




# NorthEuraLex: a wide-coverage lexical database of Northern Eurasia

Johannes Dellert<sup>1</sup>  · Thora Daneyko<sup>1</sup> ·  
Alla Münch<sup>1</sup> · Alina Ladygina<sup>1</sup> · Armin Buch<sup>1</sup> ·  
Natalie Clarius<sup>1</sup> · Ilja Grigorjew<sup>1</sup> ·  
Mohamed Balabel<sup>1</sup> · Hizniye Isabella Boga<sup>1</sup> ·  
Zalina Baysarova<sup>1</sup> · Roland Mühlenbernd<sup>1</sup> ·  
Johannes Wahle<sup>1</sup> · Gerhard Jäger<sup>1</sup>

Published online: 30 November 2019  
© The Author(s) 2019

**Abstract** This article describes the first release version of a new lexicostatistical database of Northern Eurasia, which includes Europe as the most well-researched linguistic area. Unlike in other areas of the world, where databases are restricted to covering a small number of concepts as far as possible based on often sparse documentation, good lexical resources providing wide coverage of the lexicon are available even for many smaller languages in our target area. This makes it possible to attain near-completeness for a substantial number of concepts. The resulting database provides a basis for rich benchmarks that can be used to test automated methods which aim to derive new knowledge about language history in underresearched areas.

**Keywords** Lexical database · Northern Eurasia · Indo-European languages · Uralic languages · Turkic languages · Siberian languages · Caucasian languages

## 1 Introduction

The most basic prerequisite for any computational study in historical linguistics is an electronic database which contains the information a linguist would look up in dictionaries or other sources in a machine-readable format. The absence of such databases has been one of the limiting factors to the development of the field, but some very useful resources have become available during the past decade (Dunn 2015; Greenhill et al. 2008; Wichmann et al. 2016), and the pace at which new

---

✉ Johannes Dellert  
johannes.dellert@uni-tuebingen.de

<sup>1</sup> Seminar für Sprachwissenschaft, Universität Tübingen, Wilhelmstraße 19, 72074 Tübingen, Germany

resources appear is accelerating (Greenhill 2015; Bowerman 2016; Kaiping and Klamer 2018).

While the number of large-scale lexical databases is increasing, most of them systematically cover only about 100 or 200 very basic concepts for each language. An extreme case among these narrow-coverage databases is the database produced by the Automated Similarity Judgment Program (ASJP) (Wichmann et al. 2016), which due to its global coverage of more than 5000 languages is by far the largest in terms of the number of languages covered, but only includes data for a mere 40 concepts per language. The rationale for this reductionist approach is presented in Holman et al. (2008), arguing that longer lists do not improve performance in simple language classification and phylogenetic inference. Beyond these tasks, the ASJP database has been used to investigate many issues of general interest to historical linguists, like the stability of concepts against borrowing and semantic change, the question whether sound symbolism creates problematic amounts of lexical similarity between unrelated languages, and whether there are correlations between phoneme inventory sizes and extralinguistic factors such as population size or geographic isolation.

While it is of course much more feasible to achieve a good global coverage with such short lists, more complex tasks in tools for computational historical linguistics tend to require more data. Such tasks include the extraction of regular sound correspondences as are necessary to actually prove language relationships, the automated detection of loanwords or cognates which have undergone semantic shifts, and models for detecting contact events between languages.

Some interesting datasets have arisen from studies which investigate automated approaches to such tasks. For instance, the Austronesian Basic Vocabulary Database (ABVD) by Greenhill et al. (2008) was used by Bouchard-Côté et al. (2013) to evaluate their system for Bayesian reconstruction of proto-forms, and the IELex database of Indo-European basic vocabulary (Dunn 2015) grew out of an early lexicostatistical study by Dyen et al. (1992). Both databases primarily provide cognacy annotations which have been used for phylogenetic inference, but use either the official orthographies or often idiosyncratic transliterations instead of providing a phonetic transcription for those languages which have a standardized written form. At 1400 languages, the ABVD provides virtually complete coverage of the world's largest language family, and IELex is popular due to the central role of Indo-European in historical linguistics. Both ABVD and IELex cover lists of about 200 concepts, which is substantially more than the ASJP database, but far from enough for work on regular sound correspondences. Moreover, since each of these databases only covers a single language family, they are not adequate for the study of cross-family patterns.

More comprehensive lexical resources which aim to cover a substantial part of the basic vocabulary across more than one language family, have recently started to appear for some linguistic areas in the world. For instance, TransNewGuinea.org (Greenhill 2015) includes data about more than 800 languages and dialects of New Guinea, the least well-studied linguistic region of the world, and aims to provide a unified phonetic format that can be processed by computational tools. Due to the very sparse documentation of many languages, the total size of this database will not

be able to expand far beyond its current 125,000 entries, or an average of just over 150 words per language. The Chirila database of Australian languages by Bower (2016) is similar in scope and structure, with the goal of eventually making all known lexical data available. Due to the complicated legal situation when publishing full resources, and history-induced hesitancy of many linguistic groups when it comes to giving outsiders access to their languages, only 150,000 of 780,000 database entries are freely available at the moment. The LexiRumah database by Kaiping and Klamer (2018) summarizes the result of extensive fieldwork on the languages of the Indonesian islands of Alor and Pantar, many of which are Papuan (i.e. non-Austronesian) languages, with an average of 337 concepts across 101 language varieties.

Currently, no comparable cross-family database exists for any of the more well-researched linguistic areas in the world. The only wide-coverage lexical database which covers a multitude of language families from different regions is the Intercontinental Dictionary Series (IDS) edited by Key and Comrie (2015). This collection of dictionaries has the advantage of consisting of expert contributions, but has not been extended for years, and remains at just over 329 different languages from all over the world, with a focus on some families like Nakh-Daghestanian, Tai-Kadai, and Austroasiatic. The disadvantages of this database are that it does not cover any larger geographical area which could be used for cross-family contact models, and that there are large gaps in lexical coverage even for languages where more complete resources would be available.

NorthEuraLex, the preliminary release version of which we present in this article, currently spans a large list of 1016 concepts across 107 languages predominantly from Northern Eurasia. Across the total of 121,614 dictionary forms contained in the database, a uniform phonetic transcription in IPA is available. The vast majority of these transcriptions has been generated automatically based on phonological descriptions for the languages, except for about a dozen languages whose complex historically grown orthographies did not allow for that (e.g. English, French, Danish, and Irish), or whose writing systems do not completely represent the phonology (e.g. Chinese, Japanese, and Persian). For some of these languages, like Arabic and Hebrew, we are able to make use of extended dictionary orthographies which provide full information about pronunciation. Because a consistent IPA transcription is used for all of our word lists, they can be converted automatically into a range of other less fine-grained transcription formats, such as ASJP encoding, in order to combine them with other resources.

NorthEuraLex occupies a unique position in a fast-growing landscape of basic vocabulary databases which by now cover many linguistic areas. Due to the scarcity of documentation and unfeasibility of fieldwork covering hundreds of languages for a single institution that could ensure a uniform treatment, most of the other databases will always remain gappy. Northern Eurasia is very different in this respect, because the vast majority of its languages is well-documented, making it possible to reach near-complete coverage across an entire linguistic area. The studies building on the ASJP database exemplify the type of research that will be facilitated with the availability of our wide-coverage cross-family database in a uniform transcription format.

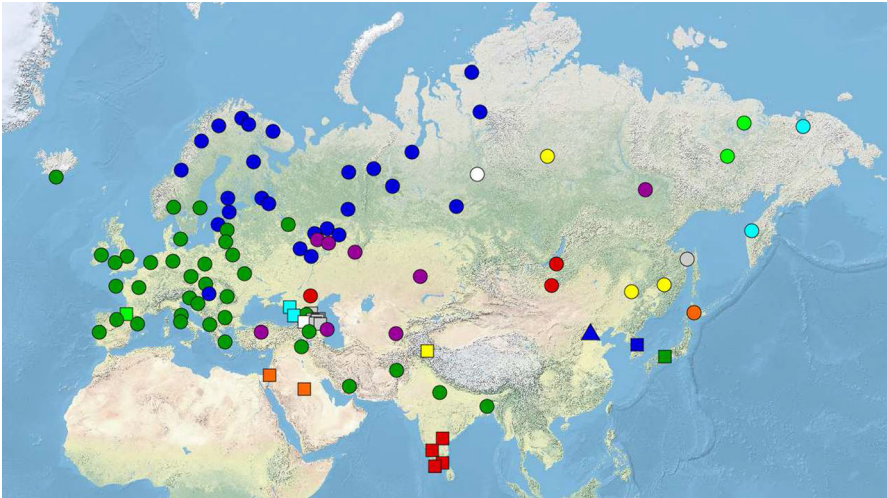
The remainder of this paper is divided into three main sections. The purpose of Sect. 2 is to describe and motivate our choices for the languages we included, and the procedure by which we arrived at our concept list. Our technical design decisions as well as the data collection procedure are detailed in Sect. 3, which also includes a general description of our approach to generating IPA transcriptions. Section 4 describes our initial evaluation of data quality for a sample of six languages with the help of native speakers, discussing and providing figures on the frequency of different types of errors. The paper concludes with an overview of our ongoing expansion efforts as well as plans for future development.

## 2 Scope

### 2.1 Language sample

The focus of NorthEuraLex lies on Northern Eurasia, more precisely on the Uralic family and surrounding languages. A prime reason for this choice is that these languages have already been extensively documented and analyzed, so there are enough sources available to find equivalents of the concepts in our long list largely without the help of experts or native speakers. The comparatively small size of Uralic as opposed to e.g. Indo-European, as well as the limited number of language families which have historically been in contact with Uralic, makes it feasible to achieve complete coverage for Uralic and all relevant neighboring families. Having a more or less complete sample of a family and its contact languages provides an ideal data set for computational models of lexical influence, which has been the primary use case of the database within our project.

In addition to all the 26 Uralic languages for which sufficient lexical resources were initially available to us, we have collected data for all language families of the area, which includes Indo-European, Turkic, Mongolic, Tungusic, Korean, Japanese as well as all the five Paleo-Siberian and the three Caucasian families, plus prominent isolates like Basque or Burushaski. In the course of the data collection process, we also started to include samples from some adjacent families outside Northern Eurasia, such as Dravidian, Afro-Asiatic, and Eskimo–Aleut. Overall, our first release version contains data for 107 languages from 21 families, which we are in the process of expanding substantially over the coming years. Table 2 in Appendix A lists all the languages included in NorthEuraLex 0.9, and our coverage of the Eurasian continent is visualized (with one symbol per language family) in Fig. 1. This initial choice of languages represents a pragmatic sample based on giving preference to large languages for which dictionaries are readily available, as well as to particularly interesting languages like language isolates or members of very small families. We estimate that a total of 350 languages and language varieties could (and are intended to) eventually be included in NorthEuraLex.

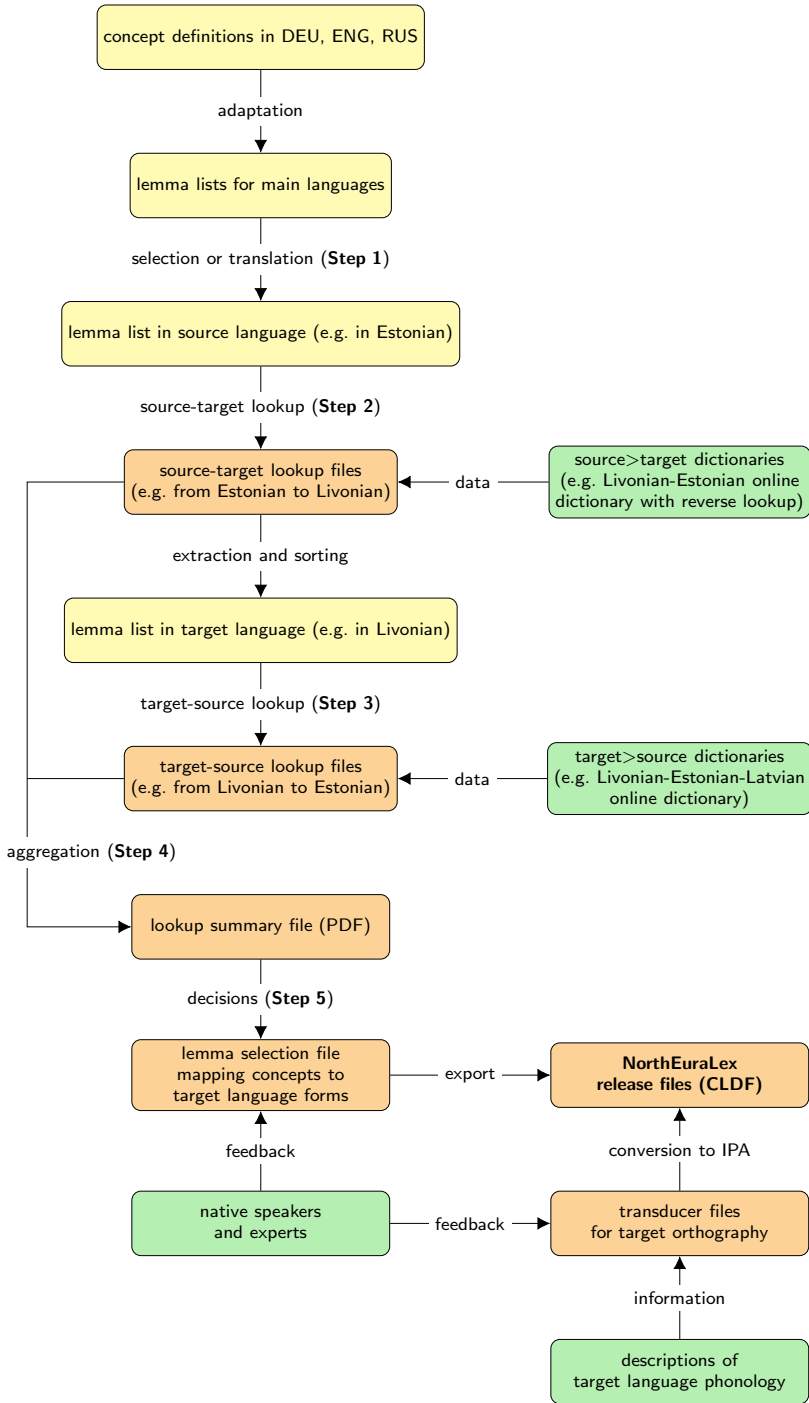


**Fig. 1** Map of languages in NorthEuraLex 0.9 (except Eskimo-Aleut). Each language family is encoded by a different combination of color and shape. (Color figure online)

## 2.2 Concept selection

To decide for which concepts to collect data, we could have adopted one of two extant long lists of basic concepts, i.e. the list used by the IDS dictionaries, or the Loanword Typology meaning list of the World Loanword Database (Haspelmath and Tadmor 2009), which is designed to find patterns of borrowability in a range of different semantic domains. The problem with both lists is that they include many concepts which are not current to indigenous cultures of our area, or at least difficult to find in the available lexical resources. This would have compromised our goal of a gapless database too much. Moreover, the two lists are only very distantly and informally grounded in data. Tracing back their history, one finds that the WOLD list is an extended version of the IDS list, which in turn is a revision of a list developed by Buck (1949) for the purpose of tracing the diachrony of Indo-European synonyms. The two existing long concept lists are therefore not only not independent, but they are very likely to share a bias towards including concepts which are of specific relevance to Indo-European languages only.

For these reasons, our database is based on a new concept list which has an empirical basis in a subset of the relevant languages. In the initial stage of the data collection project, we used a pre-existing private database of about 20,000 entries each from 12 major languages of Eurasia (with German as the common gloss language) which were collected by the first author for the purposes of language learning, along with data from a sample of five languages for which only dictionaries from a series of Soviet-era school dictionaries are available. These dictionaries, such as Menovščikov (1988) for Siberian Yupik and Volodin and Halojmova (1989) for Itelmen, are sometimes the only published lexical resource



**Fig. 2** Data collection and processing workflow for NorthEuraLex 0.9. Green nodes represent our sources of information, yellow nodes stand for informal auxiliary files, and orange nodes represent data files in machine-readable standardized formats. (Color figure online)

for the languages in question. Based on these data, our task was to choose a list of 1000 concepts that are both old enough to be of value for historical linguistics and relevant for the geographical area, while still being realistically extractable from available dictionaries.

On this initial dataset, we computed an early version of the basicness score later presented in Dellert and Buch (2018), which combines a cross-linguistically applicable measure of form simplicity with a form-distance based measure of stability. While the form simplicity measure could be applied to any orthography which encodes pronunciation, our form distance measure presupposes uniform encoding across languages, and we derive it from automatically derived phonemic IPA transcriptions. The simplicity measure computes the information content of each dictionary form in a language-specific way, generalizing word length by correcting for differences in sound systems, and morphological material in the dictionary forms. The stability measure builds on a definition of pairwise word form distances that takes some types of sound correspondences into account, and is an estimate of the correlation between the pairwise distances between the realizations of the concept in question and the aggregate language distances across all concepts.

From the top-1000 list in the resulting rough ranking of about 6000 concepts described by sets of German lemmas, we manually removed some near-synonyms as well as a range of concepts which could not be found in several of the minority-language dictionaries. This did away with many concepts that tend to be expressed by loans from Russian in the minority languages, often covering all aspects of life that are not tied to the traditional cultures. Also, since our concepts were not sufficiently disambiguated for automated cross-language lookup, some pairs of concepts turned out to be practically synonymous. For example, our initial ranking included one concept described by the German glosses *Boot* “boat” and *Kahn* “barge”, and a second concept described by the glosses *Boot* “boat” and *Schiff* “ship”. While the first concept could be used to refer to a smaller vessel than the second one, maintaining a clear separation of such concepts while mapping lexical entries across many gloss languages would be very difficult. Faced with many such examples, we decided to remove most of the concept pairs leading to many duplicate entries in our automated extraction, while keeping some of them in because of our impression that these concepts are more clearly separated in relevant languages than they are in our primary gloss languages Russian, English, or German. One such instance is the separation of the concept of understanding into being able to hear what is said, and grasping the meaning of the message. In a final step, we added some very frequently borrowed concepts, such as days of the week and month names, as simple test cases for loanword detection methods.

The resulting concept list can be separated into 480 nominal concepts which are expressed as nouns in the vast majority of the world’s languages, and 340 verbal concepts. A third category contains 102 qualities which are expressed by adjectives in English and many other languages where this category exists, and by verbs in



Italian	arcobaleno	arkobaleno	validate
Iselman	таг'ам	taq'am	validate
Japanese	虹	nikō	validate
Kalashit sut	nenusaq	nenusaq	review
Kalmyk	сонгъ	soŋ'ŋhē	validate
Kannada	ಇಳಿಮೆಣಸು	ilijemēṣa	validate
Kazakh	көкпарқосақ	kōmparqosaq	validate
Ket	улуот	uluot	validate
Khalika Mongolian	сонгоно	soŋ'ono	validate
Kildin Sami	тӳрмӧсӧдӧк	tōrmōsōdōk	validate
Komi-Permyak	ӧшӧк	oŝ'ok	validate
Komi-Zyrian	ӧшӧк	oŝ'ok	validate
Korean	무지개	mujiŋae	validate
Lak	сӱрӱн кӱртӱа	suriŋ kurt'a	review

**Fig. 3** Sample of words for ‘rainbow’ in the web interface

some others (such as Korean, and to a certain extent Chinese). A final category contains 94 additional concepts of miscellaneous types, such as pronouns, simple adverbs, numbers, and some spatial relations.

Our full list is given in Table 3 in Appendix B, grouped into 36 semantic categories in order to make it easier to get an impression of the overall structure and coverage. The annotations are sometimes needed to distinguish word senses in English (e.g. ‘to show (*let someone see*)’ vs. ‘to show (*be visible*)’), and often in order to select one sense as basic if languages make lexical distinctions within the region of semantic space denoted by the English term. For instance, many languages lexically distinguish the substance one breathes from the space in which birds move, making it necessary to further specify the English equivalent “air”. For some concepts (especially kinship terms), the specification also defines which term is to be mentioned first in our list of equivalents if a more fine-grained distinction is made in the target language, e.g. ‘grandfather (*e.g. father’s father*)’.

The overlap of our 1016-concept list with the 1329-concept IDS list is only 582. In contrast, the only items missing from the WOLD-derived 100-item Leipzig-Jakarta list of stable concepts (Tadmor 2009) are “in” (which is frequently expressed by case in North Eurasian languages), “to grind” (which is not as prominent in the absence of agriculture), and “to suck” (which is difficult to find in Russian dictionaries, presumably for taboo reasons). Of the 207 items contained in the original Swadesh lists (Swadesh 1952, 1955), the starting point of and still a popular choice in lexicostatistics, 184 concepts are also on our list. The missing concepts are predominantly from the cultural sphere (which implies low stability), tool names and agricultural terms, as well as animal and plant names that are not relevant for traditional cultures of Northern Eurasia. These differences highlight that it makes sense not to use one of the existing lists that are tuned towards Indo-European languages, but to create concept lists by means of reproducible methods



on a broad sample of languages relevant to the region. While the different sources agree on a core of about 100 items of very stable vocabulary, it is worthwhile to put more effort into deciding on a concept list for a wide-coverage lexical database.

### 3 Data collection and processing

In this section, we describe our data collection and processing workflow. In this description, we will frequently use the terms ‘source language’ and ‘target language’. Unlike in lexicography, the target language will always be the language for which we want to collect the data, and the source languages are the languages in which dictionaries into the target language are available. These names remain the same even if we refer to data from a dictionary that translates target language lemmas back into one of the source languages.

#### 3.1 Design decisions

Many similar large-scale databases (e.g. IDS or WOLD) rely on language experts or native speakers for their data. While this ensures the quality of the result, it is often difficult to get enough experts involved for getting good coverage, and there are only few or no experts or native speakers for many of the smaller languages. Because of this, we decided to initially collect all of the data ourselves in a small team from written sources, and to only ask experts and native speakers later for confirmation and help with missing and unclear data points.

Of course, this means that our initial version is not perfect and still requires corrections and revisions. Also, documentation for many languages is sparse, especially for verbal concepts, and we sometimes had to rely on small individual or old and possibly outdated sources. At times, two sources for the same language employed different orthographies which had to be bridged to maintain compatibility. These orthographies would sometimes not adequately render the language’s phonology, and phonological information was often missing or presented in a variety of different transcription schemes.

As primary data sources, we rely exclusively on published dictionaries on paper or in digital formats, from which information is typically extracted manually by going through all of the relevant entries. We initially experimented with using OCR for pre-processing, but found that OCR models perform very poorly especially on bilingual dictionaries, because they tend to be trained on running text for a single language, and do not know how to handle the very peculiar formatting and mixed-language nature of scanned dictionary pages. Post-processing the output of OCR turned out to be just as time-consuming as typing in the relevant information, especially because the extraction process involves various standardization steps, such as normalizing part-of-speech annotations and domain labels.

All lexical items are extracted in the native orthography whenever possible to preserve the information provided by the sources, and to enable automated integration of comparable data from different sources. A phonetic representation is later inferred automatically from the orthography, or stored in addition where

necessary (see Sect. 3.3). In all cases, we took the pragmatic approach of using the form used in our dictionary sources to translate our source language lemmas. Typically no attempt was made to reduce lemmas to stems, or to normalize forms in order to represent underlying representations. Only when different sources used different quotation forms we worked into the grammatical system of the respective language, and looked up or derived the desired form in order to ensure consistent treatment. For instance, in the case of qualities in Korean, we opted for using the present determiner form instead of the non-past indicative used for other verbs.

Compared to projects which are composed out of individual expert contributions, our method has the advantage that we are familiar with all the data, and have a good overview over the current status of our word lists, allowing us as the data managers to confidently implement corrections and additions that would require too much long-term commitment from expert contributors. Because we do not solely rely on external help, we can achieve complete coverage of the desired languages and were able to quickly progress in our initial data collection steps. The vast majority of our entries are filled, and the current version of our database, even though not yet reviewed by experts, is already fully functional for its intended purposes.

### 3.2 Data collection procedure

During the 4 years (2013–2017) that the data collection process lasted, we have developed systematic data collection procedures tuned towards mass processing of dictionary sources by non-experts in the respective target languages. Since to our knowledge, this type of data collection has never before been performed on such a scale, we describe what have shaped up to be our best practices in some detail here. The discussion of our five-step process, which is also outlined in a workflow diagram in Fig. 2, also describes the motivation behind some of our non-obvious design decisions.

When selecting our lexical resources, for target languages where several dictionaries were available, we did not base our decision on the immediate accessibility of the respective source language, but on a preference for work describing the standard language. For many Uralic languages, this implied not to rely on scientific dictionaries in German or English (as seems to be common practice in Western Europe), but on Russian dictionaries of the sometimes very recently established standard languages. Moreover, we always preferred lexical resources where both translation directions were developed independently to resources where one of the two directions is only available as an index or a mechanical inversion of the translation pairs in the other.

(1) *Create lookup lemma list* In order to look up the best equivalents of our 1016 concepts across all the languages of Northern Eurasia, we had to bridge different source languages, for which we needed to develop lookup lemma lists based on our original list. This original list is in German, which became the main language of the database due to the nature of our pre-existing data, and because we found it to be much more efficient and less error-prone to write dictionary entries in the native language of most project members. While all the major source languages we needed to bridge were also target languages in our database, it would not have been optimal

to simply re-use the prepared wordlists as lemma lists for lookup, because some most natural choices for a given concept are polysemous, while much more explicit alternative glosses exist. For instance, the concept of shedding tears is most naturally expressed by the verb “to cry” in contemporary English. However, especially in older dictionaries, the translations one finds under “cry” are centered around the concept of shouting, whereas the target concept can much more reliably be looked up under “weep”. A very difficult challenge faced us in the selection of the best Russian equivalents for basic verbs, especially in the domains of movement and manipulation. It is largely unpredictable which of the many corresponding Russian verbs, which lexicalize slight differences in grammatical aspect and other meaning components, is used for this purpose in small dictionaries. In larger dictionaries, the nuances expressed by the different verbs are often quite faithfully, but unnaturally modeled by regular derivational morphology in the target language. Comparing different sources and getting an impression of the most common practices, in most cases we decided to use the perfective forms (often with a disambiguating prefix) as the most useful Russian lookup lemma. Many such considerations were involved in the development of our lookup lemma lists for English, German, and Russian. These three we consider the official gloss languages of the database because together, they provide enough sources to cover the bulk of North Eurasian languages. In addition to these three primary languages, our intention to only use the best available resources for each language created a need for lookup in a surprisingly large number of smaller languages. For these smaller languages, we did not produce independent lookup lemma lists as for the major languages, but took the less ideal decision of relying on the previously collected data for the source language instead. This was our strategy for the following source languages: Norwegian (for the Western Saami languages), Swedish (for some information on South Saami), Finnish (for Inari Saami and Skolt Saami), Estonian and Latvian (for Livonian), Hungarian (for some information on Northern Mansi and Nganasan), French (for Breton), Japanese (for some information on Ainu), and Chinese (for some information on Manchu).

(2) *Lookup in source–target direction* The second step is to look up all the lemmas in a source–target dictionary (e.g. Estonian–Livonian) and digitalize the relevant parts of the target entries for each lemma as faithfully as possible while adapting them to our internal formats. This means that all equivalents are stored in the order in which they appear in the source, annotations given in the source (e.g. abbreviations such as *fig.* for ‘figurative’, and disambiguating information such as prototypical objects for verbs) are extracted, and care is taken to distinguish the separators between different senses of a lemma (frequently a semicolon) from those between alternative translations (frequently a comma). Working with a small team of data collectors coordinated by a single person made it possible to ensure largely consistent and standardized annotations across languages, a prerequisite for the later Step 4 in which all the extracted information is pre-processed automatically. Representation in the original standard orthographies ensures ease of automated retrieval, and compatibility across different resources.

(3) *Lookup in target–source direction* The third step is the reverse lookup stage, where all lemmas in the target language that were collected in the previous step,

which can amount to several thousand depending on the size of the dictionary, are looked up in a target–source dictionary (e.g. Livonian–Estonian). The lookup list for this step is produced by automatically inverting the completed lookup list from Step 2, and sorting it alphabetically in the target language for more efficient lookup. Otherwise procedures and formats are exactly the same as in the preceding lookup step, creating mirror lists which model the lexical correspondences from the viewpoint of both languages, often making it possible to resolve possible polysemies. Usage information especially on verbs which can be extracted from example sentences found in good dictionaries, is frequently encoded in additional annotations to further enrich the decision basis.

(4) *Automated aggregation* In the fourth step, the gathered lemmas are mapped back onto the primary concept list. First, the looked-up source language translations are automatically mapped to one of the gloss languages (typically German, the native language of most contributors), which helps to ensure consistency of translations across target languages. The lemmas of the target language are then assigned to the corresponding concepts in a preliminary version of a selection file. To facilitate the work for the human collector, the system already discards some lemmas based on mismatches between their translations in the two lookup steps. However, the entire data is also compiled into a PDF summary file, which lists all possible translations for each concept, even those discarded in the selection file, together with the translations from both lookups and the annotations provided by the dictionaries. This document, containing up to 500 pages of information about the 1016 concepts in our final list, provides the data compiler with a compact view of all the relevant information to efficiently perform the subsequent lemma selection decisions for each language–concept pair.

(5) *Lemma selection* The fifth and final step consists in manually reviewing the pre-generated selection files which store the selection decisions concerning the best equivalent of each concept in the target language. While automated mapping into a gloss language is used for the automated pre-filtering and to create indices bridging different source languages, the decision process itself is always performed on the original data by a data collector with at least a good passive command of the source languages. If multiple translations seem fitting, one will be selected based on disambiguation information provided by the stored dictionary annotations, consistency across multiple dictionaries (if available), and their order in the dictionary entries (assuming that the most widely used word is listed first in the sources). The latter criterion provides another good reason to prefer school dictionaries over scientific ones, because they tend to focus on a single most natural translation, instead of trying to cover all senses, in the worst case in alphabetical order. If the dictionaries themselves do not provide enough information to make a decision, the data collector will consult other sources, such as additional dictionaries, grammars or websites. For additional example sentences, we sometimes relied on the collaborative database Tatoeba (Ho and Simon 2016), and Google phrase searches in the target language often helped us to clarify the contexts in which words are used. Image searches have proved to be particularly useful for nouns, and even the word picked by translation tools such as Google Translate (as unreliable as automated translation generally is) for a gloss language lemma in the context of a

sentence are sometimes helpful in deciding on a lemma. If no translation was found in the source dictionaries, the same additional sources are consulted to cover as many concepts as possible. Concepts for which two or more target translations are required because the target language makes more semantic distinctions than the source language (e.g. ‘older brother’ vs. ‘younger brother’) or where the available resources are not sufficient to make a decision for one term, multiple translations can also be given. These are only used sparingly, however, and one of our rules is to never use more than three translations. Each selection decision is annotated by one of four status values describing our level of certainty. The possible statuses are “Questionable” for concept-language pairs for which no information or only suspiciously-looking data from unreliable sources like Google Translate was available, “Review” for decisions that we are uncertain of, typically due to ambiguous sources, and which would need to be reviewed by an expert like a native speaker or a linguist specializing in the language. “Validate” is the status of decisions where the sources seemed quite clear and we have no evidence contradicting our choice. On decisions with this status (which comprise almost 90% of the released database), we would still like to get confirmation by experts, but we do not consider these to be a very high priority. Finally, “Validated” is the status of selection decisions that have already been checked and confirmed by experts or native speakers. To facilitate further review, all data collected in the individual steps (source–target lookup, target–source lookup, selection) is retained and archived. In addition to their key role in the generation of lookup reports, separate machine-readable files for each type of information also facilitate the automation of consistency checking. Also, these files help to remove the need to go back to the primary sources on paper during subsequent steps of the revision process.

### 3.3 Deriving phonetic representations

For computational approaches that do not start on the level of cognacy decisions, the main problem of many existing lexical databases is that they primarily focus on cognacy judgments, and that little effort is put into standardized and detailed phonetic transcriptions. The differences between the transcriptions employed by the different databases also makes it quite difficult to combine their data in order to derive larger aggregate databases with better coverage.

The Austronesian Basic Vocabulary Database, for instance, while providing quite uniform transcriptions for languages which do not have an official orthography, only contains the written forms for those languages that do. ASJP consistently uses its own transcription scheme, which however reduces the sounds of the world’s languages to 41 equivalence classes, which do not suffice to adequately transcribe central phonological distinctions in many languages. While in principle, the ASJP encoding defines diacritics which would be able to express many of these distinctions, in practice only the 41 basic symbols are used consistently. IELex provides full IPA transcription, but only for some languages, whereas it relies on a mixture of original orthographies and transliteration for others. The dictionaries contained in the IDS do not even aim to include a uniform phonetic representation

across all languages, favoring orthographic forms for most languages instead, and leaving the decision how to represent the words to the expert contributors otherwise.

To retrieve a phonetic representation of our lexical data, we developed a simple transcription system that can automatically transcribe orthographic input to IPA in Unicode with the help of language-specific conversion rules. While phonetic transcriptions for individual words are hard to come by especially for smaller and less well-documented languages, most grammars include at least an overview of the phonology, providing us with information on how words in the language are pronounced.

An obvious challenge for such an approach is that its only source of information from which everything needs to be derived are the dictionary forms. This reliance on written forms causes problems whenever the standard orthography does not fully represent pronunciation. Examples include the non-phonemic weakly voiced vowels which are not represented in the orthographies of Tundra Nenets and Skolt Saami, 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 Armenian and other languages. While substantial effort was put into predicting and implementing these phenomena whenever we became aware of them, in others we opted to reduce complexity by aiming for a phonemic notation that more closely corresponds to the orthography. 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. While some of the resulting transcriptions have a hybrid status somewhere between the phonetic and phonemic levels, the level of detail usually suffices to accurately represent distinctions which are relevant e.g. for sound correspondences. Pending the expert feedback leading to even better transcriptions, the automated transcriptions generally provide a good approximation of each word's pronunciation which at least makes uniform application of algorithms across languages feasible.

The main advantage of an automatic transcription system, even if it does not always yield a perfect output, is that expert feedback on the results does not have to be applied manually to each affected word, but can usually be integrated as a new or modified rule to systematically adjust faulty transcriptions. Also, different contributors who manually write IPA transcriptions for words in a language will unavoidably disagree on some details, whereas using an automated system ensures consistency across an entire wordlist. If there are exceptions to the language's general pronunciation rules, a very common phenomenon in loanwords, our infrastructure allows to override the automatic conversion by specifying the transcription directly in the word list. These mechanisms make our design much more flexible than the approach taken by other databases where phonetic transcriptions are considered primary data which are maintained separately. The effort required for manual revisions makes it much more unlikely that existing transcriptions will be revised and adapted if e.g. it turns out the same sound was

---

<sup>1</sup> <http://northeuralex.org>.

<sup>2</sup> <https://zenodo.org/record/1312849#.XFIP-sZCdhE>.

represented by different experts in an incompatible way. Our impression is that these difficulties are the main reason why databases of expert contributions like IDS or WOLD do not contain uniform phonetic representations.

In our system, the typical transcriptor for a language is defined by one or more plain text files containing lists of simple rewrite rules. In order to facilitate human editing of these rules, all input is first converted to X-SAMPA (Wells 1995), and these X-SAMPA transcriptions are then transformed into IPA in a second step. The rules can have the form  $sch \rightarrow S$  to model simple letter-to-sound correspondences. To represent more complex phonological processes, it is possible to specify symbol classes such as  $frontVowel = [e\ i\ \ddot{a}\ \ddot{o}\ \ddot{u}]$  and  $backVowel = [a\ o\ u]$ . These classes can then be referenced in a rule to systematically change symbols in certain environments, as in properly converting German *ch* into the *ich* and *ach* sound:  $[frontVowel]ch \rightarrow [.]C$  and  $[backVowel]ch \rightarrow [.]X$ , where  $[.]$  represents the class on the left side. There can be arbitrarily many classes in a rule and the classes can contain an arbitrary amount of symbols and strings. We found that these two rule types are sufficient to concisely model the grapheme–phoneme correspondences of most languages.

The transcriptor program tries to apply each of these rules in a given file in order and greedily consumes a substring once it has been matched by a rule. Within the same rule file, that string cannot be matched again (e.g. to convert the front and back vowels of the previous example to their X-SAMPA equivalents). However, the output of the application of each rule file serves as the input to the next, so any number of files may be chained together to achieve the desired end result. It is often practical to place each type of phonetic process in a separate file, so that the generation of the final form proceeds in logically separate steps (e.g. Icelandic *öngull*  $\rightarrow \ddot{o}Nkudl \rightarrow 9yNkYdl \rightarrow 9yNkYt1\_0 \rightarrow \ddot{a}y\eta kvtl$ ).

In the future, this simple transcription system is going to be replaced by a finite-state based system, which is an interface between our previously developed transcription rule files and the Helsinki Finite State Toolkit (HFST; Koskenniemi and Yli-Jyrä 2008). It converts the rule files to a series of regular expressions, from which HFST is able to construct the corresponding finite state transducer. Once this transducer has been created, it can be reused to quickly transcribe any input in the given language. This system is faster, especially on longer words and sentences, and could thus also be employed for transcribing whole texts. Also, the underlying transducer works independently of our system and can be distributed on its own. Additionally, HFST offers convenient tools to convert transducers into the internal formats of various other finite state tools.

These HFST-based orthography-to-IPA transducers, and the code for compiling them from our rule files, will be made publicly available as part of an additional publication, which also describes transducer development in more detail. Together with the code, we also plan to release all the transducer definition files for the 103 languages for which we have automated transcription modules from either the orthography or standard transcriptions (e.g. Persian, Pashto, Japanese, Chinese).

In order to illustrate what the result of the entire workflow looks like in the current database release, we display a sample from the table containing the words for ‘rainbow’ in Fig. 3. In this snippet, we see many different writing systems, one



language where we need to generate the IPA out of a standard phonetic transcription (e.g. Japanese), one where some additional phonetic information from the dictionary needed to be modeled (e.g. Kalmyk), and many others where the pronunciation is quite predictable from the orthography, allowing us to automate the mapping using a chain of transducers. The last column contains the mentioned status values for each selection decision, in this case implying that we are quite uncertain about the words in Kalaallisut and Lak, and would prioritize these words when getting into contact with an expert and native speaker, whereas we are reasonably certain already about our choices for all of the other words.

The preliminary release version 0.9 of NorthEuraLex has been available via the project webpage<sup>1</sup> for inspection and download since July 2017, and we are officially releasing it with this article. The web interface builds on the CLLD framework by Forkel et al. (2018a), which Pavel Sofroniev adopted for the purposes of this project. The database is licensed under a CC-BY-SA license, allowing anyone to use or extend it however they wish, as long as the original version is attributed to us by citing this article, and any extensions are published under the same terms. This policy goes a long way towards ensuring long-term availability, as evidenced by the fact that it has already been added to the Zenodo online repository in a repackaged form.<sup>2</sup> This version is in the Cross-Linguistic Data Format (CLDF) as described by Forkel et al. (2018b), a linked data format which is quickly becoming the standard for lexicostatistical databases, and will also be the release format for all future versions of our database. Along with the web interface, the CLDF specification is also best reference for readers seeking to explore the possibilities of using the NorthEuraLex database in their own research.

We cannot cite all of our primary sources in this article, but they are documented both in the web interface (in table format as well as next to the wordlists) and in the `sources.tsv` file of the CLDF release. The web interface includes a warning that our database does not represent a primary resource for any of the languages concerned, and that users who are interested in the data for a particular language, and intend to use lexical data in a context where some erroneous datapoints would lead to problems, should consult and cite these primary sources instead.

## 4 Evaluation

Due to its unconventional approach of data collection by a small group of non-experts, quality control in our database can be expected to be more difficult than in the expert-contribution model. However, looking at expert-contribution databases like IDS, it becomes clear that the quality of expert contributions, even if there are no obviously wrong entries, can suffer from misunderstandings about the concepts, or the intentions of a lexicostatistical database. For instance, experts have a tendency to provide long lists of all the words that can be used for some concept, instead of focusing on the single most salient one. Also, it can be difficult to convince an expert contributor to adhere to a cross-linguistic standard of representation, because research traditions differ vastly from family to family.

These are issues which can be handled much more smoothly in our centralized approach.

On the negative side, our data collection strategy inevitably leads to some erroneous entries if measured against the goal of retrieving the most salient or natural lexeme for each concept. Missing frequency information only aggravates the problems caused by misinterpretation of dictionary entries in some of the less familiar source languages. Better results could be achieved by considering parallel corpora. However, while small corpora for many minority languages exist, and are extremely valuable resources for many linguistic questions, they typically contain only a few thousand sentences with translations into a more widespread language. For reliable lexicographic decisions based on concordancing, each lexeme from a set of alternatives should ideally occur in dozens of example sentences, which presupposes a corpus of millions of sentences. To improve data quality in NorthEuraLex beyond its current state, it will therefore be vital to rely on the assistance of experts and native speakers of the individual languages.

While focusing on the extraction of large amounts of lexical data from written sources, our work group has been getting in contact with experts and native speakers in order to get an impression of the level of quality we achieved so far, and to predict which level of quality we can expect to be achievable in data collected by linguistically informed non-experts. For this article, we systematically analyzed all our data for a sample of six languages (Italian, Lithuanian, Ukrainian, Hungarian, Udmurt, and Japanese) from three different families. After having every wordlist checked by a native speaker, followed by a thorough discussion of each problematic concept to ensure we established the best possible equivalents, we classified each lexical item that ended up being deleted (or had to be added) into one of eight error classes. The least severe type of error, which in our sample only occurred for the Japanese data, is the choice of an uncommon orthographic variant, in this case picking the Kanji representations for some terms which are most commonly written using Katakana. The error category “wrong register” is assigned to terms which are indeed used for the concept in question, but are not part of the literary standard language, usually because they are considered colloquial or outdated. The error label “low frequency” is used for forms which can be used for the concept in question, but are used a lot less frequently than the other equivalents, and are therefore considered suboptimal choices according to our quality standards. The “added form” tag is applied whenever the correction resulted in more equivalents than were in our originally extracted data, indicating that an important equivalent was absent from our data. The “wrong form” tag is used for words with spelling mistakes, which are not very frequent because many typos are fixed due to lookup in both directions, or where we chose a correct lemma or stem, but in a suboptimal shape, e.g. when we picked the intransitive verb for ‘to break’ when a causative equivalent would have been needed. We consider these five types of errors as not very severe, because the datapoints are either not technically wrong, or unproblematic in a database that is mainly used for lexicostatistics. Three additional error categories are used for errors that we would consider more severe. The first one concerns forms that were rejected by our native speaker informants as dialectal, which only occurred for some Italian and Udmurt words in our experiment. The

**Table 1** Classification of errors on our six-language evaluation sample

Error type	ITA (%)	UKR (%)	LIT (%)	HUN (%)	UDM (%)	JPN (%)
Correct form	91.9	92.0	84.7	93.8	83.7	91.8
Orthographic variant	0.0	0.0	0.0	0.0	0.0	1.4
Wrong register	0.8	0.5	1.6	0.2	3.1	0.5
Low frequency	0.7	0.8	1.5	0.4	2.4	0.1
Added form	1.9	1.8	2.1	1.4	0.7	2.5
Wrong form	0.4	0.4	1.1	0.4	1.4	0.3
Dialectal	0.3	0.0	0.0	0.0	0.8	0.0
Imprecise match	2.4	2.9	5.3	2.8	4.4	1.7
Wrong choice	1.6	1.6	3.6	1.0	3.5	1.7

most common type of severe error is labeled “imprecise match”, indicating that while there are some contexts in which the word could be used to express the intended concept, its meaning is either too general or too specialized to qualify as a good equivalent. Finally, the “wrong choice” tag is used for words that were rejected as not fitting the concept in question under any circumstances. The results for these eight error types across our six-language sample are summarized as percentages in Table 1.

Except for the error types which only occurred in certain languages (orthographic variants and dialectal forms), the figures show similar tendencies across all types. This makes it possible to limit the discussion to the most severe error categories. In those national languages for which high-quality dictionaries were available (Italian, Ukrainian, Hungarian, Japanese), only between 1.0 and 1.7% of entries were considered clearly wrong by native speakers, which we consider a good sign. Wrong choices were typically caused by polysemies of the gloss languages which the dictionaries did not sufficiently disambiguate, or in some cases by misunderstandings in interpreting the available information. The number of suboptimal choices due to imprecise matching typically stay below three percent. The much worse figures for Lithuanian and Udmurt are explainable by differences in source quality, and a lack of readily available example sentences for validating our choices. For a typical minority language, we would expect the error rate to be somewhere in the middle between these values, with Siberian languages (due to the scarcity of resources) and Caucasian languages (due to our lack of grammatical knowledge) being the most problematic in the current version.

Concerning coverage, version 0.9 contains data for 97.8% of all language-concept pairs. For 89.6% of these entries we are quite confident that no changes will be necessary, although we would still like to validate all of the data at some point, while the other 8.2% will be prioritized for review by native speakers or experts when preparing future releases. Among our language sample, only the dataset for Udmurt, one of the larger Uralic languages of Russia, contained lexical gaps due to lack of dictionary coverage. An interesting observation was that all these gaps turned out to be difficult to fill even for a native speaker, indicating that words not

listed in the dictionaries will not be very prominent in the target languages. We expect that this situation, in which older dictionaries contain more lexical information than a native speaker will actively remember, will frequently be the case in the typical bilingual situation with a very dominant state language, especially for the many moribund languages which have ceased to function as media of everyday communication.

It is also interesting to observe how the errors are distributed across concepts. If we classify concepts by the worst errors which occurred in any of our six languages, we find that the worst three error categories only ever occur in 316 of our concepts, whereas we would expect between 348 and 375 concepts to be affected by at least one of these errors if our 447 errors of this severity were completely randomly distributed across the 1016 concepts (95% CI based on 1000 simulation runs). In contrast, the data for 511 concepts was found to be without errors across our sample of six languages, whereas we would only have expected between 435 and 472 such concepts if the total of 819 errors of all types were randomly distributed. This indicates that the errors are clustering in a set of concepts which are particularly difficult to reliably extract from dictionaries. We conclude that if needed, a higher-quality subset could be extracted by discarding the data for the most problematic concepts. Based on what we know so far, the most obvious candidates for this would be the concepts with the highest number of languages for which problems were found. The maximum of four out of the six sampled languages was reached by five concepts: the two nominal concepts ‘thigh’ (due to the polysemy of both German *Schenkel* and Russian *bedro*) and ‘business (*commercial activity*)’ (due to diverging attitudes and unwanted connotations connected with loans from English *business*), and the three verbal concepts ‘to turn out (*result, end up*)’, ‘to rise (*e.g. water level*)’ and ‘to sting (*e.g. with a needle*)’, which are all difficult to look up due to a lack of relevant examples in dictionaries.

## 5 Future development

Within our group, the database has already been used for a wide range of applications in computational historical linguistics. The part which is already cognacy-annotated provides us with a large benchmark for automated sound correspondence and cognacy detection. The development of new methods for the detection of contact events on the language level (Dellert 2017) have been another focus of our work with the database. Colexifications extracted from the lexical data have also proved worthwhile in a graph-based computational model of semantic change (Münch and Dellert 2015).

A second project phase, which has started in October 2018, is going to bring NorthEuraLex towards the next release version 1.0. For this version, we have started to collect cognacy and loanword information from etymological dictionaries, with the long-term goal of providing these annotations across the database. At the same time, expansion by 96 additional languages (many ancient languages such as Ancient Greek, Sanskrit, and Hittite, plus additional smaller Indo-European, Turkic, Mongolic, and Tungusic languages) is under way, which will allow version 1.0 to

reach almost twice the coverage of the first release version which we have described in this article. These expansion measures will be complemented by continued systematic efforts to improve on the quality of the database with the help of native speakers and experts.

Many of the infrastructure tools we developed for this project could be of value to similar projects. Currently the tools are under intensive continual development due to the ongoing second project phase, but we are already distributing parts of our code to other database projects on request. After the release of version 1.0, which is currently scheduled for some time in late 2020, work on the database will switch to an open development paradigm for the further refinement stages. This includes plans to release the source code for all of our development and data management tools under an open license, and to continue further development in an openly accessible repository in order to facilitate external contributions, and to make development versions between scheduled releases available.

In a more long-term perspective for further expansions, we are planning to add morphological information to all entries in a machine-assisted way, much as we have been doing it for the transcriptions. For this, we will build on our prototype implementation of a system which extracts stem-like segments from the dictionary forms based on information content (Dellert 2018). Since some at least rudimentary morphological analysis tools even for quite a few minority languages in our area do already exist, we will employ these tools whenever possible. We are also evaluating the potential of further improvements to our phonetic representations with the help of audio recordings. Such recordings exist even for many minority languages of our target area, which would help to resolve some questions, especially in the realm of phonotactics, that we could not answer based on the available literature describing the languages' sound systems.

## 6 Conclusion

With this article, we officially release NorthEuraLex, a new wide-coverage lexical database which aims to cover a substantial core vocabulary across all languages of Northern Eurasia. The database differs from similar endeavours by initially not building on expert contributions of wordlists for individual languages, but a systematic process for extracting the necessary lexical data from existing resources, and then only relying on expert and native speaker knowledge for continual revision and improvement. This model has the disadvantage of a higher error rate in early versions, but makes maintenance and coordination of future revisions and improvements considerably easier. This is of course only possible because in the relevant geographical region, the lexicon of most languages is well-documented by published resources, making it possible to complete much preparatory work before needing to switch to unpublished field notes or new fieldwork.

The first release version 0.9 already provides more than 100,000 words in a unified IPA encoding, covering 107 languages from 21 families. To our knowledge, it is currently the only wide-coverage IPA-encoded database spanning an entire geographical region instead of focusing on a single language family. Coverage of

the target area is still far from complete, which is why NorthEuraLex is going to grow to almost twice its current size in the next release version, and pending further funding, we are planning to expand it further in order to achieve full coverage of our inventory of 350 sufficiently well-documented languages across the entire linguistic area within the next 5 years.

**Acknowledgements** The authors would like to thank Fabrício Gerardi and Amit Vrabel for their work on languages which did not make it into the current release, Pavel Sofroniev for his work on the web interface, as well as three anonymous reviewers for many helpful comments and suggestions. In addition, we would like to thank our native speaker informants Marta Berardi, Maxim Korniyenko, Živilė Rasimaitė, Anna Bródy, Olga A. Perevozčikova, and Akari Osaka for the many hours of work they put into helping us to evaluate the quality of our database in detail.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

**Funding** This work has been supported by the ERC Advanced Grant 324246 ‘EVOLAEMP’, which is gratefully acknowledged.

## Appendix A: List of languages

See Table 2.

**Table 2** List of languages in NorthEuraLex 0.9

Family	Subfamily	Languages (with ISO 639-3 codes)
Abkhaz-Adyge	Abkhaz-Abaza	Abkhaz (abk)
	Circassian	Adyghe (ady)
Afro-Asiatic	Semitic	Standard Arabic (arb), Modern Hebrew (heb)
Ainu	Hokkaido-Kuril Ainu	Hokkaido Ainu (ain)
Basque	Basque	Basque (eus)
Burushaski	Burushaski	Burushaski (bsk)
Chukotko-Kamchatkan	Chukotian	Chukchi (ckt)
	Itelmen	Itelmen (itl)
Dravidian	South Dravidian	Kannada (kan), Malayalam (mal), Tamil (tam), Telugu (tel)
Eskimo-Aleut	Aleut	Aleut (ale)
	Eskimo	Central Siberian Yupik (ess), Kalaallisut (kal)

**Table 2** continued

Family	Subfamily	Languages (with ISO 639-3 codes)	
Indo-European	Albanian	Standard Albanian ( <i>sqi</i> )	
	Armenic	Armenian ( <i>hye</i> )	
	Baltic	Latvian ( <i>lav</i> ), Lithuanian ( <i>lit</i> )	
	Celtic	Breton ( <i>bre</i> ), Irish ( <i>gle</i> ), Welsh ( <i>cym</i> )	
	Germanic		Danish ( <i>dan</i> ), Dutch ( <i>nld</i> ), English ( <i>eng</i> ), German ( <i>deu</i> ), Icelandic ( <i>isl</i> ), Norwegian Bokmål ( <i>nob</i> ), Swedish ( <i>swe</i> )
		Graeco-Phrygian	Modern Greek ( <i>ell</i> )
		Indo-Aryan	Bengali ( <i>ben</i> ), Hindi ( <i>hin</i> )
	Iranian		Northern Kurdish ( <i>kmr</i> ), Pashto ( <i>pbu</i> ), Ossetian ( <i>oss</i> ), Western Farsi ( <i>pes</i> )
		Italic	Catalan ( <i>cat</i> ), French ( <i>fra</i> ), Italian ( <i>ita</i> ), Latin ( <i>lat</i> ), Portuguese ( <i>por</i> ), Romanian ( <i>ron</i> ), Spanish ( <i>spa</i> )
	Slavic		Belarusian ( <i>bel</i> ), Bulgarian ( <i>bul</i> ), Croatian ( <i>hrv</i> ), Czech ( <i>ces</i> ), Polish ( <i>pol</i> ), Russian ( <i>rus</i> ), Slovak ( <i>slk</i> ), Slovene ( <i>slv</i> ), Ukrainian ( <i>ukr</i> )
		Japonic	Japanese ( <i>jpn</i> )
		Kartvelian	Georgian-Zan ( <i>kat</i> )
		Koreanic	Korean ( <i>kor</i> )
Mongolic		Eastern Mongolic	Buryat ( <i>bxr</i> ), Khalkha ( <i>khk</i> ), Kalmyk ( <i>xal</i> )
Nakh-Daghestanian	Daghestanian	Avar ( <i>ava</i> ), Dargwa ( <i>dar</i> ), Lak ( <i>lbe</i> ), Lezgian ( <i>lez</i> ), Tsez ( <i>ddo</i> )	
		Nakh	Chechen ( <i>che</i> )
Nivkh	Nivkh	Nivkh ( <i>niv</i> )	
Sino-Tibetan	Sinitic	Mandarin Chinese ( <i>cmn</i> )	
Tungusic	Central Tungusic	Nanai ( <i>gld</i> )	
	Northern Tungusic	Evenki ( <i>evn</i> )	
	Manchu-Jurchen	Manchu ( <i>mnc</i> )	
Turkic	Bolgar	Chuvash ( <i>chv</i> )	
	Kipchak	Bashkir ( <i>bak</i> ), Kazakh ( <i>kaz</i> ), Tatar ( <i>tat</i> )	
	North Siberian Turkic	Sakha ( <i>sah</i> )	
	Oghuz	North Azerbaijani ( <i>azj</i> ), Turkish ( <i>tur</i> )	
	Turkestan Turkic	Southern Uzbek ( <i>uzs</i> )	



**Table 2** continued

Family	Subfamily	Languages (with ISO 639-3 codes)
Uralic	Finnic	Estonian (ekk), Finnish (fin), Livonian (liv), North Karelian (krl), Olonets Karelian (olo), Veps (vep)
	Hungarian	Hungarian (hun)
	Khantyic	Northern Khanty (kca)
	Mansic	Northern Mansi (mns)
	Mari	Hill Mari (mrj), Meadow Mari (mhr)
	Mordvin	Erzya (myv), Moksha (mdf)
	Permian	Komi-Permyak (koi), Komi-Zyrian (kpv), Udmurt (udm)
	Saami	Inari Saami (smn), Kildin Saami (sjd), Lule Saami (smj), Northern Saami (sme), Skolt Saami (sms), Southern Saami (sma)
	Samoyedic	Forest Enets (enf), Nganasan (nio), Northern Selkup (sel), Tundra Nenets (yrk)
	Yeniseian	Northern Yeniseian
Yukaghir	Kolymic	Southern Yukaghir (yux)
	Northern Yukaghir	Northern Yukaghir (ykg)

## Appendix B: List of concepts

See Table 3.

**Table 3** List of concepts in NorthEuraLex 0.9, split into categories

Domain	Concepts
Animals	animal, ant, bear, bird, bull, butterfly, cat, chicken, claw ( <i>e.g. of a bird</i> ), cock ( <i>male chicken</i> ), cow, crane, crow, cuckoo, dog, duck, eagle, egg ( <i>e.g. of a bird</i> ), elk, feather, fish, flock ( <i>e.g. of sheep</i> ), fly ( <i>insect</i> ), fox, fur, gnat, goose, hare, horn, horse, lair ( <i>e.g. of a fox</i> ), louse, mouse, nest ( <i>e.g. of a bird</i> ), owl, paw ( <i>e.g. of a cat</i> ), perch, pig, pike, sheep, snake, spider, squirrel, swan, swarm ( <i>e.g. of birds</i> ), tail ( <i>e.g. of a dog</i> ), track ( <i>e.g. of an animal</i> ), wing, wolf, worm, to bark ( <i>of dog</i> ), to howl ( <i>e.g. of wolf</i> )
Artifacts	broom, bag ( <i>container for carrying</i> ), bucket, bundle ( <i>group of objects tied together</i> ), cauldron, cover/lid ( <i>of a container</i> ), cup, dishes ( <i>dishware, crockery</i> ), doll, figure ( <i>e.g. representing a god</i> ), fork ( <i>for eating</i> ), handle ( <i>part of an object which is held</i> ), hook, knife, leash ( <i>e.g. dog leash</i> ) lock, nail ( <i>spike-shaped metal fastener</i> ), net ( <i>mesh of string</i> ), picture, pot ( <i>vessel for cooking</i> ), pouch, sack, shovel, spade, spoon
Calendar	april, August, December, February, Friday, January, July, June, March, May, Monday, November, October, Saturday, September, Sunday Thursday, Tuesday, Wednesday

Table 3 continued

Domain	Concepts
Classification	angle ( <i>internal, e.g. of a room</i> ), area ( <i>extent of surface, region</i> ), circle ( <i>geometric figure</i> ), corner ( <i>external angle</i> ), count ( <i>quantity counted</i> ), cross ( <i>geometric figure</i> ), distance ( <i>between two points</i> ), edge ( <i>sharp terminating border</i> ), end ( <i>of a long object</i> ), fringe ( <i>peripheral part</i> ), gap ( <i>between objects</i> ), half ( <i>of a quantity</i> ), heap, hole ( <i>in an object</i> ), item ( <i>physical object</i> ), line ( <i>geometric figure</i> ), matter ( <i>affair</i> ), middle ( <i>central part of something</i> ), part ( <i>fraction of a whole</i> ), pattern ( <i>regular repeated elements</i> ), piece ( <i>separable part of a larger whole</i> ), row ( <i>line of objects</i> ), side ( <i>of an object</i> ), sort ( <i>type</i> ), space ( <i>available space, e.g. in a closet</i> ), thing ( <i>distinct entity</i> ), tip ( <i>of a pointy object</i> ), to alter ( <i>make different</i> )
Clothing	belt ( <i>clothing</i> ), blanket ( <i>for sleeping</i> ), boot ( <i>heavy shoe covering part of the leg</i> ), button ( <i>fastener on clothes</i> ), cap ( <i>head covering</i> ), clothes, cloth ( <i>woven fabric</i> ), collar ( <i>e.g. on a coat</i> ), knot ( <i>looping of string</i> ), leather, needle ( <i>for sewing</i> ), paint ( <i>substance for colouring</i> ), pillow ( <i>for sleeping</i> ), ribbon ( <i>strip of cloth</i> ), ring ( <i>jewellery</i> ), scarf ( <i>light, worn around the shoulders</i> ), shirt, shoe, sleeve, string ( <i>made from twisted threads</i> ), thread, to dye ( <i>e.g. hair</i> ), to get dressed, to knit ( <i>e.g. a jacket</i> ), to put on ( <i>e.g. one's coat</i> ), to sew, to take off ( <i>e.g. one's coat</i> ), trousers, wool
Crafts	ash ( <i>solid remains of a fire</i> ), board ( <i>long, wide and thin piece of wood</i> ), clay ( <i>ductile material</i> ), coal ( <i>combustible substance</i> ), glass ( <i>transparent substance</i> ), gold ( <i>metal</i> ), iron ( <i>metal</i> ), noose ( <i>e.g. on a rope</i> ), pipe ( <i>hollow conduit, tube</i> ), pole ( <i>long and slender piece of wood</i> ), sand ( <i>material</i> ), silver ( <i>metal</i> ), slab ( <i>flat piece of solid material</i> ), staff ( <i>long straight stick</i> ), stick ( <i>piece of wood used as a tool</i> ), strap ( <i>e.g. strip of leather</i> ), support/rest ( <i>something which keeps upright</i> ), wood ( <i>material</i> ), to build ( <i>e.g. a house</i> ), to make ( <i>create, bring about</i> ), to produce ( <i>manufacture</i> ), to repair/mend ( <i>e.g. a vehicle</i> )
Emotions	dear ( <i>beloved</i> ), desire ( <i>to have something</i> ), grief ( <i>sorrow, sadness</i> ), happiness, joy ( <i>emotion</i> ), laughter, sad ( <i>emotion</i> ), wish ( <i>longing</i> ), to be afraid of ( <i>fear</i> ), to be afraid ( <i>be frightened</i> ), to be annoyed ( <i>be irritated</i> ), to cry ( <i>weep, shed tears</i> ), to flee ( <i>run away</i> ), to grumble ( <i>complain in a surly manner</i> ), to laugh, to like ( <i>favor; be in favor of</i> ), to love ( <i>have a strong affection for</i> ), to move ( <i>arouse the feelings of</i> )
Housing	bed, box, bridge, chair, cradle, door, fence, floor ( <i>supporting surface of a room</i> ), gate, home ( <i>one's own dwelling place</i> ), house ( <i>building</i> ), ladder, pasture, path ( <i>trail used by pedestrians</i> ), road ( <i>way for travelling by vehicle</i> ), roof, shelf, stove ( <i>heater</i> ), table, town, village, way ( <i>connection between places</i> ), well ( <i>source of water</i> ), window, to dwell
Human body	arm, back, beard ( <i>generic</i> ), belly, blind, blood ( <i>fluid</i> ), body ( <i>of a living organism</i> ), bone, bosom ( <i>female breast</i> ), brain, breast ( <i>e.g. of a man</i> ), breath ( <i>breathing</i> ), cheek, chin, deaf, dream ( <i>while sleeping</i> ), ear, elbow, eye, face, fat ( <i>person</i> ), fever, finger, fingernail, foot, forehead, hair ( <i>on head</i> ), hand, head, health, heart, heel, hunger ( <i>condition</i> ), hungry, ill, illness, jaw, knee, leg, lip, liver, medicine ( <i>drug</i> ), moustache, mouth, nail, naked, nape ( <i>back side of neck</i> ), navel, neck ( <i>from outside</i> ), nose, pain, palm, pretty, shoulder, sinew, skin, sleep ( <i>condition</i> ), slim, stomach, strength ( <i>being strong</i> ), strong ( <i>tough</i> ), tear ( <i>drop of fluid</i> ), thigh, throat ( <i>from inside</i> ), toe, tongue, tooth ( <i>e.g. incisor</i> ), vein, weak ( <i>feeble</i> ), wound, to ache ( <i>e.g. of head</i> ), to be ill, to blow ( <i>using the mouth</i> ), to breathe, to cough, to drink, to eat, to fall asleep, to fall ill, to get tired, to get up ( <i>rise from one's bed</i> ), to groan ( <i>e.g. in pain</i> ), to recover ( <i>from an illness</i> ), to relax ( <i>recreate</i> ), to shout, to sleep, to swallow, to tremble ( <i>e.g. with fear</i> ), to wake up, to whistle

Table 3 continued

Domain	Concepts
Hygiene	comb ( <i>implement for grooming</i> ), filth ( <i>that which soils or defiles</i> ), mirror, to clean ( <i>e.g. using a brush</i> ), to comb, to decorate ( <i>e.g. a house</i> ), to have a bath ( <i>e.g. in a river</i> ), to rinse ( <i>e.g. clothes, dishes</i> ), to sweep ( <i>e.g. a room</i> ), to wash ( <i>e.g. clothes</i> ), to wash <i>oneself</i> , towel ( <i>used for wiping</i> ), to wipe
Kinship	aunt ( <i>e.g. father's sister</i> ), boy ( <i>male child</i> ), brother ( <i>e.g. elder brother</i> ), child ( <i>e.g. 10 years old</i> ), daughter, family ( <i>group of close relatives living together</i> ), father, girl ( <i>female child</i> ), grandfather ( <i>e.g. father's father</i> ), grandmother ( <i>e.g. father's mother</i> ), human ( <i>human being</i> ), husband, man, mother, parents, sister ( <i>e.g. elder sister</i> ), son, uncle ( <i>e.g. father's brother</i> ), wife, woman
Language	call ( <i>appeal</i> ), fairy tale, language, message, name ( <i>of a person</i> ), news, puzzle ( <i>riddle</i> ), sign ( <i>object bearing a message</i> ), song, speech ( <i>long oral message</i> ), story ( <i>oral narration</i> ), talk ( <i>conversation</i> ), voice ( <i>sounds uttered by vocal cords</i> ), word ( <i>unit of language</i> ), to ask ( <i>question</i> ), to brag, to call ( <i>someone to come</i> ), to chat ( <i>have a conversation</i> ), to convey ( <i>e.g. a message</i> ), to name ( <i>give a name to</i> ), to request, to say ( <i>e.g. utter</i> ), to speak ( <i>produce utterances</i> ), to talk ( <i>exchange information</i> ), to tell, to translate ( <i>e.g. text</i> )
Manipulation	force ( <i>which is applied to produce an effect</i> ), to add, to bend ( <i>e.g. a pipe</i> ), to bite ( <i>e.g. a child, of a dog</i> ), to bow ( <i>e.g. the arm</i> ), to break ( <i>e.g. a plate</i> ), to bring, to burn ( <i>e.g. paper, wood</i> ), to carry ( <i>e.g. a box</i> ), to catch ( <i>e.g. a ball</i> ), to choose ( <i>between options</i> ), to close ( <i>e.g. a window</i> ), to connect ( <i>join, e.g. using strings</i> ), to cover ( <i>e.g. a boat using a canvas</i> ), to cut off ( <i>e.g. a piece of cake</i> ), to cut ( <i>e.g. paper</i> ), to damage ( <i>e.g. a house, a vehicle</i> ), to destroy ( <i>e.g. a village</i> ), to dig ( <i>in earth</i> ), to divide ( <i>split into parts</i> ), to drag ( <i>e.g. the trunk of a tree</i> ), to drop ( <i>let fall from one's hands</i> ), to fill ( <i>e.g. a glass with water</i> ), to get ( <i>obtain, e.g. an important item</i> ), to glue ( <i>e.g. a poster to a wall</i> ), to grab ( <i>e.g. someone's arm</i> ), to hang up ( <i>e.g. a picture</i> ), to hide, to hit ( <i>on purpose, e.g. a table</i> ), to hold ( <i>in one's hands</i> ), to jog ( <i>e.g. on a door</i> ), to keep ( <i>e.g. food, money</i> ), to knock ( <i>e.g. on a wall</i> ), to lay ( <i>e.g. money onto a table</i> ), to leave ( <i>e.g. a coat at home</i> ), to lick ( <i>e.g. a wound</i> ), to lift ( <i>e.g. a box</i> ), to light ( <i>e.g. a candle</i> ), to lose ( <i>e.g. a key</i> ), to open ( <i>e.g. a window</i> ), to pick up ( <i>from the floor</i> ), to place ( <i>cause someone to sit</i> ), to pour ( <i>e.g. water into a glass</i> ), to preserve, to press ( <i>exert weight or force against</i> ), to pull ( <i>e.g. a rope</i> ), to push ( <i>shove, e.g. a box</i> ), to put ( <i>in upright position, e.g. a glass</i> ), to raise ( <i>e.g. one's hand</i> ), to rub ( <i>e.g. glass with a rag</i> ), to scrape ( <i>e.g. a plate using a knife</i> ), to seize ( <i>grab, e.g. a knife</i> ), to select ( <i>pick one option</i> ), to send ( <i>e.g. a letter</i> ), to shake ( <i>e.g. a bottle</i> ), to sharpen ( <i>e.g. a knife</i> ), to spin ( <i>e.g. a skewer</i> ), to spread ( <i>e.g. fat on a surface</i> ), to stick ( <i>e.g. a twig into a hole</i> ), to sweep ( <i>e.g. leaves</i> ), to swing ( <i>e.g. a flag, cloth</i> ), to take, to tear off ( <i>e.g. a doll's arm</i> ), to tear ( <i>e.g. a cloth</i> ), to throw ( <i>e.g. a ball</i> ), to tie ( <i>e.g. using a rope</i> ), to touch ( <i>make physical contact with</i> ), to turn around ( <i>e.g. palm</i> ), to turn over ( <i>e.g. pancakes, pieces of meat</i> ), to wrap ( <i>e.g. a fish in paper</i> )
Mathematics	a hundred, a thousand, amount ( <i>measurable quantity of a material</i> ), eight, eighty, eleven, fifty, first, five, forty, four, height, last ( <i>in a row</i> ), length, nine, ninety, one, second, seven, seventy, six, sixty, size ( <i>dimensions of an object</i> ), ten, third, thirty, three, twelve, twenty, two, weight ( <i>mass as measured</i> ), to calculate ( <i>reckon</i> ), to count ( <i>determine the number of</i> ), to decrease ( <i>of quantity</i> ), to increase ( <i>of quantity</i> ), to measure ( <i>e.g. ascertain the height of</i> )
Modality	can ( <i>be able to</i> ), to do ( <i>e.g. "what are you doing?"</i> ), to finish ( <i>e.g. a piece of work</i> ), to start ( <i>initiate, e.g. a quarrel</i> ), to stop ( <i>stop the current activity</i> ), to succeed ( <i>be successful</i> ), to turn out ( <i>result, end up</i> ), to want ( <i>desire</i> )

Table 3 continued

Domain	Concepts
Movement	to arrive ( <i>reach one's destination</i> ), to budge ( <i>change posture</i> ), to climb ( <i>of human</i> ), to come back ( <i>return to a place</i> ), to come ( <i>move here</i> ), to creep ( <i>of human</i> ), to dash ( <i>e.g. of a vehicle</i> ), to depart ( <i>set out on a journey</i> ), to dive ( <i>of human</i> ), to drop ( <i>e.g. water level</i> ), to enter ( <i>e.g. a building, a room</i> ), to escape ( <i>get away, e.g. from captivity</i> ), to fall ( <i>swiftly move downwards</i> ), to flow ( <i>e.g. a river</i> ), to fly ( <i>e.g. of bird</i> ), to go ( <i>move through space</i> ), to go away ( <i>leave a place</i> ), to go in ( <i>go into an enclosed space</i> ), to hurry ( <i>do things quickly</i> ), to jump ( <i>of human</i> ), to move ( <i>change place</i> ), to revolve ( <i>e.g. of a wheel</i> ), to rise ( <i>e.g. water level</i> ), to run ( <i>of human</i> ), to rush ( <i>move in haste</i> ), to seesaw ( <i>use a seesaw</i> ), to sink ( <i>e.g. stone in water</i> ), to sit down ( <i>assume a sitting position</i> ), to stand up ( <i>e.g. from a chair</i> ), to stay ( <i>remain in a place</i> ), to step ( <i>move the foot in walking</i> ), to stop ( <i>cease moving</i> ), to stream ( <i>to flow in a continuous manner</i> ), to sway ( <i>move backward and forward</i> ), to swim ( <i>of human</i> ), to swing ( <i>repeatedly move from side to side</i> ), to take a walk, to tumble ( <i>fall end over end</i> ), to walk ( <i>move on the feet</i> )
Nature	air ( <i>where birds fly</i> ), bay ( <i>small coastal inlet</i> ), brook ( <i>small river</i> ), cave ( <i>in a mountain</i> ), chill ( <i>low temperature</i> ), cloud ( <i>bright</i> ), coast ( <i>seashore</i> ), current ( <i>of a river</i> ), dirt ( <i>on objects</i> ), drop ( <i>e.g. water</i> ), dust ( <i>settled</i> ), earth ( <i>substance</i> ), elevation ( <i>of the ground</i> ), fire ( <i>flames</i> ), foam ( <i>on liquid</i> ), fog, forest, frost ( <i>temperature below freezing point</i> ), ground ( <i>soil, earth</i> ), heat ( <i>high temperature</i> ), hill ( <i>possibly wooded</i> ), hoarfrost, ice, island, lake, land ( <i>as opposed to sea</i> ), light ( <i>from a natural source</i> ), meadow ( <i>land covered with grass</i> ), moon ( <i>celestial body</i> ), moor ( <i>wasteland covered by heath</i> ), mountain ( <i>woodless</i> ), pit ( <i>hole in the ground</i> ), rainbow, rain ( <i>falling</i> ), river ( <i>larger river</i> ), sea, shadow ( <i>shady place</i> ), shore ( <i>of a lake</i> ), sky ( <i>visible</i> ), slope ( <i>of a mountain</i> ), smoke ( <i>from a fire</i> ), snow ( <i>on the ground</i> ), source ( <i>of a river</i> ), spark ( <i>from a fire</i> ), star ( <i>in the night sky</i> ), stone ( <i>substance</i> ), summit ( <i>of a mountain</i> ), sun ( <i>celestial body</i> ), swamp ( <i>area of impassable wet land</i> ), thunder ( <i>during a thunderstorm</i> ), water ( <i>cold water</i> ), wave ( <i>on water</i> ), weather, wind ( <i>outside</i> ), to blow ( <i>of wind</i> ), to boil ( <i>of water</i> ), to break ( <i>e.g. a bone</i> ), to burn ( <i>e.g. of wood</i> ), to decay ( <i>e.g. of wood</i> ), to dry ( <i>e.g. of clothes</i> ), to freeze ( <i>to become solid due to low temperature</i> ), to melt ( <i>e.g. of ice</i> ), to rain ( <i>e.g. "it is raining"</i> ), to rise ( <i>of the sun</i> ), to rot ( <i>e.g. of meat</i> ), to set ( <i>of the sun</i> ), to shine ( <i>of the sun</i> ), to smoke ( <i>give off smoke</i> ), to snap ( <i>e.g. of a thread</i> ), to thaw ( <i>e.g. of snow</i> ), to thunder ( <i>e.g. "it thundered"</i> )
Negative interaction	damage ( <i>material harm, detriment</i> ), fault ( <i>blame, responsibility</i> ), lie ( <i>falsehood</i> ), misfortune ( <i>undesirable condition</i> ), mistake ( <i>wrong action or decision</i> ), to annoy, to beat ( <i>someone</i> ), to bother, to brawl, to cheat ( <i>e.g. someone of money</i> ), to conceal ( <i>e.g. a fact</i> ), to disturb, to hinder, to kick ( <i>someone, a dog</i> ), to kill, to leave ( <i>someone</i> ), to separate ( <i>part ways</i> ), to shove ( <i>someone</i> ), to steal, to sting ( <i>e.g. with a needle</i> )
Personality	clever ( <i>intelligent</i> ), diligent, evil ( <i>malevolent</i> ), gentle ( <i>tender</i> ), lazy, merry, skilful, stingy, stupid
Place	above, ahead, back ( <i>to the start</i> ), backward, behind, below, between, down, everywhere, far ( <i>distant</i> ), forward, hence, here, hither, in front of, left ( <i>e.g. left leg</i> ), near ( <i>close</i> ), next to, place ( <i>location, position</i> ), right ( <i>e.g. right leg</i> ), there, thither, through ( <i>e.g. a window, a hole</i> ), under, up, to hang ( <i>e.g. "the coat is hanging"</i> ), to lie ( <i>of a person</i> ), to sit ( <i>of a person</i> ), to stand ( <i>of a person</i> )
Plants	apple, bark, berry ( <i>generic term</i> ), birch, fir, flower, grain ( <i>single grain</i> ), grass ( <i>ground cover</i> ), hay, leaf, limb ( <i>of a tree</i> ), mushroom ( <i>generic term</i> ), onion ( <i>edible</i> ), peel ( <i>e.g. of an apple</i> ), pine, root, seed ( <i>fertilized grain</i> ), tree, trunk ( <i>of a tree</i> ), twig, willow, to grow, to ripen
Politics	border ( <i>between states</i> ), country ( <i>set region of land</i> ), king, power ( <i>ability to control</i> ), state ( <i>sovereign polity</i> ), to rule ( <i>e.g. country</i> )

Table 3 continued

Domain	Concepts
Positive interaction	to assemble ( <i>e.g. people in a place</i> ), to bring up ( <i>raise, e.g. a child</i> ), to cover ( <i>e.g. a child with a blanket</i> ), to dance, to give ( <i>an object, e.g. a fruit</i> ), to give ( <i>as a present, e.g. a book</i> ), to guide ( <i>e.g. to one's destination</i> ), to hand ( <i>e.g. salt at table</i> ), to instruct ( <i>e.g. children</i> ), to invite ( <i>to one's place</i> ), to kiss, to lead, to let ( <i>cause someone to</i> ), to meet, to play ( <i>e.g. of children</i> ), to praise, to promise, to receive ( <i>be given</i> ), to rescue ( <i>someone</i> ), to rock ( <i>a child</i> ), to show ( <i>let someone see</i> ), to sing, to teach ( <i>e.g. to swim, to sew</i> ), to unite ( <i>form into a group</i> ), to urge, to visit ( <i>e.g. a friend</i> ), to wake, to wish ( <i>e.g. luck</i> )
Pronouns	this, that, everything, I, thou, he, we, they, you ( <i>plural</i> )
Qualities	beautiful ( <i>e.g. flower</i> ), big ( <i>e.g. rock</i> ), blunt ( <i>e.g. knife</i> ), clean, closed ( <i>e.g. door</i> ), cold, cool, damp, delicate/fine ( <i>e.g. yarn, web</i> ), dense ( <i>e.g. hair, forest</i> ), dirty, dry, empty ( <i>container</i> ), firm/solid ( <i>e.g. wood, wall</i> ), flat ( <i>object</i> ), fresh ( <i>e.g. air</i> ), full ( <i>container</i> ), hard ( <i>e.g. shell</i> ), heavy ( <i>of weight</i> ), hot ( <i>e.g. fire</i> ), little ( <i>e.g. rock</i> ), long ( <i>object</i> ), narrow ( <i>e.g. river, bridge</i> ), open ( <i>e.g. door</i> ), pointed ( <i>e.g. needle</i> ), powerful ( <i>forceful</i> ), raw ( <i>uncooked</i> ), ripe ( <i>e.g. fruit</i> ), round ( <i>circular</i> ), sharp ( <i>e.g. knife</i> ), short ( <i>object</i> ), smooth ( <i>surface</i> ), soft ( <i>e.g. cushion</i> ), thick ( <i>object</i> ), thin ( <i>object</i> ), warm, wet, wide ( <i>e.g. river, bridge</i> )
Questions	how, how much, what, where, where to, who, why
Religion	church, death, god, grave, life ( <i>state of being alive</i> ), living, sin, spirit ( <i>inner energy of a being</i> ), world ( <i>universe</i> ), to be born ( <i>of human</i> ), to die ( <i>of human</i> ), to live ( <i>be alive</i> ), to perish ( <i>die violently</i> )
Senses	bitter ( <i>taste</i> ), black, blue, bright ( <i>full of light</i> ), calm ( <i>absence of noise and disturbances</i> ), colourful ( <i>multicolored</i> ), dark ( <i>devoid of light</i> ), delicious ( <i>tasty</i> ), flavour ( <i>of something</i> ), green, grey, noise ( <i>unwanted loud sounds</i> ), odour ( <i>of something</i> ), red, sound ( <i>sensation perceived by the ear</i> ), sour ( <i>e.g. lemon</i> ), sweet ( <i>taste</i> ), tone ( <i>sound of a specific pitch</i> ), white, yellow, to appear ( <i>e.g. light, ghost</i> ), to be cold ( <i>feel the sensation of coldness</i> ), to be noisy ( <i>e.g. children</i> ), to chime ( <i>e.g. of bell</i> ), to disappear ( <i>e.g. light, ghost</i> ), to feel ( <i>sense</i> ), to hear ( <i>perceive with the ears</i> ), to hear ( <i>receive information about</i> ), to listen ( <i>pay attention to speech</i> ), to look at ( <i>turn the eyes toward</i> ), to ring ( <i>e.g. doorbell, phone</i> ), to roar ( <i>e.g. storm, sea</i> ), to rustle ( <i>e.g. of leaves in breeze</i> ), to seem ( <i>appear, be perceived as</i> ), to see ( <i>perceive with the eyes</i> ), to shine ( <i>e.g. of metal, shiny surface</i> ), to show ( <i>be visible</i> ), to smell ( <i>sense the smell of</i> ), to sound ( <i>e.g. of instrument</i> ), to sparkle/twinkle ( <i>e.g. of star, little light</i> ), to taste ( <i>sample the flavor of</i> ), to tinkle ( <i>e.g. glass</i> ), to watch ( <i>view for a period of time, e.g. movie</i> )
Social relations	boss ( <i>supervisor</i> ), companion ( <i>with whom one spends time</i> ), company ( <i>companionship</i> ), doctor ( <i>physician</i> ), friend ( <i>person whose company one enjoys</i> ), game ( <i>playful activity</i> ), gift, guest, help ( <i>assistance</i> ), master ( <i>expert tradesman</i> ), nation ( <i>community defined by common culture</i> ), people ( <i>many persons</i> ), teacher, worker, work ( <i>employment</i> ), to work
Subsistence	bread, butter, corn ( <i>cereal plant grown for its grain</i> ), dish ( <i>type of prepared food</i> ), fat ( <i>refined substance</i> ), food ( <i>consumable substance</i> ), honey, meal ( <i>process of food intake</i> ), meat, milk, mush, oil ( <i>liquid vegetable fat</i> ), salt, slice ( <i>e.g. of bread</i> ), soup, tea, trap ( <i>e.g. mouse trap</i> ), to bake ( <i>e.g. bread, a cake</i> ), to boil ( <i>e.g. vegetables</i> ), to catch ( <i>an animal</i> ), to chop ( <i>e.g. kale</i> ), to cook ( <i>e.g. a meal</i> ), to drive ( <i>cattle</i> ), to feed ( <i>an animal</i> ), to fish ( <i>catch fish</i> ), to fry ( <i>e.g. meat</i> ), to gather ( <i>e.g. mushrooms, berries</i> ), to herd ( <i>cattle</i> ), to hunt, to milk ( <i>a cow</i> ), to pick ( <i>e.g. an apple</i> ), to stir ( <i>e.g. soup</i> ), to tie up ( <i>an animal</i> ), to water ( <i>e.g. flowers</i> )

Table 3 continued

Domain	Concepts
Thinking	a little, alone, and, at once ( <i>in one go</i> ), bad ( <i>e.g. tool</i> ), because of, correct, different, familiar, famous, foreign, good ( <i>e.g. tool</i> ), hardly, in vain, meaning ( <i>denotation</i> ), memory ( <i>human ability to record information</i> ), mind ( <i>ability for rational thought</i> ), nasty/dire, not, only ( <i>e.g. only one cow</i> ), or, other, reason ( <i>motive, rationale</i> ), so, thought, together, true ( <i>truthful</i> ), truth, very, to await ( <i>wait for someone</i> ), to believe, to deceive ( <i>create the wrong impression</i> ), to endure ( <i>e.g. pain</i> ), to find ( <i>locate something searched for</i> ), to forget ( <i>lose remembrance of</i> ), to grasp ( <i>understand the meaning of</i> ), to improve ( <i>make better</i> ), to know ( <i>be aware of</i> ), to learn ( <i>e.g. to sew</i> ), to look for ( <i>seek, attempt to find</i> ), to notice, to prepare ( <i>make ready, set up</i> ), to recognize, to remember ( <i>recall from one's memory</i> ), to spoil ( <i>e.g. the mood</i> ), to think, to try ( <i>attempt</i> ), to understand ( <i>acoustically, e.g. a call</i> ), to wait ( <i>delay action until something happens</i> )
Time	afterwards, again, age ( <i>number of years since birth</i> ), already, always, at first, at that time, autumn, before, day, end ( <i>conclusion</i> ), evening, former ( <i>erstwhile</i> ), instantly, later, long ( <i>a long time</i> ), month, morning, never, new, night, noon, now, often, old ( <i>e.g. house</i> ), old ( <i>person</i> ), once ( <i>a single time</i> ), once ( <i>once upon a time</i> ), sometimes, soon, spring, still, suddenly, summer, then, time ( <i>constant passing of events</i> ), today, tomorrow, week, winter, year, yesterday, young, to become ( <i>turn into</i> ), to begin ( <i>start existing</i> ), to change ( <i>become something different</i> ), to end ( <i>stop existing</i> ), to happen ( <i>occur</i> ), to pass ( <i>of time</i> )
Trade	benefit ( <i>advantage</i> ), business ( <i>commercial activity</i> ), cheap, expensive ( <i>goods</i> ), money, poor ( <i>needy</i> ), precious, price, rich, shop, to buy ( <i>e.g. goods</i> ), to own ( <i>possess</i> ), to pay for ( <i>e.g. goods</i> ), to pay ( <i>e.g. in a restaurant</i> ), to sell ( <i>e.g. goods</i> ), ware, wealth
Travel	boat, campfire, east, load ( <i>on a vehicle</i> ), north, oar ( <i>rowing implement</i> ), ski, sleigh, south, step ( <i>single pace</i> ), to drive ( <i>travel by a vehicle</i> ), to row ( <i>propel a boat using oars</i> ), walk ( <i>trip made by walking</i> ), west
Warfare	arrow ( <i>projectile shot from a bow</i> ), bow ( <i>weapon</i> ), enemy ( <i>hostile person</i> ), fight ( <i>physical confrontation</i> ), gun ( <i>firearm</i> ), violence ( <i>use of force</i> ), war, to defend ( <i>e.g. city, state</i> ), to guard ( <i>e.g. a building</i> ), to protect ( <i>a person</i> ), to shoot ( <i>e.g. with a gun</i> ), to win ( <i>achieve victory</i> )
Writing	book, character ( <i>written symbol, letter</i> ), letter ( <i>written message</i> ), newspaper, stroke ( <i>drawn with a writing implement</i> ), to draw ( <i>e.g. a line</i> ), to paint ( <i>e.g. a picture</i> ), to read ( <i>e.g. a book</i> ), to write ( <i>e.g. a letter</i> )

## References

- Bouchard-Côté, A., Hall, D., Griffiths, T. L., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1204678110>.
- Bowern, C. (2016). *Chirila: contemporary and historical resources for the indigenous languages of Australia*. Language Documentation and Conservation (Vol. 10). <http://nflrc.hawaii.edu/ldc/>.
- Buck, C. D. (1949). *A dictionary of selected synonyms in the principal Indo-European languages: A contribution to the history of ideas*. Chicago: University of Chicago Press.
- Dellert, J. (2017). *Information-theoretic causal inference of lexical flow*. PhD thesis, Eberhard Karls Universität Tübingen.

- Dellert, J. (2018). *Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection*. In 27th International Conference on Computational Linguistics (COLING 2018).
- Dellert, J., & Buch, A. (2018). A new approach to concept basicness and stability as a window to the robustness of concept list rankings. *Language Dynamics and Change*, 8(2), 157–181.
- Dunn, M. (2015). Indo-European Lexical Cognacy Database. <http://ielex.mpi.nl/>.
- Dyen, I., Kruskal, J. B., & Black, P. (1992). An Indoeuropean classification A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5), iii-132.
- Forkel, R., Bank, S., Rzymiski, C., Littauer, R., Erlewine, M. Y. (2018a). cld/cldd: cldd—a toolkit for cross-linguistic databases. Zenodo. <https://doi.org/10.5281/zenodo.1436382>.
- Forkel, R., List, J. M., Greenhill, S. J., Rzymiski, C., Bank, S., Cysouw, M., et al. (2018b). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5, 180205. <https://www.nature.com/articles/sdata2018205>.
- Greenhill, S. J. (2015). TransNewGuinea.org: An online database of New Guinea Languages. *PLoS ONE*, 10, e0141563.
- Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The Austronesian Basic Vocabulary Database: from bioinformatics to lexicomics. *Evolutionary Bioinformatics*, 4, 271–283.
- Haspelmath, M., Tadmor, U. (Eds.) (2009). WOLD. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wold.cldd.org/>.
- Ho, T., Simon, A. (2016). Tatoeba: Collection of sentences and translations. <http://tatoeba.org/eng/>.
- Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, 42(3–4), 331–354.
- Kaiping, G. A., & Klamer, M. (2018). LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLoS ONE*, 13, 10.
- Key, M. R., Comrie, B. (Eds.). (2015). IDS. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ids.cldd.org/>.
- Koskeniemi, K., Yli-Jyrä, A. (2008). *CLARIN and free open source finite-state tools*. In FSMNLP (pp. 3–13).
- Menovščikov, G. A. (1988). *Slovar' èskimossko-russkij i russko-èskimosskij* (2nd ed.). Leningrad: Prosvěšćenie.
- Münch, A., Dellert, J. (2015). *Evaluating the potential of a large-scale polysemy network as a model of plausible semantic shifts*. In 6th Conference on Quantitative Investigations in Theoretical Linguistics (QITL-6). November 4–6, Tübingen, Germany.
- Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4), 452–463.
- Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 2, 121–137.
- Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the world's languages. A comparative handbook* (pp. 55–75). Berlin: Mouton de Gruyter.
- Volodin, A. P., Halojmova, K.N. (1989). *Slovar' itel'mensko-russkij i russko-itel'menskij*. Leningrad: Prosvěšćenie.
- Wells, J.C. (1995). Computer-coding the IPA: A proposed extension of SAMPA. <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>.
- Wichmann, S., Holman, E. W., Brown, C. H. (Eds.) (2016). The ASJP Database (version 17).