Dynamic Topic Modeling of Russian Prose of the First Third of the XXth Century by Means of Non-Negative Matrix Factorization*

Ekaterina Zamiraylova
e.zamiraylova@gmail.com

Olga Mitrofanova
o.mitrofanova@spbu.ru

Saint Petersburg State University, Saint Petersburg, Russian Federation

Abstract

This paper describes automatic topic spotting of literary texts based on the Russian short stories corpus, compiling stories written in the first third of the XXth century. Non-negative matrix factorization (NMF) is a valuable alternative to existing approaches of dynamic topic modeling and it can find niche topics and related vocabularies that are not captured by existent methods. The experiments were conducted on text samples extracted from the corpus, the given samples contain texts of 300 different authors. This approach allows to trace the topic dynamics of Russian prose for 30 years — from 1900 to 1930.

Keywords:*computational linguistics, dynamic topic modeling, Russian literature, Russian short stories*

1 Introduction

In the last decade topic modeling has become one of the most popular issues of computer linguistics. Topic modeling is usually understood as building a model that shows which topics appear in each document [Daud et al., 2010]. The topic model of a collection of text documents determines whether each document belongs to a different topic and it generates a list of words (terms) from which each topic is formed [Blei, Lafferty, 2006]. With this method, it is possible to process large amounts of data (fiction texts, magazine articles, news reports, social media, reviews, etc.) and automatically receive information about the topics of texts. Knowing what people are talking about and understanding their concerns and opinions is very valuable for science, business, political campaigns, etc.

Currently a large number of methods for topic modeling have been created. The most common in modern applications are methods based on Bayesian networks, which are probabilistic models on oriented graphs. Probabilistic topic models belong to a relatively young

field of research in an unsupervised theory. Probabilistic latent semantic analysis (PLSA) based on the principle of maximum likelihood was one of the first proposed as an alternative to classical clustering methods based on the calculation of distance functions. Next to PLSA latent Dirichlet allocation (LDA) and its numerous generalizations were proposed.

Following on from LDA methods similar probabilistic approaches have been consistently developed to track the evolution of topics over time in a sequentially organized corpus of documents, such as the dynamic topic model (DTM) [Blei, Lafferty, 2006]. Alternative algorithms, such as non-negative matrix factorization (NMF) [Lee and Seung, 1999] considered in this paper, have proven effective in finding underlying topics in text corpora [Wang et al., 2012]. For this reason, that algorithm was chosen for this study, which is aimed to automatic selection of dynamic topics in the Russian short stories corpus of the first third of the XXth century [see Sherstinova, Martynenko in this volume].

2 Selection rationale

Non-negative matrix factorization (NMF) is an unsupervised algorithm of machine learning that aims to detect useful features [Müller and Guido, 2017]. It is utilized for dimensionality of non-negative matrices, because it decomposes the data into factors in such a way that there are no negative values in them. Therefore, this method can be applied only to those data where features have non-negative values, as a non-negative sum of non-negative components cannot become negative [the same].

One of the advantages of NMF over existing LDA methods is that fewer parameter variants are used in the modelling process [Darek and Cross, 2016]. In addition, another benefit is that NMF can identify niche topics that tend to be under-reported in traditional LDA approaches [O'Callaghan et al., 2015]. Niche topics are sub-topics that can be identified within a dynamic topic.

The ability of NMF to consider how significant a word is to a document in a text collection, based on weighted term frequency values, is particularly useful. In particular, the application of log-based TF-IDF weighting factor to the data before the construction of the topic model contributed to diverse but semantically coherent topics that are less likely to be represented by the same high-frequency terms [Darek and Cross, 2016]. This makes NMF a suitable method for identifying both broad groups of high-level documents and niche topics with specialized dictionaries [O'Callaghan et al., 2015].

3 Experimental design

The experiment is based on a two-level strategy of topic modeling within the framework of non-negative matrix factorization to the Russian short stories corpus of the first third of the XXth century. This strategy is that the first stage is an application of the topic modeling NMF to one set of texts from a fixed period of time, the second stage is a combination of results of topic modeling from successive periods of time for detecting a set of dynamic topics related to a particular time window or the whole corpus.

4 Linguistic data set

The material for this paper is a selected data from the Russian short stories corpus of the first third of the XXth century, which is developed at the Philology Department of Saint Petersburg State University in cooperation with Philology Department of the National Research University Higher School of Economics, Saint Petersburg [Martynenko et al., 2018a; Sherstinova, Martynenko, 2019]. The data set consists of 300 stories of 300 unique writers — both world-famous and barely known. The corpus is a homogeneous resource, which is focused on one of the most common genres of fiction — the short story. This genre is the most popular among prose writers, its presence may be found in almost all kind of literary of almost all writers.

The corpus under development covers one of the most dramatic periods in the development of the Russian language and literature. The central point that divides the first third of the twentieth century into different time periods is the October Revolution of 1917. All other events are considered either as leading to it or as arising from it. It allows to make the quantitative analysis of language changes in rather wide chronological frameworks and to estimate what of the arisen language changes were fixed in language, were started to be often used by speakers or were disappeared after the revolutionary epoch [Martynenko et al., 2018b].

The base of the corpus provides a means for exploring the language of the first third of the twentieth century (1900-1930) divided into three main periods: 1) the beginning of the XXth century and the prerevolutionary years, including the First world war, 2) the revolutionary years — the February and the October revolutions and the Civil war, and 3) postrevolutionary years from the end of the Civil war to 1930. Each of these time periods will be analyzed separately and the results will be combined into an overall picture, reflecting the development of the Russian language in the first third of the XXth century [Martynenko et al., 2018b].

5 Experimental procedure

The experimental setup is pre-processing of texts that included: removal of non-text symbols, abbreviations, stop words and lemmatization. The volumes of the data sets are shown (in tokens) below:

Table 1. Volumes of the data sets before and after pre-processing

Before pre-processing			After pre-processing		
1900–1913	1914–1922	1923–1930	1900–1913	1914–1922	1923–1930
391736	315820	399550	195695	162381	217013

In the first step a document-term matrix is created to which TF-IDF and normalization of the document length are applied before each matrix is written. It includes marking documents

and creation of a document matrix for the window topic where the topic model is created by applying NMF to each time window.

Determining the number of topics is a nontrivial task because the choice of too few of them leads to overly generalized results while choosing too many topics entails too many small, highly similar topics [Green and Cross, 2015]. For the cases when this number is not known in advance there are different strategies for automatic or semi-automatic selection of number of topics. In particular it is proposed to build a Word2Vec Skipgram model using the Gensim library (https://radimrehurek.com/gensim/) from all documents in the case. The TC-W2V measure is used to compare different topic models and then select a model with a suitable number of topics. More details on the TC-W2V are in [O'Callaghan, 2015].

Applying the method mentioned in [Green and Cross, 2015] to determine the number of topics the following results were obtained:

- Top recommendations for number of topics for '1900–1913': 10 (Table 2)
- Top recommendations for number of topics for '1914–1922': 4 (Table 3)
- Top recommendations for number of topics for '1923–1930': 10 (Table 4)
- Top recommendations for number of dynamic topics: 4 (Table 5).

The ability of NMF to apply TF-IDF weighting to data before the topic modeling creates diverse but nonetheless coherent topics that are less likely to be represented by the same high-frequency terms allowing identification of both broad and niche topics with specialized vocabularies [O'Callaghan et al., 2015]. In the context of the study of the first third of the XXth century the discovery of these niche topics is an advantage that helps to consider the components of the topics and analyze the realities of the period in more detail. To illustrate this idea table 5 shows the top 10 terms for 4 dynamic topics. Terms in bold are unique to a topic; terms in italics are met in an overall description of a topic and in time windows (or even in one time window), terms in bold and italics are found within time windows but are not in an overall description of a topic.

The above list of words created by NMF to describe topics is rich and various, moreover each time window has its own unique words. If to compare it with the most common LDA method as the authors of the model do [Darek and Cross, 2016] NMF is more suitable for niche content analysis while LDA offers only a good general description of broader topics.

6 Linguistic interpretation of experimental results

The highest interest for linguistic analysis is the content of dynamic topics. Table 5 lists the top 4 dynamic topics penetrating all time periods (1900–1913, 1914–1922, 1923–1930). Table 6 shows niche topics and vocabularies of each dynamic topic in a specific time period. For instance, the first broad topic in the first time window is represented by 40 words with the biggest amount of unique terms (*писать (to write), любовь (love), любить (to love), сцена (scene), роль (role), ребенок (child), муж (husband), жена (wife), счастье (happiness), кабинет (room, office), деньга (money), русский (Russian), пароход (steamer), город (city, town)* and etc.). Inference should be drawn that writers at the beginning of the century

wrote a lot about a mode of life: about family (*муж (husband), ребенок (child), жена (wife), мама (mom), сестра (sister), отец (father)*), work (*кабинет (room, office), деньга (money)*) and events that occurred with the main characters, which interacted with people of different professions (*купец (merchant), приказчик (manager), извозчик (horse-cab driver), доктор (doctor)*). The number of unique terms is not numerous in the second time window (1914–1922), which is due to the fact that this is a revolutionary time and the description of life is minimal. In the post-revolutionary period the vocabulary increases again, there are unique words that reflect the «new life» (*товарищ (comrade), завод (factory), гражданин (citizen), рабочий (working)*, etc.).

There are more words describing nature in the second dynamic topic: 1900–1913 (*пруд (pond), река (river), ночь (night), солнце (sun), куст (bush), лес (forest), волк (wolf)*), during 1914–1922 - more abstract (*ветер (wind), море (sea), небо (sky), солнце (sun), ночь (night), берег (bank)), 1923–1930 (сосна (pine), птица (bird), зверь(beast), лес (forest), болото (swamp)*). The third dynamic topic is filled with words related to the military sphere. However, if to look through the niche topics and vocabularies of each period at the beginning of the century only a few words can be attributed to the military theme (*солдат (soldier), офицер (officer), пост (post)), the rest of unique terms are more related to the usual way of life (барин (lord), старик (old man), деревня (village), благородие (honour)*, etc.)), which can indicate to the maintenance of order and regulation of people relations . In the second time window (1914–1922) two unique words «немецкий» (German) and «немец» (German) appeared, and there are no abstract words, almost all content refer to the military (*солдат (soldier), офицер (officer), стрелять (to shoot), рота (troop)*, etc.), which fully reflects the revolutionary period. There is a large number of unique words in the third time window where there are the following niche topics: movement by train (*вагон (coach), пассажир (passenger), поезд (train), станция (station), ход (motion), курс (course)), family (муж (husband), мама (mom), ребенок (child), мальчик (boy)), house/home (дом (house, home), кухня (kitchen)*). Only the word «солдат» (soldier) can be attributed to the military topic.

The fourth dynamic topic has several niche topics: *village (телега (telega, horse wagon), народ (folk), изба (hut, house), etc.), religion (поп (priest), батюшка (priest), церковь (church).* It is worth paying attention to the dynamics of changes in the religious topic: more words in the revolutionary time (*батюшка (priest), Бог (God), святой (saint)) compared to the beginning of the century (батюшка (priest)), and the postrevolutionary period (поп (priest), церковь (church)*).

The above analysis shows that the internal organization of topics described as a bundle of paradigmatic and syntagmatic connections between the words of the same topic which vary in different time intervals within the same dynamic topic, change significantly over time and reflect the external events.

If we consider the components of topics from the linguistic viewpoint the largest number of words belongs to the nominative class, which is represented by common nouns. Proper names (*Александр (Alexander), Алексей (Alexey), Анна (Anna), Владимир (Vladimir), Володя (Volodya), Мишка (Mishka), Вера (Vera)*, etc.) are deliberately removed since there are many dialogues in the data. The frequency of names is very high and its topic distribution is not conditioned by anything.

Table 2: Top recommendations for number of topics for '1900–1913': 10

Rank	1900–1913_01	1900–1913_02	1900–1913_03	1900–1913_04	1900–1913_05
1	письмо (letter)	тюрьма (prison)	солдат (soldier)	леса (forest)	ребенок (child)
2	рука (hand, arm)	жизнь (life)	барин (lord)	лес (forest)	муж (husband)
3	писать (to write)	мысль (thought)	старик (old man)	лошадь (horse)	жизнь (life)
4	знать (to know)	рука (hand, arm)	офицер (officer)	изба (hut, house)	друг (friend)
5	любовь (love)	звук (sound)	деревня (village)	дед (grandfather, old man)	год (year)
6	любить (to love)	окно (window)	благородие (honour)	куст (bush)	жена (wife)
7	комната (room)	капли (drops)	крестьянин (peasant)	солнце (sun)	женщина (woman)
8	женщина (woman)	смерть (death)	пост (post)	волк (wolf)	счастье (happiness)
9	сцена (scene)	казаться (to seem)	огонек (little fire)	нога (leg)	кабинет (room, office)
10	роль (role)	слово (word)	изба (hut, house)	темный (dark)	деньга (money)

Rank	1900–1913_06	1900–1913_07	1900–1913_08	1900–1913_09	1900–1913_10
1	толпа (crowd)	студент (student)	батюшка (father, priest)	русский (Russian)	отец (father)
2	улица (street)	девушка (young lady)	рука (hand, arm)	пароход (steamer)	сестра (sister)
3	рабочий (working)	пруд (pond)	жена (wife)	город (city, town)	мама (mom)
4	крик (shout)	река (river)	дело (affair)	купец (merchant)	комната (room)
5	стоять (to stand)	лодка (boat)	хозяин (landlord)	приказчик (manager)	доктор (doctor)
6	кричать (to shout)	ночь (night)	баба (country woman, peasant's wife)	ночь (night)	дом (house, home)
7	ребенок (child)	дядя (uncle)	пойти (to go)	знать (to know)	мать (mother)
8	голос (voice)	дорога (road)	деньга (money)	старик (old man)	суп (soup)
9	друг (friend)	мгновение (instant)	матушка (mother, priest's wife)	извозчик (horse-cab driver)	обедать (to dine)
10	бежать (to run)	дом (house, home)	тетка (aunt)	часы (clock, watch)	думать (to think)

Table 3: Top recommendations for number of topics for '1914–1922': 4

Rank	1914–1922_01	1914–1922_02	1914–1922_03	1914–1922_04
1	письмо (letter)	солдат (soldier)	отец (father)	рука (hand, arm)
2	знать (to know)	немец (German)	мужик (country man, peasant man)	ветер (wind)
3	комната (room)	офицер (officer)	баба (country woman, peasant's wife)	море (sea)
4	жизнь (life)	товарищ (comrade)	старик (old man)	небо (sky)
5	рука (hand, arm)	винтовка (gun)	батюшка (father, priest)	солнце (sun)
6	слово (word)	немецкий (German)	Бог (God)	ночь (night)
7	студент (student)	стрелять (to shoot)	девка (maid)	берег (bank)
8	думать (to think)	рука (hand, arm)	святой (saint)	нога (leg)
9	женщина (woman)	команда (command, team)	изба (hut, house)	река (river)
10	дверь (door)	рота (troop)	учитель (teacher)	тело (body)

7 Results

Most nouns refer to the description of people (*ребенок (child), девушка (young lady), женщина (woman), старик (old man), дед (grandfather, old man), баба (country woman, peasant's wife), отец (father), мать (mother), муж (husband), жена (wife), father*

Table 4: Top recommendations for number of topics for '1923–1930': 10

Rank	1923 –1930_01	1923–1930_02	1923–1930_03	1923–1930_04	1923–1930_05
1	комната (room)	мужик (country man, peasant man)	дед (grandfather, old man)	комиссар (commissioner)	товарищ (comrade)
2	рука (hand, arm)	баба (country woman, peasants wife)	большой (big)	командир (commanding officer)	фабрика (plant)
3	девушка (young lady)	лошадь (horse)	сосна (pine)	пароход (steamer)	дело (affair)
4	дверь (door)	рука (hand, arm)	птица (bird)	отряд (squad)	жить (to live)
5	окно (window)	телега (telega, horse wagon)	горбатый (hump-backed)	винтовка (gun)	ребенок (child)
6	доктор (doctor)	нога (leg)	борода (beard)	команда (command, team)	гражданин (citizen)
7	лампа (lamp)	изба (hut, house)	черный (black)	капитан (captain)	сидеть (to sit)
8	хотеть (to want)	тело (body)	рука (hand, arm)	штаб (headquarter)	рабочий (working)
9	письмо (letter)	дорога (road)	глянуть (to peep)	солдат (soldier)	хороший (good)
10	любить (to love)	хлеб (bread)	шапка (cap)	начальник (chief)	слово (word)

Rank	1923–1930_06	1923–1930_07	1923–1930_08	1923–1930_09	1923–1930_10
1	старик (old man)	вагон (coach)	поп (priest)	работа (work)	мальчик (boy)
2	снег (snow)	пассажир (passenger)	старуха (old woman)	рабочий (working)	мама (mom)
3	ружье (rifle)	поезд (train)	отец (father)	работать (to work)	вещий (fatidical)
4	зверь (beast)	мешок (bag)	церковь (church)	машина (machine)	ребенок (child)
5	леса (forest)	станция (station)	мужик (country man, peasant man)	жизнь (life)	дом (house, home)
6	дерево (tree)	ученый (scientist)	народ (folk)	большой (big)	старший (senior)
7	лес (forest)	рука (hand, arm)	телега (telega, horse wagon)	пространство (space)	кухня (kitchen)
8	дядя (uncle)	вскочить (to jump up)	праздник (holiday)	завод (factory)	солдат (soldier)
9	огонь (fire)	муж (husband)	свадьба (wedding)	год (year)	рука (hand, arm)
10	болото (swamp)	ход (motion)	старик (old man)	думать (to think)	курс (course)

Table 5: Top recommendations for number of dynamic topics: 4

Rank	D01	D02	D03	D04
1	рука (hand, arm)	дед (grandfather, old man)	поп (priest)	вагон (coach)
2	комната (room)	леса (forest)	мужик (country man, peasant man)	комиссар (commissioner)
3	жизнь (life)	старик (old man)	баба (country woman, peasant's wife)	солдат (soldier)
4	знать (to know)	ружье (rifle)	телега (telega, horse wagon)	пароход (steamer)
5	дом (house, home)	лес (forest)	изба (hut, house)	командир (commanding officer)
6	хотеть (to want)	снег (snow)	старуха (old woman)	толпа (crowd)
7	думать (to think)	птица (bird)	народ (folk)	рука (hand, arm)
8	год (year)	сосна (pine)	отец (father)	винтовка (gun)
9	большой (big)	куст (bush)	лошадь (horse)	пассажир (passenger)
10	ребенок (child)	большой (big)	рука (hand, arm)	отряд (squad)

(батюшка), матушка (mother), сестра (sister), дядя (uncle), etc.), profession (солдат (soldier), барин (lord), офицер (officer), студент (student), крестьянин (peasant), доктор (doctor), etc.), body parts (рука (hand, arm), нога (leg), грудь (chest)). Another group of nouns - everyday realities (*комната (room), улица (street), пост (post), изба (hut, house), дом (house, home), письмо (letter), etc.), nature and animals (лес (forest), куст (bush), волк*

Table 6: Niche topics of 4 dynamic topics (D1-D4)

D1

Overall	1900–1913	1900–1913 (2)	1900–1913 (3)	1900–1913 (4)	1914–1922	1923–1930	1923–1930 (2)	1923–1930 (3)
рука (hand, arm)	*письмо (letter)*	***ребенок (child)***	**русский (Russian)**	**отец (father)**	*письмо (letter)*	*комната (room)*	**товарищ (comrade)**	**работа (work)**
жизнь (life)	*рука (hand, arm)*	**муж (husband)**	**пароход (steamer)**	**сестра (sister)**	*знать (to know)*	*рука (hand, arm)*	**фабрика (plant)**	***рабочий (working)***
комната (room)	**писать (to write)**	*жизнь (life)*	**город (city, town)**	**мама (mom)**	*комната (room)*	**девушка (young lady)**	**дело (affair)**	**работать (to work)**
знать (to know)	*знать (to know)*	**друг (friend)**	**купец (merchant)**	*комната (room)*	*жизнь (life)*	**дверь (door)**	**жить (to live)**	**машина (machine)**
год (year)	**любовь (love)**	*год (year)*	**приказчик (manager)**	***доктор (doctor)***	*рука (hand, arm)*	**окно (window)**	***ребенок (child)***	*жизнь (life)*
думать (to think)	***любить (to love)***	жена (wife)	**ночь (night)**	*дом (house, home)*	*слово (word)*	***доктор (doctor)***	**гражданин (citizen)**	**большой (big)**
дом (house, home)	*комната (room)*	***женщина (woman)***	*знать (to know)*	**мать (mother)**	**студент (student)**	**лампа (lamp)**	**сидеть (to sit)**	**простран-ство (space)**
письмо (letter)	***женщина (woman)***	**счастье (happiness)**	**старик (old man)**	**суп (soup)**	*думать (to think)*	*хотеть (to want)*	***рабочий (working)***	**завод (factory)**
хотеть (to want)	**сцена (scene)**	**кабинет (room, office)**	**извозчик (horse-cab driver)**	**обедать (to dine)**	***женщи-на (woman)***	*письмо (letter)*	**хороший (good)**	*год (year)*
слово (word)	**роль (role)**	**деньга (money)**	**часы (clock, watch)**	*думать (to think)*	***дверь (door)***	***любить (to love)***	*слово (word)*	*думать (to think)*

D2

Overall	1900–1913	1900–1913 (2)	1900–1913 (3)	1914–1922	1923–1930	1923–1930 (2)
дед (grandfather, old man)	*рука (hand, arm)*	**студент (student)**	*леса (forest)*	***рука (hand, arm)***	*дед (grandfather, old man)*	*старик (old man)*
леса (forest)	**жизнь (life)**	**девушка (young lady)**	*лес (forest)*	**ветер (wind)**	**большой (big)**	*снег (snow)*
старик (old man)	**мысль (thought)**	**пруд (pond)**	**лошадь (horse)**	**море (sea)**	**сосна (pine)**	**ружье (rifle)**
нога (leg)	**тюрьма (prison)**	*река (river)*	**изба (hut, house)**	**небо (sky)**	**птица (bird)**	**зверь (beast)**
лес (forest)	**звук (sound)**	***ночь (night)***	**куст (bush)**	*солнце (sun)*	**горбатый (humpbacked)**	*леса (forest)*
снег (snow)	**окно (window)**	**лодка (boat)**	*дед (grandfather, old man)*	**ночь (night)**	**борода (beard)**	**дерево (tree)**
солнце (sun)	**голос (voice)**	**пароход (steamer)**	*солнце (sun)*	**берег (bank)**	*черный (black)*	*лес (forest)*
черный (black)	*черный (black)*	**спать (to sleep)**	**волк (wolf)**	*нога (leg)*	*рука (hand, arm)*	**дядя (uncle)**
река (river)	**грудь (chest)**	*старик (old man)*	*нога (leg)*	*река (river)*	**глянуть (to peep)**	**огонь (fire)**
темный (dark)	**казаться (to seem)**	**город (city, town)**	*темный (dark)*	**тело (body)**	**шапка (cap)**	**болото (swamp)**

(wolf), солнце (sun), лошадь (horse), pond (пруд), река (river), ночь (night), снег (snow)), abstract (жизнь (life), счастье (happiness), мысль (thought), смерть (death), etc.), as well as collective nouns (*толпа (crowd), рота (troop), народ (folk)*).

The predicative class is represented by verbs: *писать (to write), знать (to know),*

D3

Overall	1900–1913	1900–1913 (2)	1914–1922	1922–1930	1922–1930 (2)	1922–1930 (3)
солдат (soldier)	*солдат (soldier)*	*толпа (crowd)*	*солдат (soldier)*	*комиссар (commissioner)*	**вагон (coach)**	**мальчик (boy)**
офицер (officer)	**барин (lord)**	**улица (street)**	**немец (German)**	*командир (commanding officer)*	**пассажир (passenger)**	**мама (mom)**
комиссар (commissioner)	**старик (old man)**	**рабочий (working)**	*офицер (officer)*	**пароход (steamer)**	**поезд (train)**	**вешний (fatidical)**
толпа (crowd)	*офицер (officer)*	**крик (shout)**	*товарищ (comrade)*	**отряд (squad)**	**мешок (bag)**	*ребенок (child)*
товарищ (comrade)	**деревня (village)**	**стоять (to stand)**	*винтовка (gun)*	*винтовка (gun)*	**станция (station)**	**дом (house, home)**
вагон (coach)	**благородие (honour)**	**кричать (to shout)**	**немецкий (German)**	*команда (command, team)*	**ученый(scientist)**	**старший (senior)**
командир (commanding officer)	**крестьянин (peasant)**	*ребенок (child)*	*стрелять (to shoot)*	**капитан (captain)**	*рука (hand, arm)*	**кухня (kitchen)**
винтовка (gun)	**пост (post)**	**голос (voice)**	*рука (hand, arm)*	**штаб (headquarter)**	**вскочить (to jump up)**	*солдат (soldier)*
рука (hand, arm)	**огонек (little fire)**	**друг (friend)**	*команда (command, team)*	*солдат (soldier)*	**муж (husband)**	*рука (hand, arm)*
стрелять (to shoot)	*изба (hut, house)*	**бежать (to run)**	**рота (troop)**	**начальник (chief)**	**ход (motion)**	**курс (course)**

D4

Overall	1900–1913	1914–1922	1922–1930	1922–1930 (2)
поп (priest)	*батюшка (father, priest)*	*отец (father)*	*мужик (country man, peasant man)*	*поп (priest)*
мужик (country man, peasant man)	*рука (hand, arm)*	*мужик (country man, peasant man)*	*баба (country woman, peasants wife)*	*старуха (old woman)*
баба (country woman, peasants wife)	**жена (wife)**	*баба (country woman, peasants wife)*	**лошадь (horse)**	*отец (father)*
отец (father)	**дело (affair)**	*старик (old man)*	*рука (hand, arm)*	*церковь (church)*
старуха (old woman)	**хозяин (landlord)**	*батюшка (father, priest)*	*телега (telega, horse wagon)*	*мужик (country man, peasant man)*
телега (telega, horse wagon)	*баба (country woman, peasant's wife)*	**Бог (God)**	**нога (leg)**	*народ (folk)*
народ (folk)	**пойти (to go)**	**девка (maid)**	*изба (hut, house)*	*телега (telega, horse wagon)*
изба (hut, house)	**деньга (money)**	**святой (saint)**	**тело (body)**	**праздник (holiday)**
церковь (church)	**матушка (mother, priest's wife)**	*изба (hut, house)*	**дорога (road)**	**свадьба (wedding)**
старик (old man)	**тетка (aunt)**	**учитель (teacher)**	**хлеб (bread)**	*старик (old man)*

любить (to love), хотеть (to want), стоять (to stand), кричать (to shout), бежать (to run), думать (to think), обедать (to dine), прийти (to come), сидеть (to sit), жить (to live), работать (to work), спать (to sleep), глянуть (to peep), стрелять (to shoot), вскочить (to jump up), пойти (to go). From the point of view of semantic classification of verbs developed by V. V. Vinogradov and supplemented by G. A. Zolotova [Zolotova, 2004] the present verbs belong to the main semantic classes:

1) verbs of movement: *стоять (to stand), сидеть (to sit), глянуть (to peep), бежать (to run), прийти (to come), вскочить (to jump up), идти (to go)*;

2) verbs of speech action: *кричать (to shout)*;

3) verbs of mental actions: *знать (to know), думать (to think), казаться (to seem)*;

4) verbs of emotional action: *любить (to love)*;

5) verbs of physiological action: *жить (to live), обедать (to dine), спать (to sleep)*;

6) verbs of activity or occupation: *работать (to work), писать (to write), стрелять (to shoot)*;

7) modal verb: *хотеть (to want).*

The attributive class is the narrowest — it includes qualitative (*хороший (good), большой (big), черный (black), темный (dark), горбатый (humpbacked), старший (senior))* and relative adjectives *(рабочий (working), русский (Russian), немецкий (German)*).

The correlation of words in the topics reflects the diversity of paradigmatic and syntagmatic relations that organize the text [Mitrofanova et al., 2014; Mitrofanova, 2014]. The language connections within the topics may be describes with lexical functions in the model «Meaning $<=>$ Text» [Melchuk, 1974/1999] which allows to cover the predictable, idiomatic connections of the word and its lexical correlates.

Among paradigmatic relations in topics the following are prevailed: synonymy (Syn), antonymy (Anti) and derivational (Der) relations, etc. For example, Syn: *мама (mom) — мать (mother), друг (friend) — товарищ (comrade), фабрика (plant) — завод (factory), черный (black) — темный (dark), поп (priest) — батюшка (priest),* etc. Anti: *деревня (village) — город (city, town)*, Der: *работа (work) — рабочий (working) — работать (to work), немец (German) — немецкий (German), любовь (love) — любить (to love), крик (shout) — кричать(to shout), команда (command, team) — командир (commanding officer)*, etc. Partitive relations: *семья (family) — ребенок (child), мама (mom), отец (father), дядя (uncle), сестра (sister), муж (husband), жена (wife), отец (father), мать (mother), тетка (aunt); армия (military) — офицер (officer), солдат (soldier), рота (troop), штаб (headquarter), капитан (captain), команда (command, team), etc.; природа (nature) — пруд (pond), река (river), берег (bank), лес (forest), снег (snow), солнце (sun), ветер (wind), дерево (tree), куст (bush), болото (swamp), etc..; лес (forest) — дерево (tree), куст (bush), болото (swamp); охота (hunt) — ружье (rifle), зверь (beast), лес (forest), огонь (fire); деревня (village) — изба (hut, house), телега (telega, horse wagon), крестьянин (peasant), барин (lord); дом (house, home) — комната (room), дверь (door), окно (window), лампа (lamp), кухня (kitchen), кабинет (room, office); передвижение на поезде (go by train) – вагон (coach), пассажир (passenger), поезд (train), станция (station), ход (motion); завод (factory) – работа (work), рабочий (working), работать (to work), машина (machine),* etc.

Syntagmatic relations are realized at the level of valence frames filled with words from the topic. Among lexical functions Oper1,2 may be selected, which connect a verb, the name of the first or the second actant in the role of subject and the name of the situation as additions: *суп (soup) – обедать (to dine), письмо (letter) - писать (to write), винтовка (gun) – стрелять (to shoot), ребенок (child) – кричать (to shout)*, etc. In addition, there are a number of examples for the implementation of the lexical function Cap: *(команда (command, team) – командир (commanding officer); штаб (headquarter) – начальник (chief); отряд (squad) – командир (commanding officer), церковь (church) – поп (priest), пароход (steamer) - капитан (captain)*, etc. The lexical function Equip ("personnel, staff"): *people (folk) – man (country man, peasant man), etc., lexical function Doc (res) ("document that is the result"): write (to write) – letter (letter); draw – drawing,* etc.

8 Conclusion

The observations made during the experiments confirm the expediency of using non-negative matrix factorization for the issues of topic modeling, including the evaluation of the content of texts as a result of semantic compression.

The results obtained in the processing of the selected data from the Russian short stories corpus of the first third of the XXth century indicate the diversity of the implementation of dynamic topic in different time periods.

The research data makes it possible to interpret the received results from the perspective of the theory of lexical functions, as well as to use historical and literary approaches for this purpose. The content of the topics allows to draw conclusions about the topic dynamics of Russian prose for 30 years — from 1900 to 1930.

Acknowledgements

The research is supported by the Russian Foundation for Basic Research, project # 17-29-09173 "The Russian language on the edge of radical historical changes: the study of language and style in prerevolutionary, revolutionary and post-revolutionary artistic prose by the methods of mathematical and computer linguistics (a corpus-based research on Russian short stories)".

References

[Daud et al. 2010] Daud A., Li J., Zhou L., Muhammad F. (2010) Knowledge Discovery through Directed Probabilistic Topic Models: a Survey // Proceedings of Frontiers of Computer Science in China.

[Blei, Lafferty, 2006] Blei D. M., Lafferty J. D. (2006) Dynamic topic models. In Proc. 23rd International Conference on Machine Learning, pp. 113–120.

[Lee and Seung, 1999] Lee D. D. and Seung H. S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–91.

[Wang et al., 2012] Wang, Q., Z. Cao, J. Xu, and H. Li (2012). Group matrix factorization for scalable topic modeling. In Proc. 35th SIGIR Conf. on Research and Development in Information Retrieval, pp. 375–384. ACM.

[Sherstinova, Martynenko, 2019] Sherstinova T., Martynenko G. (2019) Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. This volume.

[Müller and Guido, 2017] Müller. A. and Guido. S. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly., 2016

[Darek and Cross, 2016] Greene D. and Cross J. P. (2016) Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. ArXiv abs/1607.03055

[O'Callaghan et al., 2015] O'Callaghan, D., Greene D, Carthy J., and Cunningham P. (2015) An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications (ESWA).

[Martynenko et al., 2018a] Martynenko G.Ya., Sherstinova T.Yu., Melnik A. G., Popova T.I. (2018) Methodological problems of creating a computer anthology of the Russian short story as a language resource for the study of the language and style of Russian prose in the era of revolutionary changes (the first third of the XX century) / Computational linguistics and computational ontologies. Issue 2 (Proceedings of the XXI international joint conference "Internet and modern society, IMS-2018, St. Petersburg, May 30-June 2, 2018 Collection of scientific articles"). - St. Petersburg: ITMO University, 2018. P. 99-104. (In Rus.) = Metodologicheskiye problemy sozdaniya Kompyuternoy antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoy khudozhestvennoy prozy v epokhu revolyutsionnykh peremen (pervoy treti XX veka) / Kompyuternaya lingvistika i vychislitelnyye ontologii. Vypusk 2 (Trudy XXI Mezhdunarodnoy obyedinennoy konferentsii "Internet i sovremennoye obshchestvo. IMS-2018. Sankt-Peterburg. 30 Maya - 2 Iyunya 2018 g. Sbornik nauchnykh statey"). — SPb: Universitet ITMO. 2018. S. 99 -104.

[Martynenko et al., 2018b] Martynenko G.Ya., Sherstinova T.Yu., Popova T.I., Melnik A.G., Zamiraylova E.V. (2018) On the principles of creation of the Russian short stories corpus of the first third of the 20th century. Proceedings of the XV International conference on computer and cognitive linguistics "TEL 2018". - Kazan, 2018. Pp. 180-197. (In Rus.) = O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka // Trudy XV Mezhdunarodnoy konferentsii po kompyuternoy i kognitivnoy lingvistike «TEL 2018». – Kazan. 2018. – S. 180–197.

[Green and Cross, 2015] Greene D., and Cross J. P. (2015) Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis ACM Web Science 2015, 28 June - 1 July, 2015 Oxford, UK.

[Zolotova et al., 2004] olotova G. A., Onipenko N. T., Sidorova M. Y. Communicative grammar of the Russian language. Ed. - Moscow: Nauka, 2004. — 544 p (In Rus.) = Kommunikativnaya grammatika russkogo yazyka. — M.: Nauka. 2004. — 544 s. — ISBN 5-88744-050-3

[Mitrofanova et al., 2014] Mitrofanova O. A., Shimorina, A. S., Koltsov S. N., Koltsova O. Yu. (2014) Modeling semantic links in social media texts using the LDA algorithm (based on the Russian-language segment of the LiveJournal). Structural and applied linguistics, Vol. 10, 151-168. (in Rus.) = Modelirovaniye semanticheskikh svyazey v tekstakh sotsialnykh setey s pomoshchyu algoritma LDA (na materiale russkoyazychnogo segmenta Zhivogo Zhurnala). Strukturnaya i prikladnaya lingvistka. Vyp. 10. 151-168.

[Mitrofanova, 2014] Mitrofanova O. A. (2014) Topic modeling of special texts based on LDA algorithm.XLII International philological conference. March 11-16, 2013. Selected works. SPb. (in Rus.) = Modelirovanije tematiki special'nyh tekstov na osnove algoritma LDA XLII Mezhdunarodnaya filologicheskaya konferencija. 11–16 marta 2013. Izbrannyje trudy. SPb

[Melchuk, 1974/1999] Melchuk I. A. (1974/1999) Experience of the theory of the linguistic models «Meaning Text». Moscow, 1974/199. 1974 (In Russ.) = Opyt teorii lingvistitcheskix modelej Smysl Tekst, Moskva, 1974/1999.