Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2022”

Moscow, June 15–18, 2022

Setting Up A Complex Model Of Speech Analysis: Pilot Study Of Late Bilingual Speech

Alina Lozovskaya
Independent Researcher
Saint-Petersburg, Russia
alinailozovskaya@gmail.com

Danil Pitolin
Ural Federal University
Ekaterinburg, Russia
d.v.pitolin@urfu.ru

Semyon Bessonov
Independent Researcher
Saint-Petersburg, Russia
semeon-bessonov@ya.ru

Abstract

Modern applied linguistics includes a substantial amount of spoken language analysis, and current Natural Language Processing techniques allow working with large amounts of audio data. This article demonstrates a method for conducting primary research of late bilinguals' speech, which is particularly relevant in the context of modern globalization. Using a small sample as an example, the paper presents the methodology testing the informants' speech, including technical approaches for speech collecting, processing, and interpretation. The dataset for the analysis is the interview recordings, which took place after the informants watched a silent film.

Keywords: sociolinguistics, natural language processing, speech analysis, speech-to-text, transformers, hugging face, late bilinguals, language interference

DOI: 10.28995/2075-7182-2022-21-1122-1128

Формирование комплексной модели анализа речи: пилотное исследование речи поздних билингвов

Лозовская А. И.
Независимый исследователь
Санкт-Петербург, Россия
alinailozovskaya@gmail.com

Питолин Д. В.
Уральский федеральный университет
Екатеринбург, Россия
d.v.pitolin@urfu.ru

Бессонов С. А.
Независимый исследователь
Санкт-Петербург, Россия
semeon-bessonov@ya.ru

Аннотация

Анализ устной речи составляет большую часть современной прикладной лингвистики, а актуальные походы Natural Language Processing позволяют анализировать большие объемы аудио информации. Данная статья показывает подход к первичному анализу звучащей речи информантов – поздних билингвов, что наиболее актуально в ситуации современной глобализации. В статье представлено тестирование методологии исследования речи информантов, в том числе технических подходов для сбора, обработки и анализа речи, на примере небольшой выборки. Материалом для анализа служит запись интервью по итогам просмотра информантами немого фильма.

Ключевые слова: социолингвистика, natural language processing, анализ речи, speech-to-text, transformers, hugging face, поздние билингвы, языковая интерференция

1. Introduction

With ongoing globalization and mobility people all over the world migrate like never before. Moreover, this trend is projected to become even more prominent [2]. Consequently, the first generation of migrants usually has to learn a language of the destination country. However, these people do not fit the quite tight category of natural or balanced bilinguals and for decades stayed unnoticed by the scholars studying bilingualism [7]. More recent studies though have examined the subject matter in more detail [3, 5, 6, 9]. The research at large aims to continue closing the gap and study late Russian English bilinguals once it goes full-scale. It is to be done through a series of online interviews with adult informants who moved to an English-speaking country in the post-puberty period (for the purpose of the research it is roughly estimated as 18 years). In a later large-scale research such factors as age, age of acquisition, age of relocation, occupation and others are to be analyzed. However, the aim of the research described was merely to test if the chosen speech analysis model was suitable for the data obtained.

Among other activities they have watched an excerpt from a silent film to elicit the production of monologue speech. Such films have been noticed to be well described as being suitable for speech elicitation as they do not influence the informant with sound of speech of any language [8]. For instance, the piece of Charlie Chaplin's *Modern Times* (1936) served as elicitation material to dozens of experiments (for more in-depth description of the piece see [1]). The authors of the article were to find a piece of video resembling Modern Times' easy and humorous narrative depicting daily life. The first part of *Wake Up Lenochka [Разбудите Леночку]* (1934) was chosen as it has approximately the same duration, is silent, and was produced in Russian language.

After watching the part of the film the informants were to retell the story and describe their general impression of it in Russian. With late bilinguals it was hypothesized that the interference of English (L2) would manifest itself in informants' native Russian.

The article presents the results of a pilot study that was aimed at testing and adopting the practices most suitable for a future large scale research. The amount of the monologous speech elicited from the informants that is described does not let the authors draw any linguistic conclusions yet appears to be sufficient to test the future model of transcribing larger data amounts.

2. Task Description

At this stage the task is to test the methodology on a small volume to see what problems will arise:

- at the data collection stage;
- at the data processing stage;
- at the data interpretation stage.

Therefore, the authors of this article decided to conduct an EDA on the audio material of eight informants, to see what questions they were asked after watching the film and how they responded to them.

One of the most essential tasks, apart from decoding audio into text, was to find a model that can decode the audio chunks into text. For initial EDA, we chose Yandex SpeechKit, which produces adequate results for primary analytics.

3. Exploratory Data Analysis

As mentioned above, Russian-language interviews were chosen for the initial analysis, focusing on the questions asked in the informants' native language. The purpose of this stage is to check the efficiency of the questions, to what extent they urge the informants to talk about the film fragment.

The first to be done for the datastudy is to mark up the questions as timestamps, where the first column is the person's ID, the second column is the question the person answers, the third column is when the person started answering the question, and the fourth column is when the person finished answering the question. The first three lines will look like this:

ID	question	start	end
1	What was the film about	01:13:32	01:14:04
1	What caught your attention	01:14:11	01:14:22
1	Was it difficult to watch the film	01:14:30	01:14:38

Table 1: Initial data

Next step is calculating the duration of each question. All the columns are converted to timedelta64 except for one numeric column.

ID	question	start	end	duration	duration_sec
1	What was the film about	0 days 01:13:32	0 days 01:14:04	0 days 00:00:32	32
1	What caught your attention	0 days 01:14:11	0 days 01:14:22	0 days 00:00:11	11
1	Was it difficult to watch the film	0 days 01:14:30	0 days 01:14:38	0 days 00:00:08	8

Table 2: Updated data

The next step is to find out how many questions on the film each informant received. The maximum number was six questions and the minimum was three questions. See fig.1 with the questions frequency.

Questions with frequencies

	question	freq
0	What was the film about	8
1	What are your impressions of the film	5
2	Was it difficult to watch the film	4
3	What is going to happen next in the film	3
4	Do you watch old movies	2
5	Do you want to see the film up to the end	2
6	In what years would the situation take place	2
7	What caught your attention	1
8	Why was it hard to watch the film	1
9	What's the main idea of the film	1
10	Do you remember the details of the characters' appearance	1

Figure 1: Questions with frequencies

The distribution of questions was uneven, see fig. 1. Initially, there was no strict list of questions to be asked during the interview, so the discussion of the film was done as a free dialogue, in order to further assess which questions might be the most relevant to the discussion of the film fragment.

Next, it is useful to examine the average, maximum, and minimum length of each question, where the level of answer detail is directly proportional to the length of that answer (fig. 2).

Duration statistics

	question	seconds_mean	seconds_max	seconds_min
0	What was the film about	102	216	6
1	Why was it hard to watch the film	74	74	74
2	What's the main idea of the film	43	43	43
3	What time is in the film	32	37	26
4	What happens next	22	30	10
5	What are your impressions of the film	19	36	6
6	Do you want to see the film up to the end	15	18	12
7	Do you watch old movies	15	26	4
8	What caught your attention	11	11	11
9	Was it difficult to watch the film	11	18	7
10	Do you remember the details of the characters' appearance	9	9	9

Figure 2: Duration statistics

According to statistics, the following three questions should be left for the next interview, both in Russian and English:

- What was the film about?
- What time is in the film?
- What happens next?

The first question is the most essential one. With this question, one can assess how the informants perceived the film and, by doing a semantic analysis, find out what caught their attention without asking them an additional question about the details. The question about the time in which the film takes place provides more detailed speech elicitation beyond mundane topics. The last question about what might happen next on one hand leads them into a particular and fairly uniform way of storytelling when most of them describe the main character being late for school. On the other hand, the elicited speech still holds most individual peculiarities of each informant. All the questions above have proved to be fruitful in provoking long responses, yet they do not lead the informant to any particular pattern of language choice that would distort the natural flow of speech.

3.1. Speech-to-Text

We are interacting more and more with technology through voice, yet, as Daniel Jurafsky and James Martin note, automatic transcription of speech by any speaker in any environment is still far from solved, but ASR technology has matured to the point where it is now viable for many practical tasks [4].

Therefore, after collecting initial record data, it needs to be converted into text format for extensive exploratory data analysis. This task requires some easy to use but yet accurate tools. At this stage of the project Yandex Speech Kit proved to be suitable for the task. It provides robust and accurate speech to text service for Russian and English languages.

Sound processing pipeline at this stage roughly is:

- Convert collected audio into Ogg Opus format via ffmpeg;
- Split audio files into small chunks (one per answer);
- Use speech recognition service on each audio chunk;
- Save each result separately;
- Manually validate all collected results to correct spelling, grammar, and punctuations;
- Aggregate results into initial dataset for EDA.

We hope that in future by using more advanced state-of-the-art models, and with better audio collecting procedures we will eliminate most of the manual labor, thus giving us the ability to work with much bigger amounts of data. However, current implementation is enough for bootstrapping.

3.2. Keywords extraction

The next step is the analysis of the text itself and highlighting the key bigrams for each of the questions asked. In order to highlight bigrams, it was decided to use the open-source KeyBERT[1] project, since it gives fine results and supports many embedding models. It is useful as we can load the Russian language model from the Hugging Face community for keyword extraction. The authors use Flair together with rugpt3medium_sum_gazeta[2]. This model was created for summarization and is based on rugpt3medium_based_on_gpt2[3], which in our opinion can be suitable for keyword extraction.

The major stage of text preprocessing is stop words cleaning, which are the most frequent in the text and usually have less lexical content, besides such words do not hold much of a semantic meaning. Consequently, it was important that these words are not included in the key phrases. It is worth noting that when working with a large amount of text data, BERT does not need to pre-clean the texts from stop words, because such models are perfect to perceive the context, so in this case the cleaning is dictated by the small text volume. The library spaCy[4] which contains a sufficient set of stop words for the Russian language was chosen for this purpose. The raw material being spoken language lead to the expansion of the word stops,, but the negative "no" in the two versions "не" and "нет", as well as "everyone" (все) and "everything" (всё) have been removed from the original list.

When extracting keywords, the diversity setting was 0.1, since the texts are non-voluminous and there will be little variation in meaning. To select the most relevant keywords the most relevant 10 key words of each response were selected.

Obviously, the extraction of key bigrams from a small amount of text causes errors and may produce irrelevant keywords, so it was decided to analyze the resulting bigrams and choose the most appropriate. Besides, prepositions and negation, if implied, have been restored in extracted bigrams.

The keywords presented in the table are ordered by their probability prediction (not specified in the table) from KeyBERT.

question	**keywords**
What was the film about	mom has come [мама пришла]; ran to school [побежала в школу]; got ready and ran out [собралась и выбежала]; realized she was late [поняла, что опоздала]; all gone [все ушли]; ran out of the house [выбежала из дома];
Why was it hard to watch the film	no dialogues [диалогов нет]; they write quickly [пишут быстро]; no language [языка нет]; not interesting [не интересно]; old movie [старый фильм]; picture visualization [визуализация картинки]

[1] https://github.com/MaartenGr/KeyBERT

[2] https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta

[3] https://huggingface.co/sberbank-ai/rugpt3medium_based_on_gpt2

[4] https://spacy.io

What's the main idea of the film	before bedtime [перед сном]; understandable phenomenon [понятное явление]; strong immersion [сильное погружение]; sleep problems [проблемы со сном]
What time is in the film	no idea [не разбираюсь]; silent movies [немое кино]; thirties to forties [тридцатые – сороковые]; twenties to forties [двадцатые – сороковые];
What happens next	ran late [побежала с опозданием]; will run to school [побежит в школу]; i think she's late [думаю опоздала]; she is going to wash up [соберется умываться]; she will wait for her brother [дождется брата]
What are your impressions of the film	enjoyed it [все понравилось]; Lenochka is infantile [леночка инфантильная]; strange family [семья странная]; loved the authenticity [понравилась аутентичность]; not interesting [не интересно]
Do you want to see the film up to the end	Lenochka is worried [леночка переживает]; what she will do to the guy [сделает парнем]; what she will do next [дальше сделает]; I watch everything to the end [досматриваю все]
Do you watch old movies	don't watch [не смотрю]; used to watch [раньше смотрел]; occasionally [бывает редко]; Charlie Chaplin [чарли чаплин]; new movies [новые фильмы]
What caught your attention	note to her brother [записка брату]; ask for a wake-up [просьба разбудить]
Was it difficult to watch the film	everything is clear [все понятно]; no fine [норм нет]; no words [без слов]; simple story [простой сюжет]
Do you remember the details of the characters' appearance	they looked strange [выглядели странно]; twenties [двадцатые годы]

Table 3: Keywords extraction

Keywords contributed to the text content analysis When answering the question "What was the film about", the informants paid attention to the fact that the main character overslept, woke up when everyone had already left, and rushed to get ready. Understanding differences in informants' perception of time can be useful for the research: the most predictable one here is "no idea", but the informants also thought that the action takes place either in the thirties or the forties, and it is worth noting that they were right. In answering the question "What happens next" almost all informants suggested that Lenochka would run late to school.

4. Conclusion

This article presents an approach to analyzing the speech of late bilinguals using the example of an interview after watching a silent film. A timestamps partitioning was performed and the approach to STT was tested, splitting audio files into small chunks. The next step was to select a SOTA approach for extracting keywords from the resulting text.

For further work we note that informants are interviewed in two languages: in their native Russian language and in English. In the nearest perspective, we would like to analyze how informants described the silent film in English. The next major stage of the research is to track phonetics changes of late bilinguals, to understand whether this change is typical for all late bilinguals who have English as a second language, or whether phonetic changes in speech are individual in nature. Moreover, gradually over the year, it will be possible to come up with more data, which will allow building our own model for ASR and phoneme analysis, taking into account the specificity of speech.

References

1. Bergmann C. (2015). Collecting and Analyzing Spontaneous Speech Data. SpringerBriefs in Linguistics, pp. 37–53. Access mode: doi:10.1007/978-3-319-11529-0_4
2. Dao T. H., Docquier F., Maurel M., & Schaus P. (2021). Global migration in the twentieth and twen-ty-first centuries: the unstoppable force of demography. Review of World Economics, 157(2), pp. 417–449. Access mode: doi:10.1007/s10290-020-00402-1
3. Isurin L. (2021). Does language transfer explain it all? The case of first language change in Russian-English bilinguals. Russian Journal of Linguistics (Online), 25(4), pp. 908–930.
4. Jurafsky D., Martin J. H. Speech and Language Processing. — 2021. — Access mode: https://web.stanford.edu/~jurafsky/slp3/
5. Novitskiy N., Myachykov A., & Shtyrov Y. (2018). Crosslinguistic interplay between semantics and phonology in late bilinguals: neurophysiological evidence. Bilingualism: Language and Cognition, 1–19. Access mode: doi:10.1017/s1366728918000627
6. Novitskiy N., Shtyrov, Y., & Myachykov A. (2019). Conflict Resolution Ability in Late Bilinguals Improves with Increased Second-Language Proficiency: ANT Evidence. Frontiers in Psychology, 10. Access mode: doi:10.3389/fpsyg.2019.02825
7. Pavlenko A. (2000), L2 Influence on L1 in Late Bilingualism. Issues of Applied Linguistics, Vol. 11, pp. 17–205.
8. Perdue C. (ed.) (1993). Adult Language Acquisition. Vol 1: Field Methods. Cambridge University Press.
9. Shishkin E., & Ecke P. (2018). Language Dominance, Verbal Fluency, and Language Control in two Groups of Russian–English Bilinguals. Languages, 3(3), 27. Access mode: doi:10.3390/languages3030027