EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
LANGUAGE EVOLUTION:
THE EMPIRICAL TURN

NorthEuraLex: A deep-coverage lexical database of Northern Eurasia

Poznań, September 16, 2016

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle and Gerhard Jäger

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP.

Table of Contents

NorthEuraLex: Goals

Goals of our data collection project:

- cover a substantial part of the basic vocabulary in a large continuous area that spans many language families
- aim at high coverage (few gaps in the database)
- unified phonetic format

Motivation for high number of concepts:

- enough to find regular sound correspondences
- enough to make multiple layers of loans visible
- finding cognates which have undergone semantic change

NorthEuraLex: Scope

- goal: collect lexical data for all languages of Northern Eurasia
- core families: Uralic, Indo-European, Turkic, Mongolic, Tungusic, Korean, Japanese, all Paleo-Siberian and Caucasian families, plus isolates (Basque, Burushaski, ...)
- some important languages from neighboring families: Afroasiatic, Dravidian, Eskimo-Aleut
- now covering 104 languages, expansion is under way
- initial sample: Uralic and its contact languages
- a perfect version would contain data for about 300 languages (some of which are too sparsely documented)

NorthEuraLex: Current Languages

- **Uralic**: Finnic (6), Saami (6), Mordvin (2), Mari (2), Permian (3), Hungarian, Mansi (1), Khanty (1), Samoyedic (4)
- **Indo-European**: Indo-Iranian (6), Balto-Slavic (8), Germanic (6), Celtic (2), Romance (5), Greek (1), Armenian (1), Albanian (1)
- **Turkic**: Turkish, Uzbek, Kazakh, Bashkir, Tatar, Sakha, Chuvash
- **Mongolic**: Khalkha, Buryat, Kalmyk
- **Tungusic**: Evenki, Nanai, Manchu
- **Eskimo-Aleut**: Aleut, Siberian Yupik, Greenlandic
- **Afroasiatic**: Arabic, Hebrew, Coptic, Tamasheq, Hausa, Somali
- **Dravidian**: Telugu, Tamil, Kannada, Malayalam
- Abkhaz, Adyghe, Chechen, Avar, Tsez, Lak, Lezgian, Dargwa
- Ket, Yukaghir (2), Chukchi, Itelmen, Nivkh
- Korean, Japanese, Ainu
- Georgian, Basque, Burushaski

Table of Contents

Design Decisions: Selecting the Concepts

- joint work with Armin Buch (forthcoming): use automated criteria (information content, correlation of overall and concept-specific realization distance) to rank candidate concepts on the basis of available data; first version used 12 languages
- initial list manually filtered and extended to include some more concepts which are well-documented in smaller minority languages of Russia (based on a sample of five school dictionaries)
- 480 nominal and 304 verbal concepts, 102 qualities
- 94 additional concepts of miscellaneous types (pronouns, simple adverbs, numbers, some spatial relations)
- overlap with WOLD list: about 800
- Swadesh-207 and Leipzig-Jakarta are subsets

Design Decisions: Data Collection

A **five-stage process** of data collection from dictionaries:

- create list of target glosses in the relevant gloss language (e.g. Norwegian for Western Saami languages)
- look up all target glosses, create list of relevant target-language lemmas (e.g. Lule Saami)
- look up all target-language lemmas, extract glosses, semi-automatically translate into German
- compile the information into a report file, create selection file defining the map from concepts to target-language lemmas
- fill gaps by using other sources (grammars, Wikipedia, example sentences, ...)

Design Decisions: Data Collection

Challenges:

- bridging 10 different gloss languages:
  German, English, French, Norwegian, Swedish, Russian, Latvian, Finnish, Estonian, and Hungarian (so far)
- making the selection decisions based on the sparse information in some dictionaries (especially for verbal concepts)
- unifying different sources targeted at different audiences, covering different dialects, using incompatible transcription systems (e.g. the Uralic Phonetic Alphabet)
- phonetic differences not represented by imperfect orthographies

Design Decisions: Data Representation

- most recent **native orthography** whenever possible (ensuring comparability across sources)
- **dictionary forms**, not stems (easier for non-expert data collectors, and we have methods for detecting the relevant segments based on information content)
- **digitalize all lookup information** for later reference

Design Decisions: Phonetic Representation

- in principle, we are using **IPA** in Unicode
- direct specification of pronunciation in X-SAMPA is possible (and necessary for some languages), but typically rely on **automated converters** from orthography or standard transcriptions
- support for automated conversions into other formats:
  - Dolgopolsky sound classes
  - LingPy's internal model ("List classes")
  - ASJP sound classes
  - reduced versions of IPA (e.g. without coarticulations)

Design Decisions: Workflow

- in contrast to comparable efforts (e.g. IDS, WOLD), we do not rely on experts providing us with data
- instead: do the manual work in exactly the format we want, ask experts for confirmation on semi-final version
- ask native speakers or experts for help on specific points

Disadvantages:

- potentially lower-quality data in initial version
- requires working into many grammars and writing systems
- comprehensive documentation must be available

Advantages:

- faster initial progress, possibility of complete coverage
- full control over and familiarity with the data, easier to update

Table of Contents

Data Handling: Selection Decisions

The selection decisions (which lexemes to include for each concept) are made based on a combination of criteria:

- order of translations in both directions
- additional disambiguating information (e.g. argument restrictions)
- example sentences given in dictionaries
- consistency across dictionaries (if several were available)
- additional sources (textbooks, grammars, websites)
- phrase searches in the target language
- image searches (e.g. for disambiguating household items)

Data Handling: IPA conversion

- builds on text files defining simple greedy replacement rules
- each file defines one transducer pass
- grapheme-to-phoneme conversion works in several passes: Icelandic *öngull* ⇒ `öNkudl` ⇒ `9yNkYdl` ⇒ `9yNkYtl_0` ⇒ œyŋkʏtl̥
- disadvantage: a complex task, there will always be gaps in coverage which need to be fixed manually (in our database: override automated conversion by adding X-SAMPA)
- advantage: expert feedback on the transcriptions can often be applied mechanically, no need to manually edit every transcription; incremental refinement possible
- recent work by Thora Daneyko: automated conversion of our transducer files into more mainstream and highly efficient finite-state transducers, will be made publicly available

Table of Contents

NorthEuraLex: Current Status

- some data was found for **97% of all language-concept pairs**
- for 87% of selection decisions, sources were clear enough to give us some confidence that no changes will be necessary
- the remaining 10% of assignments are tentative, and need to be clarified in collaboration with native speakers and/or experts
- we have first versions of **IPA converters for all languages** where it was feasible (exceptions: English, Danish, Irish, French)

NorthEuraLex: What we are doing with it

Current applications within our project:

- sound correspondence and cognacy detection (forthcoming)
- determining the directionality of lexical flow between languages (my dissertation, to be published next year)
- loanword detection (Köllner & Dellert, forthcoming)
- models of semantic change (see e.g. Münch & Dellert 2015)

NorthEuraLex: Future

- during 2017: correcting selection decisions and filling the last remaining gaps with the help of native speakers and experts
- in progress: expansion by about 30 additional languages (mainly Indo-European and Turkic)
- in the future: further languages, with a special focus on all remaining minority languages of Russia

Conclusions

NorthEuraLex

- is a new deep-coverage lexical database which attempts to cover all of Northern Eurasia
- already provides about 100.000 words from about 100 languages spanning 20 families in a unified IPA encoding
- is subject to continual revision and improvement
- is partially available (Uralic data) to other researchers: `http://www.sfs.uni-tuebingen.de/~jdellert/northeuralex/`
- will be made publicly available in its entirety in early 2018

Acknowledgments

Thanks are due to everyone who participated in data collection:

- Thora Daneyko (student assistant)
- Alla Münch (student assistant)
- Alina Ladygina (student assistant)
- Armin Buch (postdoc)
- Natalie Clarius (student assistant)
- Ilja Grigorjew (student assistant)
- Mohamed Balabel (student assistant)
- Isabella Boga (student assistant)
- Zalina Baysarova (student assistant)
- Roland Mühlenbernd (postdoc)
- Johannes Wahle (PhD student)
- Gerhard Jäger (principal investigator of EVOLAEMP)

References

Dellert, J. (2015). Compiling the Uralic Dataset for NorthEuraLex, a Lexicostatistical Database of Northern Eurasia. First International Workshop on Computational Linguistics for Uralic Languages. January 16, Tromsø, Norway.

Dellert, J. and Buch, A. (2015). Using computational criteria to extract large Swadesh lists for lexicostatistics. Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. October 26-30, Leiden, The Netherlands.

Münch, A. and Dellert, J. (2015). Evaluating the Potential of a Large-Scale Polysemy Network as a Model of Plausible Semantic Shifts. 6th Conference on Quantitative Investigations in Theoretical Linguistics (QITL-6). November 4-6, Tübingen, Germany.