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»

Кошевой Алексей Глебович

ПАРАДОКС НАБЛЮДАТЕЛЯ НА МАТЕРИАЛЕ ИНТЕРВЬЮ КОРПУСА БАССЕЙНА РЕКИ УСТЬЯ

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ152

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« » _____ 2019 г.

Научный руководитель

К.ф.н., профессор
М. А. Даниэль

_____

ОГЛАВЛЕНИЕ

1. Introduction

In this paper[1] I will present the analysis of the interviews from the Ustja River Basin Corpus (von Waldenfels et al., 2014) — a collection of field recordings of the speakers of Mikhalevskaja dialect spoken in Arkhangelsk Oblast'— in order to detect the effect of the so-called Observer's Paradox (Labov, 1972: 209); the idea that the recorded interviews can possibly differ from the normal day-to-day unobserved speech by the parameters that are studied. I will analyse possible correlations between the frequencies of the dialectal and standard realizations of a given linguistic variable with the time of the interview, the gender of the subject, his/her year of birth and other factors. The possible findings may help to better understand the nature of the Observer's Paradox as well as to come up with new methodologies as to how control for its effect in the analysis of linguistic data.

Interviews are an important source of data in different domains of research dealing with (socio-)linguistic variation. For many years, there has been a debate about ecological validity of the data obtained by recording. In sociolinguistic domain, some researchers argue that, while recorded interviews are not inherently bad as a source of linguistic data, they may represent a special speech style (see, for example Labov, 1972; Wilson, 1994: Wetheim, 2002 etc). This idea is an important starting point in the analysis of the data obtained using interviews.

Since the introduction of the Observer's Paradox (cf Labov, 1972: 209, the idea that all the recorded speech data is different from the unobservable 'regular' speech, and possible solutions to that problem proposed in the same work) there had been many publications addressing the same issue. For example, in a recent work (MacLeod & Grant, 2016) where some examples of dialogues in the online messengers, the participant of which were aware of the fact that their dialogue will be read by the

[1] I am very grateful to all the people who have helped me with this study. I would like to especially thank my supervisor Michael Daniel for his support and enormous help, all the staff of the Linguistic Convergence Laboratory and especially George Moroz, Ilya Chechuro and Nina Dobrushina for their criticism, fruitful comments and suggestions. I am also grateful to Katya Gerasimenko who provided me with all the data and helped with my understanding of the Ustja River Basin corpus database.

researchers, were discussed in the light of the Observer's Paradox. The authors have noticed that when the participants of a dialogue are aware of them potentially being observed by someone else, they try to adapt their speech style to the image of this third-party observer (usually it was known by both speakers). This observation is somehow parallel to the audience design framework (Bell 1984), in which the main thesis is based on the idea that the person's speech style is formed according to his/her listener. This framework will be discussed later in the present study. In the literature which addresses the effect of the recording and the presence of the fieldworker on the speech production of the observed individual, the change of speech style are interpreted as recorded speech simply represents another style (different from other styles such as casual, formal etc., see Wertheim, 2002; Wilson, 1994)

The main goal of this study is to test the current theoretical assumptions about the validity of the data obtained during the sociolinguistic interviews. In the case of the Ustja River Basin corpus, there is a big variation on the speaker level of the usage of dialectal or standard realization within the same variable (see Daniel et al., in press: 15). The presence of this variation arises the question of how exactly the possible change in the distribution of dialectal and standard realizations of a given variable changes across time. If there is such time-related change, it is interesting to see, whether it shaped primarily by the fact that the speaker can lose self-caution during the long interviews and start to use the style of speech which is more casual than the one he had used in the beginning in the interview. Or whether this possible change is motivated by the presence of the interviewer who was usually directly involved in the conversation, so he/she can motivate the possible change of style towards the more standard, or more dialect (same phenomenon as the "on-stage" style in (Wertheim, 2002) study of Russian-Tatar bilinguals). Another idea that will be tested using this data was proposed in (Sankoff & Blondeau 2007: 9), where the apparent-time study of /r/ in Montreal dialect of French was presented. They argued that the first 10 minutes of the interviews should be discarded, as they often provided unstable data.

Here, I will introduce the notion of dialectness – the probability of an observation of a particular variable being dialectal. For example, the realizations of a variable X can be

either x-1 (dialect) or x-2 (standard). In a given file (or subset of files for given speaker) Z, there is n observations on X being dialectal (x-1), and k observations being standard (x-2), so, the dialectness for a given file Z is computed as follows[2]:

$$dialecntess(Z) = P(dialect) = \frac{N(dialect)}{N(dialect) + N(standard)}$$

I will propose that the dialectness of the observed person will vary across time, especially towards the end, when the observed person will potentially become more self-aware and will start to speak more casually, meaning that he/she can possibly use more standard realizations than the dialectal ones. Another hypothesis will be based on the audience design principles and will propose the change towards more standard style in the end. This hypothetical change in dialectness of a certain person could be possibly motivated by different factors including but not limited to prestige of standard Russian and negative attitude towards Russian dialects formed in the Soviet society, as dialects have been identified as pertaining to the peasant culture which was highly disliked by Soviet officials in the early stages of the USSR[3].

In Section 2 I will discuss the theoretical background for my work. In Section 3 the data which is used in the present study would be described. In section 4 will be the analysis. In Section 5 the analysis results would be summarized and discussed. In section 6, the summary of all the work done will be given. Section 7 will contain the litterature used and the appendix will be presented in the Section 8.

2. Background

One of the main challenges of the sociolinguistic research are the possible differences between spontaneous speech and recorded speech produced during sociolinguistic interviews. This challenge is described by the so-called 'Observer's Paradox', formulated in (Labov, 1972: 209);

[2] The dialectness can be also computed for all the variables in a given file
[3] Alexandra Ter-Avanesova, personal communication

"*We are then left with the **Observer's Paradox**: the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation.*"

Labov also notes that this paradox can be partially overcome: the interview may be done with some pauses, so that the interviewee may become less focused on the fact that he or she is being recorded, etc (see Labov, 1972: 92). The effects similar to the Observer's Paradox (although the author did not address this formulation of the problem) can be interpreted in terms of style, using the Style Axion formulated in (Bell, 1984: 151);

"Variation on the style dimension within the speech of a single speaker derives from and echoes the variation which exists between speakers on the "social" dimension."

Bell argues that the image of the listener created by the speaker is reflected in his (speaker's) speech. In order to describe these effects, Bell proposed that the speakers are taking their hearers into account when producing their speech. He proposes the hierarchy of different audience roles based on the following factors: how the person is addressed by the speaker, is the person known to the speaker, how is he/she ratified by the speaker. The resulting hierarchy can be shown using the table below:

Table 1: audience role hierarchy (adapted from pell Bell 1984: 160):

role	known	ratified	addressed	person choice
speaker	+	+	-	1-st
addressee	+	+	+	2-nd
auditor	+	+	-	3-rd
overhearer	+	-	-	-
eavesdropper	-	-	-	-

The role which can strongly affect the variation of the speaker during the interaction, is, presumably, the addressee role as the addressee is the person which is directly involved in the communication with the speaker. Bell indicates that the persons who are occupying the highest position on this hierarchy, can strongly affect the variation of speech produced by the speaker.

This idea could be potentially used when analysing the data examined in this study, as it may explain the possible change in dialectness over time of the interview. This change could be influenced by the fact that the interviewers were usually speakers of standard Russian which is more prestigious over the dialectal variety spoken in the area. So we can expect two hypothesis of change on the global level:

(a) Speakers are affected by the presence of the interviewers which results in the adaptation of their speech style to the image of the listener (as the interviewers can be considered as addresses or auditors which makes them highly influential on the observed person's style). This can hypothetically result in a negative change of the dialectness (i.e. from the less dialect in the beginning and to the more dialect from the mid-section of the interview).
(b) Speakers will lose self-caution during the time of the interview, as they will try to adapt to the listener, but then start to speak more casually which will result in a positive change in dialectness.

Here, the works which are dealing with direct examination of the Observer's Paradox and the effects of Interviewers on the speech production of individuals. One example is (Wilson, 1994). The author notes that the recording device is usually personalized as a human recipient by the speakers. This results in various sorts of auto-correction and lexical choices being different than usual ones for the speaker. He notes that in most cases he had studied, the recording device was personalized as a researcher which resulted in a more formal style of speech, auto-correction of swear words etc. Wilson concludes that these effects should be studied separately as instances of style variation and the data should not be considered as "dirty" by any means.

(Wertheim, 2002) analyses the data collected in Tatarstan and contains the speech obtained from Tatar-Russian bilingual subjects. In Tatarstan two equally prestigious major languages are present: Russian and Tatar, so the data could not be analyzed in the image of the vernacular-standard variants (the paradox formulated in Labow works was based on his observations on the distinction of the standard American English and the African American vernacular variant). The first crucial finding is the fact that all speech analyzed differed significantly from the casual, day-to-day speech of the observed individuals. The author uses (Bell 1984) audience design framework in order to explain the observed variation of styles. Wertheim noticed that in most of the interviews she had recorded, the speakers were usually highly affected by her presence when she was in a role of the addressee which resulted in the observed speakers choosing the simplified, "on-stage" Tatar over Russian (the language also known to the author). However, she notices that the observed persons are using different styles for different situations, following the audience design principles, in the situation when the observer is not an addressee, but the auditor or an overhearer. The speech was always adjusted accordingly to the addressee. When the observed persons were speaking to their family, they used to do code-switching between Russian and Tatar. When they talked to the service personnel, they used Russian with some addition of the Tatar. However, when the author was in the role of the auditor and observed the family conversations, she noticed that the speaker was using mostly on the "on-stage" Tatar, as in the situation when she was the addressee. These observations lead to a conclusion that the observation of style

changes depending on the audience role attributed to the fieldworker are crucial to the collection and the interpretation of the sociolinguistic data. She also indicates that the speech style used when the fieldworker is in the role of addressee or auditor is as unmarked, as other styles are.

Another way of analyzing the data gathered under the observation is presented in (MacLeod & Grant 2016). The data described in this study are similar to the data from (Wertheim, 2002), for example. The authors analysed different types of dialogues where the participants were aware of the fact that they are being recorded and then someone will access their dialogue. They indicated that the participants have constructed the image of the potential reader in order to regulate their speech styles and their theme choice. They indicate that the fact that the person is being observed during the recording should not produce any constraints on "natural" language production, meaning that the speech produced under those circumstances is as "natural" as the causal speech. However, the data should be analysing bearing in mind that there could be some potential style changes, based on the principles of the audience design theory (Bell, 1984).

All of the studies discussed above are indicating that the person's speech is shaped accordingly to the listener which results in a difference of styles from setting to setting. The potential style changes occurring during a interview are usually explained by the frameworks describing the effects of the potential listeners based on their roles in the conversation (see Bell, 1984: Giles & Ogay, 2007). However, these changes were never studied quantitatively or linked with the other parameters of the interview, such as the elapsed time and the timestamp in which the observed variables are occuring in the interview.

3. Data

The data used in this study is a subset of the data from the Ustja River Basin corpus (see von Waldenfels et al., 2014). It is a moderately small dialectal corpora (around 215.000 tokens, of which around 180.000 belongs to speakers of the Ustja dialect of Russian

language). The dataset used in this project contains a sample of 178 interviews from the corpora (belonging to 135 different speakers) which were annotated by different people in order to retrieve information about particular variables, 13 in total, concerning primarily phonetics (see full list in Daniel et al., in press). For each variable, the type of its realization (being either dialectal or standard) was recorded. The overall dataframe contains 24.563 observations.

3.1. Structure of the initial dataset

Initial dataset contained standardized observations on usage of 13 variables discussed above. The file annotated for each variable file had the following structure:

Fig. 1: ae-variable description

	A	B	C	D	E	F	G	H	I
	N	speaker	audio	link	context	token	realiz	realizati	class
148606		авм1922	20130626d-avt	http://www.pa	Дак а больше никого не допускали, там их #обучали# .	обучали	e	cons	
306582		авм1922	N20130624b2	http://www.pa	О - ой, всяко... Сто человеков ходило, из ни = ой... ой, #бл	блядь	a	inn	
53203		авм1922	20130701d-avt	http://www.pa	Сколько годов - то робит - то, уж четыре года, не #пять# л	пять	a	inn	
53864		авм1922	20130701d-avt	http://www.pa	С #матерями# , да внучата, да ведь.	матерями	a	inn	
53917		авм1922	20130701d-avt	http://www.pa	Ну, #глянется# ли старому человеку самому уж что в чужс	глянется	e	cons	
54124		авм1922	20130701d-avt	http://www.pa	Как отдельно дак эти бы #взяли# да увезли.	взяли	e	cons	
54399		авм1922	20130701d-avt	http://www.pa	Мне - то уж нужда #заставляет# - то, и даю денежек - то к	заставляет	e	cons	
54809		авм1922	20130701d-avt	http://www.pa	Петю, того из Ленинграда #взяли# в Чечню.	взяли	e	cons	
55625		авм1922	20130701d-avt	http://www.pa	Женка, у неё осталось #пятеро# робят.	пятеро	e	cons	
56743		авм1922	20130701d-avt	http://www.pa	#Глянется# ?	Глянется	?	ncl	
56745		авм1922	20130701d-avt	http://www.pa	#Глянется# - глянется?	Глянется	?	ncl	
56747		авм1922	20130701d-avt	http://www.pa	Глянется - #глянется# ?	глянется	e	cons	
57300		авм1922	20130701d-avt	http://www.pa	Ой, дак не сплю - то я - то #ночами# - то.	ночами	a	inn	

One row in the dataset shown in Fig. 1 represents one particular observation of this variable in a given text. Overall, the tables contained data about the speaker who had produced the given observation; the token, in which this variable is present; the realization class of the variable (dialect, standard or non-classified); the wider context etc. However, this data did not fully respond my needs, so some preprocessing was performed.

3.2. Data preprocessing

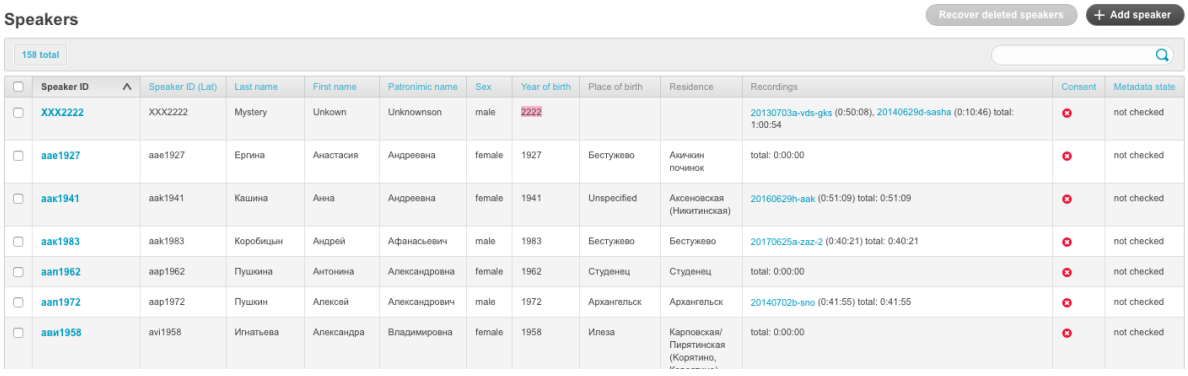
Firstly, the timestamp of the exact realization was obtained from the link to the audio in the corpus. For example, the following link *http://www.parasolcorpus.org/Pushkino/OUT/20130701d-avm-1549155-1551459.wav* contains two timestamps (*1549155-1551459*); the offset from the the beginning of the audio file in ms for the beginning and the end of the realization of a variable. Firstly, these two numbers were retrieved using the following regular expression:

(\d+)-(\d+).wav

Then, the mean between these two numbers were computed in order to retrieve the relative value (in ms) of where the realization happens during the interview.

Additional speakers metadata, such as year of birth, place of birth and gender was retrieved from the Ustja River Basin corpus database. The data contained there looks as follows:

Fig. 2: Ustja River Basin corpus metadata



Also, as the longer files were sometimes separated in parts, those parts were found and their timestamps were added up accordingly (if there was 2 separate files the timestamps of the second were added to the timestamp of the ending of the first and so on and so forth). The encodings of the realization class which were tripartite in the

initial data (innovative (dialectal), conservative (standard) and non-classified) were transformed to binary (dialectal, standard) by deleting the non classified variants. The length of the files was added by finding the max values of timestamp for each separate file (all observations were taken into consideration).

These preprocessing resulted in the following dataset:

1. Speaker — individual code of the speaker
2. Year of birth – year of birth of the speaker
3. Gender – gender of the speaker
4. Audio — name of the audio file containing the recording (as the data contains splitted files this field will be useful for gluing the cuted files together)
5. Context — context, in which certain variable was pronounced
6. Token — the exact token which contains the variable
7. Realization class — dialectal or standard
8. Variable — exact variable used
9. Timestamp — offset from the beginning of the file of the position, in which given variable occurs in the interview (in milliseconds)
10. Interviewer — the person who has taken the following interview
11. Length – length of the audio file

3.3. Interviewers involvement in the recordings

As I will use the audience design principles in order to explain the findings, I will show in this section that the interviewers were usually involved in the conversation in the role of the addressee (they were known to the speaker, they were addressed using the 2nd person etc.). Firstly, some extraits from the corpus will be analyzed (the translations will be given in brackets, Sp. denotes “speaker”, and Int. denotes “interviewer”):

(1) Sp: Архангельск-то знаете, наверное? (You know Arhangelsk, right?)
Int: Конечно, знаем. (Yes, we sure know it.)
Sp: Или **вы** от= или **вы** оттуда? (And you are.. and you are from there?)

Int: С Питера. (From St. Petersburg.)

Sp: А, **вы** с Питера? Так а где **вы** преподаете, это самое, в филологическом университете что ли? (Ah, you are from St. Petersburg? So you are teaching at, how is it called, the university of litterature studies?)

Int: Да. (Yes.)

<...>

The example (1) shows that the interviewer is fully engaged in the interview. He is answering different questions of the speaker, and the speaker addresses him using the 2nd person plural (*вы*). Another example of speaker-interviewer interaction could be the observations where the interviewer is adding some clarification question to the speaker's presentation:

(2) <...>

Sp: Назову-ка я девушку-то Любой. В честь этого врача меня и назвала меня Любой. (I will name the girl Lyba. I was named Lyba in honor of this doctor.)

Int: Так они не расписаны были да? со Степаном с этим? (So they were not formally married with Stepan?)

Sp: Нет. Нет, нет, нет. Не расписана. У меня фамилия была отцова. (No. They were not. I had my father's surname.)

<...>

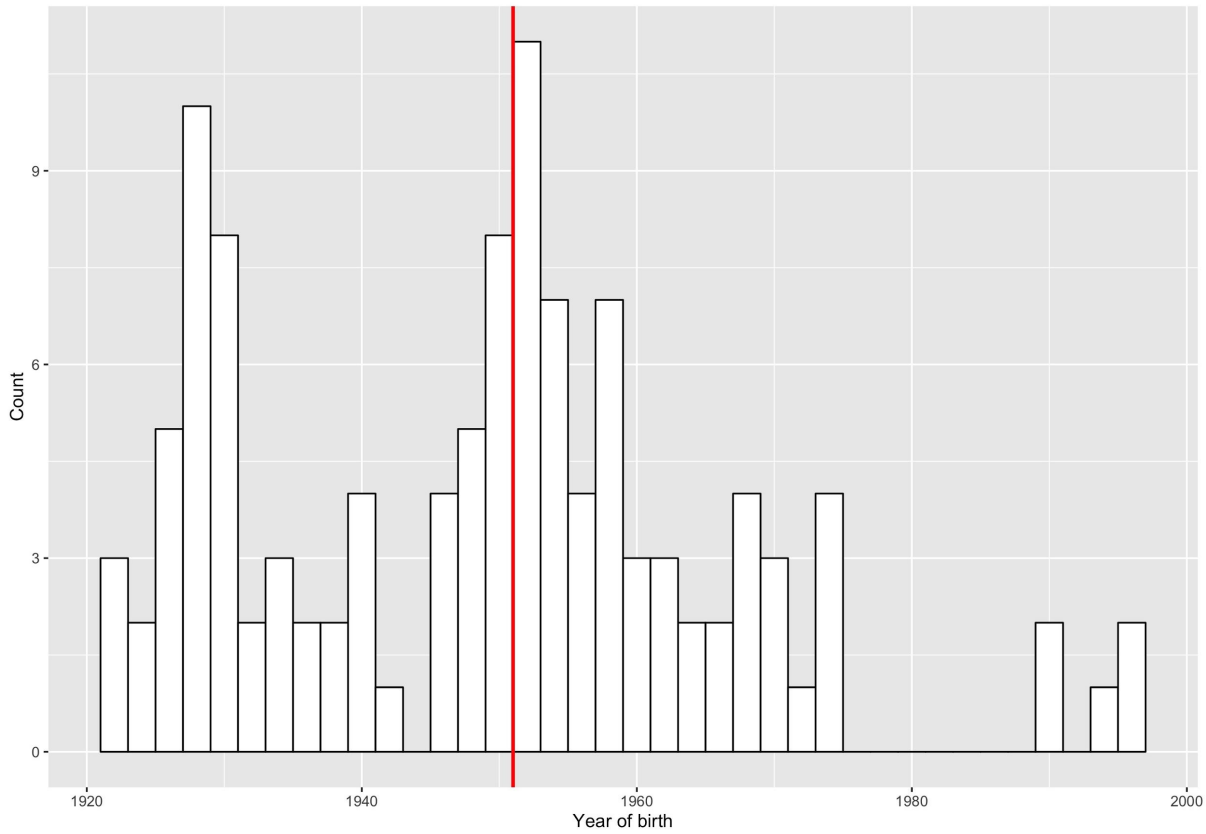
In the example (2), the interviewer tries to clarify some details of the story. The examples are illustrating that usually interviewers were actively involved in the conversation, either by asking some clarification questions (2) or by being directly engaged in a dialog as in (1). This fact may indicate that in those files the interviewer is usually in the role of addressee, following the Bell's audience design framework. Another argument for the assumption that usually all the interviewers were in the position of the addressee could be obtained from the corpus statistics. In the Ustja River Basin Corpus, roughly 16% of tokens (35.000 out of 215.000, for more detailed description see von Waldenfels et al., 2014) belong to the interviewers. This means that

they were strongly involved in the interviews, and usually were not in the roles such as overhearrer (the person who is not directly involved in the talk and is not known to the speaker(s)), but rather as the auditor or the adressee.

3.4. Descriptive aspects

In order to provide a better understanding of the data, in this section I will propose its descriptive analysis. Firstly, the distribution of the year of birth was plotted:

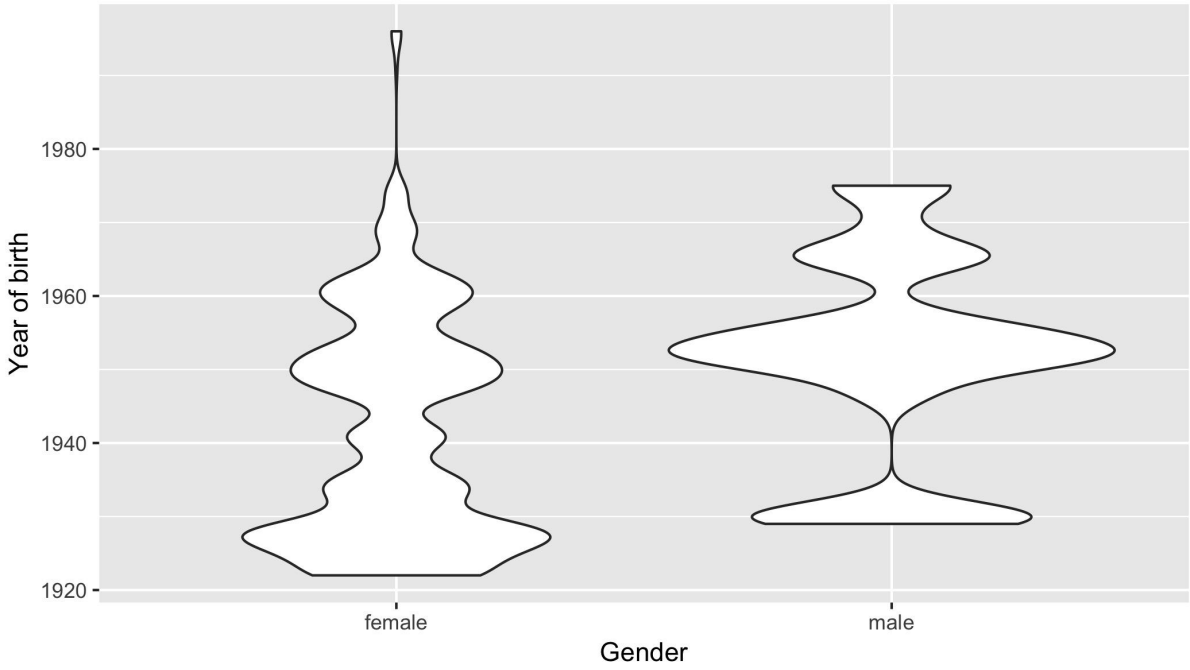
Fig. 3: Year of birth distribution (median plotted in red)



As can be seen from Fig. 3, there are clearly three different age cohorts; one in the range 1920-1940, second in the range 1945-1970 and third in the years between 1990 and 2000's. This will be used in the proposed analysis.

The gender distribution is not balanced. There were 18,121 observations of different variables coming from female speakers and 6,442 observations from male speakers. The distribution of genders alongside the year of birth is shown at Fig. 4:

Fig.4: Gender and year of birth violin plot

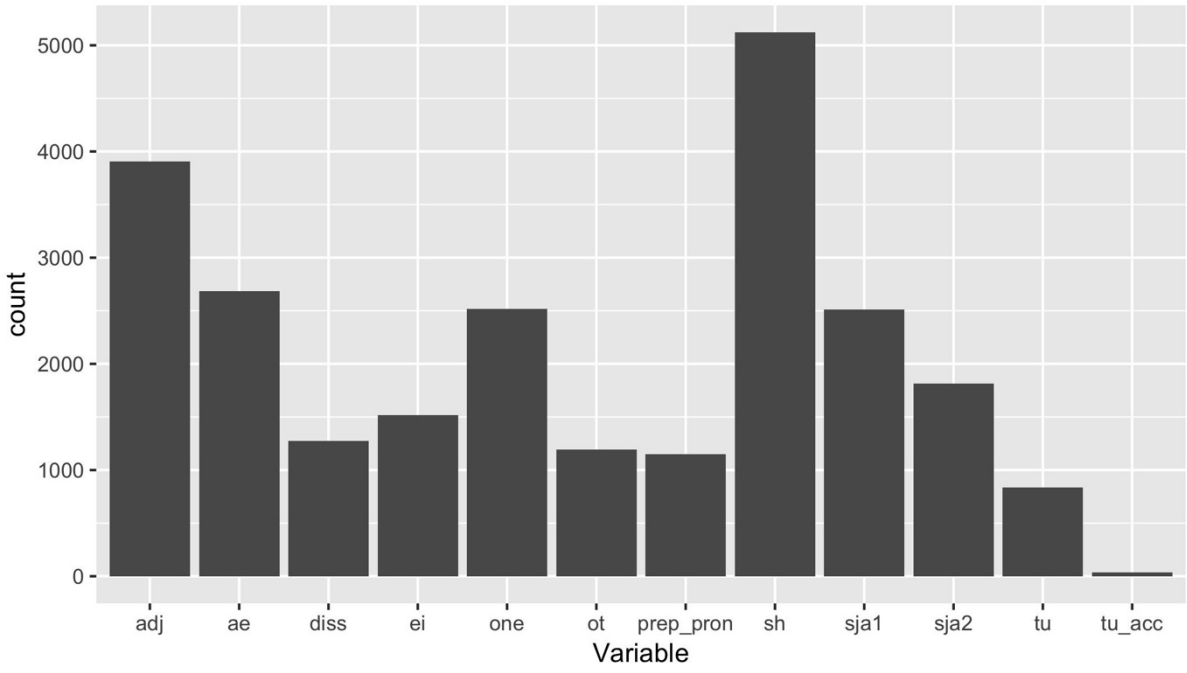


This figure clearly shows that the data obtained from female speakers is distributed more evenly across the year of birth as opposed to the male speakers.

In the list below, I will briefly describe the variables which are presented in the data (the detailed description of variables is provided in (in Daniel et al., in press)):

The variables presented in the survey are distributed as follows:

Fig. 5: variable distribution



The prevailing quantity of variables are usually phonetical by their nature; the difference in sound realization (sh), the differences in adjectival endings (adj) and others have most occurences. The morphological variables are usually much infrequent; the differences in postpositions such as tu and tu_acc are much less common.

4. Analysis

4.1. Individual interview file analysis

The main goal of this section is to demonstrate that the timestamp-related effects may appear on a small-scale level while it may possible be hard for the models which are making generalizations over all of the data (such as logistic regression) to catch the tendencies present in the dataset. I will address the hypothesis that the first 10 minutes of a recording are usually different from the rest of the recording, so this part should be discarded (see Sankoff & Blondeau 2007: 9).

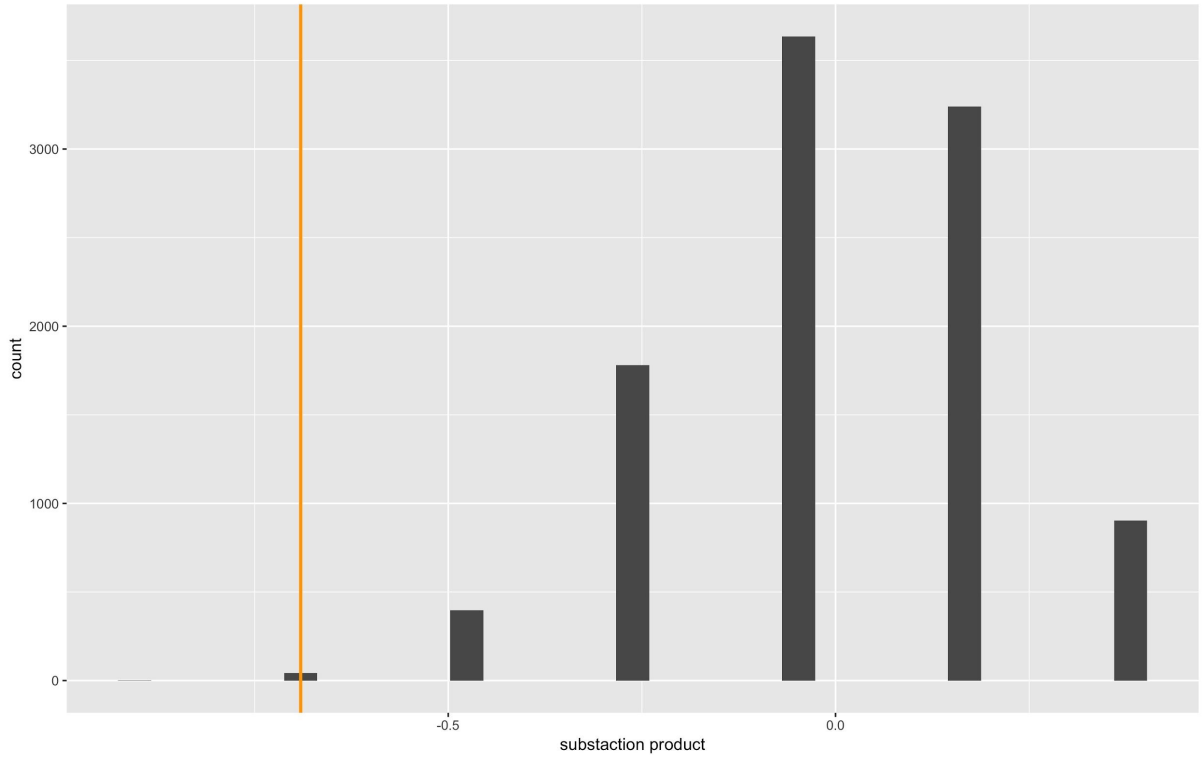
In order to determine which observations are significant the following method is used. For each file, the first 10 minutes entries and the entries after the 30 minutes mark are shuffled between each other. For each randomized group, the dialectness is computed and then subtracted in this manner:

result = dialectness of the first part - dialectness of the second part

This procedure is repeated for 10000 times. The observed subtraction of dialectness of first 10 minutes and 30 minutes plus part will also be computed. Then, the distribution of possible subtractions is plotted against the observed value: if the real subtraction result fall within the subtraction products with a low probability of occurrence (less than 0.05), it indicates that the given observation has a big likelihood of occuring in a random sample. If not, then the observed change in dialectness could be considered as credible.

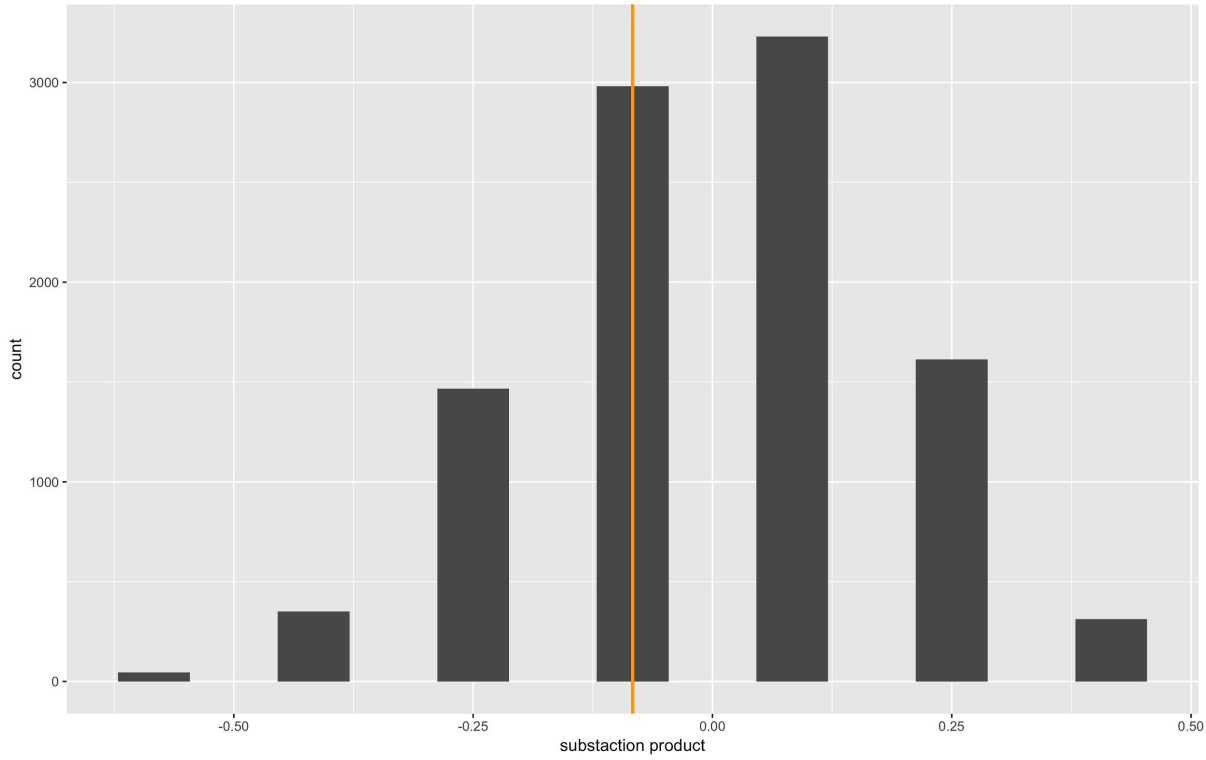
I will give two examples in order to illustrate how the algorithm works. In the first part (0-10 minutes) of the example file, there is 6 observations, 5 of them being standard and 1 being dialect. In the second part (30+ minutes), there are 21 observations, 3 of which are standard and 18 dialect. The dialectness of the first part is ~0.167, the dialectness of the second is ~0.868. This observation is suggesting that the speaker is clearly altering his behaviour during the interview, as he is more dialectal in the second part, than in the first. It is possible to check whether these results can be random:

Fig. 6: example subtraction shuffling plot: different means (fake data



The results of this experiment are indicating that the observed value is falling into the value which is less likely to be random than all other possible outcomes (subtraction products). The probability of the randomized subtraction value to be equal to the observed one is 42/1000 = 0.0042, which is very low. So, the difference between dialectness of the second and first parts could be considered significant, and it can be concluded that the speaker is more dialect in the 30+ minutes part than he is in the beginning of the interview. Another example will have the subsets with similar values of dialectness in order to show that the given method will not signal the difference in two subsets. The first subset will contain 8 observations 5 of which will be dialect and 3 standard. The second subset will contain 24 observations: 17 dialect and 7 standard. The dialectness first subset is 0.625 and the dialectness of the second -- ~0.709 . So, the subtraction product will be equal to –0.084 which in its turn will mean that there is a little style-change.

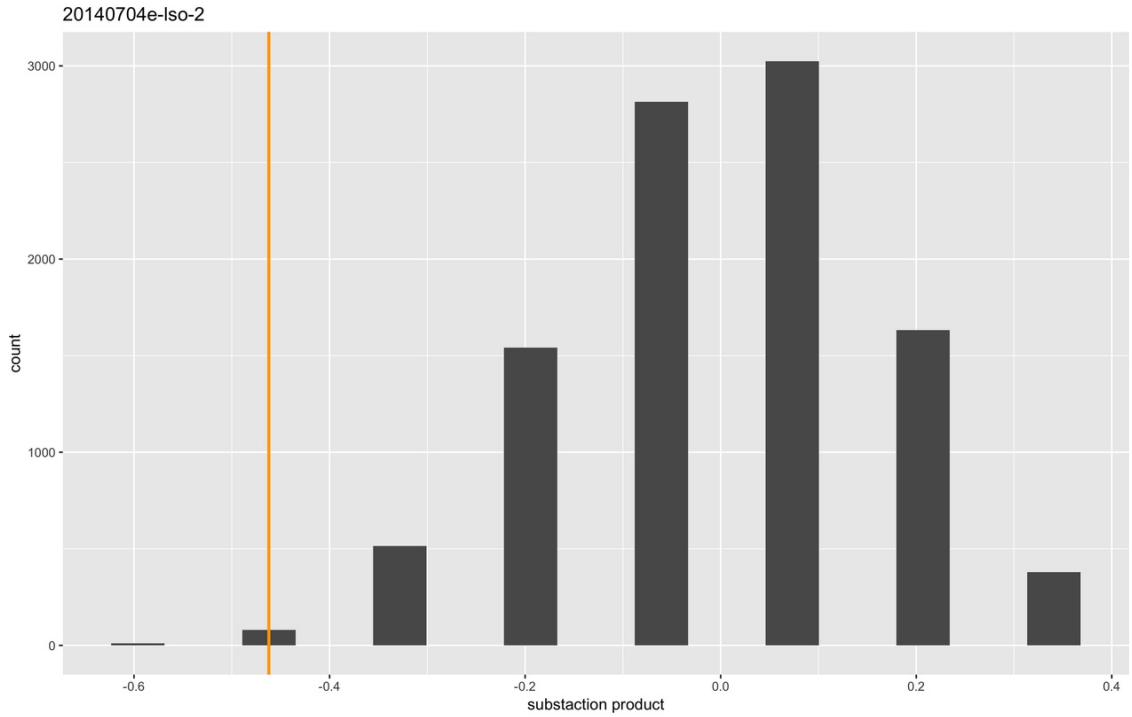
Fig. 7: example subtraction shuffling plot: same means (fake data)



The example plot is also indicating that the observed subtraction product is most likely random, as the probability of observed subtraction value being random is 2982/10000 = 0.2982. Therefore, for this file, the significant change in dialectness between the two subsets could not be postulated.

Overall, files that have shown significant difference between first 10 minutes and the 30+ minutes segment are containing 8 different variables: *sh, one, adj, prep_pron, sja1, sja2, ei* and *ot*. Firstly, the *sja1* variable will be analyzed. Among all the files containing observations about this variable, only one has shown credibility according to the test proposed above. The results can be observed in the plot below:

Fig. 8: distribution of generated subtraction products for the file 20140704-lso-2



The file 20140704-lso-2 features the female speaker born in 1941 (individual code лсо1941). The figure above suggests that the real subtraction value is offset from the main random distribution peaks which suggests that this value has a small chance of occurring in a random setting. The real value is equal to -0.4621212 which indicates that in the given file the speaker tends to be less dialectal during the first 10 minutes than in the rest of the recording.

Then, the *adj* variable is analyzed. In the analyzed subset the difference between means have been found significant in six different files. I will analyze two of them in greater detail. The files are recorded from female speakers born in 1952 (I20130623b1) and 1922 (20130701d-avm). The significance test results are presented below for some of the interviews:

Fig. 9: distribution of generated subtraction products for the file I20130623b1

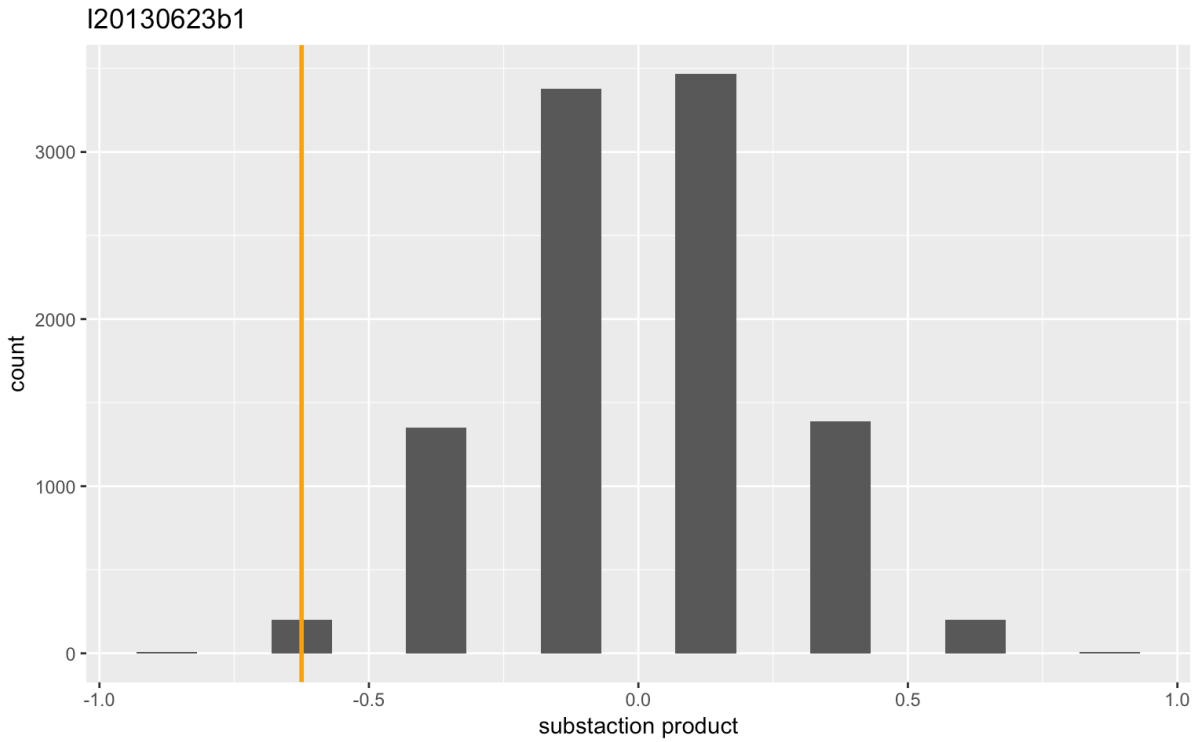
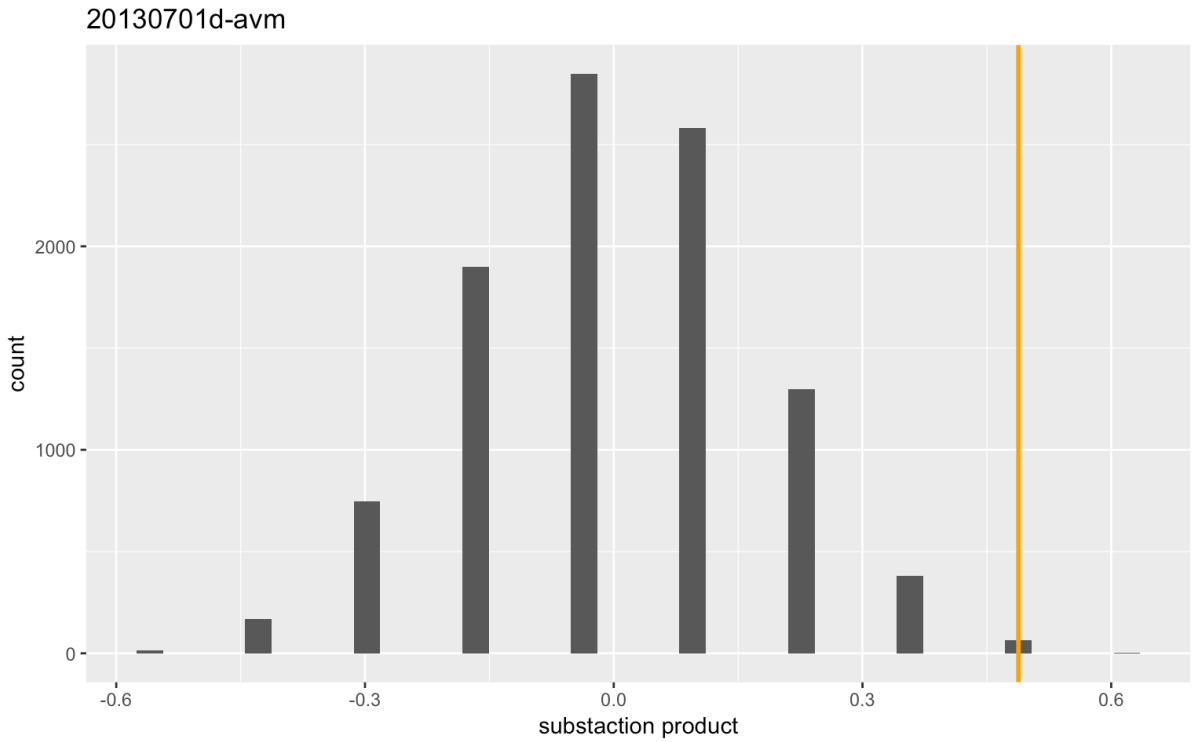


Fig. 10: distribution of generated subtraction products for the file 20130701d-avm



Overall, this plots show that when observing the *adj* subset, two different strategies could be found. One of them is to be more dialect in the first part, than in the second (I20130623b1), and the other represents the inverse strategie (20130701d-avm). For the rest of the files, only the observations on the subtraction product will be given in order to summarize the data[4]:

Table 3: results of the subtraction of dialectness values

variable	audio	gender of the speaker	year of birth	observed difference
adj	20140624e-lso	female	1941	-0.3111111
	20130702b-nvt	female	1952	-0.3589744
	20130701d-avm	female	1922	0.4880952
	20140626c-npo-1	male	1954	0.5
	20150715a-nnt-1	female	1960	0.1764706
	I20130623b1	female	1952	-0.625
	20160624c-nfm	female	1961	0.2857143
prep_pron	20130701d-avm	female	1922	-0.375
ot	20150712a-ops-1	female	1935	-0.375
ei	20130704c-lpp	female	1928	0.625
	20130701c-epr	female	1926	-0.5
sh	20130702b-nvt	male	1952	-0.125
	20150716b-sek-1	male	1930	-0.3194444

[4] All the experiment results are presented in the GitHub repository of the project

	20160624c-nfm	female	1961	-0.4509804
	20160630g-ans-1	female	1952	-0.25
sja1	20140704-lso-2	female	1941	-0.4621212

The results are indicating that the strategies chosen by speakers are varying from variable to variable. When using variables such as sh (the most frequent variable), speakers are usually less dialect in the first 10 minutes. But in a variable such as *adj* there is a great variability across speakers which seems not to be explainable by gender differences or by year of birth differences (as there is some speakers born in the same decade which have different strategies: female speakers born in 1952 and 1954, one is more dialect in the beginning, and the other is more dialect towards the end). However, the findings made for *sh* variable may interpreted in favor of the self awareness hypothesis. They may indicate that the speakers are trying to be less dialect to the beginning, as they are trying to minimize the variation between them and the interviewer (speaker of the standard Russian) and then at some point they are losing self-awareness and start to speak more casually. This results in higher dialectness in the second part.

4.2. Generalized linear model with mixed effects: preliminary analysis

In Section 6.1, individual interviews were analyzed. The results are indicating that in some of them there may be a trend of positive change in dialectness over time (although a few interviews show a negative trend). I will now try to analyze the data in such a way as to include sociolinguistic variables such as gender, year of birth etc. into consideration. Generalized Linear Model with random effects will be used[5]. All numerical data will be scaled. As it was shown in (Daniel et al. in press), there are some strong predictors of dialectness such as year of birth and variable. In order to analyse the predicting power of the timestamp, the following test will be applied. First, two GLM's with mixed effects will be compared, with and without timestamp as a predictor. The

[5] The R-package lme4 has been used in order to perform all the following computations

AIC (Akaike information criterion) metric will be used in order to compare these two models. This metric deals with the probability of over- or under- fitting the model, so it will help to choose the optimal as well as the most informative model with a greater explanatory power than the others. The models are as follows:

1. *realization class ~ timestamp + age_group + age + sex + var + (1 | speaker) + (1 | audio)*
2. *realization class ~ age_group + age + sex + var + (1 | speaker) + (1 | audio)*

The model which contains the timestamp as one of the predictor has the AIC value of 20223.23 and the one that has not 20222.72. This indicates that the presence of the timestamp as an independent predictor has not enhanced the quality of the model. I will now try to analyze interaction of the timestamp with other variables such as gender and age-group.

1. *realization class ~ timestamp:sex + age_group + age + sex + var + (1 | speaker) + (1 | audio)*
2. *realization class ~ timestamp:var + age_group + age + sex + var + (1 | speaker) + (1 | audio)*
3. *realization class ~ timestamp:var:sex + age_group + age + sex + var + (1 | speaker) + (1 | audio)*
4. *realization class ~ timestamp:age_group+ age_group + age + sex + var + (1 | speaker) + (1 | audio)*
5. *realization class ~ timestamp:age_group:sex + age_group + age + sex + var + (1 | speaker) + (1 | audio)*

The AIC scores for each model are presented in the Table below:

Table. 4: AIC scores of given models

model	1	2	3	4	5
AIC	20220.75	20212.94	20201.22	20223.28	20225.93

The results are indicating that the models 1, 2 and 3 have AIC scores which are higher than the score of the model without the timestamp as predictor (20222.72). This means that these model are more informative. The results are also indicating that the models with the tripartite interactions with gender (*timestamp:age_group:sex* and *timestamp:var:sex*) have higher scores than the models with binary interactions (*timestamp:age_group* and *timestamp:var*) so they will be used in the future examinations.

4.3. Generalized linear model with mixed effects analysis

In this section, I will test the assumptions made with the small-scale analysis. I will demonstrate, how the timestamp intercacts with the realization class on the gender, age-group and variable levels.The general procedure will be as follows; firstly, many models are presented for each relation which is modelled (for example, gender, position and timestamp), based on the sociolinguistic expectations. Those models will have different formulas, in order to determine which of them is more accurate when fitting the data. Then, they are compared with each other using the AIC (Akaike information criterion) metric. This metrics deals with the probability of over- or under- fitting the model, so it will help in choosing the optimal as well as the most informative model with a greater explanatory power than the others. Then, the significance of the timestamp when predicting the realization class is retrieved and analysed. The slopes produced by the model will help to test my previous assumptions based on the small-scale analysis.

As one of the possible model, I will try to predict the realization class using the timestamp without any interactions, adding different random factors which will be in various ways nested in each other. The formulas for the first test will be as follows:

1. realization class ~ timestamp
2. realization class ~ timestamp + (1 | audio/sex/speaker) + (1 | token/var) + (1 | year of birth)
3. realization class ~ timestamp + (1 | audio/sex/speaker) + (1 | token/var) + (1 | age group)
4. realization class ~ timestamp + (1 | sex/speaker) + (1 | token/var) + (1 | year of birth) + (1 | audio)
5. realization class ~ timestamp + (1 | sex/speaker) + (1 | token/var) + (1 | age group) + (1 | audio)
6. realization class ~ timestamp + (1 | audio/sex/speaker/year of birth) + (1 | token/var)
7. realization class ~ timestamp + (1 | audio/sex/speaker/age group) + (1 | token/var)
8. realization class ~ timestamp + (1 | audio) + (1 | sex) + (1 | speaker) + (1 | age-group) + (1 | token) + (1 | var)

Firstly, each model containing year of birth as a random effect will be tested against the model which is containing the age group instead. This is made in order to see, whether the choice of a simpler categorical variable (age group contains only 3 levels) will increase the simplicity and the power of the model or not. As well I will test, whether year of birth/age group can be nested with audio, sex etc. The models 2 and 3 will test the assumption, whether audio can be nested in the sex and speaker or could be used as a separate effect. The model 8 will test, whether the nesting of variables is important. The resulting AIC scores are presented below:

Table 5: model comparison results (all the data)

model number	1	2	3	4	5	6	7	8
AIC	34042.22	22007.94	22010.89	21876.46	21908.31	22145.98	22145.98	19255.41

Thus, the model 8, with has the lowest AIC, will be used in this analysis. The results are the follow:

Table 6: model with lowest AIC results (all the data)

	Estimate	Std. Error	z value	p-value
Intercept	-0.65170	1.43163	-0.455	0.649
timestamp	0.03500	0.02547	1.374	0.169

This model indicates that the timestamp is not significant, as the p-value of it being significant is bigger than 0.05. This is correlated with the observation made in the previous Section, where the deletion of the timestamp from the model did not change its performance.

Then I will address the interactions with gender, as if one of the genders will have significance together with the timestamp, it can be used for later examinations[6]. The best model (AIC=22079.29) formula is as follows:

[6] For now on, all the formulas tested are presented in the appendix

realization class ~ timestamp:sex + (1 | speaker) + (1 | token) + (1 | age_group) + (1 | audio)

The results of this model are presented in the table below:

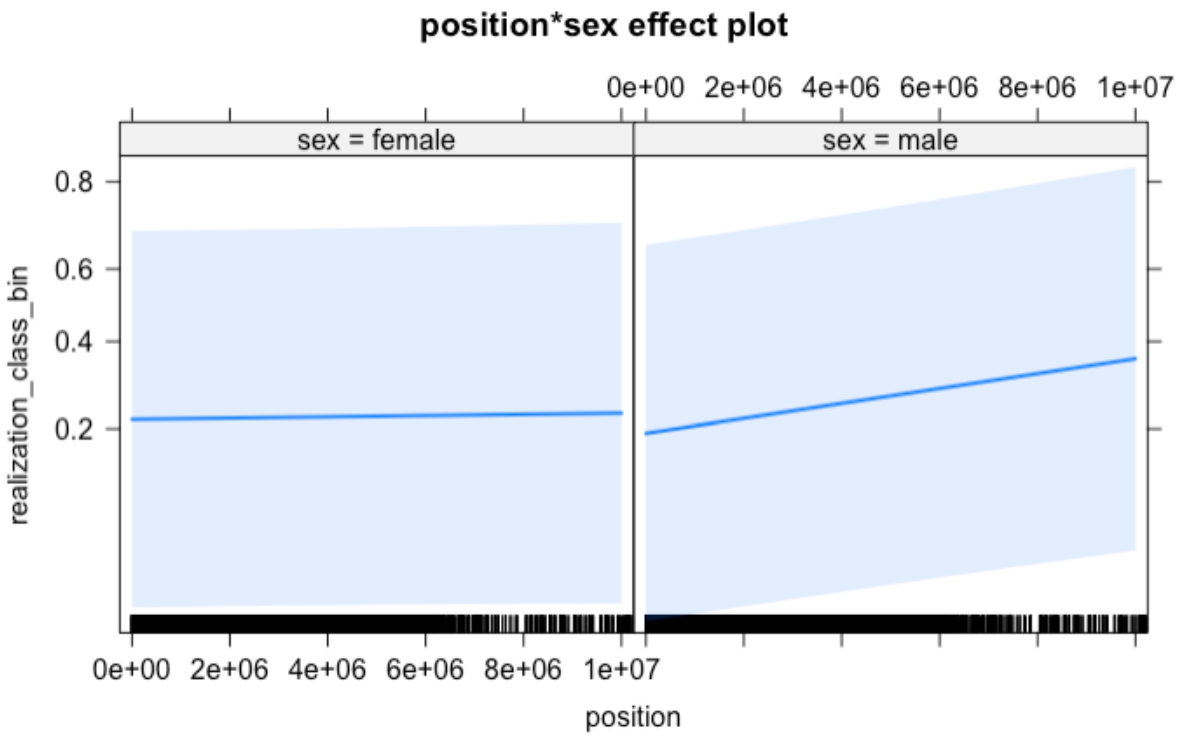
Table 7: model with lowest AIC results (timestamp and gender interaction)

	Estimate	Std. Error	z value	p-value
Intercept	-1.26792	1.07703	-1.177	0.23910
timestamp:female	0.01128	0.03280	0.344	0.73090
timestamp:male	0.13620	0.04243	3.210	0.00133*

As the table suggests, this model indicates that the timestamp is more significant among male speakers, then among the female ones (p-value<0.05). The slope of change is also positive which is indicating that the male speakers tend to be more dialect towards the end. The slight positive can be also spotted for the female speakers, but the interaction of female gender and timestamp has shown to be not significant in the observed model. Those observations are supported with the plots of the predicted probabilities[7]

[7] R-package effects was used in order to construct this plot.

Fig. 11: predicted probabilities



The next model will address the differences on the variable levels. For that, we will use the interaction between the variable 'variable' and the timestamp, as well as with gender, when predicting the realization class.The best model (AIC=22016.37) formula looks as follows:

realization class ~ timestamp:var:sex + (1 | speaker) + (1 | token) + (1 | age_group) + (1 | audio)

For now, I will be including only the significant ($p<0.05$) variable interactions, as there is 24 possibilities total (12 * 2 = 24). The results table is presented below:

Table 8: model with lowest AIC results (timestamp, variable and sex interaction, only $p<0.05$)

variable	gender	Estimate	Std. Error	z value	p-value
sh	male	0.50718	0.10724	4.729	2.25e-06

	female	0.18280	0.07392	2.473	0.013396
ae	female	-0.20036	0.09499	-2.109	0.034924
ei	female	-0.35125	0.10063	-3.491	0.000482
sja1	female	-0.37069	0.10517	-3.525	0.000424
sja2	female	0.29101	0.11401	2.553	0.010693
one	male	0.44462	0.12990	3.423	0.000620
ot	male	0.33241	0.14640	2.270	0.023177

Overall, 7 variables (sh, ae, ei, sja1, sja2, one and ot) have shown the significance of the timestamp when predicting the realization class. Those results are indicating that the slope of change in dialectness of *sh* during the interview is bigger for male speakers (0.50718) than for female speakers (0.18280). The model is also predicting negative slopes of change for variables *ae, ei* and sja1. The negative change in dialectness of *sja1* was also spotted on the small scale level (see Section 6.1).

Another models which were used in this survey tend to identify any effects on the age-group level. In order to do so, the speakers will be clustered into 3 distinct categories:

The best model (AIC 19316.74) has the following formula:

realization class ~ timestamp:age group:sex + (1 | speaker) + (1 | token) + (1 | var) + (1 | audio)

The overall results are indicating that only in the timestamp is significant for the male speakers within the people who were born before 1940, the results are presented in the table below:

Table 9: model with lowest AIC results (age-group and timestamp interaction, only $p<0.05$)

	Estimate	Std. Error	z value	p-value
age_group< :male	0.12874	0.05675	2.269	0.0233*

Overall, these observations have shown that there is some differences between the male and the female behaviour, as well as it has indicated that the overall slope of change seems to be positive. When analyzing the interactions between timestamp, variable and gender, the variable *sh* has been shown to have positive slope (i.e. going from less dialect to more dialect) for both men and women as it was shown with the probabilistic analysis.

4.4. Variable importance analysis[8]

Previously I tried only to identify whether the timestamp does predict the realization class. In this section, however, new approach will be introduced. In order to understand the overall importance of the variable, or the combination of different variables (such as sex, year of birth etc.) I will run the following experiment; using the Logistic

[8] The scikit-learn Python package was used (Pedregosa et. al 2011)

Regression modelling with cross validation, I will test each variable, used in this survey, on its predictive power, when predicting the realization class of a variable. Then I will collect the scores from all the models and compare them. The metric of choice will be the roc-auc score, as it presents a threshold which indicates that the predictions made with such models, can be achieved with a random-guessing classifier (score < 0.5). The choice of the cross-validation technique is motivated to the fact that it can help produce the best model possible for each variable which will make the comparison more adequate. The accuracy score will be recorded as well in order to provide more adequate comparison. The resulting roc-auc scores plot is represented below:

Fig. 12: roc-auc score plot for each variable

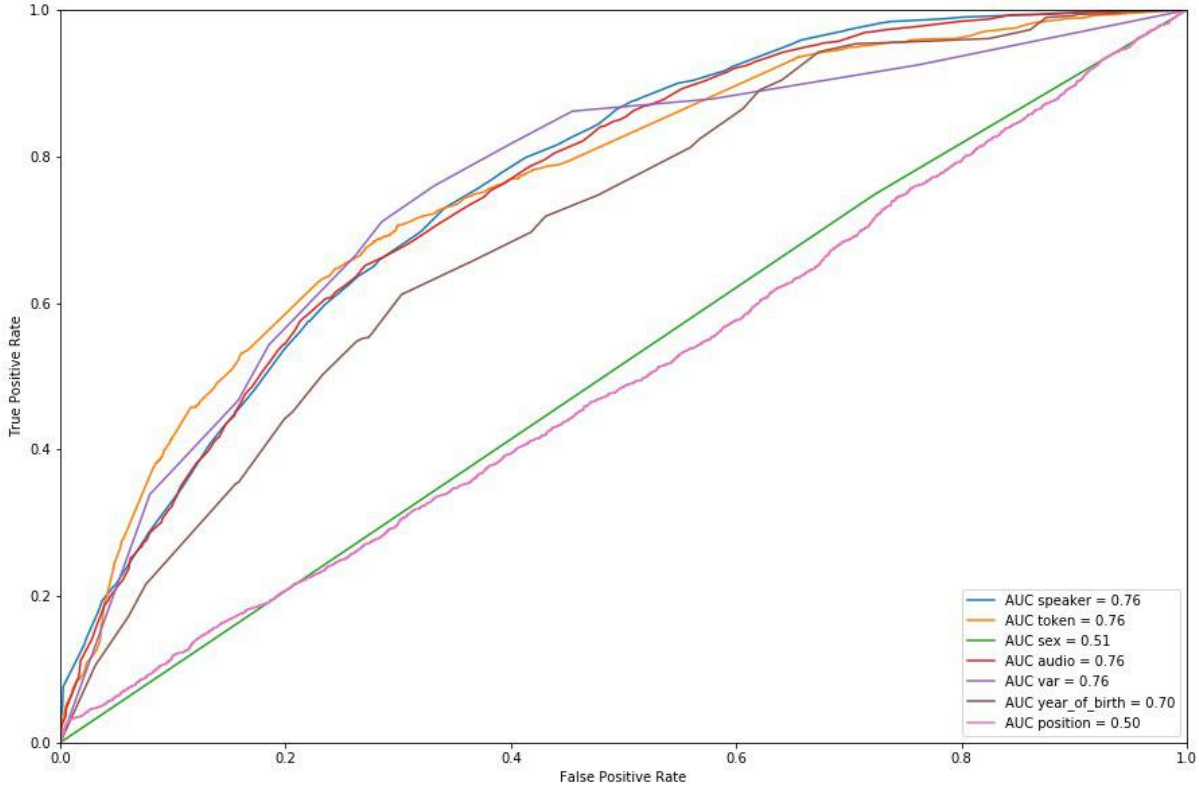


Fig. … indicates that some variables have nearly indistinguishable roc-curves which may indicate that they have nearly the same predicting power (variables such as speaker, token, variable and audio). The interesting part being that the variables token and audio, as well as speaker and audio are forming closely related pairs (see Section 6.1.), so there is no surprise that those pairs are both presented among the significant variables. The overall results are presented in a table below:

Table. 10: metric values for each variable

feature	roc-auc score	accuracy
speaker	0.7649513879567051	0.693313594867999
token	0.7668210514909306	0.7009622501850481
gender	0.5127723612203098	0.5138169257340242
audio	0.7592689184569918	0.6860350357759685
variable	0.757570685390088	0.7142857142857143
year of birth	0.7038243726518485	0.654206760424377
timestamp	0.4974395313422195	0.4988897113249445

The overall hierarchy of variable importance can be represented as follows:

token > speaker > audio > variable > year of birth > gender > timestamp

Overall, once again, this indicates that timestamp is not important by itself on the global level, however that the realization class is much better predicted by other factors such as token, speaker and audio etc.

5. Results

Different approaches to analysis of the dataset described in Section 4 gave birth to many outcomes, some of which need further analysis and theoretical grounding. One of the main observations shows that there is a slow positive change in dialectness over the time of the interview. This observation was first suggested by the small-scale testing and then confirmed by Generalized Linear Models examination. However, this positive

change in dialectness has been found significant only for male speakers and some of the variables.

Small-scale analysis used to test the assumption of the first 10 minutes of the recording being different from the rest of the recording (see Sankoff & Blondeau 2007: 9) has shown interesting results. Firstly, it indicates that in some files the 10 minutes part is indeed different from the rest. Secondly, it has shown that the speakers have not any consistent strategies on the variable level. In the first 10 minutes the observed speaker can either be more dialectal, than in the rest of the audiofile, or less dialectal (the second variant is prevailing). The variable *sh* being an exception, as in all the observed cases speakers were less dialect in the beginning (0-10 minutes) and more dialect in the last part of the interview. This needs further examination, as this fact seems not to be related neither to the interviewers or to the year of birth and sex of the speaker.

When the data was examined using the GLM's with mixed effects, the timestamp has found to be more significant for predicting the realization class on the data obtained from male speakers. The slope of change is positive, which corresponds to the observations made in small-scale analysis. The variable *sh*, which demonstrated increasing dialectness in the small-scale analysis was also found to have a positive slope of change. The age group factor, however, has shown no or little effect on the model.

6. Conclusion

The aim of this study was to test two proposed hypothesis: first of them used the ideas from the audience design framework and suggested that the dialectness will degrade over time of the interview, as the speakers will adapt to the interviewers style (a). The second one suggested the positive change in dialectness, which is explained by the loss of self-awareness and hence the increase of the dialectness over time (b).

The analysis proposed in this study has shown that the hypothesis (b) can explain the effects observed in this survey. Firstly, the small-scale variable analysis indicates that

there are some audio-files that are showing significant difference between the dialectness in the first 10 minutes and the other part of the file. In the case of the variable *sh* (one of the most frequent variables in the dataset) the trend is positive. In other words., speakers start their recorded speech using the standard realization of this variable more, but then tend to be more dialectal. This observation was also supported with the generalized linear model analysis, where the probability of *sh* being dialectal changes positively over time. The analysis has also shown that the positive trend is more clear for male speakers (when the interaction between timestamp and gender was used). This result seems very promising, as there are less observations for mlale speakers.

Althrought, when the plain timestamp was used as one of the predictors, the effect was not found significant. Overall, this study shows that the possible timestamps-effects are appearing on the small-scale level, even if the effect is not so strong when the data is observed as a whole. This suggests that the timestamp of a variable as well as other factors such as role of the interviewer and the length of the interview should be taken into consideration when studying speaker variation.

7. References

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145-204.

Cheshire, J. (1982). Dialect Features and Linguistic Conflict in Schools. *Educational Review*, *34*(1), 53–67. http://doi.org/10.1080/0013191820340106

Daniel, M.; Kazakova, P.; Ter-Avanesova, A.; von Waldenfels, R.; Gerasimenko., E.; Ignatenko, D.; Makhlina, E.; Ovsjannikova, M.; Say, S.; Schurov, I.; Tsfasman, M.; Verhees, M.; Vinyar, A.; Zhigulskaya, V.; Dobrushina, N. (to appear). Dialect loss in the Russian North: modelling change across variables.

Giles, H., & Ogay, T. (2007). Communication accommodation theory. *Explaining communication: Contemporary theories and exemplars*, 293-310.

Roberts, S., & Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PloS one*, *8*(8), e70902.

Macleod, N.; Grant, T. (2016). "You have ruined this entire experiment…shall we stop talking now?" Orientations to the experimental setting as an interactional resource. *Discourse, Context & Media*, *14*, 63–70. http://doi.org/10.1016/j.dcm.2016.10.001

Wertheim, S. (2002). Rethinking the Observer's Paradox and Data "Purity". *Annual Meeting of the Berkeley Linguistics Society*, *28*(1), 511. http://doi.org/10.3765/bls.v28i1.3862

Wilson, J. (1994). Paradoxes, sociolinguistics and everyday accounts. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, *13*(3), 285–300. http://doi.org/10.1515/mult.1994.13.3.285

von Waldenfels, R., Daniel, M., & Dobrushina, N. (2014). Why standard orthography? Building the Ustya River Basin Corpus, an online corpus of a Russian dialect. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue* (Vol. 13, pp. 720-728).

Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 560-588.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

8. Appendix

The appendix for this project is avalaible online:
https://alexeykosh.github.io/diploma/docs/

The Git-Hub repository of the project:
https://github.com/alexeykosh/diploma