View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repository of the Academy's Library

Sound-aligned corpus of Udmurt dialectal texts

Timofey Arkhangelskiy
Universität Hamburg / Alexander von Humboldt Foundation
timarkh@gmail.com

Ekaterina Georgieva
Research Institute for Linguistics
Hungarian Academy of Sciences
ekaterina.georgieva@nytud.mta.hu

Abstract

The paper describes an ongoing effort aiming at building a sound-aligned corpus of Udmurt spoken texts. The corpus currently consists of about 3.5 hours of recordings, collected during fieldwork trips between 2014 and 2016. The recordings represent three dialect groups of Udmurt (Northern, Central and Southern). The recordings were transcribed with the help of native speakers. All morphological peculiarities characteristic of spoken or dialectal Udmurt were faithfully reflected, however, the transcription was somewhat normalized in order to facilitate morphological annotation and cross-dialectal search. The pipeline of our project includes aligning the texts with the sound in ELAN and annotating them with a morphological analyzer developed for standard Udmurt. We use automatic annotation as a much less time-consuming alternative of manual glossing and explore the resulting quality and the downsides of such annotation. We are specifically investigating how much and what kind of change the standard analyzer requires in order to achieve sufficiently good annotation of spoken/dialectal texts. The corpus has a web interface where the users may execute search queries and listen to the audio. The online interface will be made publicly available in 2018.

Kivonat

Ezen tanulmányban egy pilot projektet mutatunk be, amely célja egy hanganyagot tartalmazó udmurt nyelvjárási korpusz építése. A készülő korpusz 2014 és 2016 között végzett terepmunkák során gyűjtött, jelenleg körülbelül 3,5 órányi lejegyzett hanganyagból áll, amely az udmurt nyelv fő nyelvjáráscsoportjait (északi, közép- és déli nyelvjárásait) mutatja be. A hangfelvételek lejegyzése udmurt anyanyelvi beszélők segítségével történt. A lejegyzés hűen tükrözi a hangfelvételeken előforduló, az udmurt nyelvjárásokra vagy az udmurt beszélt nyelvre jellemző morfológiai jelenségeket. A lejegyzés azonban fonetikai szempontból bizonyos mértékben sztenderdizálva lett annak érdekében, hogy megkönnyítse a szövegek morfológiai elemzését és a több nyelvjárásra kiterjedő keresést. A szövegek feldolgozása a következő lépésekből áll: a szövegek ELAN-nal való lejegyzése (amelynek során a legjegyzett szöveg időben illesztve lesz a hanganyaghoz),

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), pages 26–38, Helsinki, Finland, January 8–9, 2018. ©2018 Association for Computational Linguistics

majd az udmurt irodalmi nyelvre fejlesztett morfológiai elemzővel való annotálása. A korpuszépítés során az automatikus annotálás mellett döntöttünk, amellyel sok idő megspórolható a manuális annotáláshoz képest. Cikkünkben megvizsgáljuk az automatikus annotálás alkalmazhatóságát, különös tekintettel arra, hogy milyen mértékű és típusú módosításokat kell elvégezni az irodalmi udmurt nyelvre fejlesztett morfológiai elemzőn, hogy az a beszélt nyelvi és nyelvjárási szövegek elemzésére is alkalmas legyen. A korpusz online felülettel rendelkezik, amely lehetővé teszi a felhasználók számára az adatok lekérdezését és a hanganyag meghallgatását. Az online felületet a 2018-as év folyamán nyilvánossá tervezzük tenni.

Аннотация

В этой статье описывается текущий проект, в рамках которого планируется создать звуковой корпус устных текстов на удмуртских диалектах. Наш корпус в настоящий момент включает около 3,5 часов расшифрованных записей на трёх группах удмуртских диалектов (северные, срединные и южные), которые были собраны в ходе экспедиций 2014–2016 гг. Все тексты были расшифрованы с помощью носителей. Все морфологические особенности устных/диалектных текстов точно отражены в расшифровке, однако с фонетической точки зрения расшифровки были стандартизованы, чтобы облегчить морфологическую разметку и одновременный поиск в текстах на разных диалектах. Обработка данных в нашем проекте включает в себя выравнивание расшифровок со звуком с помощью ELAN и их автоматическую морфологическую разметку с помощью стандартного удмуртского анализатора. Мы рассматриваем автоматическую разметку как намного менее затратную альтернативу ручному глоссированию и проводим оценку качества и минусов такой разметки. В особенности мы рассматриваем вопрос о том, насколько сильно и как именно необходимо изменить стандартный анализатор, чтобы добиться достаточно качественной разметки устных/диалектных текстов. Корпус имеет веб-интерфейс, через который пользователи могут задавать поисковые запросы и прослушивать фрагменты аудио. Этот интерфейс будет открыт для общего доступа в 2018 году.

1 Introduction

This paper summarizes the preliminary results and the future directions of building a linguistic corpus of spoken texts from different Udmurt dialects.

Udmurt belongs to the Permic branch of the Uralic language family. Udmurt is spoken mainly in the Udmurt Republic, but also in the Republic of Tatarstan, the Republic of Bashkortostan, Perm Krai, Sverdlovsk Oblast and Kirov Oblast. According to the Russian Census of 2010, there are about 325,000 speakers of Udmurt.[1] The EGIDS level of Udmurt is 5, i.e. it is a *developing language*, which means that "[t]he language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable".[2] Standard Udmurt has an official orthography based on the Cyrillic script. This orthography is taught in schools and is familiar to most Udmurt speakers, regardless of their dialect.

As far as the existing corpora of Udmurt are concerned, we would like to mention three recent projects aiming at building linguistic databases for Udmurt, namely

[1] http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf

[2] https://www.ethnologue.com/language/udm

the Udmurt Corpus, the Beserman Corpus and the UraLUID database. The Udmurt Corpus[3] contains about 7.3 million tokens from mostly newspaper texts written in standard Udmurt. The Beserman Corpus[4] consists of transcribed oral texts in the Beserman dialect of Udmurt (currently, its size is about 65,000 tokens). The UraLUID database[5] encompasses both Udmurt texts collected in the 19^{th} century as well as text samples from two Udmurt blogs (the aim of this project was to create a database containing at least 4000 tokens of these two types of Udmurt texts, see Simon and Mus 2017).

Our goal is to process fieldwork recordings collected between 2014 and 2016. During these fieldwork trips, we collected contemporary spoken language material, hence, the corpus is aimed to represent the spoken varieties of Udmurt. Additionally, it is noteworthy that the recordings do not exemplify the standard Udmurt language but rather its dialects. In this way, our corpus is a further step in the corpus building efforts for Udmurt.

Needless to say, spoken texts are an irreplaceable source of valuable data for linguists. This is especially true in the case of endangered and under-documented languages like Udmurt. This highlights the importance of making fieldwork data open and reusable for the researchers, and possibly, for the native speaker community as well.

However, collecting and transcribing recordings is an extremely long and expensive process. The traditional approach to compiling spoken corpora includes aligning the transcription with the recording and then manually annotating the texts in Toolbox or FLEX (for Beserman Udmurt see Arkhangelskiy et al. 2017; for Enets see Khanina 2017). In this case, the annotation is by itself quite time-consuming and, in case of small-scale project like ours, could keep researchers and fieldwork linguists from processing and sharing their data. Hence, we took a different approach, namely to sound-align the recordings manually, but annotate them automatically. This approach to text processing has been advocated in several recent language documentation projects, e.g. in the Ustya Basin Russian project (Waldenfels et al., 2014) as well as in the Saami and Komi documentation project (Blokland et al., 2015), since it has two advantages: first, it is much less time-consuming, and second, the corpus built in this way can still be used for many research purposes.

In this paper, we discuss the workflow of our ongoing project and the obstacles we faced in the process. The paper is organized as follows. In Section 2, we present the Udmurt data used in the corpus with special reference to the metadata of the recordings. Then, in Section 3, we turn to the text processing steps, namely transcription and morphological annotation. We present the transcription used in the corpus and discuss several problematic cases in connection to it. As for the morphological annotation, we used a morphological analyzer originally developed for standard Udmurt. Given the fact that the corpus contains dialectal data, we had to make some adjustments to the analyzer, which is another novelty of the project presented in this paper. We specifically discuss the difficulties dialectal data pose with respect to morphological annotation (Section 4). Finally, we briefly describe the main features of the user interface of the corpus in Section 5.

[3] http://web-corpora.net/UdmurtCorpus/search/

[4] http://beserman.ru/corpus/search/

[5] http://www.nytud.hu/depts/tlp/uralic/dbases.html

Collection point	Dialect	Speaker(s)	Duration
Alnash district, Udmurtia	Southern	VE; EE	33:30
Alnash district, Udmurtia	Southern	LP; EE	29:02
Grakh district, Udmurtia	Southern	OK; IK; MK	53:54
Grakh district, Udmurtia	Southern	ESj	09:40
Grakh district, Udmurtia	Southern	VK; IK	27:09
Izhevsk, Udmurtia	Central	EL	08:28
Izhevsk, Udmurtia	Central	SSh	06:10
Izhevsk, Udmurtia	Northern	OS	04:53
Balezino district, Udmurtia	Northern	TS; TaS	09:15
Balezino district, Udmurtia	Northern	TS; ES	34:24
Kukmor district, Tatarstan	Southern Peripheral	EK; KK	22:33

Table 1: The Udmurt fieldwork recordings used in the corpus

2 Data

The Udmurt recordings used in this corpus were collected by Ekaterina Georgieva during three fieldwork trips conducted between 2014 and 2016 in the Republic of Udmurtia (and partly in the Republic of Tatarstan). All audio data were recorded in `.wav` format. The data represent different dialects of Udmurt, which we briefly overview below.

The dialects of the Udmurt language are divided into four main groups, namely the Northern, Central, Southern and Beserman dialect groups (Kelmakov, 1998, p. 41–44). The Southern dialect group is further divided into Southern dialects (spoken in the southern parts of the Republic of Udmurtia) and Southern Peripheral dialects (spoken in the Udmurt diasporas in Tatarstan, Bashkortostan, etc.).

Additionally, a division is made between "standard Udmurt" and its vernacular varieties (Edygarova, 2014). Standard Udmurt is used mainly in written form. As for the vernacular varieties of Udmurt, Edygarova (2014) distinguishes between local and cross-local vernacular varieties. The local varieties of Udmurt show features of a particular dialect, while in the cross-local varieties, features of mixed dialect and standard forms occur (Edygarova, 2014, p. 379).

Taking into account these facts, we assume that the fieldwork recordings used in the corpus represent the spoken varieties of Udmurt (that differ from its written variety, i.e. standard Udmurt) as well as exemplify certain dialectal features characteristic of the speakers' dialects. Nevertheless, they often contain standard forms alongside the dialectal ones. For example, in the texts, we find infinitives in both *-n* (dialectal) and *-ni̯* (standard), with the standard variant being slightly more frequent.

Now let us now take a closer look at the recordings used in our corpus. At present, the recordings collected during the fieldwork conducted in July and August 2014 are being processed. Below, we overview some basic metadata of these recordings, such as the place of recording, the dialect recorded, the speakers participating in the interviews and the duration of the recording, see Table 1. As can be seen from the table, the corpus is meant to cover (to a varying degree) the main dialects of Udmurt: Northern, Central, Southern and Southern Peripheral.

During the fieldwork trips, the semi-structured interview method was chosen. This format gave the speakers some freedom in the course of the interview. This was needed in order to ensure the right settings for a natural recording. It should be also emphasized that Udmurt was the only medium of the interviews. The interviews cover different genres, such as narratives, informal conversations between speakers, description of customs, etc.

Furthermore, the recordings fall into two groups regarding the number of speakers participating in the interview. In some of the recordings, only one native speaker was interviewed by the (non-native) fieldwork linguist, while in other recordings, the informant(s) was/were interviewed with the help of another native speaker. In the latter case, the result was a group conversation (featuring two or three native speakers and the linguist).

3 Text processing

In this section, we present the steps of processing the recordings. First, we discuss transcription and related issues. Then we proceed to the morphological analysis, for which we used the analyzer developed for standard Udmurt with some necessary adjustments. More specifically, we evaluate the applicability of this analyzer to dialectal data.

3.1 Transcription

The recordings were transcribed and time-aligned in ELAN[6]. ELAN allows to create complex annotations for audio and video files. The annotations are organized in a layered structure, in so-called *tiers*. The audio files were utterance/sentence-level time-aligned. Currently, the annotation of the recordings consists of two types of tiers: transcription and fieldwork notes. In each recording, there is a separate transcription and notes tier for each of the speakers (including the interviewer).

The first step of processing the audio files was the transcription. The transcription was carried out with the help of native speakers (in some cases, a speaker of the relevant dialect or one of the participants in the recording in question). Given the fact that we are dealing with spoken language recordings showing dialectal features, we had to make some principal decisions regarding the transcription we used. Let us mention a couple of problems in oral texts: assimilations, colloquial forms, unfinished words, hesitations, dialectal morphological features, etc. Below, we summarize the the decisions we have made regarding the transcription used in the corpus. Our goal was to apply these principles throughout, as consistency is one of the keys properties of corpus building.

First of all, it should be emphasized that we did not aim at providing a phonetic transcription. Hence, we chose the Cyrillic script used in the case of standard Udmurt, and not the Finno-Ugric Transcription System/Uralic Phonetic Alphabet[7] or the International Phonetic Alphabet. This made the transcribed texts consistent with the standard Udmurt texts and also facilitated the morphological analysis of the transcribed files. This choice was also motivated by the fact that our aim was to create a valuable and useful corpus with limited resources and within a relatively short period of time.

In contrast to other Cyrillic-based transcriptions of Udmurt oral texts (Kelmakov, 1998), we do not mark certain phonological processes, such as assimilations and non-standard stress patterns. For example, in this text collection, the regressive assim-

[6] https://tla.mpi.nl/tools/tla-tools/elan/

[7] http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2419a.pdf

ilation is transcribed, as in the verb forms like *tod-sko* (know-PRS.1SG) used in the Northern dialects, which is realized as *totsko* or *tocko*. Based on our data, it seems that the devoicing always applies, hence, we prefer to transcribe this verb form simply as *тодско* instead of *тотско* or *тоцко*. This transcription has the advantage that we do not need to add a stem allomorph *tot* of the verb *todi̮ni̮* (to know). Additionally, we also normalize certain colloquial forms, such as *бенэть* and *капказьын* to *бен ведь* and *капка азьын*, respectively. However, we do mark the actual realization of these colloquial forms in angle brackets (see in Table (2)). Hence, our decision in most cases is to adhere to the standard orthography with some exceptions that we discuss below.

The most important exception is the transcription of dialectal morphological features. Since our goal was to test whether the standard Udmurt morphological analyzer can deal with dialectal morphology, morphological features were always transcribed according to their actual realization in the recordings. Moreover, our data show that both the standard and the dialectal forms can be used in the same dialect or even by the same speaker, for example, as mentioned above, we find infinitives in both *-n* (dialectal form) and *-ni̮* (standard form). Hence, it was necessary to mark the actual realization of the infinitive suffix.

Furthermore, we did not normalize dialectal lexical items, such as *gid'* (pigsty) and *ti̮riśen* (since), corresponding to *gid* and *di̮ri̮śen*, respectively. In Section (4), we will discuss how these lexical items have to be processed morphologically.

The third major deviation from the standard Udmurt orthography concerns the transcription of compounds. We would like to stress that the standard orthography is very inconsistent with respect to compounds: some of them are written as one word, others are written with a hyphen, but most of them are written as two words. Moreover, the descriptive studies are also inconclusive of what exactly counts as a compound in Udmurt (Fejes, 2005). Hence, we decided to hyphenate all potential instances of compounding in the corpus.

Our transcription approach resembles the one adopted by (Waldenfels et al., 2014). On the one hand, we do not standardize the text on morphological level, so that the users can search for dialectal morphological features. On the other hand, we standardize the spelling to a certain extent to make it consistent throughout the corpus. This approach gives us two advantages. First, it minimizes the changes we have to make to the standard Udmurt morphological analyzer in order to apply it to our data. Second, it allows the users to search certain morphemes, words and lemmata in all dialects at once, while otherwise they would have to take into account all possible phonetic variants. Since the corpus has sound alignment, the users can still research dialectal phonetics by listening to the examples they find, regardless of the simplifications in the transcription.

Additionally, we chose to mark certain discourse and extralinguistic elements in the transcription.[8] This was motivated not only by the fact that we aimed at transcribing the recordings as precisely as possible, but also by the fact that we aimed at building a multi-purpose corpus. The conventions we adopted are listed in Table (2). It should be emphasized that we transcribed only those discourse elements that occurred inside the utterances. External noise, such as coughing, laughing, etc. that occurred between the utterances were not transcribed. A further convention of our transcription is that sentences start with lower case letters, and capitals are used only to mark proper names.

[8] We are grateful to Katalin Mády and Uwe Reichel for their suggestions regarding this part of the transcription.

Symbol	Description
<P>	unfilled pause
<B>	breathing
<S>	lipsmack
<L>	laugh
<N>	non-human noise
<H:xx>	filled pause, such as ыы, öö, aaa
<H:>	verbal realization of hesitation that cannot be captured with phonemes
<H>	hesitation, interruption after a word
<%>	non-understandable speech
<CF:xx>	colloquial form
xx<A>	aborted articulation of a word; written without a blank
xx<F>	foreign word (not used for Russian loanwords); written without a blank
.	finished utterance
no punctuation mark	unfinished utterance
,	used according to the intuition of the annotator
?	question
!	exclamation

Table 2: Transcription symbols

Below we provide an example of an utterance from the corpus. In this utterance, several discourse tags can be seen, as well as the compound *ǯe̮k-ki̮šet* (tablecloth) which is transcribed with a hyphen as *жӧк-кышет.*[9]

(1) 20140811; Balezino district; TS

кыше<A> жӧк-кш<A> жӧк-кышетъёсыз вань-а ӧвӧл-а шуса <B> тӥнь озь <CF:тнёзь>, кӧня штука, тӥнь та сюанлэн, мынам, жӧк-кышетэ.

3.2 Morphological analysis

The morphological analysis was carried out using an open-source rule-based morphological analyzer used previously for processing written texts in standard Udmurt[10]. The rules it uses consist of a dictionary, where the lexemes are listed together with their stems, part-of-speech tags and Russian translations, and a formalized description of the morphology.

Before processing the texts, we compiled a frequency list of word forms from our texts and manually added to the dictionary about 30 lexemes that were absent there, but frequent in the texts. This list included dialectal variants of several frequent words, such as the postposition *śajen* instead of *śamen* (in some way/language), the particle *bon*, several place names discussed in the texts, as well as the deictic adverb series with the stem *so-* (these correspond to the *o*-adverbs in standard Udmurt).

[9] The question clitic *a* is also hyphenated, as required by the standard orthography.

[10] https://github.com/timarkh/uniparser-grammar-udm/

After the morphological analysis, part of the ambiguity was removed with the help of a small set of Constraint Grammar rules (Bick and Didriksen, 2015). These rules have also been developed for standard Udmurt and cover only several prominent cases where the ambiguity can be eliminated with near-total accuracy.

Statistics regarding the quality of morphological analysis were calculated based on a pilot portion of texts, which contains about 2,500 words in the Southern dialect (30 minutes of sound). Initially, the proportion of the tokens that did not receive any analysis reached 13.9%. However, after performing the small dictionary enhancement described above, this proportion fell to 10.1%, which nearly equaled the proportion for the written texts in standard Udmurt (9.5%)[11]. In accordance with the Zipf's law, half of this improvement could have been achieved by adding only four new lexical entries.

The results of the morphological annotation in the case of Udmurt dialectal texts can be explained by two opposite trends. On the one hand, our texts have higher proportion of lexemes and features characteristic of spoken/dialectal Udmurt, which are not recognized by the analyzer. On the other, the speakers use more basic vocabulary without the complex neologisms one often encounters in standard Udmurt, and especially, in the Udmurt newspaper texts, which makes it easier for the analyzer. The ambiguity rate, about 1.4 analyses per analyzed token, was also approximately equal to that of the written texts. Although these results are preliminary and may be imprecise due to the size of the test corpus, they show that in general, it is possible to use the standard analyzer with minimal additions to process dialectal oral texts.

According to a very rough estimate, dialect and spoken features account for around 25% of the unanalyzed tokens. The rest consists of Russian loanwords (45%), proper names (15%) and standard Udmurt vocabulary handled incorrectly by the parser due to the incompleteness of the dictionary or the morphological description (also 15%). The 25% dialect-specific unanalyzed tokens will be discussed in more detail in the next section.

4 Problems with processing dialectal data

During text processing of dialectal data, several problems might arise. These concern the dialectal vocabulary and morphology, and to some extent, the orthography. Below, we summarize the main obstacles we had to face while processing Udmurt dialectal texts, and the solutions we came up with.

4.1 Dialectal vocabulary

One of the most obvious obstacles to processing dialectal data is the vocabulary, which may differ from that of the standard language. However, in reality, differences in vocabulary do not constitute a big problem. Dialectal nouns and verbs occur sporadically in standard Udmurt texts, they usually appear in standard Udmurt dictionaries, thus, many of them have been already included in the dictionary of the analyzer. For instance, when processing Southern texts, we added the noun *ajšet* (apron). Although this noun is marked as dialectal in Kirillova's (2008) dictionary, it could and should

[11] The figures for both corpora were calculated by the authors in November 2017. Currently, the proportion of unanalyzed tokens is less than 5% in both of them due to an enhancement of the analyzer dictionary, which was performed later.

have already been added to the dictionary of the analyzer because it appears 193 times in the corpus of standard Udmurt.

Dialectal variants of words are more problematic. Dialectal variants in our case included stems that were slightly different from their standard Udmurt counterparts, such as *ńil'* instead of *ńi̮l'* (four); *gid'* instead of *gid* (pigsty); and *nal/nnal* instead of *nunal* (day). There are two possible approaches regarding these cases. The first is to include them in the dictionary as separate entries. The second one is to list them in the existing entries as stem variants. We chose to adhere to the latter strategy. For example, the word *nal* is assigned the lemma *nunal* and can be found as one of its forms in the corpus. Nevertheless, it is still possible to find all words with the stem *nal-* because the segmentation of words into morphemes is stored in the analyzed files.

Furthermore, adverbs and words that belong to closed grammatical classes are especially problematic. We added to the dictionary words belonging to several closed classes, such as series of deictic adverbs in *so-*, particles (e.g. *gi̮nek* (only)) and postpositions (e.g. *ti̮riś* (since)). All of them have standard Udmurt counterparts, but are better analyzed as separate entries. Although these words constitute less than half of the lexemes we needed to add, they account for the majority of unanalyzed words in terms of token frequency.

4.2 Dialectal morphology

Dialectal morphology constitutes a double problem for the annotation. On the one hand, most words with dialectal suffixes simply will be left unanalyzed because they are absent from the morphological description of the analyzer. On the other hand, adding all of these suffixes may give rise to another problem, namely morphological ambiguity. This happens in cases when the dialectal morpheme homophonous with another morpheme used in standard Udmurt.

The solution that we applied in most of these cases was to add the lacking suffix to the grammatical description of the analyzer. In some of the cases, we also had to make changes to the dictionary, by adding stem allomorphs to certain lexemes. Finally, when we were dealing with the features that could increase ambiguity, we introduced additional constraints. Below, we list the suffixes we had to add in order to analyze the Southern texts, with special reference to the potential ambiguity.

- **Epenthetic *-j-* in an intervocalic position**. This concerns primarily the plural marker, which has the form *-os* after vowels and *-jos* after consonants in the standard language. In Southern texts, due to the epenthesis, the suffix *-jos* might be used in both cases. This variation never leads to ambiguity.

- **Dialectal variants of case markers**: *-iś* (*-i̮ś* in standard Udmurt) for the elative and *-ťi* instead of *-ti* for the prolative. Both do not lead to ambiguity.

- **Converb in *-ki̮*** (*-ku* in standard Udmurt). Does not lead to ambiguity.

- **Converb in *-sa*** instead of *-i̮sa*. Both variants exist in the standard language, but, just as with the infinitive marker, *-sa* is restricted to the non-*a*-stems. In Southern texts, *-sa* can be used with all stems. This variant does not lead to ambiguity.

- **Plural negative verbal form in *-ele*** (*-e* in standard Udmurt), used with non-*a*-stems. Does not lead to ambiguity.

- **Assimilated iterative suffix *-ća*** instead of *-ja*, following *-t*. This variant should not lead to ambiguity. Since this suffix is not fully productive, it occurs in both the dictionary and the grammar components of the analyzer. The combinations of verbal stems with the *-ja* suffix are stored in the dictionary. However, its quite frequent combination with the causative suffix, *-(e)t-ja*, is stored in the grammar component. Therefore, unlike other cases on this list, this dialectal morphological feature should be handled by both adding the suffix to the grammar and adding stem variants to the dictionary.

- **Colloquial verb forms.** A handful of frequent verb forms have widespread colloquial versions, e.g. *šuko/ško* instead of *šu-iśko* (say-PRS.1SG). We added these forms to the dictionary as separate entries with standard lemmata.

- **Infinitive in *-n/-i̮n*** instead of the standard Udmurt suffix *-ni̮/-i̮ni̮*. The *-n/-ni̮* variant attaches to the stems ending in *-a*, while *-i̮n/-i̮ni̮* attaches to all other stems. Since *-i̮n* is the standard Udmurt locative marker, this can lead to ambiguity in cases when there is a nominal stem homonymous with a verbal stem. There are quite few such pairs, but they include frequent words, such as *ul-i̮n* (live-INF vs. under-LOC), or *zor-i̮n* (rain:V-INF vs. rain:N-LOC). The situation could be partly amended by the Constraint Grammar rules, since there are not so many frequent contexts where an infinitive could appear. There is, however, a bigger problem with the *a*-stems because *-n* is the standard Udmurt nominalization suffix for these stems, cf. *uža-n* (work-INF vs. work-NMLZ). The derivation of *n*-nominalizations in Udmurt is fully productive and very frequent.

- **Non-standard morphophonology of *-śk-*.** In Udmurt, *-śk-* is used as a present tense marker as well as a passivizing/intransitivizing suffix. Descriptively, it has been observed that this suffix can have different morphological realizations in the Udmurt dialects (Kelmakov, 1998, p. 147–150). We will explore the consequences of this dialectal variation with respect to the processing of the oral texts.

  In the presence of the suffix *-śk-*, the preceding *-d/-t* is elided regardless of its morphological status. This can give rise to several problems:

  1. First, we have to add stem variants for all verbs ending in *-d/-t*: *tod+śko* might be realized as *to-śko* (know-PRS.1SG).
  2. Second, the biggest problem stems from the fact that passive can be – and frequently is – preceded by the causative suffix *-(e)t-*. The causative suffix might be elided, too, which leads to ambiguity. For instance, the verb form *todma-śk-i-di̮* can represent two different cases: (i) recognize-PASS-PST-2PL or (ii) recognize-CAUS:PASS-PST-2PL, in the latter of which the causative suffix has been elided. This kind of ambiguity can be partly resolved by rules if there is transitivity information in the dictionary, since intransitive verbs can only have impersonal 3SG passive forms. We have manually added transitivity information to the dictionary. However, in the case of transitive stems or 3SG forms, the ambiguity will remain. Currently, the analyzer does not annotate these transitive verb forms as containing a zero causative to avoid significant amount of ambiguity that would follow. Consequently, the form *todmaśkidi̮*, which actually contains a causative suffix in our corpus, is treated incorrectly by the analyzer.

> It should be noted that the same phonological process sometimes works even in the cases where the segment *-śk-* is part of the stem and not a suffix, as with the dialectal verb *uśki̮ni̮* (look), standard form of which is *ućki̮ni̮* . Therefore, we have to locate such verbs in the dictionary and add stem variants for them as well.

Below, we summarize our preliminary notes regarding the morphological peculiarities of the Northern texts, which we have started processing. The Northern dialects share some of the non-standard features described above (epenthetic *-j*, infinitive in *-n*, shortened verbal forms, converbs in *-sa*, consonant assimilation before the suffix *-ja*). Here we only touch upon some of the features absent in Southern texts.

- **A series of personal-local case markers, e.g. *-ńe*.** Personal-local suffixes are a combination of the marker *-ń-* with one of the spatial cases, and convey the meaning 'at/to/from/through one's place' (Teplyashina, 1981). They do not lead to ambiguity.

- **Frequentative suffix *-i̮ll-*** instead of *-i̮l-*. Does not lead to ambiguity.

- **Limitative converb in *-ććoź*** instead of *-toź*. Does not lead to ambiguity.

- **Non-standard forms of *-śk-*.** The suffix *-śk-*, as it was stated above, has several dialectal variants. In the Northern texts, it often appears as *-sk-* and devoices the preceding consonant. Unlike its Southern counterpart, the Northern variant does not pose a problem to the analysis if the transcription is somewhat standardized, e.g. when *todsko* (know-PRS.1SG) is spelled as *тодско* rather than *тоцко* or *томско*, cf. Section 3.

- **Non-standard forms of personal and reflexive pronouns.** Certain case forms of personal and, especially, reflexive pronouns used in the Northern texts are non-standard, such as *mil'emesti̮* (we.ACC) with the standard form being *mil'emi̮z*. Since these forms of the pronouns are morphologically irregular, all of them need to be stored in the dictionary. We added the dialectal form to the dictionary as well, which did not lead to any ambiguity.

4.3 Orthography

Apart from lexical and grammatical challenges, there are also challenges related to orthography. The cases that involve non-standard tokenization are especially problematic. A large share of these cases is represented by compounds which consist of two nominal stems, or by complex numerals. Nominal compounds are frequent in Udmurt, and all but a relatively small number of them are written as two separate words according to the standard orthography. When we transcribed compounds, we wrote them with a hyphen regardless of their lexicalization degree and stress pattern. We took a similar approach in the case of numerals like *ki̮ź odig* (twenty-one), which are always written as separate words in standard Udmurt, but are hyphenated in the corpus. Although hyphenation of all compounds and numerals diverges from the official orthography and does not capture the difference between lexicalized and non-lexicalized compounds, it allows the users to easily find all such instances and decide for themselves.

5 User interface

Our corpus will be accessible for the linguistic community through an online web-interface.[12] We use the open source *tsakorpus* platform, which was developed by Timofey Arkhangelskiy and is available under an MIT license[13]. Each ELAN file is passed through morphological analyzer and Constraint Grammar disambiguator. The analyzed file is stored in JSON format, which is then uploaded to an Elasticsearch database. The functions of the web interface include: search by word, lemma, Russian translation, grammatical tags and their combinations; search with regular expressions; search for specific allomorphs of a morpheme; multi-word search; and selecting texts based on metadata values. Users are able to see or download sentences that contain the requested words, listen to the sound aligned with the sentence, get frequency lists of words, and chart word distribution e.g. by dialect. Source Cyrillic orthography and automatically transliterated Uralic Phonetic Alphabet representation are supported. The interface is available in English and Russian, but other languages can be easily added to this list. Currently, we are testing the interface. We intend to make it publicly available in 2018.

6 Conclusions

In this paper, we presented an ongoing project in the course of which we are going to develop a sound-aligned corpus of spoken texts in Udmurt dialects. Our workflow consists of transcribing and aligning the texts in ELAN using (mainly) standard orthography, automatic morphological analysis with partial rule-based disambiguation, and publishing the recordings online using a publicly available web interface. Standardization of orthography enables cross-dialectal search and facilitates automatic processing of the texts. We demonstrated that a morphological analyzer for the standard language works sufficiently well for our data, and can be relatively easily adjusted for the annotation to be comparable in quality to that of the standard Udmurt texts. We also outlined the obstacles we faced in the process. Some of them are caused by the inconsistencies of the standard orthography, while others stem from the ambiguity introduced by dialectal variants of morphemes. We believe that the same workflow can be applied by other researchers who have dialectal recordings at hand to efficiently produce valuable dialectal corpora with relatively small investments of time or resources.

Acknowledgments

We are immensely indebted to the native speakers who participated in the fieldworks and helped our work.

References

Timofey Arkhangelskiy, Natalia Serdobolskaya, and Maria Usacheva. 2017. Corpus-oriented lexicographic database for Beserman Udmurt. *Acta Linguistica Academica* (64):397–415.

[12] For the moment, we do not intend to make the source ELAN files available to the general public.

[13] https://bitbucket.org/tsakorpus/tsakonian_corpus_platform/overview

Eckhard Bick and Tino Didriksen. 2015. CG-3 — Beyond Classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. Linköping University Electronic Press, Linköpings universitet, 109, pages 31–39.

Rogier Blokland, Marina Fedina, Ciprian-Virgil Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2015. Language Documentation meets Language Technology. In *Septentrio Conference Series*. pages 8–18.

Svetlana Edygarova. 2014. The varieties of the modern Udmurt language. *Finnish-Ugrische Forschungen* (62):376–398.

László Fejes. 2005. *Összetett szavak finnugor nyelvekben*. Ph.D. thesis, Eötvös Loránd Tudományegyetem.

Valentin V. Kelmakov. 1998. *Kratkij kurs udmurtskoj dialektologii*. Izdatelstvo Udmurtskogo universiteta, Izhevsk.

Olesya Khanina. 2017. Digital resources for Enets. *Acta Linguistica Academica* (64):417–433.

L. E. Kirillova, editor. 2008. *Udmurtsko-russkij slovar*. Izhevsk.

Eszter Simon and Nikolett Mus. 2017. Languages under the influence: Building a database of Uralic languages. In *Proceedings of the Third International Workshop for Computational Linguistics of Uralic Languages*. pages 10–24.

Tamara I. Teplyashina. 1981. O novyx udmurtskix padezhax. In *Congressus Quintus Internationalis Fenno-Ugristarum, Turku*. 20 (27), Pars VI, pages 285–292.

Ruprecht von Waldenfels, Michael Daniel, and Nina Dobrushina. 2014. Why standard orthography? Building the Ustya River Basin corpus, an online corpus of a Russian dialect. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*. 13, pages 720–728.