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 64 доклада 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 были отобраны 64 доклада 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 reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

ПРАГМАТИКА В ИНТЕРПРЕТАЦИИ СФЕР ДЕЙСТВИЯ (НА МАТЕРИАЛЕ ПИСЬМЕННЫХ РУССКИХ ТЕКСТОВ)[1]

Апресян В. Ю. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

Национальный исследовательский университет «Высшая школа экономики»,
Институт русского языка им. В. В. Виноградова РАН

PRAGMATICS IN THE INTERPRETATION OF SCOPE IN WRITTEN RUSSIAN TEXTS

Apresyan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

National Research University Higher School of Economics,
Vinogradov Russian Language Institute of the Russian Academy of Sciences

The paper is a corpus study of pragmatic factors involved in disambiguating sentences with negation and universal quantifier in written Russian and English, such as *Ja ne pozval vseh svoih dal'nih rodstvennikov,* 'I haven't invited all of my distant relatives.' Ambiguity results from differences in scope. If negation scopes over the quantifier, we get partial negation: 'I have invited some, but not all of my distant relatives.' If negation scopes over the verb, we get total negation: 'I haven't invited any of my distant relatives.' Our study is based on Russian and English data extracted from a variety of corpora.

We demonstrate that despite syntactic differences, Russian and English rely on similar mechanisms of disambiguation via pragmatic reasoning. We show that quantifier 'all' has different interpretations with verb vs. quantifier negation: emphatic in the former case and quantificational in the latter. Contextual markers for each reading are consistent with this difference. V-negation occurs with demonstrative pronouns, negatively connoted nouns and temporal modifiers, which add emphasis (*I don't want to talk to all these idiots*; *I haven't eaten all day*), while Q-negation occurs in the context of quantitative verbs that consolidate the interpretation of quantity (*I haven't listed all the options*).

[1] Публикация подготовлена в ходе проведения исследования по проекту «Factors in resolving scope ambiguity» (№ 18-01-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018–2019 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5–100».

THE CORPUS OF CONTACT-INFLUENCED RUSSIAN OF NORTHERN SIBERIA AND THE RUSSIAN FAR EAST[1, 2]

Khomchenkova I. A. (irina.khomchenkova@yandex.ru)

Lomonosov Moscow State University; Vinogradov Russian Language Institute & Institute of Linguistics, RAS; Moscow, Russia

Pleshak P. S. (polinapleshak@yandex.ru)

Lomonosov Moscow State University; Institute of Linguistics, RAS; Moscow, Russia

Stoynova N. M. (stoynova@yandex.ru)

Vinogradov Russian Language Institute & Institute of Linguistics, RAS; NRU HSE; Moscow, Russia

The paper presents a spoken corpus of contact-influenced Russian, which consists of oral spontaneous Russian speech of bilingual speakers of indigenous languages of Northern Siberia and the Russian Far East (Samoyedic, Tungusic, Chukotko-Kamchatkan). The texts included in the corpus were transcribed in ELAN in Standard Russian orthography and provided with a special system of manual annotation of contact-induced features developed for the corpus. The paper focuses mainly on this system of annotation, which is relevant in a wider context of annotating any kind of speech with "deviations" from the standard language variety (bilinguals', learners', dialectal speech etc.). The annotation tags are grouped in several separate levels: contact-induced morphological, syntactic, phonetic, lexical features etc. The exact meanings for the annotation tags were proposed on empirical grounds. Transcribed and annotated texts gain morphological annotation and search implementation based on the Tsakorpus platform. The aim of the project is to provide a useful resource for linguistic studies on language contact.

Key words: corpus linguistics, spoken corpora, Russian, minor languages of Russia, language contact

[1] The research was conducted with support of RSF grant No. 17-18-01649 (Dynamics of language contact in the circumpolar region).

[2] Many thanks to our colleagues who granted us their field records to include in the corpus and to the anonymous reviewers of "Dialogue-2019".

КОРПУС КОНТАКТНО-ОБУСЛОВЛЕННОЙ РУССКОЙ РЕЧИ БИЛИНГВОВ-НОСИТЕЛЕЙ МАЛЫХ ЯЗЫКОВ СЕВЕРА СИБИРИ И ДАЛЬНЕГО ВОСТОКА

Плешак П. С. (polinapleshak@yandex.ru)
МГУ им. М. В. Ломоносова;
Институт языкознания РАН; Москва, Россия

Стойнова Н. М. (stoynova@yandex.ru)
ИРЯ им. В. В. Виноградова; Институт языкознания, РАН;
НИУ ВШЭ; Москва, Россия

Хомченкова И. А. (irina.khomchenkova@yandex.ru)
МГУ им. М. В. Ломоносова; ИРЯ им. В. В. Виноградова;
Институт языкознания, РАН; Москва, Россия

В статье описан создаваемый нами корпус контактно-обусловленной русской речи, который состоит из устных спонтанных текстов на русском языке, записанных от билингвов Севера Сибири и Дальнего Востока, носителей самодийских, тунгусских и чукотско-камчатских языков. Тексты расшифрованы в стандартной русской орфографии и снабжены специально разработанной ручной разметкой контактно-обусловленных грамматических особенностей в программе ELAN. Наиболее подробно в работе обсуждается опыт разметки, который может быть интересен в более широком контексте аннотирования речи, так или иначе отклоняющейся от литературной нормы (речи билингвов, изучающих иностранный язык, диалектной речи и т. д.). Разметка разделена на несколько уровней: контактно-обусловленные морфологические, синтаксические, лексические, фонетические особенности и т. д. Корпус частично доступен онлайн на платформе Tsakorpus с возможностью поиска по разработанной нами разметке контактно-обусловленных черт, морфологической разметке и метаданным. Цель проекта — создание удобного ресурса для исследований в области языковых контактов.

Ключевые слова: корпусная лингвистика, корпуса звучащей речи, русский язык, малые языки России, языковые контакты

1. Introduction

In the paper, we will present a new corpus of Russian spoken by bilinguals and discuss some problems of annotating "deviations" from the standard language variety, relevant for corpora of speech of bilinguals, learners, heritage speakers, people with speech disorders, as well as for child speech and dialectal corpora.

The corpus constitutes a transcribed and annotated collection of oral spontaneous Russian speech of bilingual speakers of indigenous languages of Northern Siberia and Russian Far East (Samoyedic, Tungusic and, to a smaller extent, Chukotko-Kamchatkan). The majority of the texts are short narratives.

The transcription is made in ELAN in standard Russian orthography with a simplified intonation marking and with the manual annotation of contact-induced features.

The text collection, which is planned to be included in the corpus, consists by the moment of ca. 100 hours of records. Ca. 29 hours of records have been already transcribed and annotated, these texts are available offline in the ELAN-format. A small test text sample was added to the online resource, which is being created for the corpus on the Tsakorpus platform: http://web-corpora.net/tsakorpus_russian_nonst/corpus.html. Transcribed and annotated texts gain morphological annotation and search implementation based on the platform.

The resource is aimed to be used by specialists on language contact to trace the influence of indigenous languages of the area on the Russian speech of their speakers. The corpus is the most convenient to study contact-induced morphosyntactic features. However, it also can be used in other studies on language contact, e. g. studies on lexicon and phonetics.

The paper is comprised of 6 parts. In **Section 2**, we discuss some corpus projects which are similar to ours. **Section 3** presents the text collection included in the corpus: the amount of data, types and genres of texts, the narrators and languages they speak. **Section 4** describes our conventions of transcription (**4.1**), the system of annotation of contact-induced grammatical features used in the corpus (**4.2**) and the online searching interface (**4.3**). In **Section 5**, we list some studies on language contact based on our corpus data. Section 6 contains brief concluding remarks and plans on further development and use of this corpus.

2. Similar projects on bilinguals' Russian

There are some parallel projects, devoted to other varieties of Russian, spoken by bilinguals or learners. For example, resources the most close to ours are corpora made by Linguistic Convergence Laboratory, HSE—the corpus of Daghestanian Russian (DagRus, http://www.parasolcorpus.org/dagrus/#, cf. [Daniel & Dobrushina 2013]) and the corpus of Chuvash Russian (ChuvashRus, http://www.parasolcorpus.org/chuvashrus/). These corpora consist of oral spontaneous texts collected in the form of sociolinguistic interviews. In contrast to our corpus, they do not include any special annotation of contact-induced features.

One more similar resource is Russian Learner Corpus created in Linguistic Laboratory for Corpus Studies, HSE (http://www.web-corpora.net/RLC/), cf. [Rakhilina 2016];

[Rakhilina et al. 2016]. It consists of texts of speakers who learn Russian as their second language and of heritage speakers of Russian. The texts are mostly written. They are provided with the annotation of non-standard grammatical features ("errors" in terms of its creators), similar to ours.

Texts of bilingual speakers were also included in the spoken subcorpus of Russian National Corpus (http://ruscorpora.ru/search-spoken.html), see [Savchuk 2018] for more detail. They are provided with standard grammatical annotation of Russian National Corpus. Unfortunately, by the moment, the user has no possibility to separate this text collection from oral texts of monolinguals.

3. Text collection

The text collection consists of spontaneous oral texts, mostly short narratives (folklore, biographies) and descriptions (ethnographic texts, recipes etc.); some texts are everyday dialogues with linguists. They were collected by us and by our colleagues as a "by-product" of current language documentation projects. For many of them we have also parallel (or near-parallel) versions in the indigenous language.

The overall text collection includes ca. 100 h. of records. Tungusic and Samoyedic varieties are the best represented by the moment. We also have modest collections from speakers of Chukchi, Yakut and Yukaghir.

By now, we have transcribed and annotated ca. 29 h. (out of 100 h.), which is approximately 117,000 words. The **Table 1** represents the total amount of textual data in hours and words.

Table 1. Text collection

	all in hours	annotated in hours	annotated in words
Enets (Forest and Tundra)	26.5	12.5	49,128
Nenets	9	1.5	9,292
Nganasan	10	6	19,072
Nanai	42	8	29,076
Ulch	8.5	1	10,564
Even	1	0	0
Chukchi	1.5	0	0
multilingual speakers from Lower Kolyma	2	0	0
total amount	100.5	29	117,132

The majority of these languages have a comparable sociolinguistic situation: they are endangered; the typical speaker acquired Russian at school age, but now they use actively both Russian and the indigenous language or almost only Russian.

For each text we also collected some metadata: 1) technical information on the record: file name, record date, record place, duration; 2) information on the text: type, genre, content, existence of a parallel version in the indigenous language; 3) information on the narrator: name, code, indigenous language(s) (s)he speaks; 4) information

on transcription and annotation: annotator, date, size (in clauses). For narrators, we also have separate more detailed metadata: name, birth place, birth date, place of residence, level of education, acquisition age (for Russian), indigenous language(s) (s)he speaks, and short sociolinguistic biography. Unfortunately, there remain a lot of gaps by the moment.

4. Transcription and annotation

4.1. Transcription in ELAN and the structure of tiers

The text for the corpus are transcribed in ELAN in standard Russian orthography with a simplified intonation marking (rising and falling tones indicated after words bearing phrasal accents) and a small number of special marks for the features of oral spontaneous speech (self-corrections, pauses, non-speech sounds, fragments in the indigenous language), see **Fig. 1**. The standard orthography was chosen for technical reasons, cf. [von Waldenfels et al. 2014] for the same decision and its reasons. Contact-induced or dialectal phonetic features are not reflected in the transcription, some of them are annotated in special tiers (see **Section 4.2**).

Fig. 1. Transcription and annotation in ELAN

Proper names are marked with square brackets (e. g. *M*[*ауш*]*а*) to become automatically anonymized in the web version of the corpus[3]. Fragments in the indigenous language, which sometimes occur in our texts, are transcribed if the annotator is familiar with the language enough or remain untranscribed (in this case we use a special mark *CS*).

The texts are segmented into clauses, or intonation units, more or less corresponding to clauses in oral speech. Ideally, 1 ELAN-annotation ≈ 1 clause. In practice, we rely more on intonation and pauses than on syntactic structure. In case of discrepancy between clausal boundaries and pausation, annotation boundaries correspond to pauses.

[3] The corresponding audio-fragments have not been anonymized by the moment.

We have separate transcription tiers for each participant of the conversation. Besides the transcription tier, which is synchronized with the audio data, there is a word tier, 6 special tiers for the annotation of contact-induced features (see **Section 4.2**), having the word tier as a parent, and some technical tiers. The latter include tiers for comments and for the translation of code-switching fragments. See the structure of the tiers in **Fig. 2**.

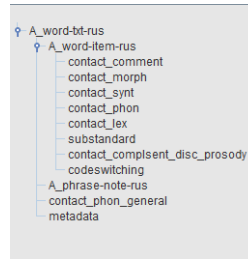

Fig. 2. Structure of tiers in ELAN

4.2. Annotation of contact-induced features

We use 5 ELAN-tiers for annotation of contact-induced features on different levels: phonetics, lexicon (loanwords and calques), morphology (including productive derivation, inflection and the use of grammatical categories), syntax (clause-level), one general tier is reserved for complex sentences, discourse and prosody.

One more tier ("substandard") is used for peculiarities that are presumably of non-contact nature (see some examples below and the discussion on the choice between particular tags).

To make the manual annotation of contact-induced features more structured and convenient for the search, we use Controlled Vocabulary incorporated into ELAN. For each tier (level) we have a set of tags among which the annotator can choose, see **Fig. 1**. The particular features and their values were chosen on empirical grounds. After a preliminary set had been proposed, it was used in the annotation during the testing period. Afterwards, the tags were discussed, some tags were added, which is considered to be enough for the text collection so far. Some tags are specific for texts produced by speakers of a concrete indigenous language. However, most of them are general enough to be used throughout the whole Northern Siberian corpus and even to be applied to other text collections. The tool is flexible and more tags can be added if needed.

The tags are ascribed to the words. There can be more than one tag ascribed to one word. Syntactic tags are ascribed to the word that manifests the syntactic relation. Usually, this is the dependent. For instance, the non-standard agreement tag is attached to adjectives, the non-standard argument encoding tag is attached to nouns etc. Intonation tags are attached to the accent-holder.

One of the problems with the Controlled Vocabulary is that it lacks hierarchical structure. So, in one tier, the annotator chooses among several possibilities without any further subdivision. We resolved this problem first, setting up separate tiers for

each level (phonetics, lexicon, morphology, syntax, complex sentences & discourse & prosody), and second, introducing complex names for tags: e. g. *agr_adj_gender*, *agr_adj_num*, *agr_adj_case*. All the three tags are used to indicate phenomena of disagreement but only for adjectives (in contrast to the verbal or anaphoric disagreement). Moreover, each tag is specified for the features that are involved. Therefore, we have a large set of disagreement tags within the syntactic tier, which are the following: *agr_adj_gender*, *agr_adj_num*, *agr_adj_case*; *agr_verb_gender*, *agr_verb_num*, *agr_verb_pers*; *gr_anaph_num*, *agr_anaph_gender*.

Such a fine-grained subdivision in the disagreement domain is due to the fact that it is one of the most frequent features in the non-standard speech. Having this powerful inventory, one can search choosing different sets of tags, in accordance to the purposes (see the description of a corpus-based study on gender disagreement in **Section 5**). The inventory of morphological features, which are more rare, is smaller.

By the moment, we use 73 tags in total. The level of morphology (including word-formation, inflection, use of grammatical categories) contains 10 tags[4]. The level of syntax (only within the clause) is the most elaborated and it contains 23 tags. The level of complex sentences, discourse and prosody contains 12 tags. The level of lexicon contains 3 simple tags: one for loanwords, one for calques, and one for non-evident cases. The level of phonetics includes 19 tags, almost all of them are very specific (the non-standard realization of a particular phoneme or a small group of phonemes) and the inventory of tags in use varies a lot across particular local varieties of bilingual Russian included in our sample. The phonetic and prosodic features, in contrast to morphological and syntactic ones, are marked with special tags not very consistently, since they are too frequent to mark them all and not clear enough for perception to mark them appropriately during the transcription without any additional instrumental analysis. So phonetic and prosodic tags are used only to mark striking clear cases just for an easy search of illustrative examples.

The level of "substandard" (non-contact) features contains 6 tags (one for each level: phonetics, morphology, syntax, lexicon etc.). These (dialectal, regional, register) features are not in our main focus, so they are not annotated very consistently either. The main reason to annotate them in our corpus is to make it possible for a user to differentiate between these features and contact-induced ones. In less clear cases, we use the corresponding "contact" tag, the "substandard" tags are reserved for more evident cases of non-contact features. However, since we cannot attribute all cases for

[4] Non-standard inflection and derivation patterns must be interpreted as under-acquisition of Russian rather than copying of the corresponding indigenous patterns (such as lexical calques or argument encoding patterns inherited from the indigenous language). In our annotation we do not differentiate between these two types of features, marking all of them as contact-induced. Another problem is to differentiate between contact-induced under-acquisition and non-standard inflectional and derivational patterns that can be produced also by monolinguals as occasional speech errors or as features of uneducated speech. There is no clear borderline between them. Our technical decision is to provide with tags as many cases as possible, ranking them according to the probability to be contact-induced. The annotator distributes them between the "contact" tier (= probable to be contact-induced, cf. the form *стаёт* 'becomes') and the "substandard" one (= less probable to be contact-induced, cf. *радовается* 'is glad'), basing on his/her intuition, see below on the substandard tier.

sure, we try to annotate everything that deviates from standard monolingual Russian[5] not to miss any relevant information. The aim of the annotator is not to make a right choice in all particular cases (it is generally impossible without a special investigation), but rather to rank the attested peculiarities roughly according the probability to be of a contact nature[6]. Therefore, our "contact" tags mark cases that are likely to be of contact nature and our "substandard" tags mark cases that have a chance to be interpreted as contact-induced. We leave the final decision to users of the corpus, giving them the access to both types of cases.

Table 2. Contact-induced features and tags

level (tier)	N of tags	examples (tags)
phonetics	19	***с***аръ 'king' (affr), ***т***арик 'oldman' (clust)
lexicon	3	*крупы* ***налила*** (calque)
morphology	10	*за неделю* ***пил*** (asp), *обитает****ся*** (refl)
syntax	23	*икра нету* (neg), *сетк****а*** *кинет* (gov_dom)
complex sentences & discourse & prosody	12	*попросили кто-нибудь увез* (subord_compl), *А чо* ***грит*** *мне от тебе надо\ Я\ ей говорю* (disc_word), *А там [****красивый****\ девушки]RHEME гоняют\ их* (pros_accent)
substandard	6	*у мен****е*** (morph), ***с*** *города* (synt), *балдой* (lex)

The full list of tags with short descriptions and illustrative examples is available at http://web-corpora.net/tsakorpus_russian_nonst/corpus.html.

4.3. Web-interface

The online interface for the corpus was implemented on the platform Tsakorpus (https://bitbucket.org/tsakorpus/tsakonian_corpus_platform), which had been developed by T. Arkhangelsky for small spoken corpora created in ELAN. The platform provides the possibility of search on grammatical features (the annotation mystem), search and filtration on metadata and search on any specific tag set used in a particular corpus (the annotation of contact-induced grammatical features in our case), see **Fig. 3**. Search results are given per clauses (with possibility of enlarging the context), both the transcribed fragment and the audio-fragment are available, see **Fig. 4**.

[5] The question arises, which monolingual variety must be chosen as tertium comparationis. The best option would be to use a text sample, the most comparable to ours: oral narratives produced by monolinguals of the same area and of the same sociolinguistic background as our narrators. Having no access to such texts, we rely on the intuition of the annotator, trying to mark with any tag as many "non-standard" features as possible.

[6] A more "honest" and simple way would be to mark on equal terms everything that the annotator assesses as non-standard, without any further differentiation during the annotation process. But in this case too much useless information would fall into the annotation.

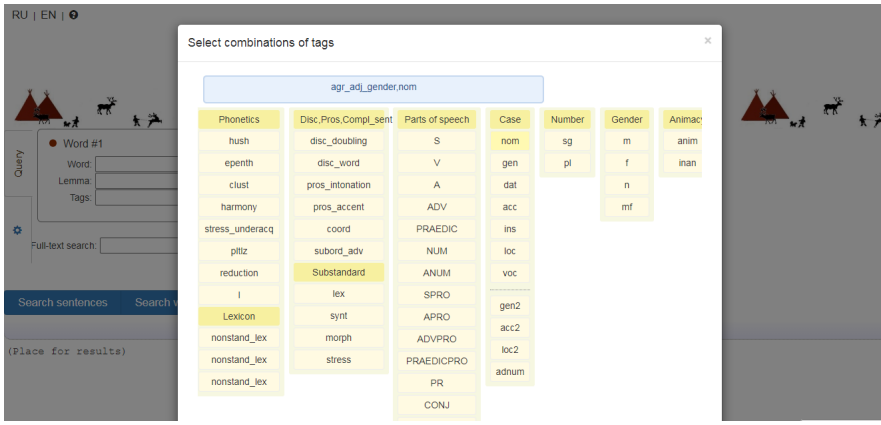

Fig. 3. Web-interface: search on grammatical features and contact-induced features

Fig. 4. Web-interface: search results

The user can find comparable samples of “standard” and “non-standard” uses, combining the search on grammatical tags provided with the platform and the search on our tags of contact-influenced features. For instance, one can find all occurrences of standard prepositional phrases (such as *в доме* ‘in the house’), using grammatical tags (the query “PR”), and then all non-standard occurrences with preposition drop (such as *доме* ‘(in) the house’), using tags of contact-induced features (the query “prep_drop”), see **Section 5** for the study based on these data.

At the moment, the online resource is working in a test mode at http://web-corpora.net/tsakorpus_russian_nonst/corpus.html. Only a small part of our transcribed and annotated text collection has been placed on the web. We are planning to enlarge the range of metadata types available for search, to make the search on contact-induced features more user-friendly and then to add the whole text collection. The next step is disambiguation of the grammatical annotation, which will make the search much more effective.

5. Using the corpus

The aim of the project is to provide a useful resource for linguistic studies of contact-induced language changes. In this section, we present studies conducted on the data of this corpus to illustrate possibilities of its application.

In some of them, the corpus served just as a source of examples, which were used to describe non-standard grammatical features attested in bilingual Russian in detail. Basing on the data of the Tungusic subcorpus, [Oskolskaya and Stoynova 2017] proposed a classification of uses of the construction *делал был*, *делал было* (V.PST + *be*.PST) in Nanai Russian and compared them to those of the similar construction in monolingual Russian and the pluperfect construction with the verb 'be' in Nanai.

One more way of using corpus data in the research of contact features in grammar is to calculate the frequency of "standard" (typical of monolingual Russian) and "non-standard" uses in the Russian speech of bilingual speakers and to reveal correlations with the grammatical context. In [Khomchenkova et al. 2018], gender disagreement in Russian speech of speakers of Southern Tungusic and Samoyedic languages of the elder generation was investigated (*бабка помер* 'old woman die.PST.MASC', *моя папка* 'my.FEM father'). The corpus data show that bilingual speakers are less likely to follow the standard agreement pattern for adjectives and more likely to choose the standard form of verbs and especially of anaphoric elements.

The data of the corpus can also be used to get a complex picture on some particular variety of bilingual Russian. For instance, in the grammatical description of Souther Tungusic Russian (Stoynova, to appear) the author gives some quantitative data on the relative frequency of different contact-induced grammatical features attested in this variety.

The list of some other studies on the data of this corpus is available at http://web-corpora.net/tsakorpus_russian_nonst/publ.html.

6. Conclusion

The present project has three main advances. First, it contributes to the overall collection of spoken corpora of Russian that are open source and can be used in linguistic studies. Second, it represents the speech of bilingual speakers and can serve as a representative data source for studies on language contact. Third, an important point, which was described in the paper in great detail, is a special system of annotation of contact-induced grammatical and lexical features created for the corpus. It reflects the peculiarities attested in particular varieties of Russian we deal with. However, it is quite flexible to be adapted for other contact-influenced varieties of Russian. The presence of such annotation gives the possibility to apply quantitative methods in studies of contact-induced features as these are difficult to search using only morphological tagging, concrete lemmas and regular expressions.

This experience also contributes to a more general problem, relevant for corpus linguistics, namely the problem of annotating any kind of speech, "deviating" anyhow from the standard language variety, including speech of learners, heritage speakers, children, people with speech disorders, as well as speech with regional and dialectal features.

We are planning to develop the project in the following directions. First, we will continue transcribing and annotating the existing text collection; the expansion to other bilingual varieties is in further plans as well. Second, we will continue the work on the online resource. The whole transcribed text collection will be placed on the web. The search interface will be improved—particularly, the search on different types of metadata will be added and the search on contact-induced features will become more user-friendly. One of our current plans is manual disambiguation of the automatic morphological annotation, which is used in the corpus.

References

1. *Bayda K., Kholodilova M., Kozhemjakina A., Romanova E., Remizova T., Storozheva A., Tarasova N., Zorina A., Morozova V., Panova A., Dobrushina N.* (2018) ChuvashRus Corpus, Moscow: Linguistic Convergence Laboratory, NRU HSE, available online at http://www.parasolcorpus.org/chuvashrus/.
2. *Daniel M. A., Dobrushina N. R.* (2013), A corpus of Russian as L2: the case of Daghestan [Russkij jazyk v Dagestane: problemy jazykovoj interferencii], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", 12(1), Moscow: RSUH, pp. 186–211.
3. *Dobrushina, N., Daniel M., von Waldenfels R., Maisak T., Panova A.* (2018), Corpus of Russian spoken in Daghestan, Moscow, Linguistic Convergence Laboratory, NRU HSE, available online at http://www.parasolcorpus.org/dagrus/.
4. *Khomchenkova I. A., Pleshak P. S., Stoynova N. M.* (2018), Gender disagreement in the contact-influenced Russian of Northern Siberia and the Russian Far East, presented at the conference "TheGen", Berlin, 14–15.06.2018.
5. *Oskolskaya S. A., Stoynova N. M.* (2017), Nanai verb categories in the Russian speech of Nanai speakers: 'be'-constructions [Nanajskije glagoljnyje kategorii v russkom jazyke nanajcev: konstrukcii tipa "byli delali"], presented at the conference "Russian grammar: describing, teaching, testing [Russkaja grammatika: opisanije, prepodavanije, testirovanije], Helsinki, June 7–9, 2017.
6. *Rakhilina E., Vyrenkova A. et al.* (2016), Russian Learner Corpus. Moscow: Linguistic Laboratory for Corpus Studies, NRU HSE, available online at http://www.web-corpora.net/RLC/.
7. *Rakhilina E., Vyrenkova A., Mustakimova E., Ladygina A., Smirnov I.* (2016), Building a learner corpus for Russian, Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016, pp. 66–75.
8. *Rakhilina E. V.* (2016), On a new instrumentary for describing Russian grammar: the corpus of errors [O novyx instrumentax opisanija russkoj grammatiki: korpus ošibok], Russkij jazyk za rubežom, 3, pp. 20–25.
9. *Russian National Corpus*, available online at http://www.ruscorpora.ru.
10. *Savchuk S. O.* (2018), Russian speech in polyethnic regions [Russkaja rečj v polietničeskix regionax], presented at the conference "Indigenous languages in contact with Russian: morphosyntactic and semantic interference", 30.11–01.12, 2018, Moscow: Russian Language Institute RAS, available at: https://drive.google.com/file/d/1GmJ4wxGM4DIWdSZftzrggcsnVvoWIyd3/view?usp=sharing.

11. *Stoynova N.* (to appear), Russian in contact with Southern Tungusic languages: evidence from Contact Russian Corpus of Northern Siberia and the Russian Far East, Slavica Helsingiensia, in print.
12. *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: Kompjuternaja lingvistika i intellektualjnye tekhnologii: Po materialam jezhegodnoj Mezhdunarodnoj konferentsii "Dialog" (Bekasovo, 4–8 June 2014) / V. Selegey. (ed.) № 13(20). M.: RSUH,. pp. 720–728.