EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
FB Neuphilologie
Seminar für Sprachwissenschaft

Towards richer multi-source machine-readable etymologies

Johannes Dellert

February 25, 2021

MaEiQCL workshop at the DGfS annual meeting

European Research Council
Established by the European Commission

Table of Contents

Motivation: Current State of Etymological Databases

Existing etymological databases can be classified into two types:

- **source-oriented**: faithful digitalization of published etymological sources in order to make them searchable, but often not fully machine-readable (e.g. UraloNet, Bakró-Nagy et al. 2013)
- **data-oriented**: fully machine-readable etymological annotations built on top of lexical databases, limited to connecting the relevant subset of the lexicon in the relevant languages only (e.g. LexiRumah, Kaiping and Klamer 2018)

Focus of presentation: steps towards fully machine-readable source-oriented etymology models (with the option of converting them into annotations if desired)

Motivation: Relations vs. Annotations

Data-oriented etymological annotations tend to

- be limited to some rather flat aspects (e.g. cognacy, loanword status, morphological splits) which fall short of full etymologies
- avoid modeling intermediate stages which are not directly expressible as annotations to the data (e.g. historical derivations)
- encode decisions which go beyond what is explicitly stated in the sources, and therefore often constitute primary research, which tends to be less scrutinized than etymologies published on paper, but can nevertheless have enormous impact

Motivation: Relations vs. Annotations

A more source-oriented toolbox for modeling etymologies will

- take the **relational nature** of etymological data more seriously (e.g. inheritance events are primary, cognate sets secondary)
- allow to **underspecify** every aspect of the underlying relational model (e.g. "likely a borrowing from a Turkic language, cf. Chuvash *xy*")
- provide **consistency checks** to prevent contradictory combinations of links (much harder than for annotations!)
- be **fully transparent about sources** of each etymological judgment, model only what is explicit in those sources, and ideally allow configurable mixing of sources, e.g. in the shape of a preference ordering

Motivation: Issues of Representation

problems of representing sources:

- while in the ideal case, our inventory of relation symbols should be expressive enough to represent all of the information contained in any etymological dictionary, some compromise will be needed
- very difficult to ensure consistent treatment of many complex structures when only collecting them as relation tuples
- information is often given only implicitly by sources

our approach: format with wide variety of syntactic elements geared towards direct representation of the knowledge as provided in etymological sources, supported by a complex parser for factoring the format into a relational model which supports underspecification

Table of Contents

SOrEty: Guiding Principles

- **source-oriented** format
  - ▷ one file per etymological source
  - ▷ can mirror typical structures of current etymological literature (i.e. organization by etymon or by modern dictionary form)
  - ▷ non-trivial steps performed by the data collector during lookup are explicitly documented in separate index files (not covered today)
- **faithful** representation
  - ▷ documents mapping to original language identifiers
  - ▷ emulates original formatting of cited forms as far as possible by strings of Unicode characters
  - ▷ allows addition of explicit links to connect e.g. with orthography or alternative forms used by a database to be annotated
- **plain text** to ensure sustainability and facilitate processing

SOrEty Basic Syntax: Form Identification

- a **form** in the data model minimally consists of:
  - ▷ a language ID (might be a proto-language)
  - ▷ a form representation (exactly as in the original)
  - ▷ a morphological category
    (can include e.g. gender or transitivity)
  - ▷ a string of glosses in any language
    (used for distinguishing homographs)
- in SorEty form specifications, category (after colon) and gloss (in single quotes) are optional and can be inherited from a pivot form (typically the lemma in the etymological source)
- examples:
  - ▷ `ENG window`
  - ▷ `DEU Gans:Nf 'goose'`
  - ▷ `PSAM *äŋ 'suu, aukko'`

SOrEty Basic Syntax: Form Bridging

- in order to facilitate conversion of a source model into database annotations, SOrEty supports explicit links for forms by attaching alternative representations connected by ==
  - ▷ `SEL qętti̯==ҕэтты 'linnoitus, kaupunki'`
  - ▷ `GRC therɯmós==θερμός 'warm'`
- this mechanism can also be used to bridge different citation forms between sources and databases:
  - ▷ `EVN имӯ-==имӯ-мӣ 'мазать (жиром, мазью); топить жир'`
  - ▷ `BAK ышан-==ышаныу 'верить; доверять'`
- as a first example of underspecification, ~= is used to represent a partial match (e.g. in case derivates or inflected forms are not found in source)
  - ▷ `UZS åpa~=oralik 'промежуток'`

SOrEty Structures: Explicit Relations

Form specifications can be explicitly linked by relation symbols:

- inheritance and internal change:
  - ▷ `SME vuoggjâ-==vuodjat 'swim' < PURA *uxï- 'swim'`
  - ▷ `PITA *feχʷri 'fever' < PIEU *dʰegʷʰ-ri- 'burn, fever'`
  - ▷ `HUN jön 'to come' < HUN *jäβ/*jöü`
- borrowing:
  - ▷ `SQI treg Nm 'market' << PSLA *trъgъ 'market'`
  - ▷ `MNC xefeli==hefeli 'брюхо, живот' << KHK хэвлий`
- unspecified derivation (plus inheritance):
  - ▷ `EKK hingama 'to breathe' <D EKK hing 'soul; breath'`
  - ▷ `FRA nuage 'cloud' <D LAT nŭbes *cloud'`

SOrEty Structures: Chain Structures

- relation symbols can be chained together to represent direct etymological paths of arbitrary length:
  - ▷ `SQI shkurt:Nm 'February' <D SQI shkurt:A 'short' << PGER *skurtaz 'short'`
  - ▷ `SQI vajze 'girl' < SQI *varë < SQI *vëharë < PALB *swesarā < PIEU *sue(-)sr- 'sister'`
  - ▷ `HUN csizma 'boot' << HRV čizma 'boot' << OTA çizme 'boot' <D OTA çizmek 'to tear' < PTUR *čiŕ- 'to tear'`

SOrEty Structures: Morphological Processes

- derivational processes can also be expressed explicitly via the notation `Result <D Form1 + Form2 + ...`
  - ▷ `FIN herkullinen:A 'tasty'`
    `<D FIN herkku 'treat' + FIN -llinen`
  - ▷ `PALB *banti 'dwelling'`
    `<D PALB *banja 'to make, to do' + PALB *-nti`
- the same format is used for representation of compounding:
  - ▷ `FIN helmikuu 'February'`
    `<C FIN helmi 'pearl' + FIN kuu 'moon'`
  - ▷ `PITA *awizdj[eo]- 'to hear'`
    `<C PIEU *h₂eu-is 'clearly' + PIEU *dʰh₁-ie/o- 'to render'`
- occasionally, it is also useful to model inflection in this way:
  - ▷ `FIN aikanaan:ADV 'once'`
    `<I FIN aika 'time' + FIN -nA + FIN -Vn 'POSS.3SG'`

SOrEty Structures: Relation Symbols

- inventory of etymological link types we currently use:

`<`	inherited from X under regular sound change
`<A`	inherited from X under analogical change
`<A[Y]`	inherited from X under analogy with Y
`<B`	inherited/derived from X through backformation
`<C`	inherited/derived from X through compounding
`<D`	inherited/derived from X through derivation
`<I`	inherited/derived from X through inflection
`<P`	inherited/derived from X under phrase formation
`<R`	inherited/derived from X under regularization
`<U`	inherited/derived from X under other type of reshaping (e.g. folk etymology, re-analysis)
`<<`	borrowed from (description of donor language form follows)

SOrEty Line Format

The default source format consists of **six tab-separated columns**, where each line represents an etymon:

- **etymon ID** (from the source, or page number plus running index)
- **lemma** (reconstructed or attested, connects etymologies)
- **category** (part-of-speech for the lemma)
- **gloss** for lemma (in single quotes, as provided by source)
- **descendant links**: a semicolon-separated list of etymological comparisons (single forms, chains or formations) which express forward development through time (e.g. reflexes of lemma, targets of borrowing)
- **antecedent links**: a semicolon-separated list of etymological comparisons (single forms, chains or formations) which require backtracing through time (e.g. proto-form, cognates from other branches, sources of borrowing)

SOrEty Semantics: Pivot Forms

- SOrEty implements the common convention of treating the lemma as a **pivot form** to which the remainder of the entry is put in relation
- default semantics on the descendant list: list of reflexes
  - ▷ `CHU běgati 'run, flee'; RUS bégať 'run'; CES běhati 'run'` (for pivot form `PSLA *bě̀gati:V 'run, flee'`)
  - ▷ `KHK нохой; BXR нохой; XAL ноха; PEH ноŋуi==nөgəi; ...` (for pivot form `PMON noqai:N 'собака'`)
- default semantics on the antecedent list: full cognacy
  - ▷ `XCL ayl 'other'; GRC ἄλλος 'other'; LAT alius 'other'; SGA aile 'other'` (for pivot `PGER *alja-:A 'someone else'`)
  - ▷ `FIN ikä; SME âkke==ahki; MYV ije==ие; MHR ii(j)==ий; HUN év` (for pivot form `PUGR *ĭkä 'year, age'`)

SOrEty Semantics: Phylogeny-Based Semantics

Lists of reflexes and cognates often have additional hidden structure which makes tree-based interpretation necessary:

- a reflex is always a direct descendant of the pivot entry, but not necessarily an immediate child:
  - ▷ `GOT andeis 'end'; ANG ende 'end, back'; ENG end`
- interpretation of cognates depends on whether the quoted language is a direct ancestor of the pivot language!
  - ▷ `PIEU *Heh₃l-én-eh₂-; GRC ώλένη 'elbow, underarm'; LAT ulna 'forearm'` (with pivot form `PGER *alīnō- 'forearm'`)

The (author-specific) language tree against which etymologies need to be interpreted, can be specified in a SOrEty source file (default: Glottolog).

SOrEty Structures: Relations and Chains in Lists

- default semantics can be overriden by starting a list entry with a relation symbol (pivot form will be inserted to the left of the symbol):
  - `EVN эр 'this'; >D EVE эрэк 'this'; ULC эj==эй 'this'`
  - `SPA almohada; POR almofada << ARA muḫadda 'cushion'`
- in a chain, the final element of the chain is the form that is relevant for attachment to the pivot form:
  - `PALB *mazdnja <D PIEU *mazd- 'feeding'; GOH mast`
  - `ITA padrone 'master'; ITA patrono << LAT patrōnus`

SOrEty Structures: Subetymologies

- any form can be replaced by subetymology, i.e. a list of forms in additional parentheses, with the default semantics applied locally
- in many cases, some partial cognates of the pivot form are more closely connected, in which case it is useful to group them into subetymologies:
  - ▷ `PIEU *h₂eus-n- <D (PIEU *h₂ous; SQI vesh; GRC οὖς)`
    (with pivot form `PGER *auzōn- 'ear'`)
- they can also be used to trace the parts in word formation events:
  - ▷ `SQI vëlla 'brother' < PALB *swelaudā`
    `<C (PALB *swe <I PIEU *sṷo- 'self') +`
    `(PALB *laudā < PIEU *leudh-; GOH liut 'people')`

Underspecification in SOrEty

- question marks can be used to underspecify forms:
  - ▷ unknown language
  - ▷ unknown form
- special relation symbol `X ~ Y` directly encodes partial cognacy (i.e. effectively a shorthand for `X <D ? ? >D Y`);
  this can also connect entire subetymologies: `X ~ (Y; Z)`
- Examples:
  - ▷ `< ? *apa << (PTUR ?; AZJ aba; CHV upá 'bear')`
    'from earlier **apa*, a borrowing from a Turkic language;
    cf. Azeri *aba*, Chuvash *upá* bear"'
  - ▷ `PITA *bak-(k)elo-:N 'stick, staff'`
    `~ (GLE bacc 'hook'; CYM bach) < PCEL *bakko-;`
    `~ LAV bakstít:V==bakstīt 'to poke'`

Representing Alternatives in SOrEty: Disjunction

- (sub)etymologies can be joined together using the symbol | to express disjunction (i.e. alternative etymologies)
- examples:
  - ▷ `FIN hanka 'fork (of tree); rowlock'`
    `(HUN ág 'branch, twig') | << (PGER *hangu- < PGER *hanhu;`
    `NON hár 'rowlock'; SWE-DIA hå 'rowlock')`
  - ▷ `PGER *aban-:Nm 'man, husband'`
    `< (<NURSERY>; ETT apa 'father')`
    `| <D ( PIEU *h₃ep- 'to labor, be powerful')`

SOrEty: Epistemic Modifiers

- information, especially in the best sources, often comes with a stance expressed by the author; sometimes etymologies are explicitly rejected
- SOrEty allows to model these stances by **epistemic modifiers**, which can be prefixed to (almost) any structure:

symbol	meaning
`?+`	probably, likely
`?`	possibly, maybe, perhaps, could be
`??`	not very likely, might be
`???`	improbable, questionable
`¬`	rejected etymology

- Examples:

```
▷ < ? *apa ? << (PTUR ?; AZE aba; ? CHV upá 'bear')
▷ PSLA *bedrò    PIEU *bʰedʰ-róm; ¬ PIEU *bed- 'swell'
```

Representing Sources: Example from Kroonen (2013)

***blēwa-** adj. 'blue' — ON *blár* adj. 'blue, livid, black', Far. *bláur* adj. 'blue; dark', Elfd. *blå* adj. 'blue', OE *blāw* adj. 'id.', E Scot. *blae* adj. 'blackish; livid, pale', OFri. *blāw* adj. 'id.', Du. *blauw* adj. 'id.', OHG *blāo* adj. 'blue, dark, grey', G *blau* adj. 'id.' ⇒ *b^{h}léh$_1$-uo- (EUR) — Lat. *flāvus* adj. 'blond', OIr. *blá* adj. 'yellow', W *blaw* adj. 'grey' < *b^{h}lh$_1$-uo-.

```
#68.4  *blēwa- A       'blue'
  NON blár 'blue, livid, black'; FAO bláur 'blue; dark'; OVD blå 'blue';
    ANG blāw 'blue'; SCO blae:A 'blackish; livid, pale'; OFS blāw; NLD blauw;
    GOH blāo:A 'blue, dark, grey'; DEU blau
  PIEU *bʰléh1-uo-;
    ~ (LAT flāvus 'blond'; SGA blá 'yellow'; CYM blaw 'grey') < PIEU *bʰlh1-uo-
```

Representing Sources: Example from De Vaan (2008)

ānser 'goose' [m. (f.) *r*] (Pl.+)
PIt. **χans-*.
PIE *ǵ^{h}h$_2$ens [nom.], *ǵ^{h}h$_2$ns-os [gen.] 'goose'. IE cognates: OIr. *gēiss* 'swan', Skt. *haṁsá-* [m.], Gr. χήν, -ός [m. f.], Dor. Boeot. χᾶν, OPr. *sansy*, Lith. *žąsìs* [f.], acc. *žąsi*, Ru. *gus'*, Po. *gęs'* (< PSl. **gǫsь*), OHG *gans*, OE *gōs* 'goose'.

Initial **h*- has been dropped. The length of *ā* is automatic in front of *ns*. Leumann 1977: 380 reconstructs **hāns*, **hānesem*, **hāns-os* > acc.sg. **hānerem*, which was replaced by **hānserem* on the analogy with the gen.sg. **hāns*-. From the acc.sg., *-er-* would have been introduced into the other case forms.
Bibl.: WH I: 52, EM 36, IEW 412, Kortlandt 1985a: 119, Schrijver 1991: 113.

```
#44.2 *χans-  N       'goose'
  LAT ānser 'goose' <A[*hāns] LAT *hānerem:N[AkkSg] < LAT *hānesem:N[AkkSg] <I LAT *hāns
  PIEU *ǵĥh2ens 'goose';
    SGA gēiss 'swan'; SAN haṁsá-:Nm; GRC χήν:Nmf; GRC-DOR χᾶν́; GRC-BOE χᾶν́;
    PRG sansy; LIT žąsìs:Nf; (RUS guś==гусь; POL gęś) < PSLA *gǫsь; GOH gans; ANG gōs
```

Table of Contents

Processing SOrEty: Extracting etymological paths

Core principles of **path extraction algorithm**:

- start at each given form and try to work way backwards recursively
- chained relations take precedence over list neighbors
- in each step, collect equivalent classes (local cognate sets) by graph search through list neighbors and explicit cognacy links
- within equivalence classes, always move to lowest attested ancestor; create virtual node for the lowest common ancestor if none exists
- sublists create boundaries for creating equivalence classes
- handle disjunction by factoring out paths
- assign any epistemic modifier on the path to the entire path (unifying towards the lowest confidence)

Many complex boundary cases, especially due to underspecification.

Processing SOrEty: Extracting etymological paths

```
FRM lang_id="PITA" form="*χans-" category="N" gloss="goose"
  LNK symbol=">"
    LST
      FRM lang_id="LAT" form="ānser" gloss="'goose'"
        ooo
        LNK symbol="<A[*hāns]"
          FRM lang_id="LAT" category="N[AkkSg]" form="*hānerem"
            ooo
            LNK symbol="<"
              FRM lang_id="LAT" category="N[AkkSg]" form="*hānesem"
                ooo
                LNK symbol="<I"
                  FRM lang_id="LAT" form="*hāns"
  LNK symbol="<>"
    FRM lang_id="PIEU" form="*ǵʰh2ens" gloss="'goose'"
  LNK symbol="<>"
    FRM lang_id="SGA" form="gēiss" gloss="'swan'"
  LNK symbol="<>"
    FRM lang_id="SAN" category="Nm" form="haṁsá-"
  LNK symbol="<>"
    FRM lang_id="GRC" category="Nmf" form="χήν"
  LNK symbol="<>"
    FRM lang_id="GRC-DOR" form="χᾶ́ν"
  LNK symbol="<>"
    FRM lang_id="GRC-BOE" form="χᾶ́ν"
  LNK symbol="<>"
    FRM lang_id="PRG" form="sansy"
  LNK symbol="<>"
    FRM lang_id="LIT" category="Nf" form="žąsìs"
  LST
    FRM lang_id="RUS" nelex_form="гусь" form="guś"
    FRM lang_id="POL" form="gęś"
    ooo
    LNK symbol="<"
      FRM lang_id="PSLA" form="*gǫsь"
  LNK symbol="<>"
    FRM lang_id="GOH" form="gans"
  LNK symbol="<>"
    FRM lang_id="ANG" form="gōs"
```

- for RUS form, search for ancestor among list neighbors; none found (only POL), follow chain which list is part of ⇒ link to PSLA found
- build equivalence class of PSLA, found first PITA and then, by following <> links, PIEU, SGA, SAN, GRC, GRC-DOR, GRC-BOE, PRG, GOH, ANG; lowest attested ancestor of PSLA is PIEU ⇒ link to PIEU found
- equivalence class for the PIEU form is identical to the previous one, form is common ancestor without further path upwards
  ⇒ path complete, terminate

Processing SOrEty: Exporting annotations

- exporting **cognate set annotations**
  - ▷ extract paths for every form up to first link whose relation is not <
  - ▷ attempt to unify each pair of paths (with some optimizations);
    if unification is successful, the forms belong to the same cognate set;
    if unification fails due to a conflict, they do not
  - ▷ for homologue sets, step across << links as well when extracting paths
  - ▷ for partial cognacy, step across <D links as well when extracting paths
- exporting **binary loanword annotations**
  - ▷ extract full path for each form (stepping across any type of link)
  - ▷ if path contains << link anywhere, label as borrowed
  - ▷ else, if path traces back to proto-language, label as not borrowed

Table of Contents

Source Aggregation: General Considerations

Considerations concerning the aggregation of source-based models

- creating a data-driven etymological database of any significant size will require aggregation across sources, and we would like to automate this process as far as possible
- problem: sources do contradict each other, aggregation implies making decisions which we are not necessarily qualified to make
- idea: a user-configurable **trust ranking of sources**, the way source models are unified is determined solely by such decisions
- needed: another unavoidably quite complex algorithm

Issues of Source Aggregation: Unifying Etyma

Problem 1: During path unification, how can etyma be determined to be equivalent (or complementary) if both **reconstructions and the selection of attested forms can vary**?

- ▷ Sammallahti (1988):
  `PFIU *åśk[iå]li 'step'; PFPE *aśkili; FIN askel==askel; MYV eśḱil´a-~=эскельдямо; MHR aaškël==ӧшкыл; UDM ućkyl==?; KPV voškal==?; PUGR *åskal-; ? MNS *uusel==ӯсыл`
  - ▷ Itkonen and Kulonen (1992):
    `FIN askel; MYV eśkeľks==эскелькс; MRJ aškə̑l==ашкыл; MHR oškə̑l==ӧшкыл; UDM utśkil==?; KPV voškol==воськов; MNS ūsil==ӯсыл; SEL āsel-==? 'astua yli'`
- Solution: identify etyma not based on proto-form, but on overlap of mappable attested forms (in example: FIN, MHR)

Issues of Source Aggregation: Unifying Etyma

Problem 2: what if a less reliable source quotes **more languages for a cognate set** than a more reliable one?

- Example: if we trust Sammallahti (1988) more, we would miss the equivalent in Hill Mari (MRJ), but we also would not want to take over the Selkup (SEL) reflex from Itkonen and Kulonen (1992)
- Solution: check whether according to the more reliable source, the language is a descendant of a proto-language for which the existence of the cognate set is considered established
- in example, the aggregation would contain the MRJ reflex (because it is a descendant of PFPE in Sammallahti's model), but not the SEL reflex (because Samoyedic is not a descendant of PFIU)
- this logic could potentially be applied to automatically inferred cognate sets as well (though initial results are not encouraging)

Issues of Source Aggregation: Contradictions

Difficult situations where sources contradict each other:

- one source may align a word with one cognate set, and one with another, but both would reject joining the cognate sets!
- a form can be partially linked to one cognate set, but also with additional forms e.g. in a neighboring branch
- one source can quote a form as a loanword, whereas the other connects it to related forms assuming cognacy

Step towards consistent treatment of such situations:

- by default, reject any relation from a lower-ranked source that contradicts a higher-ranked source
- do not propagate cognate relations from one source into the donor language if it the word is described as a loanword by another source

Table of Contents

Issues of Source Coverage

Absence of a link in a source can have many meanings:

- absent links are rejected, and therefore not included (**"complete theory" interpretation**)
- links that are not mentioned are assumed to be "inherited" from previous literature (**"no objection" interpretation**)
- it is assumed to be obvious based on the given links (**"fill in the gaps" interpretation**)

Ideas for handling absence of information (feedback very welcome!):

- if the most reliable source claims complete coverage for certain languages, adopt "complete theory" semantics?
- otherwise, use similar criterion as during source unification to catch and process instances of "no objection" semantics?
- for "fill in the gaps", model our judgments as additional sources that the database user can choose to reject or edit?

State of the Infrastructure

- Python implementation of format parser is finished,
  though error feedback for data collectors could be improved
- Python implementation of path extraction algorithm still not feature-complete, but already handles the vast majority of the tens of thousands of etymologies we digitalized
- Python implementation of annotation export exists,
  and will soon be plugged together with CLDF module
- cross-source aggregation, compensating lack of explicit coverage in sources: still foci of ongoing work
- planned for open-source release as `pysorety` package on GitHub
  (will accompany a future publication)

State of Data Collection

All examples were from ongoing etymology work on NorthEuraLex:

- **63,481 etymological paths** from five language families
- 41,565 paths to NorthEuraLex languages, **16,788 forms covered**
- current coverage of the five families: **13.7%** of 97,023 forms

Main issues making it difficult to achieve high coverage:

- the most efficient sources, i.e. etymological dictionaries covering entire families, tend to only cover the inherited lexicon
- very often, only some "key languages" are mentioned (e.g. no modern Scandinavian languages; fill in the gaps?)
- derivates and compounds are not covered comprehensively
- wide coverage of loanwords will require processing large amounts of specialized literature that is frequently only available in the target language ($\Rightarrow$ standards for accompanying digital materials would help!)

Acknowledgments: Data contributors

- Anna Bródy (Hungarian, Samoyedic, Saami languages)
- Živilė Rasimaitė (Baltic languages)
- Yuliya Mkhayan (Slavic languages)
- Rahel Albicker (Germanic languages)
- Karina Hensel (Romance languages)
- Anna Karnysheva (Turkic languages)
- Arne Rubehn (Italic und Tungusic languages)

Acknowledgments: Funding

My work is being funded by the project CrossLingference, a grant from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme
(grant agreement no. 834050, awarded to Gerhard Jäger).

European Research Council
Established by the European Commission

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT,
FORSCHUNG UND KUNST

Work on data collection was financed primarily by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63), complemented by a RiSC grant from the MWK Baden-Württemberg, for developing the prototype of an etymological inference engine (EtInEn).

References

Bakró-Nagy, M., Mus, N., Oszkó, B., Sipos, M., Takács, D., and Várnai, Z. (2013). Uráli etimológiák a világhálón. In *Obi-ugor és szamojéd kutatások, magyar őstörténet. Hajdú Péter és Schmidt Éva emlékkonferencia 2012, Pécs.*

De Vaan, M. (2008). *Etymological Dictionary of Latin and the other Italic Languages*. Brill, Leiden.

Derksen, R. (2008). *Etymological Dictionary of the Slavic Inherited Lexicon*. Brill, Leiden.

Itkonen, E. and Kulonen, U.-M. (1992). *Suomen sanojen alkuperä: etymologinen sanakirja*. Suomalaisen kirjallisuuden seura.

Kaiping, G. A. and Klamer, M. (2018). LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLoS One*, 13,10.

Kroonen, G. (2013). *Etymological Dictionary of Proto-Germanic*. Brill, Leiden.

Orel, V. E. (1998). *Albanian Etymological Dictionary*. Brill, Leiden.

Sammallahti, P. (1988). Historical phonology of the Uralic languages (With Special Reference to Permic, Ugric and Samoyedic). In Sinor, D., editor, *The Uralic Languages*.