatio, who is supposed to be one of the main characters, does not affect the connections between characters that much — due to the fact that in a high clustered graph concerning the traditionally main characters, removing one of them does not have a big impact on the network in general. In fact, *Hamlet* graph without Horatio only misses several characters. The resting main ones are still present, which means that Horatio does not much affect a play. This proved that using quantitative methods on drama material is reasonable, therefore giving ground to many other research papers. Since that time, network analysis (and drama analysis, too) had become popular among digital humanists.

Stage directions have rarely been a main object of study. In fact, (Rasmussen 2003) even claimed that stage directions do not carry any sense and do not matter when analyzing theatrical text. However, there is numerous research supporting the opposing view.

The importance of stage directions for the plays was one of the main arguments

[2] Moretti was talking about reading (hence the term name includes "reading"), nowadays the term and the concept is also applied to other types of data.

in (Detken 2009). In her work, Anke Detken claimed that stage directions are not only in the text for the interpretation of a play but also for conveying author's thoughts and ideas. She split the directions in German drama into technical ones and more fictional ones, thus creating the classification of her own. The analysis was conducted in a philological manner and tradition. She read the plays closely and attentively to get an impression of stage directions in the plays and interpreted those plays based solely on the directions. In fact, *Im Nebenraum...* is a qualitative study, while the present paper is meant to be quantitative.

There are some studies on the other dramas, too. For example, Simone Dompeyre provides a detailed and thorough classification with big types and their subtypes (Dompeyre 1992). The majority of the bigger types correspond to TEI <stage> types. Another study (Issacharoff 1981) describes stage direction in French drama from a syntagmatic view, developing a typology of for types: out-of-text directions (French: *le hors-texte didascalique*), which are generally big pieces of text such as decorations description, or big descriptions at the beginning of a scene/act, or a description of multiple actions at once. The second type is self-contained directions (French: *didascalies autonomes*), which describe the way to read the text — that is, the way actors should speak on stage. The third type is normal directions (French: *didascalies normales*) that address both the actors and the audience. The last type is "non-readable" directions (French: *indications scéniques "illisibles"*) — these are the instructions for the play director. They describe stage business which does not involve the actors — curtains drawn or fallen, items on stage interacting with each other, etc.

Italian drama has been researched, too. In (Titomanlio 2011), the author describes the way Italian theatre was becoming more and more Brechtian. This idea is proved with several arguments, stage directions getting bigger and bigger and describing the evolution of stage directions from static descriptions to involving action on the stage among these points. This corresponds to one of the ideas in (Maximova, Fischer, and Skorinkin 2018), where the same trend is observed in Russian drama.

Russian drama first got into spotlight with (Sperantov 1998). In fact, this was the first attempt to perform quantitative analysis on Russian drama. Having taken tragedies from late 18th–19th centuries, Sperantov developed and calculated the following metrics based solely on stage directions:

- density coefficient, i.e. frequency of stage directions occurrence in a given play,
- average direction length,
- cooperation intensity — frequency of the cases when a stage direction occurs "inside" a verse or a character's speech,
- lexical variety coefficient — how many different lexemes were used in stage directions of a given play,
- emotionality coefficient.

All the metrics were normalized; after that, a so-called "classicality" metric was calculated for each play. This metric showed how close a play was to the classic canon of tragedy. Based on that metric, Sperantov divided Russian tragedies into those that follow the canon quite strictly, those with minor retreats, plays with major retreats from the canon, and decidedly non-classic plays.

As can be seen, Sperantov used stage directions to describe plays in general (to be precise, only tragedies). All the calculations and annotations were made manually — for instance, Sperantov decided himself which words should be considered emotional and which should not. More than 20 years later, it is now possible to reproduce this research, use the results of this research, or make it scalable. It is quite important to understand that this work uses the same material (that it, stage directions) in a slightly different way — rather than using directions to describe plays, I will not try or implement any play classification.

However, (Sperantov 1998) is not the only work on Russian drama. In PhD thesis (Usovski-Rosa 2010), a detailed and close analysis is provided for the staged play design elements (German: *die szenischen Gestaltungsmittel*) in Anton Chekhov's plays. Stage directions are the first element to be discussed there. In a separate chapter, the author articulates the importance of stage directions for the text, highlights the range of ideas Chekhov brought to his plays with the help of directions and makes a classification of her own. Not only she distinguishes “technical” and “fictional” directions but also divides fictional directions into categories.

This is an important milestone for the present research due to a number of reasons. To begin with, (Usovski-Rosa 2010) places Russian drama in the European context, which allows to make comparisons and claim that some linguistic and/or structural drama features may be shared — either they might have been borrowed from European literary tradition to Russia, or vice versa. Secondly, it is clearly seen that there had been numerous attempts to classify the directions. This indicates that the idea of categorizing directions is still relevant as no united and agreed-upon system has been introduced yet and there is an open question on that.

As a foreword to the present work, (Maximova, Fischer, and Skorinkin 2018) used Sperantov’s strategy and described the plays and the general trends. They observed an epification trend (German: *Epifizierungstendenz*) which indicates that Russian drama had been borrowing structures, vocabulary, and other features from prose and other literary genres (cf. (Titomanlio 2011)).

2.1.2 Short text classification

The classification problem is one of the major problems in machine learning. Many algorithms were developed for this task, including those regarding textual data. Under “textual data” the researchers usually meant quite big texts (several sentences, or even paragraphs) that allowed to extract a large number of linguistic and non-linguistic features. On the contrary, short texts present an additional challenge: due to their size, it is problematic to extract or calculate any features. Development of algorithms and approaches to short text classification problem were developed on two well-known datasets: *Reuters-21578* and *20 Newsgroups*.

The first one, *Reuters-21578*, is probably one of the best-known datasets for classification, especially when taking text classification (also called text categorization) in consideration. It represents 21578 text pieces that appeared in Reuters news feed in 1987. The dataset used to be bigger (22173 texts); this collection was used and discussed at ACM SIGIR[3] conference in 1996, where the data was cleaned, some

[3] ACM Special Interest Group on Information Retrieval

features were dropped out, or replaced. The finalized dataset was released later in 1996, with some duplicates removed.

Rule-based systems were applied to the dataset. Normally it would be necessary to create a universal dictionary of words, select those that are present in a given text, and derive rules from these words. Another approach, introduced in (Apté, Damerau, and Weiss 1994) suggests to create local vocabularies for each topic in the collection, then analyze texts for the presence of those words, creating a boolean feature table. After that, a set of rules had to be set. It was rather general at first; then, the rules were chosen heuristically. If a given rule would outperform other rules for that class, the latter were dropped from a rule set. Next, finalized sets were "decomposed" with a reference to the travelling salesman problem, which allowed to make the rules much less complex with a small decrease in quality. As a result, their rule-based models outperformed those built on decision trees and probabilistic Bayes.

Another dataset, called *20 Newsgroups*, consists of ca. 20000 articles collected from Usenet. The articles were placed in 20 different groups, with each article assigned exactly one group. One of the most popular approaches to this dataset is (Joachims 1996). For each document, the most informative words were chosen with the help of mutual information metric: the less a given word reduced the entropy, the more informative it was considered. After that, only 15 most informative words were left for each document. TF-IDF algorithm was used to create a vector space for all the documents. The general idea of using the algorithm is that the closer the document vectors are, the more they are similar to each other. Those features were passed into a Naïve Bayes classifier which uses an idea that the probability of a document belonging to a certain category was calculated from a probability of the features of that document (that is, TF-IDF vectors of most informative words) belonging to that class. As a result, the Naïve Bayes outperformed models which did not use any probabilities.

Previously mentioned studies dealt with longer (in a way, more traditional) texts. Apart from that, short text classification has also been heavily researched. Some works approach this problem from Information Retrieval aspects. For instance, (Sun 2012) uses TF-IDF and a purpose-built metric of Clarity to select the words which represent the text the most. After selection, all other tokens are dismissed. This allows to create a denser feature space and therefore use it more effectively to perform the classification.

Newer approaches involve neural networks. An example of those, (Lee and Dernoncourt 2016) proposes using artificial neural networks. Authors" main idea is using the information about preceding texts as well as the text to classify. They feed this information, as well as 300-dimensional semantic vectors, into a long short-term memory recurrent neural network. Nevertheless, the results were unstable: with some data sets, the suggested model was the most successful, with others, simpler models showed better accuracy.

There has also been an ongoing discussion on implementing linguistic features. In (Moschitti and Basili 2004), authors claim that adding complex features such as morphology information, part of speech, n-grams, syntax information, and/or semantics, does not result in a big improvement. They tested their hypotheses on 5 datasets in English and Italian. To do so, they used different combinations of features with state-of-the-art models. Unfortunately, the best enhancement they got was 1.5%. On the other hand, Yang in (Yang et al. 2013) ran topic modelling on the documents from the data set, chose the most discriminative features and used them to improve text classification accuracy.

2.2 Russian Drama Corpus and its analogues

This research was conducted on the material of Russian Drama Corpus. It is a corpus of 144[4] plays written by Russian playwrights from the middle of 18th to the middle of the 20th century. The corpus is TEI-P5 encoded. This allows to understand the structure of the play even if the person does not speak Russian.

However, Russian Drama Corpus is not the only drama corpus, there are similar projects for other European literary traditions. I have analyzed an overall of 10 corpora which use TEI standards.

When speaking of a corpus size, Russian Drama Corpus rather belongs to

[4] As of May 2, 2019

the bigger corpora of the selection than to the smaller ones. On the one hand, there are multiple corpora with Shakespearean texts — these (*Shakespeare Folger Library* and *Shakespeare His Contemporaries*) are limited to the original 37 plays. Another "closed-material" corpus is Danish *Ludvig Holbergs skrifter*, based solely on the works of Ludvig Holberg, Dano-Norwegian/Scandinavian playwright of the Northern Renaissance; more than that, this corpus is the only one in the collection consisting entirely of comedies. On the other hand, the resting five corpora are not limited to a certain author, era, or any other feature. Amongst these, Russian Drama Corpus is relatively small — the majority of the corpora are already bigger, French *Théâtre Classique* unequivocally being the leader with more than a thousand plays. As can be seen, the corpus used for this study is relatively small — especially after I've compared it to the other "multi-author" corpora — yet it is important to mention that it covers a range of plays from mid-18th to mid-20th centuries.

The majority of the corpora are encoded in TEI standard, though *Théâtre Classique* uses a previous revision of the standard P4 rather than the latest P5. This can be explained by the fact that the overall project of *Théâtre Classique* started earlier than the other projects. This might also be an explanation to the fact that the annotation is inconsistent: for example, some plays have stage direction types while others do not. More than that, *Shakespeare Folger Library* uses some tags and attributes which are not present in the standard description, e.g. sounds. Such annotation is, without any doubt, motivated by researchers' goals.

The annotation of Russian Drama Corpus is very similar to that of German Drama Corpus. Both are collected and maintained by the Drama Corpora project, which allows researchers to compare the two.

Comparisons and more figures are presented in the table below.

	Corpus	Language	Plays	Encoding	<stage> types	Other comments
1	Dramawebben	Swedish	62	TEI-P5	yes	
2	German Drama Corpus	German	472	TEI-P5	no	
3	La Biblioteca Electrónica Textual del Teatro en Español	Spanish	25	TEI-P5	no	
4	Letteratura teatrale nella Biblioteca italiana	Italian	171	TEI-P5	no	
5	Ludvig Holbergs skrifter	Danish	472	TEI-P5	no	only comedies
6	Shakespeare Folger Library	English	37	TEI-P5	yes	not all types are in TEI documentation
7	Shakespeare His Contemporaries	English	800			
8	Théâtre Classique	French	1260	TEI-P4	yes	inconsistent annotation
8	Russian Drama Corpus	Russian	144	TEI-P5	no	

Table 1: Drama corpora comparison

2.3 TEI element <stage> and its documentation

In TEI-P5 documentation stage directions are marked with a special tag <stage> (TEI Consortium 2019b). It is obligatory for the genres where such type of text is present — mainly theatrical text. This tag has an optional attribute *type* that has several values proposed by the TEI community:

- setting,
- entrance,
- exit,
- business,
- novelistic,
- delivery,
- modifier,
- location,

- mixed.

A stage direction can have multiple types except for *mixed* — this type excludes any others. The documentation provides very short and simple descriptions for each type. For example, *setting* type "describes a setting" (TEI Consortium 2019b).

On the other hand, TEI also proposes a tag <move>, which describes character movement on the stage (TEI Consortium 2019a). Its optional attributes are *type* (entrance/exit/onStage), *where* (left/right/stage), and *perf* (any other actions a character performs while moving). The main difference of <stage> types is that the values for the attributes are a closed set; apart from that, <move> seems to be a nested tag for <stage>, i.e. all the <move> entities are a subset of <stage> entities. Having taken this into consideration, <move> tag is not considered throughout this paper.

The main theoretical question of performing such classification would be its applicability. In (Maximova, Fischer, and Skorinkin 2018) I decided that *novelistic* type is not present in Russian stage directions, based on the description provided by the TEI community and TEI Consortium, though the resting types are present in Russian drama. There is also a chance that Russian drama might need some classes which are not listed in TEI documentation.

3 Working with corpus and its data

For any research and any machine learning task one needs a gold standard to use for fitting and testing the models. In this case, I had to extract the stage directions from several plays and annotate them manually.

The extraction was performed with the help of RusDraCor API (application programming interface), which allowed to extract directions for a given play with a single GET request.

3.1 Play selection

Another issue to address was setting criteria for play selection. One of the main principles was to save the corpus representation — that is, to make the play selection balanced: if there are many plays from a given time span, there should be more plays of that time in the annotation sample. The second principle was to be versatile with authors. We know for sure that A. Chekhov, known for his short stories, also made an impact as a playwright, so the presence of a big number of his plays is expectable. However, the annotation sample should not consist of 90% Chekhov plays and 10% of other authors.

Having taken that into consideration, 18 plays from 1747 to 1903 were chosen. Overall, they contained 6569 stage directions. The most popular (presented) playwrights are A. Ostrovskiy and A. Chekhov — each of them is represented with 3 plays.

The full list of the plays is:

- *Horev*, by A. Sumarokov (1747),
- *Nedorosl*, by D. Fonvizin (1781),
- *Urok dochkam*, by I. Krylov (1785),
- *Gore ot uma*, by A. Griboyedov (1824),
- *Boris Godunov*, by A. Pushkin (1825),
- *Revizor*, by N. Gogol (1835),

- *Maskarad*, by M. Lermontov (1836),
- *Svoi ljudi – sochtjomsja*, by A. Ostrovskiy (1849),
- *Holostjak*, by I. Turgenev (1849),
- *Les*, by A. Ostrovskiy (1849),
- *Tsar Boris*, by A. Tolstoy (1870),
- *Bespridannitsa*, by A. Ostrovskiy (1878),
- *Chaika*, by A. Chekhov (1895),
- *Djadja Vanja*, by A. Chekhov (1896),
- *Tri sestry*, by A. Chekhov (1900),
- *Zhivoj trup*, by L. Tolstoy (1900),
- *Vishnevyj sad*, by A. Chekhov (1903).

This list appears to be balanced due to the fact it represents each time span with an adequate amount of plays (e.g., not so many 18th century plays are present in the corpus, hence there are only 3 plays from that period in the gold standard). More than that, it does not show a visible preference to a particular author or time period.

3.2 Annotation

After having extracted stage directions for every play, they were annotated manually. Each direction could be assigned any of the direction types present in (TEI Consortium 2019b) except for *narrative*. Instead, *unknown* type was introduced for directions which did not meet any criteria for the other types. General rules for annotation were the following:

1. each direction may have several types,
2. if a direction is assigned *mixed* or *unknown*, no other types shall be used,

3. type *unknown* is reserved for the most complicated cases; it is in our own best interest not to use it.

Sometimes a direction like *Занавес.* 'The curtain falls.' or *Конец.* 'The end.' would appear; according to TEI, such texts are rather instances of <trailer> tag denoting information at the end of the text or its part (for example, an act or a scene). Such directions were first annotated as *unknown*, then dropped from the dataset used for the classification task.

Further information and examples of <stage> types are listed in Appendix 2.

3.3 Data analysis annotation-wise

With all the plays annotated, some quantitative analysis became available. Having done that, several questions regarding TEI classification arose.

3.3.1 Ambiguous cases

Russian drama has several specific traits which were addressed while annotating several first plays; after that, it was decided to write an annotation guide to fix all the conventions.

First, Russian playwrights indicate who is on stage in the given moment. That resembles *entrance* type, however, these characters might already be on stage, therefore they are not really entering the stage. Such a direction rather draws attention to the mentioned characters than really speaks of an entrance.

(1) Липочка и Аграфена Кондратьевна.
[A. Ostrovskij, *Svoi ljudi – sochtjomsja*]

(2) Г-жа Простакова, Простаков, Скотинин.
[D. Fonvizin, *Nedorosl*]

Sometimes these directions would start with words *Те же и...* "Same ones and... ", which indicates that in this part of the scene is performed by characters mentioned in the direction.

(3) Те же и Устинья Наумовна.
[A. Ostrovskij, *Svoi ljudi – sochtjomsja*]

Being limited by the direction types provided by TEI standard, these directions were still annotated as *entrance*. However, it seems logical to introduce another type, *presence*. This name is neutral and describes precisely what this type means. Another "candidate name" would logically be *attention*, but this is a worse choice due to the fact that attention is not the only purpose such directions might be used for.

Secondly, it appeared that *modifier* type is not really present in Russian drama. According to examples in (TEI Consortium 2019b), this type is used to describe any changes in character's appearance — mostly disguise. Such type of change is widely present in Shakespearean texts (Kreider 1934) — for example, Rosalind (a female character) in *As you like it* changes her appearance and disguises herself as a male character, Ganymede.

(4) ROSALIND, to Touchstone. Peace, I say.
As Ganymede, to Corin. Good even to you, friend.
[W. Shakespeare, *As you like it*]

There also are multiple examples of characters disguising in *King Lear*; sometimes these disguises are shown in character's speech, like Edgar, who disguises himself as Poor Tom after a wrong accuse of plotting a murder of his father.

(5) I heard myself proclaimed,
And by the happy hollow of a tree
Escaped the hunt. <...>
"Poor Turlygod! Poor Tom!
"That's something yet. "Edgar" I nothing am.
[W. Shakespeare, *King Lear*, Act 2, Scene 3]

In Russian drama, disguising is not a frequent action — quite on the contrary, it is very rare. In this work's sample, no disguises were found, therefore type *modifier* is not present at all.

3.3.2 Type distribution

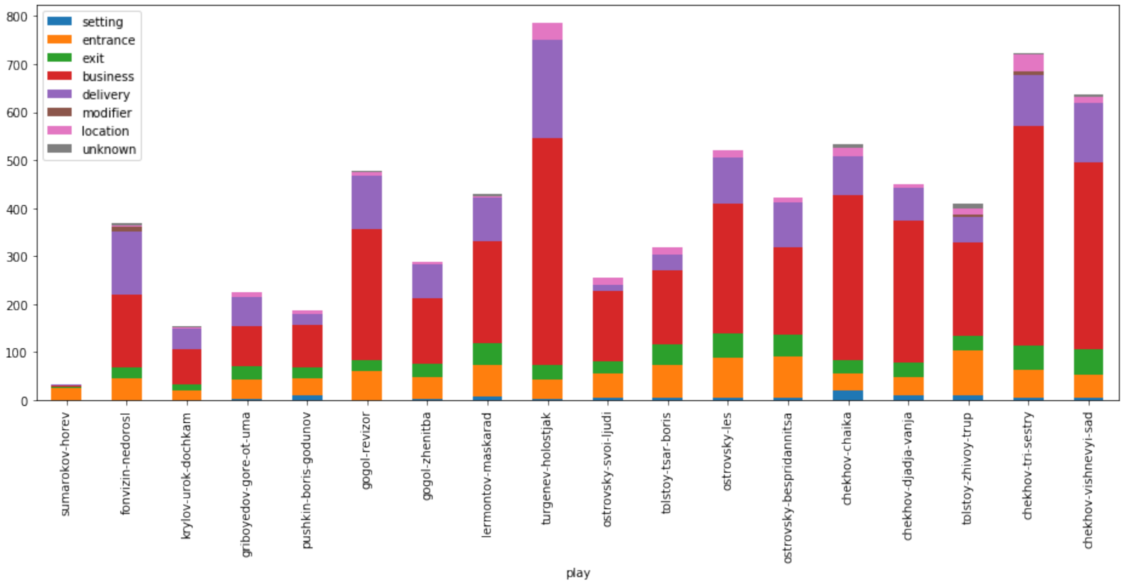

Figure 1: Direction types distribution in plays (absolute values)

In Figure 1, which shows the proportions of different direction types in plays, it is clearly seen that business type prevails over others. However, there is a semi-exception where the dominance is not drastic — it is Fonvizin's *Nedorosl*, where the amount of *business* directions is almost equal to those of *delivery*.

Sumarokov's *Horev* consists prevalently of *entrance* type, which can be explained by the influence of classic plays. This is the oldest play in the corpus (dated 1747) which explains a small number of directions — in that time, classic canons were followed strictly, and a small number of stage directions was typical for a classic play. Frequency of type entrance may be is a inherited from Ancient Greek drama, which occasionally (yet not on a regular basis) marked entrances or exits (Carlson 1991).

Second place goes to either delivery (*Gore ot uma*, *Revizor*, all Chekhov's plays) or entrance (*Boris Godunov*, *Svoi ljudi — sochtjomsja*, *Zhivoy trup*). There is no visible correlation with the year the play was written. The only hypothesis is that this correlated with the number of characters in the play — it is quite logical to suppose that a play with many characters would enter the scene more than in that with fewer characters. On the other hand, the former are relatively small regarding cast size (*Gore ot uma* and *Revizor* — 31, *Chaika* — 12, *Tri sestry* — 14, *Vishnevyj sad* — 14), the latter have bigger ones. For instance, *Boris Godunov* has 79 characters — more than two times more if compared to *Gore ot uma* and *Revizor*. *Zhivoy trup* has 40 characters, comparable with other delivery plays. *Svoi ljudi — sochtjomsja* has the least amount of characters in the whole sample — just 8. In case of these plays, we can rather admit that the characters are more "fast-moving" — for instance, in *Svoi ljudi — sochtjomsja* a character may leave the stage and then enter it back several (many) times throughout a single scene.

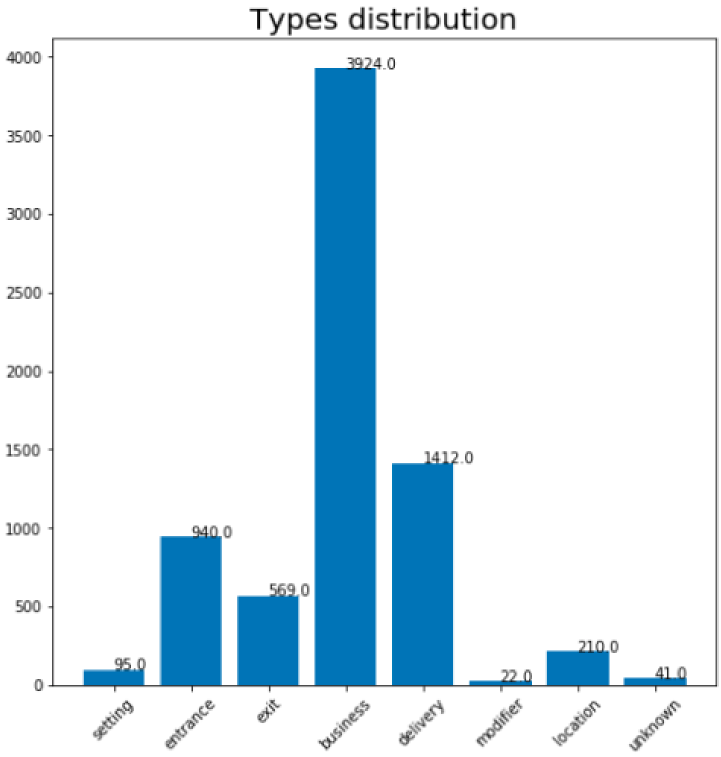

Figure 2: Overall type distribution (absolute values)

Figure 2 displays another aspect of distribution — overall type distribution throughout all annotated plays. The sum of columns' values gives a number greater than a total number of annotated directions which may occur due to the fact that one direction may be assigned more than one value (with an exception for types *mixed* and *unknown*).

One of the most evident observations is the absence of type *mixed*. When the annotation guide allowed to choose several types at once, mixed was not necessary anymore. In (Maximova, Fischer, and Skorinkin 2018) annotation strategy was different and did not allow multiple types at once, therefore type *mixed* was well represented. The following example would receive a single *mixed* tag in 2018 paper; instead, it is now annotated as *entrance* and *business*.

(6) Маски, Арбенин, потом князь Звездич. Толпа проходит взад и вперед по сцене. налево канапе.
[M. Lermontov, *Maskarad*]

The most popular type without any doubt is *business*. This is quite expected, as Russian drama became more and more "active", it used a bigger amount of verbs to indicate the actions (Maximova, Fischer, and Skorinkin 2018). In fact, such a predominance of this type may also be explained by the fact that bigger (direction-wise) plays include many business directions, thus raising the bar very high. Nevertheless, if these values are turned into relative numbers, the trend stays the same.

In contrast with Chapter 3.3.1, there is a little number of directions annotated with *modifier*. These primarily are from Lermontov's *Maskarad*, where some of the characters wear masks. It is highly important to understand that this is only an outlying case, therefore on the contrary with Shakespeare's plays.

(7) 1-я маска входит быстро в волнении и падает на канапе.
[M. Lermontov, *Maskarad*]

The small amount of *unknown* directions represents other types of functional text which were unintentionally annotated as stage directions.

4 Classification

For each TEI direction type, several models were trained. The reason for this decision was No Free Lunch theorem which — in terms of machine learning — states that a single algorithm cannot be good enough to perform well on all problems and/or all types of data. If a model is highly specialised, then it is expected to give a good result on the specialised data. However, if several specialised models are taken and tested on all the data (both the one the models were coordinately created and fitted and all data for other models), the overall result will be as good as guessing the outcome randomly (Wolpert 1996). Stage direction types are different in their purpose and structure, therefore it is quite strange to believe that a single model will put the directions in all the classes correctly.

Train/test data had already been annotated — it comprises 18 plays (see Section 3.1 for more details). This gave approximately 6,500 stage directions with their types assigned manually. 66% of these directions will be used for training the models. After training, the models were tested on the resulting 34% of the dataset. To ensure that models do not overfit while learning, 20% of the test set was left out of the training as a validation set. The validation set is relatively small, and if a model performed badly on it, it was expected to have a low score on the test set.

For some types, a rule-based approach was expected to have quality comparable with machine learning models. For others, different algorithms were tried and implemented.

4.1 Preprocessing and feature extraction

4.1.1 Workflow

For each direction, feature vector was computed as follows:

Morphological and part-of-speech analysis performed by Mystem Mystem is a rule-based morphological analyser developed in the 1990s. It uses Grammatical dictionary of Russian originally created by A. Zaliznyak in 1977 (the analyser is built on the 1980 revision). Even though the original source seems to be outdated, it still covers the majority of Russian inflectional and derivational morphology. With this in its core, Mystem is able to analyse unknown words (Segalovich 2003). Such analysis is based on the idea that any new word still has the same morphological properties and features as other, “regular” words of the language. This fact is also applicable to the words which are not used anymore. Consider the following example:

(8) обробев и иструсясь
[D. Fonvizin, *Nedorosl*]

Neither the first nor the third word is used in contemporary Russian. In fact, these words were found in a 1840s dictionary yet nowhere else. Speakers of Russian (both native and those who know Russian at the quite advanced level) may guess their meanings by finding similar words: *обробев* seems to be similar to the word *оробеть* ‘to lose one’s nerve’, and *иструсясь* looks like a close relative to *трусить* ‘to lose courage’. However, it is immediately clear that both are past participles, derived from verbs *обробеть* and *иструсить* respectively, which also are not used in standard Russian nowadays. Speakers of Russian know that past participles may be formed by adding *-ив* to the original verb — this model was also described in the Grammatical dictionary of Russian, and Mystem knows how to parse this word from a morphological point of view.

Mystem is also able to evaluate several analyses of the same entity by their probability. By default, the most probable analysis was taken.

Named entities recognition Named entity recognition (NER) is a task of identifying all named entries in a given, raw (i.e. as-is, without any preprocessing) text. These entries might be personal names, company/organization names, places, etc. NER is generally considered a subtask in Information Retrieval domain. There are several algorithms developed specifically for solving this problem in any given text. In the case of stage directions, the majority of named entities are just personal names. Specific NER algorithms are quite complicated and they are retrieving more information that might be required, so such algorithms are “overkills” of a certain kind. Instead, Mystem was used.

Another option provided by the tool is predicting whether a given token can be interpreted as a name, surname, or a patronymic name. This is simple enough for a text like a stage direction, and such an approach does not make the final model and preprocessing stages more complicated than they might be.

These NER-like name grammemes appear in the resulting token analysis. In general, one of the preprocessing functions analysed a pair of lemma and its Mystem analysis. In case any of the name grammemes (*имя* 'name', *отч* 'patronymic', *фам* 'surname') are found in the analysis, item's part of speech is assigned the value PERSN (person) and its lemma is converted to *Имя* (name).

The goal of performing this type of NER is to reduce the sparsity of the feature array. Having many different names distorts any model, when in fact all these different names have exactly the same meaning in the context of stage directions — they denote a character who is involved in some kind of stage business. This business may either be "active" (a character performs an action) or "passive" (they are addressed) yet anyways this business concerns characters in general, there is no need to save the names themselves.

Semantic vectors Another important (or even critical) feature In order to gain more information and more features from stage directions. One of the approaches to this is TF-IDF algorithm (term frequency, inverse document frequency). TF-IDF allows to indicate the most important words of the dataset. A word weight is computed as its relevant frequency in the texts (i.e. the ratio of word occurrence amount and overall corpus size) multiplied by an inverse number of documents it is present in. These words weights allow to present documents as vectors of words which are present in those documents. However, TF-IDF has several limitations: for instance, it is not able to handle (or compute a vector for) unknown words in an adequate manner — these new words are assigned a vector which is not similar to any other vector. Apart from that, the resulting TF-IDF result relies heavily on the documents it was trained on. If input corpus is unbalanced, the outcome would not represent standard language and it could not be used for work with any other data which is unbalanced in a different manner; neither could it be used with the standard balanced corpus.

An alternative to TF-IDF is distributional semantics. It is widely used when addressing linguistic problems and tasks computationally. The general idea behind it was formulated by John Rupert Firth, an American linguist: “You shall know a word by the company it keeps”. It is possible to gather a large number of texts and analyse which words tend to stick together and which do not. If two words appear in similar contexts (i.e., if they co-occur with the same set of words), they are supposed to be semantically similar.

One of the main implementations of this idea is word2vec, a group of models that use a large corpus to create word embeddings (“vectors” for every word in the corpus) and turn them into a many-dimensional vector space. After training, the model receives a word and returns its vector. Word2vec may also show how similar a pair of words is. Firstly, the model computes vectors for each word. Secondly, the cosine similarity between the two vectors is calculated. The closer cosine similarity is to 1, the more similar is the given pair of words; on the contrary, if cosine similarity is close to -1, the words are as different as possible. In fact, similarity of 1 between two words means that these words are complete synonyms of each other, similarity of -1 shows that words are antonyms.

If there is need to compute a word2vec vector for a text which contains several words, a model calculates vectors for each word in the texts, sums them, and then divides this “sum-vector” by the word count. The resulting vector represents the whole text.

There are two algorithms to train a word2vec model: continuous bag-of-words (CBOW) and continuous skip-gram. They are different in the ways they interpret the initial corpus passed for learning the words and computing vectors. CBOW defines a word by its context (that is, words will be similar if they have similar contexts) and skip-gram, on the contrary, learns words and tries to predict the context starting from the word in question. Some stage direction types are believed to be quite typical: for example, *delivery* often consists of a preposition and a name or a pronoun: *к нему* ‘to him’; in other cases, the preposition is omitted: *Ирине* ‘to

Irina'. The similarity of the contexts leads to the hypothesis that a CBOW model is more applicable due to the fact it mainly takes the context into consideration, instead of the words (like skip-gram).

The model of choice was downloaded from RusVectores, a publicly available source of distributional semantic tools for Russian (Kutuzov and Kuzmenko 2017). It has the following characteristics:

- algorithm: CBOW,
- trained on: Russian National Corpus,
- vocabulary: 270 000 000 tokens, 189 183 unique,
- tagset: Universal Tags,
- min frequency: 5,
- vector size: 300.

4.1.2 Final dataset description

After processing annotated directions with morphology and semantics, the final dataset was compiled. It consisted of the following features:

1. direction text, lemmatized, processed through NER, and converted to an entity compatible with word2vec model of choice;

   Lemmatization is the process which converts a given word to its normal form — singular and nominative for nouns, infinitive form for verbs, etc.

   Word2vec model chosen for this work accepts the input values formatted as "lemma_UD part of speech", so the direction text was converted to a list of such formatted items.

2. labels for all direction types — 1 if the type was assigned to the direction, 0 if not;

3. total counts for UD parts of speech in a given direction.

4.2 Problem statement

It is impossible to solve a task without knowing the given values, the methods that may be used, and the goal. Therefore, a clear statement of the problem has to be given.

A dataset consisting of semantic vectors and part of speech amounts is available. Using this dataset against labels assigned for each direction type, a model that predicts labels for those types is required. Each direction type is taken and concerned separately — therefore one problem is split into six binary classification problems.

Possible approaches to the problem are discussed in the next subsection.

4.3 Possible approaches

4.3.1 Rule-based

Among all TEI types, there are two which are the most consistent and predictable. These types are *entrance* and *exit*. A typical *entrance* direction would have *входить* 'to enter' or its synonyms; *exit*, in a similar manner, is expected to contain the word *уходить* 'to leave' or its synonyms. This is the main principle behind the rule-based model for checking a stage direction for *entrance* and *exit*.

The set of rules and actions is the following:

1. Drop every lemma that is not a verb.

   In such types of directions, it is only verbs that matter, so all words which belong to other parts of speech may be dropped.

2. Check whether there is any top-10 verb from *entrance* or *exit*.

   If there is, assign a corresponding type.

   If not, proceed.

3. Compute a word2vec vector for the direction.

4. Compute cosine similarity of the "direction vector — *входить* vector" and "direction vector — *уходить* vector" pairs. The winning type assigned to the direction is the one that is closer to 1 (synonyms) vector-wise.

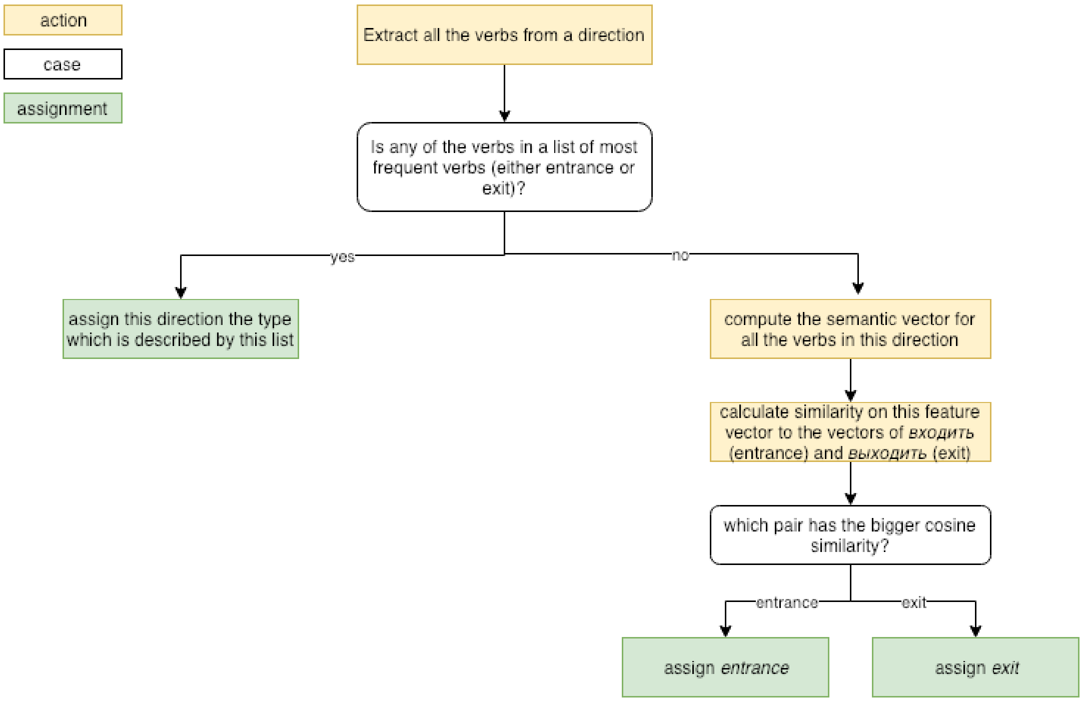
action
case
assignment
Extract all the verbs from a direction
Is any of the verbs in a list of most frequent verbs (either entrance or exit)?
yes
no
assign this direction the type which is described by this list
compute the semantic vector for all the verbs in this direction
calculate similarity on this feature vector to the vectors of *входить* (entrance) and *выходить* (exit)
which pair has the bigger cosine similarity?
entrance
exit
assign *entrance*
assign *exit*

Figure 3: Rule-based model visualization

4.3.2 Logistic regression

One of the algorithms that come to mind when thinking of a binary classification task is logistic regression. In fact, this is a linear regression with certain enhancements. A linear regression model tries to find such a linear space that separates feature representations of different label entries. Linear regression decision function is:

$$p = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_n \cdot x_n$$

Logistic regression model adds a logit function of the linear regression decision function:

$$p = \frac{exp(b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_n \cdot x_n)}{1 + exp(b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_n \cdot x_n)}$$

The result of a logit function is always in the interval $[0; 1]$, which allows to use it for binary classification. Given that each direction type gets its own classifier, it makes sense to try logistic regression on all types.

Using (or, at least, trying) logistic regression is also motivated by its renowned successfulness when addressing text classification problems. Combined with semantic data, it becomes a powerful instrument. Sometimes it is enhanced by other approximations — for instance, probabilistic Bayesian algorithms in (Genkin, Lewis, and Madigan 2007). When logistic regression is not a primary model, it may still be used for setting a baseline or for comparing logistic regression and other solutions (Zhang and Oles 2001).

4.3.3 Other algorithms

Logistic regression is not, of course, the only approach to the text classification problem. For this work, I have also chosen random forest (as an "upgrade" of decision tree) and Support Vector Machine.

The choice is based on the belief that stage direction types are distinct from one another, consequently having a distinct combination of features that may separate directions themselves. Decision tree, being an algorithm which is based on the conditions that divide features and assign labels, might show a good performance. Random forest is an enhancement of decision trees: a forest consists of multiple trees that have weights assigned to them. Each tree "makes a decision" and assigns a label; the labels are multiplied by the weights of the trees that generated the labels, and the result of this voting is passed as an overall model prediction.

Support Vector Machine (SVM) approaches the problem differently. It does not use any probabilities (as logistic regression) or feature selection (like random forest). A typical SVM classifier maps the object feature representations and labels the classes. Then it tries to determine a gap that would be able to separate the maximum amount of objects possible. After having established a gap, an SVM classifier starts to broaden it as much as possible. When the model is shown a new object, the classifier maps in the feature space and assigns the object class by recognizing which part of the feature space this object belongs to.

For *entrance* and *exit* types, a rule-based classifier is also taken into consideration. There might have been an opportunity to build such classifiers for all types. However, entities of the resting types are not as consistent and similar to each other as are those of *entrance/exit*.

Overall, three different algorithms were used for each direction type: a probabilistic linear one (logistic regression), a distinguishing one (random forest), and a mapping one (SVM). *Entrance* and *exit* type classification also involved a rule-based classifier. This allows to approach the same data from different perspectives, thus creating highly specialized models. According to the No Free Lunch theorem, one of these models might turn out to be better at classifying that certain type of directions.

4.3.4 Neural networks and their perspectives

More and more problems are now solved with the help of neural networks; in fact, many state-of-the-art results were outperformed by neural networks. There is numerous research on the usage of convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to classify texts.

Sometimes the two are combined: (Lee and Dernoncourt 2016) built a model for short text classification which involved previous short text. They first used an RNN on a preceding short text, then the result was passed into a two-layer CNN which returned the probability of a given type. When speaking about RNNs, a subtype called Long Term Short Memory (LSTM) is popular for categorizing: as an example, (Pushp and Srivastava 2017) assembled a big corpus, converted its tokens into word embeddings with the help of Google News Embedding, then fitted an LSTM so it would be able to predict the classes even in the case it meets new, unseen words.

However, fitting a neural network is not possible without big amounts of data. Given that Russian Drama Corpus has approximately 33,000 stage directions overall, developing a neural network model seems senseless — if such a model was to be designed, the annotators would have to annotate all currently present directions just in order to form the train and test sets.

5 Results

From the very beginning, the hypothesis was that some classes should be easy to classify. On the other hand, it would be unrealistic to believe that each class will be detected easily. Due to dataset imbalance and possible lack of features, some types will inevitably be a bigger challenge.

5.1 Metrics

All the models were evaluated with F1 metric. Generally, the two most popular metrics for classification tasks are F1 and ROC-AUC (area under the receiving operating characteristic curve).

F1 score is based on two other metrics: precision and recall. The former one indicates how many objects are assigned the correct label among all the objects which received that label. The latter shows how many documents were assigned the correct label among all objects that actually have that label. F1 score balances the two metrics out and is computed as follows:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

ROC-AUC is another popular metric. It is based on receiving operating characteristic curve that shows how well a binary classifier performs. On the horizontal axis is the false positives rate (documents that received a label whereas they should not have), on vertical — true positives (correctly labelled documents). Mapping a random guess model on such a plot would result in a straight line from point $(0;0)$ to $(1;1)$, therefore the ROC-AUC score is 0.5 — that straight line separates the plot area in equal shares. If a model performs better than a random guess, it is supposed to be plotted as an arched curve, consequently, ROC-AUC score is greater than 0.5. The closer the score is to 1, the better the model is; any value greater than 1 is impossible.

The main difference between the metrics is the dataset balance. ROC-AUC displays accurate data when the rates increase slowly and there is no drastic difference between objects. However, the present dataset is quite imbalanced (see chapter

3.3.2), hence ROC-AUC metric would not give an adequate and interpretable result. F1 score was chosen due to the fact that it does not rely on dataset balance.

5.2 Best models and their performance

Type	Amount / Share	Best model	Test score
business	3924 / 54,88%	Random Forest	0,905702
delivery	1412 / 19,75%	Random Forest	0,732673
entrance	940 / 13,15%	SVC	0,725389
exit	569 / 7,96%	LogReg	0,725552
location	210 / 2,94%	LogReg	0,272727
setting	95 / 1,33%	SVC	0,642857

Table 2: Best models for each type and their results

As can be seen from the table, different models worked best for different types. This points out that machine learning does not provide a single tool applicable (and well-performing) for all types at once, thus proving the No Free Lunch theorem.

Location type looks like an obvious fail: even the best model only scored 0,(27). This might be explained by the fact that there had only been 210 examples of such directions, which is less than 3%. Such a small amount of data could have been insufficient for fitting any model.

Another conclusion is the trend of model quality decreasing with the share of directions in the dataset. The only exception is the *setting* type, which is relatively easy to distinguish taking into consideration the length of an average *setting* direction — it may take several lines or even consist of several paragraphs. Other directions are much less in size, so even though settings are the least represented type in the dataset they were identified accurately enough.

It is interesting that each model showed the best results on two types. Random forest was the most effective on *business* and *delivery* types. Such performance may be caused by a large amount of these directions. Random forests had enough data to process through the decision trees it consists of, therefore it is able to show good results on test data. This might be the only case when dataset imbalance turned out to be the advantage rather than a drawback.

Support vector classifier turned out to be the most productive on *entrance* and *setting* types. These types are easily distinguishable from the others: the former due to a limited vocabulary, the latter because of the average direction size. This is exactly the type of difference that is interpretable for any support vector model. It notices that difference between true positives and everything else and establishes a (relatively) correct boundary. It is interesting that SVC was almost as good as logistic regression with *exit* type — the difference between models quality is just 0,000164 (see Table 8, Appendix 3). In fact, logistic regression model may be replaced with a support vector classifier with no significant loss in quality.

Both *entrance* and *exit* types were classified with approximately the same quality of approx. 0,755; however, different algorithms achieved that result. Support vector classifier for *entrance* was just described. *Exit* is predicted best by logistic regression. Turning back to examples for each type, it is possible to notice that with the scope of time exits started to be combined with other stage business. Consider the following examples:

(9) Еремеевна отходит.
[D. Fonvizin, *Nedorosl* (1781)]

(10) Уходит и рассуждает сам с собою.
[M. Lermontov, *Maskarad* (1836)]

(11) Слышно: «Прощайте! Будьте здоровы!» Слышен веселый смех Тузенбаха. Все уходят. Анфиса и горничная убирают со стола, тушат огни. Слышно, как поет нянька. Андрей в пальто и шляпе и Чебутыкин тихо входят.
[A. Chekhov, *Tri sestry* (1901)]

It is clearly seen that an *exit* action starts to mix with other ones. A logistic regression model predicts the probability of an exit in certain actions; an assumption may be made that the contexts for leaving might be relatively similar.

At Chapter 4.3.1., rule-based models were introduced. However, they are not present in the table of the best models. In fact, they had the lowest scores — 0,370744 for *entrance* and 0,284533 for *exit* (scores for all models on all types are in Appendix 2). The most obvious explanation would be that the models used the wrong words. Rule-based algorithm dropped everything except for verbs, though it should have probably concerned other types of speech.

6 Conclusions

6.1 Classification improvement

The previous version of the classifier tool (Maximova, Fischer, and Skorinkin 2018) was working on approx. 1500 annotated stage directions, each assigned exactly one type; used TF-IDF for direction representation; the problem was stated as a multi-class classification task. Even though the cross-validation results were promising, that solution had an F1 score of 0,2 when it was tried on the test set.

Models developed in this work have an average quality of 0,667 on the test set. An average score does not fully represent the overall state due to the fact that it includes both *business* with an F1 score of 0,9 and *location* with that equal to 0,(27). Median of the values — 0,725 — is a better description. Aside from the two outlying cases, classifiers for other types show quality close to the one represented by the median.

None of this would have been possible without a specified annotation guide. Developed in accordance with the data, it made the annotation consistent and reproducible. Apart from that, the guide might be used and/or adapted for annotating other corpora; in this case, several corpora may be analyzed and compared based on the fact that they had the same algorithm or procedure for annotating and assigning values.

As described in chapter 2.1, stage directions have already been researched. The impact of this work is a quantitative approach used to the material that was thoroughly studied and “close-read”.

Finally, this paper adds new angles to describing Russian drama through the prism of stage directions. Exploring a well-known material with the help of methods that were never used on that material gives a new perspective which may — probably, even should — lead to further discoveries.

6.2 TEI standard and Russian stage directions

In chapter 3.3 I have analyzed annotated directions through a quantitative rather than qualitative prism. The result was quite expected: provided types did not entirely match the material of Russian drama. This was clearly seen with the so-called *presence* directions which are special for Russian material; other dramas tend to indicate entrances and exits rather than the characters being on stage at the given moment.

TEI is an international standard, which means that it is meant to cover as many common features as possible. However, any international standard starts with some materials of a certain culture (in this case — dramatic and literary tradition). As already mentioned, direction type *modifier* is not present in Russian drama but is common in Shakespeare plays. That may result in a hypothesis that the standard was made on the basis of English drama.

Each literary tradition is different from each other; it is especially seen in Europe, where neighbouring countries may have absolutely different approaches to the same genre. For instance, the same genre of novel in France evolved into a philosophical novel, whereas German culture turned it into a growing-up novel (German: *Bildungsroman*). Nevertheless, the standard, especially if it claims to be international, should cover both genres as much as possible. This rather brings to an idea that the standard should be quite small — so that it could identify the general parts — and the genre- or culture-specific traits should have a tagset of their own. On the other hand, a small tagset is not as informative as a larger one. This is the point where a border between "specific" and "general" should be drawn — even though there will always be contradictions, rejections, and discussion.

So, if the suggested types are not enough, why not invent some special types and use them? This seems to be a compromise between sticking to a specific standard and not losing cultural features. It appears that this approach was also chosen by the TEI community which stated that the set of *type* attribute values is neither closed nor proposed for the time being. Having taken all that into concern, closing the list of values is out of the question as counterproductive and unconstructive.

In summary, this work attentively analyzed a specific short text type, its limitations and characteristic properties. After that, several approaches to classify those texts according to the predetermined types were undertaken, with a different success rate. The results and the types gave ground to new assumptions both about the directions and the standard, thus stating new questions and opening new horizons.

References

Apté, Chidanand, Fred Damerau, and Sholom M. Weiss (July 1, 1994). "Automated Learning of Decision Rules for Text Categorization". In: *ACM Transactions on Information Systems* 12.3, pp. 233–251. ISSN: 10468188. DOI: 10.1145/183422.183423. URL: http://portal.acm.org/citation.cfm?doid=183422.183423 (visited on 12/27/2018).

Carlson, Matvin (1991). "The Status of Stage Directions". In: *Studies in Literary Imagination* 24.2, pp. 37–48.

Detken, Anke (2009). *Im Nebenraum des Textes: Regiebemerkungen in Dramen des 18. Jahrhunderts.* Vol. 54. Walter de Gruyter. ISBN: 3-11-023003-8.

Dompeyre, Simone (1992). "Étude des fonctions et du fonctionnement des didascalies". In: *Pratiques* 74.1, pp. 77–104. ISSN: 0338-2389. DOI: 10.3406/prati.1992.1665. URL: https://www.persee.fr/doc/prati_0338-2389_1992_num_74_1_1665 (visited on 05/16/2019).

Genkin, Alexander, David D Lewis, and David Madigan (Aug. 2007). "Large-Scale Bayesian Logistic Regression for Text Categorization". In: *Technometrics* 49.3, pp. 291–304. ISSN: 0040-1706, 1537-2723. DOI: 10.1198/004017007000000245. URL: http://www.tandfonline.com/doi/abs/10.1198/004017007000000245 (visited on 05/29/2019).

Issacharoff, Michael (May 1981). "Texte théâtral et didascaleture". In: *MLN* 96 (French Issue), pp. 809–823.

Joachims, Thorsten (1996). *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Tech. rep. Carnegie-Mellon University (Pittsburgh, PA), Dept. of Computer Science.

Kreider, P. V. (Oct. 1934). "The Mechanics of Disguise in Shakespeare's Plays". In: *The Shakespeare Association Bulletin* 9.4. Publisher: Oxford University Press, pp. 167–180. URL: https://www.jstor.org/stable/23675558.

Kutuzov, Andrey and Elizaveta Kuzmenko (2017). "WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models". In: *Communications in Com-*

puter and Information Science. Analysis of Images, Social Networks and Texts (AIST). Ed. by D Ignatov. Vol. 661. Springer.

Lee, Ji Young and Franck Dernoncourt (Mar. 11, 2016). "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks". In: *arXiv:1603.03827 [cs, stat]*. arXiv: `1603.03827`. URL: `http://arxiv.org/abs/1603.03827` (visited on 12/27/2018).

Maximova, Daria, Frank Fischer, and Daniil Skorinkin (Dec. 8, 2018). "A Quantitative Study of Stage Directions in Russian Drama". In: EADH 2018: Data in Digital Humanities. Galway, Ireland.

Moretti, Franco (Jan. 1, 2000). "Conjectures on World Literature". In: *New Left Review* 1.1, pp. 54–68.

— (2011). "Network Theory, Plot Analysis". In: *New Left Review* 68.

Moschitti, Alessandro and Roberto Basili (2004). "Complex Linguistic Features for Text Classification: A Comprehensive Study". In: *Advances in Information Retrieval*. Ed. by Sharon McDonald and John Tait. Red. by Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen. Vol. 2997. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 181–196. ISBN: 978-3-540-21382-6 978-3-540-24752-4. DOI: `10.1007/978-3-540-24752-4_14`. URL: `http://link.springer.com/10.1007/978-3-540-24752-4_14` (visited on 12/27/2018).

Pushp, Pushpankar Kumar and Muktabh Mayank Srivastava (Dec. 16, 2017). "Train Once, Test Anywhere: Zero-Shot Learning for Text Classification". In: *arXiv:1712.05972 [cs]*. arXiv: `1712.05972`. URL: `http://arxiv.org/abs/1712.05972` (visited on 12/27/2018).

Rasmussen, Eric (2003). "Afterword". In: *Stage Directions in Hamlet. New essays and new directions*. Ed. by L. Aasand Hardin. Fairleigh Dickinson Univ Press, p. 226.

Segalovich, Ilya (Jan. 2003). "A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine". In: *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications*. MLMTA. Las Vegas, USA.

Sperantov, V. V. (1998). “The Poetics of a Stage Direction in Russian Tragedies of the 18th–beginning of the 19th Centuries (Towards the Typology of Literary Genres) / Поэтика ремарки в русской трагедии XVIII-начала XIX века (К типологии литературных направлений)”. In: *Philologica* 5.11, pp. 9–48.

Sun, Aixin (2012). “Short Text Classification Using Very Few Words”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’12*. the 35th international ACM SIGIR conference. Portland, Oregon, USA: ACM Press, p. 1145. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348511. URL: http://dl.acm.org/citation.cfm?doid=2348283.2348511 (visited on 12/27/2018).

TEI Consortium (Jan. 29, 2019a). *TEI element <move>*. P5: Guidelines for Electronic Text Encoding and Interchange, version 3.5.0. URL: https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-move.html (visited on 03/22/2019).

— (Jan. 29, 2019b). *TEI element <stage>*. P5: Guidelines for Electronic Text Encoding and Interchange, version 3.5.0. URL: https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-stage.html (visited on 03/22/2019).

Titomanlio, Carlo (Oct. 14, 2011). “Dalla parola all’azione. Forme della didascalia drammaturgica negli anni Venti del Novecento italiano”. PhD thesis. Università di Pisa.

Usovski-Rosa, Irina (2010). “Die szenischen Gestaltungsmittel in Čechovs Dramenwerk”. Inaugural Dissertation. Cologne: Universität zu Köln. 438 pp.

Wolpert, David H. (Oct. 1996). “The Lack of A Priori Distinctions Between Learning Algorithms”. In: *Neural Computation* 8.7, pp. 1341–1390. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1996.8.7.1341. URL: http://www.mitpressjournals.org/doi/10.1162/neco.1996.8.7.1341 (visited on 05/29/2019).

Yang, Lili et al. (2013). “Combining Lexical and Semantic Features for Short Text Classification”. In: *Procedia Computer Science* 22, pp. 78–86. ISSN: 18770509. DOI: 10.1016/j.procs.2013.09.083. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877050913008764 (visited on 12/27/2018).

Zhang, Tong and Frank J Oles (2001). “Text categorization based on regularized linear classification methods”. In: *Information retrieval* 4.1, pp. 5–31.

Appendix 1: Links

Code for the project and all data is available at the GitHub repository: creaciond/russian-stage-classification

Russian Drama Corpus is available at: https://dracor.org/rus

Appendix 2: Annotation guide

This is the guide that was used to assign direction types correctly.

A2.1 General ideas

TEI documentation page: can be found here.

Important points:

- If you are unsure about the direction, mark it as *unknown*.
- Several types might be present within the very same direction.
- Types are limited to the list: entrance, exit, business, delivery, modifier, location, setting + unknown. No other types should be used.
- Trailers — занавес, конец — are not to be considered directions. Mark them as *unknown*.

A2.2 Direction types with examples

Entrance *TEI documentation:* describes an entrance.

Characters entering the stage. For the time being, including the line-up of already present characters (*те же и...*). Realistically, this should be "presence", in opposition to "entrance".

- Чацкий, Репетилов (вбегает с крыльца, при самом входе падает со всех ног и поспешно оправляется).
  (Griboyedov, *Gore ot uma*)
- Входит Лариса, за ней человек с бутылкой шампанского в руках и стаканами на подносе.
  (Griboyedov, *Gore ot uma*)
- Входят Огудалова и Лариса слева.
  (Ostrovsky, *Bespridannitsa*)

Exit *TEI documentation:* describes an exit.

- Уезжает.

  (Griboyedov, *Gore ot uma*)

- Лиза свечку роняет с испугу; Молчалин скрывается к себе в комнату.

  (Griboyedov, *Gore ot uma*)

- Вожеватов и Кнуров уходят.Лариса подходит к Карандышеву.

  (Ostrovsky, *Bespridannitsa*)

Business *TEI documentation:* describes stage business.

An action undertaken by a character. Might be connected to a verb.

- Осматривается.

  (Griboyedov, *Gore ot uma*)

- глядит в дверь налево

  (Ostrovsky, *Bespridannitsa*)

- припадает к ее руке

  (Chekhov, *Djadja Vanja*)

Delivery *TEI documentation:* describes how a character speaks.

May be a single adverb describing the manner of speaking, or a single noun in Dative (who does the character address?). Concerns a communicative act (voice/silence etc.) — how something is delivered in a communicative way.

- со вздохом

  (Griboyedov, *Gore ot uma*)

- Тарелкин (сконфуженный). Скажи ему, что некогда.... занят.

  (Sukhovo-Kobylin, *Delo*)

Modifier *TEI documentation:* gives some detail about a character.

Usually, result of other character's actions. Might also be changes of appearance (disguise, dressing as another character).

Location *TEI documentation:* describes a location.

Usually present in later plays.

How to distinguish from type *setting*: *location* pinpoints a place at the scene where the action takes place; *setting* is more general.

- еще в дверях
  (Gogol, *Revizor*)

- в зале у стола, сердито
  (Chekhov, *Tri sestry*)

Setting *TEI documentation:* describes a setting.

Such a direction would generally appear at the beginning of a scene or an act; in some cases, it just names the characters present on a stage — nevertheless, it's still a setting. It doesn't usually include a verb (an active one), but may contain some "static" verbs (those of standing, lying, etc.) Might also include indicators: sounds, weather, ...

- У Фамусова в доме парадные сени; большая лестница из второго жилья, к которой примыкают многие побочные из антресолей; внизу справа (от действующих лиц) выход на крыльцо и швейцарская ложа; слева, на одном же плане, комната Молчалина. Ночь. Слабое освещение. Лакеи иные суетятся, иные спят в ожидании господ своих.
  (Griboyedov, *Gore ot uma*)

- В доме Прозоровых. Гостиная с колоннами, за которыми виден большой зал. Полдень; на дворе солнечно, весело. В зале накрывают стол для завтрака.
  (Chekhov, *Tri sestry*)

- За сценой цыгане запевают песню.
  (Ostrovsky, *Bespridannitsa*)

Appendix 3: Model performance on direction types

Model	**Cross-validation**	**Validation**	**Test**
LogReg	0,891841	0,906188	0,884453
Random Forest	0,90022	0,96	0,905702
SVC	0,526805	1	0,575758
average	*0,772955*	*0,955396*	*0,788638*

Table 3: Overall performance quality for type *business*

Model	**Cross-validation**	**Validation**	**Test**
LogReg	0,732372	0,775	0,72158
Random Forest	0,707697	0,96	0,905702
SVC	0,721186	0,874419	0,72042
average	*0,720418*	*0,845225*	*0,724891*

Table 4: Overall performance quality for type *delivery*

Model	**Cross-validation**	**Validation**	**Test**
LogReg	0,342365	0,470588	0,272727
Random Forest	0,352769	0,825397	0,232558
SVC	0,372803	0,911765	0,25
average	*0,355979*	*0,735917*	*0,251762*

Table 5: Overall performance quality for type *location*

Model	**Cross-validation**	**Validation**	**Test**
LogReg	0,536318	1	0,6
Random Forest	0,29141	1	0,45
SVC	0,613119	1	0,642857
average	*0,413864*	*1*	*0,525*

Table 6: Overall performance quality for type setting

Model	**Cross-validation**	**Validation**	**Test**
LogReg	0,67583	0,834008	0,70814
Random Forest	0,64848	0,92607	0,689139
Rule-based	0,320973	0,380682	0,370744
SVC	0,700364	0,878661	0,725389
average	*0,484727*	*0,653376*	*0,529942*

Table 7: Overall performance quality for type *entrance*

Model	**Cross-validation**	**Validation**	**Test**
LogReg	0,707451	0,753846	0,725552
Random Forest	0,717889	0,966887	0,710526
Rule-based	0,309677	0,311512	0,284533
SVC	0,700364	0,878661	0,725388
average	*0,513783*	*0,6392*	*0,49753*

Table 8: Overall performance quality for type exit