Past Voices, Present Insights: Sociolinguistic Research through Literary Artifacts

Tatiana Sherstinova, Elizaveta Ziulkova, Margarita Kirina
National Research University Higher School of Economics
Saint Petersburg, Russia
tsherstinova@hse.ru, alisiacats99@gmail.com, mkirina@hse.ru

Abstract—**Oral speech, historically the foundational mode of human communication, has not been explored as extensively as its written counterpart. This disparity underscores the necessity of examining sociolinguistic characteristics of speech across time. Current analyses often rely on data from contemporary speech corpora, yet understanding historical speech patterns is equally vital. Literary works, particularly from periods with scarce speech resources, offer invaluable insights into the sociolinguistic landscape of the past. Extracting speech fragments from literary texts creates a sub-corpus, approximating the spoken word as "heard" by writers of the time. This study, leveraging the Corpus of Russian Short Stories of the early 20th century, provides preliminary statistical insights into the speech patterns of various social groups. Its contribution is twofold: firstly, in pioneering a methodology for investigating sociolinguistic variability through literary analysis; secondly, in laying the groundwork for a theoretical model of dynamic sociolinguistic variability. These findings not only enhance our understanding of historical speech patterns but also aid in forecasting sociolinguistic trends, thereby informing the development of future speech technology applications tailored to evolving language use.**

I. Introduction

Oral speech represents a unique genre of speech behavior, the most natural for humans. Yet, the ability to record spoken word using sound recording devices emerged relatively recently – less than 150 years ago. Initially, recordings were made in exceptional cases, and only today has it become possible to record speech en masse, thanks to the widespread availability of mobile phones and personal audio equipment. Therefore, despite oral speech being historically primary and remaining the leading form of communication for human interaction, it remains understudied compared to written speech, for which centuries of civilization have amassed large arrays of textual data, now transformed into text corpora.

Given that language is inherently social, it is imperative to consider speech within its sociocultural context. Communication occurs within the fabric of society, emphasizing the need for analyzing speech through not just psychological or individualistic lenses, but also through sociological perspectives. This idea aligns with Baudouin de Courtenay's assertion that the foundation of linguistics should encompass both individual psychology and societal influences [1]. The study of how various aspects of society, including cultural norms and contexts, influence language usage and its variations is undertaken within the framework of sociolinguistics. This interdisciplinary field combines linguistics and sociology to explore how social factors shape linguistic behavior [2]. For example, such social factors as gender, age, social standing, profession, etc., have a significant impact of the linguistic characteristics of one's speech—from intonation and speech fluency to individual's lexical choices, sentence lengths, and phraseological diversity [3].

To study the sociological characteristics of speech in dynamics, it is crucial not only to have a current snapshot of data obtained from speech corpora but also to look a bit back to see how stable, for example, male or female speech characteristics are over time. This task becomes particularly challenging for the everyday register of spoken language since non-public, everyday speech seldom became the object of recording. In situations where speech resources are scarce, literary fiction can serve as a source for sociolinguistic research of speech for earlier periods, with the extraction of speech fragments providing a sub-corpus that serves as an approximation of speech—as it was heard and conveyed by writers of those times.

The research described in this article has allowed us to obtain preliminary statistical characteristics for different social groups of speakers based on the representative corpus of the Russian prose. The significance of the work lies in the development and testing of a methodology for researching sociolinguistic variability based on literary text.

The research was conducted using the material from the Corpus of Russian Short Stories of 1900-1930s (https://russian-short-stories.ru), a specialized linguistic resource designed for extensive studies of the Russian language in artistic prose and Russian literature as a whole [16], [25]. The corpus was automatically segmented into author's speech and characters' speech, similar to the well-known CLiC Dickens corpus [27], followed by manual error correction and tagging of each utterance with its speaker. A database was created identifying all speaking characters in the annotated sample, with each receiving a sociological description based on the information contained in the literary text, enabling sociolinguistic research of character speech.

The characters in Russian stories from the first third of the 20th century are multidimensional, with unique individual traits as well as sociodemographic attributes like gender, age, occupation, social class, and marital status. These sociological factors provide a framework for analyzing the characters' spoken language, offering a lens through which to understand how different social strata, educational backgrounds, genders, and age groups articulate themselves in similar contexts [2].

The study of characters' spoken language in narratives offers insights into how their linguistic traits correlate with sociodemographic variables like gender, age, and occupation. When working with a corpus of characters in Russian short stories, one key task is to identify the sociological variables that will be instrumental in a comparative analysis of their spoken language. The characters in this corpus hail from diverse social backgrounds, including the high-class intelligentsia, middle class, and lower-class peasants. Each group brings its unique social and economic status, as well as education levels, into the mix. These sociodemographic factors contribute to the nuances of each character's speech, as substantiated by studies focusing on the linguistic markers associated with an individual's economic or social class [4], [5]. Therefore, an in-depth understanding of characters' speech requires a multi-layered approach that takes into account both individual and sociological determinants. In this research we consider two main sociolinguistic parameters—speaker's gender and age.

II. Sociolinguistic Variables and Language Studies

A. Gender

Gender serves as a significant sociological variable affecting speech patterns, with distinctions in language use between men and women being a focal point in contemporary sociolinguistic research. American studies from the 20th century have extensively explored this relationship, highlighting various facets of gendered language use [7], [8], [9].

Deborah Tannen's seminal work delves into the communication breakdowns that can occur between men and women, attributing these disparities to societal influences and differing communication styles conditioned by gender [28]. Tannen posits that women's speech tends to be more emotionally charged, conciliatory, and less confrontational. Conversely, men often adopt a more competitive stance in conversations, frequently steering dialogues and changing topics to establish dominance.

Robin Lakoff's influential article "Language and Woman's Place" offers further nuances, demonstrating that women tend to use adjectives related to color nuances and intensifiers like "*so*" and "*very*" more frequently than men [10]. From a politeness perspective, women generally use the imperative mood less and incorporate courtesy markers like "*please*" more frequently. The article also introduces the hypothesis that women's vocabulary may display greater levels of uncertainty or hesitation, employing words that reflect doubt more often.

Russian and soviet research has also delved into linguistic genderology, particularly as it pertains to forensic applications. These studies aim to differentiate between male and female speech for purposes like identity verification through written or oral language, or to detect imitation [11], [12].

According to these studies, male written language tends to incorporate prison and army slang, and often employs obscenities—in a monotonous fashion. Emotional expression in men's writing is generally subdued, opting for words with "the least emotional indexing", leading to a more restrained and monotone emotional lexicon. Additionally, male writing frequently borrows clichés from journalistic sources [11].

In contrast, women's written speech often employs phrases and constructs that express uncertainty, such as "*maybe*", "*apparently*," or "*in my opinion*" [ibid.]. Women are also noted for incorporating bookish vocabulary and literary expressions, aiming for a more refined and intelligent linguistic presentation. When expressing emotions, women's lexical choices are generally more varied and nuanced, a notion supported by many other linguistic studies on gender and language.

In applying these insights to a comparative analysis of characters in the Corpus of Russian Short Stories, researchers can test the hypotheses regarding gender-specific linguistic characteristics within narrative contexts. This interdisciplinary approach offers a nuanced understanding of how gender, as both an individual and sociological variable, shapes language in both real-world and fictional settings.

B. Age

Age is another key variable that significantly shapes language usage, leading to distinct linguistic patterns among different age groups. Middle-aged speakers are typically seen as representing the "standard" in language use. While teenagers and young adults are more likely to employ jargon, neologisms, and vocabulary that diverges from standardized language norms, older individuals often use archaic or outdated terms.

Youth language has been specifically examined in various studies addressing contemporary language trends, youth-specific jargon, and the state of modern spoken language [13]. One noteworthy aspect of youth speech is its type of speech behavior. The age factor also often predetermines a prevailing "anti-behavior" among young people, characterized by protest, negativism, and even aggression [29].

In the context of analyzing a corpus of characters, whether in real-world studies or literary works like the Corpus of Russian Short Stories, one can expect to identify distinctive linguistic features corresponding to different age groups. Researchers may look into the frequency of speech patterns, types of vocabulary used, and other behavioral markers to understand the age-specific characteristics in the speech of characters. Thus, age not only influences language on an individual level but also serves as a sociological variable that impacts speech across various contexts.

C. Features of the historical context

When examining the oral speech of characters in Russian stories in the first third of the 20th century, it's crucial to consider the historical and linguistic contexts of the era spanning 1900-1930. This period was marked by significant linguistic changes and innovations, as highlighted by Soviet linguist A. M. Selishchev in "The Language of the Revolutionary Era" [14].

One prominent feature of the language during this time was the rise in the usage of obscene or vulgar language, termed as "cynical swearing" [ibid., p. 48]. This was indicative of broader societal shifts, including social and political upheavals.

Moreover, certain linguistic features were specific to particular social classes or professional groups. For instance, revolutionary figures frequently employed military terminology like "struggle," "merciless struggle," "decisive battle," "army," "vanguard," "lines," "ranks," "review," and "banner" ("red") [ibid., p. 65]. Such language not only embodied the revolutionary zeal of the era but also served as an identifying marker for individuals belonging to specific social or political groups. Additionally, the period witnessed the emergence of new lexical formations, compound words, and abbreviations. Examples include "agitprop," which is shorthand for "agitation propaganda," as well as the creation of new words through suffixes like *-nik*, *-schik*, and *-yak* [14, p. 158].

Therefore, any comparative analysis of characters' oral speech in stories from this period must take into account these historical linguistic trends. The language is not just a reflection of individual traits but is also deeply rooted in the sociological and historical contexts of the time. By recognizing these factors, researchers can offer a deeper understanding of characters' language use within the broader tapestry of early 20th-century society.

III. Data Description

A. Corpus of Russian Short Stories

The annotated sub-corpus of the Corpus of Russian Short Stories, which is the basis for computational analysis in the study, comprises 310 stories from Russian authors, who wrote during the tumultuous first third of the 20th century. This period encapsulates a range of pivotal events in Russian history—from the Russo-Japanese War and the First Russian Revolution to World War I, the February and October Revolutions of 1917, the Civil War, the establishment of the Soviet state, the New Economic Policy (NEP), collectivization, and the onset of industrialization [26].

These historical developments had profound impacts not only on Russia's political and social landscape but also on its language. As outlined in [16], there was a significant turnover in vocabulary, with new words replacing 'obsolete' ones, shifts in meanings and connotations, and overarching stylistic and structural changes in accepted speech patterns.

These linguistic shifts are vividly captured in the literature of the era, in particular — in short stories. Given their shorter format and quicker "time-to-reader" publications [16], short stories are especially sensitive to societal changes, including evolving language norms. This genre thus serves as a particularly reflective mirror of its times.

The corpus is diverse in its inclusion of various authors, lending objectivity to any ensuing research. It avoids skewing towards the idiosyncrasies of any single author's style or the broader prose tendencies of the era. The study focuses on the oral speech of characters in these stories, encompassing dialogues, monologues, and isolated phrases. This spoken language acts as a living document of the linguistic dynamics of the period [26].

Therefore, analyses based on this corpus offer invaluable insights into the interplay between historical events, societal changes, and linguistic evolution as captured through the lens of literary characters. Their dialogues and monologues serve as a reflection of the linguistic zeitgeist of this complex and transformative period in Russian history.

B. Characters speech annotation

The initial step of the research involved constructing a specialized sub-corpus. This sub-corpus is designed to aggregate all character statements across the stories in the main corpus. To do this, two distinct datasets are merged.

The first dataset organizes textual elements based on the type of utterance—labeled either as 'NAR' for the author's narrative speech or 'SAY' for the characters' dialogues, phrases, and remarks. This dataset also includes metadata identifying the speaker (who is speaking) and the addresser (to whom they are speaking). Given that the research aims to conduct a comparative analysis of characters' speech, this corpus was filtered to exclusively include entries falling under the 'SAY' category, thereby focusing solely on the spoken words of the characters.

The second dataset concentrates on sociological aspects, capturing a variety of sociodemographic attributes associated with each character. The second dataset includes the sociological characteristics of the character, namely:

- Full name
- Main character (yes/no)
- Profession
- Family status
- Gender
- Age
- Social background

By combining these two datasets, the research aims to comprehensively analyze the linguistic characteristics of the characters in their historical and sociocultural context. This integrated sub-corpus allows for a nuanced study that considers not just the spoken words themselves but also the social and demographic factors influencing those words. This dual focus provides a multi-layered approach to understanding language use within the scope of the Russian short stories from the early 20th century.

To explore the impact of sociological variables on characters' oral speech, both datasets were merged into a unified database. This was facilitated by the presence of a shared variable, *code_name*, which combines the story code and the character's name and exists in both datasets. The datasets were integrated using the *inner_join* function within the R 'dplyr' package, resulting in a unified database containing both characters' speech and their sociological attributes. This integrated approach allows for a nuanced sociolinguistic analysis by enabling filtering based on various sociological variables.

Selected sociological variables for this comparative speech analysis include gender and age. The gender division in the Corpus of Russian Short Stories features 1,354 male characters compared to 480 female characters.

With regards to age, the five-tier categorization is strategically designed to be well-balanced for computational analysis:

- *under 12 years* (children),
- *12-18 years* (teenagers),
- *18-30 years* (young adults),
- *30-50 years* (middle-aged),
- *50+ years* (pre-elderly and elderly people).

With each age group featuring over 20 characters, the sample size is sufficiently large for constructing representative frequency dictionaries of their speech. This methodological approach enables a thorough and statistically robust investigation into the impact of sociological variables on the speech patterns within early 20th-century Russian stories.

It is noteworthy to mention that, while characters' ages in stories are not always explicitly stated, they can often be inferred contextually. For instance, descriptors like "child", "children", and "baby" typically point to characters in the under-12 age group, while terms like "middle-school students," "high-school students," or "teenagers" indicate the 12-18 age group. The 18-30 age group typically includes students and young adults, whereas middle-aged characters usually have their age specified directly in the text, falling within the 30-50 years range. The 50+ age group is less often explicitly defined but can generally be identified through keywords like "old man," "elderly," "old," "grandfather," or "old woman."

IV. Results

In this section, the results of the comparison of speech differentiation in terms of gender and age of the fictional characters for each sociolinguistic variable will be present in the following manner. Firstly, the average sentence length, measured as a mean total number of words per sentence, is considered. This quantitative parameter provides insights into the verbosity or brevity of characters' statements. Secondly, the frequency dictionaries of social groups, with the prime focus on relative frequencies (in terms of instances per million words, or IPM) rather than absolute numbers, are investigated. Frequency dictionaries are built for each of the social groups in the form of tables, each of which consists of three columns: *lemma*, *absolute frequency*, *relative frequency (IPM)*.

When considering the data obtained, we focus on the following questions — How does the speech of male characters differ from that of females? How does the speech of young characters differ from older ones? What other distinctive features are observed when analyzing the characters’ oral speech?

A. Gender

1) Average sentence length analysis

Table I reflects the relationship between gender and the average sentence length. Based on the means analysis, it can be noted that this parameter does not differ in dependency from the gender of the speaker. The average length of utterances for male and female characters is comparatively similar in values.

TABLE I. Relationship between the average length of an utterance and the gender of the character.

Character gender	**Average number of words in a sentence**
male	5.51
female	5.41

This data, despite the common thinking that women averagely talk more than men, corresponds with the results obtained in the course of the comparative analysis based on the oral corpus of the National Corpus of the Russian Language. It was revealed that the average length of utterances for men is higher than for women [21]. This finding suggests that oral speech of fictional characters may serve as a reliable source of linguistics data in terms of gender-determined specificities representation.

2) Frequency lists analysis

To begin with, let us consider the size of the vocabulary. The number of words in the frequency lists for subgroups is following:

- female characters' speech — 44,057 tokens,
- male characters' speech — 145,212 tokens.

Due to the wider presence of male characters, the corresponding sample is three-times larger and more divers (15,461 lexical types for males vs. 6,424 lexical types for females). With the most frequent lexis not differing much between the lists and consisting mainly from pronouns, prepositions, and conjunctions, we focus the comparison of linguistic features on the word classes that contribute to the content. We hypothesize that in gender-varied frequency lists of the characters’ speech the peculiarities found will resemble ones described in section II of this paper.

Taking into account the sociohistorical background of the stories written between the years 1900 and 1930—the Russo-Japanese War, the First Russian Revolution, the First World War, the February and October Revolutions of 1917, the Civil War—we propose that frequency dictionaries should contain the lexical examples referring to the military theme [14]. Moreover, following the findings discussed above [11], we suggest that military discourse will be more characteristic for men's speech than for women's.

As expected, most words related to military topics are found in the frequency dictionaries of male characters (note that frequencies hereinafter are given in IPM): "*voyna*" (*war*) (m: 234, f: 68), "*oficer*" (*officer*) (m: 254, f: 90), "*soldat*" (*soldier*) (m: 468 , f: 158), "*service*" (*sluzhba*) (m: 227, f: 68), "*front*" (*front*) (m: 151, f — not found), "*polkovnik*" (*colonel*) (m: 151, f — not found), "*oruzhie*" (*weapon*) (m: 110, f: 22), "*ruzh'e*" (*rifle*) (m: 185, f — not found), "*voennyj*" (*military*) (m: 96, f: 45).

Another difference between the frequency dictionaries of male and female characters that prompts interest regards the historically induced address "*tovarishch*" (*comrade*), which was widespread in the first third of the 20th century. It is found that this address is much more common in male frequency dictionaries (1535) than in women's (453). The relative frequency in male frequency dictionaries is more than three-times higher, which suggests that, despite the fact that the lemma "*tovarishch*" (*comrade*) is gender-neutral and can apply to both men and women, it is male characters who use it most often. Interestingly, terms related to revolution in general also have higher frequencies in men's speech dictionaries: "*kommunist*" (*communist*) (m: 137, f: 113), "*revolyuciya*" (*revolution*) (m: 172, f: 158).

Next, the higher frequencies of invective lexis were found in frequency dictionary of male characters, including such lemmas as "*chert*" (*devil*) (m: 1198, f: 658), "*durak*" (*fool*) (m: 557, f: 226), "*merzavets*" (*scoundrel*) (m: 165, f: 45). The frequencies of adjectives that can be classified as rude forms or insults are also higher in the frequency vocabularies of male characters "*sukin*" (*of a bitch*) (m: 117, f: 22), "*chertov*" (*damned*) (m: 123, f: 22), but exceptions make up the adjectives like "*prokliatyj*" (*cursed*), the frequency of which is higher in women's frequency dictionaries (m: 185, f: 226). An interesting feature is also a pair of offensive noun "*dura*" (feminine) – "*durak*" (masculine), the relative frequencies of which differ in the following manner: the lemma "*dura*" (feminine) is more often used by women (m: 137, f: 249), and the adjective "*durak*" (masculine) is more often used in their speech by men (m: 557, f: 226). It can be assumed that women in their speech in short stories can afford to make offensive statements specifically in relation to female characters, and men — to male characters.

At last, by analyzing frequency dictionaries, one can find nouns, adjectives and verbs that can thematically be attributed to the description of feelings. In studies devoted to the comparative analysis of male and female speech, it was elaborated that women more often express feelings and emotions in speech than men [11]. Our findings concur these suggestions; thus, the frequency of emotional lexis being used by females significantly exceeds when compared to males: "*lyubit'*" (*to love*) (m: 1356, f: 2928), "*chuvstvovat'*" (*to feel*) (m: 185, f: 249), "*lyubov'*" (*love*) (m: 351, f: 567).

B. Age

1) Average sentence length analysis

Table II displays the age groups of characters and their corresponding average sentence lengths.

The data obtained indicates distinct patterns in average sentence length across various age groups. Both children (under 12 years old) and adolescents (12-18 years old) tend to use shorter sentences compared to adults. Interestingly, the average sentence length for young children is slightly higher than for adolescents. One explanation for this could be the frequent repetition of words in children's speech in the stories, often used for emphasis. For instance, in Zinovieva-Annibal's "Wolves" (1907), a child character says, "*Fedor, a Fedor, znaesh' chto?*" ("*Fedor, and Fedor, you know what?*"), or in Shaginyan's "Agitvagon" (1923), "*Dyaden'ka, dyaden'ka, za vami soldaty prishli*" ("*Uncle, uncle, the soldiers have come for you*"). These repetitions might also serve to heighten emotional intensity, as seen in exclamatory sentences like "*Zachem, zachem*!" ("*Why, why!*") from Militsyn's "In the Forest" (1902).

TABLE II. RELATIONSHIP BETWEEN THE AVERAGE LENGTH OF AN UTTERANCE AND THE AGE OF THE CHARACTER.

Character age	Average word count in a sentence
< 12 years	4.5 words
12-18 years	4.3 words
18-30 years	5.4 words
30-50 years	5.9 words
50+	6.22 words

In contrast, the average sentence length for the 18-30 age group is notably higher than that for children and adolescents. This suggests greater linguistic diversity and complexity in the speech of young adults (18-30 years old) and middle-aged adults (30-50 years old). Moreover, the data shows a gradual increase in average sentence length as characters age, culminating in the 50+ age group, which has an average sentence length of 6.22 words. This implies that older characters tend to use longer, more complex sentences, and their narratives are more varied compared to younger characters.

2) Frequency lists analysis

The extracted oral speech size in the corresponding frequency dictionaries per speaker's age group varies in the following manner:

- up to 12 years old — 2,134 tokens,
- 12-18 years old — 4,738 tokens,
- 18-30 years old — 25,934 tokens,
- 30-50 years old — 25,762 tokens,
- 50+ years — 27,853 tokens.

The high-frequency list of words for children's speech (up to 12 years) reflects a number of features characteristic of this particular childhood age: 1) excessive address to the parental figures and relatives, 2) usage of diminutive suffixes to create new words and utilization of non-literary forms, and 3) children's babble.

First interesting feature that derives from children group's characters analysis is the prevailing usage nouns denoting parents in their speech. Notably, the words "*papa*" (*father*) (12599) and "*mama*" (*mother*) (7932) are of high rank in the corresponding frequency lists. Other relatives are also present in children's speech, e.g. "*tyotka*" (*auntie*) (1866), "*dedushka*" (*grandfather*) (1399). In addition, one can note the high variability of the forms referring to a parent. For example, to address a mother, there are found such forms as "*mamochka*" (*mommy*) (3266), "*mat'*" (*mother*) (933), "*mamachen*" (*mother*) (933), "*mamanya*" (mother) (466), "*mamasha*" (*mother*) (466), "*mamka*" (*mother*) (466) and to address a father — "*batya*" (*father*) (2799), "*papochka*" (daddy) (933), "*otec*" (father) (467), "*papasha*" (*father*) (467), "*papus'ka*" (*daddy*) (467). This emphasizes that parents play the most important role in a child’s life, being the ones children of this age communicate at most.

Another linguistic specificity that stands out is the active usage of different affixes to create different variants of word forms, including irregular ones, by children of this age group. It can be noted not only from the examples above but also in more challenging communicative situations. For example, in Kropotkina's "Polar Christmas tree" one of the younger characters recalls, "*Vsya* ***elochka*** *byla uveshena zolotymi* ***oreshkami****,* ***yablochkami****, takimi malen'kimi,* ***krasnen'kimi****, vkusnymi, vkusnymi!*" ("*The entire Christmas tree was adorned with golden nuts and apples, so tiny, red, and delicious, delicious!*"). This phenomenon is explained by the child's attempt to learn the language he speaks, since, unlike the speech of adults, children's speech has not yet been established, and it is not limited by any strict boundaries [17].

Additionally, in children's speech one can find irregular word forms. For instance, the word form "*tama*" (*there*) is not a literary norm, but it is characteristic of children's speech, which is spontaneous and unstandardized (see "*Mat', vit', umirat tama…*" ("*Mother, viti, they are dying there...*")). This tendency for morphological "inventions" correlates with studying of children's speech within ontolinguistics framework [18].

At last, the following variants of syllable repetition are observed in the speech of the children under the age of 12: "*O-o-ogo-go-o!*", "*Tya-tya... tya-tya...*", which resembles babble talk widely documented in children's speech by ontolinguists [18]. In addition, frequently noticeable, one stressed vowel is repeated in words to recreate the "drawn-out" speech of a child: "*Ma-a-am, where is the calf?*", "*Mom-ko-oo!*", "*Bah-cha-ya!*", "*Oh-oh-oh-oh!*", "*Who's eta-ah?*", "*Pa-a-pa!*" This artistic element recreates the child's oral speech; the reader seems to hear the child character pronouncing his phrases.

In the frequency dictionary of adolescents (12-18 years old), new animate nouns appear that can be attributed to the description of people. If in childhood these nouns mainly refer to parents, in the frequency dictionary of adolescents they are replaced by other of the higher frequency: "*lenin*" (*Lenin*) (4643), "*tovarishch*" (*comrade*) (1477), "*brat*" (*brother*) (1688), "*mamka*" (*mother*) (1688). In addition, obscene language and rude statements like "*svoloch'*" (*bastard*) (2743), "*chert*" (*devil*) (1266) are found. When considering verbs, one can also note colloquial rude verbs: "*zhrat'*" (*to fress*) (663), "*sblevyvat'*" (*to vomit*) (844). In the high-frequency list of nouns, a new lexis appears, characteristic of a given historical period — "*revolyuciya*" (*revolution*), "*tovarishch*" (*comrade*). The latter acquires especially high frequencies in the high-frequency vocabularies of teenagers 12-18 years old and young people 18-30 years old.

Interestingly, the change of the form of address — from "*gospodiv*" (*mister*) to "*tovarishch*" (*comrade*) [20] — is more prominent in the speech of adolescents. Compared to others, characters of 12-18 years old have the highest relative frequency of the lemma "*tovarishch*" (*comrade*) (1477). Young people aged 18-30 also have a high relative frequency (1311). At the same time, the frequency of the word "comrade" decreases sharply among adults 30-50 (426) and elderly people 50+ (394). The most popular title among older people is "*gospodin*" (*mister*) (574) which proves that older people often use a large layer of outdated vocabulary in their speech. In the speech of adults, there is no particularly clear preponderance in favor of any of the variants of address, since both variants are not high-frequency in this age group. This may suggest that young people and adolescents are being more depicted in the short stories of the "new era" and often used by authors to be the "voices of the revolution" and parts of new groups and movements (e.g. pioneer movement which was founded in 1922). As a result, their language appears to be more responsive to sociocultural changes.

Frequency dictionaries for young people (18-30 years old) seem to be more standardized than adolescent speech. High frequency nouns include words such as: "*chelovek*" (*man*) (3778), "*zhizn'*" (*life*) (2891), "*delo*" (*matter*) (1735), "bog" (*god*) (1388), "god" (year) (1388), "*tovarishch*" (*comrade*) (1311). Unlike the speech of adolescents, young people tend to use a lesser number of rude statements: the lemma "*svoloch'*" (*bastard*) (269), the lemma "*chert*" (*devil*) (1002). The speech of young people, unlike the speech of adolescents, children and the elderly, does not have any special distinctive features; it represents a language standard.

The frequency dictionary for adults (30-50 years old) is in many ways similar to the data obtained for young people. Frequency dictionaries do not contain high-frequency archaisms, colloquialisms, or profanity. High-frequency nouns for adults are "*chelovek*" (*man*) (3959), "*delo*" (*matter*) (2755), "*zhizn'*" (*life*) (1746), "*rebenok*" (*child*) (1707), "*bog*" (*god*) (1707), "*otec*" (*father*) (1436). The discussion of topics related to life, family, fatherhood may indicate that the characters are depicted as being more mature at this point of their life.

A distinctive feature of the frequency dictionary of elderly characters is that it contains many high-frequency words referring to the belief in God. Thus, the lemma "*vera*" (*faith*) is the second most popular high-frequency noun (2728). The words "*bog*" (*god*) (2225), "*batjushka*" (*priest*) (1112), "*khristos*" (*Christ*) (502), "*cerkov'*" (*church*) (503), "*gospod'*" (god) (682) are also of the high frequency. In the high-frequency list of adjectives for older people aged 50+, there is the lemma "*svyatoj*" (*saint*) (466), which also emphasizes the suggestion that the lexis referring to the topic of faith and religion appears to be more popular among older people.

The correlation between age and frequency of religious lexis is further supported by the comparative analysis of data from the frequency lists, showing that older speakers tend to use religious vocabulary more frequently. For instance, the relative frequency of the lemma "*bog*" (*god*) for speakers under the age of 12 equals to 933, at the age of 12-18 — 1266, at the age of 18-30 — 1388, at the age of 30-50 —1707, at the age of 50+ — 2225. Overall, these findings suggest that older individuals are more likely to incorporate religious language and themes into their narratives compared to younger characters or speakers.

V. Has the language changed over 100 years?

In this section, we will attempt to compare how much the frequency dictionaries of our contemporaries — men and women — differ from the statistics obtained from the Corpus of Russian short stories. As contemporary reference material, we will use data obtained for the speech of men and women from the "One Day of Speech" sound corpus of Russian everyday speech (the ORD corpus), which was recorded in 2007-2016, i.e., practically 100 years later compared to the speech of literary characters "living" in the early 20th century described in this study. The work [30] presents the upper zone of the frequency dictionary for men and women. Let's compare these data with the literary ones (see Tables III-IV).

First and foremost, it should be noted that these results allow only for a partial comparison. The reason is that the statistics for character's speech were conducted on lemmatized texts, while the statistics for everyday modern speech is based on the non-lemmatized ones. Therefore, a comparison of only unchangeable words (particles, conjunctions, prepositions), which, however, predominate in the upper zone of the frequency dictionary presented in the tables, would be correct. This partially explains the higher volume of the relative frequency indicator (IPM) for such frequent words as "*ya*" (*I*), "*ty*" (*you*), "*on*" (*he*), "*ona*" (*she*) and especially "*byt'*" (*be*).

In general, the tables show that the differences between the speech of male and female characters in the upper zone of the frequency dictionary are less pronounced than in the everyday spoken language of our contemporaries. One explanation for this fact is that the spoken language of characters still turns out to be more literary than authentic spontaneous speech and does not contain certain discourse vocabulary; moreover, the share of discursive words in real spoken language is higher than in the speech of literary characters (these elements are marked with an asterisk *). Furthermore, the lexical composition of the language's dynamics over 100 years cannot be disregarded. But whether and how to separate these factors is a methodological question that deserves special consideration.

Thus, to the question posed in the title of this section, we cannot yet give a definitive answer based on the data obtained. Moreover, based on the research results, it would be logical to assume that the share of discursive words in the speech of characters will be significantly lower than in authentic oral speech, which abounds in "irregularities", discursive and pragmatic markers. It would be more appropriate to compare the speech of characters with the texts of interviews — also reflecting spontaneous speech but subjected to editorial processing.

VI. Conclusion

In the presented research, a comparative analysis of the oral speech of characters in the first third of the 20th century was carried out, which revealed the speech characteristics of the characters according to such parameters as age and gender.

It is these two parameters that provide the most material for research. As a result, the frequency dictionaries of Russian short stories character's speech were obtained and mean length of sentences analysis was conducted.

TABLE III. The upper zone of frequency words for women's speech.

Literary texts, early 20th century			Contemporary speech, early 21th c.		
Rank	**Word**	**IPM**	**Rank**	**Word**	**IPM**
1	ya	41947	1	ya	27000
2	ne	31029	2	ne	23600
3	i	29349	3	vot*	23400
4	ty	25649	4	da*	22500
5	chto	21246	5	nu*	22400
6	a	18885	6	chto	19300
7	vy	17387	7	a	18900
8	byt'	14413	8	i	18400
9	v	14141	9	eto	16700
10	on	11622	10	v	15400
11	eto	10986	11	tak	14100
12	na	10623	12	u	13600
13	da	10396	13	tam*	12900
14	to	10192	14	na	10200
15	kak	9647	15	kak	9900
16	s	9261	16	ty	8700
17	u	8512	17	vsyo	8100
18	tak	8194	18	s	7900
19	ona	7944	19	ugu	7300
20	zhe	7377	20	to	7200
21	vot	7059	21	net	6900
22	my	6991	22	ona	6800
23	nu	6968	23	mne	6500
24	vse	6446	24	on	6400
25	moy	5765	25	(e)	5900

TABLE IV. The upper zone of frequency words for men's speech.

Literary texts, early 20th century			Contemporary speech, early 21th c.		
Rank	**Word**	**IPM**	**Rank**	**Word**	**IPM**
1	ya	34569	1	nu*	24700
2	ne	28820	2	ya	24000
3	i	27650	3	ne	22800
4	a	20627	4	vot*	22600
5	v	18789	5	v	20000
6	chto	17970	6	da*	19300
7	ty	17026	7	a	18600
8	vy	14630	8	i	18400
9	byt'	13667	9	tam*	17500
10	na	11794	10	eto	16000
11	eto	10018	11	chto	16000
12	to	9487	12	na	13300
13	s	9391	13	u	11800
14	kak	8868	14	tak	10400
15	da	8696	15	(e)	9100
16	on	8241	16	on	8300
17	u	8062	17	to	8200
18	my	7904	18	vsyo	8000
19	tak	6795	19	kak	7900
20	nu	6733	20	ty	7800
21	vot	6733	21	s	7300
22	ona	6589	22	net	6400
23	zhe	6059	23	est'	5700
24	vse	5129	24	bl**d'*	5100
25	moy	5765	25	seychas*	4900

The initial hypothesis that the data obtained could be used to study the sociolinguistic variability of everyday speech over time was only partially confirmed. On one hand, indeed, the obtained frequency dictionaries vividly characterize the lexical features of significant vocabulary used by a particular social group. On the other hand, comparing these frequency lists with that of the real oral speech shows that in terms of the use of hesitations, pragmatic, and discursive markers, the speech of characters appears more "correct", or more "literaturized". The approximate scale of such adaptation can be assessed by comparing contemporary speech and the speech of contemporary literary characters. Then it will be possible to talk about the possibility of modeling the speech features of past epochs based on the speech of literary characters. Although to some extent, this model will still have a probabilistic nature.

Despite the preliminary results of the conducted research, its contribution is twofold: firstly, in pioneering a methodology for investigating sociolinguistic variability through literary analysis; secondly, in laying the groundwork for a theoretical model of dynamic sociolinguistic variability. These findings not only enhance our understanding of historical speech patterns but also aid in forecasting sociolinguistic trends, thereby informing the development of future speech technology applications tailored to evolving language use.

Acknowledgment

The publication was prepared in the framework of the project “Literature and Society: Digital Platform SOCIOLIT”, supported by the Scientific Fund program of the National Research University Higher School of Economics in 2024.

References

[1] I. A. Baudouin de Courtenay, ‘On the mixed nature of all language’, *Selected works on general linguistics, vol. I, I. A. Baudouin de Courtenay*, 1963.

[2] N.B. Vakhtin, ‘Sociolinguistics and sociology of language’, *Publishing house of the European University*, 2012.

[3] E. A. Naiman, ‘Sociolinguistics: a course of lectures’, 2004.

[4] V.I. Karasik et al., ‘The language of social status’, ITDGK "Gnosis", 2002.

[5] L.P. Krysin, ‘Speech communication and social roles of speakers’, *Socio-linguistic studies*, 1976.

[6] Firebox L.V., ‘Linguistic marking of the social status of the speaking subject in languages of different types’, *Sociosphere*, 2015.

[7] Merchant K., ‘How men and women differ: Gender differences in communication styles, influence tactics, and leadership styles’, 2012.

[8] J. C. Pearson, ‘Gender and communication’, *Dubuque, I A: William C. Brown,* 1985.

[9] B. Thorne, N. Henley, ‘Difference and dominance: An overview of language, gender, and society’, *B. Thorne & N. Henley (Eds.), Language and sex: Difference and dominance, Rowley,* 1975.

[10] R. T. Lakoff. Language and woman's place: Text and commentaries, Oxford University Press, USA, 2004, vol. 3.

[11] A.V. Kirilina, M. Tomskaya, ‘Linguistic gender studies’, *Otechestvennye zapiski*, 2, 2005.

[12] T.V. Gomon, ‘Study of documents with deformed internal structure’, *Diss. Ph.D. legal Sciences*, 1990.

[13] V.V. Khimik, ‘The language of modern youth // Modern Russian speech: state and functioning: Collection of analytical materials’, St. Petersburg: Faculty of Philology of St. Petersburg State University, 2004, pp. 7-66.

[14] A. M. Selishchev, ‘Language of the revolutionary era’. URSS, 2003.

[15] T. Sherstinova, G. Martynenko, ‘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’, R. Piotrowski's Readings in Language Engineering and Applied Linguistics, *Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019)*, 2019, pp. 105-120.

[16] G.Ya. Martynenko, T.Yu. Sherstinova, T.I. Popova, A.G. Melnik, E.V. Zamirajlova, ‘On the principles of creation of corpus of Russian short stories of the first third of the 20th century’, *Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics ‘TEL 2018’*, Kazan, 2018, pp. 180–197.

[17] A.K. Vaganova. ‘Functions of children’s occasional names with specifically used affixes’, *Scientific thought of the Caucasus*, 2013, 1 (73).

[18] A. N. Gvozdev, ‘From first words to first grade’, 2005.

[19] L. A. Berbeshkina, N. V. Chernikova, ‘Addresses in Russian etiquette XX-XXI (history of addresses citizen and comrade)’, *Current problems of education and upbringing: integration of theory and practice*, 2019, pp. 252-255.

[20] M.A. Korotkevic, ‘Forms of addressing a stranger in modern Russian speech etiquette’, *Theory and practice of linguistic communication*, 2016, pp. 161-165.

[21] Yu. G. Zelenkov, ‘NKRY as a tool for sociolinguistic research. Episode IV. Speaker's gender and utterance length’.

[22] F.P. Sorokoletov, ‘History of military vocabulary in the Russian language (XI-XVII centuries)’, *Limited Liability Company Book House Librocom*, 2009.

[23] B. L. Boyko, ‘Military vocabulary in speech communication’, *Issues of psycholinguistics*, 25, 2015.

[24] B.B. Ayusheev, ‘Lexico-semantic and structural characteristics of soldier’s jargon’, ‘World of Science. Sociology, philology, cultural studies’, 2019, vol. 10 (4), pp. 31-31.

[25] G. Ya. Martynenko, T. Yu. Sherstinova, A.G. Melnik, T.I. Popova, ‘Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era of revolutionary changes (first third of the 20th century)’, *Computational linguistics and computational ontologies*, 2, ITMO University, St. Petersburg, 2018, pp. 99–104.

[26] G.Ya. Martynenko, T.Yu. Sherstinova, ‘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’, *CEUR Workshop Proceedings*, vol. 2552, 2020, pp. 105–120. http://ceur-ws.org/Vol-2552/

[27] Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., & O’Donnell, M. B., ‘CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, 2016, 11(3). Pp. 433–463.

[28] D. Tannen, ‘Gender differences in topical coherence: Creating involvement in best friends' talk’, *Discourse processes*, 1990, vol. 13 (1), pp. 73-90.

[29] E. V. Ataeva, ‘On some areas of research into the language of youth: towards the formulation of the problem’, *Bulletin of the Faculty of Humanities of ISUTU*, 4, 2009, pp. 204-209.

[30] T. Sherstinova, ‘Speech acts annotation of everyday conversations in the ORD corpus of spoken Russian’, *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings 18, Springer International Publishing*, 2016, pp. 627-635.