Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2019)

Выпуск 18

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2019)

Issue 18

УДК 80/81; 004
ББК 81.1
К63

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 29 мая — 1 июня 2019 г.). Вып. 18 (25), 2019.

Сборник включает 61 доклад международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2019», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

18-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 25-й международной конференции «Диалог». На основании мнений нашего рецензентского корпуса для публикации в ежегоднике редколлегией был отобран 61 доклад из ста работ, которые были приняты к представлению на конференции в 2019 году.

Работы в сборнике отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на конференции:

- Компьютерные лингвистические ресурсы
- Компьютерный анализ документов (классификация, перевод, поиск, саммаризация, генерация, анализ тональности и т. д.)
- Глубокое обучение в NLP (методики применения, содержательная интерпретация)
- Компьютерный анализ Social Media
- Корпусная лингвистика и корпусометрия (методики создания, использования и оценки корпусов)
- Лингвистический анализ текста (морфология, синтаксис, семантика)
- Лингвистические онтологии и автоматическое извлечение знаний
- Мультимодальная коммуникация (включая лингвистический анализ речи)
- Модели общения и диалоговые агенты
- Компьютерная лексикография

В соответствии с традициями «Диалога», старейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка.

Одно из ключевых событий «Диалога» — подведение итогов технологических соревнований между разработчиками систем лингвистического анализа текстов, Dialogue Evaluation. В этом году состоялись четыре соревнования:

- автоматическая генерация заголовков новостей;
- автоматический анализ малоресурсных языков (для которых очень мало данных для машинного обучения);
- автоматическое разрешение анафоры и определение референциальных цепочек (различных упоминаний одного и того же объекта в тексте),
- автоматическое восстановление слов по контексту (гэппинг-эллипсис).

В сборник включены наиболее оригинальные работы участников этих соревнований.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редсовет отказался от печати сборника на бумаге, поскольку бумажный вариант пользуется все меньшей популярностью. Сборник, как и в прошлые годы, размещается на сайте конференции и индексируется Scopus.

Программный комитет конференции «Диалог»

Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Филологический факультет МГУ
- Школа прикладной математики и информатики МФТИ

Международный программный комитет

Богуславский Игорь Михайлович	ИППИ РАН, Россия; Мадридский политехнический университет, Испания
Буате Кристиан	Университет Джозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мексика
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт системного анализа РАН, Россия
Райгородский Андрей Михайлович	МФТИ, Школа прикладной математики и информатики, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания ABBYY, МФТИ, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Захаров Леонид Михайлович	МГУ им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Пиперски Александр Чедович	РГГУ
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Федорова Ольга Викторовна	МГУ им. М. В. Ломоносова
Шаров Сергей Александрович	Университет Лидса

Секретариат

Родионова Ольга Игоревна, *координатор оргкомитета*	Компания ABBYY
Ульянова Анна Вячеславовна, *секретарь оргкомитета*	РГГУ

Рецензенты

Tania Avgustinova
Vladimir Benko
Anatoly Gersman
Diana Macartney
Preslav Nakov
Piek Vossen
Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Артемова (Черняк) Екатерина Леонидовна
Архангельский Тимофей Александрович
Байтин Алексей Владимирович
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Галинская Ирина Евгеньевна
Галицкий Борис Александрович
Гельбух Александр Феликсович
Гращенков Павел Валерьевич
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Добрынин Владимир Юрьевич
Дроганова Кира Андреевна
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Катинская Анисья Юрьевна
Кибрик Андрей Александрович
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кронгауз Максим Анисимович
Кутузов Андрей
Левонтина Ирина Борисовна
Леонтьев Алексей Петрович
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Марков Александр Юрьевич
Мисюрев Алексей Владимирович
Недолужко Анна Юрьевна
Новицкий Валерий Игоревич
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Пивоварова Лидия
Пиперски Александр Чедович
Подлесская Вера Исааковна
Смирнов Иван Валентинович
Смуров Иван Михайлович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Сорокин Алексей Андреевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Усталов Дмитрий Алексеевич
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Шелманов Артём Олегович

Contents*

---

* The papers are ordered by the surname of the first author in compliance with the English alphabet.

A SIMPLE FINGERPRINT APPROACH TO EXTRACTING THE GLOBAL PROSODIC PROPERTIES FROM FIELD DATA[1]

Chechuro I. Yu. (ilyachechuro@gmail.com)[1],
Lyashevskaya O. N. (olesar@yandex.ru)[1, 2]

[1]National Research University Higher School of Economics, Moscow, Russia;
[2]Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper reports a method to create a speaker's prosodic fingerprint based on the global characteristics of the pitch movement. Prosodic fingerprint is the distribution of f0 in the low, middle, and high ranges and the distribution of pitch movements from one range into other [Šimko et al. 2017]. This fully automated method can be used to classify the records and to provide the reference level for more sophisticated analysis of the pitch movement and intonation strategies. We evaluate the method by applying it to the spontaneous Russian spoken data recorded in different regions. We model the correlation between the fingerprint and sociolinguistic features such as age, gender, and region. The results of this analysis allow to formulate several sociolinguistic hypotheses that can further be tested with a more detailed analytic technique.

Key words: prosodic fingerprint, speaker's prosodic portrait, pitch movement, unigram fingerprint, delta fingerprint, Russian prosody

1 The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

ПРОСТОЙ МЕТОД ПРОСОДИЧЕСКОГО ОТПЕЧАТКА ДЛЯ АНАЛИЗА ОБЩИХ СВОЙСТВ ИНТОНАЦИИ НА МАТЕРИАЛЕ ПОЛЕВЫХ ДАННЫХ

Чечуро И. Ю. (ilyachechuro@gmail.com)[1],
Ляшевская О. Н. (olesar@yandex.ru)[1, 2]

[1]Национальный исследовательский униветситет «Высшая школа экономики», Москва, Россия; [2]Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье обсуждается применение метода создания просодического отпечатка говорящего на основе общих характеристик движения основного тона. Просодический отпечаток—это распределение f0 в нижнем, среднем и верхнем диапазонах и распределение движений высоты тона из одного диапазона в другой [Šimko et al. 2017]. Этот полностью автоматизированный метод может использоваться для классификации записей в корпусе и получения представления о фоне, с которым будут сравниваться данные при дальнейшем, более сложном анализе стратегий интонирования. Мы применили метод к спонтанным русскоязычным данным, записанным в разных регионах. Разработаны модели анализа зависимости между данными просодического отпечатка и социолингвистическими характеристиками, такими как возраст, пол и регион. Результаты проведенного нами анализа данных позволяют сформулировать ряд социолингвистических гипотез, которые впоследствии могут быть проверены с использованием более глубоких методов анализа.

Ключевые слова: просодический отпечаток, просодический портрет говорящего, движение частоты основного тона, русская просодия

1. Introduction

The notion of the speaker's prosodic portrait [Kibrik 2009], [Kibrik, Fedorova 2018] is important as it assumes the existence of certain intrinsic, or neutral, properties of a particular speaker's voice. [Kibrik and Fedorova 2018] includes in the representation of the portrait features such as the f0 range (minimal and maximal f0 values), the standard level of the elementary discourse unit onsets, the level of fallings in the final and non-final positions, and the level of rises (and fallings on post-accent syllables) characteristic of the comma and "three dots" intonation. More detailed representations may contain hundreds of features. [Hönig et al. 2015] create fingerprints of recordings rather than speakers and include as many as 167 features. However, all such representations rely heavily on the manual (or partially automated) annotation of the underlying linguistic and perceptive information.

In this study, we propose a simple automated approach to creating primary speaker's prosodic fingerprint based on the global characteristics of the pitch movement. Prosodic fingerprint is the distribution of f0 in the low, middle, and high ranges and the distribution of pitch movements from one range into other [Šimko et al. 2017]. Using the pitch data extracted with the Praat software, we create fingerprints of 26 Russian speakers from different regions of Russia. We then conduct a case study and illustrate how this technique can be used to help formulate primary hypotheses regarding the relations between pitch movement and the sociolinguistic features of speakers such as age, gender, and region. This study is a part of a larger research project dealing with the regional aspects of the Russian intonation previously, in which data from different regions of Russia and ex-USSR are collected and documented.

Our study is complementary to the qualitative studies of Russian intonation that suggest the classifications of intonation patterns and their functional interpretations such as [Bryzgunova 1977], [Odé 1989], [Kodzasov & Krivnova 2001], [Korotaev & Podlesskaya 2008], [Kodzasov 2009, Grammatchikova et al. 2014], [Vol'skaya 2014], [Podlesskaya 2017], [Yanko 2017], [Korotaev 2018]. We use a quantitative approach, which makes our study more in line with the studies of [Skrelin and Volskaya 2006, 2008], and those based on the corpus of One Speaker's Day [Stepanova et al. 2008] and CORUSS [Kachkovskaia et al. 2016]) that aim at modeling the speakers' behaviour rather than propose overall generalizations. The study reported here is an attempt to automatically investigate the structure of the field data in order to formulate primary research hypotheses based on the observed phenomena.

Using the recordings made in four different regions of Russia (Krasnoyarsk, Moscow, Nakhodka and Novosibirsk), we analyze the pitch movement in spontaneous speech of the native speakers of regional Standard Russian. For each speaker, we recorded three samples of spontaneous speech: an interview, a dialogue and a retell of the Pear movie [Chafe 1980]. With the help of linear mixed effect modelling, we explore the correlation between the shape of the fingerprints and the biological sex, age and place of residence of the speakers.

The remainder of this paper is organized as follows. Section 2 discusses the experimental settings, and Section 3 addresses the data sampling. Section 4 presents the method to collecting the fingerprints based on the distribution of the pitch values. Section 5 reports the statistical analysis of the data, and in Section 6, a discussion of the obtained results is provided.

2. Participants and Experimental Conditions

All participants are monolingual native speakers of Russian born in Krasnoyarsk, Novosibirsk, Nakhodka, and Moscow. Krasnoyarsk and Novosibirsk represent Siberia, Moscow—Standard Russian and Nakhodka—Far East (the city population of which usually originates from different regions of the ex-USSR and is highly mixed). At the moment of the recordings, all the participants lived in their home regions or have recently moved to Moscow to study at the university (1st year students in the beginning of the 1st semester). All regional participants were divided into two age groups: from 25 to 40 years old vs. 45 years old and older. This division was made in order to balance the sample; in the analysis presented in this paper age was used as a numeric and not as a categorical variable. In each age group, there were two male and two female participants. The speakers from Moscow were represented by two females from the lower age group.

Each recording has been taken from two participants. In all pairs, the interlocutors knew each other relatively well (they were classmates, friends or relatives) and belonged to the same age and social group.

The spontaneous dialogues were recorded in the "fieldtrip" conditions in a quiet room using a recorder that supports .WAV format with no compression. The recordings made in Moscow, including those with the regional respondents, were made with a professional recorder and individual headset microphones for each speaker.

The experiment began with setting up the recording devices and instructing the participants. This stage took from 5 to 10 minutes. During this time, the participants could talk to each other freely and simultaneously get used to the recording equipment and the experimental environment.

2.1. Tasks for the Participants

There were three types of tasks. In the first task, the participants had to tell a small story about their life (e.g. parents and family, school, favorite teachers, hometown, etc.). The second task was an experiment with a map based on [Usacheva 2017], in which two participants, the instructor and the follower, were given a map of the Moscow Zoo printed on an A2 sheet and a set of objects (coins, pencils, dices, etc.) to place on the map. The experimenter placed objects on the Zoo map in front of the instructor and the instructor had to explain the positions of these objects to the follower so that the follower could repeat it on his map. During the experiment, the speakers communicated using mobile phones. The third part of the experiment implied retelling the Pear Movie [Chafe 1980] that was presented to the participants on the screen of the experimenter's laptop.

3. Data Sampling and Annotation

Each speaker in the dataset was represented by 40 randomly selected utterances. The utterances were extracted from each recording type using the following proportion: 15 recordings from the interview, 15 recordings from the experiment, and ten recordings from the pear story. The length of the recordings was not normalized. The pitch values were extracted from each recording with a 10 ms step. The pitch values were extracted with the standard functions of Praat. The pitch range (maximal and minimal pitch values) was defined for each speaker separately.

The data have been annotated in Praat. The first tier contained the boundaries of the speech units defined by pauses on the oscillogram. The parts of the recordings that contained sounds other than the participant's speech (experimenters' instructions, random noises) or were technically problematic to analyze (e.g. distortion or low volume units) were marked on a separate tier. These parts of the recordings were not used in the analysis.

4. Constructing the fingerprints

The analysis of the data presented in this section partly adopts the approach introduced in [Šimko et al. 2017]. Smoothing and wavelet transformations were omitted in the analysis. For each of the speakers, the pitch range was defined as the difference between the minimal and the maximal pitch values in all 40 recordings, with the exclusion of 5% of the observations: 2.5% with the minimal values and 2.5% with the maximal values [Fig. 1]. This type of range narrowing lowers the probability of including octave jumps and other artefacts into the analysis. Then, the pitch values were normalized by the z-score. The remaining range (95% observations) was divided into three equal parts that were coded, respectively, with −1 (Low), 0 (Medium) и 1 (High), which correspond to the commonly used division of the pitch range into Low, Medium and High [Keijsper 2003], [Odé 1989].

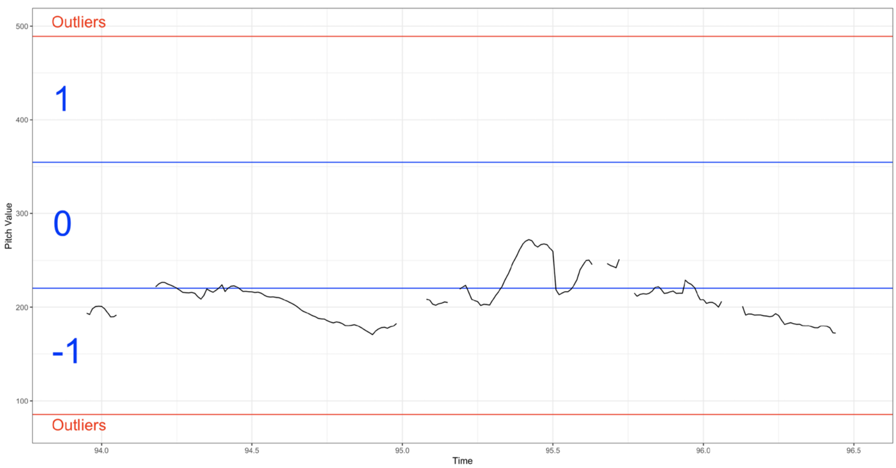

Fig. 1. Pitch sub-ranges in a recording sample

At the second step of annotation the transitions between the −1, 0 and 1 levels were encoded. The data were annotated as follows: if the points N and N+1 (taken with a 10 ms interval) are in the same pitch level, we interpret this as no-change in pitch shape and code it as 0. If the points are in different sub-ranges, the transition is coded as the difference between the levels: −2 (High to Low), −1 (Medium to Low, High to Medium), +1 (Low to Medium, Medium to High) или +2 (Low to High). This annotation was designed in order to distinguish substantial pitch movements from its minor fluctuations within a single sub-range.

The two types of the annotation were used to compose four datasets: the distribution of observations by the sub-ranges, the transitions between the sub-ranges, and the number of observations in each sub-range and the transitions of each type.

Each line of the dataset corresponded to one pitch value and contained the information about the speaker's name, their place of living, biological sex, age, text type, sentence ID from 1 to 40, time on the recording the observation corresponds to, the sub-range value −1, 0 or 1 (further called unigram) or the transition value −2, −1, 0, 1, 2 (further called delta). Upon this dataset, we created a new one, where each line corresponded to a sentence and the rows contained the meta information about the text and the number of unigrams or deltas of each type in this sentence.

Due to the size of the datasets of the first type (the size of each dataset in the .csv format was over 300 megabytes) and the limited resources of the personal computer used for statistical modelling, the data of the first dataset were not used in the current study. Nevertheless, we plan to use these data for statistical modelling using specifically designed systems with a better performance. The analysis presented in this paper was conducted using only the second type of datasets.

5. Data Analysis

The preliminary analysis of the data was conducted using histograms of the unigrams and deltas values per speaker. Figure 2 illustrates the distribution of unigrams per speaker, the histograms are colored by the region:

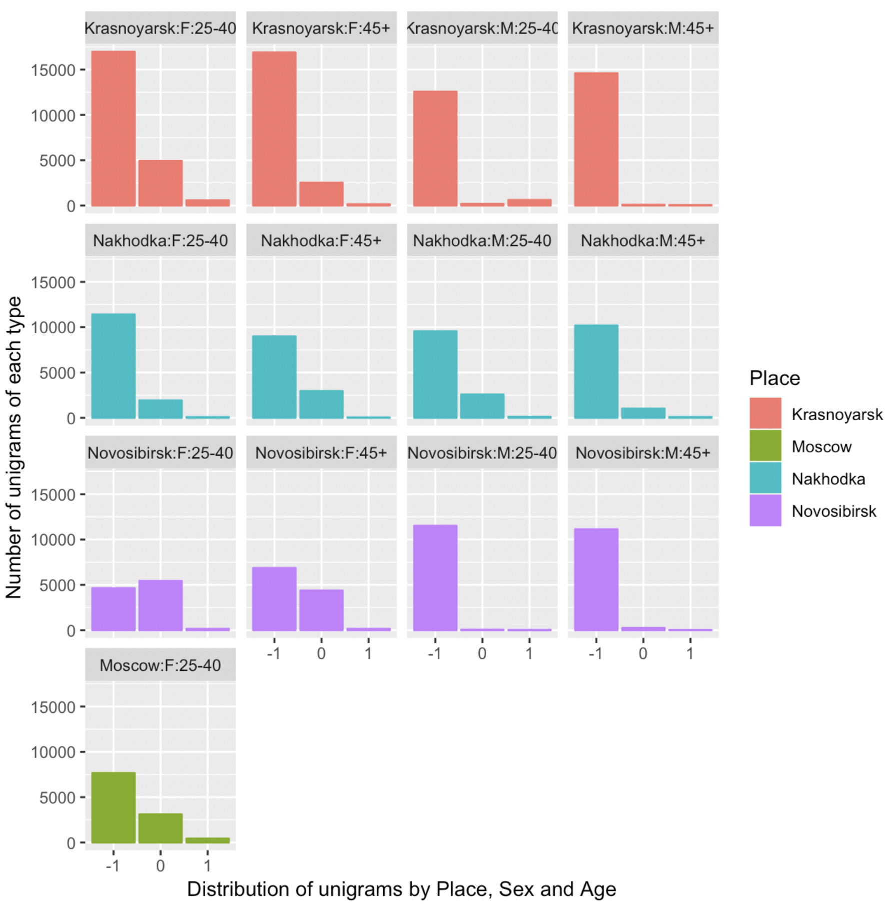

Fig. 2. The Distribution of the z-scored Unigram Values by Place, Sex and Age

The shape of the histograms in Figure 2 suggests that the difference between male and female respondents may play a significant role in Novosibirsk and Krasnoyarsk with women having a greater proportion of "0" unigrams, while in Nakhodka this difference is less pronounced and both sexes are similar. The histograms also indicate that the number of "1" unigrams may not be of much significance and the opposition can be viewed as "−1" vs. "not −1" values.

Figure 3 illustrates the distribution of deltas by speaker. The bar for "0" deltas is intentionally omitted since its relative size did not allow to observe the "1" and "−1" in men and women (the range for this bar is roughly between 10,000 and 20,000). The shape of the histograms suggests that the main difference between male and female respondents is in how often the "−1" and "1" deltas occur in their recordings. Again, in Nakhodka this difference is less pronounced than in other regions. Apart from three speakers, the number of −2 and 2 deltas is imperceptible and the main opposition appears to exist between 0 vs. −1 and 1 deltas. The data can thus be coded as "0" vs. "not 0" deltas indicating the presence and the absence of pitch movement.

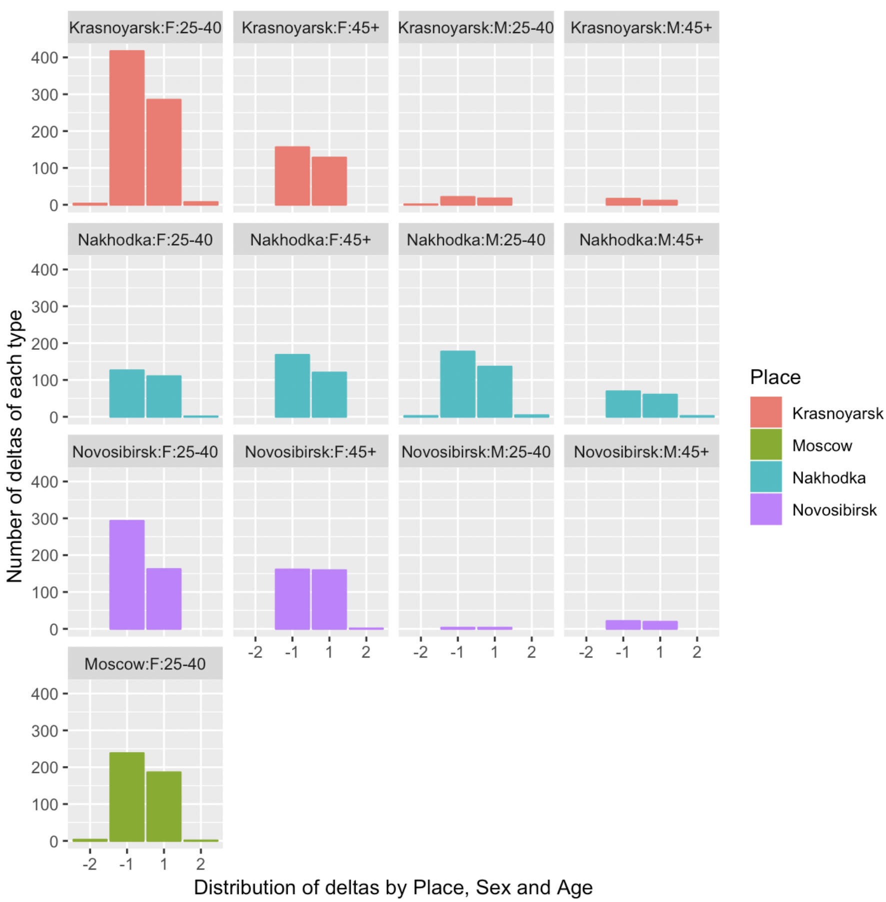

Fig. 3. The Distribution of the z-scored delta Values by Place, Sex and Age

In the following analysis of the data we use linear mixed effect modeling to explore the effects of biological sex, age, place of origin and type of text (dialogue vs. monologue) on the distribution of unigrams and deltas. We fitted the following models. The first model predicted the proportion of "–1" to "not –1" unigrams on the basis of biological sex, age, place of origin of the speakers, the type of the text and the speaker identity as a random effect lmer(X1Prop ~ Sex*Age + Place + TextType + (1 | Speaker_ID)). The second model predicted the same value on the basis of sex, age and the type of the text with the place of origin as a random effect lmer(X1Prop ~ Sex*Age + TextType + (1 | Place)). Then, we fitted two models with the same predictors for the proportion of "0" deltas to the "not 0" deltas: lmer(X0Prop ~ Sex*Age + Place + TextType + (1 | Speaker_ID)) and lmer(X0Prop ~ Sex*Age + TextType + (1 | Place)).

The predicted value in the first model was the proportion of "–1" unigrams to the "not –1" values. The controlled variables were biological sex, age, place of origin and type of text (dialogue vs. monologue) and the speaker ID as a random effect. The stepwise regression model selection with backward elimination has shown that the only significant variable are sex (p-value = 0.000244) and text type (p-value = 0.009803) with random intercepts by speakers. The effect of the two variables is provided in Fig. 4.

Figure 4 shows that the proportion of "–1" unigrams is significantly lower in females than in males. Similarly, the same proportion to a smaller degree is observed with respect to the text type. The factors of place and age turned out to be insignificant.

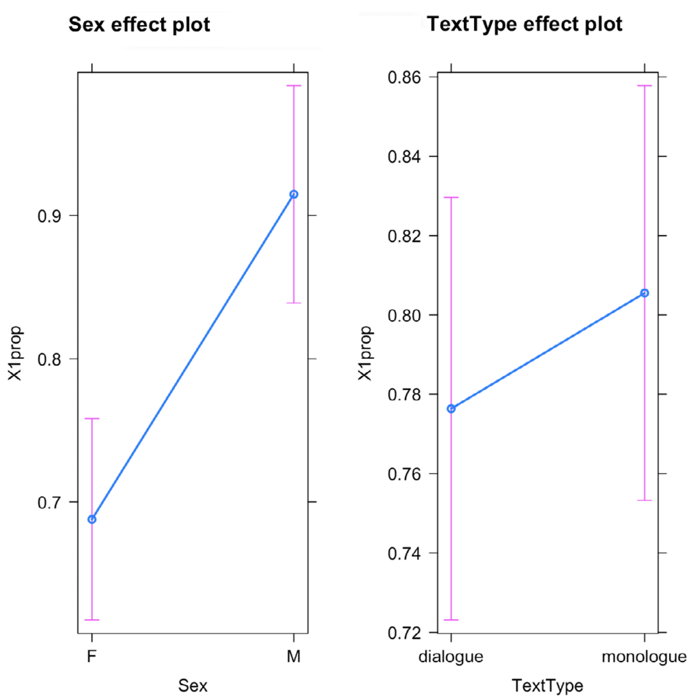

Fig. 4. The effect of biological sex and text type on the proportion of "–1" unigrams to "not –1" unigrams

The second model was delta-based and predicted the proportion of "0" deltas to the "not 0" deltas. The set of predictors was the same as in the first model. The backward model selection has shown that the only significant predictors is biological sex. Figure 5 illustrates that the amount of pitch movement in male participants is significantly lower than that in female participants.

The unigram-based model we fitted next was similar to the one described above with the only change being made to the random effect structure: instead of fitting random intercepts for speakers, we used random intercepts for different places. The backward stepwise regression model validation has shown that the significant predictors are sex, age and text type with the age-related change in pitch use being significant in both sexes.

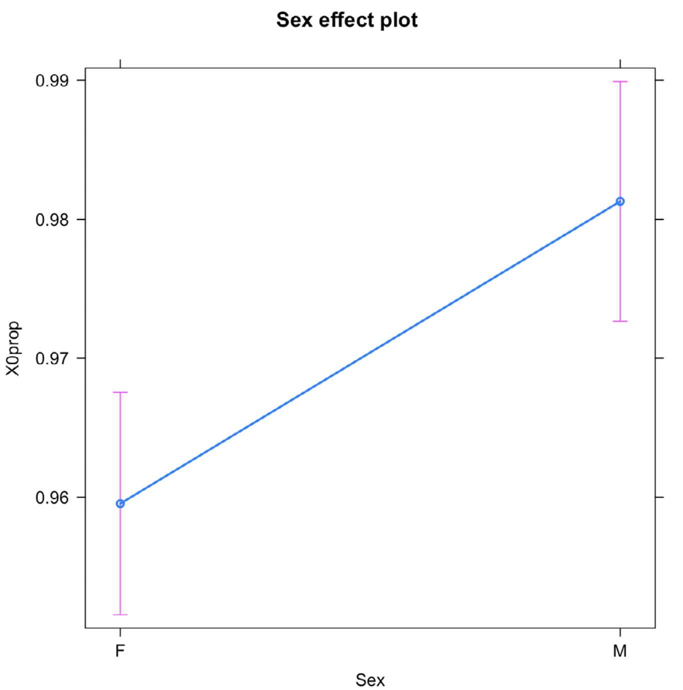

Fig. 5. The effects of biological sex and age on the proportion of "0" and "not 0" deltas

Similar changes were made to the delta-based model. The significant effects turned out to be the same as in the unigram-based model. Figure 6 and Figure 7 illustrate the effect of age with respect to biological sex. The effect of age is significant in both genders and is more pronounced in women (the older female respondents use less pitch movement than the younger ones). The models thus suggest that though there is no global effect of age as suggested by the first two models, it does exist within each city separately.

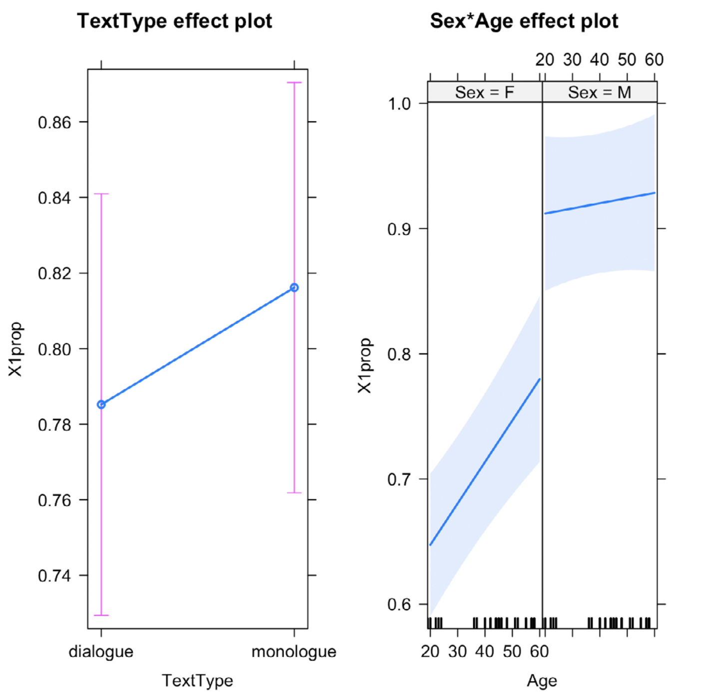

Fig. 6. Effect of age by sex and text type on the proportion of "–1" unigrams

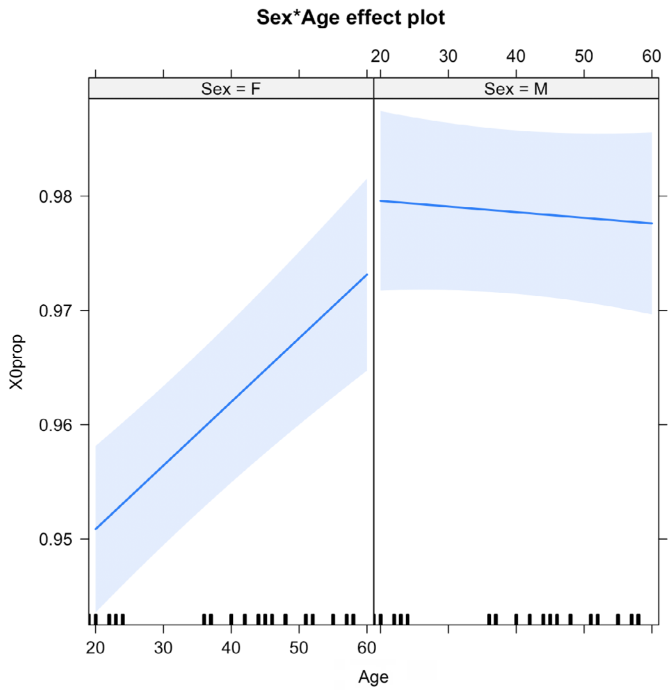

Fig. 7. Effect of age by sex on the proportion of "0" deltas

The last pair of models that we have fitted predicted the proportion of −1 unigrams and 0 deltas on the basis of the interaction between sex, age and place: *lm(formula = X0prop ~ Sex:Age:Place)* and *lm(formula = X1prop ~ Sex:Age:Place)*. The data for Moscow were removed from the dataframe since they only correspond to one age group and one biological sex.

Both models suggest that there is an age-related change in men in all regions (p-values < 0.01 in both models for all regions), while for women it is attested only in Krasnoyarsk (for “0” deltas, p-value = 0.007) and in Novosibirsk (for “–1” unigrams, p-value = 0.002). Figure 8 and Figure 9 illustrate the effect of age and sex on the proportion of –1 unigrams and 0 deltas.

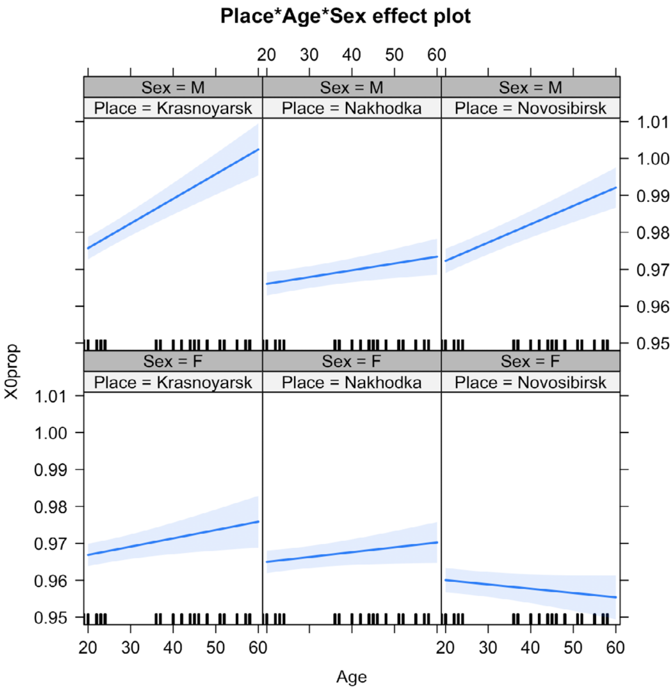

Fig. 8. Effect of age and sex by region on the proportion of “–1” unigrams

The last two models also allow to hypothesize that there is a major difference between the Siberian cities and Nakhodka: while in Novosibirsk and Krasnoyarsk the differences between speakers of different sexes are relatively clear, in Nakhodka men and women appear to use pitch more similarly. Another possible interpretation of this result may be that the biological sex in Nakhodka does play a role but our current annotation system does not track these differences.

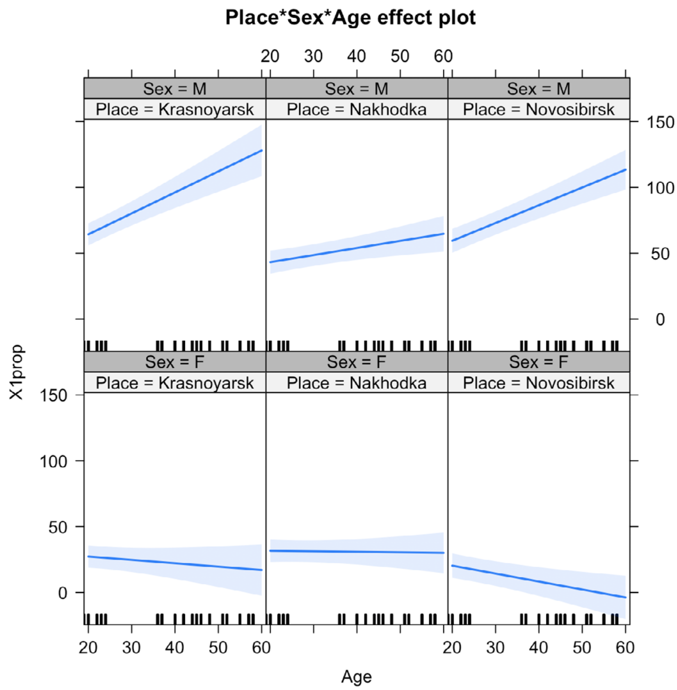

Fig. 9. Effect of age and sex by region on the proportion of "0" deltas

6. Discussion

The results of the regression analysis can be interpreted as follows. First, the first unigram-based regression model has shown that male and female speakers use the available pitch range differently. While males mostly use the "Low" part of the range, women use other parts of the range more often. Linguistically, this means that men may use less pitch movement in their speech, which may be related to a more expressive function of pitch in male speech (significant pitch changes are rare and therefore more noticeable). The first delta-based model has shown that male speakers cross the sub-range boundaries significantly less often than females, which supports the hypothesis of the comparably lower pitch use in their speech.

Another possible explanation of these results is that men divide their pitch range differently than women and our version of the tripartite division of the pitch range is not sensitive enough to track the pitch changes. The pitch movement of males may occur within a single sub-range (e.g. within "–1") and the remaining part of the range will be reserved for the rare utterances with an extreme degree of expression. There are thus two possible scenarios: (a) males use pitch movement more rarely than females and (b) the pitch movement in males has a lower amplitude in males but it is not necessarily less frequent. Both interpretations, however, suggest that there is a major difference between male and female speech and only differ in the nature of these differences, which means that the use of the pitch range that can be tracked automatically using a relatively simple technique. These results are interesting in the sense that they contradict the findings reported in the previous studies [cf. Skrelin, Volskaya 2006, 2008], where no gender-related differences have been reported. The difference between genders that is observed in our study may most likely be explained by the different choice of measured parameters. While in the previous studies the models measured and predicted the proportions of pitch curves of different types in different speakers, we model the overall amount of pitch movement regardless of the particular contours. The second reason may be that the previous studies did not consider regional variation, which appears to have an impact on the prosodic portraits of male and female speakers. In our study, gender-related differences were observed in Novosibirsk and Krasnoyarsk, while in Nakhodka they were not attested. This hypothesis, however, requires additional testing with the use of a larger data sample from each region.

Interestingly, the models with place as a random effect suggest that there is an age-related difference in male speakers with older speakers having a different amount pitch movement. Thus, though the age-related differences are not seen globally, they exist within each city. From the linguistic point of view, this means that older men use pitch differently from the younger ones but there is no such tendency in women. It may also mean that the parts of the pitch range get re-organised with the increase age and the available range starts to be used differently.

The regional difference between Nakhodka and other cities tracked by the last two models allows to hypothesize the existence of an areal comparative concept, namely of the difference between men and women in the intensity of the pitch use. This means that different regions of Russia may differ with respect to whether men intonate somehow differently than women or not. Another possible interpretation of this result may be that the biological sex in Nakhodka plays a role but our current annotation system does not track these differences. Both results, however, suggest that the regions of Russia differ with respect to how men and women use pitch.

7. Conclusion

In this paper, we proposed a simple fully automated method of portraying a speaker's pitch usage. We created fingerprints of 26 speakers from different regions of Russia and conducted statistical analysis of these data. The regression analysis has shown that even though our representation of data is very simplistic, it may reveal some significant correlations with the sociolinguistic features. The results of this analysis allow one to formulate several sociolinguistic hypotheses that can further be tested with a more detailed analytic technique and a larger data sample. The first hypothesis regards the differences in the use of the available pitch range in male and female speakers and the frequency of pitch change in the speakers of different biological sex. The second hypothesis is related to the age-related differences in the pitch use in male speakers. Finally, the third hypothesis concerns differences in pitch use between men and women across different regions.

8. Supplementary Materials

Data used in the analysis and source code for the R scripts are available at: https://github.com/author/screenedrepository.

References

1. *Bryzgunova E. A.* (1977), Sounds and intonation of Russian speech [Zvuki i intonatsiia russkoi rechi]. Moscow.
2. *Chafe W. L.* (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood, NJ: Ablex.
3. *Féry, C.* (2017), Intonation and Prosodic Structure. CUP.
4. *Grammatchikova E. V., Knyazev S. V., Luk'yanova L. V., Pozharitskaya S. K.* (2014), The rhythmic structure of the word and the place of realization of the tonal accent in regional variants of the modern standard Russian [Ritmicheskaya struktura slova i mesto realizatsii tonal'nogo aktsenta v regional'nykh variantakh sovremennogo russkogo literaturnogo yazyka], in Issues of theoretical and applied phonetics. Collection of articles in honor of O. F. Kryvnova [Aktual'nyje voprosy teoreticheskoj i prikladnoj fonetiki. Sbornik statej k jubileju O. F. Krivnovoy], Moscow, Buki Vedi.
5. *Hönig F., Batliner A., Nöth A.* (2015), How many speakers, how many texts—the automatic assessment of non-native prosody. In Proceedings of SLaTE 2015, Leipzig, September 4–5, 2015.
6. *Keijsper, C. E.* (2003), Notes on intonation and voice in modern Russian. Studies in Slavic and General Linguistics, 30, 141–214.
7. *Kachkovskaia T., Kocharov D., Skrelin P. A., Volskaya N. B.* (2016), CoRuSS-a New Prosodically Annotated Corpus of Russian Spontaneous Speech, Proceedings of LREC-2016.
8. *Kibrik A. A.* (2009), A speaker's prosodic portrait, Kibrik A. A., Podlesskaja V. I. (eds.), Night Dream Stories: A corpus study of spoken Russian discourse. Moscow: Jazyki slavjanskikh kul'tur.
9. *Kibrik A., Fedorova O.* (2018), A «Portrait» Approach to Multichannel Discourse, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan.
10. *Kodzasov S. V.* (2009). Studies in Russian prosody [Issledovanija v oblasti russkoj prosodii]. Moscow: Jazyki slavjanskikh kul'tur.
11. *Kodzasov S. V., Krivnova O. F.* (2001) General Phonetics [Obschaya fonetika]. Moscow.
12. *Korotaev N. A.* (2018), How intonation structures spoken narratives: non-final phase contexts [Intonatsionnaja struktura ustnogo rasskaza v kontekste nezavershennosti], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2018", Issue 17 (24), Moscow.

13. *Korotaev N. A., Podlesskaya V. I.* (2008), Prosody of clause-combining in Russian: A corpus-based study [Frazovaja akcentuacija v složnyx predloženijax s postpozitivnym pridatočnym v russkom jazyke: opyt ispol'zovanija ustnogo korpusa s prosodičeskoj razmetkoj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008", Issue 7 (14), Moscow, pp. 234–240.
14. *Odé, C.* (1989), Russian intonation: a perceptual description, Series: Studies in Slavic and General Linguistics, Vol. 13, Amsterdam–Atlanta, Rodopi.
15. *Podlesskaya V. I.* (2017), "Ja skazhu tebe s poslednej prjamotoj": Direct and indirect speech viewed through the prism of prosodically annotated corpus data [«Ja skažu tebe s poslednej prjamotoj»: prjamaja i kosvennaja reč' po dannym korpusa s prosodičeskoj razmetkoj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2017", Issue 16 (23), Vol. 2, Moscow, pp. 355–371.
16. *Šimko J., Suni A., Hiovain K., Vainio M.* (2017), Comparing languages using hierarchical prosodic analysis, Proceedings of Interspeech-2017, Stockholm, Sweden, Interspeech, ISCA, Baixas, pp. 1213–1217.
17. *Skrelin P., Volskaya, N.* (2006), Russian read and spontaneous speech: prosodic data analysis, Proceedings of Nordic Prosody 2006.
18. *Skrelin P., Volskaya N.* (2008), Prosodic model for Russian, Proceedings of Nordic Prosody 2008.
19. *Stepanova S. B., Asinovsky A. S., Bogdanova N. V., Rusakova M. V., Sherstinova T. Ju.* (2008), Speech corpus of the Russian everyday communication "One Speaker's Day": basic conception and current state [Zvukovoj korpus russkogo jazyka povsednevnogo obschenija «Odin rechevoj den'»: kontseptsija i sostojanie formirovaniya], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", Issue 7 (14).
20. *Usacheva M.* (2017), Experiments in the field: the experience of describing Permian languages [Eksperimenty v pole: opyt opisaniya permskikh yazykov]. A talk given at the workshop on Minor languages in the big linguistics [Malyje jazyki v bol'shoj lingvistike], Moscow, Russia, 2–3 November 2017.
21. *Vol'skaya N. B.* (2014), Intonation and language contact: pragmatic aspects of within- and cross-language interference [Intonatsija i jazykovoj kontakt: pragmaticheskij aspekt vnutri- i mezhjazykovoj interferentsii], Proceedings of XLII Int. philological conference, Selected papers, St.-Petersburg.
22. *Yanko T. E.* (2008), Intonation strategies of Russian speech in comparative aspect [Intonatsionnyje strategii russkoj rechi v sopostavitel'nom aspekte], Moscow, Jazyki slavjanskikh kul'tur.

APPLYING AN AUTOMATIC FTD CLASSIFIER TO THE ANNOTATION OF THE GICR CORPUS

Bulygin M. V. (bulyginmv1996@gmail.com), Radboud University, Nijmegen, Netherlands,
Sharoff S. A. (s.sharoff@leeds.ac.uk), Russian State University of Humanities, Moscow, Russia; Leeds University, Leeds, UK

This paper addresses the task of automatic genre classification for Russian within the Functional Text Dimensions (FTD) framework. Our aim in this study was to build the optimate FTD classification model to annotate web texts from the GICR corpus. For training data, we used an extended GICR dataset. We used the Support Vector Machine method with linear kernel for classification and converted training data to lower case to increase accuracy. During our research we experimented with several classification parameters, such as types of features, C-value and feature filtering to determine the best option for the classification model of the GICR dataset. The resulting model was able to achieve satisfactory classification accuracy and was used for GICR annotation. We also looked at the most significant features for each FTD in our best performing model and compared them to the most frequent words in which these features occur. Finally, we applied our model to segments of the GICR and looked at the FTD components in these segments.

A SIMPLE FINGERPRINT APPROACH TO EXTRACTING THE GLOBAL PROSODIC PROPERTIES FROM FIELD DATA

Chechuro I. Yu. (ilyachechuro@gmail.com)[1], **Lyashevskaya O. N.** (olesar@yandex.ru)[1, 2];
[1]National Research University Higher School of Economics, Moscow, Russia; [2]Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper reports a method to create a speaker's prosodic fingerprint based on the global characteristics of the pitch movement. Prosodic fingerprint is the distribution of f0 in the low, middle, and high ranges and the distribution of pitch movements from one range into other [Šimko et al. 2017]. This fully automated method can be used to classify the records and to provide the reference level for more sophisticated analysis of the pitch movement and intonation strategies. We evaluate the method by applying it to the spontaneous Russian spoken data recorded in different regions. We model the correlation between the fingerprint and sociolinguistic features such as age, gender, and region. The results of this analysis allow to formulate several sociolinguistic hypotheses that can further be tested with a more detailed analytic technique.

CLASSIFICATION MODELS FOR RST DISCOURSE PARSING OF TEXTS IN RUSSIAN

Chistova E. V. (chistova@isa.ru)FRC CSC RAS, Moscow, Russia;
RUDN University, Moscow, Russia, **Shelmanov A. O.** (shelmanov@isa.ru), Skoltech, Moscow, Russia, FRC CSC RAS, Moscow, Russia, **Kobozeva M. V.** (kobozeva@isa.ru), **Pisarevskaya D. B.** (dinabpr@gmail.com), **Smirnov I. V.** (ivs@isa.ru), FRC CSC RAS, Moscow, Russia, **Toldova S. Yu.** (toldova@yandex.ru), NRU Higher School of Economics, Moscow, Russia

The paper considers the task of automatic discourse parsing of texts in Russian. Discourse parsing is a well-known approach to capturing text semantics across boundaries of single sentences. Discourse annotation was found to be useful for various tasks including summarization, sentiment analysis, question-answering. Recently, the release of manually annotated Ru-RSTreebank corpus unlocked the possibility of leveraging supervised machine learning techniques for creating such parsers for Russian language. The corpus provides the discourse annotation in a widely adopted formalisation—Rhetorical Structure Theory. In this work, we develop feature sets for rhetorical relation classification in Russian-language texts, investigate importance of various types of features, and report results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank corpus. We consider various machine learning methods including gradient boosting, neural network, and ensembling of several models by soft voting.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 18 (25). 2019

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06