Corpus of usage examples: What is it good for?

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

Abstract

Lexicography and corpus studies of grammar have a long history of fruitful interaction. For the most part, however, this has been a one-way relationship. Lexicographers have extensively used corpora to identify previously undetected word senses or find natural usage examples; using lexicographic materials when conducting data-driven investigations of grammar, on the other hand, is hardly commonplace. In this paper, I present a Beserman Udmurt corpus made out of “artificial” dictionary examples. I argue that, although such a corpus can not be used for certain kinds of corpus-based research, it is nevertheless a very useful tool for writing a reference grammar of a language. This is particularly important in the case of underresourced endangered varieties, which Beserman is, because of the scarcity of available corpus data. The paper describes the process of developing the Beserman usage example corpus, explores its differences compared to traditional text corpora, and discusses how those can be beneficial for grammar research.

1 Introduction

Following a widely acknowledged idea that the language above the phonological level can be roughly split into lexicon and grammar, language documentation is divided into two interconnected subfields, lexicography and grammar studies. Corpora have been used since their advent in both these domains; one of the first studies of English based on the Brown corpus (Francis and Kučera, 1982) contained frequency analysis for both words and parts of speech. It was recognized early on in the history of corpora that they are an excellent source of usage examples for both dictionaries and reference grammars. This gave rise to a data-driven approach to language documentation, which prescribes using only “real” examples taken from corpora in descriptive work (Francis, 1993). Corpora have become standard providers of dictionary examples, which can even be searched for automatically (Kilgarriff et al., 2008). However, this approach can hardly be applied to underresourced endangered languages because it relies on large representative corpora, which are normally unavailable for such languages. Grammatical studies based on small spoken corpora with minimal help of additional elicitation are often possible and have been conducted, e.g. by Khanina (2017) for Enets or by Klumpp (2005) for Kamas. The same cannot be said about lexicography. While even small corpora are a valuable source of usage examples for dictionaries, the lexicographer has to elicit examples for non-frequent or obsolete entries or word senses. These examples usually stay in the dictionary and are not used for any non-lexicographic research.

I argue that such elicited usage examples can be turned into a corpus, which can actually prove to be helpful for a variety of grammar studies, especially in the absence of large “natural” corpora. An example of a feature that cannot be available in traditional corpora, but can appear in elicited examples, is negative linguistic material. The paper describes a corpus of usage examples I developed for the documentation of Beserman Udmurt. It presents the data and the methods I used to develop the corpus. After that, its frequency characteristics are compared to those of traditional spoken corpora. Finally, its benefits and disadvantages for certain kinds of grammar research are discussed.

2 The data

Beserman is classified by Kelmakov (1998) as one of the dialects of Udmurt (Uralic > Permic). It is spoken by approximately 2200 people who belong to the Beserman ethnic group, mostly in North-

Western Udmurtia, Russia. Having no official orthography, it remains almost entirely spoken. This, together with the fact that transmission to children has virtually stopped, makes it an endangered variety. It differs significantly from Standard Udmurt in phonology, lexicon and grammar, which justifies the need for separate dictionaries and grammatical descriptions. The two existing grammatical descriptions, by Teplyashina (1970) and Lyukina (2008), deal primarily with phonology and morphemics, leaving out grammatical semantics and syntax.

Beserman is the object of an ongoing documentation project, whose goal is to produce a dictionary and a reference grammar, accompanied by a spoken corpus. The dictionary, which nears completion, contains approximately 5500 entries, most of which have elaborate descriptions of word senses and phraseology, and illustrated with examples. The annotated texts currently count about 103,000 words, which are split into two collections, sound-aligned corpus (labeled further SpC) and not sound-aligned corpus.[1]

3 Development of the corpus

The Beserman usage example corpus (labeled further ExC) currently contains about 82,000 words in 14,000 sentences. Each usage example is aligned with its translation into Russian[2] and comments. The comments tier contains the information about the number of speakers the example was verified with, and remarks on possible meanings or context. The main source of examples for the corpus was the dictionary, however a small part (5% of the corpus) comes from grammatical questionnaires. The dictionary examples were collected using several different techniques. First, there are translations from Russian (often modified by the speakers in order to be more informative and better describe their culture). Second, many examples were generated by the speakers themselves, who were trained to do so by the linguists. Finally, some of the examples were produced by the linguists (none of who is a native Beserman speaker) and subsequently checked and corrected by the speakers. The development of the corpus included filtering and converting the source data, performing automatic morphological annotation, and indexing in a corpus platform with a web interface.

3.1 Data preparation

The dictionary is stored in a database created by the TLex dictionary editor, which allows export of the data to XML. The XML version of the dictionary was used as the source file for filtering and conversion.

The filtering included three steps. First, the Beserman dictionary contains examples taken from the spoken corpus. Such examples had to be filtered out to avoid duplicate representation of the same data in the two corpora. Identifying such examples might be challenging because they are not consistently marked in the dictionary database, and because some of them were slightly changed (e.g. punctuation was added or personal data was removed). To find usage examples that come from the spoken corpus, each of them was compared to the sentences of the corpus. Before the comparison, all sentences were transformed to lowercase, and all whitespaces were removed. If no exact matches were found, the Levenshtein distance to the corpus sentences of comparable length was measured. The cut-off parameters were obtained empirically to achieve a balance between precision and recall. If a corpus sentence with a distance of $max(2, len/8)$ was found, where len is the length of the example in question, the example was discarded. Additionally, the “comment” field was checked. The example was discarded if it contained an explicit indication that it had been taken from a text.

Second part of the filtering involved deduplication of examples. Deduplication was necessary because some usage examples appeared in multiple entries. In this case, their versions could have minor differences as well, e.g. because of typo corrections that were made in only one of them. Two sentences were considered identical if both their lowercase versions and their lowercase translations had the Damerau–Levenshtein distance not greater than $len/10$.

Finally, the dictionary is a work in progress. As such, it contains a number of examples that

[1] Both the dictionary in its current version and the corpora are publicly accessible at http://beserman.ru and http://multimedia-corpus.beserman.ru/search.

[2] Since the native speakers, including heritage speakers, are the primary target audience of the dictionary, and they are bilingual in Russian, the examples are translated into Russian. Translation of the dictionary into English has started recently, but no English translations of examples are available at the moment.

have not been proofread or have not passed sufficient verification, which means checking with three speakers, as a rule. The goal of the third step of filtering was to include in the corpus only reliable examples that do not require additional verification. This was done using a combination of comments and available metadata. The examples from the sections that have been entirely proofread for publishing were all included. Examples from other sections were only included if they had indication in the comment that they have been checked with at least two speakers. This excluded a considerable number of examples from the verbal part of the dictionary, which is under construction now. The total share of examples that come from the verbal part is currently just above 9%, despite the fact that verbal examples in the dictionary actually outnumber examples from all other parts taken together. As a consequence, the total size of the corpus is likely to increase significantly after the verbal part has been finalized, probably reaching 150,000 words.

An important difference between a traditional corpus and a collection of usage examples is that the latter may contain negative examples, i.e. sentences which are considered ungrammatical by the speakers. About 9.5% of sentences in the Beserman corpus of usage examples contain negative material. Following linguistic tradition, the sentences or ungrammatical parts of sentences are marked with asterisks. Other gradations of grammatical acceptability include $*?$ (ungrammatical for most speakers) and ?? (marginal / ungrammatical for many speakers). Negative examples are identified by the converter based on the presence of such marks in their texts and by the metadata.

The resulting examples are stored as tab-delimited plain text files. Each file contains examples from one dictionary entry or one grammatical questionnaire; positive and negative examples are kept in separate files. Each example is stored on a separate line and contains the original text, its Russian translation, and comments. Additionally, all filenames are listed in a tab-delimited file together with their metadata.

3.2 Morphological annotation

A workflow used in the Beserman documentation project prior to 2017 involved transcription of recordings for the spoken corpus in a simple text editor and manual annotation in SIL FLEX. That workflow was abandoned, mainly because manual annotation required too much resources, given the amount of recorded data that had to be processed. The new workflow comprises transcription, translation and alignment of recordings in ELAN, subsequent automatic morphological annotation and automatic disambiguation. Such an approach has been demonstrated to be well suited for processing spoken corpora of comparable size by Gerstenberger et al. (2017).

Developing a rule-based morphological analyzer would be a time-consuming task in itself, which is why Waldenfels et al. (2014) and Gerstenberger et al. (2017) advocate for transcribing in standard orthography and applying an analyzer for the standardized variety of the language. The Beserman case is different though because there already exists a digital dictionary. Using the dictionary XML as a source, I was able to produce a grammatical dictionary with the description of lemmata and inflection types. The dictionary was manually enhanced with additional information relevant for disambiguation, such as animacy for nouns or transitivity for verbs. Apart from that, several dozen frequent Russian borrowings, absent from the dictionary, were added to the list. Coupled with a formalized description of morphology, which I compiled manually, it became the basis for the automatic Beserman morphological analyzer. This analyzer is used for processing both new transcriptions in ELAN and usage examples.

After applying the analyzer to the texts, each token is annotated with a lemma, a part of speech, additional dictionary characteristics, and all inflectional morphological features, such as case and number. Annotated texts are stored in JSON. If there are several analyses possible for a token, they are all stored. The resulting ambiguity is then reduced from 1.7 to 1.35 analyses per analyzed token by a set of 87 Constraint Grammar rules (Bick and Didriksen, 2015). The idea was to apply only those rules which demonstrate near-absolute precision (at least 98%), while leaving more complex ambiguity cases untouched. The resulting coverage of the analyzer is 96.3% on the corpus of usage examples. While this number might look unusually high, it is in fact quite expected, given that there should be a dictionary entry for any non-borrowed word that occurs in any dictionary example.

3.3 Indexing

The annotated texts are uploaded to a server and indexed, after which they are available for search through a web interface using an open-source Tsakoprus corpus platform.[3] The interface allows making simple and multiword queries. The queries may include constraints on word, lemma, part of speech, morphological tags, glosses, or any combinations. In word and lemma search, wildcards and regular expressions are available. The search can produce a list of sentences or a list of words that conform to the query, together with some statistical data. The Russian translations are also morphologically analyzed and searchable. The interface is available in English and Russian, and several transliteration options are available for displaying the texts.

4 Differences from spoken corpora

For comparison, I will use the sound-aligned part of the Beserman spoken corpus (SpC), which was morphologically annotated and indexed the same way the Corpus of usage examples (ExC) was processed. At the moment, it contains about 38,000 words in monologues, natural dialogues and dialogues recorded during referential communication experiments. To account for the difference in size, all frequencies will be presented in items per million (ipm).

Two obvious factors make ExC quite different from SpC frequency-wise. First, the former contains comparable amount of usage examples for both frequent and non-frequent words. As a consequence, it has a different frequency distribution of words and lemmata. Its lexical diversity, measured as overall type/token ratio, is visibly higher than that of SpC (0.224 vs. 0.179), despite the fact that it is more than twice as large.

Second, elicited examples constitute a genre very different from natural narratives and dialogues. For example, they contain a much smaller number of discourse particles or personal pronouns. Table 1 demonstrates how different the frequencies of certain lexical classes are in the two corpora.

POS / Lexical class	SpC	ExC
noun	206K	**400K**
verb	213K	**289K**
adjective	52K	**68K**
pronoun	**177K**	123K
discourse ptcl.	**72K**	22K

Table 1: Frequencies (in ipm) of certain parts of speech and lexical classes in SpC and ExC.

POS	SpC	ExC
noun	3.15	**4.44**
verb	4.75	**6.08**
adjective	2.16	**2.56**
pronoun	2.72	**2.92**

Table 2: Average number of different forms per lemma for certain parts of speech in SpC and ExC.

Additionally, there are more different forms attested for a single lemma on average in ExC than in SpC (Table 2). Although unequal size of the corpora being compared could play a role here, the difference could be explained at least in part by the fact that the lexicographers had a goal of providing each word with a handful of examples containing that word in different forms and in different syntactic positions.

However, the analysis of value distributions of individual grammatical categories within a given part of speech reveals that they are usually not drastically different in the two corpora. Let us take nouns as an example. Nominal categories in Udmurt are case, number and possessiveness. Table 3[4] shows the distribution of case forms in the two corpora (all 8 spatial cases were collated in the last row). Table 4 shows the distribution of number forms. Table 5 shows the distribution of possessive suffixes.[5]

Case and number distributions only have minor differences in SpC and ExC. Moreover, an analysis of combinations of case and number suffixes shows that the distributions of their combinations also look very much alike. Possessiveness presents a somewhat different picture. Although not entirely different from SpC, the distribution in ExC shows lower figures for 2sg, 3pl, and especially 3sg, and higher ones for the first person possessors. Lower numbers for 3sg and 2sg have a straightforward explanation. Apart from being

[3] https://bitbucket.org/tsakorpus/tsakonian_corpus_platform/src

[4] The numbers in each column add up to slightly more than 100% because of the remaining ambiguity.

[5] Nouns may occur without possessive marking; only nouns marked for possessiveness were included. Figures for 2pl were verified manually because of large-scale ambiguity between nom,2pl and pl,acc.

case	SpC	ExC
nominative / unmarked	60%	65.7%
accusative (marked)	4.7%	9.9%
dative	2%	1.2%
genitive	1.8%	2.2%
2nd genitive (ablative)	0.74%	1.6%
instrumental	7%	4.7%
caritive	0.025%	0.024%
adverbial	0.25%	0.13%
all spatial cases	24.6%	20%

Table 3: Share of different case forms for nouns in SpC and ExC.

number	SpC	ExC
sg	94.6%	92.7%
pl	5.4%	7.3%

Table 4: Share of different number forms for nouns in SpC and ExC.

possessor	SpC	ExC
1sg	13.2%	28.2%
1pl	1.7%	5.6%
2sg	13.5%	7.7%
2pl	0.2%	0.4%
3sg	63%	54.8%
3pl	8.4%	3.3%

Table 5: Share of different possessive forms for possessive-marked nouns in SpC and ExC.

used in the direct, possessive sense, these particular suffixes have a range of "discourse", non-possessive functions. This is also true for Standard Udmurt (Winkler, 2001) and other Uralic languages (Simonenko, 2014). The 3sg possessive marks, among other things, contrastive topics and semi-activated topics that have to be reactivated in the discourse. The 2sg possessive is also used with non-possessive meanings, although less often, only in dialogues and in other contexts than the 3sg. Its primary discourse function is marking a new referent that is located physically or metaphorically in the domain of the addressee. Example 1, taken from a dialogue, shows both suffixes in non-possessive functions:

(1) Vaj so-ize=no
bring.IMP that-P.3SG.ACC=ADD
gozɘ̑-de!
rope-P.2SG.ACC
'Bring me that rope as well!'

tense	SpC	ExC
present	43.8%	39.4%
past (direct)	30%	43.2%
future	26.2%	17.4%

Table 6: Share of different tense forms for finite verbs in SpC and ExC.

The 3sg possessive on "that" marks contrast: another rope was discussed earlier in the conversation. The 2sg possessive on "rope" indicates that this particular rope has not been mentioned in the previous discourse and should be identified by the addressee. The rope is located next to the addressee. That the direct possessive sense is ruled out here follows from the fact that the whole dialogue happens on the speaker's property, so the rope belongs to him, rather than to the addressee. Such "discourse" use of the possessives in elicited examples is quite rare because they normally require a wider context to appear. A possible explanation for the lower 3pl figure in ExC is that it is often used when the plural possessor was mentioned earlier and is recoverable from the context.

A quick look at the distribution of verbal tenses in Table 6[6] shows a picture similar to that of the possessives. The distributions are not wildly different, but the past tense is clearly more frequent in ExC. This is expected because a lot of usage examples, especially for obsolete words, contain information about cultural practices connected to the item being described that are no longer followed. In this respect, the tense distribution in ExC resembles the one in the narratives, which usually describe past events.

5 Fitness for grammar studies

The way a corpus of usage examples is different from traditional corpora makes it unfit for some kinds of linguistic research. It cannot be used in any study that looks into long-range discourse effects, requires a context or a conversation involving several participants. This includes, for example, studies of discourse particles, use of anaphoric pronouns, or information structure (topic/comment; given/new information, etc.). Similarly, it is useless for studies that rely on

[6] Only finite forms were counted. The fourth tense, the second (evidential) past, was not included because most of its forms are ambiguous with a much more frequent nominalization, to which it is historically related.

word or lemma frequency counts, or on type/token ratios, as those may differ significantly from traditional corpora.

All downsides listed above are quite predictable; in fact, it was hardly necessary to build a corpus of usage examples to arrive at those conclusions. Whether such a corpus could be of any value in other kinds of linguistic research, is a less trivial question. The observations in Section 4 suggest that the answer to that question is positive. While the Beserman corpus of usage examples differs from the spoken corpus in the places where it could be predicted to differ, it is remarkable how statistically similar the two corpora are in all other respects. The relative frequencies of grammatical forms are only different (although not extremely different) if these forms convey deictic or discourse-related meanings. In other cases, the forms have very similar distributions. This means that corpora of examples in principle can be used in linguistic research that involves comparing frequencies of certain grammatical forms or constructions, with necessary precautions.

Only research that involves comparing frequency counts or distributions of linguistic phenomena has been discussed so far. A lot of grammar studies, however, only or primarily need the information about (un)grammaticality or productivity of a certain phenomenon. It turns out that a corpus of usage examples could be actually superior to a traditional corpus of a comparable size for such studies. The reason for that is higher lexical diversity and more uniform frequency distribution of lemmata. This means that for any form or construction being studied, the researcher will see more *different* contexts involving it than in a traditional corpus, on average.

Let us take the approximative case as an example. This case, which has the marker *-lań* in all Udmurt varieties where it exists, marks the Ground in the approximate direction of which the Figure is moving. Although it is claimed to exist in literary Udmurt by all existing grammars, corpus studies reveal that it only functions as an unproductive derivational suffix compatible with a handful of nominal, pronominal and adjectival stems. To learn whether it can be considered a productive case suffix in Beserman, its compatibility with a wider range of nouns should be established.

The approximative has similar relative frequencies in ExC and SpC: 1858 ipm and 1750 ipm, respectively. In both corpora, the approximative was not a primary focus of investigation. The number of different contexts it is encountered in, however, is much higher in ExC. In SpC, there are 16 different types that contain an approximative suffix, featuring 10 different stems. Only one of those stems (*ulća* 'street') does not attach the derivational approximative suffix in Standard Udmurt. This is definitely insufficient to establish its productiveness in Beserman. ExC, however, contains 32 different types that belong to 25 different stems. Such a difference can hardly be ascribed to the larger size of ExC because the number of different types is expected to have slower-than-linear growth with respect to the corpus size, and the type/stem ratio is expected to go up rather than down. Out of these stems, at least 5 are incompatible with the approximative in Standard Udmurt, including *reka* 'river', *korka* 'house' and *šundə̂* 'sun'.[7] Another 5 come from negative examples that highlight the incompatibility of the approximative with certain inflected postpositions (relational nouns). All this proves that it is most probably a productive suffix, while outlining the limits of its productivity.

Comparison of the figures for the recessive suffix, which also was not the focus of investigation in either corpus, yields similar results. The recessive case, with a marker *-laśen*, is the semantic opposite of the approximative and does not exist in the literary language even as a derivational suffix. In SpC, there are 7 different types that contain it. All of them have different lemmata, but again, only one of the types (*kətlaśen* 'from the side of the belly') suggests that they might not constitute a closed set of denominal adverbs. By contrast, ExC has 26 different types, containing 26 different stems, 4 out of which come from negative examples.

The skewed distribution of parts of speech could be beneficial too. For example, there is a construction in Udmurt that involves juxtaposition of an unmarked nominal stem to another noun, e.g. *t'ir nə̂d* 'axe handle'. It exists in other Uralic languages and has been analyzed as compounding by Fejes (2005). However, its productivity, allowed

[7] The 65,000-word FLEX part of the spoken corpus, not counted here, contains 21 different stems. However, this is only because it includes transcriptions of referential communication experiments specifically designed to make the participants use spatial cases and postpositions with a variety of nouns.

possessive relations and constraints on the dependent noun (animacy, referential status, etc.) vary significantly across Uralic languages and dialects. A detailed investigation of that construction would therefore require looking at a large number of instances featuring a variety of dependent nouns. A search for such a construction yields 77 different dependent nouns in SpC (total frequency 7525 ipm) and 449 different nouns in ExC (total frequency 25132 ipm).[8] In this case, the tremendous increase in numbers stems from the increased share of nouns in ExC, rather than from an increased lexical diversity. Nouns are almost twice as likely to appear in ExC than in SpC. Therefore meeting a sequence of two nouns would be almost 4 times higher if they were generated independently. The independence condition does not hold for words in a sentence, of course, but the observed factor of 3.34 is still large enough to make the corpus of examples a more attractive tool for studying the juxtaposition construction.

Research questions that arise when writing a reference grammar mostly belong to the type discussed above. Concise description of morphological and syntactic properties of an affix or a construction mostly require looking at a range of diverse examples to determine their productivity and constraints. The three case studies above show that this is exactly the situation where a corpus of usage examples could outperform a similarly sized spoken corpus.

Apart from the aforementioned considerations, which would probably be valid for a corpus based on any comprehensive dictionary, there are specific features of the Beserman usage example corpus that may become beneficial for research. Most importantly, it contains elicited examples from grammatical questionnaires, both positive and negative. Although their share is too small to reach any conclusions about their usefulness at the moment, this could be a first step to the reuse of elicited data in subsequent research. Currently, data collected through questionnaires is virtually always used only once. At best, the examples are archived to support the analysis based on them and ensure that the findings are reproducible; at worst, they are discarded and forgotten. Of course, each questionnaire is tailored to the particular research question of its author. However, it is probable that if the corpus is large enough, it will contain examples that could prove useful for the research questions other than their author had in mind. Presence of negative material could partially reduce the general aversion to corpora that the linguists working in the generative paradigm tend to have. Availability of such corpora for multiple languages will also facilitate typologically oriented research on syntax, which otherwise relies on manual work with reference grammars and other secondary sources.

Finally, the Beserman corpus of usage examples, as well as any corpus based on a bilingual dictionary, is in essence a parallel corpus. This could be used for both linguistic needs (see e.g. Volk et al. (2014) for the list of possibilities) and for developing or evaluating machine translation systems.

6 Conclusion

Having a comprehensive dictionary in a machine-readable form and a morphological analyzer allows one to create a corpus of usage examples rather quickly. Its size could be comparable to those of large spoken corpora, which tend to have a maximum of 150,000 tokens for unwritten endangered languages. Even if a spoken corpus is available, this is a significant addition. Sometimes, however, dictionary examples could be the only large source of linguistic data, e.g. in the case of extinct languages. It is therefore important to know how the data of usage examples corresponds to that of spoken texts, and for what kind of linguistic research they are suitable. An analysis of the Beserman corpus of usage examples reveals that it can actually be used as a reliable tool in a wide range of research on grammar. Obvious limitations prevent its use in studies that involve word and lemma counts, discourse or information structure. However, outside of these areas, usage examples are quite similar to natural texts, which justifies use of such corpora. Increased lexical diversity and more uniform word distributions make corpora of usage examples even more useful for some kinds of research than traditional corpora of similar size. Finally, such corpora can additionally contain data from questionnaires and negative material, which could facilitate their reuse.

Acknowledgements

I am grateful to all my colleagues who have worked on the documentation of Beserman Ud-

[8] To avoid as much ambiguity as possible without manual filtering, the search was narrowed down to unambiguously annotated tokens.

murt and especially to Maria Usacheva. This research would have not been possible without their prolonged lexicographic work.

References

Eckhard Bick and Tino Didriksen. 2015. CG–3 - Beyond Classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39. Linköping University Electronic Press.

László Fejes. 2005. *Összetett szavak finnugor nyelvekben [Compound words in Finno-Ugric languages].* PhD, Eötvös Loránd Tudományegyetem, Budapest.

Gill Francis. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and technology: In honour of John Sinclair*, pages 137–156. John Benjamins, Amsterdam & Philadelphia.

W. Nelson Francis and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, Boston.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 57–66. Association for Computational Linguistics.

Valentin Kelmakov. 1998. *Kratkij kurs udmurtskoj dialektologii [Brief course of Udmurt dialectology].* Izdatel'stvo Udmurtskogo Universiteta, Izhevsk.

Olesya Khanina. 2017. Digital resources for Enets: A descriptive linguist's view. *Acta Linguistica Academica*, 64(3):417–433.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pave Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, pages 425–432, Barcelona.

Gerson Klumpp. 2005. Aspect markers grammaticalized from verbs in Kamas. *Acta Linguistica Hungarica*, 52(4):397–409.

Nadezhda Lyukina. 2008. *Osobennosti jazyka balezinskix i jukamenskix besermjan [The peculiarities of the language of Balezino and Yukamenskoye Besermans].* PhD, Udmurt State University, Izhevsk.

Aleksandra Simonenko. 2014. Microvariation in Finno-Ugric possessive markers. In *Proceedings of the Forty-Third Annual Meeting of the North East Linguistic Society (NELS 43)*, volume 2, pages 127–140.

Tamara Teplyashina. 1970. *Jazyk besermjan [The language of the Beserman].* Nauka, Moscow.

Martin Volk, Johannes Graën, and Elena Callegaro. 2014. Innovations in Parallel Corpus Search Tools. In *LREC 2014 Proceedings*, pages 3172–3178.

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"*, volume 13, pages 720–728.

Eberhard Winkler. 2001. *Udmurt.* Number 212 in Languages of the world. Lincom Europa, Munich.