ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Терехина Мария Дмитриевна

Исследование скорости речи на материале корпусов звучащей речи языков России

(Speech Rate Study Based on Data from Spoken Corpora of Languages of Russia)

Выпускная квалификационная работа студентки 4 курса бакалавриата группы БКЛ-152

Академический руководитель	Научный руководитель
канд. филологических наук,	Старший преподаватель
доц. Ю.А. Ландер	Г.А. Мороз
___________________________	_____________________________
« » __________ 2019 г.	

Москва 2019

CONTENTS

1. Introduction

Speech rate is usually defined as ‘the rate at which phonetic events occur over time’ (Quené, 2008, p. 1104). It is most often measured in syllables or phonemes per second. In the middle of the 20th century, studies on speech rate began to appear. The first of them were dedicated to individual differences in ‘the speed of talking’ both within speakers and across them (Frieda Goldman-Eisler, 1954). Some studies found a dependency between the structure of an utterance and its speed of pronunciation. For example, it was found that the longer the phrase, the faster it will be pronounced (Fonagy & Magdics, 1960; Stepanova, 2011).

Later, works, that were focused on the more general analysis of the speech rate came out. Some of them were dedicated to the analysis of speech rate of a language as a whole (for example, see Pimsleur, Hancock, & Furey, 1977; Tauroza & Allison, 1990 about English; Stepanova, 2011 about Russian; Zellner, 1998 about French), there were even cross-linguistic studies on speech rate (for example, Barik, 1977; Ding et al., 2017).

At the same time, there is a wide range of studies showing that the difference in speech rate that we *perceive* may be based not on a real difference in the pronunciation tempo, but on the individual characteristics of a hearer, such as a hearer’s own speech rate (Bosker, 2016), a rhythmical class (‘stress-timed’, ‘syllable-timed’)[1] of a language being perceived, the native language of a speaker (Abercrombie, 1967), and some others.

This *perceptual* difference has given rise to the idea, that some languages are spoken faster than others. As a result, the mentioned above studies investigating the overall language speech rate occurred. These studies strengthen the concept that language speech rate can be measured and represented as a single average number and that languages can be typologically compared by this parameter. But, taking into account that there are individual differences influencing the speech rate of a speaker under different conditions, it may be irrelevant to compare overall language speech rates because of a great variability within a language and even within a speaker.

[1] ‘...it usually seems that syllable-timed speech sounds faster than stress-timed to speakers of stress-timed languages’ (Roach, 1998, p. 153).

In this work, I will try to find out whether there is a significant difference in the speech rates of speakers of several languages of Russia, to what extent does a language influence the speech rate of the speakers, and therefore verify the validity of language speech rate parameter and its comparison between languages. The **hypothesis** is that the language speech rate, as a language-specific parameter, has a minor influence on individual speech rates of a particular language speakers.

To the best of my knowledge, the question of the validity of language speech rate estimation was never studied before, as well as there are no cross-linguistic studies on speech rate, based on the data from languages of Russia.

To testify the hypothesis, I use spoken corpora of **Russian** (Corpus of Rogovatka dialect (Ter-Avanesova et al., 2018), Ustja River Basin Corpus (Daniel, Dobrushina, & von Waldenfels, 2018)), **Azeri** (Corpus of Qakh Dialect of the Azeri Language (Linguistic Convergence Laboratory, n.d.)), **Bashkir** (Spoken corpora of the Bashkir language (Ovsyannikova, Say, Aplonova, Smetina, & Sokur, 2017)), **Beserman dialect of Udmurt** ('The Spoken Corpus of the Beserman Dialect of the Udmurt Language', 2018), and **Chukchi** ('The Multimedia Corpus of the Chukchi Language', 2018) languages. The data from the corpora is pre-divided into sentences or words by corpora compilers with the corresponding beginning and ending times. I base my analysis on this mark-up. In this work, the speech rate is measured in syllables per second. That is, to get the speech rate of an utterance, I calculate a number of syllables in it and divide it by the utterance's duration in seconds. The choice of the measurement units will be discussed further (see the Literature review section).

One observation in my sample equals to one utterance. As soon as there is always more than one utterance pronounced by one and the same speaker, the observations are not independent. As observations are not independent, it is impossible to use regular regression for the data analysis, because the independence of residuals (errors) is one of the fundamental assumptions of the regression analysis (Draper & Smith, 1998, p. 61), and if the observations are not independent, the residuals are not independent too.

In the case when independence residuals assumption is violated, Multilevel Mixed-Effects Models (Bryk & Raudenbush, 1988) are used for the data analysis. These models allow representing nested data: the data, in which values of some independent variable group within another independent variable (for more details see the Multilevel Mixed-Effects Models chapter in the Methods section). These models suit well for the present analysis as observations are nested within participants and participants are nested within languages.

To define, whether a language is an important parameter, influencing the speech rate of a person, I design several models, some of which do include language parameter and others do not, and further compare them using statistical tests. If the models with language parameter describe the data better than the models without it, it is possible to state, that language is an important factor, influencing a person's speech rate.

The main difficulty with such an approach is that the common statistical tests (which are frequentist) are either allow rejecting the null hypothesis or do not allow to do so. They do not provide any evidence to state that the null hypothesis is correct. To find out whether the null hypothesis is correct, other techniques for model selection are applied, such as, among others, the Akaike information criterion and Bayes factor (Burnham & Anderson, 2002).

Bayes factor provides evidence to support some hypothesis against another by comparison of Bayesian statistical models, representing the corresponding hypotheses (Kass & Raftery, 1995). To perform both frequentist and Bayesian hypothesis testing and model selection, I implement both frequentist and Bayesian statistical models (for more details on their difference, see the Frequentist vs. Bayesian approach to statistics chapter in the Methods section). In accordance with the hypothesis, I expect model without the language parameter to describe the data at least not worse than the models, that do take language into account.

2. Literature review

When discussing speech rate, we can consider the two principal levels: **intra-speaker** (at this level, the difference in speech rate of one speaker is analysed) and **inter-speaker** (at this level, the difference in speech rate between different speakers is analysed). While there are numerous psycholinguistic studies focused on intra-speaker variation (which will be discussed the first) in speech rate (for example, see Brown, Giles, & Thakerar, 1985; Fridea Goldman-Eisler, 1954), there is still no uniform opinion on whether the differences at this level are significant or not.

2.1. Intra-speaker speech rate variation

The Goldman-Eisler's (1954) classic paper is the first well-known experiment aimed at calculating intra-speaker variation speech rate. The second goal of this work was to find out whether there is dependence between the length of an utterance and the speech rate. The author studied four adults, three men and one woman, who were the academic and scientific staff of Maudsley Hospital, and four young neurotics (patients of the hospital), all female. The patients were interviewed separately by the author of the paper. The staff-participants were talking with each other separately and in groups. The author recorded the conversations and then counted the speech rate as a number of syllables per minute. It was found that the speech rate discriminates significantly between staff-participants in all lengths of utterance ($p = 0.001$). For patients, discrimination was significant in all but the shortest lengths of utterance. It was also found, that communicating with different interlocutors, staff-participants talk with significantly different speech rates ($p < 0.05$). Table 1 summarizes the data of this part of the experiment.

Table 1. Results of analysis of variance testing the differences in speech rates in different conversational situations for staff-participants (Goldman-Eisler, 1954, p. 98).

Subject	Conversation	Means	S.D.	F-ratio	P
A	With B	216.7	57.5	4.8	0.01
	With C	254.5	58.3		
	With D	206.0	80.4		
	With B, C	209.5	23.2		

	With B, C, D	244.8	24.4		
B	With A With C With D With A, C With A, C, D	206.1 193.8 252.2 237.2 275.7	31.3 36.5 56.9 52.5 42.9	16.3	0.001
C	With A With B With D With A, B With A, B, D	181.0 186.6 187.5 206.2 214.5	46.1 29.1 48.5 96.4 39.6	1.5	N.S.
D	With A With B With C With A, B, C	244.8 273.0 272.7 274.1	52.1 44.3 37.3 41.3	3.18	0.05

Miller, Grosjean, and Lomanto (1984) showed that speech rate variations at the intra-speaker level are significant even if we look at a single speech event both within and across speakers. They have studied 30 people, who were participants of ‘The World at One’ BBC programme. The average standard deviation for the speakers’ average syllable duration was 67 ms with an average of 216 ms, which results in the variation of 31%. During the analysis of single question response, they have also found a considerable variation: the average difference in rate between fastest and slowest run[2] in a single response was 309 ms for the speaker with the highest variation and 169 ms for the speaker with the lowest variation. On the other hand, Grosjean (1980) in his early study detected almost no variance at the intra-speaker level: the rate across individual participant was nearly constant.

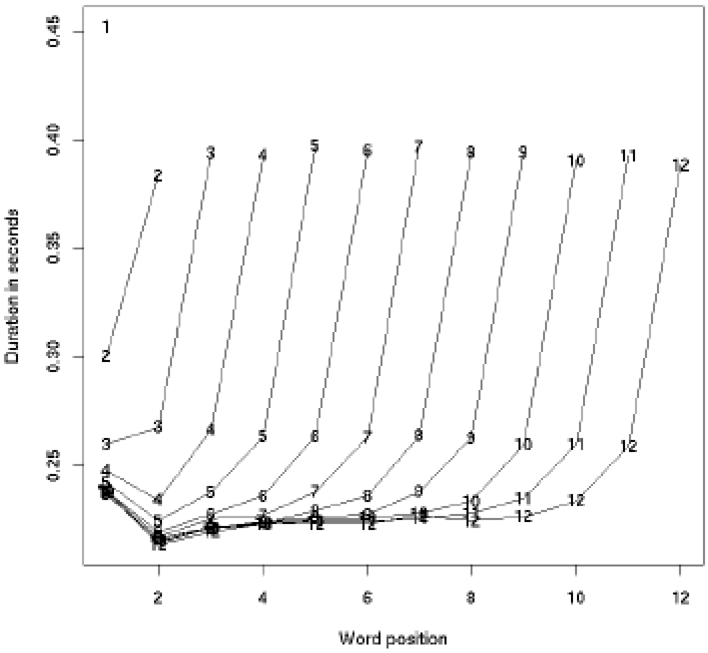
In subsequent studies, researchers tried to find out how the speech rate relates to

[2] Speech run is a pause-free stretch of speech, where pause is a salient interval of 250 ms or more (Miller, Grosjean, & Lomanto, 1984, p. 219).

the syntagmatic division of a sentence, but the results were also controversial. For example, Deese (1984), reported that the average speech rate of conversational speech is from 5 to 6 syllables per second and that speakers tend to speed up by the end of an utterance.

By contrast, Yuan, Lieberman, and Cieri (2006) showed that speakers tend to slow down by the end of an utterance, that is, the last word is the slowest in an utterance. In this study, corpora of telephone conversations of English and Chinese were used. The analysis was performed only for sentences consisting of 12 or words or less. The results are summarized in Plot 1.

Plot 1. Mean word duration by its position in a sentence (Yuan et al., 2006, p. 2).



So, why are the results so conflicting? The reason may lie in the methodological differences in evaluating the intra-speaker speech rate. The first difference, called 'the window effect,' is the choice of the domain within which speech rate variations will be calculated (Cedergren & Perreault, 1994, p. 1087). For example, in (Miller et al., 1984, p. 222) it was argued, that 'articulation rate typically has been measured over large stretches of speech, such that the local variation characterized by the peaks and troughs … is neutralized.' That is, the longer speech unit is estimated, the slighter variation in speech rate will be observed because we will compare the average values of long utterances and local deviations will be missed there. However, the authors assert that even if they analyse an entire run of pause-free speech, the speech articulation tempo steel varies considerably (Miller et al., 1984, p. 222). The experiment details were described above.

In (Cedergren & Perreault, 1994) it is suggested to use syllable timing (one syllable pronunciation time) to constraint this issue, but here we will face the problem of syllable division: duration of syllables will depend on a set of rules we choose for syllables division.

For example, for the Russian language, there are several theories of syllable division. The most recognized one is the sonoristic theory of syllables based on acoustic criteria. With reference to the Russian language, it was developed by Avanesov (1956). But there also other theories, such as Scherba's stress-based theory (Zinder & Maslov, 1982) and sometimes they provide different results, which means that the syllable structure of a word depends on the theory we choose.

Finally, it is also possible to base the speech rate calculation on the number of sounds pronounced in a period of time, as was suggested in (Roach, 1998). Explaining this idea, he claims, that languages with simple syllable structure (such as Japanese) can fit more syllables in a time period than languages with complex syllable structure (such as English or Polish), so while compared, the former will be treated as faster and the later as slower. So, if we compare the number of sounds, we will avoid this issue. But the author immediately says, that 'the faster we speak, the more sounds we leave out' (Roach, 1998, p. 152). The omission of sound segments is known to reach 20-22% (Greenberg, 1999; Johnson, 2004), while syllable omission, by different estimations, fluctuates from 1% (Greenberg, 1999) to 5.1% (Johnson, 2004), which is appreciably lower.

The second methodological difference that may give rise to mixed results is the so-called 'measurement issue.' This term describes the difference in approaches to what is exactly considered a variation in the speech rate. For example, we can calculate an average syllable duration over a stretch of speech (as, for example, in Crystal & House, 1990), or we can build a language model and then calculate the difference between the predicted and the observed speech rates (as in Campbell, 1992), or we can encode the speech rate effects by computing a tempo adjustment metric (as in Van Santen, 1992;

Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). While the first method is the most simple and obvious, the last two seem to be more complex, but all of them are valid.

Finally, the effects of some unobvious independent variables can also be the cause of such contradictory results. One group of such effects is interviewer and co-participant (interlocutor) effects. In (Kendall, 2009), a range of these effects was investigated. He found that subjects, who were interviewed by female interviewers only, had significantly lower speech rate than participants, interviewed by male only or mixed interviewers. Importantly, the difference in speech rate between male and female subjects interviewed by females was not significant, which means that the only important factor here is interviewer's gender. This factor is of such significance, that as soon as it is included in the model, other factors, that were significant earlier (region, median pause duration, gender, and ethnicity) lose their significance (Kendall, 2009, pp. 183–184). It was also found, that speakers tend to speak more slowly with people of different ethnicity (Kendall, 2009, p. 187).

In (Yuan et al., 2006), in addition, it was discovered that people normally use longer segments (the length was counted in words and characters) but pronounce them with lower speech rates when talking with strangers comparing to friends and family members. This finding contradicts with the 'anticipatory shortening' effect, stating that longer phrases should be pronounced faster. Yuan, Liberman, & Cieri explain that it may be an effect of politeness that people show while interacting with strangers (Yuan et al., 2006, p. 3). Thus, even well-known principles of speech rate organization may be violated when interviewer and co-participants effects interfere.

2.2. Inter-speaker speech rate variation

Let us now consider the inter-speaker level, which is usually viewed from the sociolinguistic perspective. Quite interestingly, at this level, there is also no consensus about whether the variation of speech rate is significant or not. For example, while Goldman-Eisler (1968), shows that native English speakers demonstrate considerable variation in their overall speech rates, Deese (1984) declares the opposite: 'few native-born speakers of the standard dialect of English vary a lot in their rate of speaking' (Deese, 1984, p. 105).

According to (Kendall, 2009), one reason for such disagreement may be the strong correlation between the speech rate and the length of an utterance. As was shown in (Quené, 2008), the more syllables an utterance contains, the faster they will be pronounced (this phenomenon is called ‘anticipatory shortening’). This effect is of such significance, that if we include the length of an utterance as an intra-speaker factor in a mixed-effect model, then inter-speaker effects of age and gender become insignificant.

The second reason of such contradictory results of evaluation of intra-speaker speech rate variation, according to (Kendall, 2009), is that what in early studies was described as ‘a very small range of variation (4.4 to 5.9 syllables per second)’ (Goldman-Eisler, 1961, p. 171), may actually be not so small. To clarify this issue, we should understand what difference in the speech rate is perceptible to hearers. As was stated by (Quené, 2008), listeners percept speech rate difference at the level of about 5%. That is, if we take the average speech rate for 5 syllables per second (something between 4.4 and 5.9), then 5% of this will be 0.25 syllables per second. Thus, the variation of 1.5 syllables per second is quite noticeable to listeners and cannot be considered as ‘a very small range of variation’ (Goldman-Eisler, 1961, p. 171).

If we address inter-speaker variation in speech rate from the sociolinguistic angle, we will probably expect, that there will be almost no variation if we examine speakers with the same sociolinguistic parameters. In (Kendall, 2009) one exemplary case was studied. He examined the speech rate of 12 children aged from 12 to 17, interviewed by one person, whose name was Clarisa. She conducted an ethnographic study; the children were its participants. All of them were of the same ethnicity and from the same region (Washington, DC). As Clarisa stated, all of them had a similar educational and socio-economic background. There were two boys and ten girls. Despite all the similar characteristics and the fact, that they were interviewed by one and the same person, the children varied significantly in their speech rate (ANOVA p-value < 0.001) (Kendall, 2009, pp. 194–199). Plot 2 represents the experiment results.

Plot 2. Results of the ‘12 children’ experiment (Kendall, 2009, p. 198).

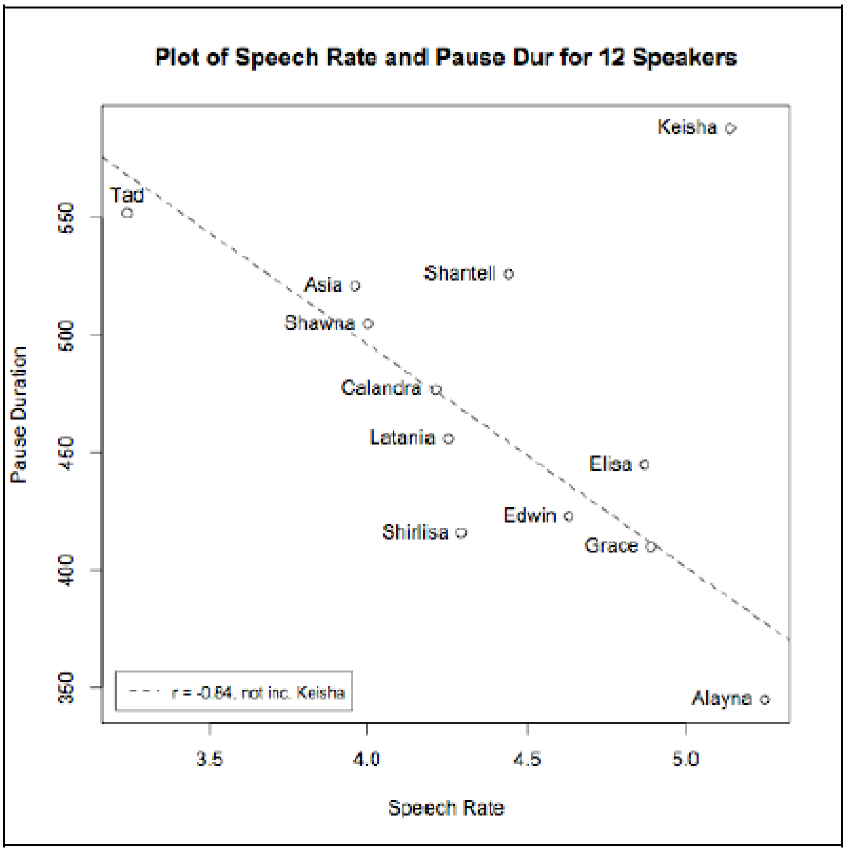


Figure 9.5.2. Plot of speech rate and pause duration for Washington, DC interviewees

Consequently, we can argue that even equality in sociolinguistic parameters does not guarantee equality in speech rate. Thus, in this case, we can explain the difference in the speech rate only by postulating it an individual characteristic of a speaker, which can be influenced by various parameters.

But to what extent can speech rate be influenced? In the same work, Kendall (2009) studied the data from the experiment with the 12 children from another perspective: from Clarisa’s. He found that Clarisa accommodates to her interlocutor’s speech rate and the rank orders of her speech rate measures correlate with her interviewers rank orders (r^2= 0.75), but this is not the case for values of the measures (r^2= 0.19). That is, an interlocutor’s pace of speech does influence a speaker’s speech rate, but not dramatically. Plot 3 represents the change in Clarisa’s speech rate.

Plot 3. Change in Clarisa’s speech rate (Kendall, 2009, p. 202).

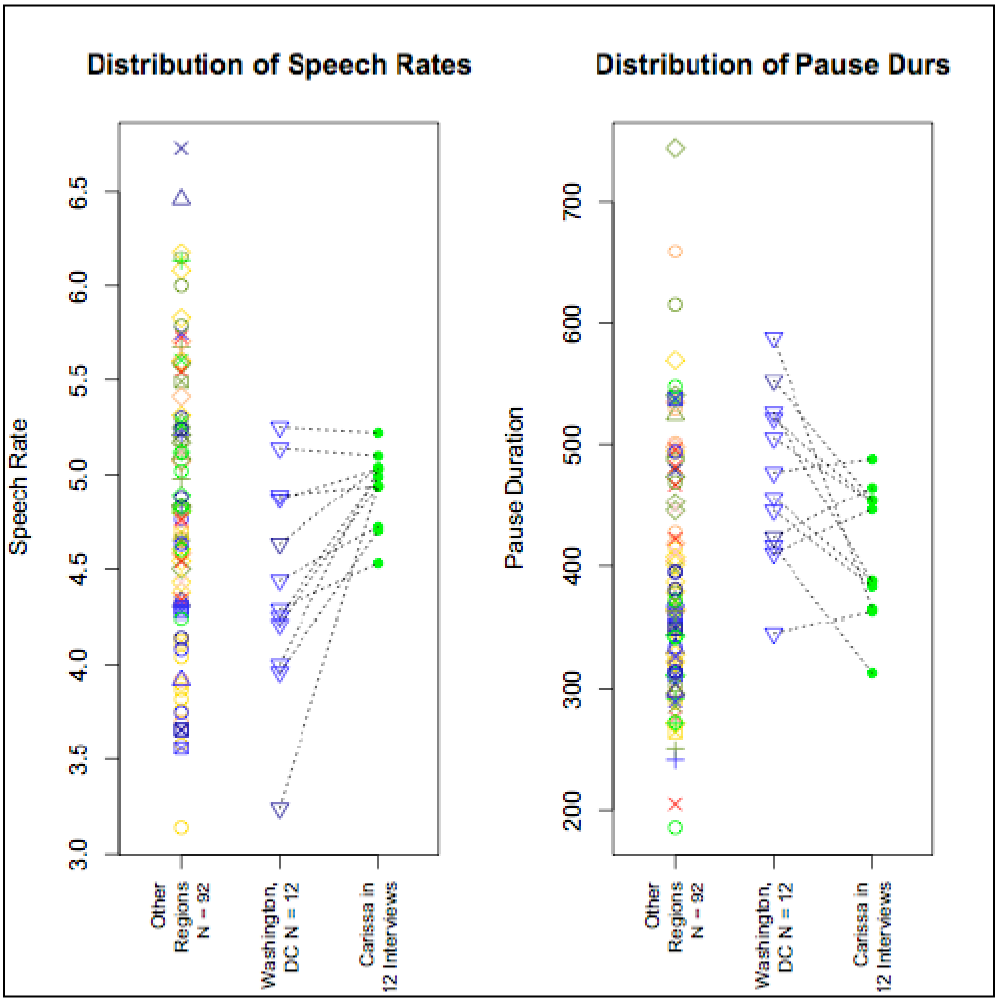


Figure 9.5.4. Distributions of speech rate and pause data, including Carissa

Despite the numerous works showing that speech rate discriminates between and within speakers, there are studies considering temporal characteristics of speech as a universal parameter for all speakers of a language. The main part of them consists of studies estimating the average language speech rate. For example, in (Tauroza & Allison, 1990) researchers studied average speech rate of 5 types of native English speech: radio monologues, conversations, interviews, lectures, and combined. From the first four categories, 30-minutes and from the fifth 120-minutes samples were chosen. Obtained results were compared with the results of (Pimsleur et al., 1977) experiment. It turned out, that in all the speech types, Tauroza & Allison got considerably lower speech rate than Pimsleur, Hancock, & Furey’s great average. The results summarized in Table 2.

Table 2. Mean number of words per minute (w.p.m.), syllables per minute (s.p.m.), and syllables per word (s.p.w.) in the different categories of speech (Tauroza & Allison, 1990, p. 97).

Category	w.p.m.	s.p.m.	s.p.w
Radio	160	259	1.6
Conversation	210	260	1.3
Interview	190	250	1.3
Lecture	140	190	1.4
Combined	170	240	1.4
Pimsleur et al.	180	300	1.7

In (Stepanova, 2011) the average rate of Russian speech is estimated based on data from Speech Corpus of Russian Everyday Interaction ‘One Day of Speech’ (ORD Corpus). Stepanova says: ‘It is assumed that there is an average generally accepted tempo in each linguistic community, and deviations from this rate may be attributed to numerous factors: the speaker’s age, gender, education, and so on’ (Stepanova, 2011, p. 1902). In other words, it is stated, that language speech rate is a single-value parameter of a language and the deviations from it should be explained by sociolinguistic parameters. But as we have seen before, in the (Kendall, 2009), it was found, that even within a group with similar sociolinguistic parameters there is significant variation in speech rate.

2.3. Cross-linguistic studies

The next level of generalization is represented by cross-linguistic studies. In (Barik, 1977) a cross-linguistic analysis of different speech types of English and French is presented. He studied spontaneous and semi-prepared speech, prepared or formal speech for both oral and for written delivery. For each type, a passage from only one speaker was recorded for each language. The content of passages was comparable: for spontaneous speech – film discussion, for semi-prepared speech – ‘live’ lecture recordings, both on a general theme, for oral prepared – ‘formal non-technical speech in

English' and its oral translation from French interlocutor, for written prepared – reading of a 'not very technical' article, available both in English and French. For both languages, two adult male speakers were involved, recording different types of speech. The part of results, related to speech and articulation rate, is presented in Table 3.

Table 3. Data for English and French texts (Barik, 1977, p. 120).

	English texts					French texts				
	Story	Film	Lecture	Speech	Written text	Story	Film	Lecture	Speech	Written text
Speech rate (syll./min.)	149.9	194.1	189.7	155.6	193.2	131.8	174.0	209.1	183.3	204.8
Articulation rate (syll./min.)	292.3	273.0	252.7	280.0	275.0	303.1	305.9	256.1	272.2	285.8

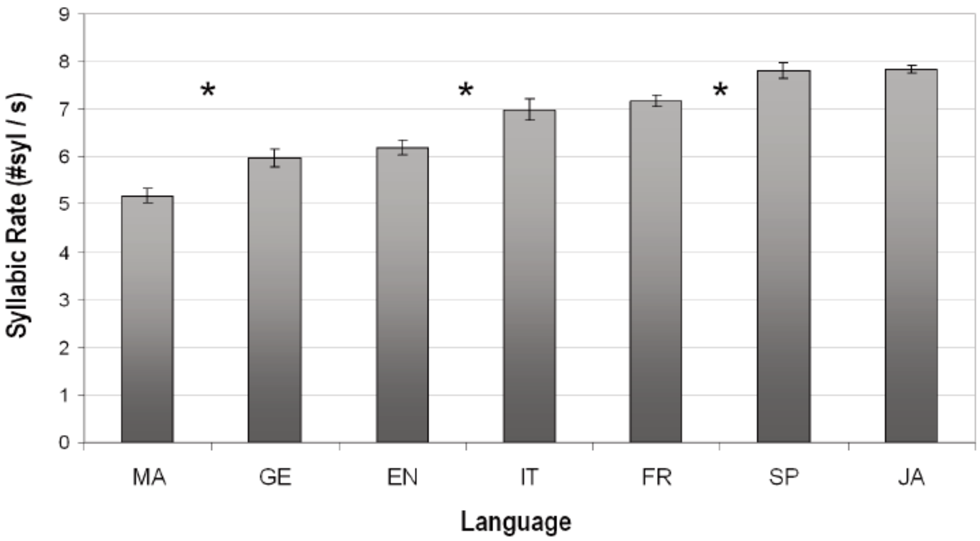
On the base of these results, the author concludes, that 'articulation rate is similar across languages in relation to each type of material separately, and is a relatively constant speech parameter' (Barik, 1977, p. 116). But if we calculate the articulation rate in syllables per second for FILM category, we get 4.55 and 5.09 syll./sec for English and French accordingly. If we then calculate 5% of the English rate we get 0.23 syll./sec and 4.55 + 0.23 = 4.78 syll./sec, which is less than 5.09. That is, this difference in speech rate is quite considerable for a hearer[3]. If we consider speech rate values from Table 3 above, we will get an even greater difference, which refutes the author's conclusions about the cross-linguistic similarity of this parameter.

The paper by Pellegrino, Coupé, & Marsico (2011) is also a cross-linguistic study of speech rate and, more generally, the rate of information transmission. The goal of the study was to find out, whether there are differences in information transmission rate across seven languages: French, German, Italian, Japanese, Mandarin Chinese and Spanish. The total of 20 texts was collected from MULTEXT multilingual corpus (Campione & Véronis, 1998). The original texts in British English and were then translated into other languages. The texts were then recorded by several native speakers.

[3] In (Quené, 2008), it was found, that hearer percept 5% difference in the speech rate.

The main effect of the Language parameter was found, except between English and German, French and Italian, and Japanese and Spanish. The results are presented in Figure 1.

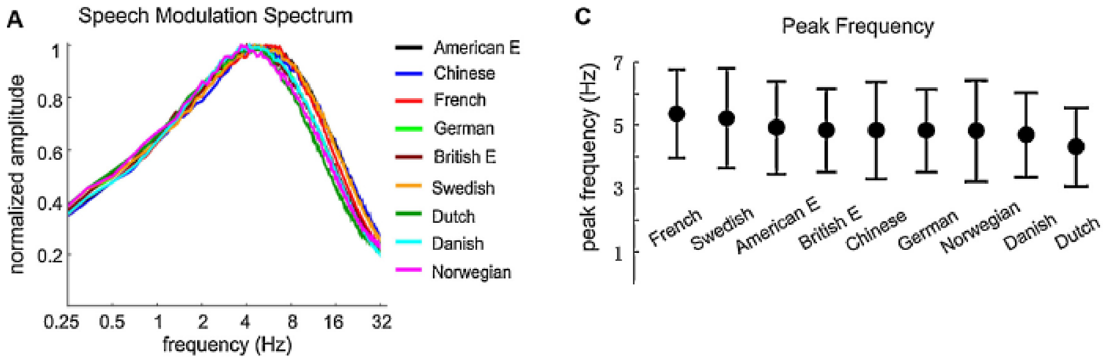
Figure 1. Speech rate of the studied languages (Pellegrino et al., 2011, p. 546).



On the other hand, Text and Speaker parameters were also of high significance (for both $p < 0.0001$) (Pellegrino et al., 2011, p. 546). It means, that there is considerable variation in speech rate on the inter-speaker (Speaker) and intra-speaker (Text) levels.

In recent years, works, comparing not only languages between themselves by temporal characteristics came out. For instance, Ding et al. (2017) studied slow temporal modulation structure of nine languages: Chinese, English (British and American), German, Swedish, Dutch, Danish, Norwegian and French, and compared it with slow temporal modulation structure of music. Temporal modulation reflects ‘how sound intensity fluctuates over time’ and is ‘a primary acoustic correlate of perceived rhythm’ (Ding et al., 2017, p. 182). Temporal modulations below 16 Hz are known to be caused by syllabic rhythm (Goswami & Leong, 2013; Greenberg, Carvey, Hitchcock, & Chang, 2003). It was found, that the speech modulation spectrum is of high consistency for all the nine languages and has a peak between 4 and 5 Hz. The results are presented in Figure 2.

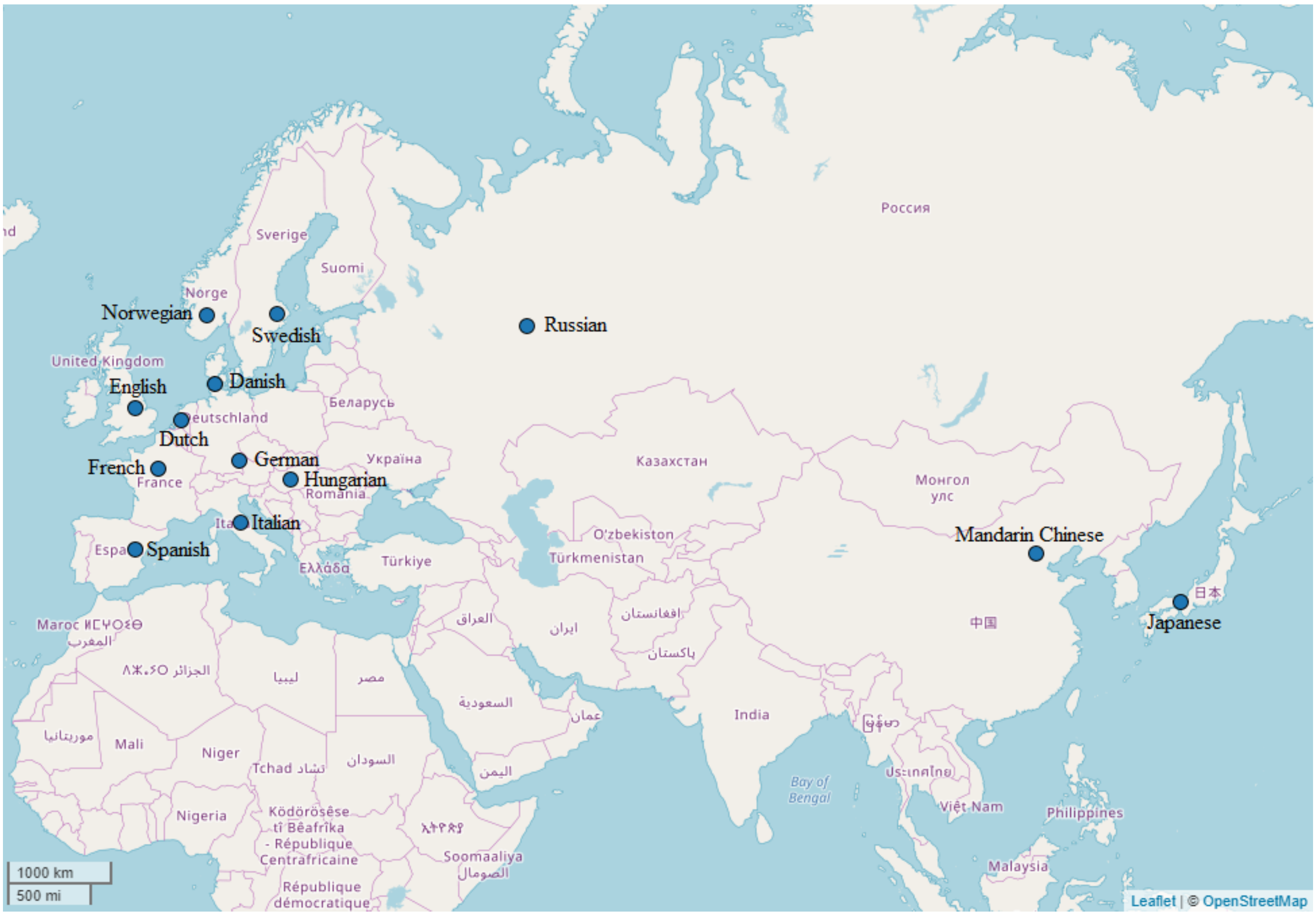
Figure 2. The modulation spectrum of speech (Ding et al., 2017, p. 184).



Authors compare this consistency with 5-8 Hz syllabic rate, observed in (Pellegrino et al., 2011). Authors claim, that 'these two perspectives indicate likely universal rhythmic properties of human speech' (Ding et al., 2017, p. 184). But as was discussed before, 5-8 Hz is highly considerable variation for speech rate. This exact variation of 5-8 Hz was discussed above, and the authors of that research claim that the differences in speech rate between most of the languages studied are statistically significant (Pellegrino et al., 2011). Therefore, it is incorrect to describe this variation as a universal characteristic.

In the discussed studies, a limited number of languages were studied. All in all, there were 15 of them: English (British and American), German, Swedish, Dutch, Danish, Norwegian, French, Italian, Japanese, Mandarin Chinese, Spanish, Russian, Hungarian. If we mark them on a map (see Map 1), we will see, that most of them are in Europe. In other words, it is possible to call the discussed studies Eurocentric, which also contributes to scepticism regarding the universality of the figures presented and the conclusions drawn from them.

Map 1. Languages studied in the reviewed literature[4].



Eurocentrism in linguistic studies can usually be explained by the availability of the European languages data and unavailability of others. Indeed, the European languages have large corpora, including multimedia subcorpora, which is needed for speech rate studies. As a result, it is much easier to conduct studies on these languages since one can skip the data collection process, which is indispensable in other studies.

For European languages there is another way for data acquisition — secured big data from technological companies. Since the Internet segments of minor languages are very little (or absent at all), there is few data or no data at all that can be used for linguistic research. In this research, I use data from languages of Russia to testify some assumptions made on the basis of major languages data.

[4] The map is drawn using **lingtypology** R package (Moroz, 2017a).

3. Methods

3.1. Materials

In this study, to analyse the difference in the speech rate at the intra-speaker and inter-speaker levels, I use the following spoken corpora:

- Corpus of Rogovatka dialect (Ter-Avanesova et al., 2018);
- Ustja River Basin Corpus (Daniel, Dobrushina, & von Waldenfels, 2018);
- Corpus of Qakh Dialect of the Azeri Language (Linguistic Convergence Laboratory, n.d.);
- Spoken corpora of the Bashkir language (Ovsyannikova, Say, Aplonova, Smetina, & Sokur, 2017);
- Beserman (‘ The Spoken Corpus of the Beserman Dialect of the Udmurt Language’, 2018);
- Chukchi (‘The Multimedia Corpus of the Chukchi Language’, 2018).

All the data was collected during the expeditions by students and staff members of the National Research University Higher School of Economics[5].

3.2. Data processing

Each of the listed corpora contains a number of sound files and corresponding text files including transcripts of utterances. The transcripts include timestamps and speaker identifiers. The utterances in all corpora are pre-divided into sentences (Beserman, Rogovatka, Ustja, Chukchi, Azeri) or words (Bashkir) by the corpora compilers; the subsequent analysis is based on this division. For Bashkir language, I put words back into sentences for the consistency of the analysis, as for all the other corpora speech rate of an utterance include the pauses between words. Otherwise, for Bashkir language instead of the speech rate, I would calculate the articulation rate which is always higher than the speech rate.

I carry out all the preliminary data processing using the Python programming

[5] I thank Alexandra Ter-Avanesova, Svetlana Dyachenko, Elena Kolesnikova, Anna Malysheva, Daria Ignatenko, Anastasia Panova, Nina Dobrushina, Timofey Arkhangelskiy, Oleg Volkov, Maria Ovsjannikova, Sergey Say, Ekaterina Aplonova, Anna Smetina, Elena Sokur, Ruprecht von Waldenfels, Michael Daniel, Yury Lander for the provided data and the possibility to use it in the present research.

language (Van Rossum & Drake, 2011). The link to the programming code can be found in Appendix 1.

Firstly, for each language, I chose speakers, whose overall speech duration in a particular sound file is no less than 70% of the overall speech duration of the file. If there is no such a speaker in a sound file, I exclude this file from the analysis. This constraint is based on the idea that the monologue form of speech is more suitable for the speech rate evaluation. That is, if we do not differentiate between the monologue and dialogue (or even polylogue), our evaluation will be biased by the effects related to the accommodation of the speech rate of a participant to that of an interviewer or co-participant (for more details, see Literature review section). However, if one person's speech comprises more than 70% of the overall speech duration of a sound file, we can be almost certain that it is a monologue (with, possibly, some short questions from an interviewer).

Notwithstanding in above-mentioned works it was found that longer utterances are pronounced faster (see Fonagy & Magdics, 1960 about Hungarian; Stepanova, 2011 about Russian), I do not exclude observations from the sample based on their very short or very long duration. It may seem natural to do if we want to uniform the data and to avoid the length effect, but there are two reasons for keeping them. Firstly, in other works, for example in (Moroz, 2017b), the tendency to utter longer phrases faster was not observed in the Kabardian language. Secondly, the speech of a person consists of utterances of various lengths, and his/her speech rate, as well as our perception of it, is formed by all of them.

Secondly, for each sentence, I calculate its duration in milliseconds according to the time of its beginning and ending which marked in the corresponding file. Then, using the list of vowels for each language, I count the number of syllables in each sentence. After that, I divide the number of syllables in a sentence by its duration in seconds and get the speakers speech rate in syllables per second. The structure of the summary table is presented in Table 4.

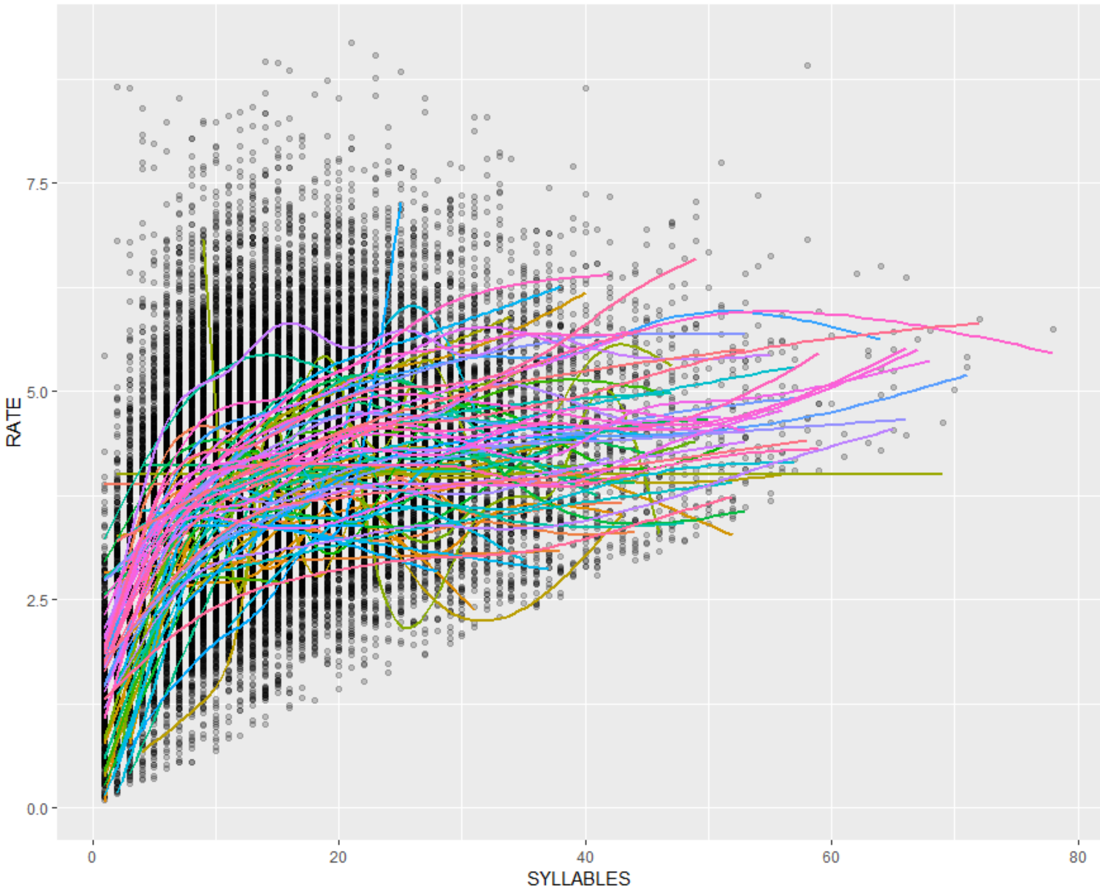
Table 4. The structure of the summary table.

Id	Language	Participant	Filename	Age	Gender	Text of utterance	Duration of utterance, ms	Number of syllables	Speech rate, syllables/second

After all the data is summarized, I exclude observations, in which the speech rate is 10 syllables per seconds or higher, as they are obvious outliers. Their presence may be explained by markup errors. All in all, there were 31 446 observations and after the exclusion, this number reduced to 31 434. The observations, which duration was longer than 15000 ms were also excluded, as they are abnormally long. Most likely, they correspond to several sentences, which, for some reasons, were not divided into separate segments.

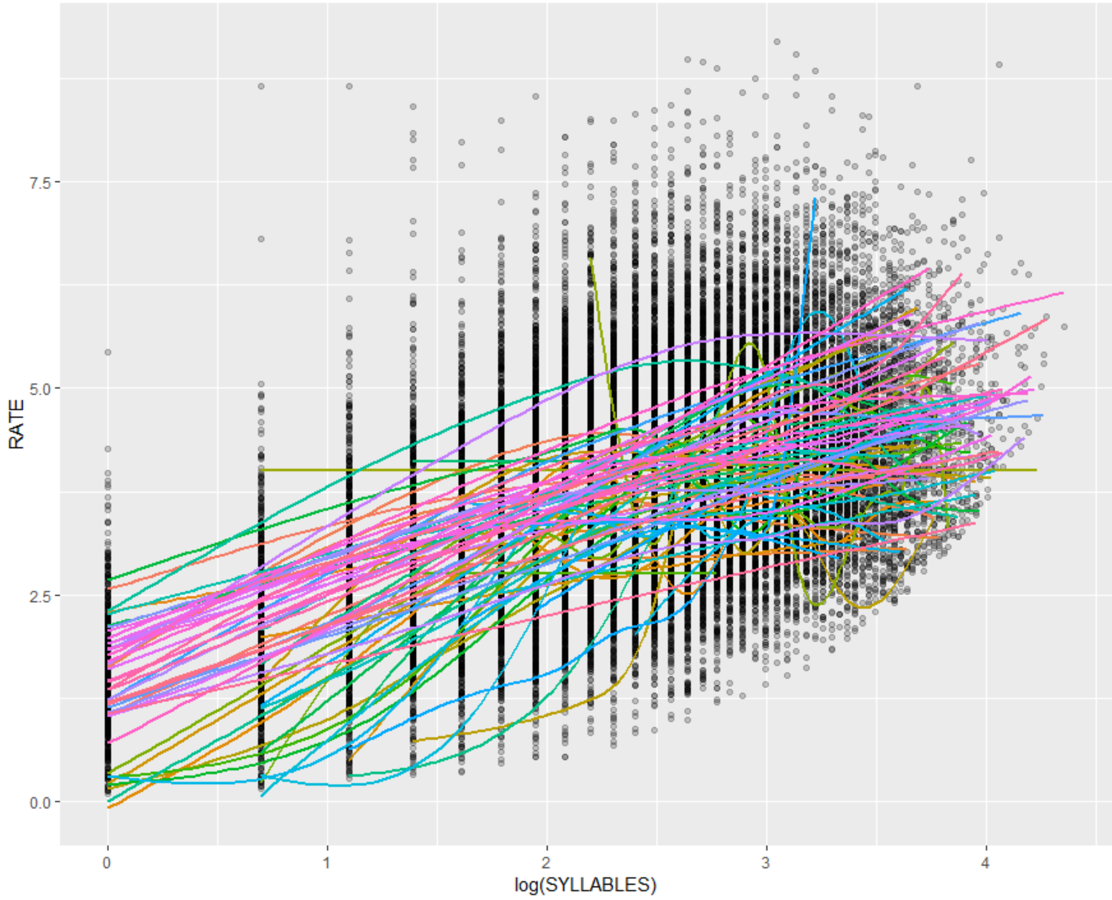
As a next step, I visualize the data to control the linear dependency of the dependent variable (speech rate) on independent variables (length of an utterance in syllables, age and gender of participants), as the linear dependency of variables is one of the assumptions of linear regression. Plot 4 contains this visualization (different colours mark different participants).

Plot 4. Dependence of speech rate on length of an utterance in syllables.



It is now obvious from the Plot 4, that the dependence of speech rate on the length of an utterance is non-linear. To avoid the notorious inaccuracy of models' predictions, it is necessary to transform the data. One of the most convenient ways to do so that preserve the ability to straight forward interpret a model's parameters is logarithmic transformation. This transformation for each value returns its natural logarithm. Plot 5 represents the results of this transformation, applied to values of lengths of utterances (different colours mark different participants).

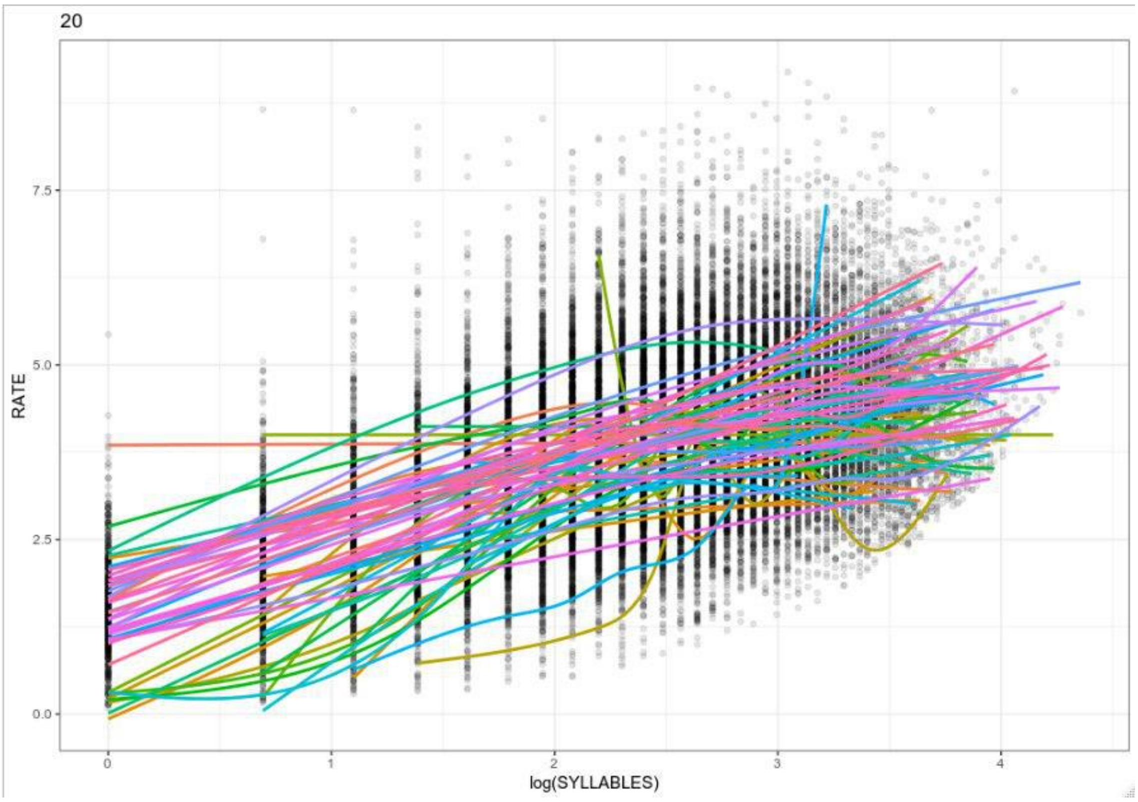
Plot 5. The dependency of speech rate on logarithmically transformed length of an utterance.



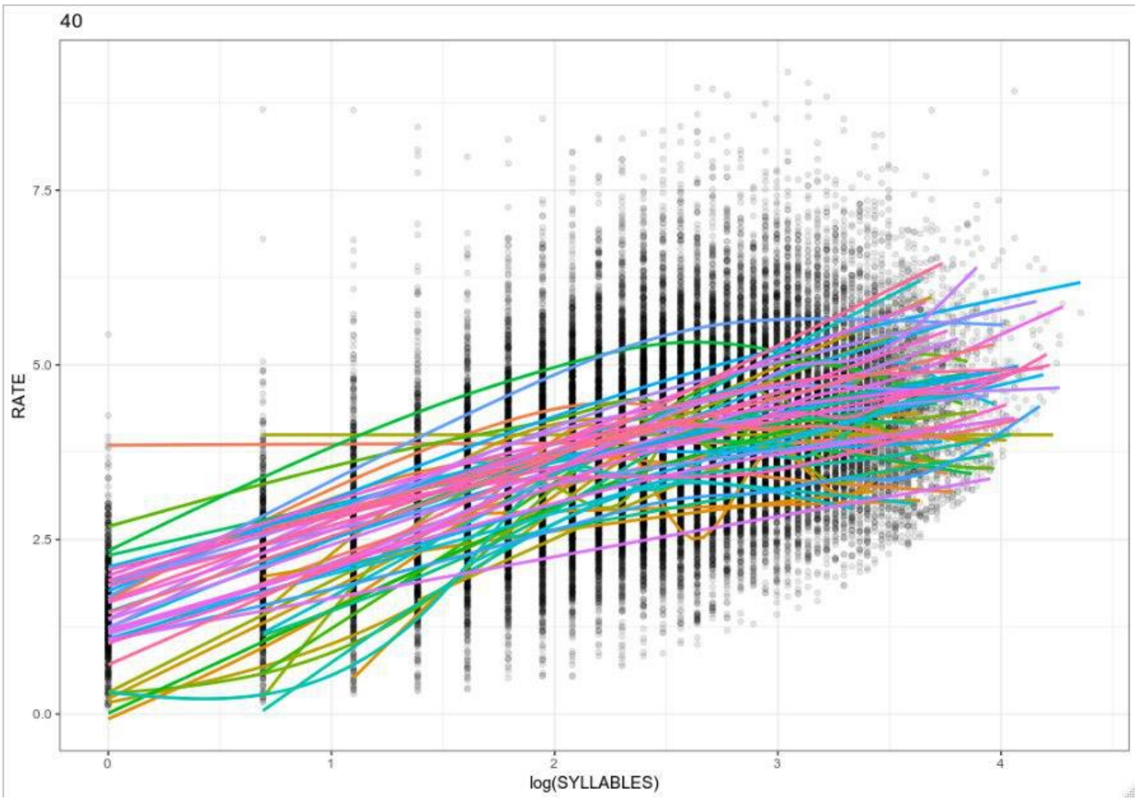
For most participants now, the dependency became more linear. After the data examination, that non-linear dependency preserves for participants with a relatively small number of observations. The gradual elimination of such participants showed that the number of observations, necessary for the linearity of the dependence is 100. Plots 6a-e represent the process of the elimination.

Plots 6a-e. The dependency of speech tare on logarithmically transformed length of an utterance for participants with more than a) 20 observations, b) 40 observations, c) 60 observations, d) 80 observations, e) 100 observations.

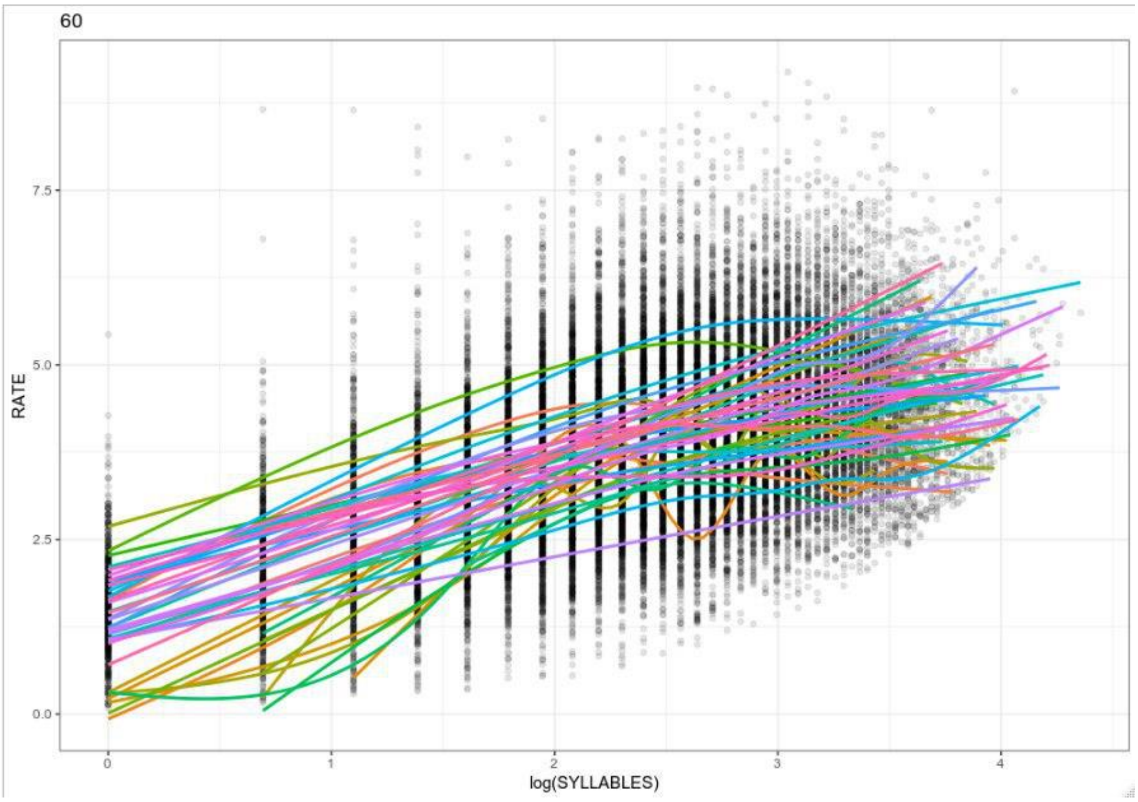
6a.


20
RATE
7.5
5.0
2.5
0.0
0
1
2
3
4
log(SYLLABLES)

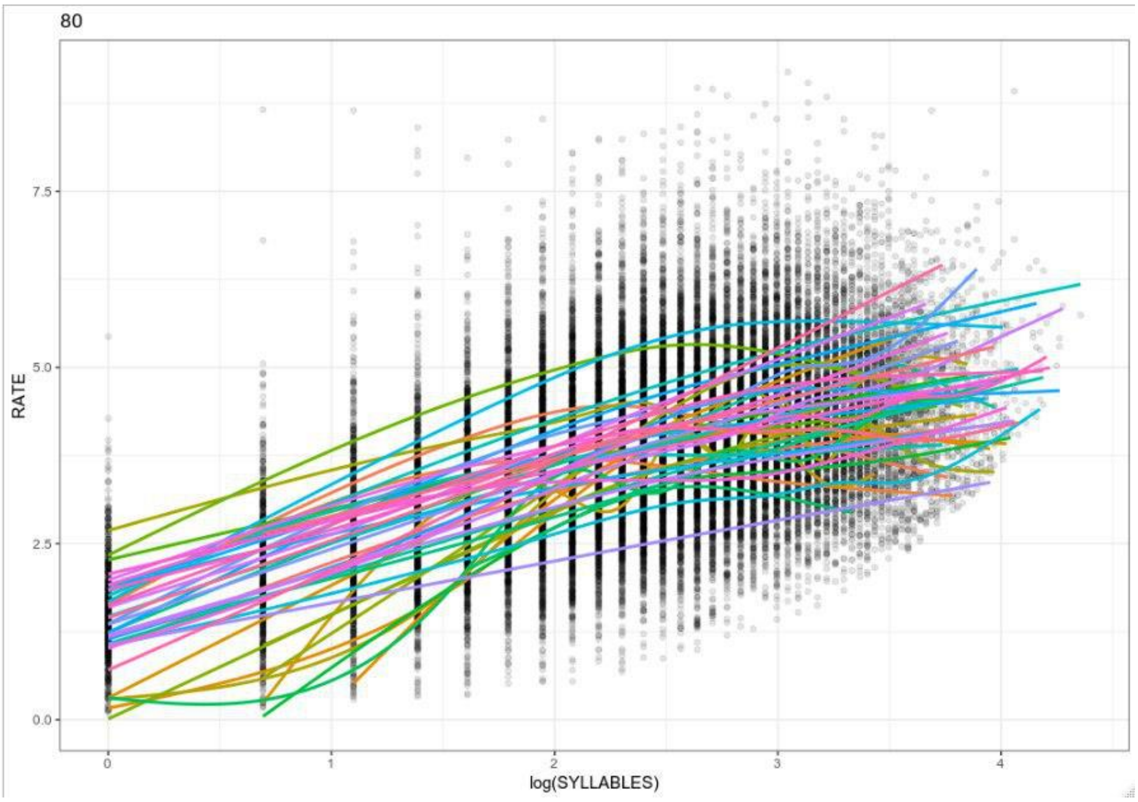
6b.


40
RATE
7.5
5.0
2.5
0.0
0
1
2
3
4
log(SYLLABLES)

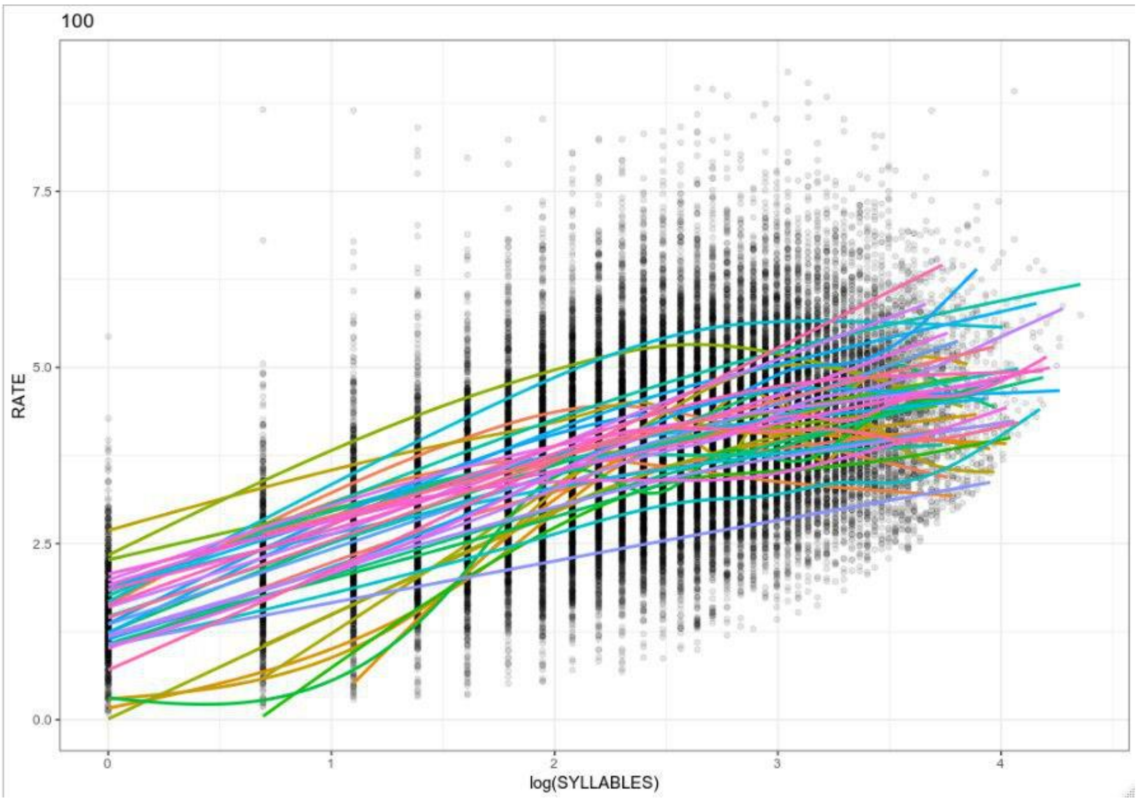
6c.


60
RATE
7.5
5.0
2.5
0.0
0
1
2
3
4
log(SYLLABLES)

6d.


80
RATE
7.5
5.0
2.5
0.0
0
1
2
3
4
log(SYLLABLES)

6e.



After all the exclusions, a total of 29 632 observations left. The quantitative information by languages is presented in Table 5. The quantitative information by age and gender is presented in Table 6.

Table 5. Quantitative information by languages.

Language	Number of speakers	Number of observations	Total duration, sec.
Beserman	5	1 241	4 541
Bashkir	9	1 607	8 712
Chukchi	1	121	437
Azeri	0	0	0
Ustja	21	18 940	67 933
Rogovatka	6	7 723	31 885
Total	**42**	**29 632**	**113 508**

Unfortunately, all the speakers of the Azeri language had less than 100 observations, so they are not presented in the analysed sample. And for the Chukchi language, only one speaker met the criterions for inclusion in the sample.

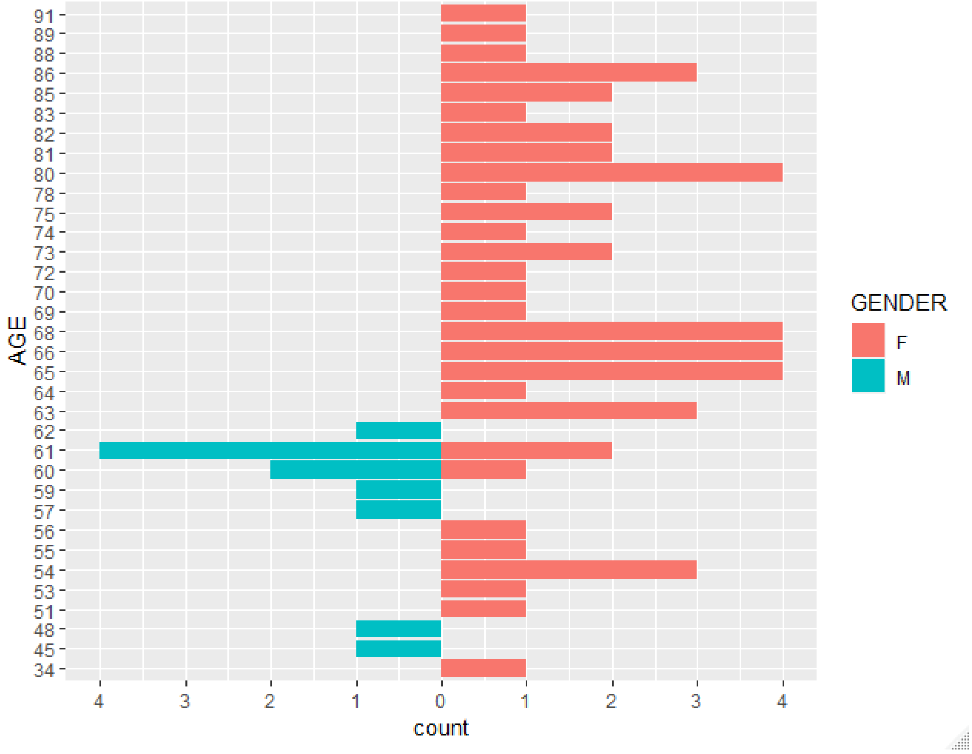
Table 6. Quantitative information by gender.

Gender	**Number of participants**	**Age[6]**	**Number of observations**	**Total duration, sec.**
Female	34	mean = 70.30 std = 11.89	26 722	101 794
Male	6	mean = 57.72 std = 5.75	2 910	11 714
Total	**42**	**mean = 68.15 std = 12.03**	**29 632**	**113 508**

From Table 6, we can see, that the number of female and male participants is dramatically different. Taking this into account, it may be of low efficiency to study the dependency of speech rate on these parameters. Plot 7 contents population plot of age by gender, which illustrates Table 6.

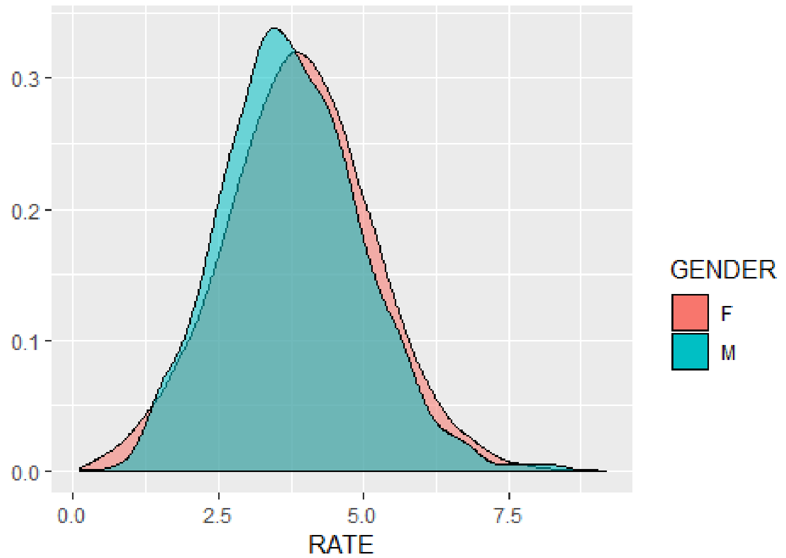
[6] For some participants, there were recordings from different years, hence they are of different age on these recordings. When calculating mean and standard deviations in such cases, one participants of different age was treated as different participants.

Plot 7. Population plot of age by gender[7].



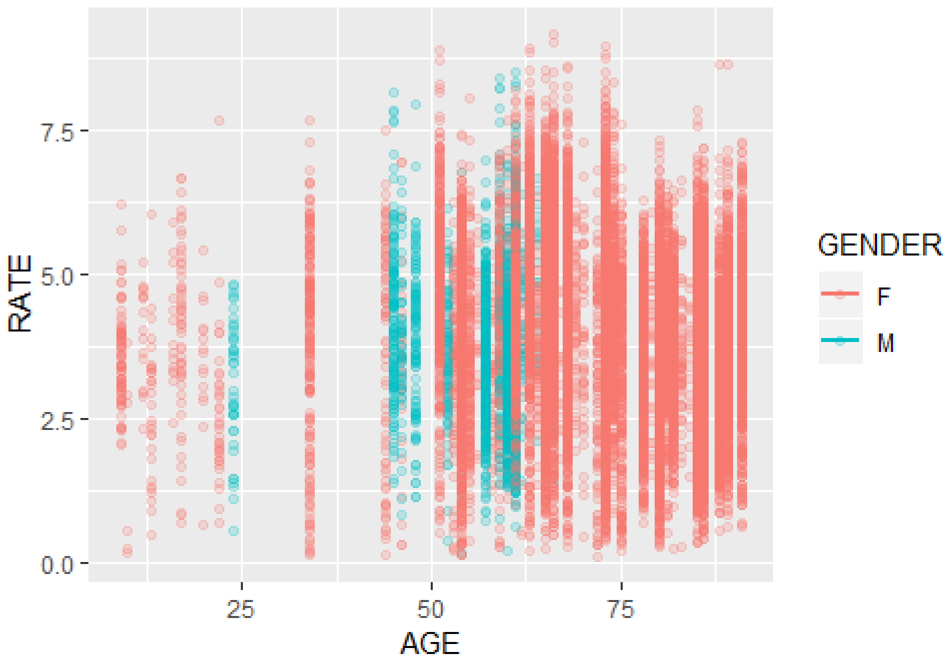
To finally refuse the idea of studying the dependency of the speech rate on age and gender, it is necessary to visualize it. Plot 8 represents the distribution of speech rate by gender, and Plot 9 represents the dependence of speech rate on age.

Plot 8. Distribution of speech rate by gender.



[7] In Plot 7, there are more bars representing a participants age, than the number of participants, indicated in the Table 6, again due to the fact, that fore some participants there were recording from different years, so on the plots they are treated as different participants.

Plot 9. Dependence of speech rate on AGE.



As can be seen from the Plot 8, the distribution of speech rate has almost no difference between the two gender groups: their density plots are nearly identical. Plot 9 also does not provide any evidence for out-of-level linear dependency between the age and the speech rate, which means, that the change in age does not induce any change in speech rate. Hence, the usage of age and gender as the main effects in regression will not provide any satisfactory results.

3.3. Statistical analysis

3.3.1. Multilevel Mixed-Effects Models

All the statistical analysis is performed in the R programming language (R Core Team, 2019). The link to the programming code may be found in Appendix 1.

For the analysis purposes of this work, I use Multilevel Mixed-Effects Models (Bryk & Raudenbush, 1988). These models allow taking into account complex interactions of the studied variables on more than one level (that is why ‘multilevel’). Using them, it is possible to model nested data: the data, in which values of some independent variable group within another independent variable, and to consider that some parameter may vary differently across these groups. For example, if we study the dependency of GPA of a student on his/her attendance, the students will be nested in classes, nested in schools. The ‘mixed-effects’ part of the name comes from the possibility to study two types of variables (which are called ‘effects’ in this paradigm) using these models: population- and group-level (these two groups are often called ‘fixed’ and ‘random’ in the literature, but since these terms are not used unambiguously, I avoid them in this work (Gelman & Hill, 2006)). The population-level effects are those independent variables, which are usually intentionally set and controlled in experiments, and group-level are those, which influence cannot be controlled, and which values do not represent all possible levels of the factor in population.

The main difference between the population- and the group-level effect is that the latter is estimated using partial pooling, while the former is not. It means that in the situation of a small number of representatives of some group of a group-level factor, an estimate of this level will be partially based on the data from other levels. This allows avoiding two desperate measures: a complete pooling of all levels, which masks group-level effects and a separate estimation, which may turn out to provide low-quality estimations for underrepresented levels (Paul (https://stats.stackexchange.com/users/11646/paul), 2017).

This way of estimation leads to the fact, that group-level effects have an associated variance, while population-level ones do not. As a result, group-level effects estimate not only the population mean but also variance, while population-level estimate a single parameter: population mean. The other important outcome is that partial pooling makes group-level parameters shrink towards the population mean, which helps to avoid overestimation of outliers and thus makes the regression more robust. Mathematical formulas of simple linear regression and Mixed-Effects linear regression are the following:

$y = X\beta + \varepsilon$ (simple linear regression)

$y = X\beta + Z\gamma + \varepsilon$ (Mixed-Effects linear regression),

where y is a dependent matrix, X is an independent matrix, β is a vector of coefficients that we pick up to better explain y(intercept and slope), $\varepsilon$is a vector of residuals (part of variance that we unable to explain using model), Z is group-level effects matrix, and $\gamma$is a corresponding vector of coefficients. Training of a model consists of step-by-step adjusting of β and γ in order to predict y as accurate as possible.

The Mixed-Effects Models are useful when you try to model a situation when your subgroups are a part of a greater common group with a common mean. That is, you presuppose that your groups are samples from a parent group and the means of these subgroups are not deviate crucially from the parent group mean. The deviations are usually thought to be normally distributed (to form the Gaussian curve). That is why the group-level effects are called 'random': deviations of their means from a parent group mean form the distribution of a random variable. The misunderstanding of the term comes from the fact, that sometimes people think, that if a variable called 'random', it is 'randomly-sampled', which is not true (Paul (https://stats.stackexchange.com/users/11646/paul), 2017).

Group-level effects are also helpful for the 'nested' data in the following sense: most statistics are designed for the independent residuals since the independence of the observations is one of the basic assumptions of the regression analysis (along with the assumption that a dependent variable is distributed normally). It means, that knowing one residual, you will not be able to predict any other above chance. But since we have groups within the data, this restriction is violated. If we know that most of the residuals of a participant are positive, we can predict above chance that one another residual of the same participant will be also positive. Adding group-level parameters helps to decorrelate the residuals by modelling them as a covariance matrix thus avoids the independent assumption for observations within groups.

The Mixed-Effects Models are designed to automatically define the level of partial pooling for each group based on the following parameters:

- the number of observations in a group;
- total number of observations;
- mean and variance within a group;
- mean and variance within the whole sample.

The Multilevel Mixed-Effects Models suit perfectly for the data of the present work. Indeed, all the observations are nested, firstly, within a participant and, secondly, all the participants are nested within languages. It is also the case that some participants have shorter recordings than others and, consequently, represented by fewer observations. The same situation with languages: some languages have a smaller number of participants than others (the quantitative information by languages was presented in Table 5).

3.3.2. Frequentist vs. Bayesian approach to statistics

In this work, I use two statistical paradigms, frequentist and Bayesian, in regard to the Multilevel Mixed-Effects Models. The main difference between frequentist and Bayesian paradigms is an approach to what the frequency is. Bayesian statistics define probability as *a degree of belief* that some event takes (/will take/has taken) place, while frequentist interpretation is that probability is *a limit of the relative frequency* of some event after numerous (infinite) number of trials (Gelman et al., 2013). The degree of belief in the Bayesian approach can be based on the results of previous experiments or on the subjective opinion of a person. This information or opinion is called *prior knowledge* (or just *prior*). To update the probabilities in order to take into account new data, Bayes' theorem is used. The process of updating the probabilities is called Bayesian inference. The mathematical formula of the probability though is the following:

(1) $P(H \vee D) = \frac{P(D \vee H) \cdot P(H)}{P(D)}$, where

$P(H)$ — is a prior probability of a hypothesis, before obtaining new data D.

$P(D)$ — is what is called 'marginal likelihood', the probability of the data regardless of the hypothesis (equivalent for all the possible hypotheses).

$P(D \vee H)$ — the probability to observe data D given that the hypothesis H is true.

$P(H \vee D)$ — is a posterior probability, the probability of the hypothesis H given the data D.

The second difference between frequentist and Bayesian statistics to be noted is an approach to parameter estimation. From the frequentist point of view, the true parameter value ($\tilde{x}$) is the one fixed value, that could be approached by an infinite number of experiments and that provides the highest probability for the data D:

(2) $\tilde{x} := arg \max_{x} p(D|x)$

Such an estimator is called 'maximum-likelihood estimator'. This approach does not provide the probability of some parameter value to be the true value.

From the Bayesian perspective, unlike the frequentist one, it is valid to speak of the probability distribution of some parameter x being estimated. This distribution is written as:

(3) $p(x \vee D)dx \propto p(D \vee x)p(x)dx.$

As can be noted, the form of the probability distribution is based on the Bayes' theorem (1), but the hypothesis H is replaced here by the parameter value. The distribution provides information on a probability for each possible value of x to be the true value of an estimated parameter given the data D.

3.3.3. Models description

For the frequentist approach, I use **lme4** R package (Bates, Mächler, Bolker, & Walker, 2014) and for the Bayesian one **brms** package (Bürkner, 2017), which provides the lme4-like interface for Stan models (Stan Development Team, 2014).

According to the hypothesis of the research, the speech rate is not a language-specific parameter, which means, that a language as a parameter should not influence the speech rate of a particular speaker. To testify the hypothesis, I designed several Mixed-Effects models, proceeding from different assumptions regarding the interaction of population- and group-level effects. If the hypothesis is correct, the models, not including language as a parameter, will describe the data at least not worse than those including it. I will first describe the general rules of the lme4-like notation used in the models, then proceed to the structure of the models I designed, and finally to the methods used for model comparison.

The general form of a simple linear model in R is the following:

lm(Y ~ A),

where Y is a studied parameter and A is an effect affecting Y. We assume, that there is a linear dependency between them. While the left side of the formula is always the same (one parameter that we study), the right side may differ. Some of the possible variants are described in Table 7.

Table 7. Variants of notation of a simple linear model.

Notation	Meaning
A	There is one effect, the influence of which we study.
A + B	There are two effects which affect the studied parameter and they do not interact

	with each other (i.e. there is no correlation between them).
A:B	There are two effects which interaction affect the studied parameter.
A * B = A + B + A:B	There are two effects which affect the studied parameter separately and they do interact with each other (i.e. we assume, that there is a correlation between them).

In addition to these parameters (which are called population-level effects in the Mixed-Effects models), in the Multilevel Mixed-Effects Models we have group-level effects. The notations used for them in the designed models are represented by Table 8.

Table 8. Notations for the group-level effects.

Notation	Meaning
(1|A)	The mean of the studied parameter may vary on different levels of effect A.
(B|A) = (1 + B|A)	The mean of the studied parameter and the slope of the regression line vary on the different levels of the effect A. The variation of mean and slope interact with each other (i.e. there is a correlation between them).
(B||A) = (0 + B|A) + (1|A)	The mean of the studied parameter and the slope of the regression line vary on the different levels of effect A. The variation of mean and slope do not interact with each other (i.e. there is no correlation between them).

(1|A/C)	The mean of the studied parameter varies on the different levels of effect A. The mean also vary between different levels of B which levels are sublevels of the levels of A (C is nested within A).
(B|A/C)	The mean of the studied parameter and the slope of the regression line vary on the different levels of effect A. The mean and the slope of the regression line also vary between different levels of C which levels are sublevels of the levels of A (C is nested within A). The variation of mean and slope interact with each other (i.e. there is a correlation between them).
(B||A/C) = (0 + B|A/C) + (1|A/C)	The mean of the studied parameter and the angle of slope of the regression line vary on the different levels of effect A. The mean and the angle of slope of the regression line also vary between different levels of C which levels are sublevels of the levels of A (C is nested within A). The variation of mean and slope do not interact with each other (i.e. there is no correlation between them).

As it was decided to not use age and gender as population-level factors (see the Materials chapter of the Methods section), only the length of an utterance in syllables was placed on this position. Table 9 summarizes the information about the designed models and provides explanations regarding their structure. The names of the factors have the following meaning:

- RATE — speech rate of observation (1 utterance = 1 observation)
- GENDER — gender of a speaker
- AGE — age of a speaker
- LANGUAGE — a language of an observation
- PARTICIPANT — id of a speaker
- SYLLABLES — length on an utterance measured in the number of syllables

Table 9. Multilevel Mixed-Effects Models.

Model 1	
Notation	RATE ~ 1 + (1|LANGUAGE/GENDER/PARTICIPANT)
Effects	Population-level: - Group-level: PARTICIPANT nested in GENDER, which nested in LANGUAGE
Meaning	1. The mean of RATE may vary on the different level of LANGUAGE, its sublevels — GENDER, and sublevels of GENDER — PARTICIPANTS.
Model 2	
Notation	RATE ~ 1 + (1|GENDER/PARTICIPANT)
Effects	Population-level: - Group-level: PARTICIPANT nested in GENDER
Meaning	1. The mean of RATE may vary on the different level of GENDER, and its sublevels GENDER — PARTICIPANTS.
Model 3	
Notation	RATE ~ SYLLABLES + (SYLLABLES|LANGUAGE/PARTICIPANT)
Effects	Population-level: SYLLABLES Group-level: LANGUAGE and PARTICIPANT, nested in it
Meaning	2. The studied parameter RATE is affected by the population-level effect of

	SYLLABLES. 3. The mean of RATE may vary on the different level of LANGUAGE and its sublevels — PARTICIPANTS. 4. The SYLLABLES may differently affect (the regression line may have a different angle of a slope) the RATE on the different levels of LANGUAGE and its sublevels — PARTICIPANTS. 5. The effects from points 2 and 3 may be correlated.
Model 4	
Notation	RATE ~SYLLABLES + (SYLLABLES|PARTICIPANT)
Effects	Population-level: SYLLABLES Group-level: PARTICIPANT
Meaning	1. The studied parameter RATE is affected by the population-level effect of SYLLABLES. 2. The mean of RATE may vary on the different level of PARTICIPANTS. 3. The SYLLABLES may differently affect (the regression line may have a different angle of a slope) the RATE on the different levels of PARTICIPANTS. 4. The effects from points 2 and 3 may be correlated.
Model 5	
Notation	RATE ~ 1 + (1|LANGUAGE/PARTICIPANT)
Effects	Population-level: - (intercept-only model) Group-level: PARTICIPANT nested in LANGUAGE
Meaning	1. The mean of RATE may vary on the different level of LANGUAGE and its sublevels — PARTICIPANTS.
Model 6	
Notation	RATE ~ 1 + (1|LANGUAGE/PARTICIPANT)

Effects	Population-level: - (intercept-only model) Group-level: PARTICIPANT
Meaning	1. The mean of RATE may vary on the different level of PARTICIPANTS.

The models are grouped in pairs (1-2, 3-4, 5-6). In all the pairs, the first model does include LANGUAGE as a group-level factor and the second does not. This design allows testifying importance of the LANGUAGE factor. The groups itself differ in what parameters are on the population-level and group-level position. The 1-2 models 3-4 have SYLLABLES, 1-2 and 5-6 do not have population-level parameters at all. In the 1-2 and 5-6 models, all the variance is explained by group-level parameters. These two models allow testifying significance of population-level effects used in 3-4 models.

There was an attempt to design models with several population-level parameters (with the interaction of AGE and GENDER and SYLLABLES) together, but, unfortunately, there was not enough data for such models to converge (as the more complex structure of a model requires the more complex structure of data samples and, therefore, mode observations in each group).

3.3.4. Models comparison procedure

All six models were implemented using both **lme4** and **brms** R packages. As soon as the dependent variable (RATE) is distributed almost normally (see Plot 8), it is possible to use regular regression. After all the models are fitted, for **lme4** models I perform analysis of variance using **anova()** standard R function, and for **brms** models, I perform Bayes factor analysis to find out, which model describes the given data better. If the hypothesis is correct, the model, not including LANGUAGE as a group-level factor will describe the given data at least not worse than the models, that do include it.

ANOVA (Analysis of variance) is a set of statistical models and their estimation procedures. It is used to analyse the means of groups in a sample and difference between them. ANOVA is usually used to perform a statistical test, aimed at defining whether two (or more) population means are equal, which is similar to t-test. The null hypothesis is that the means are equal. The analysis usually results in some p-value, which is the probability of the null hypothesis given the observed distributions of means (Navarro, 2019).

Bayes factor is a ratio of a probability of the data D given that the hypothesis $H_1$is true (likelihood of the hypothesis H_1) to a probability of the data D given that the hypothesis $H_2$is true (likelihood of the hypothesis H_2):

$$B_{12} = \frac{P(H_1 \vee D)}{P(H_2 \vee D)}$$ (Kass & Raftery, 1995, p. 776)

In other words, the Bayes factor tells how many times one hypothesis is more probable than another given the data D. The scale of evidence, proposed by Lee & Wagenmakers (2013) which is based on the classic book of Jeffreys (1998) is presented in Table 10.

Table 10. The scale of evidence for the Bayes factor (Lee & Wagenmakers, 2013).

Value of Bayes factor B_{12}	Interpretation
> 100	Extreme evidence for H_1
30 – 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
3 – 10	Moderate evidence for H_1
1 – 3	Anecdotal evidence for H_1
1	No evidence
1/3 – 1	Anecdotal evidence for H_2
1/3 – 1/10	Moderate evidence for H_2
1/10 – 1/30	Strong evidence for H_2
1/30 – 1/100	Very strong evidence for H_2
< 1/100	Extreme evidence for H_2

4. Results

4.1. The lme4 (frequentist) models

The results of the ANOVA test of lme4 models are presented in Table 11. The p-value column provides the p-value for comparison a model with a model from the previous position in the rating. If two models do not differ at the level of statistical significance (p-value = 0.05 or less), they have the same position in the rating (marked as n (=k)).

Table 11. ANOVA test results for lme4 models.

	Population-level effects	**Group-level effects**	**AIC[8]**	**Position in rating**	**p-value (χ^2test)**
Model 1	-	LANGUAGE/ GENDER/ PARTICIPANT	94 870	3 (=4)	< 0.001
Model 2	-	GENDER/ PARTICIPANT	94 868	5 (=6)	< 0.001
Model 3	SYLLABLES	LANGUAGE/ PARTICIPANT	83721	1	
Model 4	SYLLABLES	PARTICIPANT	83 743	2	< 0.001
Model 5	-	LANGUAGE/ PARTICIPANT	94 868	4 (=3)	0.61
Model 6	-	PARTICIPANT	94 866	6 (=5)	1

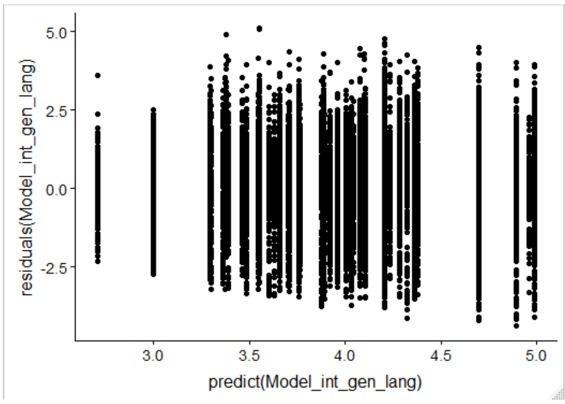
In the pairs Model 1-2 and Model 5-6 ANOVA test not found a statistically significant difference between models with and without LANGUAGE group-level factor (Model 1-2[9]: p-value = 0.47, Model 5-6: p-value = 0.61). In the pair Model 3-4

[8] AIC (Akaike Information Criterion) estimates explanatory power of statistical models for a given data. The estimation is relative, the **better model has lower AIC value**. The absolute value of the criterion does not have any sense.

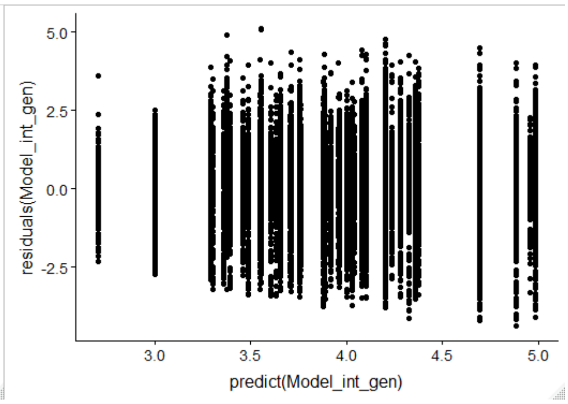
[9] Model x-y stands for comparison of Model x with Model y using ANOVA.

the model with LANGUAGE was significantly better than the model without this factor (p-value < 0.001). There was no statistical difference between models with and without GENDER group-level factor (Model 1-5 p-value = 0.61, Model 2-6 p-value = 1). Plots 10a-f represent the residuals plots of the models.

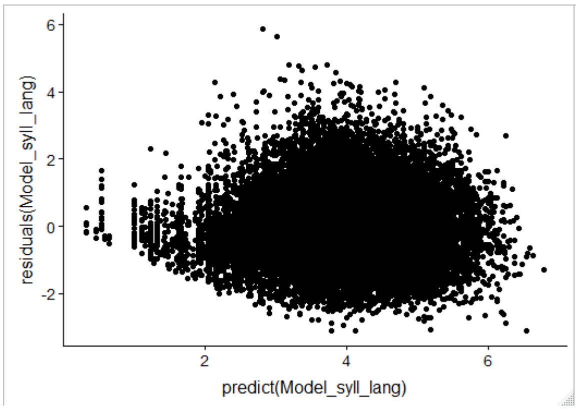
Plot 10a. Model 1 residuals plot.



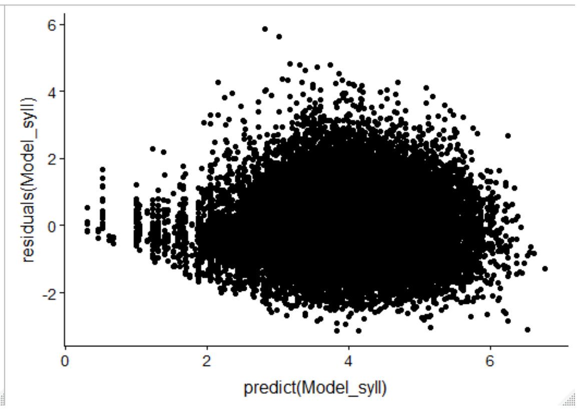
Plot 10b. Model 2 residuals plot.



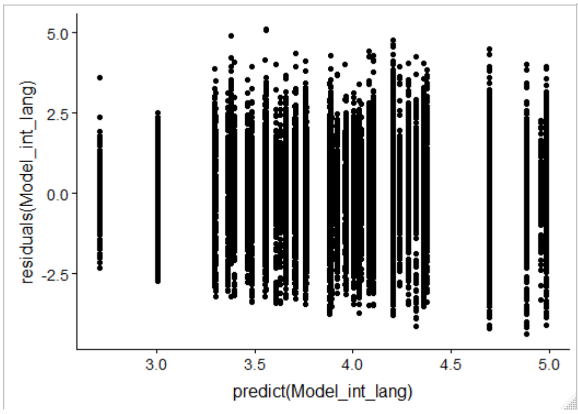
Plot 10c. Model 3 residuals plot.



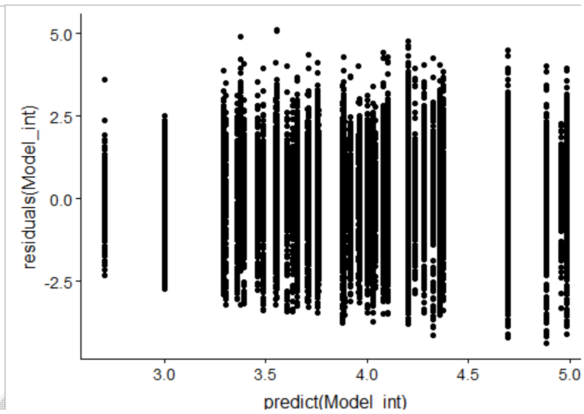
Plot 10d. Model 4 residuals plot.



Plot 10e. Model 5 residuals plot.



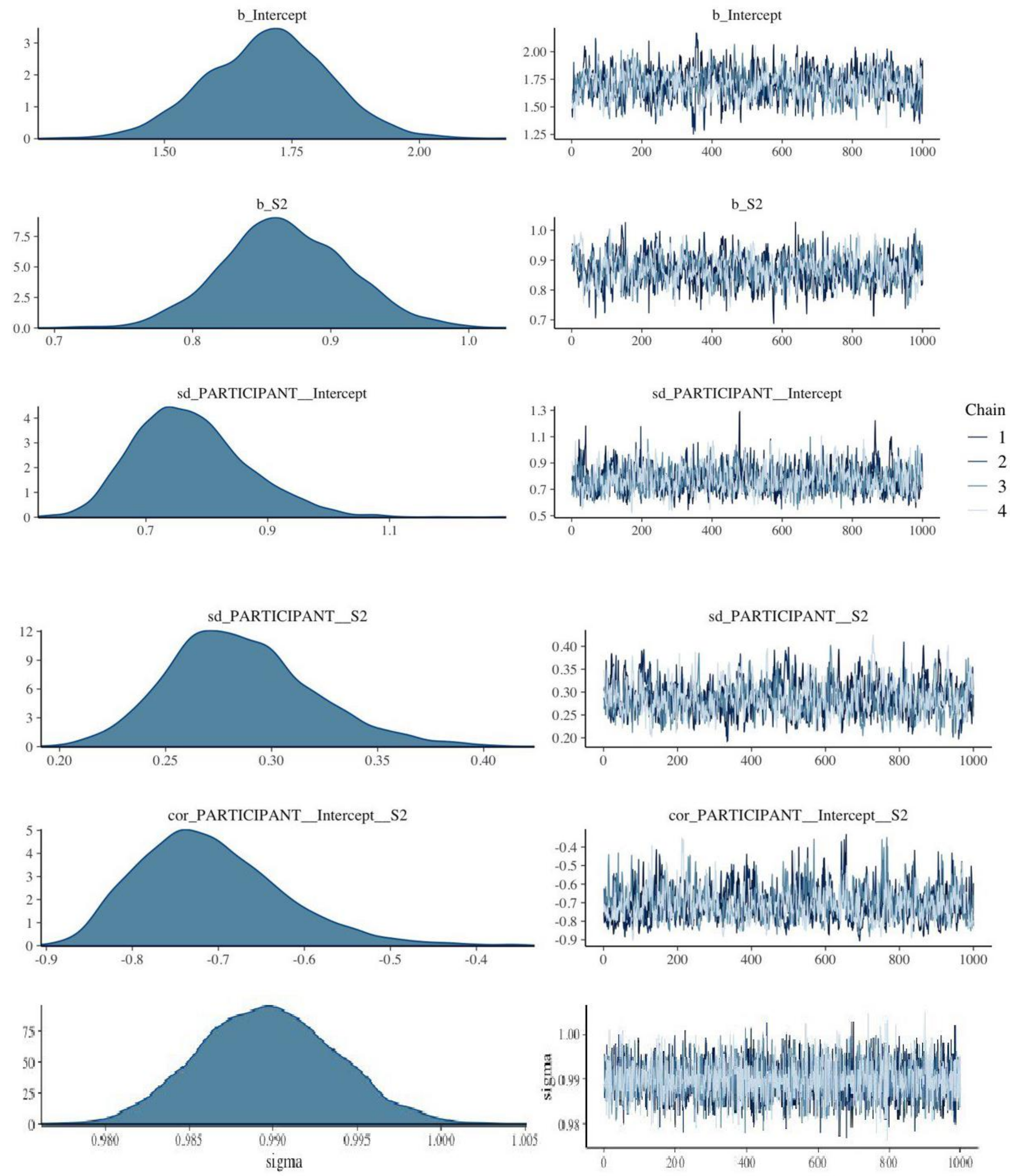
Plot 10f. Model 6 residuals plot.



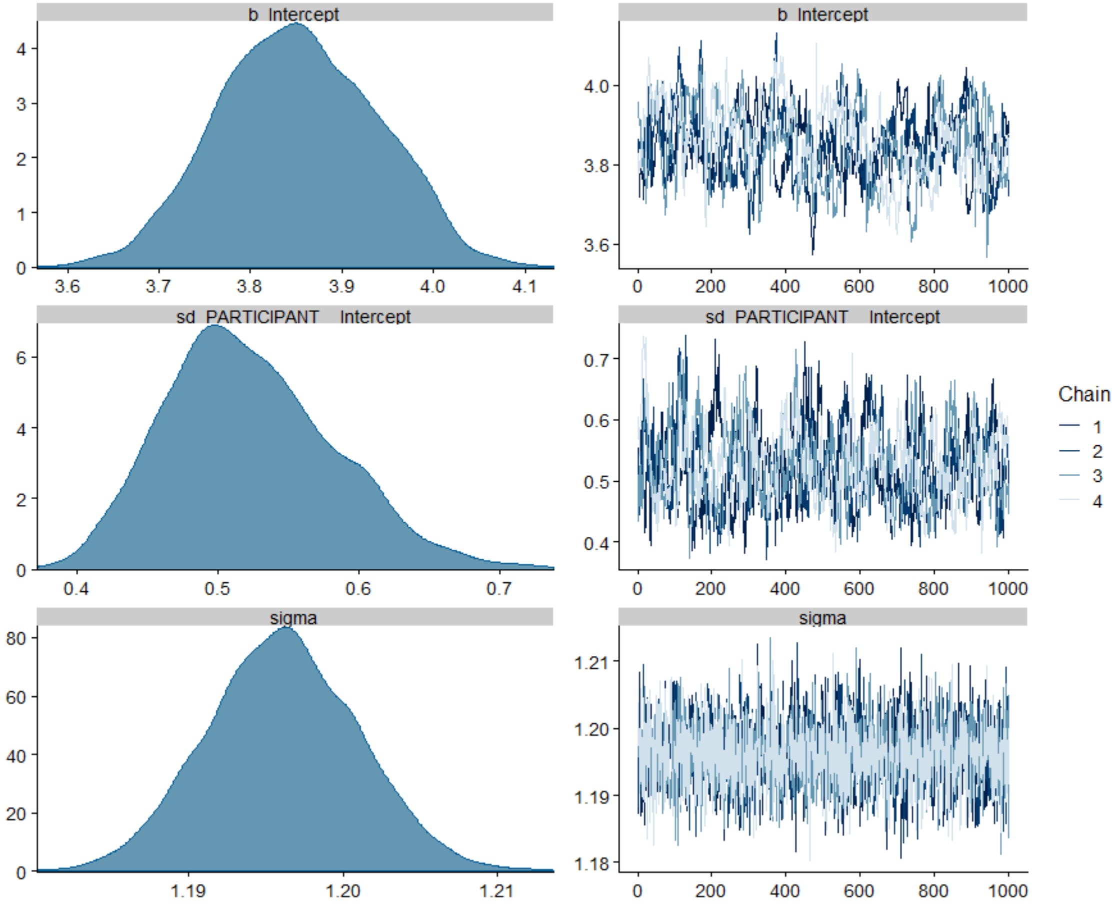
4.2. The brms (Bayesian) models

Bayes factor provides strong evidence that the model with SYLLABLES population-level factor represents the data better than the model with no population-level factors (intercept-only model) (Model 4-6 Bayes factor $\gg 100$). Plots 11a-b represent the process of Markov chains fitting and resulting distribution of the factors for Models 4 and 6 accordingly (S2 label stands for the logarithmical length of an utterance in syllables, i.e. log(SYLLABLES)).

Plot 11a. Model 4 Markov chains fitting and resulting distribution.



Plot 11b. Model 4 Markov chains fitting and resulting distribution.



Unfortunately, models having LANGUAGE or GENDER among group-level factor failed to converge. Most probably, it happened due to insufficient representation of some levels of LANGUAGE factor and, therefore, its excessive imbalance (see the Data processing in the Methods section). So, only two models:

Model 4 — RATE ~ log(SYLLABLES) + (log(SYLLABLES)|PARTICIPANT)

and

Model 6 — RATE ~ 1 + (1|PARTICIPANT) converged.

In these models, I specified no prior distributions for effects, because, as was stated in the Literature review sections, there is no general consensus on the type of dependency of speech rate on the factors, used in the models. As priors are not specified, non-informative priors, that have a minor influence on the resulting distributions, were automatically used.

4.3. Predictors' effects

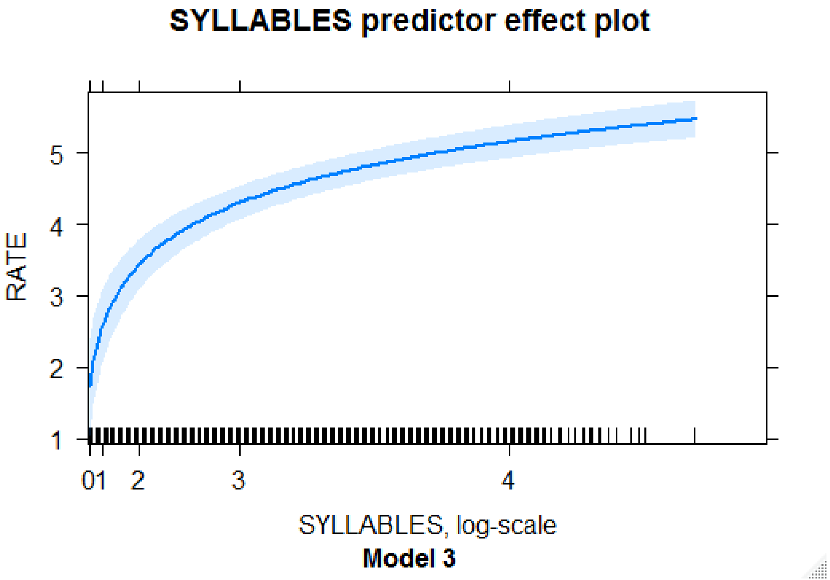
Besides the relative position in rating, there is another parameter describing the relevance of a model — effects of its predictors. The predictors' effects answer the following question: how will change the independent variable in case of unit change of a predictor. In Table 12, I provide an estimate for log(SYLLABLES) effect of the above-described models and its interpretation. All the other models do not have population-level predictors.

Table 12. Predictors effects of the designed models.

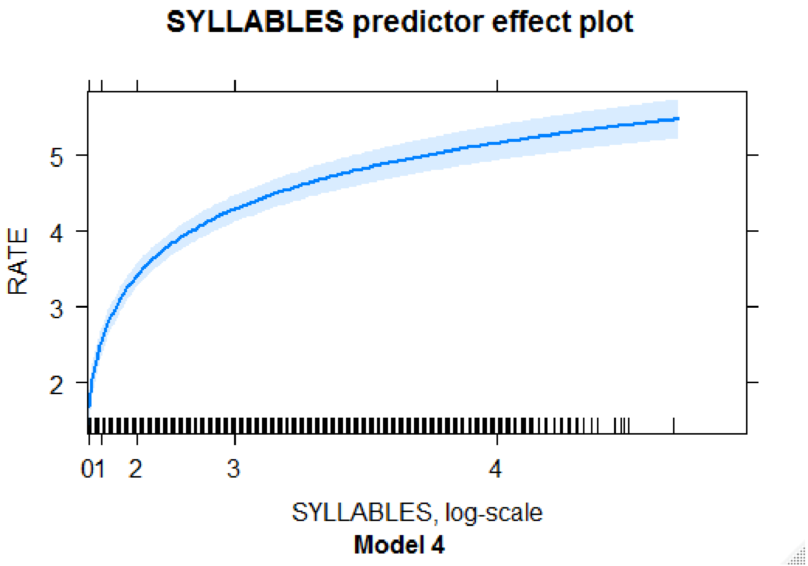
	Model	Predictor	Estimate	Interpretation
lme4	Model 3	log(SYLLABLES)	0.86116	When increasing SYLLABLES by 1%, the speech rate increases by $0.01 \cdot 0.86116 = 0.0086116$ syllables per second.
	Model 4	log(SYLLABLES)	0.86960	When increasing SYLLABLES by 1%, the speech rate increases by $0.01 \cdot 0.86960 = 0.0086960$ syllables per second.
brms	Model 4	log(SYLLABLES)	0.87 (0.78-0.96 95% credible intervals)	When increasing SYLLABLES by 1%, the speech rate increases by $0.01 \cdot 0.87 = 0.0087$ syllables per second.

For the lme4 models, it is also possible to visualize the fitted regression models predictors using the effects R package (Fox & Weisberg, 2019). This package allows visualizing of complex interactions of predictions, for which a straight forward interpretation of estimations is not possible. Although the present analysis is based on one effect — SYLLABLES — it may be also useful to visualize the fitted dependencies. For example, it is possible to show the original SYLLABLES logarithmic structure, which was made at the plots below. Plots 12a-b shows the predictor effects of SYLLABLES.

Plot 12a. Predictor effect of Model 3.



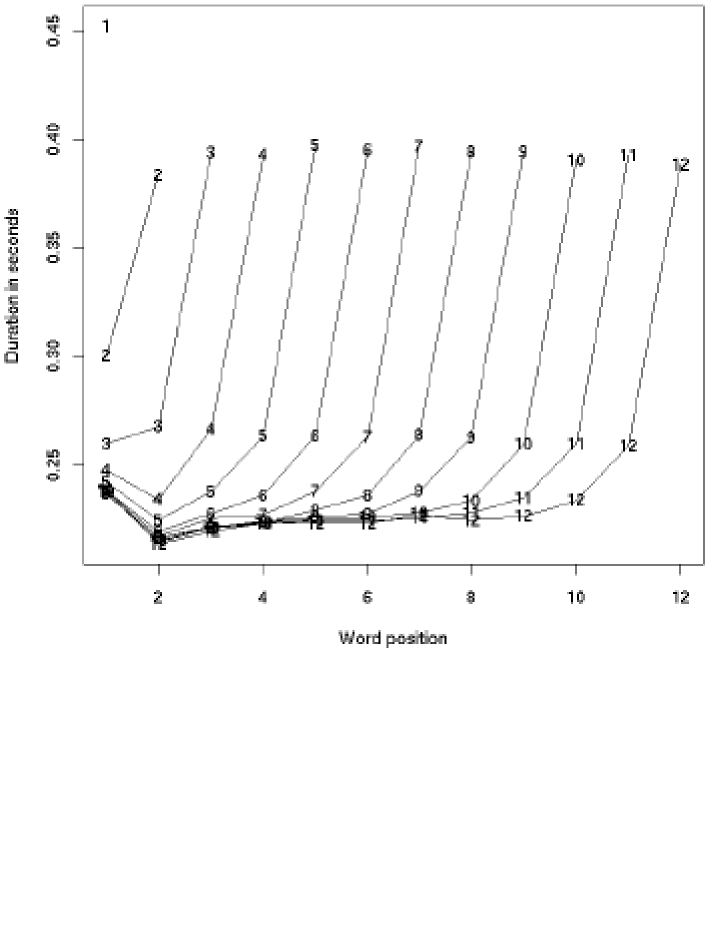
Plot 12b. Predictor effect of Model 4.



The light-blue area represents the pointwise confidence band for the predicted values. The bands are calculated ‘based on standard errors computed from the covariance matrix of the fitted regression coefficients’ (John Fox & Weisberg, 2018, p. 4). It can be seen, that with the increase of utterances’ length, the speech rate starts to increase less and less dramatically.

This phenomenon shows agreement with Yuan et al. (2006) results, who studied the dependency between the position of a word in a sentence and its duration. The results of their experiment are presented in Plot 13. Indeed, if in long utterances the words placed in the middle have almost the same articulation rate and, then each additional word contributes less to the change in the speech rate of an utterance. The Plots 12a-b show the same tendency.

Plot 13. Mean word duration by its position in a sentence (Yuan et al., 2006, p. 2).



5. Discussion and conclusions

5.1. Overview

The goal of this work was to assess the degree of influence of language as a parameter on a speaker's speech rate and, as a result, to answer the question: does it make any sense to calculate a language speech rate as an average number and to compare languages on this parameter.

To achieve this goal, I used the data from six corpora of the languages of Russia: **Russian** (Corpus of Rogovatka dialect (Ter-Avanesova et al., 2018), Ustja River Basin Corpus (Daniel, Dobrushina, & von Waldenfels, 2018)), **Azeri** (Corpus of Qakh Dialect of the Azeri Language (Linguistic Convergence Laboratory, n.d.)), **Bashkir** (Spoken corpora of the Bashkir language (Ovsyannikova, Say, Aplonova, Smetina, & Sokur, 2017)), **Beserman dialect of Udmurt** (' The Spoken Corpus of the Beserman Dialect of the Udmurt Language', 2018), and **Chukchi** ('The Multimedia Corpus of the Chukchi Language', 2018). With this data, I have designed several statistical Multilevel Mixed-Effects Models, representing different assumptions regarding the structure of the data and the importance of factors. The Multilevel Mixed-Effects Models were used because of the lack of independence between observations, as one observation in the sample equals to one utterance and there are multiple observations from one speaker. The independence of observations in one of the fundamental assumptions of the regression analysis (Draper & Smith, 1998, p. 61). The Multilevel Mixed-Effects Models allow using regression analysis in a situation of independence violation.

There were six models, grouped in pairs. Each pair had a different combination of population-level and group-level factors. Within a pair, models differed only by the parameter of presence/absence of language group-level factor. Each model was implemented both in frequentist and Bayesian statistical paradigms. In accordance with the hypothesis, it was expected, that models without language group-level factor will describe the data at least not worse than models with this factor. The models were compared using ANOVA (for frequentist models) and Bayes factor (for Bayesian models).

5.2. Discussion

The results of the statistical analysis are presented in the Results section. The results of the ANOVA comparison of lme4 models do not support the advanced hypothesis. Although in the two pairs of models (Model 1-2 and Model 5-6) there was no statistical difference between model with and without language group-level factor, these two pairs of models itself were significantly worse in their predictive ability than Models 3-4, which had SYLLABLES (the length of an utterance in syllables) on the position of the population-level factor. In this pair, the model with language group-level effect was better than the model without this effect at the level of statistical significance (p-value < 0.001). That is, the best of the six models is the one that does have language group-level factor.

From the statistical point of view, the fact that models with SYLLABLES population-level factors (Models 3-4) are significantly better describe the data than the intercept-only models with the same structure of the group-level factors (Models 5-6) also provides the evidence, that the length of an utterance is a factor, that significantly affects the speech rate. On the other hand, the value of the estimate, provided for the SYLLABLES factor by the Models 3-4 is 0.86 and 0.87 for lme4 Models 3 and 4 accordingly, and 0.87 for the brms Model 4, which is a highly consistent result. Here it is necessary to remember, that SYLLABLES factor was logarithmically transformed to gain the linear dependency between the speech rate and the length of an utterance (for more details, see the Materials chapter of the Methods section). As it was logarithmically transformed, interpretation of the estimates in the following: when the length of an utterance increases by 1%, the speech rate increases by $0.86 \cdot 0.01 = 0.0086$ syllables per second (for lme4 Model 4). How to understand, whether this change is significant or not?

To answer this question, I will return to the data and simulate the change of the length of an utterance for each of the observed lengths. Firstly, I identify all the possible lengths of an utterance, that were observed in the data. Secondly, I add to each length one syllable (as it is impossible to add less) and calculate by which amount of per cents did the length increase. Then, for each length, I calculate the mean speech rate, observed in the sample for this length and adjust this value by adding $0.0086 \cdot \%\ of\ the\ length\ change$ syllables per second. Finally, I calculate, by which per cent did the speech rate increased.

To illustrate the procedure, I will calculate the speech rate change for the length increase from 5 to 6 syllables. Change from 5 to 6 syllables is a 20% change. The mean speech rate for the 5-syllables-long utterances in the data is 3.132238 syllables per second. To adjust it, we need to add $0.0086 \cdot 20 = 0.172$ syllables per second to the rate and get 3.304238 syllables per second. Then we calculate the percentual difference: (3.304238 - 3.132238) / 3.132238 ·100% = 5.491281%. According to (Quené, 2008), this is a significant difference, as hearers can perceive the difference in the speech rate at the level of about 5%. The case is that the change from 5 to 6 syllables is the last change that can be precepted, as all the changes with the higher lengths result into the change of the speech rate by less than the boundary 5%. All in all, it is only 14.6% of the total amount of observations. The Table 13 represents the calculation for the first ten changes. The whole Table can be found in Appendix 2.

Table 13. The calculations for speech rate change in accordance with the Model 3 predictor effect estimation.

Initial length, syll.	Adjusted length, syll.	Change in length, %	Value to add, syll./sec	Initial speech rate, syll./sec	Adjusted speech rate, syll./sec	Change in speech rate, %
1.0	2.0	100.0	0.86	1.55	2.41	55.35
2.0	3.0	50.0	0.43	2.14	2.57	20.08
3.0	4.0	33.33	0.29	2.61	2.9	10.99
4.0	5.0	25.0	0.22	2.9	3.12	7.41
5.0	6.0	20.0	0.17	3.13	3.3	5.49
6.0	7.0	16.67	0.14	3.41	3.56	4.2
7.0	8.0	14.29	0.12	3.57	3.69	3.44
8.0	9.0	12.5	0.11	3.69	3.8	2.91
9.0	10.0	11.11	0.1	3.83	3.92	2.5
10.0	11.0	10.0	0.09	3.9	3.98	2.21

5.3. Conclusions

To sum up, although the models with SYLLABLES factor represent the data significantly better than other models, from the linguistic point of view, they still provide the dependency that under the level of perceptual significance for the major part of the data. Therefore, these models have to be improved by an additional amount of data, that may strengthen the already discovered dependency or reveal new ones. Until then, I do not consider it possible to definitively confirm or disprove the hypothesis of the irrelevance of a language speech rate as a cross-linguistic parameter.

Nevertheless, the present work provides an algorithm to study various linguistic parameters in terms of their influence on individual speaker's characteristics. Using it, it is possible to provide strong evidence for or against the linguistic generalizations of different levels.

References

Abercrombie, D. (1967). *Elements of general phonetics* (Vol. 203). Edinburgh University Press Edinburgh.

Avanesov, R. (1956). *Фонетика современного русского языка [Phonetics of modern Russian language]*. МГУ.

Barik, H. C. (1977). Cross-linguistic study of temporal characteristics of different types of speech materials. *Language and Speech*, *20*(2), 116–126.

Bosker, H. R. (2016). Our own speech rate influences speech perception. *Speech Prosody 2016*, 227–231.

Brown, B. L., Giles, H., & Thakerar, J. N. (1985). Speaker evaluations as a function of speech rate, accent and context. *Language & Communication*, *5*(3), 207–220. https://doi.org/10.1016/0271-5309(85)90011-4

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396.

Burnham, K. P., & Anderson, Dav. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York: Springer-Verlag.

Campbell, W. (1992). *Multi-level speech timing control.* (PhD Thesis). University of Sussex.

Campione, E., & Véronis, J. (1998). A multilingual prosodic database. *Fifth International Conference on Spoken Language Processing*.

Cedergren, H. J., & Perreault, H. (1994). Speech rate and syllable timing in spontaneous speech. *Third International Conference on Spoken Language Processing*.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, *88*(1), 101–112.

Daniel, M., Dobrushina, N., & von Waldenfels, R. (2018). The language of the Ustja river basin. A corpus of North Russian dialectal speech. Bern, Moscow. [Corpus]. Retrieved from Ustja River Basin Corpus Query interface website: http://www.parasolcorpus.org/Pushkino/www.parasolcorpus.org/Pushkino

Deese, J. (1984). *Thought into speech: The psychology of a language*. Prentice Hall.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons.

Fonagy, I., & Magdics, K. (1960). Speed of utterance in phrases of different lengths. *Language and Speech*, *3*(4), 179–192.

Fox, John, & Weisberg, S. (2018). *Predictor Effects Graphics Gallery*.

Fox, John, & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Goldman-Eisler, Fridea. (1954). On the variability of the speed of talking and on its relation to the length of utterances in conversations. *British Journal of Psychology. General Section*, *45*(2), 94–107.

Goldman-Eisler, Frieda. (1954). A study of individual differences and of interaction in the behaviour of some aspects of language in interviews. *Journal of Mental Science*, *100*(418), 177–197.

Goldman-Eisler, Frieda. (1961). The significance of changes in the rate of articulation. *Language and Speech*, *4*(3), 171–174.

Goldman-Eisler, Frieda. (1968). *Psycholinguistics: Experiments in spontaneous speech*.

Goswami, U., & Leong, V. (2013). Speech rhythm and temporal structure: converging perspectives. *Laboratory Phonology*, *4*(1), 67–92.

Greenberg, S. (1999). Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*(2–4), 159–176.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, *31*(3–4), 465–485.

Grosjean, F. (1980). Comparative studies of temporal variables in spoken and sign languages: A short review. *Temporal Variables in Speech*, 307–312.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Johnson, K. (2004). Massive reduction in conversational American English. *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, 29–54. Citeseer.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kendall, T. S. (2009). *Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project* (PhD Thesis).

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York, NY: Cambridge University Press.

Linguistic Convergence Laboratory. (n.d.). *Corpus of Qakh Dialect of the Azeri Language*. the School of Linguistics of NRU HSE.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*(4), 215–225.

Moroz, G. (2017a). lingtypology: easy mapping for Linguistic Typology. Retrieved from https://CRAN.R-project.org/package=lingtypology

Moroz, G. (2017b). Скорости речи носителей кубанского диалекта кабардино-черкесского языка: устный дискурс vs. чтение текста [Speech rate of speakers of Kuban dialect of the Kabardian language: oral discourse vs. text reading]. *Томский Журнал ЛИНГ и АНТР. Tomsk Journal LING & ANTHRO*, *2*(16), 9–17.

Navarro, D. (2019, January 11). Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6.1). Retrieved 4 June 2019, from https://learningstatisticswithr.com/book/

Ovsyannikova, M., Say, S., Aplonova, E., Smetina, A., & Sokur, E. (2017). The spoken corpus of the Bashkir language of Rakhmetov and Baimovo villages. Retrieved 11 March 2019, from https://linghub.ru/oral_bashkir_corpus/

Paul (https://stats.stackexchange.com/users/11646/paul). (2017, April 17). What is the difference between fixed effect, random effect and mixed effect models? Retrieved from https://stats.stackexchange.com/q/151800

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, *87*(3), 539–558.

Pimsleur, P., Hancock, C., & Furey, P. (1977). Speech rate and listening comprehension. *Viewpoints on English as a Second Language*, 27–34.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, *123*(2), 1104–1113.

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. Retrieved from https://www.R-project.org/

Roach, P. (1998). Some Languages are Spoken More Quickly Than Others. In *Language Myths* (pp. 150–158). Retrieved from http://w3.salemstate.edu/~jaske/courses/readings/Some_Languages_are_Spoken_More_Quickly_Than_Others_By_Peter_Roach.htm

Stepanova, S. (2011). Russian Spontaneous Speech Rate-Based on the Speech Corpus of Russian Everyday Interaction. *ICPhS*, 1902–1905.

Tauroza, S., & Allison, D. (1990). Speech Rates in British English. *Applied Linguistics*, *11*(1), 90–105. https://doi.org/10.1093/applin/11.1.90

The Multimedia Corpus of the Chukchi Language. (2018). Retrieved from http://chuklang.ru/corpus

The Spoken Corpus of the Beserman Dialect of the Udmurt Language. (2018). Retrieved from http://multimedia-corpus.beserman.ru/search

Van Rossum, G., & Drake, F. L. (2011). *The python language reference manual*. Network Theory Ltd.

Van Santen, J. P. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*(6), 513–546.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. *Ninth International Conference on Spoken Language Processing*.

Zellner, B. (1998). *Fast and Slow Speech Rate: a Characterisation for French*. *7,* 3159–3163. Retrieved from http://cogprints.org/893/

Zinder, L., & Maslov, J. (1982). *Л.В.Щерба – лингвист-теоретик и педагог [L.V.Shcherba – the linguist theorist and teacher]*. 'Наука', Ленинградское отд-ние ['Nauka', St. Petersburg department].

Appendix 1

Link to the repository of the present work on GitHub: https://github.com/maria-terekhina/Diploma. Via the link, you can find programming code used for processing and analysis of the data.

Appendix 2

The calculation for speech rate change in accordance with the Model 3 predictor effect estimation. The full version of Table 13.

Initial length, syll.	Adjusted length, syll.	Change in length, %	Value to add, syll./sec	Initial speech rate, syll./sec	Adjusted speech rate, syll./sec	Change in speech rate, %
1.0	2.0	100.0	0.86	1.55	2.41	55.35
2.0	3.0	50.0	0.43	2.14	2.57	20.08
3.0	4.0	33.33	0.29	2.61	2.9	10.99
4.0	5.0	25.0	0.22	2.9	3.12	7.41
5.0	6.0	20.0	0.17	3.13	3.3	5.49
6.0	7.0	16.67	0.14	3.41	3.56	4.2
7.0	8.0	14.29	0.12	3.57	3.69	3.44
8.0	9.0	12.5	0.11	3.69	3.8	2.91
9.0	10.0	11.11	0.1	3.83	3.92	2.5
10.0	11.0	10.0	0.09	3.9	3.98	2.21
11.0	12.0	9.09	0.08	3.99	4.07	1.96
12.0	13.0	8.33	0.07	4.04	4.11	1.77
13.0	14.0	7.69	0.07	4.1	4.16	1.61
14.0	15.0	7.14	0.06	4.1	4.16	1.5
15.0	16.0	6.67	0.06	4.15	4.2	1.38
16.0	17.0	6.25	0.05	4.16	4.21	1.29
17.0	18.0	5.88	0.05	4.19	4.24	1.21
18.0	19.0	5.56	0.05	4.28	4.33	1.12

19.0	20.0	5.26	0.05	4.39	4.43	1.03
20.0	21.0	5.0	0.04	4.4	4.44	0.98
21.0	22.0	4.76	0.04	4.45	4.49	0.92
22.0	23.0	4.55	0.04	4.41	4.45	0.89
23.0	24.0	4.35	0.04	4.41	4.45	0.85
24.0	25.0	4.17	0.04	4.48	4.51	0.8
25.0	26.0	4.0	0.03	4.45	4.49	0.77
26.0	27.0	3.85	0.03	4.48	4.52	0.74
27.0	28.0	3.7	0.03	4.47	4.5	0.71
28.0	29.0	3.57	0.03	4.43	4.46	0.69
29.0	30.0	3.45	0.03	4.48	4.51	0.66
30.0	31.0	3.33	0.03	4.47	4.5	0.64
31.0	32.0	3.23	0.03	4.57	4.6	0.61
32.0	33.0	3.12	0.03	4.56	4.59	0.59
33.0	34.0	3.03	0.03	4.59	4.61	0.57
34.0	35.0	2.94	0.03	4.31	4.33	0.59
35.0	36.0	2.86	0.02	4.48	4.51	0.55
36.0	37.0	2.78	0.02	4.4	4.43	0.54
37.0	38.0	2.7	0.02	4.57	4.6	0.51
38.0	39.0	2.63	0.02	4.51	4.53	0.5
39.0	40.0	2.56	0.02	4.61	4.63	0.48
40.0	41.0	2.5	0.02	4.52	4.54	0.48
41.0	42.0	2.44	0.02	4.33	4.35	0.48
42.0	43.0	2.38	0.02	4.68	4.7	0.44
43.0	44.0	2.33	0.02	4.39	4.41	0.46
44.0	45.0	2.27	0.02	4.68	4.7	0.42
45.0	46.0	2.22	0.02	4.51	4.53	0.42

46.0	47.0	2.17	0.02	4.5	4.51	0.42
47.0	48.0	2.13	0.02	4.73	4.75	0.39
48.0	49.0	2.08	0.02	4.57	4.59	0.39
49.0	50.0	2.04	0.02	4.87	4.89	0.36
50.0	51.0	2.0	0.02	4.41	4.42	0.39
51.0	52.0	1.96	0.02	4.81	4.83	0.35
52.0	53.0	1.92	0.02	4.5	4.51	0.37
53.0	54.0	1.89	0.02	4.82	4.83	0.34
54.0	55.0	1.85	0.02	5.18	5.19	0.31
55.0	56.0	1.82	0.02	4.74	4.76	0.33
56.0	57.0	1.79	0.02	4.86	4.88	0.32
57.0	58.0	1.75	0.02	4.7	4.72	0.32
58.0	59.0	1.72	0.01	5.77	5.78	0.26
59.0	60.0	1.69	0.01	4.82	4.83	0.3
60.0	61.0	1.67	0.01	5.26	5.27	0.27
61.0	62.0	1.64	0.01	4.35	4.37	0.32
62.0	63.0	1.61	0.01	6.41	6.42	0.22
63.0	64.0	1.59	0.01	5.1	5.11	0.27
64.0	65.0	1.56	0.01	5.26	5.28	0.26
65.0	66.0	1.54	0.01	5.21	5.23	0.25
66.0	67.0	1.52	0.01	5.42	5.43	0.24
67.0	68.0	1.49	0.01	5.63	5.64	0.23
68.0	69.0	1.47	0.01	5.73	5.74	0.22
70.0	71.0	1.43	0.01	5.02	5.04	0.24
71.0	72.0	1.41	0.01	5.36	5.37	0.23
72.0	73.0	1.39	0.01	5.87	5.89	0.2
78.0	79.0	1.28	0.01	5.75	5.76	0.19