Compiling the Uralic Dataset for NorthEuraLex, a Lexicostatistical Database of Northern Eurasia

Johannes Dellert
Seminar für Sprachwissenschaft
Universität Tübingen
jdellert@sfs.uni-tuebingen.de

December 16, 2014

Abstract

This paper presents a large comparative lexical database which covers about a thousand concepts across twenty Uralic languages. The dataset will be released as the first part of NorthEuraLex, a lexicostatistical database of Northern Eurasia which is being compiled within the EVOLAEMP project.

The chief purpose of the lexical database is to serve as a basis of benchmarks for different tasks within computational historical linguistics, but it might also be valuable to researchers who work on the application of computational methods to open research questions within the language family.

The paper describes and motivates the decisions taken concerning data collection methodology, also discussing some of the problems involved in compiling and unifying data from lexical resources in six different gloss languages.

The dataset is already publicly available in various PDF formats for inspection and review, and is scheduled for release in machine-readable form in early 2015.

1 Introduction

Recent years have seen a surge of interest in computational historical linguistics, where computational methods are used to analyze phenomena of interest to historical linguistics, such as language relationships, language contacts, and language change.

The main focus of the field has been on applying phylogenetic methods from bioinformatics to automated language classification. These methods can be separated into two basic approaches which differ in the type of data they operate on.

In the more popular character-based approaches, expert cognacy judgments are encoded as binary matrices where each line represents the presence or absence of some cognate set in each language. Character-based approaches have recently started to be applied to Uralic languages [1].

In the competing distance-based approaches, phonetic representations of word forms are used to compute a measure of lexical distance between languages. While in general, these approaches perform worse than character-based approaches in the language classification task, they have the advantage of much wider applicability because they do not rely on expert cognacy judgments. Furthermore, work on actual realizations is closer to the interests and goals of mainstream historical linguistics. For instance, work in the field includes the automation of some parts of the comparative method, such as the detection of regular sound correspondences.

A major weakness of the datasets commonly used in research on distance-based methods is that the concept lists they build on are very short (40 concepts in the case of the global-coverage ASJP database [2], and Swadesh lists of not more than 200 concepts in sources for individual language families). While longer lists provably do not lead to higher performance in current approaches to language classification [3], such short lists are clearly not sufficient for more advanced comparative tasks.

One of the key goals of the EVOLAEMP project at the University of Tübingen is to improve the state of the art in distance-based approaches. The small size of the word lists in existing databases necessarily makes large-scale data collection a major part of our work. We are collecting substantial amounts of lexicostatistical data for more than a hundred languages of Northern Eurasia, which we plan to release under the name NorthEuraLex.

Since the Uralic family is among the most thoroughly investigated language families, we do not expect to find out anything new about their internal classification based on this dataset. Given the extensive knowledge of cognate sets gained in more than two centuries of Uralic studies, character-based methods as applied by [1] will remain the preferred methods for automated classification within the family.

However, the wealth of established knowledge also makes the Uralic languages an ideal benchmark for other methods, since automatically extracted sound correspondences and cognacy judgments can be evaluated against comprehensive expert knowledge. The central role of Uralic in the EVOLAEMP project is motivated by this fact, and the advantages of choosing a relatively small and therefore tractable family compared to the even more thoroughly researched Indo-European languages.

In addition to Uralic, NorthEuraLex is intended to also cover the contact languages

of all branches of Uralic throughout their development (including samples from all branches of Indo-European, and from all the language families sometimes summarized as Altaic and Paleosiberian), with promising perspectives for evaluating computational models of language contact.

With the subproject of compiling the Uralic dataset close to completion, the author (who is responsible for Uralic and Paleosiberian data collection within the project) has decided to release this data set to the Uralist community in order to receive feedback on errors and possible improvements. Collaboration with experts in the individual languages as well as in computational methods for Uralic languages are the necessary next steps for improving the quality of the dataset.

2 Data Collection

This section describes and motivates the many decisions taken during data collection, including the choice of languages, concepts, and sources. As an introductory remark, it should be emphasized that our goal is not to compile high-coverage general-purpose electronic dictionaries, but a lexicostatistical database. This means that only the most common or natural realization for each concept is to be included, not all the synonyms commonly listed in large dictionaries. Multiple translations are only included in the very few cases where there is no clear preference for one realization. As an example, the database only contains the Finnish word *koira* for "dog", rather than the dialectal *hunttu*, the poetic *hurtta*, and even the archaic *peni*, even though the latter is most interesting to the Uralist because of its cognates in many other branches.

2.1 The Language Sample

The database covers a sample of twenty Uralic languages, chosen mainly according to the availability of high-coverage lexical resources, but also in order to cover all branches of the family. In the following list, the twenty selected languages are given with the corresponding ISO 639-3 language codes, which are used throughout the released resources as well as this paper as shorthands:

Finnish (`fin`), Karelian (`krl`), Veps (`vep`), Standard Estonian (`ekk`), Livonian (`liv`), Southern Sami (`sma`), Northern Sami (`sme`), Skolt Sami (`sms`), Kildin Sami (`sjd`), Meadow Mari (`mhr`), Moksha (`mdf`), Erzya (`myv`), Udmurt (`udm`), Komi-Zyrian (`kpv`), Hungarian (`hun`), Northern Mansi (`mns`), Northern Khanty (`kca`), Northern Selkup (`sel`), Tundra Nenets (`yrk`), and Nganasan (`nio`).

2.2 Selecting the Concepts

Since the concept lists used in lexicostatistical work are typically much shorter than our goal of 1.000 concepts, we decided to build our own concept list for data collection. The concept list used by [1] mainly builds on Swadesh-type lists of stable concepts such as the Leipzig-Jakarta list [4], and a list of 100 concepts especially selected for their stability within Uralic. However, they also sample a hundred concepts from the 1.460-item basic vocabulary list used by the World Loanwoard Database WOLD [5] to assess the consequences of using a set of less stable concepts.

While our list includes all concepts from the various Swadesh-type lists, basing the extension to the desired size on the full WOLD list would have led to the inclusion of many concepts which have appeared in the material culture of many Uralic minorities only very recently, so that the frequency of loanwords in a database based on such a list would be very high.

Moreover, for some of the languages in our sample (`sjd,mns,kca,sel,nio`), the best available dictionaries only contain a few thousand entries. Often, the space which recent Russian loans would have occupied is saved to provide better coverage of the inherited lexicon. In the interest of achieving full coverage for as large a concept set as possible, our concept set is therefore based partly on data availability.

Even within the inherited lexicon, many concepts which are only relevant to some climate zones or modes of subsistence (“tent”, “reindeer”, different types of boats and sleighs), had to be left out because of low overall dictionary coverage. Conversely, we opted to include some very commonly borrowed words with near-complete coverage (e.g. “bread”, “book”, “church”, names of weekdays and months) as useful benchmarks for loanword detection.

Our current list of 1.016 concepts consists of about 480 nouns (including 48 parts of the body and 35 animal names), 100 adjectives, 45 adverbs, 340 verbs, and 50 words from smaller word classes (pronouns, question words, numerals). For inspection of the concept list, the reader is referred to the table of contents in the preliminary PDF release of the database (see below).

2.3 Sources

While sufficent lexical resources are available for all the languages in the sample, a major obstacle to data collection was that to build on the best resources for each language, six different gloss languages had to be bridged. In order of relevance, the six most important gloss languages were Russian, German, Norwegian, Finnish, English, and Estonian. In order to clarify translations, some additional information needed to be retrieved from sources in Swedish, Hungarian, and Latvian.

Within the EVOLAEMP project, German serves as the pivot language for the entire lexical infrastructure. Choosing the native language of most project members minimizes the risk of additional errors and imprecisions which occur when glosses are written by non-native speakers of the gloss language. German is an especially useful choice for Uralic because of its long tradition as the primary language of the field, leading to high availability of dictionaries for target languages across the sample. To make the data more accessible, English and Russian glosses are included in our databases, sometimes also serving as an informal way of disambiguating polysemous German glosses.

To avoid duplication of effort, freely available electronic resources were used wherever possible. This includes downloadable lexical databases such as the `mhr-eng` dictionary developed by the Mari Web Project[1], and the electronic dictionaries maintained by the Giellatekno project[2]. For four languages (`krl`, `liv`, `sjd`, `mhr`), online dictionaries were used either because they constitute the highest-coverage resources for their respective languages, or because their contents are identical to the best published dictionaries, making them convenient tools for access to printed material.

While these electronic resources were extremely helpful for data collection, the bulk of the data still needed to be extracted from dictionaries published on paper. We avoided using etymological dictionaries as primary sources because of their possible bias towards including lexemes with parallels in related languages (see the *peni* example above), instead of the most natural realizations in the current standard languages. Whenever available, school dictionaries were preferred over these more comparatively oriented dictionaries. Despite their shortcomings in accurately representing pronunciation, the standardized orthographies employed by these dictionaries make it easier to aggregate information across resources, and to abstract away from dialectal distinctions that would otherwise make it very hard for a non-specialist to compile a word list which consistently represents a single variant. School dictionaries, which tend to only give a few translations sorted by frequency or salience, are the most reliable source if one wants to find the most common realization for some basic concept in an unfamiliar target language. Still, large dictionaries of the standard language and scientific dictionaries were found invaluable as fallback options for problematic cases. A full list of all sources employed can be found on the author's website, and is also distributed together with the data.

[1] `http://www.univie.ac.at/maridict/site-2014/`
[2] `http://giellatekno.uit.no/words/dicts/`

2.4 Methodology

The process of compiling the wordlist for a language (notation: `lng`) in NorthEuraLex is organized in five stages. First, the initial German concept list is translated into the relevant gloss language (notation: `glo`), with the requirement that at least one resource which allows lookup in both directions (`glo-lng` and `lng-glo`) must be available. Resources where both directions are treated separately are always preferred to resources where one of the two directions is only available as an index or a mechanical reordering of the translation pairs in the other.

In the second stage, the gloss language lemmas are looked up in the `glo-lng` direction. All `lng` lemmas occurring in the relevant dictionary equations are collected, possibly extended by annotations given in the sources. The purpose of this stage is to collect a set of `lng` lemmas which cover the concept set as completely as possible.

The third stage consists in looking up the collected lemmas in the `lng-glo` resource. The `glo` lemmas corresponding to each `lng` lemma are collected, including all polysemies and annotations. In many cases, the example sentences given by the dictionaries are interpreted for further usage hints, which are stored as additional annotations for later reference.

The fourth stage is where information in different gloss languages is aggregated by translating the `lng-glo` entries into `lng-deu` entries. This process is assisted by additional electronic dictionaries in both relevant directions (`glo-deu` and `deu-glo`) to ensure consistency of translations across target languages.

In the fifth stage, the decisions on which `lng` lemmas to include for each concept are made in a final pass over the entire data. This step involves a complex and not fully formalizable decision process based on best fit of glosses, position of the translation pairs in the `lng-glo` and `glo-lng` resources, the disambiguating information collected in stages 2 and 3, some ad-hoc research in online resources such as Wikipedia entries in various gloss and target languages, and the author's varying levels of familiarity with the languages involved. In order to reduce the impact of translation errors, the glosses in the translated `lng-deu` entries are only used as indices bridging the different gloss languages, while the decision process itself only relies on the `lng-glo` and `glo-lng` entries extracted from the original resources.

2.5 Issues and Difficulties

Of the many problems encountered in the various stages of data collection, the problem of picking a small set of good gloss language lemmas for each concept turned out as being the most crucial and problematic. Beyond the expected problem of polysemous glosses that overlap between gloss languages, one of the most challenging

problems for unifiability is the different granularity of meanings lexicalized in a given domain in different gloss languages.

As a case in point, consider the representation of verbal meanings in Russian. In stage 2 of our lookup process, we used the following six Russian equivalents of the German verb *legen* "to lay": класть, положить, укладывать, уложить, складывать, сложить. While German can also express many of the distinctions expressed by the Russian equivalents using prefixes to the basic verb (*weglegen*, *ablegen*, etc.), the prefixless verb form is clearly perceived as most basic, and can therefore be expected to be used in any `deu-lng` dictionary to index the `lng` equivalents of "to lay".

By contrast, it is largely unpredictable which of the corresponding Russian verbs is used for this purpose in a small `rus-lng` dictionary. While this merely complicates the lookup process, the real problem is the treatment of verbs in larger dictionaries. Where the Russian verbs lexicalize slight differences in grammatical aspect and other meaning components, regular derivational morphology of the target language is often used to faithfully represent these nuances in `rus-lng` dictionary equations. For a significant part of the verbal concepts in our list, this leads to a very large number of candidate lexemes. For instance, our lookup process leads to not less than 16 possible `kpv` equivalents for *legen*.

The task of selecting the single most natural equivalent of the German verb from such a list is a challenge even with good knowledge of the target language, and cannot reliably be accomplished by a non-expert. Our preliminary solution is to adapt a selection process based on a hierarchy of preferences. Preference is first given to glosses mentioned earlier in the `rus-lng` direction, then to non-derived `lng` verbs, then to lexemes where multiple equivalents of the intended concept are mentioned early in the `lng-rus` direction, and finally, to less polysemous glosses. For instance, *положить* receives more weight than *сложить* as an equivalent of "to lay" because the latter can also mean "to fold up".

Measured against the goal of retrieving the most natural lexeme for each concept, our data collection strategy inevitably leads to many erroneous entries due to missing frequency information, in addition to the problems caused by misinterpretation of dictionary entries in one of the less familiar gloss languages. Without good knowledge of all the target languages, better results could only be achieved by considering parallel texts which describe many prototypical situations. However, in the absence of parallel corpora of any useful size for any of the minority languages, this would be a large project in itself, requiring extensive fieldwork. Contributions by experts on all the target languages are therefore needed if the quality of the dataset, especially in the verbal domain, is to be improved much beyond the current state.

3 Phonetic Transcription

In a lexicostatistical database, some form of unified phonetic description is needed to achieve comparability across languages. The aim for such a description is to be as faithful as possible, while still abstracting away from dialectal and speaker-specific phenomena. While within Uralic studies, there are good reasons for using the traditional Uralic Phonetic Alphabet (UPA) for such purposes, our database requires a unified phonetic notation that also covers all the relevant contact languages, and is as accessible as possible to non-Uralists. For these reasons, we decided to use the International Phonetic Alphabet (IPA) as the common representation system.

To implement the orthography-to-IPA transducers, sets of very simple greedy replacement rules were compiled for each language. For many languages, a second processing stage for treating palatalization as represented in Russian-based Cyrillic orthographies was needed. While simple toolchains of this type work reasonably well in most cases, a number of imperfections and challenges remain.

An obvious challenge is any information that is not immediately visible in the dictionary forms because the standard orthography does not fully specify or faithfully represent pronunciation. Examples include the non-representation of non-phonemic weakly voiced vowels in Tundra Nenets and Skolt Sami, palatalization in the nominative case of some Estonian nouns (caused by an elided front vowel which is only visible in other case forms), and epenthetic vowels which split up consonant clusters in Northern Sami and other languages. While in many cases, some effort was made to predict and implement these phenomena, in others we opted to reduce complexity by aiming for a phonemic notation that more closely corresponds to the orthography.

Another area of difficulty is the treatment of suprasegmental phenomena. This includes issues like the impact of stress on vowel quality in Moksha and other languages, and suprasegmental palatalization in Skolt Sami. Since a correct implementation of these phenomena presupposes a level of understanding which is difficult to derive from literature alone without the help of experts, they are not yet fully covered by the current version of our transliterators.

Since many of these distinctions are not of central importance to historical linguistics, we decided that even a transcription which does not fully cover these phenomena was good enough for a first release. Even though the shortcomings leave the current transcriptions in an imperfect state somewhere between the phonetic and phonemic levels, already the current version is detailed enough to accurately represent many of the relevant sound correspondences between languages.

Still, there is considerable room for improvement. In addition to extending the coverage of the described phenomena, the phonetic transcription pipeline could be streamlined a lot more by using full-blown finite state technology. While some work

Language	fin	krl	vep	ekk	liv	sma	sme	sms	sjd	mhr
Certain	1015	976	914	1014	894	874	889	956	951	913
Uncertain	1	39	81	2	96	72	73	23	26	101
Missing	0	1	21	0	26	70	54	37	39	2

Language	mdf	myv	udm	kpv	hun	mns	kca	sel	yrk	nio
Certain	920	955	934	883	1015	941	938	877	945	888
Uncertain	64	53	72	119	1	52	45	56	50	47
Missing	32	8	10	14	0	23	33	83	21	81

Figure 1: Current coverage of the concept list (on December 15th, 2014).

in this direction will eventually be done within our project, we would be happy to collaborate with any researchers who work on grapheme-to-phoneme conversion for any Uralic language. Feedback about errors of and possible improvements to our transliterators is very welcome as well.

4 The Dataset

From the full set of 1.016 concepts across 20 languages, we have so far been able to retrieve some information for 97% of all concept-language pairs. For about 92%, we are reasonably confident that the extracted translations are correct. For 5% of the data, our translations are still classified as uncertain, the main causes being lack of precision in or insecurity about the interpretation of source entries. Figure 1 gives an overview of the current coverage for each language. Because some work is still being done, we expect a slightly smaller proportion of uncertain entries in the final release.

Current snapshots can be retrieved in different PDF formats from the author's webpage[3]. The main file contains all the parallel translations and their transliterations in a one-concept-per-page format. Alternatively, single-language wordlists with glosses in German, English, and Russian are available as well. The purpose of these pre-release snaphots is to make it as easy as possible for specialists in the individual languages to inspect the data and to give us feedback on our unavoidably many errors.

Releases in various machine-readable formats will follow after some feedback was received. New versions will continually be released based on expert feedback or additional knowledge gained by the author. All materials will be published under an open license, allowing other researchers to build upon the database as they wish.

[3] http://www.sfs.uni-tuebingen.de/~jdellert/northeuralex

5 Future Work

One of the long-term goals of the author is to further improve the coverage of the Uralic language family by collecting data on some additional languages. At least six languages are planned for future inclusion: Enets, Hill Mari, Lule Sami, Inari Sami, Ingrian, and Võro. Work on producing the datasets for major contact languages is already under way, with the current focus of data collection being on the languages of Siberia. The release of this part of the NorthEuraLex database is currently scheduled for the end of 2015, with the remaining parts following until the end of 2017.

To turn the database into an attractive benchmark for cognate detection, we are also starting to enhance the dataset by cognacy judgments. This subproject will be pursued further in collaboration with other researchers, based on written sources, or on existing databases such as the one developed by the Etymon project [6].

6 Conclusion

This paper describes the design decisions behind and the data collection process for a large lexicostatistical database that spans about a thousand concepts in twenty Uralic languages. The envisioned primary use case of the database is as a benchmark for different tasks in computational historical linguistics. Once it is enhanced by cognacy judgments, the database will become one of the largest available testsets for automated cognate detection. The unusually high coverage of the database will also allow markup of quite a few cross-semantic cognates, providing a first test case for advanced cognate detection methods that also attempt to model semantic change. By including variants of Proto-Uralic as reconstructed by Károly Rédei [7] or Pekka Sammallahti [8], the database could also become a very interesting test case for the challenging task of automated proto-language reconstruction (see e.g. [9]).

Beyond its relevance for the field of computational historical linguistics, an open and readily available lexical database is likely to increase the attractiveness of interdisciplinary work on Uralic languages, generating more interest in the field and hopefully leading to new discoveries.

Acknowledgments

This work has been supported by the ERC Advanced Grant 324246 'EVOLAEMP', which is gratefully acknowledged.

References

[1] Kaj Syrjänen, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski, and Niklas Wahlberg. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica*, 30(3):323–352, 2013.

[2] Søren Wichmann, André Müller, Annkathrin Wett, Viveka Velupillai, Julia Bischoffberger, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Zarina Molochieva, Pamela Brown, Harald Hammarström, Oleg Belyaev, Johann-Mattis List, Dik Bakker, Dmitry Egorov, Matthias Urban, Robert Mailhammer, Agustina Carrizo, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Pattie Epps, Anthony Grant, and Pilar Valenzuela. The ASJP Database (version 16), 2013. http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm.

[3] Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354, 2009.

[4] Uri Tadmor. Loanwords in the world's languages: Findings and results. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the World's Languages: A Comparative Handbook*, page 55–75. Mouton de Gruyter, 2009.

[5] Martin Haspelmath and Uri Tadmor, editors. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2009.

[6] R. Yangarber, M. Salmenkivi, and M. Välisalo. A database of the uralic language family for etymological research. Technical report, University of Helsinki, 2008. Technical Report Series C; C-2008-38.

[7] Károly Rédei, editor. *Uralisches etymologisches Wörterbuch*. Akadémiai Kiadó, Budapest, 1988.

[8] Pekka Sammallahti. Historical phonology of the Uralic languages (With Special Reference to Permic, Ugric and Samoyedic). In Denis Sinor, editor, *The Uralic Languages*. 1988.

[9] Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 10.1073/pnas.1204678110, 2013.