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»

КУРСОВАЯ РАБОТА

Лингвистический анализ авторских ремарок в русских драматических текстах XVIII–XX вв.

Linguistic Analysis of Stage Directions in Russian Drama from the 18th to the 20th Century

Студентка 3 курса
группы БКЛ151
Максимова Дарья Максимовна

Научный руководитель
Фишер Франк
доцент Школы лингвистики

Консультант
Скоринкин Даниил Андреевич
преподаватель Школы лингвистики

Москва, 2018 г.

Contents

“Stage directions, quite literally, don’t count."

— Eric Rasmussen [Rasmussen, 2003]

1 Introduction

A **stage direction** is an instruction in the text of a play which helps actors and all the cast to interpret the text correctly. It may control the flow of the play, the way actors speak, the lightning, or decorations. In this work, stage directions in Russian dramas are analyzed.

Even though stage directions are often treated as functional text (see [Rasmussen 2003], [Carlson, 1991], and other research), they may also be a representative feature of the text in general, which means that the researchers may use less resources to analyze the play and its text. They also might be useful when exploring the play, for instance, in the “distant reading” cases [Moretti, 2013], when perceiving all the text is not necessary (or even counterproductive) for understanding the trends of the play.

Analysis of stage directions may also shed some light on drama evolution in Russian literature.

2 Corpus at hand

All the research is conducted on RusDraCor, a corpus of Russian drama. It comprises plays from the 18th to the 20th century, written by various Russian authors. The corpus comprises 90 plays, dating from 1747 (*Horev* by Alexander Sumarokov) to 1943 (*Poslednie dni (Pushkin)* by Mikhail Bulgakov).

The corpus is encoded in XML technical standard called TEI-P5. This is a primarily semantic standard which pays special attention to the written text and its peculiarities. It is also the most well-known and maintained standard for historic documents at the moment. As for drama, TEI-P5 has everything set and there is no need to invent any new tags or attributes: everything is well-documented and regularly revised (see [TEI Consortium, 2018b]).

Study group which created the corpus focused on extra-linguistic annotations and social relations, whereas this work pays special attention to the text and to its annotation, also trying to enhance it.

The corpus is available online at https://dracor.org/rus.

3 Former works on stage directions

This work deals particularly with stage directions. As mentioned earlier, stage directions are small parts of the text, which tell actors what to do, how to speak, etc.

Stage directions combine the stage business (their technical purpose) with symbolic significance and author's own style [Munkelt, 1987]. In this way, they are a very special part of the play. Any irregularities, seeming errors, or misunderstandings while acting or reading the play may turn out to be meaningful, as they may reveal some important information about the author, their position, or intentions.

There were different approaches to stage directions. Dessen pointed out that the thesaurus of words referring to the so-called body of stage directions is quite limited (see [Dessen, 2001]), so there was an attempt to create a dictionary of words used in Shakespeare's plays as stage directions. The resulting dictionary [Dessen and Thomson, 1999] covers 500 plays dated from 1590 to 1642. It covers more than 22 000 stage directions that result in approximately 900 terms with explanations.

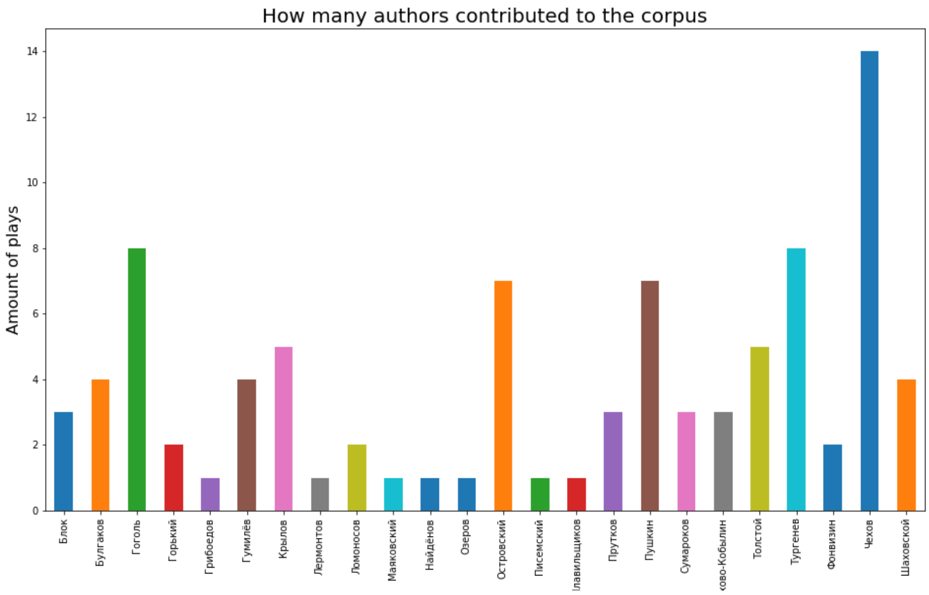
Nevertheless, stage directions are often interpreted as plain instructions, therefore dropped out of peculiar analysis [Carlson, 1991]. For the majority of literary researchers, stage directions seem to be "rather stark and unadorned simplifications".

4 Describing the corpus and general trends

4.1 Corpus in general

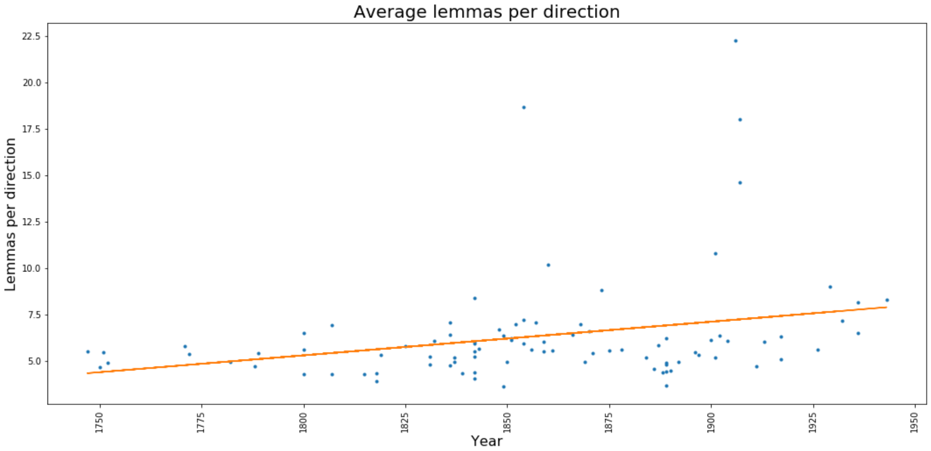
It may be rather reasonable to take a look at the corpus in general and at the general trends.

Figure 1: Authors' contribution to the corpus



From the bar plot, it is clearly visible that the biggest part of the corpus (as many as 14 plays) are the plays written by Anton Chekhov (1860–1904), probably the most famous and well-known Russian playwright. There also is a reasonably big share of plays by Nikolay Gogol (1809–1952), Ivan Turgenev (1818–1883), Alexander Ostrovsky (1823–1866) and Alexander Pushkin (1799–1837). All these playwrights lived in the 19th century, which is believed to be the so-called "golden age" of Russian literature, therefore, its share in the corpus should be the biggest.

Figure 2: Average lemma usage in different plays



It is quite clear from the figure that in the course of time stage directions grew in size — from less than 5 lemmas per direction in the middle of the 18th century to approximately 8 in the middle of the 20th century.

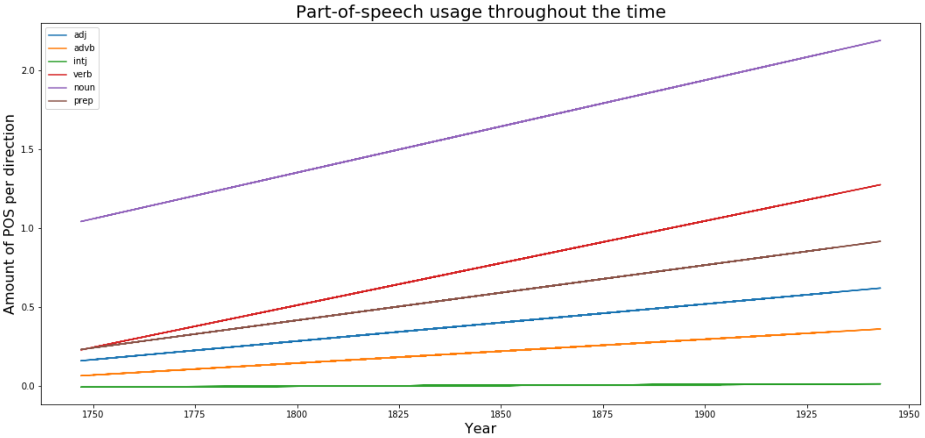
4.2 Trends

Apart from lemma count, shares of some parts of speech were calculated. They are:

- nouns,
- adjectives,
- adverbs,
- prepositions,
- verbs,
- interjections.

All the morphological analysis and tokenization was performed with help of pymorphy2, an open-source morphological analyzer of Russian for Python [Korobov, 2015].

Figure 3: Part-of-speech usage throughout the time



This figure probably is the most informative of all, as it shows the global trend of part-of-speech usage throughout the time. Horizontal axis maps the years, and the vertical shows how many words of the following parts of speech are used per direction. The legend is located in the upper-left corner.

It is apparent that with the time an average direction grows in length, as all the parts-of-speech graphs are growing. It is also easily detectable that noun usage more than doubled. It could possibly mean that authors made their descriptions (that means, directions) richer. In general, nouns are used to describe the setting, so one may assume that the settings became more complex.

This also correlates with the fact that the usage of prepositions and adjectives also appears to have doubled. This fact can be explained by the notion that these parts of speech tend to appear together with the nouns: prepositions combine with nouns to form a prepositional group [Testelets, 2001], and adjectives are prototypically used to characterize the objects, which are represented by nouns. These facts correspond flawlessly with the fact that the lines representing the three parts of speech discussed (nouns, adjectives, and prepositions) behave in the same manner.

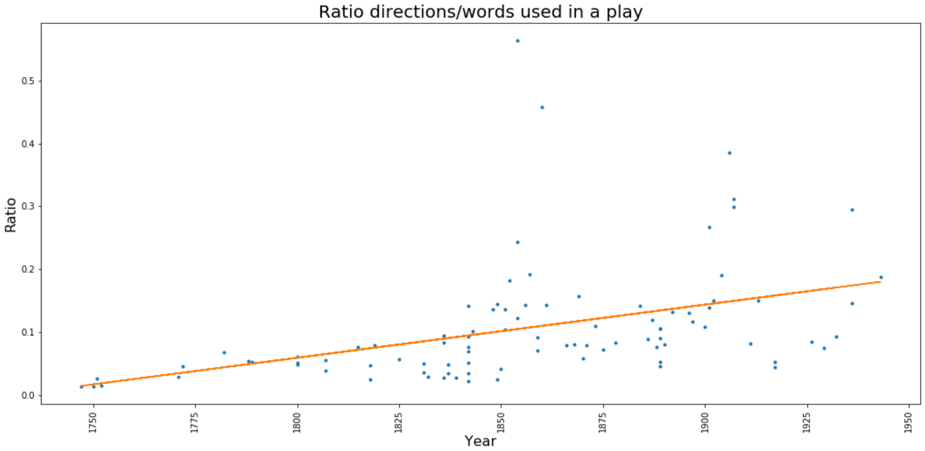
Another curious tendency is the growth of verb usage. It is the part of speech which has grown the most of all. Starting at approximately 0,25 verbs per direction (which means that out of four randomly taken directions in the middle of the 18th century, a verb would appear only at one of them), the line keeps going up, and by the beginning of the 20th century, the number tripled and became approximately 0,75 verbs per direction — which means that at that moment of time verbs could appear in three directions out of four. This could possibly mean that the behaviour of actors on stage became more active; therefore, the plays themselves became more active, flowing, or even alive.

On the other hand, there is a "stable" part of speech, which does not change over time: interjections seem to be very rare, because the line representing them is almost horizontal and it is extremely close to zero. This can be explained by the fact that the interjections, by their meaning, portray some sounds, onomatopoeiae, or unintentional exclamations which are commonplace for everyday spoken language. Stage directions, on the other hand, deal with actors' actions and other stage business, so there is no need to use the interjections.

Apart from that, it is also important to mention that the lines representing these parts of speech do not intersect. This observation indicates that, even though the authors started using more words in their directions, the structures they used did not change dramatically.

In conclusion, let us also take a look at the share of the stage directions in the text of the play. In the following figure, the X axis marks the year when the play was created, and the Y axis — the ratio of words used as stage directions to the words used in the whole play.

Figure 4: The ratio of stage directions to total number of words used in a play



The observer is able to clearly see that the ratio has also grown; one may even say that it has grown drastically. From an extremely small number in the middle of the 18th century, the share of the stage directions had grown to approximately 20% by the beginning of the 20th century. It is also important to remember that the shares of all parts of speech in the directions are growing.

Overall, the resulting figures may possibly mean that stage directions are getting bigger, and we can task of *epification* of Russian drama, as the directions tend to comprise more and more various ideas. This phenomenon ("Episierungstendenz") was pointed out by Peter Szondi (see [Marx, 2016]); this is the tendency when Naturalistic drama in Germany experienced the crisis. It drifted towards epics, narratives, or lyrics. It is visible from the data that Russian drama also adopts the features from those genres, and even more — from the lyrics (the dramas of Blok are the best example).

4.3 Frequent entries

Below is the list of 10 most used stage directions in the corpus, with the number of their respective occurrences.

Table 1: Most frequent directions

Direction text	**Count**
уходит	414
пауза	326
в сторону	313
смеётся	130
помолчав	102
кричит	90
поёт	82
входит	80
встаёт	79
молчание	75

From the table, it is clear that the most frequent directions are the ones regarding either the way a character speaks (such as ‘пауза’ or ‘в сторону’) or their appearance/disappearance on stage (e.g., ‘уходит’ or ‘входит’). This, on the one hand, proves a point that stage directions are technical text which has nothing to do with author’s intentions, beliefs, or messages to the audience.

On the contrary, this is the list of the most frequent directions does not necessarily have to prove the existence of such message. More than that, we would be surprised if the list of the most common words would not look like this. The list only highlights the most noticeable feature of any dramatic text — the presence of the instructions to the actors.

This argument can also be proved by taking into account the most frequent lemmas of various parts of speech. The top 10 lemmas of each part of speech considered in this work are provided in Appendix 2, along with the figures representing the dynamics of their usage in the corpus in the scope of time.

As for the verbs, the most frequent lemmas are ‘уходить’ and ‘входить’, which correlates perfectly with the idea of perceiving directions as the instructional texts. Each of the ten most used verbs can be easily imagined in a stage direction. Apart from marking entrances and exits, there also are the verbs depicting some finite or punctive actions, such as ‘вставать’, ‘подходить’, or ‘целовать’.

In this case, the list of top ten nouns represents the field of objects or places where the actions — denoted by the verbs — take place. Characters interact with hands (‘рука’), doors (‘дверь’) or tables (‘стол’); we can also figure out that the action may take place in a room (‘комната’) or a window (‘окно’).

It is also curious that proper name *Иван* ‘Ivan’ emerged in this list; it may probably be the particular quality of the morphological parser we used, as it tends to treat people’s proper names as nouns [Korobov, 2015].

Adverbs, being the part of speech traditionally used for modifying the verb phrase, in the case of Russian drama also comply with the rule of “being wholly instructional for actors”. They primarily involve mode of speaking, such as speed (e.g. ‘быстро’), strain (‘тихо’ and its antonym ‘громко’), or quality (‘несколько’ or ‘немного’).

5 Classifying directions

It seems noticeable that stage directions are different in their purpose. Based on that purpose, one is able to create a classification of their own. An example of such a classification may be found in [Carlson, 1991]. He divides the *didascalia* — as he calls the stage directions — into the following types:

Table 2: Types of didascalia (stage directions) according to [Carlson, 1991]

Type	**Description**
didascalia of attribution	identifies the person speaking
structural didascalia	divides the play into parts, such as acts, scenes, etc.
locational didascalia	announcements of new scenes and acts and their settings
didascalia of character description	describes the character
performance didascalia	technical directions: lightning changes, sounds, the movements of objects, stage business

This classification is not the only one. This paper tries to classify the stage directions by the way TEI-P5 standard does it. In TEI, stage directions are classified in different types, which are the following: setting, entrance, exit, business, novelistic, delivery, modifier, location, and mixed (for more information and examples, see the documentation [TEI Consortium, 2018a]).

Table 3: Types of stage directions according to the TEI-P5 standard

Type	**Description**
setting	how does the stage look like: decorations, lighting, etc.
entrance	marks the entrance of an actor
exit	marks the exit of an actor
business	actors' behaviour on the scene: movements, actions, changes of performance, etc.
novelistic	narrative direction
delivery	how a character speaks
modifier	details about the character
location	describes a location (such as a geographical point, etc.)
mixed	one or more of the above

It seems significant to point out that novelistic directions tend to describe something outside the scope of the play business, such as feelings, reasons to behave in a certain manner, or broad descriptions which in Russian literary tradition often appear in philological analysis, but not in the play itself.

Consider the following example:

(1) *<stage type="novelistic">Having had enough, and embarrassed for the family.</stage>* [TEI Consortium, 2018a]

In this stage direction, we are given the character's motivation. Russian playwrights tend to use other types, which only depict character's behaviour or speech, but the motivation is left for the readers and the spectators, like in the following example from Anton Chekhov's *Djadja Vanja* (1898):

(2) *<stage>(смеясь). </stage>Браво, браво!.*

On these grounds, in this research we left novelistic type out of the classification.

5.1 Data

The goal is to classify stage directions into these types. To get the gold standard data, several plays were annotated manually. These plays are:

- "Svoi ljudi — sochtiomsia" *(It's a Family Affair-We'll Settle It Ourselves)*, written by Alexander Ostrovsky,
- "Khorev", by Alexander Sumarokov,
- "Balaganchik", by Alexander Blok,
- "Revizor", *(The Government Inspector)* by Nikolai Gogol,
- "Djadja Vanja", *(Uncle Vanja)* by Anton Chekhov.

Overall, RusDraCor has over 22 000 textual phrases marked up by tag <stage>, which is responsible for stage directions. Of these 22 thousand directions, there are only 13 855 unique. The five plays mentioned above give us around 1 150 directions to use as the gold standard (of which 858 are unique). The distribution of the types is shown below.

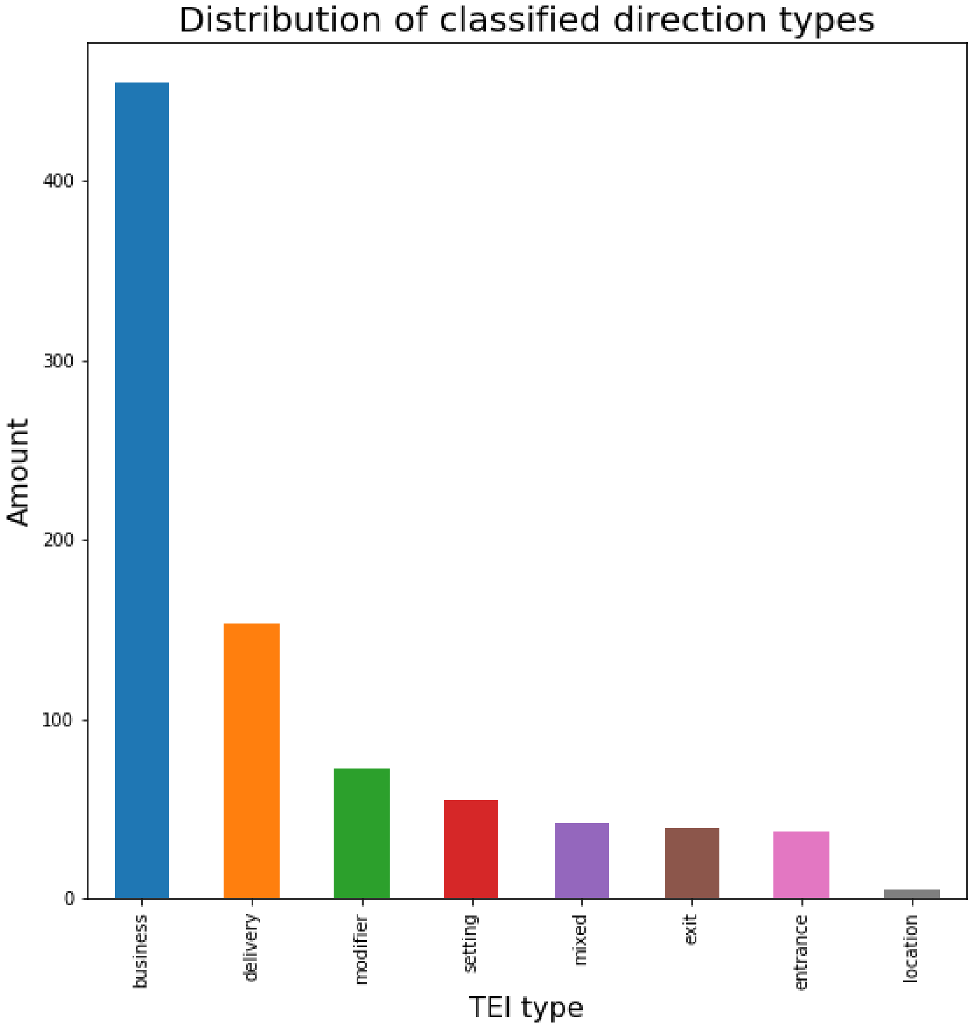


Figure 5: Distribution of stage directions in the gold standard (unique directions)

From the figure, it is clearly visible that the most varying type are the business and delivery type directions. All other types tend to be equally ‘different’ — to a certain extent. For instance, entrance and exit types seem to be equal, as the lexical fields of words which are used in these types seem to be rather closed. Therefore, these types are not topping the direction distribution, because the ways to mark stage entrances and exits are limited, which is definitely not the case for the stage business.

5.2 Task definition, algorithms, metrics

In order to complete the task, one has to address the classic machine learning **classification** problem. In this task, the model is given different objects with several values — which are known as objects' features — and the model is asked to sort these objects into N several classes. These classes are known in advance and are fixed throughout the whole experiment. The model has to predict the classes for the objects as precise as possible. In these experiments, the data is separated into train sets and test sets. The former are used for training the models, and the latter for evaluating them.

In this work, several classification algorithms were used, which are: k Nearest Neighbours (kNN), decision trees (DT), and random forest (RF).

To measure the quality of the classifiers, the F1 metric, or F1-score, was used (see below in chapter 5.2.2).

We also performed a 5-fold **cross-validation**, which means that the data was split into five parts. Five different models (using the same algorithm) were trained on four parts of the data; after training, they were tested on the remaining part. The model with the best score was saved for later comparison with other models.

Another key point was the **grid search**. For each model, there are certain parameters which can be alternated. In order to achieve the best result possible, a range of parameters was tested.

5.2.1 Algorithms

k Nearest Neighbours, or kNN, is the algorithm which relies on the data around the point. The algorithm traverses all the points, which are the objects (or, observations) in the dataset. For each point, its proximity to all the other points is calculated. It is usually performed with use of Euclidean distance, but in case of multidimensional spaces it could be different. After that, the algorithm takes into consideration k points around the one being classified at the moment — they are called the neighbours of this point. Finally, the point is assigned the class which prevails among these neighbors.

It is the easiest classification algorithm which relies on compactness hypothesis.

According to this hypothesis, the correct similarity measure will place the objects in such a way that they will be placed in the same class. On the other hand, kNN may be difficult to implement on large datasets (for instance, on those with the number of objects greater that 1000) [Friedman et al., 2001]. Furthermore, the question of metric selection is particularly different with this algorithm.

The testing grid search parameter was the number of neighbours to rely on.

Decision Tree, abbreviated as DT, is another approach, which is often used in combination with other approaches. For each class, it looks at all the examples of this class in contrast to the other classes. It creates rules for separating the data into the classes. The name of the algorithm appeared because it is human-readable: it can be represented as a tree graph, where the leaf nodes are the classes, and the root and the internal nodes are the conditions for splitting the data.

One of the greatest advantages of this approach, apart from being easy to interpret, is that decision trees may produce good results with little data and/or data which is hard for understanding. Still, they are fairly unstable, which means new data may be misinterpreted.

The most significant parameter is the depth of the tree; it was also the parameter which was grid-searched.

Random Forest (RF) is the third algorithm. It expands the idea of instability of decision trees; instead, the method suggests to use an ensemble of DTs, separated into some clusters. Each of these clusters proposes a decision which is brought to voting. Clusters' decisions are multiplied by a certain share, and the result of such voting is the output of the model. Due to the fact that decision trees are unstable and their quality may be low, having several of them in a cluster increases the quality of that cluster, hence the model.

Random forests are generally believed to be quite stable and persistent to overfitting, which is a considerable problem for machine learning engineers. They are also relatively popular on the grounds of lack of need of major data preparation. Alternatively, it is also important to mention that Random Forests are resource-consuming: one may require more time and computational resources.

The grid-searched parameter (and, obviously, the one contributing the most to the final outcome) was the number of estimators (i.e., trees) the random forest uses.

5.2.2 Metric

In order to measure the quality of the classifiers, F1-score was used.

F1 score (also F-metric, of F-score) is the metric which measures the accuracy of a given model. It takes into consideration both model's precision and recall, returning their harmonic mean.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Model's precision is the ratio of true positive objects (i.e., correctly classified objects) to all positive objects (both correctly and fallaciously appointed to the given class); its recall is the ratio of true positives to the sum of true positives and false negatives (i.e., the objects which were not assigned to this class by mistake).

5.3 Data preprocessing, model selection, and results

All the models do not take string-typed variables as object features. In order to use the texts of the directions when classifying them, we used TF-IDF (text frequency — inverse document frequency) algorithm, so the texts are represented as vectors, which are kept as sparse matrices. It is also essential to mention that the vectorizing algorithm was run on normalized directions (i.e., all the tokens were converted into their lemmas, and the punctuation marks were removed).

The directions which were manually annotated were used as both the gold standard and the training dataset. Cross-validation allowed us to calculate the F1 score on them in order to have a notion of the overall model performance.

To select the best model, all the three algorithms were tested in numerous conditions. First, the datasets only contained TF-IDF vectors. After that, the models were trained exclusively on part of speech shares calculated in advance. On the third iteration, the models could use both the vectors and the shares.

The results are presented below:

Table 4: Results of the models

Model	F1 score		
	only TF-IDF vectors	only POS shares	TF-IDF + POS shares
k Nearest Neighbours	0,6270	0,6433	0,7051
Decision Tree	0,6445	0,6422	0,7191
Random Forest	0,6667	0,6573	0,7401

It is easily observable from the table that the best result is reached by using Random Forest on all the data possible — that is, on both TF-IDF direction vectors and data on the part of speech shares. Training on either TF-IDF vectors or shares produces approximately the same results.

6 Conclusions

This work may be considered as an exhaustive corpus study of the stage directions from several angles:

- as a self-sufficing textual phrase;
- as an important part of the text;
- and also as a source for the machine learning task.

The corpus of Russian drama showed that with the course of time an average stage direction has become bigger; moreover, the shares of different parts of speech also increased gradually. This means that we are able to talk about the epification of Russian drama — a phenomenon of stage directions becoming more and more important for the text.

Apart from having conducted quantitative corpus analysis, a classifier capable of performing primary annotation was also developed.

The classification tool is able to categorize the directions into nine types according to the TEI-5 standard. This allows to enhance the markup of Russian Drama Corpus. Consequently, it may complement further analysis on the corpus.

Any neighbouring or subsequent research is now equipped with new data which may result in new discoveries.

To sum up, in traditional research stage directions might not count, as the motto of this paper suggests. On the other hand, paying attention to the stage directions allows us to keep track of Russian drama evolution without having to closely read all plays accessible. In this case, ‘distant reading’ uses a larger scale for detecting the results and dynamics which could have possibly stayed invisible when applying ‘close reading’.

6.1 Further annotation

On the grounds that annotations provided by the model are not perfect, it seems to be a reasonable idea to create a platform for annotation verification. The suggestested classifier is believed to perform with an F1-measure of 0,74, which means that out of four directions only one would be likely to be misclassified.

Nevertheless, it would be useful to have another read on these directions and double-check whether the algorithm was correct. We could also do more precise error analysis if we have the data from the annotators to compare with the data received from the models.

References

Carlson, M. (1991). The status of stage directions. *Studies in the Literary Imagination*, 24(2):37.

Dessen, A. C. (2001). The body of stage directions. *Shakespeare Studies*, 29:27.

Dessen, A. C. and Thomson, L. (1999). *A Dictionary of Stage Directions in English Drama, 1580-1642*, volume 230. Cambridge University Press Cambridge.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Korobov, M. (2015). Morphological analyzer and generator for russian and ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 320–332. Springer.

Marx, P. (2016). *Handbuch Drama: Theorie, Analyse, Geschichte*. Springer-Verlag.

Moretti, F. (2013). *Distant reading*. Verso Books.

Munkelt, M. (1987). Stage directions as part of the text. *Shakespeare Studies*, 19:253.

Rasmussen, E. (2003). Afterword. In *Stage Directions in Hamlet: New Essays and New Directions*. Fairleigh Dickinson Univ Press.

TEI Consortium (2018a). “7.2.4 stage directions.” guidelines for electronic text encoding and interchange.

TEI Consortium (2018b). Guidelines for electronic text encoding and interchange.

Testelets, Y. (2001). *Introduction to Syntax*. Russian State University for the Humanities.

Appendix 1. GitHub Repository

Code and data used in this work are available in GitHub repository at:

https://github.com/creaciond/russian-drama.

Appendix 2. Part of speech usage growth figures

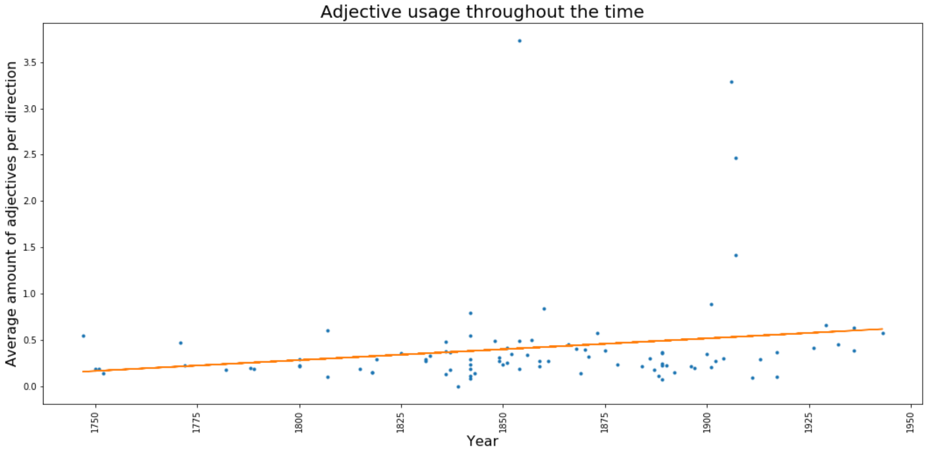


Figure 1: Adjective usage throughout the time

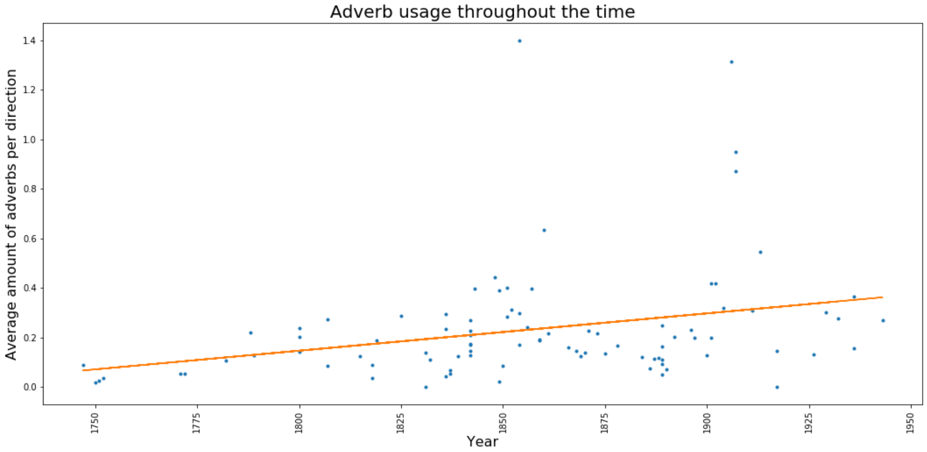


Figure 2: Adverb usage throughout the time

Table 1: Most common and frequent adverbs in the corpus

Lemma	**Count**
тихо	289
потом	240
быстро	156
несколько	134
опять	120
громко	111
вдруг	83
вслед	81
немного	74
вполголоса	71

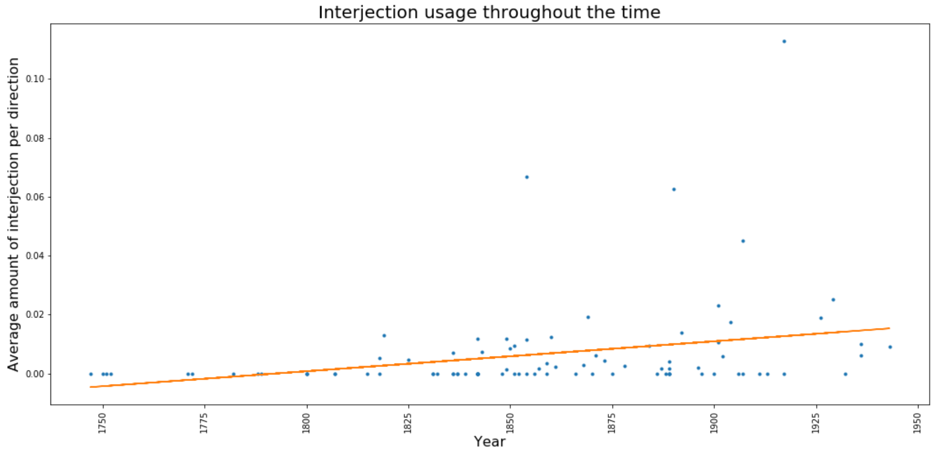


Figure 3: Interjection usage throughout the time

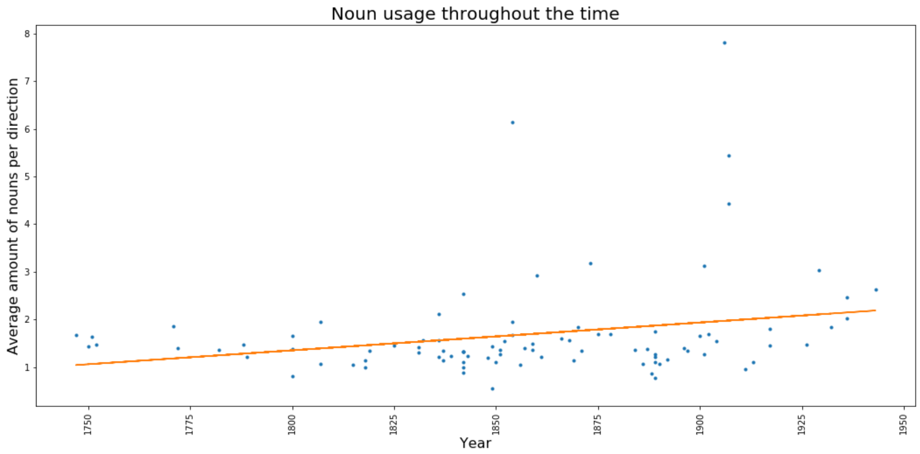


Figure 4: Noun usage throughout the time

Table 2: Most common and frequent nouns in the corpus

Lemma	**Count**
рука	1111
дверь	920
сторона	561
стол	441
пауза	389
голова	306
голос	289
комната	281
иван	281
окно	244

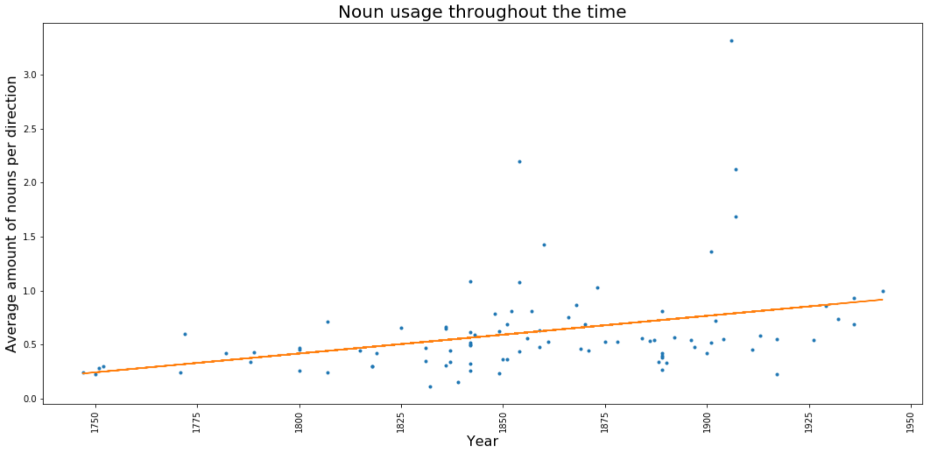


Figure 5: Preposition usage throughout the time

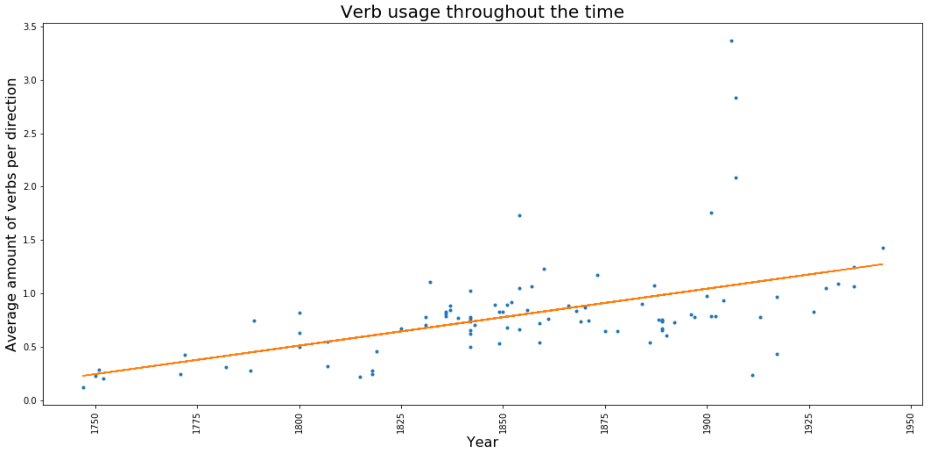


Figure 6: Verb usage throughout the time

Table 3: Most common and frequent verbs in the corpus

Lemma	**Count**
уходить	1258
входить	1024
садиться	448
идти	428
подходить	365
брать	356
выходить	274
смотреть	266
вставать	227
целовать	227