EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

EVOLAEMP
LANGUAGE EVOLUTION:
THE EMPIRICAL TURN

Current Trends: Lexicostatistical Databases

Tübingen, January 16th, 2018

Johannes Dellert

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

Table of Contents

Lexicostatistics

Lexicostatistical Databases

Data Collection Example: NorthEuraLex

Lexicostatistics: Motivation

- **historical linguistics** is about systematically analysing similarities between languages, and using them to reconstruct **proto-languages** (common ancestors)
- this is usually done on all levels of linguistic description: phonology, morphology, syntax, ...
- the **lexicon** contains the largest amount of information (the largest number of independent datapoints)
- ideally, the similarities are strong enough to perform reconstruction using the logic-based comparative method
- in most cases, probabilistic and quantitative arguments need to be made at some point, especially at high timedepths

Lexicostatistics: Glottochronology

- inspired by radiocarbon dating in archaeology
- Swadesh (1955): after measuring the ratio of shared basic vocabulary between two languages, we can compute the time of their latest common ancestor
- assumption: constant rate of lexical replacement
  (Swadesh arrives at about 14% per millennium on a list of 200 basic concepts)
- in reality: Icelandic only replaced 4% compared to Old Norse of 1000 AD, Norwegian about 20%

Lexicostatistics: Phylogenetic Inference

- do not assume constant rate of lexical replacement
- find a tree with optimal structure and replacement rates at each branch (or even better, assign a probability to each subgrouping)
- requires very sophisticated statistical techniques
- major trend: encode the basic lexicon across some language family in a machine-readable format, and use it to infer family trees
- this has recently been done for Indo-European, Uralic, Austronesian, Bantu, Pama-Nyungan, Alor-Pantar, ...

Lexicostatistics: Problems

- loanwords can cause languages to seem related which are not
- actual replacement rate seems to be more like 5% per millennium, the rest are loanwords
- need to distinguish loans from cognates to be reliable
- state of the art: manually sifting out loans before applying phylogenetic inference, or not caring about it (and claiming it doesn't make a real difference because loans are ubiquitous)
- incipient development towards phylogenetic networks instead of trees (i.e. there are lateral connections)

Table of Contents

Lexicostatistics

Lexicostatistical Databases

- Form Databases
- Cognate Databases

Data Collection Example: NorthEuraLex

Lexicostatistical Databases

A **lexicostatistical database** contains information about

- words for a set of **basic concepts**
- across **many languages**
- in a **machine-readable** format

Often included additional data:

- cognacy annotation (= common descent)
- loanword annotation

Lexicostatistical Database: Small Example (25*10)

concept	EYE	EAR	NOSE	DOG	HORSE	FISH	TWO	THREE	FOUR	NAME
Dutch	oog	oor	neus	hond	paard	vis	twee	drie	vier	naam
German	Auge	Ohr	Nase	Hund	Pferd	Fisch	zwei	drei	vier	Name
Swedish	öga	öra	näsa	hund	häst	fisk	två	tre	fyra	namn
Icelandic	auga	eyra	nef	hundur	hestur	fiskur	tveir	þrír	fjórir	nafn
Polish	oko	ucho	nos	pies	koń	ryba	dwa	trzy	cztery	imię
Czech	oko	ucho	nos	pes	kůň	ryba	dva	tři	čtyři	jméno
Croatian	oko	uho	nos	pas	konj	riba	dva	tri	četiri	ime
Latvian	acs	auss	deguns	suns	zirgs	zivs	du	trīs	četri	vārds
Lithuanian	akis	ausis	nosis	šuo	arklys	žuvis	divi	trys	keturi	vardas
French	œil	oreille	nez	chien	cheval	poisson	deux	trois	quatre	nom
Portuguese	olho	orelha	nariz	cão	caballo	peixe	dois	três	quatro	nome
Spanish	ojo	oreja	nariz	perro	cavalo	pez	dos	tres	cuatro	nombre
Italian	occhio	orecchio	naso	cane	cavallo	pesce	due	tre	quattro	nome
Romanian	ochi	ureche	nas	câine	cal	peşte	doi	trei	patru	nume
Irish	súil	cluas	soc	madra	capall	iasc	dhá	trí	ceathair	ainm
Welsh	llygad	clust	trwyn	ci	ceffyl	pysgod	dau	tri	pedwar	enw
Albanian	sy	vesh	hundë	qen	kalë	peshk	dy	tre	katër	emër
Finnish	silmä	korva	nenä	koira	hevonen	kala	kaksi	kolme	neljä	nimi
Estonian	silm	kõrv	nina	koer	hobune	kala	kaks	kolm	neli	nimi
Northern Saami	čalbmi	beallji	njunni	beana	heavuš	guolli	guokte	golbma	njeallje	namma
Hungarian	szem	fül	orr	kutya	ló	hal	kettő	három	négy	név
Turkish	göz	kulak	burun	it	at	balık	iki	üç	dört	ad
Uzbek	ko’z	quloq	burun	it	ot	baliq	ikki	uch	to’rt	ot
Basque	begi	belarri	sudur	txakur	zaldi	arrain	bi	hiru	lau	izen
Greenlandic	isi	siut	qingaq	qimmeq	hiisti	aalisagaq	marluk	pingasut	sisamat	ateq

Lexicostatistical Databases: Purpose

Use in historical linguistics:

- give a rough heuristic to determine whether languages are **related** (are words more similar than expected by chance?)
- form initial hypotheses about the **tree structure** of a language family (group of related languages), showing the development
- provide **statistical evidence** if classical method can't decide an open question (usually "which language split off first?")
- **dating** of proto-languages (disputed!)
- **location** of proto-languages (even more disputed!)

Types of Databases

- **form databases** include the forms for each language-concept pair in a unified phonetic format, allowing forms to be compared by a computer:

	Armenian	Albanian	Greek	Georgian
HAND	[d͡zɛrkʰ]	[dɔrə]	[çeri]	[χɛlɪ]

- **cognate databases** provide an encoding of cognate relationships as determined by historical linguists, but do not contain information on pronunciation

	Armenian	Albanian	Greek	Georgian
HAND_SET1	1	1	1	0
HAND_SET2	0	0	0	1

- most valuable type of database combines both!

Form Databases: Advantages

- forms are easier to extract from sources than etymologies
- more empirical: cognacy judgments are treated as secondary structures, not as elementary facts
- can also be used for language families where etymology is underdeveloped
- native speakers can help with data collection
- data can be re-used for many other purposes (e.g. comparative phonotactics)

Form Databases: Problems

- some cross-linguistic phonetic representation is necessary
- most languages have no standardized orthography
- different sources might disagree on the pronunciation
- disagreement about reconstructed forms
- dialect differences cause problems if more than one source is used
- issues of representation often introduce noise

Cognate Databases: Advantages

- cleaner data
- binary data easier to model mathematically
- abstracts away from a lot of irrelevant details (like exact pronunciation)
- previous knowledge, often the result of decades of research, is not discarded, but made good use of

Cognate Databases: Problems

- a lot of information is lost in binary cognate judgments (more closely related languages will have more similar forms)
- experts will frequently disagree on cognate judgments (the data are theories, not measurements)
- can only be compiled with the help of experts or literature
- for many language families, there are no up-to-date etymological dictionaries; information needs to be scraped together from articles
- difficult to ensure equal quality (more etymological work will have been done for some languages)
- not necessarily neutral: an etymological dictionary may give preference to interpretations that support the author’s theory

Lexicostatistical Databases: Example

In the rest of the talk, our own database is used to illustrate

- how to decide which languages to include in a database
- how to decide which concepts to include in a database
- how to collect the data in a principled way
- which challenges arise when working across many languages
- the effort necessary to arrive at a useful form database

Table of Contents

Lexicostatistics

Lexicostatistical Databases

Data Collection Example: NorthEuraLex

- Goals and Scope
- Design Decisions
- Data Handling
- Current Status & Future

NorthEuraLex: Goals

Goals of our own 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

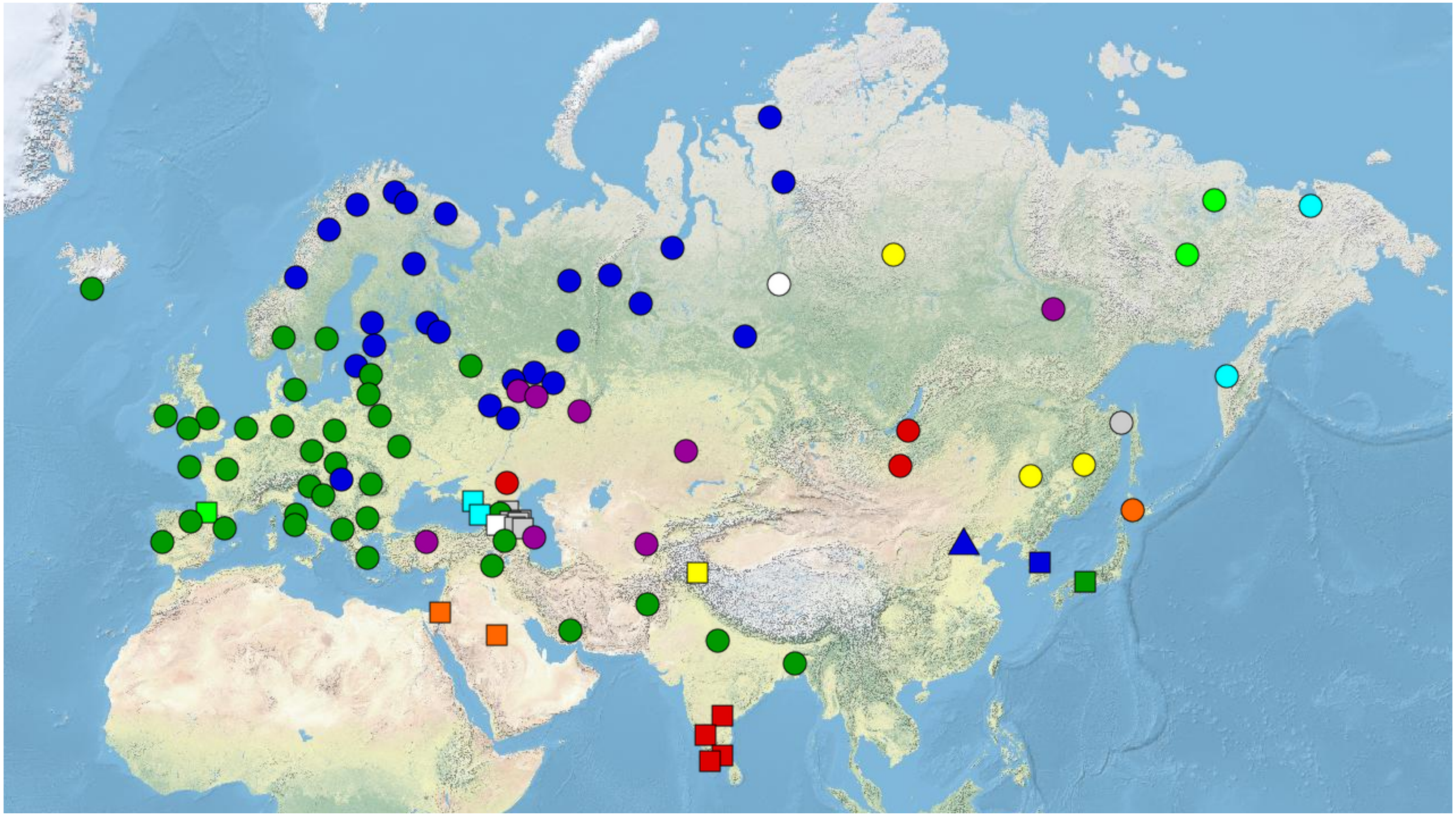
Availability:

- release version 0.9 available at `northeuralex.org`

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 107 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 Coverage



Design Decisions: Selecting the Concepts

- most databases use adapted Swadesh lists, or older wordlists
- we 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)

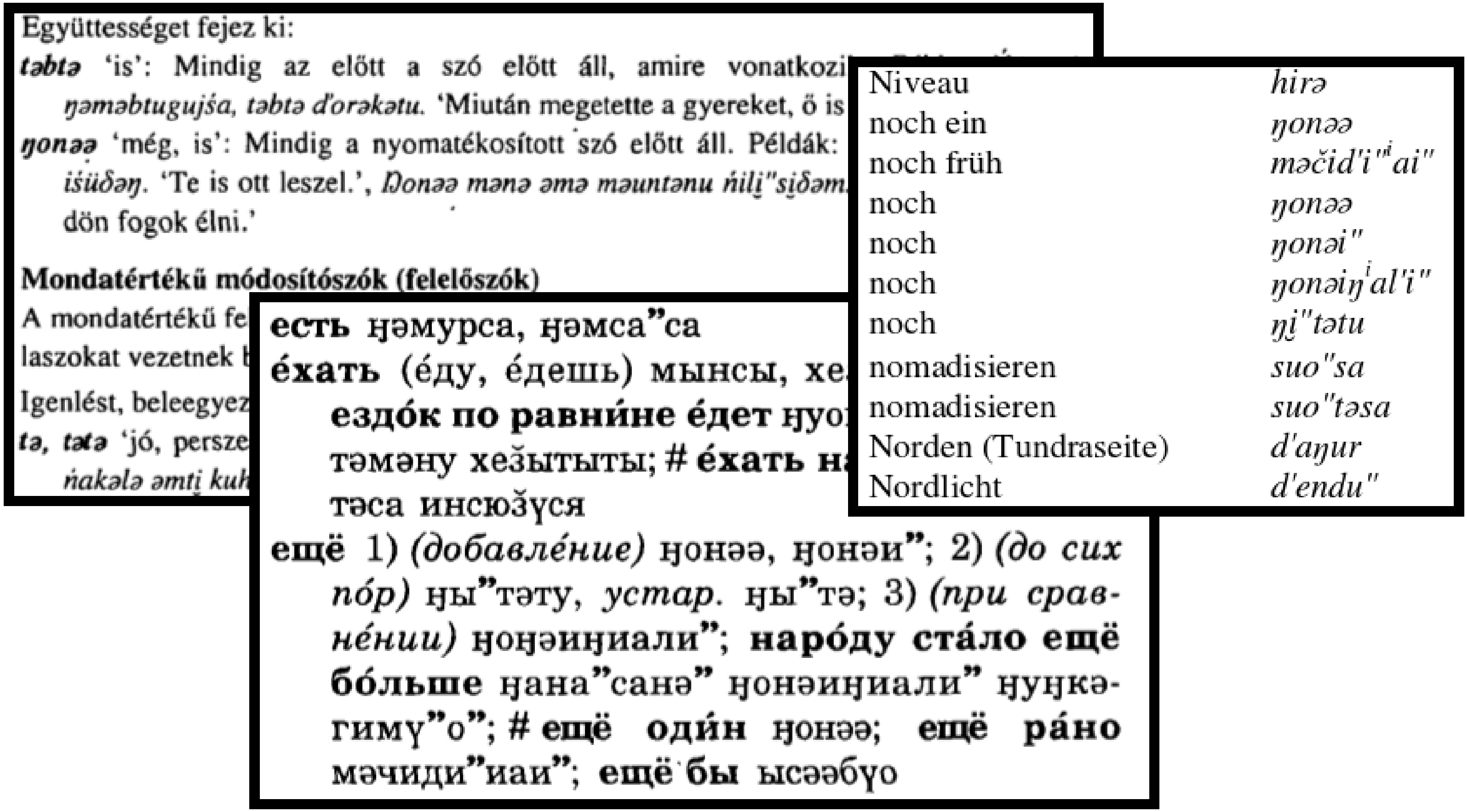
Design Decisions: Data Collection

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, ...)

Data Collection: Challenges

- bridging 10 different gloss languages



Együttességet fejez ki:

təbtə 'is': Mindig az elött a szó elött áll, amire vonatkozi
ŋəməbtugujśa, təbtə d'orəkətu. 'Miután megetette a gyereket, ő is

ŋonəə 'még, is': Mindig a nyomatékosított szó elött áll. Példák:
iśüδəŋ. 'Te is ott leszel.', *Ŋonəə mənə əmə məuntənu ńili̮"si̮δəm.*
dön fogok élni.'

Mondatértékű módosítószók (felelőszók)

A mondatértékű fe
laszokat vezetnek b

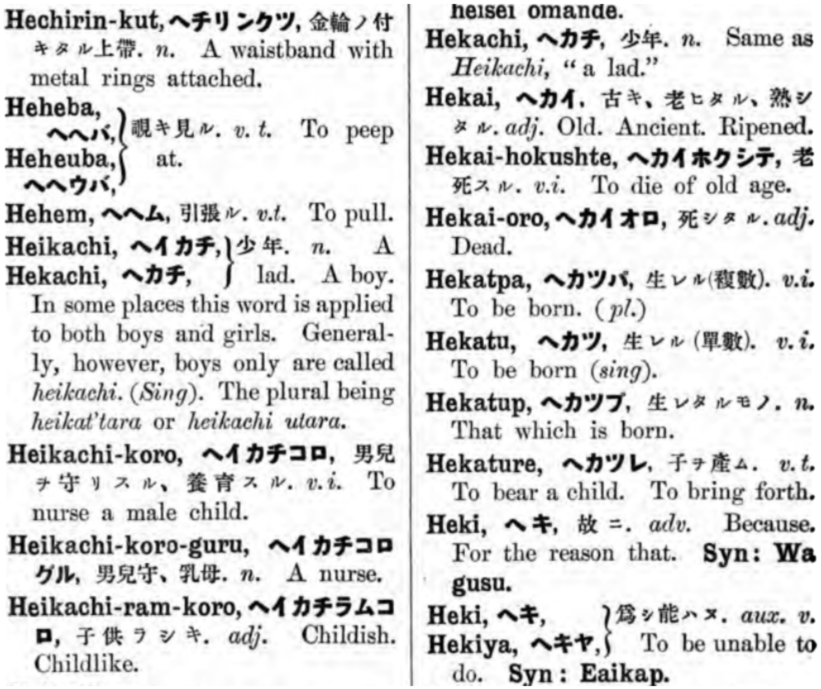
Igenlést, beleegyez
tə, tatə 'jó, persze
ńakələ əmti̮ kuh

есть ӈәмурса, ӈәмса”са
е́хать (е́ду, е́дешь) мынсы, хе
ездо́к по равни́не е́дет ӈуо
тәмәну хеўытыты; # **е́хать н**
тәса инсюўүся
ещё 1) *(добавле́ние)* ӈонәә, ӈонәи”; 2) *(до сих по́р)* ӈы”тәту, *устар.* ӈы”тә; 3) *(при сравне́нии)* ӈоӈәиӈиали”; **наро́ду ста́ло ещё бо́льше** ӈана”санә” ӈонәиӈиали” ӈуӈкәгиму”о”; # **ещё оди́н** ӈонәә; **ещё ра́но** мәчиди”иаи”; **ещё бы** ысәәбүо

Niveau	*hirə*
noch ein	*ŋonəə*
noch früh	*məčid'i"ⁱai"*
noch	*ŋonəə*
noch	*ŋonəi"*
noch	*ŋonəiŋⁱal'i"*
noch	*ŋi̮"tətu*
nomadisieren	*suo"sa*
nomadisieren	*suo"təsa*
Norden (Tundraseite)	*d'aŋur*
Nordlicht	*d'endu"*

Data Collection: Challenges

- making the selection decisions based on the sparse information in some dictionaries (especially for verbal concepts)



Hechirin-kut, ヘチリンクツ, 金輪ノ付キタル上帶. *n.* A waistband with metal rings attached.

Heheba, ヘヘバ, Heheuba, ヘヘウバ, 覗キ見ル. *v. t.* To peep at.

Hehem, ヘヘム, 引張ル. *v.t.* To pull.

Heikachi, ヘイカチ, Hekachi, ヘカチ, 少年. *n.* A lad. A boy. In some places this word is applied to both boys and girls. Generally, however, boys only are called *heikachi.* (*Sing*). The plural being *heikat'tara* or *heikachi utara.*

Heikachi-koro, ヘイカチコロ, 男兒ヲ守リスル、養育スル. *v.i.* To nurse a male child.

Heikachi-koro-guru, ヘイカチコログル, 男兒守、乳母. *n.* A nurse.

Heikachi-ram-koro, ヘイカチラムコロ, 子供ラシキ. *adj.* Childish. Childlike.

heisei omande.

Hekachi, ヘカチ, 少年. *n.* Same as *Heikachi,* "a lad."

Hekai, ヘカイ. 古キ、老ヒタル、熟シタル. *adj.* Old. Ancient. Ripened.

Hekai-hokushte, ヘカイホクシテ, 老死スル. *v.i.* To die of old age.

Hekai-oro, ヘカイオロ, 死シタル. *adj.* Dead.

Hekatpa, ヘカツパ, 生レル(複數). *v.i.* To be born. (*pl.*)

Hekatu, ヘカツ, 生レル (單數). *v. i.* To be born (*sing*).

Hekatup, ヘカツプ, 生レタルモノ. *n.* That which is born.

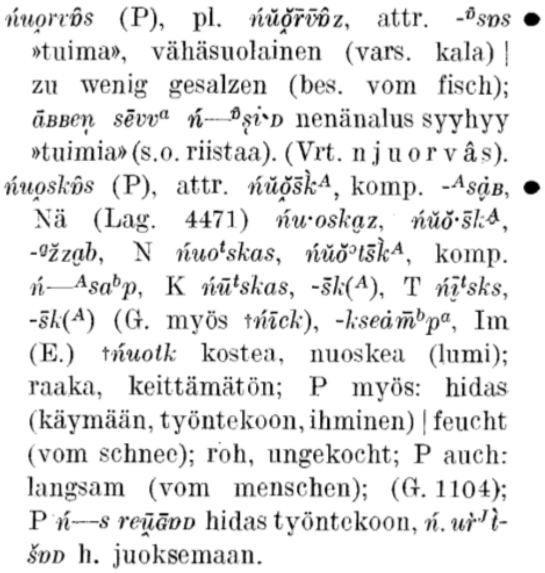
Hekature, ヘカツレ, 子ヲ產ム. *v. t.* To bear a child. To bring forth.

Heki, ヘキ, 故ニ. *adv.* Because. For the reason that. **Syn: Wagusu.**

Heki, ヘキ, Hekiya, ヘキヤ, 爲シ能ハヌ. *aux. v.* To be unable to do. **Syn: Eaikap.**

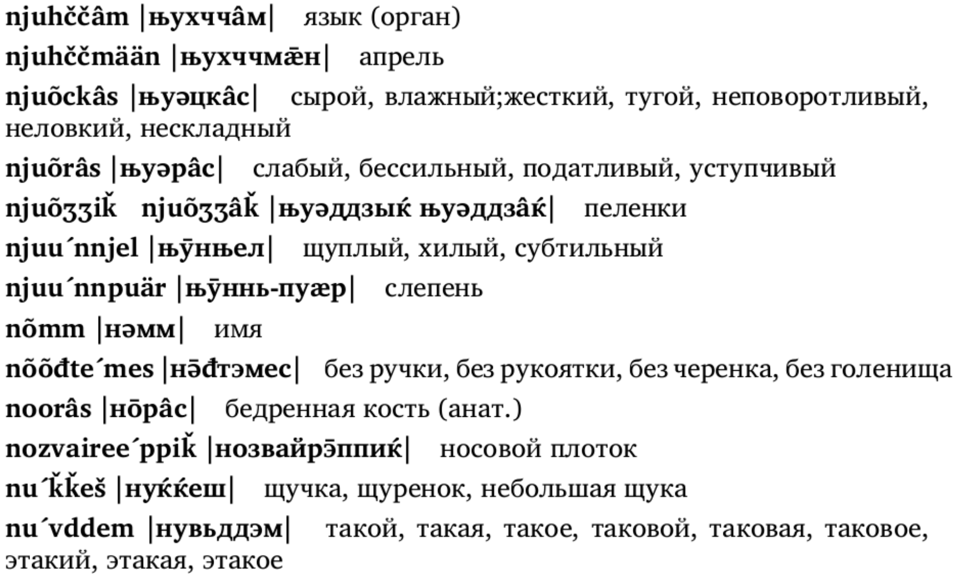
Data Collection: Challenges

- unifying different sources targeted at different audiences, covering different dialects, using incompatible transcription systems (e.g. the Uralic Phonetic Alphabet)



ńuọrvɒ̂s (P), pl. *ńŭọ̆r̄v̄ɒ̂z*, attr. -$^{ɒ̂}$*sɒs* ● »tuima», vähäsuolainen (vars. kala) | zu wenig gesalzen (bes. vom fisch); *āʙʙeṇ sēvv^{a} ń—$^{ɒ̂}$ṣi·ᴅ* nenänalus syyhyy »tuimia» (s.o. riistaa). (Vrt. njuorvâs).

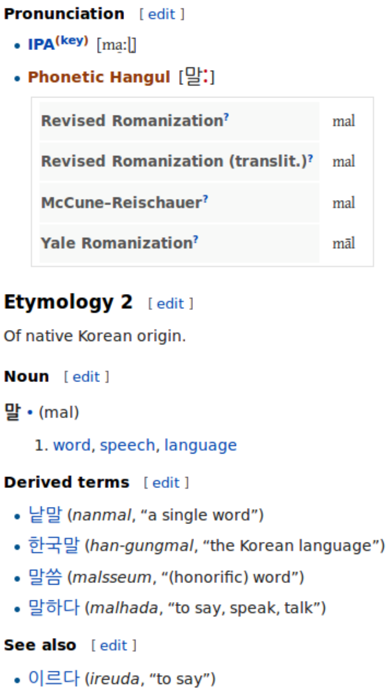
ńuọskɒ̂s (P), attr. *ńŭọ̆s̄k̄^{A}*, komp. -^{A}*sạʙ*, ● Nä (Lag. 4471) *ńu·oskạz*, *ńŭŏ·s̄k^{A}*, -^{g}*žzạb*, N *ńuo^{t}skas*, *ńŭŏ^{ɔ}ts̄k̄^{A}*, komp. *ń—^{A}sa^{b}p*, K *ńū^{t}skas*, -*s̄k*(^{A}), T *ńī̯^{t}sks*, -*s̄k*(^{A}) (G. myös †*ńīck*), -*kseȧm̄^{b}p^{a}*, Im (E.) †*ńuotk* kostea, nuoskea (lumi); raaka, keittämätön; P myös: hidas (käymään, työntekoon, ihminen) | feucht (vom schnee); roh, ungekocht; P auch: langsam (vom menschen); (G. 1104); P *ń—s reṷāɒᴅ* hidas työntekoon, *ń. uṙ^{J}ì-šɒᴅ* h. juoksemaan.



njuhččâm	**|њухччâм|**	язык (орган)
njuhččmään	**|њухччмǣн|**	апрель
njuõckâs	**|њуəцкâс|**	сырой, влажный;жесткий, тугой, неповоротливый, неловкий, нескладный
njuõrâs	**|њуəрâс|**	слабый, бессильный, податливый, уступчивый
njuõʒʒiǩ njuõʒʒâǩ	**|њуəддзыќ њуəддзâќ|**	пеленки
njuu´nnjel	**|њӯнњел|**	щуплый, хилый, субтильный
njuu´nnpuär	**|њӯннь-пуæр|**	слепень
nõmm	**|нəмм|**	имя
nõõđte´mes	**|нə̄ђтэмес|**	без ручки, без рукоятки, без черенка, без голенища
noorâs	**|нōрâс|**	бедренная кость (анат.)
nozvairee´ppiǩ	**|нозвайрэ̄ппиќ|**	носовой плоток
nu´ǩǩeš	**|нуќќеш|**	щучка, щуренок, небольшая щука
nu´vddem	**|нувьддэм|**	такой, такая, такое, таковой, таковая, таковое, этакий, этакая, этакое

Data Collection: Challenges

- phonemic differences not represented by imperfect orthographies
- example: Korean *māl* “language” vs. *mal* “horse”



Pronunciation [edit]

- **IPA**(key) [ma̠ːl]
- **Phonetic Hangul** [말ː]

Revised Romanization?	mal
Revised Romanization (translit.)?	mal
McCune-Reischauer?	mal
Yale Romanization?	māl

Etymology 2 [edit]

Of native Korean origin.

Noun [edit]

말 • (mal)

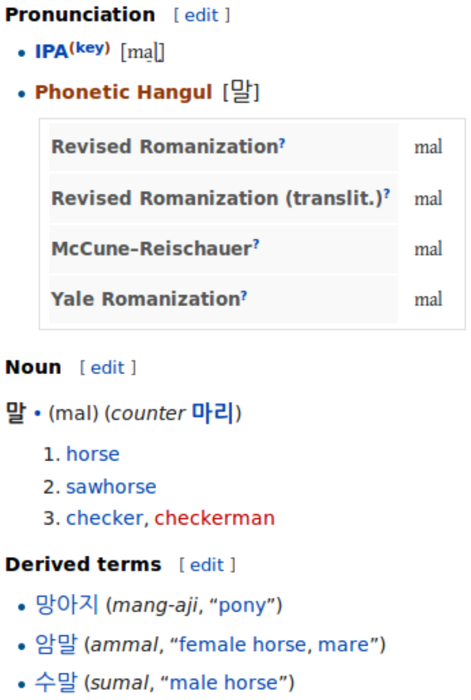
1. word, speech, language

Derived terms [edit]

- 낱말 (*nanmal*, “a single word”)
- 한국말 (*han-gungmal*, “the Korean language”)
- 말씀 (*malsseum*, “(honorific) word”)
- 말하다 (*malhada*, “to say, speak, talk”)

See also [edit]

- 이르다 (*ireuda*, “to say”)



Pronunciation [edit]

- **IPA**(key) [ma̠l]
- **Phonetic Hangul** [말]

Revised Romanization?	mal
Revised Romanization (translit.)?	mal
McCune-Reischauer?	mal
Yale Romanization?	mal

Noun [edit]

말 • (mal) (*counter* 마리)

1. horse
2. sawhorse
3. checker, checkerman

Derived terms [edit]

- 망아지 (*mang-aji*, “pony”)
- 암말 (*ammal*, “female horse, mare”)
- 수말 (*sumal*, “male horse”)

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 common formats
  - ▷ Dolgopolsky sound classes: `KWVRP`
  - ▷ LingPy's internal model ("List classes"): `CBULB`
  - ▷ ASJP sound classes: `cvElf`
  - ▷ reduced versions of IPA: `tsvœlf`

Design Decisions: Workflow

- in contrast to comparable efforts, we do not rely on experts providing us with complete wordlists
- 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

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
- automated conversion of our transducer files into more mainstream and highly efficient finite-state transducers, public release in preparation

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)
- loanword detection (Köllner & Dellert, in preparation)
- models of semantic change (see e.g. Münch & Dellert 2015)

NorthEuraLex: Future

- during 2018: 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

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.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.