# SHORT TEXT CLASSIFICATION: A CASE OF STAGE DIRECTIONS IN RUSSIAN DRAMA

*Daria Maximova, 151*
*Supervisor: Frank Fischer, Ph.D.*

# INTRODUCTION: STAGE DIRECTIONS AND TEI

**Russian Drama Corpus:** 141 play from mid-18th to mid-20th century, TEI-P5 encoded.

Specifically developed TEI tags for theatrical text elements, e.g.:

- speech,
- technical text,
- stage directions.

# INTRODUCTION: STAGE DIRECTIONS AND TEI

**Russian Drama Corpus:** 141 play from mid-18th to mid-20th century, TEI-P5 encoded.

Specifically developed TEI tags for theatrical text elements, e.g.:

- speech,
- technical text,
- stage directions.

```
<sp who="#nazarovna">
  <speaker>Назаровна</speaker>
  <stage>(Савве).</stage>
  <p>Тебе бы, старик, таперича в тепле полежать,
  ножку-то погреть.</p>
  <stage>Пауза.</stage>
  <p>Старик! Человек божий!</p>
  <stage>(Толкает Савву.)</stage>
  <p>Ай помирать собираешься?</p>
</sp>
```

(from A. Chekhov, *Na bol'shoj doroge*)

# TEI <STAGE> TYPES

TEI Consortium: optional yet recommended `@type` attribute with following values:

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

# TEI <STAGE> TYPES

TEI Consortium: optional yet recommended `@type` attribute with following values:

| | |
|---|---|
| setting | delivery |
| entrance | modifier |
| exit | location |
| business | mixed |
| novelistic | |

**Goals:**

1. create an automated classification tool for Russian Drama Corpus
2. research the applicability of TEI types to Russian drama material

# INTRODUCTION: PREVIOUS RESEARCH

Stage directions in general

- [Detken 2009] — big qualitative work on stage directions as main subject of study (German drama)
- [Issacharoff 1981] — French directions from a syntagmatic view, with a typology of its own

Russian drama

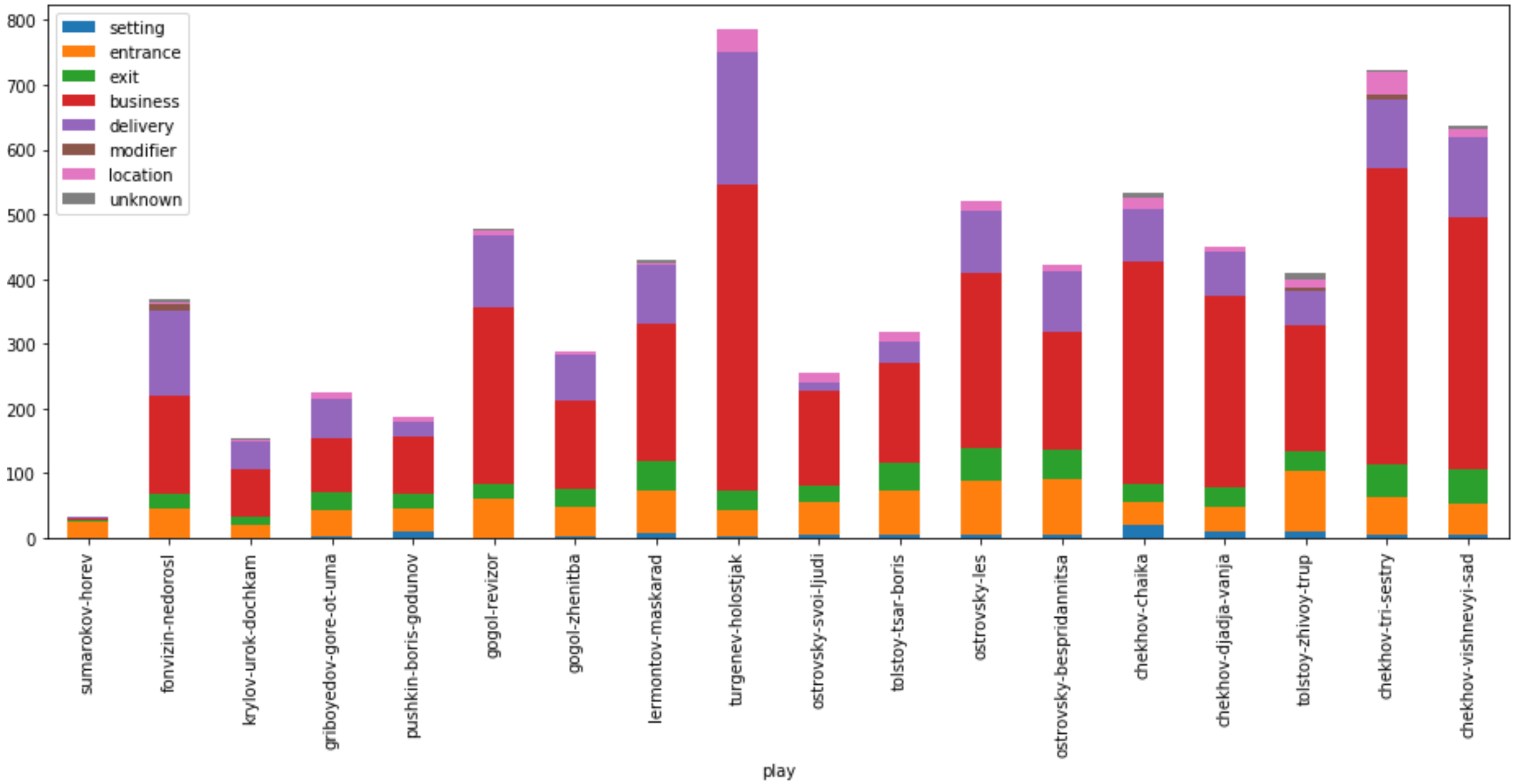
- [Sperantov 1998] — quantitative study on '*classicality*' across 70 Russian tragedies with several metrics, data annotated manually
- [Usovski-Rosa 2010] — stage directions in Chekhov's plays
- [Maximova, Fischer, and Skorinkin 2018] — linguistic analysis of Russian Drama Corpus, epification trend

# DATA ANALYSIS

# ANNOTATION TASK

- Source: 18 plays from 1747 to 1903 => 6569 stage directions (extracted with corpus API)
- Options to choose from: TEI types, excl. *mixed*
- Result: directions + their types available at GitHub repository

# TYPE DISTRIBUTION BY PLAY



## TEI TYPES AND RUSSIAN MATERIAL

Special type — **presence**, as in:

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

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

## TEI TYPES AND RUSSIAN MATERIAL

Special type — **presence**, as in:

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

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

**Modifier**: describes disguise in TEI, present mainly in Lermontov's *Maskarad*:

(3) 1-я маска входит быстро в волнении и падает на канапе.

## TEI TYPES AND RUSSIAN MATERIAL

Special type — **presence**, as in:

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

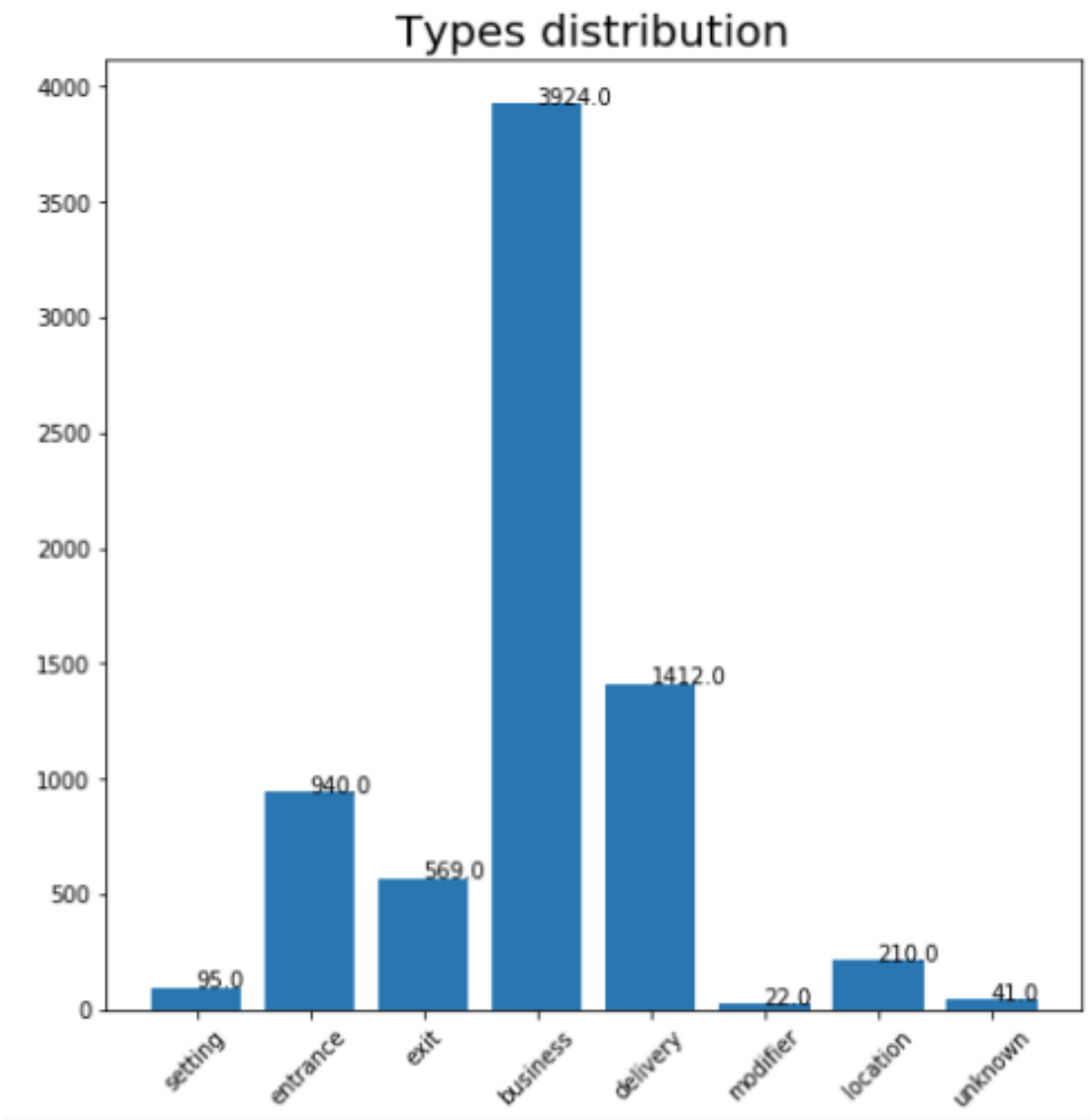
(2) Г-жа Простакова, Простаков, Скотинин.
[D. Fonvizin, *Nedorosl*]

**Modifier**: describes disguise in TEI, present mainly in Lermontov's *Maskarad*:

(3) 1-я маска входит быстро в волнении и падает на канапе.

**Novelistic:** 'a narrative, motivating stage direction'

# OVERALL TYPE DISTRIBUTION



# CLASSIFICATION

## PROBLEM STATEMENT

**Data:** 6569 annotated stage directions

**Problem:** binary (0/1) classification for each of the following classes:

*setting*, *entrance*, *exit*, *business*, *delivery*, *location*

=> **6 different classification models** (according to No Free Lunch theorem, cf. [Maximova et al., 2018])

# PREPROCESSING

1. Morphological + POS analysis, lemmatization (Mystem)
2. Named entity recognition
3. Semantic vectors (word2vec)

# FINAL DATASET DESCRIPTION

1. Lemmatized & NER-processed text
2. Direction type label: 1/0
3. POS counts in a given direction

# APPROACHES

## Non-ML

Rule-based — only for types *entrance*, *exit*

# APPROACHES

## Non-ML

Rule-based — only for types *entrance*, *exit*

## ML

Probabilistic: Logistic regression

Feature selection and combination: Random forest

Distance between feature vectors: Support vector classifier

# RESULTS

# RESULTS

Metric: F1 measure

| 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 |

Median quality: 0,725.

# RESULTS: INTERPRETATION

**Random Forest:** delivery, business

Large amount of directions => enough data for extracting features

# RESULTS: INTERPRETATION

**Random Forest:** delivery, business

Large amount of directions => enough data for extracting features

**Support vectors:** entrance, setting

Easy to distinguish (special vocabulary/more lengthy that usual) => interpretable for the model

# RESULTS: INTERPRETATION

**Random Forest:** delivery, business

Large amount of directions => enough data for extracting features

**Support vectors:** entrance, setting

Easy to distinguish (special vocabulary/more lengthy that usual) => interpretable for the model

**Logistic regression:** exit

With the scope of time, exit gets mixed with other types => probability problem

# OVERALL

**Data, code, final classification tools:** available at https://github.com/creaciond/russian-stage-classification

**TEI types:**

+ presence

– modifier
– narrative

# Thank you for your attention!

## REFERENCES


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

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

Maximova, Fischer, and Skorinkin 2018 — *A Quantitative Study of Stage Directions in Russian Drama*. Presented at EADH 2018: Data in Digital Humanities. Galway, Ireland.

Sperantov 1998 — *Поэтика ремарки в русской трагедии XVIII-начала XIX века (К типологии литературных направлений)*. In: Philologica 5.11, pp. 9–48.

TEI <stage> element documentation: http://tei-c.org/release/doc/tei-p5-doc/en/html/ref-stage.html

Usovski-Rosa 2010 — “Die szenischen Gestaltungsmittel in Cechovs Dramenwerk”. Inaugural Dissertation. Universiät zu Köln.