WHY STANDARD ORTHOGRAPHY? BUILDING THE USTYA RIVER BASIN CORPUS, AN ONLINE CORPUS OF A RUSSIAN DIALECT[1]

Waldenfels R. von (ruprecht.waldenfels@gmail.com)
Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warsaw, Poland

Daniel M. (misha.daniel@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Dobrushina N. (nina.dobrushina@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

The paper describes a corpus of dialectal Russian speech under development. The corpus relies on interviews conducted by a joint Swiss-Russian team in the summer of 2013 in a small cluster of North Russian villages with the goal of studying the local dialect from a sociolinguistic and dialectological perspective.
The interviews are transcribed into standard Russian and thus do not involve a detailed phonetic representation. The text is then lemmatized and grammatically annotated with standard tools and fed into a corpus. The corpus can be queried via a web-based interface which provides the user with access to the original sound recordings on a per-utterance level. This design, the paper argues, allows for a rapid development of the corpus without a major loss in usability, since the audio data are readily available. Future plans include more field trips as well as a more convenient interface providing, among other features, for user correction of the transcription.

Keywords: Russian; dialectology; corpus linguistics

1. Introduction

The abundance of linguistic data in corpora readily available over the internet has greatly changed the work of linguists in many subdisciplines. However, this

[1] This study (research grant No 14-05-0034) was supported by The National Research University–Higher School of Economics' Academic Fund Program in 2014. We thank the Slavic department of the University of Bern for hosting the corpus server.

development is probably least advanced in respect to the study of spoken language, and specifically corpora of dialectal or nonstandard speech. With some exceptions, corpus data for the study of dialects is still difficult to find on the web. Obviously, this is due to the specific challenges that spoken language poses in respect not only to collecting the data, but also to transcribing them to some written form, and making them available to the scientific community on the internet.

The present article describes a Russian dialect corpus project that aims to alleviate these problems by first transcribing the data into standard Russian, automatically annotating the corpus with standard tools, and making the result available on the internet together with aligned sound segments. This approach has a number of advantages, as well as some weaknesses.

The article is structured as follows. First we give a short overview of selected dialect corpora of Russian and other Slavic languages. We then introduce the place and circumstances of the Ustja River Basin Corpus data collection, before introducing the principles of and rationale for the transcription of the dialectal data in standard Russian. We then give an example analysis using the corpus data. Finally, we sketch planned further developments.

2. Corpora of Dialectal and Other Spoken Variants

Most Slavic dialect corpora make the data available in some sort of transcription that is usually situated between a faithful phonetic and a standard language representation; this is true, for example, for the Polish internet resource "Dialekty i gwary polskie" http://www.dialektologia.uw.edu.pl, for the dialectal and spoken data in the Czech National Corpus, the Slovak National Corpus, and others. The GOS corpus of spoken Slovene (http://www.korpus-gos.net/) offers both a standard and a more phonetically detailed transcription in two annotation layers. The Russian National Corpus contains a dialectal subcorpus which is mostly transcribed very near to the standard and is not very large (under 200,000 tokens), but offers the RNC's flexible search engine (powered by Yandex) for making sophisticated queries. The spoken subcorpus of the RNC is also orthography-oriented and uses a 'shallow' transcription, showing only pauses but not other discourse phenomena. The Saratov Dialect Corpus (http://www.sarteorlingv.narod.ru/projects.htm) offers detailed annotation as well as audio files.

Most of the above corpora do not include audio material. In some cases, as in the before mentioned "Dialekty i gwary polskie", the interviews are made available as full audio files alongside their transcription. In others, such as in the case of the German RuReg project (http://rureg.hs-bochum.de/), audio files of paragraph length are aligned to their transcription; the above mentioned Saratov Dialect Corpus seems to take a similar strategy (the corpus though was unavailable at the time of writing and submitting this paper).

Access to the original recording is especially important for dialect corpora, since transcription inevitably involves a loss in information that might be crucial for the analysis. We feel that for any extensive corpus based work on dialects or spoken data, it is crucial to have access to the audio files aligned to the transcribed text.

3. The Language of the Ustya River Basin: Collection and Corpus Composition

3.1. Data Collection and Transcription

The data of the Ustya Corpus was collected in the summer of 2013 by a joint group of Russian and Swiss students from the National Research University Higher School of Economics and the Slavic department of the University of Bern. The project is supported by HSE and carried out within the framework of a research and teaching cooperation between the two institutions.

The field team had its base in the village of Mikhalevskaya (locally known as Pushkino), in the Ustyan district of the Arkhangelsk Oblast, on the border with Vologda Oblast[2], but sometimes travelled a bit to the neighbouring villages. The students interviewed the villagers, asking people to tell them about their lives and other stories, and partly transcribed them on site, with more material transcribed later.

The dialects in Arkhangelsk Oblast have been object to vast research activities throughout the last century. This is why from the very beginning we intended to focus on studying variation rather than more preserved idiolects, modeling post-dialectal continuum rather than only the speech of oldest villagers (who are, as it happens today, mostly women). With some exceptions (e.g. Kochetov 2006, Krasovitsky 2013), the sociolinguistic dimension, mesolects and dialect attrition are still a rare topic in Russian dialectology.

The spoken data was transcribed using two programs: ELAN and Praat (since the formats are easily converted, transcribers were free to use either). In this type of transcription, each utterance in the audio file is marked and transcribed in one of several tiers. Informants are given separate tiers, with additional tiers for the interviewers, other speakers, and comments. The recorded speech is transcribed exclusively in standard Russian, with some provisions for marking unintelligible segments. The data is then stored in the ELAN XML format and processed further in an automatized procedure to add lemmatization and pos-tagging, and make it available over a web-based corpus interface.

Altogether, we collected some 40 hours of conversation. As of April 2014, 20 hours have been transcribed, comprising a corpus of around 215,000 tokens, of which about 180,000 tokens are informants' speech.

3.2. Why Standard Russian?

As indicated above, standard language is used, rather than a phonetic transcription as it is customary in most traditional publications. This means losing a lot of detail in comparison. Cf. the next texts from (Pozharitskaja 2005: 220, Vologda dialect), in both original transcription and the standard representation:

---

[2] We warmly thank our hosts Nikolaj Pushkin and Svetlana Pushkina for all their help in organizing our life, and work, and other practicalities.

[оп свojóй жы́з’н’е это хоц’ýгу погувур’и́т’ / жы́с’ мojá прошлá н’е óц’ен’ ва́жно / жы́ла ф-так’и́jо го́ды т’ежóлыjо / д’ит’éй у м’ен’á бы́ло п’ét’еро / подн’алá ja д’ит’éй до войны́ / фторóй сын пог’и́п на войн’é]

“Об своей жизни это хочу поговорить. Жизнь моя прошла не очень важно, жила в такие годы тяжелые. Детей у меня было пятеро, подняла я детей до войны, второй сын погиб на войне.”

Projects that adopt standardization approach are e.g., the Freiburg English Dialect Corpus (http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/), the ALCORP corpus of allemanic dialects of German, or the Nordic Dialect corpus (http://www.tekstlab.uio.no/nota/scandiasyn/). In many cases, the use of a standard language transcription is justified by focusing on (morpho)syntactic phenomena, which obviously does not assign high priority to phonetic detail (such is, for Russian, the dialectal subcorpus of the RNC). But in our case, standardization goes farther than in many other cases. At first glance, the corpus thus transcribed has nothing dialectal in it at all. What is its rationale?

In a nutshell, the standard orthography is justified by the fact that orthography is nothing but a key to the audiofiles with which the corpus is aligned. Corpora with standard transcription aligned with audio have been successfully used for phonetically oriented studies, e.g. by Streck (2012). The use of standard transcription means that we relegate a detailed phonetic analysis to a later stage (and probably to other experts). In general, this approach has the following advantages:

1. **Transcription into standard language can be done quickly.** When transcribing into standard orthography, there is no need to make difficult phonetic decisions concerning the data that involves repeatedly listening to the audio excerpt, comparing it to other segments of the main speaker, identifying phonetic variants—above all, a high expertise in dialectal phonetics. Note that even expert dialectologists may diverge on details of what they actually hear. While doing a standard transcription, it is sufficient to understand the text and identify the closest equivalent in the standard (however, below we discuss problems of defining what such an equivalent may be). This can be done much faster than phonetic transcription, and it demands by far less expertise.

2. **Transcription into the standard language effectively solves the problem of normalization and standardization.** Phonetic transcription systems used in different dialect corpora do not always coincide even for the same language, since the transcriber needs to balance readability and faithfulness to the sound shape, as well as decide what level of phonetic accuracy he or she wants to achieve for a given purpose. This is very difficult to do in a consistent way between transcribers, let alone different dialectal corpus projects. The standard language, in contrast, is well known to the transcribers and, in most cases, different transcribers will choose the same representation for a given dialect utterance without much doubt or need for consultation. This greatly reduces both systemic and non-systemic variation in the transcription of the same text by different transcribers.

3. **Transcription into the standard language makes the use of standard automatic annotation tools possible.** The automatic annotation of non-standard speech

is a difficult problem; see for example the system described in Wieczorek (2011) in the context of dialectal studies of Polish. Since we transcribe into standard Russian, however, we were able to use standard tools such as the TreeTagger (Schmid 1995) for the lemmatization and grammatical annotation of our data.

4. **Transcription into the standard language makes the data easily readable by non-linguist users.** In principle, the collected material may represent a cultural interest for a public broader than the dialectologists, including representatives of the local community, in the local towns if not in the village. Standard representation is much more suitable for the use of the interested 'lay' public, even if they are themselves speakers of the dialect (the combination of these two properties is however rare).

5. **Loss of phonetic data in transcription is made up for by aligning the transcription with the original audio.** Source audio information remains fully available to the user as the original audio is sentence-aligned to the transcription. Every user may make his or her own decision on what has been said, and how, and use examples from the corpus applying his or her own approach to dialectal transcription. For an expert, this is by far better than having to trust the transcriber.

4. An Overview of the Problems Related to the Transcription into Standard Language

The basic aim of the transcription is thus to provide the user with an easy access to the sound recordings. We do so by providing query interface based on standard automatic annotation tools. For this, the transcriber has to 'translate' or 'transpose' the dialectal text into standard language. This is far from being trivial, since many dialectal items on all linguistic levels do not have one-to-one correspondences in the standard. We will show several examples of such transpositions. Note that the transcription below (bracketed) is not intended to show the exact phonetic shape but to highlight the differences from the norm.

- If a dialectal word is different from the standard in a regular phonetic way, the standard variant is chosen: [заготовл'ёл'и] — *заготовляли*, [пр'ишóу] — *пришёл*, [пойи́с'] — *поесть*.
- Note that this includes cases where the standard correspondence may not be used in the sense in which it is used in the text; in such cases, we still use the standard word: [мои́ коро́вы шóбы не **рыч'ёл'и**] — *мои коровы чтобы не рычали* <чтобы мои коровы **не мычали**>.
- If a dialectal word is different from the standard in (the form of) the inflectional affix it takes, the standard variant is chosen: [р'евл'ý] — *реву*, [пок'исл'áje] — *покислее*, [мол'и́л'ис'е] — *молились*.
- A very frequently occurring phenomenon are postpositional particles, which correspond to the standard *-то*, but change their form depending on (the form of) the preceding word: [час'т'-ту] — *часть-то*, [тел'áта-та] — *телята-то*, [дом-от] — *дом-то*.

- If a dialectal word is different from the standard in the derivational affix it contains, the dialectal variant is chosen: [здал одно́ **кос'т'jо́**] — *сдал одно костьё* <такая худая была корова>, [бо́л'ше **н'е́как** бы́ло уч'и́ц'ц'е] — *больше некак было учиться* <больше никак не удавалось учиться>
- But if the difference in the derivational affix or in the root is trivial, the standard variant is used: [кварт'е́ра] — *квартира*, [робо́та] — *работа*, [топ'е́р'] — *теперь*.

The word 'trivial' here is not further formalized and relies very much on the intuition of the transcriber. All this boils down to the principle that, to make standard taggers applicable to the texts, we make as much phonetic adaptation as possible, reasonable and practicable without losing lexically, morphologically and syntactically relevant information—but not purely phonetic information which may be retrieved from the aligned audio. Thus, we certainly do not meddle with *meanings* and do not do *translation* of dialect texts into standards; and we do not force non-standard use of specific morphological forms into the rules of the standard language. Of course, this leaves us with many difficult cases when the transcriber has to make a decision that can hardly be formalized or generalized. For example, many dialects (including Ustja) use a standard word with different meaning—[немо́гу] corresponds to standard *не могу*, both phonetically (with an accent shift) and morphologically, but means 'to be ill'. One of our transcribers suggested that this verb should be written in the dialect as one word with the negation, as this combination has been clearly lexicalized and forms a new lexical item. We leave such subtle decisions to the transcribers, and assume that no exhaustive set of rules is possible or practicable. This may lead to some variation in transcription, but on the other hand will greatly facilitate transcribers' work.

As it is often the case in corpus building, we thus aim for a pragmatically sound, rather than for an ideal corpus, since going for such an ideal corpus would be certainly much more costly, perhaps in the end not feasible and quite conceivably unnecessary—i.e., a waste of resources. It is for this reason that in cases where several alternative solutions can be argued for and seem nearly equally plausible, we accept some variation between transcribers rather than try to achieve a completely consistent transcription and leave to the corpus user the task of dealing with potential divergences. This makes transcribers' task much more manageable, and may in fact in some cases lead to empirical solutions more robust than any theory-based inductive rule. Practice of corpus usage shows that many such theory based rules are non-intuitive anyway, and corpus users most often follow their intuitions rather than corpus descriptions. Note again that the fact that the raw data in form of audio segments are readily available means that any transcription can be checked by the user, making the transcription much less important than in traditional dialect texts.

4.1. Lemmatization, POS-Tagging and Inclusion into CWB

After transcription, the dialect text is lemmatized, tagged and imported to the Open IMS Corpus Workbench (CWB) corpus manager (http://cwb.sourceforge.net/) by a number of scripts, i.e., fully automatically once the transcribed file is entered into

the corpus repository. We use the TreeTagger (Schmid 1995) with a parameter file trained on the Multext-East tagset http://corpus.leeds.ac.uk/mocky/)

5. Using the Corpus

In this preliminary version, the data is accessible via a somewhat technical online interface that allows full CQP syntax as well as a simplified version of CQP (see Figure 1). Query results provide access to audio segments on an utterance level, so that the researcher has access to all properties that are lost under standardization—that is, to all phonetic, intonational or morphological details that are relevant for the research question the user is interested in. Expert users can check the correctness of the transcription (and, in the near future, will be able to add their comments to the texts of the corpus. Sample query results are shown in Figure 2.

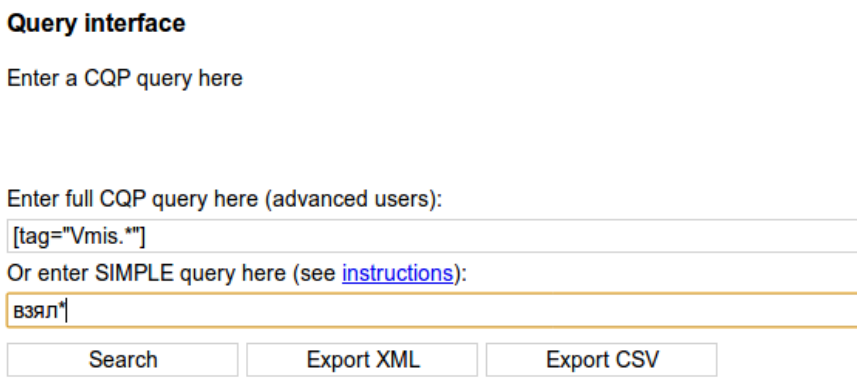

Fig. 1. Query interface

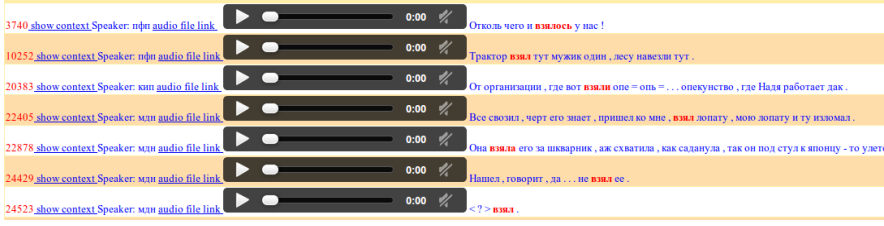

Fig. 2. Query results

The corpus interface provides access to lemmatization and grammatical information as parts of a query. For example, Seržant (2014) has used the corpus to investigate

the partitive genitive in Northern Russian. One way to find such partitive genitives is to use the query [tag="N.m.sg.*"] to look for all masculine genitive singular nouns in the standard transcription (the tagging uses the MULTEXT-EAST tagset for Russian, see http://corpus.leeds.ac.uk/mocky/msd-ru.html). The user then examines the audio files to decide which ending was used.

As a second example, consider the realization of /a/ as [e] between palatalized consonants that is found in many Russian dialects (Galinskaja 2005). To obtain all words where this change could have taken place (the envelope of variation), we look for all occurrences of "я" followed by one or more consonants, followed by a jotified vowel or soft sign in the standard transcription. We do so by using the regular expression query:

". *я([ртзпсдфгхклвбнм]*[ьеяиюё]|[цчжщй]). *"[3].

Using this query, we can quickly obtain the list of relevant word forms, which in the interviews with our oldest informant is as follows:

(30x) пятьдесят, (18x) пять, (13x) опять, (8x) объяснить, (6x) всякие, (5x) гуляет, прядет, (4x) прядешь, прясть, пятеро, пятисотку, (3x) девять, грязь, копятся, накопятся, память, представляете, пряли, тысяча, тысячи, (2x) Октябрьском, гулять, добавляли, доярки, завяжут, месяц, отправляли, отправляют, прялке, пяти, пятисотка, сеяли, сеялки, сплавляли, телятник, тысяч, тысячу, ячмень, (1x) Деревянние, Настоящая, Объяснишь, Отправляй, Прядешь, блядь, блять, вянет, гоняемся, граблями, грязе, грязи, гуляли, десять, дядя, завяжешь, заготовляли, заготовляют, запрягешь, заставляли, заставляют, кашляет, мягкие, напряде, напрядено, напрядет, настоящая, начислять, объявили, объясню, оставляют, отгоняли, поняли, пряди, прядёт, прялки, пятдесят, пятей, пятим, пятисотке, сеятся, справлялись, справляться, телятницей, телятся, трясти, тутошняя, тяжело, удобряют, яки, яме, ясли, яслях, ячмене

All the utterances with these words are displayed in the result window (cf. figure 2) and the users can examine and categorize these word forms in respect to the realization of /a/ as [e]. Preliminary analyses show that this change seems to be preserved only with older speakers; there is some evidence that it may be most resilient as a morphophonologically conditioned alternation in the language of younger speakers. But more research is necessary.

In sum, we see that for some questions, the representation in the standard transcription may be quite adequate (e.g., for some syntactic issues). For other questions, researchers need to do their own analysis of the audio data. In essence, thus, the painstaking work of a deep phonetic or other analysis is not performed in the transcription phase, as it is traditionally done, but at a later stage, and by the expert user him- or herself.

While this may seem as a drawback, note that since the analysis is done in the context of a specific research question, the accuracy of the analysis may actually be higher than in the context of a general-purpose phonetic transcription[4].

---

[3] Note that this expression is only an illustrative example that requires some later filtering, and does not cover /a/ before /j/, which is mostly lost in intervocalic surroundings.

[4] Of course, ultimately, it would be ideal if such annotations could be fed back into the corpus.

6. Further Plans

In 2014, the second field trip to Mikhalevskaya is planned, to make more recordings. The time in the field is also used as an opportunity for workshops on dialectal phonetics, morphology and syntax, for the students to exchange their ideas. As for the corpus, at this moment the interface is not yet publicly available on the web; access is granted only on an individual level. We are working on a more advanced interface that is more accessible to users that are not acquainted with the query language. Moreover, we want to enable users to correct mistakes in the transcription and in this way crowdsource some of the transcription work.

References

1. *Kochetov, Alexander* (2006). The role of social factors in the dynamics of sound change: A case study of a Russian dialect. Language Variation and Change, 18(01), 99–119.
2. *Krasovitsky, Alexander* (2013). Artikul'atsionnyj sdvig i razvitie nejtralizatsii glasnyh. In: British contributions for the XV Congress of Slavists, Minsk.
3. *Galinskaja, E. A.* (2005). Izmenenie [a] v [e] v istorii russkih dialektov. Vestnik Moskovskogo universiteta. Ser. 9, Filologija, (4), 42–54.
4. *Pozharitskaja, Sofia* (2005). Russkaja dialektologija. Moscow: Akademproekt.
5. *Schmid, Helmut* (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
6. *Seržant, Ilja* (2014) Independent partitive genitive in North Russian. In: Seržant, I. A. and B. Wiemer (eds.), Contemporary approaches to dialectology: The area of North, Northwest Russian and Belarusian vernaculars // Sovremennyje metody v dialektologii. Areal severnyh, severo-zapadnyh russkih i belorusskih govorov. Slavica Bergensia 13.
7. *Streck, Tobias* (2012) Phonologischer Wandel im Konsonantismus der alemannischen Dialekte Baden-Württembergs. Sprachatlasvergleich, Spontansprache und dialektometrische Studien. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik—Beihefte, Band 148).
8. *Wieczorek, Aleksandra* (2011). Słownictwo polskiej gwary kresowej na przykładzie Maćkowiec na Podolu. Charakterystyka funkcjonalna. Ph. D. Thesis, University of Warsaw.