Loanwords in Basic Vocabulary as an Indicator of Borrowing Profiles

Mervi de Heer
Doctoral Researcher, Department of Modern Languages, Uppsala University, Uppsala, Sweden
Corresponding author
mervi.deheer@moderna.uu.se

Rogier Blokland
Professor of Finno-Ugric Languages, Department of Modern Languages, Uppsala University, Uppsala, Sweden
rogier.blokland@moderna.uu.se

Michael Dunn
Professor of General Linguistics, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden
Michael.Dunn@lingfil.uu.se

Outi Vesakoski
Collegium Researcher, Department of Biology, University of Turku, Turku, Finland; Department of Finnish language and Finno-Ugric linguistics, University of Turku, Turku, Finland
outves@utu.fi

Abstract

Loanwords carry information on linguistic interactions, and can also reveal (pre-)historical population contacts. The contact history of a particular language family is an essential component of historical linguistics, but it is also illuminating for integrative studies of the human past. However, data availability and the time-consuming nature of etymology mean that comprehensive research on loanword layers exists for rela-

tively few languages, forcing us to rely on limited material for others. This paper compares the loanword layers in the basic and total vocabulary of six well-studied Uralic languages, assessing how accurately the borrowing profile in basic vocabulary reflects the full profile of a language. We define "borrowing profile" as the known contact history of a language reflected by its loanword layers. We demonstrate that the loanword layers in basic vocabulary provide an adequate cross-section of the full borrowing profile, although basic vocabulary manifests prehistoric contacts more strongly than more recent contacts.

Keywords

lexical borrowing – loanword typology – uralic languages

1   Introduction

The vocabulary of a language carries information about its history. Inherited words reflect genealogical, or vertical, relationships of languages within a language family. In turn, loanwords represent horizontal connections between different languages. Loanword layers reveal connections between languages and are indicative of prehistoric and more recent speaker population contacts. Research on linguistic contacts thus provides hypotheses on population contacts, and is currently an under-exploited field in the context of multidisciplinary research on human (pre-)history. Unfortunately, studies that cover contact influence of a language's entire vocabulary do not exist for most of the world's languages because etymological research is laborious and limited by existing linguistic material. This imbalance hinders the potential for cross-linguistic comparison of contact influence or interdisciplinary applications. In this paper we suggest a new approach for studies of linguistic contacts of a language by examining if the borrowing profile of a language, i.e., the representation of its contact history in linguistic material as provided by its basic vocabulary, can adequately represent the contacts that that language has historically experienced. We investigate the borrowing profiles of languages in the Uralic family, by tracing their contact history through loanword layers, by curating the basic-vocabulary lists of the languages, and then by testing whether the borrowing profile of the basic vocabulary represents the profile of the entire vocabulary. We hypothesize that such basic-vocabulary lists with borrowing information can provide an overall picture of a language's contact history. Thus, we want to discover if and how the borrowing profiles of basic vocabulary can be expected to vary systematically with regard to the overall contact history.

Loanword typology is a well-studied area of contact linguistics (Matras, 2009: 167). Loanwords are borrowed over different periods of the history of a language and if sufficiently numerous form a loanword stratum. Contacts in different periods commonly reflect different sociolinguistic situations, and the contents of these strata in turn can reflect these. This has implications for research in e.g., archaeology and genetics. Loanwords within the kinship system, for example, might be evidence for intermarriage, whereas loanwords restricted to semantic domains like trade goods suggest contact where population admixture is less likely. Thomason and Kaufman's (1988: 74–76) classic model on the intensity of the contacts offers a useful framework for extrapolating from these different patterns of borrowing to different types of sociolinguistic setting.

While we cannot observe historical sociolinguistic settings directly, synchronic contact linguistics provides the necessary framework to infer past processes from historically determined patterns. The synchronic study of language contact involves investigating the outcomes of communication between bi- and multilingual speakers in various environments. These outcomes have often been discussed from the viewpoint of language-internal constraints (e.g., Haugen, 1950), but in the past decades the role of language-external social factors as the main causes and predictors of contact-induced change has been recognized. The roots of the study of social aspects lie in dialectology and variation studies: in a classic study Labov (1963) investigates social identity signaling as a motivation for phonological variation, and Milroy and Milroy (1985) study the role of social networks in the diffusion of linguistic innovations. The psycholinguistic aspects of contact-induced change have been studied in terms of the communicative capabilities a multilingual individual possesses and employs in different settings (e.g., Croft, 2000; Matras, 2009; Ross, 2007). These studies, based on comprehensive synchronic data, offer multiple causation for linguistic change. For instance, the motivations for lexical borrowing have expanded beyond the commonly discussed dichotomy of "need" and "prestige" to include e.g., the competence, cognitive pressure and volition to use an extended linguistic repertoire made possible by multilingualism (e.g., Matras, 2009: 152; Muysken, 2013: 726). The aggregate of these synchronic processes over historical time gives rise to the particular borrowing profile of a language.

While it has long been established in contact linguistics that anything can be in contact linguistics that anything can be borrowed (Thomason and Kaufman, 1988: 14), there is likewise a consensus that constraints on borrowing nevertheless do exist (Curnow, 2001: 417–418). Certain categories within the vocabulary seem to be more prone to borrowing than others, with nouns emerging as most borrowable (Matras, 2009: 157; van Hout and Muysken, 1994: 42; Haspelmath

and Tadmor, 2009: 61). In outcomes of language shift, however, the consequences of language contact may only have a minor impact on the vocabulary and are more likely to manifest themselves in phonology and syntax (Thomason and Kaufman, 1988: 118). This study takes vocabulary as a starting point for establishing the borrowing profile of languages, because lexical loans are more readily identifiable (Thomason, 2008: 49). Lexical borrowing involves the acquisition of distinguishable linguistic matter – the adoption of both the form and function of a word (Johanson, 2002: 291; Matras, 2009: 148) – and as such, borrowing into the vocabulary leaves the most uncontroversial evidence of another language's presence in a language.

The lexicon of a language is not amenable to exhaustive listing, and therefore standardized, limited wordlists have been compiled which can be used to compare languages on an equal basis. The best known of such wordlists is the so-called "Swadesh List" of basic vocabulary (available in 100- and 200-word versions; Swadesh, 1952; 1955), but various other partially overlapping lists for basic-vocabulary meanings also exist (e.g., Dolgopolsky, 1964; Haspelmath and Tadmor, 2009; Wichmann et al., 2018; overlap discussed in Syrjänen et al., 2013). Basic vocabulary can be characterized to include meanings essential for human interaction e.g., body parts, close kin, simple actions and low numerals (e.g., Campbell and Mixco, 2007: 24–25). While there is no (and perhaps cannot be a) clear-cut definition as to which concepts define the basic vocabulary of a given language, a number of typical characteristics are commonly encountered in the literature. Words belonging to basic vocabulary are allegedly borrowed less often than words for cultural concepts, thereby retaining cognate relationships (cf. e.g., Greenberg, 1957). Further, basic meanings are described as frequent (Trask and Millar, 2015: 23), tend to be morphologically simple (Haspelmath and Tadmor, 2009: 72), and therefore semantically neutral. Due to these additional criteria described above, basic vocabulary lists do not only contain strict cognates but also words from other semantic sources, i.e., basic concepts are also acquired through semantic change and other type of linguistic innovation (Chang et al., 2015: 204). The lack of a clear-cut definition for basic vocabulary is addressed by our choice of data, which combines several word lists. Borrowing-resistance is the most important quality of the notion of basic vocabulary but in reality any linguistic trait is borrowable, including basic lexemes (Curnow, 2001: 412), and, as will be illustrated below, basic vocabulary lists definitely do contain loanwords. The critical circumstance for the purposes of this study is that the basic vocabulary of many languages is widely available, even where other linguistic documentation is lacking.

The main aim of this paper is therefore to examine whether the loanword layers detectable in basic vocabulary can be used as a proxy for the borrowing

profile of an entire language. We use Uralic languages as a test case, focusing on six well-studied languages from different subgroups (for location of these, see https://sites.utu.fi/urhia/language-maps/, Rantanen et al. 2022.) For each language a basic-vocabulary list is used with known borrowings identified. We use quantitative surveys of loanword layers as comparison material to represent the entire vocabulary, using qualitative overviews if quantitative material is not available. We explore how the basic vocabulary reflects the current understanding of the borrowing profile of the language from two perspectives:

(1)   Are the loanword layers observed in the entire vocabulary also present in the basic vocabulary?

(2)   Do the relative sizes of the loanword layers in the entire vocabulary correspond to the relative sizes of those layers (if present) in the basic vocabulary?

We do not assume that the basic vocabulary features every single loanword layer of the language since e.g., especially the borrowing-prone domain of cultural vocabulary is usually not strongly represented. Due to the supposed conservatism of basic vocabulary, specific loanword layers might be under- or over-represented, and the borrowings in basic vocabulary might not follow the general stratification of loanwords in the language. We further evaluate the fit of the borrowing profiles of basic vocabulary and whole vocabulary through three simplified scenarios:

(3)   a.   The loanword layers in the basic vocabulary follow the relative sizes of the loanword layers in the whole vocabulary.
      b.   Loanword layers are underrepresented in the basic vocabulary.
      c.   Loanword layers are overrepresented in the basic vocabulary.

We assess the accuracy of borrowing profiles inferred from basic vocabulary, discuss the results, and lay out possible explanations for the observed patterns in light of what is known about the sociolinguistic settings of the population contact. We also discuss possible applications of the borrowing profiles and how the borrowing profiles could be refined in future work.

2   Materials and Methods

2.1   *The Uralic Language Family and the Focal Languages*
The Uralic language family consists of approximately 30 languages spoken mostly in Northern Eurasia. Uralic linguistics has a long tradition of comparative research and loanwords have always been a central research interest (see
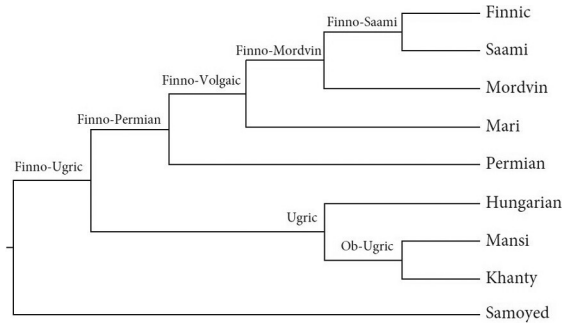
FIGURE 1   A classification of the Uralic languages according to Korhonen (1981) featuring an initial binary split into Finno-Ugric and Samoyedic
Note: The high-level subgroups (Finnic, Saami, Mordvin, Mari, Permian, Ugric and Samoyed) are uncontroversial; the exact lower-level relationships are still disputed. The intermediate Finno-Volgaic proto-language no longer has mainstream support in Uralic linguistics and the features shared by the so-called "Finno-Volgaic" subgroup likely reflect areal connections. This illustration was first published in Syrjänen et al. (2013).

Wickman, 1988). The Uralic languages are therefore well suited for a cross-linguistic quantitative study on loanword strata. We test the representativeness of basic vocabulary loanword profiles with Estonian, North Saami, Erzya Mordvin (henceforth: Erzya), Komi-Zyrian (henceforth: Komi), Meadow Mari (henceforth: Mari) and Hungarian. Each language represents one subgroup in the Finno-Ugric branch of the traditional Uralic classification (Fig. 1). The easternmost Uralic languages belonging to the Ob-Ugric and Samoyed groups are not included due to limited data availability. As the focus of this study is to assess general borrowing profiles for the Finno-Ugric branch, only one language was selected for each subgroup. These six languages were selected because their vocabulary has been thoroughly researched. We briefly discuss five other, less-researched languages in our data representing all subgroups of Uralic in the Supplement (available on Zenodo as De Heer et al. 2023, https://doi.org/10.5281/zenodo.7716522) (Table S12).

2.2   *Loanwords in UraLex Basic Vocabulary*

The analysis presented in this article is built on a dataset of Uralic basic vocabulary (Syrjänen et al., 2018) published as lexibank/UraLex: UraLex basic vocabulary dataset (Version v1.0) [Data set]. For the purposes of this analysis, the label

basic vocabulary specifically refers to this UraLex dataset. UraLex comprises lexical data from 26 Uralic languages as well as reconstructed Proto-Uralic. The language sample covers all recognized subgroups (see Fig. 1 for a classification). UraLex is composed of a union of the basic vocabulary meaning lists by Swadesh (1955), which between his 100- and 200-word lists contain 207 unique meanings, as well as the 100-item Leipzig-Jakarta list (Haspelmath and Tadmor, 2009). UraLex also includes the meanings from the World Loanword Typology with borrowability rankings from 401–500 (Haspelmath and Tadmor, 2009; Lehtinen et al., 2013). These ranks were selected for UraLex to produce a larger data sample while still being relatively resistant to replacement. The total number of meanings is 313, as the basic vocabulary lists overlap somewhat. UraLex aims for semantically neutral words, but it allows for synonymy in cases where the selection of single words is not possible and a strict requirement of a single word would otherwise lead to missing data. All words are coded as root-meaning traits, meaning that the words between languages were assigned to classes based on their shared ancestry (Syrjänen et al., 2018). The data-collection process is explained in Syrjänen et al. (2018).

For the present study, we expanded the UraLex data with borrowing information from the etymological literature, and we distinguished the root-meaning traits acquired through borrowing from inherited and other non-borrowed words. We added information regarding the donor languages from the literature on loan etymologies, using guidelines pertaining to recognizing loanwords in Uralic etymological literature (Junttila, 2015: 54–55). For source information, we utilized three types of etymological literature on the Uralic language family: primary research literature, which discusses and presents loan etymologies in their full context; etymological dictionaries; and evaluative literature, which is dedicated to critically reviewing etymologies (Junttila, 2015: 66). We used two information types following Junttila's (2015: 60–62) typology for evaluating the acceptability of loan etymologies: i) a simple statement of origin, and ii) commentary if available. A "statement of origin" refers to an etymological claim presented in literature (Junttila, 2015: 60), while "commentary" consists of comments which critique, support and analyze the existing options for the origin of a word (Junttila, 2015: 61).

We incorporated uncertainty into the loanword information with a classification based on three degrees of certainty. This allows us to update the data as and when new etymological studies appear with regard both to new etymologies and to the reassessment of old etymologies. We utilized the evaluative literature for assigning certainty when possible. This type of literature considers various kinds of arguments contributing to the reliability of a loan etymology, such as phonological substitution arguments, the meaning of words, and

the distribution of the words across language families (Junttila, 2015: 137–188). When not available, we based the certainty estimate on certainty tags (e.g., question marks in dictionaries) and the discussion comments in etymological literature. Borrowings marked as *clear* are mostly undisputed or transparent; they are explicitly noted as accepted in the literature and are not subject to certainty tagging or critical-discussion comments (Junttila, 2015: 79). *Probable* borrowings are generally accepted with some inconsistencies and critical commentary (Junttila, 2015: 80). *Possible* borrowings are still subject to discussion; they are not immediately transparent and their status is unevaluated in the literature.

The bulk of the words in the UraLex dataset are not identified as loanwords. These items include Proto-Uralic words and other endogenous items (e.g., derivations and words with a suggested onomatopoetic origin) as well as potential currently unidentifiable loanwords. All these items are assigned to a *Not borrowed* category because they show no evidence of borrowing.

The detailed loanword information compiled for and analyzed in this paper as well as all information for the languages not discussed here is published in conjunction with this publication in an updated version of the UraLex dataset called lexibank/UraLex: Uralic basic vocabulary with cognate and loanword information (Version v2.0) [Data set] (De Heer et al., 2021). The full UraLex 2.0 dataset is published as an open-access resource on the Zenodo repository; source literature is additionally listed in the Supplement.

There are earlier surveys on borrowings in Uralic basic vocabulary using different datasets. A survey on Kildin Saami (Rießler, 2009) was published in the World Loanword Typology Database (= WOLD) (Haspelmath and Tadmor, 2009) and is the only Uralic language represented there; a similar survey of the full WOLD meaning list for Finnish is Cronhamn (2018). A study focusing on loanwords and word-acquisition strategies in a strictly synonymy-free Swadesh list of Estonian, Ingrian dialects and Vote using a lexicostatistic framework is Rozhanskiy and Zhivlov (2019).

2.3   *Loanwords in the Large Vocabulary Stocks*

The dictionary data represents larger vocabulary stocks which contain all known loanword layers and capture the relative sizes of the layers in a language via absolute numbers of loanwords. We use calculations from existing surveys on the newest (etymological) dictionaries of Estonian (= EES) (Soosaar, 2013) and Hungarian (= EWUng) (Keresztes, 1998) and the large Mari–German dictionary (= TschWB) (Saarinen, 2010).

For comparison data for Komi, we counted the borrowings in the etymological dictionary of Komi (KESK, 1970) and its supplement (Lytkin and Guljaev,

1975). We obtained a numerical range of loanwords based on the number of undisputed borrowings in loanword categories and the maximum number of all possible borrowings, including uncertain ones. The KESK data was adapted according to the labels of loanword layers selected for UraLex (See Tables S6–7 for a full breakdown and labeling of the KESK data).

For Mari, comparison to a larger vocabulary stock is possible because of the categorization of items in the Mari dictionary (= TschWB) (Saarinen 2010). Although TschWB is not an etymological dictionary, it does contain information on the source languages of (newer) loanwords in the entries. In order to increase the resolution of the Mari dictionary (= TschWB) data, we counted the archaic Indo-European and Indo-Iranian borrowings in Mari suggested in the literature (Holopainen, 2019; Joki, 1973; Koivulehto, 2016; Rédei, 1986); only the items present in the dictionary were added. These counts were considered together with the handful of such borrowings already tagged in TschWB.

There are no large etymological dictionaries of Erzya (Cygankin and Mosin, 2015 and Veršinin, 2004 not reaching the reliability and the size needed for our comparison) or North Saami. We collected rough estimates on the sizes of loanword layers to provide a point of comparison for these two languages. In this paper, we refer to the large vocabulary stocks represented by the dictionary data and the vocabulary the more general loanword-layer size estimates pertain to, as *whole vocabulary*.

2.4   *Labeling the Loanword Layers*

Our focal languages acquired borrowings throughout their history, but the sources categorize and label the loanword layers inconsistently. To compare the loanword layers within and across languages, we employ three different ways of labeling: 1) labels that capture as much detail as possible on the source language; 2) rougher subgroup-based labels for examining the contact influence from a wider perspective, and 3) labels taken from large vocabulary stocks for comparison of loanword-layer sizes in whole and basic vocabulary. These cover terms capture varying levels of detail. The layers occurring in UraLex are discussed in the results focusing on the source-language groups, as the focal languages often share borrowing sources through their ancestral stages.

First, labels expressing a higher degree of resolution are used in Sections 3.2–3.4 and in the Supplement. Many loanwords present in the various target languages were borrowed from earlier stages of the donor languages (usually referred to as "Proto X") before the target languages' independent existence; e.g., the Germanic loans in Estonian refer to loans borrowed from Proto-Germanic into Proto-Finnic, not into modern Estonian. We chose the labels to retain comparability and to eliminate redundancy, as the level of detail in the anal-

ysis varies in the sources of the focal languages. The categories which cover chronologically different stages of a borrowing source or which might have less transparent labels are defined below:

Finnic includes internal borrowings from Proto-Finnic and modern Finnic languages. The Finnic items cannot usually be stratified clearly or to a comparable degree in this paper.

Indo-European includes all words that are of unclear Indo-European origin where the exact source language had not been defined. In addition, this category includes *Wanderwörter* diffused across Uralic languages with the meanings 'salt, salty' of which the Indo-European source cannot be determined. Two North Saami items in UraLex, *gutna* 'ashes' and *arvi* 'rain', are grouped in this category because the literature (Koivulehto, 2001: 288) assumes that the source of these items is not North-West Indo-European (henceforth "NWIE", see below), but another branch of Indo-European contemporaneous to NWIE.

Iranian includes all Iranian borrowings, from Proto-Iranian to Middle Iranian languages. The Iranian stages are combined because the Iranian source languages are often not specified in sources.

North Germanic includes Proto-North-Germanic words and items labeled "Scandinavian" in sources with no further details. In addition, modern North Germanic loanwords without commentary in the sources are also added to this category since the sources lack detail.

North-West Indo-European (NWIE) is a category of archaic Indo-European borrowings from the presumed ancestor of Germanic and Balto-Slavic (Koivulehto, 2001: 239). These items have a narrower distribution than the words grouped in the Proto-Indo-European category (Koivulehto, 2001: 239). These items are only found in Finnic, Saamic, Mordvinic, Mari and Permic. Items of so-called Pre-Baltic and Pre-Germanic origin refer to this source and are therefore allocated to the NWIE category.

Permic refers to borrowings from Proto-Permic borrowed into other Uralic languages.

Unknown language refers to an extinct non-IE, non-Uralic language once spoken where North Saami is now located, a so-called paleo-language (Aikio, 2004: 5). This category only concerns the North Saami data.

Volga Bulgar refers to borrowings acquired from a language of the Oghur Turkic branch of the Turkic family.

> *Western Uralic* refers to extinct languages bearing resemblances to
> Finnic (Saarikivi, 2018: 271).
>
> *West Old Turkic* refers to Oghur Turkic borrowings in Hungarian
> acquired between 500–1200 AD. The label is adopted from Róna-Tas
> and Berta (2011).

Second, we present a macro-scale borrowing profile of the focal languages
using an even broader categorization of the borrowing sources (conversely, see
a more fine-grained categorization for the individual languages in the Supple-
mentary Materials). The following macro-categories are used to summarize the
results (Fig. 3):

> The *Archaic Indo-European* label combines Proto-Indo-European, Proto-
> North-West Indo-European and other archaic Indo-European items
> of which the exact source language is not identifiable.
>
> *Indo-Iranian* includes all items from Proto-Indo-Iranian, Proto-Iranian
> and Middle Iranian languages.
>
> The *Balto-Slavic* category combines items of Proto-Balto-Slavic origin,
> loanwords from the separate Baltic and Slavic branches and mod-
> ern Balto-Slavic languages, Russian being the most relevant for the
> UraLex dataset.
>
> The *Germanic* category includes items of Proto-Germanic origin and
> Proto-North-Germanic as well as younger loanwords from West Ger-
> manic and North Germanic languages.
>
> All loanwords acquired from languages from the Turkic language family
> are conflated into the single label *Turkic*.
>
> *Western Uralic languages* include all borrowings acquired from other
> Uralic languages such as Proto-Finnic and Finnish.

Third, we compare the UraLex data to larger vocabulary stocks (see Section 3.5);
we derive this comparison data from previous dictionary surveys (see Section
2.3 for details). Where the surveys do not provide enough resolution, chronolog-
ically different UraLex categories were conflated into broader composite cate-
gories marked by an ampersand, e.g., *Indo-Iranian and Iranian*, which refers to
all items from Proto-Indo-Iranian and all stages of Iranic. Conversely, because
the borrowing sources in UraLex could not always be consistently and com-
parably tagged with regard to a specific language variety or dialect, such cate-
gories are conflated in the comparison data, e.g., for the Estonian comparison
the Swedish and German varieties are conflated under the labels *German* and
Swedish. (Fig. 4A). Proto-languages are referred to in the same way as described

above. The full labeling of the comparison data is presented in Tables S2, S4–5, S7, S9–10.

2.5 Comparison of Loanword Strata between Basic Vocabulary and Large Vocabulary Stocks

We first compare the proportions of loanwords in the basic vocabularies of the six focal languages. Next, we examine whether the borrowing profiles of the basic vocabulary reflect the loanword layers of the whole language quantitatively, comparing the percentages of borrowings in each category of the UraLex basic vocabulary lists vs. the dictionary word stocks for Estonian, Komi, Mari and Hungarian.

We visually assess the accuracy of the borrowing profiles in UraLex through scenarios 3A–C described above. To assess the scenarios 3A–C, we plot a line indicating an expected scenario where the proportions of loanwords in basic and whole vocabulary match, i.e., categories representing a high (or a low) proportions of loanwords in the dictionary word stocks also have similar relative high (or low) proportions in the UraLex basic vocabulary data. The comparisons use two types of data points due to differences in the word-stock data. First, we use averages of loanwords in a category for Estonian and Komi. The large word-stock surveys provide a range of undisputed and all possible loanwords in a loanword layer. For UraLex, a range can be established using the clear-status items and the maximum number of loanwords, which are the sums of clear, probable and possible items in a category. We calculated the average of undisputed and all borrowings to estimate the size of each loanword layer considering the uncertainty expressed by the range. Next, we calculated the percentage of borrowings that the average of loanwords in a category represents from all items in the dictionary or UraLex, respectively. The second type uses percentages calculated with the maximum counts of loanwords in a category, because the Hungarian and Mari dictionary surveys do not express a degree of certainty. Therefore, we use percentages calculated with the maximum counts of loanwords in a category for UraLex as well. A summary of all analyses and information types used in this paper is presented in Table S11.

For Estonian, Komi, Mari and Hungarian, we visually assess the fit of the test scenarios by studying the location of the data points in relation to the predicted line:

A. The data points scattered around the prediction line indicate cases where the relative sizes of the loanword layers in basic vocabulary and the whole vocabulary are congruent (Fig. 2A).
B. The data points emerging in the upper-left corner indicate loanword layers which are underrepresented in basic vocabulary (high value in the whole vocabulary, low value in basic vocabulary) (Fig. 2B).
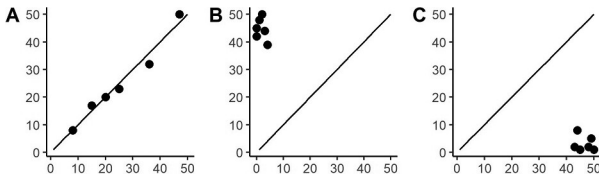
FIGURE 2   Illustrations of different scenarios based on
             whether the proportions of loanwords would be (A)
             equally large (or small) in basic and whole vocab-
             ulary, (B) underrepresented in basic vocabulary or
             (C) overrepresented in basic vocabulary

C.   The data points emerging in the lower-right corner indicate loanword lay-
        ers that are overrepresented in basic vocabulary (low value in the whole
        vocabulary, high value in basic vocabulary) (Fig. 2C).
In cases where it was unclear which scenario the data points follow, we closely
examined the percentages indicating the size of the loanword layer in basic and
whole vocabulary. This allows us to evaluate all data points according to scenar-
ios A–C. The percentages used in the quantitative comparisons are laid out in
Tables S3, S8–10. For North Saami and Erzya we could not conduct these com-
parisons as comparable whole vocabulary data is lacking. Instead, we assess the
fit of the scenarios qualitatively by comparing the prominence of loanword cat-
egories to rough estimates described in the literature (Tables S4–S5). In order
to provide a point-of-comparison with other language families which may not
have extensive basic vocabulary data, we briefly present the key results medi-
ated by the Swadesh-100 in the results and in the discussion.

3   Results

3.1   *Proportions of Borrowings in UraLex*
The number of borrowings in UraLex naturally varies from language to lan-
guage. We present here the main components of the borrowing profiles in the
focal languages; the exact counts and percentages of loanwords in a category
per focal language can be found in the Supplement (Fig. S1). General estimates
of numbers of loanwords from the literature as well as counts and percentages
of the whole-vocabulary comparison data are shown in detail in the Supple-
ment. The majority of items in the basic-vocabulary lists do not show any evi-
dence of borrowing (Fig. 3). The focal languages contain on average 24% of
borrowings in basic vocabulary. North Saami features the largest percentage of
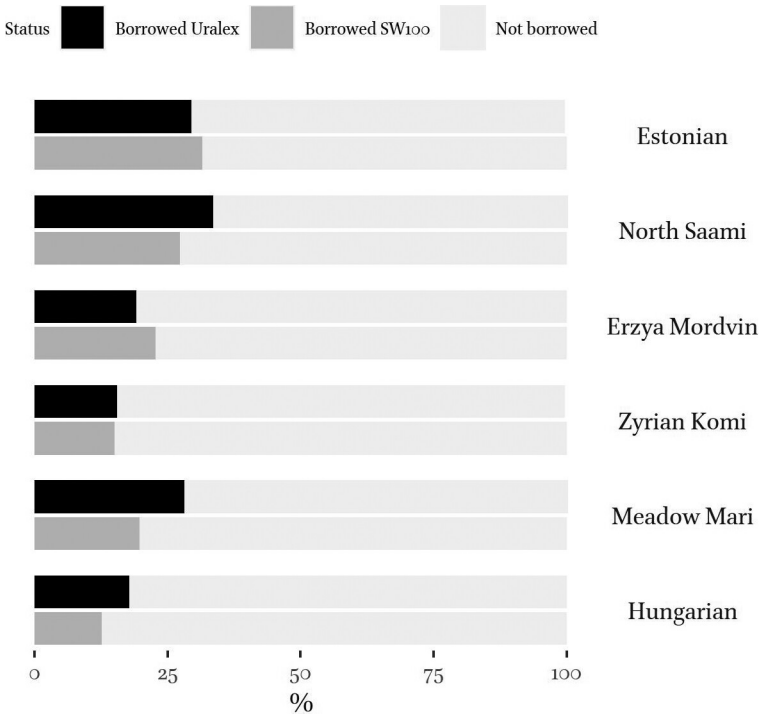borrowings; Komi the lowest (Fig. 3).

FIGURE 3   The percentages of borrowed and non-borrowed items in the basic
            vocabularies of the six focal languages
            Note: The black bars represent the borrowed proportion in UraLex
            and dark gray bars represent the borrowed part in the Swadesh-
            100 lists. The proportion of items with no evidence of borrowing is
            marked with light gray. See Fig. S1 and Table S1 for exact numbers and
            Fig. S2 for a breakdown by basic-vocabulary list.

The differences between the borrowing percentages between the full list
and Swadesh-100 can be explained by the synonymy attribute of UraLex 2.0.
The average borrowing percentage for Swadesh-100 is 22 %, slightly lower than
the average of the complete UraLex 2.0 lists. A full breakdown of the loanword
layers and their representation in the Swadesh-100 list can be found in the sup-
plementary tables (S2–5, S7, S9–10), which also give the same information for
the longer Swadesh-207 lists for additional context.

3.2   *Indo-European Loanword Layers*

Indo-European languages are the most common sources of loanwords. This
is also conveyed by the macro-scale borrowing profile (Fig. 4) of the focal
languages characterized by influence from the Balto-Slavic and Indo-Iranian
branches, as well as archaic Indo-European influence which cannot be allo-

cated to any one independent branch. Contact influence acquired from the Germanic branch is practically only seen in Estonian, representing the Finnic subgroup, and in North Saami, representing Saamic.

3.2.1 Archaic Indo-European, Indo-Iranian and Iranian layers

A handful of ancient Indo-European loanwords have been proposed to have been acquired already by Proto-Uralic (e.g., Koivulehto, 2001: 235) or even Pre-Proto-Uralic (Häkkinen, 2012: 5); there are, however, conflicting opinions on how many such loanwords exist, or even if they exist at all (see e.g., Simon 2020). Another layer of phonologically archaic words is only found in the western branches. These narrowly distributed words are categorized in the literature as so-called North-West Indo-European borrowings and reflect a Proto-Indo-European phonological level of reconstruction (Koivulehto, 2001: 239).

Basic vocabulary shows that the most archaic Indo-European borrowings are shared among all the six focal languages. In comparison to the other focal languages, the archaic Indo-European loanword layers are slightly more prominent in Finnic and Saamic than in other subgroups (Fig. 4).

Borrowings from Indo-Iranian and Iranian represent chronologically diverse layers, the earliest being acquired into dialectal Proto-Uralic and others later independently into different Uralic language sub-groups (Holopainen, 2019: 343; see also Kümmel, 2020 for an overview). Influence from the Indo-Iranian languages is prominent in the UraLex data of all the focal languages (Fig. 4).

The Iranian loanwords visible are likely from Proto-Iranian and Middle Iranian languages (see Holopainen, 2019: 34–35 for discussion on these designations). Relatively recently Iranian languages affected the languages spoken in the Volga region and Hungarian (Korenchy, 1988: 675). However, the numbers of borrowings restricted only to Erzya, Mari and Hungarian present in basic vocabulary are low (Tables S5, S9–10) and mostly belong to the possible certainty category. While the chronological stage of the Indo-Iranian source is often ambiguous, the recent evaluative work analyzing a large body of Indo-Iranian etymologies in light of current views on Uralic historical phonology (Holopainen, 2019) has allowed us to assign some of the borrowings in the basic vocabulary a clear certainty estimate. However, a critical evaluation of the proposed Indo-Iranian etymologies and challenges regarding the stratification have also reduced their numbers, especially for Saamic (Holopainen, 2018: 170).

The layers of Indo-Iranian origin are also strongly present in Komi (Fig. 4, Table S7). This was to be expected because contacts have been postulated especially between Proto-Permic and Middle Iranian languages (Holopainen, 2019:

377, Metsäranta, 2020: 193–195). The loanword layers in basic vocabulary are likely acquired by Proto-Permic, thus contributing to the Indo-Iranian proportion in the macro-scale borrowing profile together with loanwords adopted already during Pre-Permic times (Fig. 4).

3.2.2 Germanic and Balto-Slavic Layers

The Finnic and Saamic groups were both influenced by the Germanic and Balto-Slavic branches. The Proto-Finnic–Proto-Baltic and Proto-Finnic–Proto-Germanic contacts were long-lasting and contemporaneous, leaving a strong, multi-layered influence on the vocabulary of the Finnic languages (Kallio, 2006: 13–14). The Baltic influence on the Finnic subgroup eventually diminished, but contact with Germanic continued until modern times (Junttila, 2012: 265). For Saamic, the major source of Indo-European influence consists of Proto-North Germanic borrowings and loanwords from modern North Germanic languages (Aikio, 2006: 9).

Counted in their hundreds (Junttila, 2015: 255; Kallio, 2012: 231; LÄGLOS) and covering many semantic fields, the prevalence of Germanic and Baltic loanwords in the Estonian basic vocabulary is unsurprising. The most prominent loanword category is Germanic (Fig. 3). The Baltic category contains the highest number of clear items, resulting from rigorous evaluative work conducted in recent years. A few borrowings acquired independently into Estonian from more recent languages, e.g., German and Latvian, are visible in UraLex, but their influence on basic vocabulary is inconsequential.

In the North Saami basic vocabulary, influence from Germanic is strong. The large proportion of clear status items can be attributed to the intense contacts with North Germanic. The Baltic category is weakly represented in North Saami basic vocabulary (Fig. 4). The Baltic borrowings in Saami are said to be the result of instances of direct contact but also due to the mediation of loanwords via Finnic, indicating a more casual relationship between Baltic and Saamic (Aikio, 2006: 40). Further, Baltic influence decreased over time, contributing to the lower number of possible borrowings (Junttila, 2012: 265). In comparison, the long-standing and direct contacts of Saamic with Finnic and Germanic (continued by Scandinavian) have led to a large body of loanwords, both of which are strongly reflected in the basic vocabulary (Fig. 4).

Baltic influence has affected the Erzya vocabulary with a few dozen loanwords (Grünthal, 2012: 297; van Pareren, 2009: 234; further analyzed in Junttila, 2018). Baltic items form a distinct category in the Erzya basic vocabulary (Table S5) and contribute to the branch-level Balto-Slavic influence visible in basic vocabulary (Fig. 4). However, most of the Baltic loanwords in the Erzya data are assigned to the uncertain possible category (Fig. S1).

Slavic and Russian have influenced all focal languages to varying degrees. While they have affected the cultural vocabulary of all Finnic languages, in the Estonian basic vocabulary Slavic influence only weakly features (Table S2). The few Slavic borrowings in the Hungarian UraLex data are all clear cases. A majority of the Slavic loanwords were acquired after the Hungarian conquest of the Carpathian Basin in the 9th century, and contacts with the surrounding Slavic languages have continued (Gerstner, 2006: 311), but this influence has mostly not reached the basic vocabulary.

Influence from Russian on the Uralic languages spoken in Russia is ubiquitous and ever increasing. Russian is especially well-represented as the most prominent category in the basic vocabularies of Erzya and Komi (Tables S5, S7). This is expected since both languages are spoken in Russia and have thousands of Russian borrowings. In general, Russian is the most important source of influence from the Balto-Slavic branch in the full UraLex data in both focal languages (Fig. 4). Russian accounts for the majority of the clear portion for both Erzya and Komi (Fig. 4).

Within the macro-profile conveyed by the focal languages, the Balto-Slavic category has the largest clear-status proportions of which the majority can be explained with the recent and transparent Russian influence on the focal languages.

3.3 *Turkic Layers*

The second major source of loanwords in Uralic is the Turkic family, which is reflected to varying degrees in the basic vocabulary (Fig. 4). We distinguish between the two main branches of the Turkic family: Common Turkic, which includes all Turkic languages except one, and Oghur Turkic, with only one surviving member (Chuvash). Uralic has also been in contact with extinct Oghur Turkic languages, mainly Volga Bulgar. Both Oghur Turkic and Common Turkic have influenced Erzya, Mari and Hungarian, while Komi only has old Oghur Turkic loanwords. There has been no contact between Turkic and Finnic or Saamic.

Contact between Proto-Permic and Volga Bulgar is the source of basically all Turkic loanwords in the Komi basic vocabulary (Rédei and Róna-Tas, 1972: 297), (Fig. 4). Only a few Volga Bulgar words are found in the Mordvinic languages (Butylov, 2007: 32), but single items still are visible in the Erzya basic vocabulary (Table S5). The Volga Bulgar loanwords tend to be clear.

Oghur Turkic influence is the strongest in Mari with hundreds of borrowings from Volga Bulgar or from its descendant or close relative Chuvash. Contact between Mari and Oghur Turkic-speaking groups has been intensive as these groups have been living closely together for centuries (Agyagási 2012: 26).

Results of these long-lasting contacts are visible in UraLex where Volga Bulgar and Chuvash influence is most prominent, also accounting for most clear status items (Fig. 4).

Hungarian and Turkic have been in contact before and after the Hungarian conquest of the Carpathian Basin during the 9th century. The earliest Turkic borrowings are from Volga Bulgar, and more broadly, so-called West Old Turkic, referring to Oghur Turkic languages spoken in the 5th–12th centuries, has left hundreds of loanwords in Hungarian (Róna-Tas and Berta, 2011: 1143). This influence is strongly present in the full Hungarian UraLex data where West Old Turkic loanword strata are most prominent and have the largest proportion of clear items of the Hungarian basic vocabulary (Fig. 4). Later Hungarian acquired minor loanword strata from other Turkic languages; these are only weakly represented by single items in UraLex (Fig. S1).

Tatar, a Common Turkic language, is an important source of cultural vocabulary for the languages spoken in the Middle-Volga area, and Erzya has borrowed hundreds of Tatar loanwords (Bartens, 1999: 17; Butylov, 2007: 46; Zaicz, 1998: 214). As these Turkic words are mostly concepts related to e.g., housekeeping (Paasonen, 1897: 22, Butylov, 2007: 81), they are weakly visible in the Erzya basic vocabulary (Table S5). Mari has thousands of Tatar loanwords (Saarinen, 2010: 338; Table S9), with considerable differences between the dialects. The Mari UraLex data contains a strong Tatar layer, where most items have a clear status (Table S9).

3.4 *Other Loanword Layers*

Many Uralic languages have been in contact with each other during their history, but research in this area is underdeveloped. Only those family-internal relations responsible for loanwords in UraLex are briefly presented here.

Saamic and Finnic have been in close contact; the number of Finnic loanwords in Saamic can be counted in the thousands as a consequence of active contact facilitated by geographical proximity. Etymological nativization makes it difficult to stratify the Finnic and Finnish borrowings (Aikio, 2012: 68). They comprise the largest loanword stratum in the North Saami basic vocabulary and this stratum contains mostly clear borrowings (Fig. 4). However, etymological nativization makes it difficult to stratify the Finnic and Finnish borrowings (Aikio, 2012: 68). Etymological nativization, also known as 'correspondence mimicry' (e.g., Alpher and Nash, 1999; Dench, 2006: 117–118), 'loan nativization' (e.g., Leer, 1990: 88), or 'loan adaptation' (e.g., Gardiner, 1983), refers to speaker's competence in applying an awareness of regular sound correspondences between (often but not necessarily) closely related varieties: by way of this 'correspondence mimicry' loanwords reflect the same sound correspon-
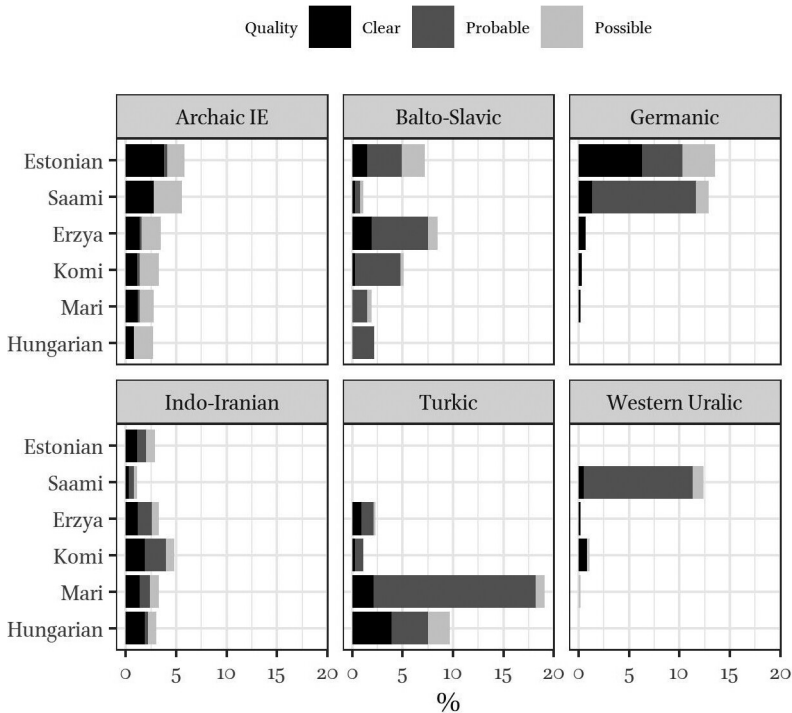
FIGURE 4   A macro-scale summary of contact influence in the basic vocabulary of
the focal languages
Note: The percentages of borrowings as well as certainty estimates in
UraLex are grouped roughly according to the Indo-European donor lan-
guage subgroups. All languages from the Turkic and Uralic families are
summarized as single labels. Categories containing single items with an
unclear source language, e.g., Finnic or Germanic, are omitted here (see
Supplement for detailed categories).

dences as inherited words, which then cannot be easily distinguished from
one another. It has also been reported, under various designations, in, e.g., Aus-
tralian languages (cf. Evans, 1998), Baale (Nilo-Saharan) (cf. Dimmendaal, 2006:
361, 363), Papua New Guinean languages (cf. Ross, 1997), Athabaskan (Leer,
1990), and Tatar (Benzing and Menges, 1959: 2). That such awareness occurs
among speakers of Uralic languages was already mentioned in the literature 70
years ago (e.g., Itkonen, 1961: 53). See Aikio (2007) for an overview, especially
with regard to Saamic and Finnish.

Hundreds of words in Saamic are of as yet unknown origin (Aikio, 2004: 8;
2012: 83); they are generally ascribed to substrates. They probably contribute
to a percentage of non-borrowed data in UraLex (Fig. 3). One item is visible in
UraLex (Fig. 4).

A few Western Uralic borrowings from extinct languages and Finnic are found in the Komi data (Fig. 4). They probably comprise several instances of borrowing to Proto-Permic and later to Komi (Saarikivi, 2018: 342). The items are uncertain, and none could be assigned as clear. Mari has been in contact with Permic and later with Udmurt, (Bereczki, 1994: 12–13). Permic influence in Mari is represented by a single item (Figure S1).

3.5   *Comparison between Basic and Whole Vocabulary*

We compare basic vocabulary and whole vocabulary data for Estonian, Komi, Mari and Hungarian and examine how the sizes of loanword layers in basic vocabulary concur with the sizes of loanword layers in the whole vocabulary data (Figs. 5A and 5B). Common for all comparisons is that basic-vocabulary data contains a smaller proportion of borrowings than the whole vocabulary.

3.5.1   Estonian

The Estonian UraLex data is compared to the data in the Estonian etymological dictionary (= EES) as surveyed in Soosaar (2013) (Fig. 5A).[1] In basic vocabulary and EES, the Latvian, North Germanic, Slavic and Old Russian, Iranian and Indo-Iranian layers are the most congruent ones in size. All these layers contain low numbers of borrowings in general (Table S2), and they represent the smallest proportions of loanwords in both vocabulary inventories. The percentage of Baltic items in basic vocabulary as well as in EES is larger than the proportions of these smaller categories. The size of the Baltic layer follows our expectation relatively well and can be interpreted as congruent between basic vocabulary and the comparison data.

The biggest difference between Estonian basic vocabulary and EES lies in the sizes of the Low German, German and Swedish layers. Both occur in basic

---

1   The scale on the y-axes stands for the proportion a category represents in the respective Estonian and Komi dictionaries. The proportions are calculated using estimates for the loan-category size which are the averages from the range of possible and clear loanwords in a category. The scale on the x-axes represents the proportions for the Estonian and Komi UraLex data obtained in the same way. The position of the data points indicate the representation of the loan-category in basic vocabulary: e.g., in Estonian, the Low German category is under-represented as it has a high value on the y-axis but a low value on the x-axis. Therefore, the data point is located in the upper left corner. If the relative sizes of the loan-category are similar between UraLex and the dictionary data, the data point is located close to the diagonal line, e.g., Russian in Komi. The data points for overrepresented categories in basic vocabulary appear under the diagonal line on the right side of the figure following the x-axis: e.g., The Indo-Iranian and Iranian category in Komi. A comparison between the Swadesh lists and comparison data is presented in Figs. S4A, S5A.

vocabulary but are clearly underrepresented in UraLex. The Low German and German are the least congruent categories in this comparison and they represent the largest loanword layers in EES. UraLex contains no items from more recent Finnish and Russian loanword layers, which, however, are prominent in EES.

The sizes of the Indo-European and Germanic layers are less congruent between basic and whole vocabulary so that both layers are overrepresented in basic vocabulary. The Indo-European category in EES pools together items of the older Indo-European layers with wide and narrow distributions in Uralic. The number of identified loanwords from Proto-Indo-European and North-West Indo-European origin is smaller to begin with; the Indo-European category is therefore not very prominent in EES. In basic vocabulary, however, it is well-represented as the third largest layer. The Germanic layer is strong in both vocabularies but the proportions are less congruent in relation to the other layers. Germanic is the largest category in UraLex which makes it overrepresented, as this is not the case in EES.

3.5.2   Komi-Zyrian

There are thousands of Russian borrowings in Komi (Bartens, 2000: 22), and Russian is clearly the most prominent category in both KESK and UraLex; meaning it is the most congruent layer in the comparison. On the contrary, the Volga Bulgar layer is relatively small in both KESK and UraLex and its size is congruent between both basic and whole vocabulary. The putative Germanic borrowings in Komi are the result of sporadic or indirect contact. This negligible influence is reflected by the size of the layer.

The Finnic and Western Uralic category is prominent in KESK; however, KESK somewhat overestimates the numbers, and newer literature takes a more cautious stance, so the actual number of these items is likely smaller (Table S7). UraLex does feature the Finnic and Western Uralic category, but in the context of KESK data it is underrepresented. A major difference between KESK and UraLex is that KESK contains Ob-Ugric (Khanty and Mansi) and Samoyedic (Nenets) borrowings. There are no traces of these family-internal interactions in the Komi basic vocabulary. In addition, the minor Old Russian layer is not represented in UraLex.

As a result, the Komi basic vocabulary is biased towards the archaic Indo-European and layers from Indo-Iranian and Iranian as they are clearly overrepresented. The latter is the least congruent category between KESK and UraLex.

3.5.3 Meadow Mari

A survey (Saarinen, 2010) of the Mari dictionary (= TschWB) provides quantitative information for our comparison (Fig. 5B). The Chuvash and Volga Bulgarian influence is noticeable in the basic and whole vocabulary. This category is the largest in both TschWB and UraLex and clearly congruent in size. This result is to be expected, considering the long-standing and intensive contact situation between Mari and the Oghur Turkic languages. The minor Permic layer is congruent as well, representing only a small fraction of loanwords. The least congruent layer in the Mari comparison is Russian, which is the largest layer in TschWB but heavily underrepresented in UraLex. The Tatar layer is prominent in basic vocabulary but, in relation to the expected size of the layer, it is still underrepresented in UraLex.

The Indo-European and the Indo-Iranian and Iranian categories constitute under one percent of the items in TschWB, but the proportions are larger in basic vocabulary making these layers less congruent between the two vocabularies (Table S9). Both categories can be interpreted as overrepresented in basic vocabulary.

3.5.4 Hungarian

Comparing the Hungarian basic vocabulary to the Hungarian etymological dictionary (= EWUng) (Fig. 5B), there is almost no congruence between the sizes of the loanword categories. Only the heterogeneous "Other category" appears as congruent. While this composite category is not informative on its own, it provides a data point with congruence that the other categories can be related to.

The Slavic category is the least congruent between EWUng and basic vocabulary; it is large in the dictionary data, but heavily underrepresented in basic vocabulary. In EWUng, loanwords from Latin, German, Romani or Romance languages acquired in the Middle Ages or later overshadow the older strata. These categories are not featured in UraLex at all. Moreover, the Turkic category is the largest in basic vocabulary, but it is noticeably less prominent in EWUng. The dictionary survey (Keresztes, 1998) treats the archaic Indo-European, Indo-Iranian and Iranian layers as one category. They represent a very small fraction of under one percent of the words in EWUng, but they are the second largest category in basic vocabulary. Both the Turkic and Indo-European categories are thus heavily overrepresented in UraLex.
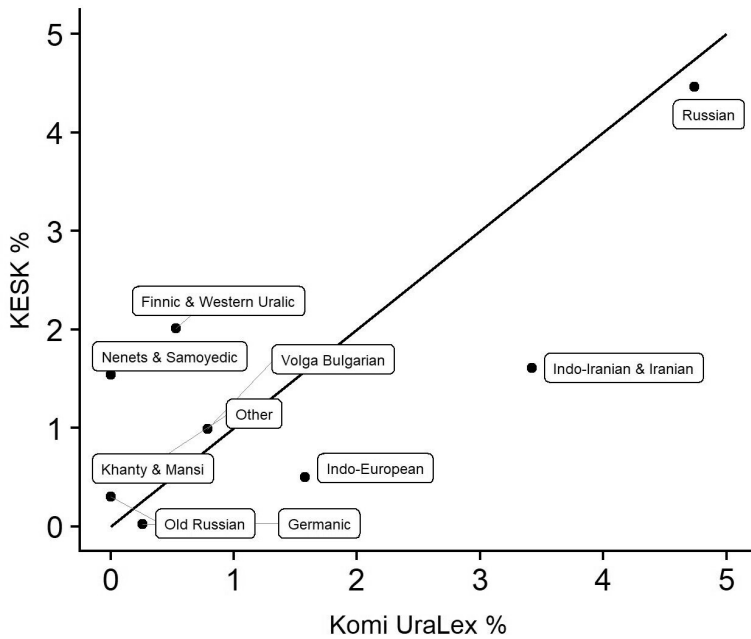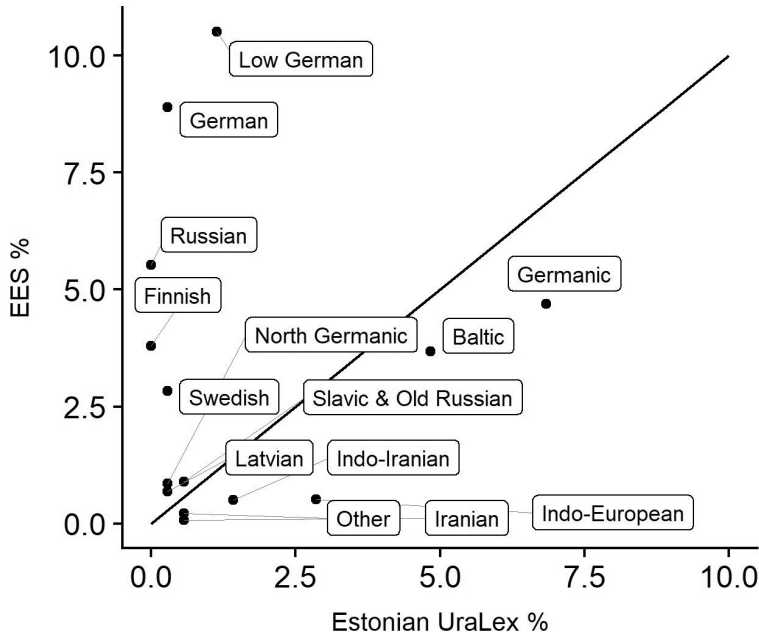
FIGURE 5A Comparison of prominent loanword categories in etymological dictionaries (EES and KESK) (y-axes) and UraLex basic vocabulary data (x-axes) for Estonian and Komi
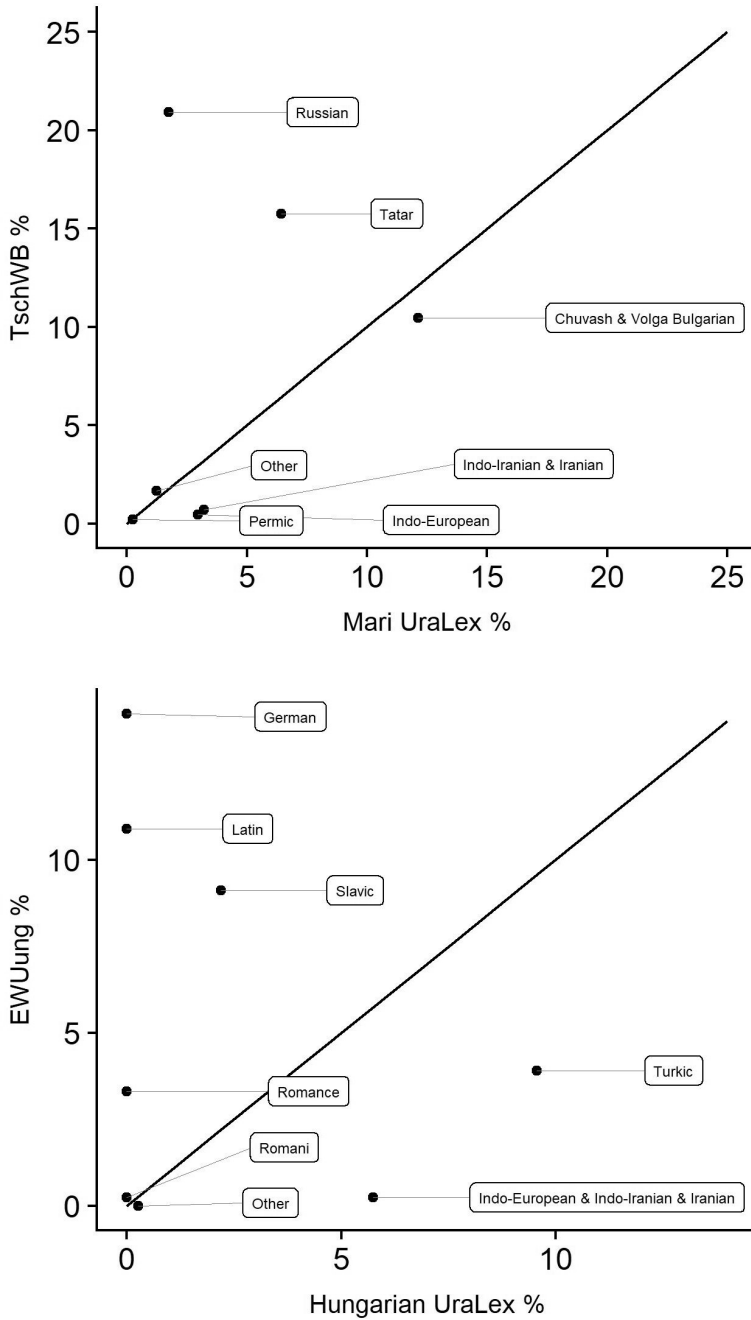
FIGURE 5B Comparison of prominent loanword categories in dictionaries (TschWB and EWUung) (y-axes) and UraLex basic vocabulary data (x-axes) for Mari and Hungarian

4   Discussion

In this paper we explored whether basic vocabulary can be used to establish a borrowing profile of a language, i.e., whether basic vocabulary captures the contact history of a language in a representative manner. We therefore examined the loanword layers in the basic vocabulary of six Uralic languages and compared them to loanword layers in the whole vocabulary. First, to assess the success of detecting the loan layers in basic vocabulary in general, we examined which loanword layers are present in the basic vocabularies. Second, to evaluate whether the basic vocabulary captures the relative sizes of the loan layers accurately, we compared the loanword layers in basic vocabulary with those in the languages as a whole, through quantitative comparison for four languages (Estonian, Komi, Mari and Hungarian) and qualitatively for two (North Saami and Erzya). We studied three possible scenarios of the relationship of these two data: A) loanword layers in basic vocabulary are representative of the whole language, i.e., we assume congruence in relative sizes of the loanword layers, B) loanword layers in the basic vocabulary underrepresent the layers in the whole vocabulary of a language, and C) loanword layers in the basic vocabulary overrepresent the loanword layers in the whole vocabulary.

4.1   *Detecting Borrowing Profiles*
We evaluate the success of establishing the borrowing profile of a language through basic vocabulary loans by considering two levels of detail in detection of contact influence: the subgroup-level of a donor language family the contact influence comes from (this section) and the level of individual donor languages the subgroup level consists of for higher resolution (Section 4.2).

When focusing on a macro-level categorization through subgroups of the main borrowing sources, it is noticeable that UraLex (i.e., the basic wordlists with synonymy allowed) shows contact influence from almost all expected branches of the Indo-European (Archaic Indo-European, Indo-Iranian, Balto-Slavic and Germanic), Turkic (Oghur and Common Turkic) and Uralic (Western Uralic but not Eastern Uralic) families. This means that on the roughest macro-scale level, basic vocabulary is highly successful in establishing a borrowing profile since the subgroup-level influence is clearly detected.

Archaic Indo-European as well as Indo-Iranian influence is retained in basic vocabulary across all focal languages. The presence of these categories reflect the oldest contacts between Uralic and Indo-European. The majority of borrowings from the Indo-Iranian branch in the basic vocabulary data are likely early borrowings into dialectal Proto-Uralic instead of late loanwords borrowed by independent language groups. Conversely, the Balto-Slavic, Germanic, Tur-

kic and Western Uralic traces accumulated through contact between later protolanguages and individual modern languages. The Turkic influence in the macro-profile mostly consists of Oghur Turkic layers.

The basic vocabularies of the focal languages also distinctly capture the borrowing profile of the specific Uralic subgroups these languages represent. The subgroup-specific loanword layers regularly appear in the focal languages, e.g., North Saami basic vocabulary features borrowings from North Germanic and Finnic sources which are also relevant for the whole Saamic subgroup.

4.2 *Detecting Individual Loanword Layers in Basic Vocabulary*

In general, basic vocabulary captures the borrowing profiles of the six focal languages relatively well because most known loanword categories are present in UraLex. However, there is obvious variation in the success of detecting individual loanword layers on the language-specific level. While influence from a certain donor branch is clearly present on a macro-scale, a micro-scale examination reveals that in some cases the influence only comprises one particular loanword layer, while others are not present in the basic vocabulary. The variation is commonly caused by the absence of loanword layers from more recent donor languages.

Basically all expected categories were represented in the basic vocabularies of North Saami, Erzya and Mari. However, in Mari the single Permic category does not allow us to differentiate between earlier (Proto-Permic) or later (Udmurt) origin. The number of potential Udmurt borrowings is very small to begin with (Saarinen, 2010: 337) and is not detectable in basic vocabulary.

Estonian basic vocabulary was missing two large categories: the influence from Slavic is weakly represented because it only comprises single Old Russian items while modern Russian is not present at all. The Estonian basic vocabulary also lacks Finnish loanwords meaning there is no visible Uralic family-internal influence. Estonian is a majority language acting donor language for smaller Finnic languages (see Björklöf, 2019), which likely explains the lack of family-internal borrowings in general.

In the Komi basic vocabulary, modern Russian is responsible for all Slavic influence; the Old Russian layer is missing. A more clear-cut deficiency in the Komi loan profile of basic vocabulary is its failure to detect any Eastern Uralic presence in the form of Samoyedic and Ob-Ugric layers. Komi has thus three undetected loanword layers. Basic vocabulary still captures the Komi borrowing profile successfully as most loanword sources still appear in UraLex.

The profiling of Hungarian was least successful of the focal languages because four expected categories are absent from the Hungarian basic vocabulary. These are the German, Latin, Romance and Romani layers. Finally, the

most recent international influence, e.g., from English, has not affected the basic vocabulary of any of the languages.

4.3   *Borrowing Percentages in the Basic Vocabulary*

The proportion of borrowings varies between the focal languages. 33% of the North Saami basic vocabulary is borrowed, whereas only 16% is in Komi. Multiple reasons are likely to be behind this variation. Until recently Uralic etymological studies tended to focus more on western Uralic languages, leading to larger numbers of loanword proposals, and reducing the number of words with unknown origin.

An uneven research history is also echoed by the certainty estimates. Existing evaluative literature allows assigning certainty estimates with more confidence, but sometimes a clear conclusion concerning the donor language cannot be reached, as is e.g., the case with the multi-layered Indo-European influence. In the future, more evaluative studies are needed to resolve the less clear etymologies, and etymological study of the eastern Uralic languages still has some catching up to do in relation to western Uralic, and need to be contextualized with regard to other (non-Uralic) Siberian languages.

The borrowing percentages in the Swadesh-100 lists reveal that borrowing has affected the studied languages comparably to the full lists: Estonian (32%) and North Saami (27%) have the highest borrowing percentages; Hungarian (13%) and Komi (15%) have the lowest borrowing percentages. As the Swadesh-100 list only has a limited number of meanings, the actual numbers of loanwords are low (see Table S1).

4.4   *Loanword Layer Sizes in Basic and Whole Vocabulary*

For the four Uralic languages Estonian, Hungarian, Komi and Mari with etymological dictionaries as useful comparative data, we examined the congruence of the relative sizes of loanword layers in the basic vocabulary and whole vocabulary. In total there were 52 language pairs to compare, e.g., Indo-Iranian in i) the basic vocabulary and ii) in the whole vocabulary, so that "pair" refers to the name of a language (or languages) as used in both the basic and the whole vocabulary. 36 of these pairs are found in the 4 languages with etymological dictionaries, 17 in the two, Erzya and North Saami, that do not. For these two we could only make a qualitative comparison of the loanword layers in the basic and whole vocabulary. We do not include the heterogeneous Other category in the pairs.

The accuracy of the borrowing profiles of our six languages varies, because not all loanword layers in basic vocabulary correctly reflect the relative sizes of the layers approximated for the whole vocabulary. Basic vocabulary tends

to overrepresent the oldest loanword layers, while influence from more recent languages tends to be lacking.

4.5 *Congruence between Basic and Whole Vocabulary*

Of the above-mentioned 52 pairs, 18 showed congruence, i.e., a clear correlation in size, between the basic vocabulary loanword layer and dictionary data. Language-wise, Estonian has six (out of 14) congruent layers, Komi has three matching pairs (out of nine) and Mari has two congruent layers (out of six). Through our qualitative comparison, we found three congruent pairs (out of eight) for North Saami and four (out of nine) instances of congruence for Erzya. For Hungarian, the sizes of loanword layers in the dictionary data and UraLex deviate from each other.

In total, six pairs represent loanword layers with very high numbers of words in both basic and whole vocabulary: the quantitative comparisons show three large and congruent layers. These are Baltic in Estonian, Russian in Komi, and Chuvash and Volga Bulgar in Mari. Comparing the general estimates of loanword-layer sizes to those in the North Saami and Erzya basic vocabulary, we identified three layers that have the highest numbers of identified loanwords in general. The same layers are also the largest ones in basic vocabulary and therefore are congruent. The matching pairs are the Finnic and North Germanic layers in North Saami and the Russian layer in Erzya.

Five of the six pairs involving large loanword layers have largely been acquired in historical times. These are the North Germanic and Finnic categories in North Saami, Russian in Erzya and Komi, as well as most of the Chuvash and Volga Bulgar category for Mari. Baltic is the large, old and congruent category in Estonian.

Twelve pairs represent only small proportions of loanwords in the whole and basic vocabulary. Thus the small expected sizes are correctly reflected by the borrowing profile. In the quantitative comparisons there are eight such pairs: the Latvian, North Germanic, Slavic and Old Russian, Iranian and Indo-Iranian pairs are small and congruent for Estonian. Similarly, for Komi, the Volga Bulgar and Germanic categories concur in smallness. The small size of the Permic layer in Mari is also evident in UraLex. We qualitatively identified three matching pairs with small numbers of loanwords: Balto-Slavic in North Saami and the sporadic Germanic layer as well as the Volga Bulgar and uncategorized Turkic borrowings in Erzya. The small and congruent pairs represent both older and newer layers without a clear-cut pattern.

4.6　　*Underrepresented Loanword Layers in Basic Vocabulary*

While the sizes of the 18 pairs discussed above can be considered as congruent between basic and whole vocabulary, it is clear that basic vocabulary does not reflect the relative sizes of all loanword categories in a language in general with perfect accuracy. All languages have loanword layers that are underrepresented in their basic vocabulary. Twelve loanword layers out of 52 were underrepresented in the basic vocabularies. The underrepresented categories tend to be the least congruent ones in size in the quantitative comparison, e.g., in Estonian Low German is the largest loanword layer in the whole vocabulary, but weakly represented in the basic vocabulary.

The Estonian basic vocabulary underrepresents three layers out of 13; the Low German, German and Swedish layers. For Komi, there is one underrepresented category, namely Finnic and Western Uralic (one out of nine pairs). This category is prominent in the etymological dictionary of Komi published in 1970, but current etymological research generally finds less evidence for Finnic and Western Uralic influence in Komi (e.g., Saarikivi, 2018). In Mari, Russian and Tatar are underrepresented layers, and in Hungarian Slavic is underrepresented. Moreover, we found in the qualitative comparisons of North Saami and Erzya 5 cases out of 17 where loanword layers in the basic vocabulary were underrepresented.

Germanic and Baltic are underrepresented in North Saami basic vocabulary. The representation of these layers is complicated because of the close genealogical and areal relationship between Finnic and Saamic. Saamic shares dozens of Proto-Germanic loanwords with Finnic; these words were previously thought to have been mediated via Finnic, but it has been shown that Saamic also has separate Proto-Germanic borrowings (Aikio, 2006: 10). The large number of North Germanic borrowings and the influx of loanwords from modern languages in Saamic indicate increasingly intensive contacts (Aikio, 2006: 13). In general, there are only a few Baltic items in Saamic not shared with Finnic; it is unclear whether they were acquired through direct contact (Aikio, 2012: 75).

In North Saami basic vocabulary there is only one attested borrowing from substrate languages, which are unattested languages spoken in the Saami language area at the time of arrival of proto-Saamic. The research on substrate borrowings in the Saamic languages is still ongoing and the number of these items can potentially be much higher (Aikio, 2012: 85). Therefore, the single appearance of such borrowings in basic vocabulary can be interpreted as underrepresentation.

Out of the twelve underrepresented loanword layers, seven represent newer layers acquired in historical times: Low German, German and Swedish in Esto-

nian, Russian in Mari, Tatar in Mari and Erzya, and Slavic in Hungarian. In addition, two pairs involving the composite category of Finnic and Western Uralic present in both Komi and Mari include some newer loanwords. The underrepresented prehistorical layers in North Saami, Germanic, Baltic and substrate words, have uncertainty as described above.

4.7 *Overrepresented Loanword Layers in Basic Vocabulary*

All six languages share a number of words borrowed from archaic Indo-European and Indo-Iranian. All except Hungarian also have a common distributionally limited and archaic North-West Indo-European layer. It can be quantitatively shown that these three layers are overrepresented in the basic vocabulary of Mari, Komi and Hungarian; for Estonian this only applies to archaic Indo-European.

Out of the languages with quantitative representation (Estonian, Mari, Komi, Hungarian) the overrepresentation of these old loan layers is strongest in Hungarian. However, this is an artefact of lumping, because the dictionary survey forces all instances of Indo-European and Indo-Iranian influence to be counted as a single category. For Komi the Indo-Iranian and Iranian category is not only overrepresented but also the least congruent category in the comparison, i.e., matches least in size between basic and whole vocabulary. This composite category includes Iranian loanwords borrowed separately into Proto-Permic; these are strongly represented in the basic vocabulary.

North Saami is claimed to be extensively influenced by archaic peripheral Indo-European varieties (Sammallahti, 2001: 413). Out of all proposed archaic Indo-European and North-West Indo-European borrowings in Saami, 33 % are present in North Saami basic vocabulary, showing clear overrepresentation of the archaic Indo-European influence in North Saami basic vocabulary. It had earlier been proposed that Finnic and Saamic acquired Indo-Iranian and Iranian loanwords separately after Proto-Uralic disintegrated (Koivulehto, 1999: 232; Koivulehto, 2001: 240–245). However, in the case of Saamic, the number of Indo-Iranian borrowings is small, and direct contacts between Indo-Iranian and Saamic are improbable (Holopainen, 2018: 170). The layers present in the Saami basic vocabulary were probably already acquired earlier at an ancestral western Uralic stage.

Low German, German and Swedish layers occur in Estonian in both its basic vocabulary and in UraLex (where they are underrepresented). However, Proto-Germanic, the oldest Germanic layer, which is slightly younger than North-West Indo-European, is underrepresented in Estonian.

The lexical influence of Turkic on Hungarian has been relatively strong, though the (newer) Common Turkic layers are small. There are many borrow-

ings from (older) Oghur Turkic in the full Hungarian UraLex data where these form the largest proportion of clear items in the basic vocabulary in which this layer appears to be overrepresented.

The Erzya basic vocabulary has a distinct Baltic layer. Of all the proposed Baltic borrowings for Erzya, 28% are found in its basic vocabulary, which we interpret as overrepresentation. It has been discussed whether these Baltic loanwords were borrowed into a common proto-language stage of Finnic and Mordvinic (the proto-language from which Erzya evolved) or independently (Grünthal, 2012: 311). The latter is considered to be more likely because of phonetic irregularities. The number of Baltic loanwords is significantly smaller in Mordvinic than in Finnic, and borrowings shared with Finnic can also be explained as parallel borrowings (Grünthal, 2012: 310).

4.8   *Summary of Results*

Our general results for estimating the contact history of a language through borrowing profiles in basic vocabulary are summarized below and the results for our test scenarios are presented in Fig. 6.

We examined the relative sizes of the loanword layers in the basic vocabulary and the whole vocabulary as 52 language pairs (i.e., name of a language (or languages) in both the basic and the whole vocabulary, see section 2.4. for labels). This was done quantitatively for Estonian, Komi, Mari and Hungarian, and qualitatively for North Saami and Erzya.

– In total, 18 out of 52 pairs are congruent, i.e., the basic vocabulary accurately reflects the size of the loanword layer in question. Six of the congruent pairs represent layers with high numbers of loanwords and twelve pairs correspond to small layers congruent in size. The large and congruent layers tend to be younger in age, whereas the age of the small and congruent layers varies.

– Twelve pairs out of 52 show underrepresentation in basic vocabulary, and in total there are nine cases of missing layers. The majority of the underrepresented and missing loanword layers are younger.

– Thirteen out of 52 pairs compared show overrepresentation in basic vocabulary. Ten of these involve the oldest Indo-European, Indo-Iranian and Iranian contacts. In addition, Estonian overrepresents the Germanic layer and Erzya the Baltic layer; these were acquired in prehistoric times. The Turkic category in Hungarian contains mostly older borrowings.

– On a macro-scale, the borrowing profiles identified from basic vocabulary for all six languages, detect influence from virtually all (except eastern Uralic) contact languages or language groups. The basic vocabularies of North Saami, Erzya, and Mari reveal the most complete borrowing profile,
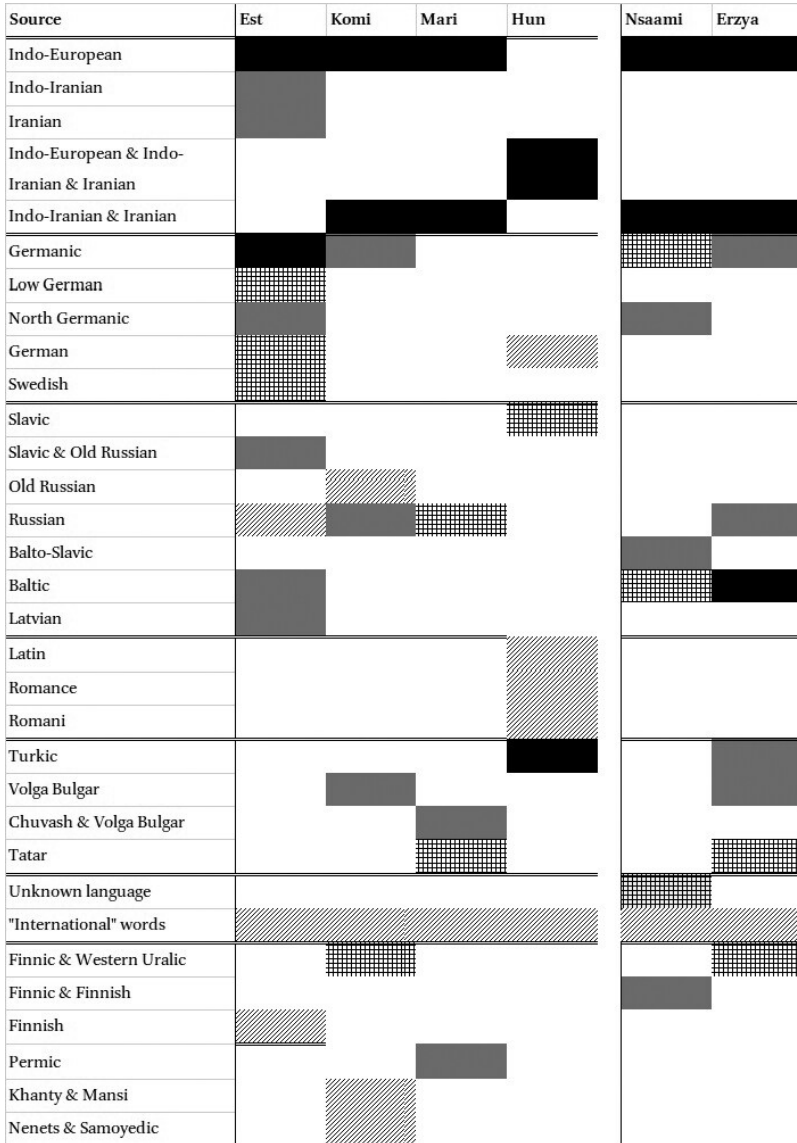
Source	Est	Komi	Mari	Hun	Nsaami	Erzya
Indo-European						
Indo-Iranian						
Iranian						
Indo-European & Indo-Iranian & Iranian						
Indo-Iranian & Iranian						
Germanic						
Low German						
North Germanic						
German						
Swedish						
Slavic						
Slavic & Old Russian						
Old Russian						
Russian						
Balto-Slavic						
Baltic						
Latvian						
Latin						
Romance						
Romani						
Turkic						
Volga Bulgar						
Chuvash & Volga Bulgar						
Tatar						
Unknown language						
"International" words						
Finnic & Western Uralic						
Finnic & Finnish						
Finnish						
Permic						
Khanty & Mansi						
Nenets & Samoyedic						

FIGURE 6   Summary of the representations of the loanword categories in basic
vocabulary
Note: Estonian, Komi, Mari and Hungarian use the results from the qual-
itative comparisons; North Saami and Erzya are presented separately
because their information is based on a qualitative assessment. Gray
colour in a cell indicates congruence in size, raster underrepresenta-
tion and black overrepresentation and diagonal lines denotes missing
categories. Non-relevant category names are blocked out with white. A
summary with the basic vocabulary sublists is presented in Fig. S3.

whereas Estonian and Komi have a few clearly discernible loanword layers which are missing from the basic vocabulary. The Hungarian borrowing profile is the most deficient.

– Although the borrowing profile of Estonian lacks a number of loanword layers, it is the most congruent because it has the highest number (6 out of its 13 loanword layers; the category labelled as Other omitted) of congruent layers. The borrowing profile of Hungarian is the least congruent because it has the most missing layers (4 out of 7 (Other category omitted) and the layers found in basic vocabulary do not match the whole vocabulary layers (overrepresented 2/7 or underrepresented 1/7).

4.9   *Summary of Results for the Swadesh-100 Lists*

We provide a brief summary of the borrowing profile mediated by the Swadesh-100 lists, a smaller selection of data, in order to increase comparability with other languages and language families (see Section S1 for a full analysis). Unsurprisingly, the complete UraLex 2.0 list captures a more complete and accurate borrowing profile a language in comparison to the short Swadesh-100 lists. This is supported by the higher numbers of loanword layers present in the complete data overall (43 in the complete list vs. 35 in Swadesh-100 profiles of the total 52 pairs of languages compared) and a higher number of categories which are accurately represented, i.e., congruent in size between basic and whole vocabulary (18 vs. 7 of 52 pairs). Twelve vs. 14 of the 52 pairs show underrepresentation and 14 vs. 13 overrepresentation.

The borrowing profiles captured by the Swadesh-100 lists have two general attributes setting them apart from the more accurate profiles. First, a common pattern for all studied languages is that especially the oldest loanword layers involving most commonly the archaic Indo-European, Indo-Iranian and Iranian donor languages tend to be even more overrepresented in Swadesh-100 than in the complete lists. Second, most loanword layers which are the result of somewhat less significant contact or have smaller numbers of loanwords in general are only weakly captured by the Swadesh-100 lists (congruent or overrepresented in the complete UraLex 2.0 list, underrepresented in Swadesh-100: 3 cases; congruent or underrepresented in the full list, missing from Swadesh-100: 7 cases).

It is evident that the Swadesh-100 list, on account of its brevity, can only capture a limited perspective of a borrowing profile. The shorter list strongly underrepresents or completely filters out evidence of less intense or limited linguistic contact, but it can still provide a snapshot of the oldest and the most influential contacts of a language. With these limitations in mind, the highly accessible Swadesh-100 list still provides a valuable opportunity for cross-linguistic com-

parison of borrowing profiles, especially when the research focus is on older linguistic contacts.

4.10 *Sociolinguistic Factors and Language Contact*

As mentioned in the introduction, the basic vocabulary of a language is usually compiled based on the criteria of frequency, semantic neutrality and stability, implying that it does not contain loanwords. However, as e.g., Thomason and Kaufman (1988: 74) point out, borrowing into basic vocabulary can commonly occur in more intense language contact. The borrowing profiles of our focal languages are therefore likely facilitated by more extensive bilingualism and intense cultural pressures rather than just casual contacts. The sizes of the loanword layers are further affected by other factors governing the intensity of contacts: time and length of contact, geographical closeness and majority/minority dynamics between speaker groups (Thomason and Kaufman, 1988: 66). For loanword layers representing more recent contacts, the exact circumstances are well documented, but for the older layers heavily featuring in basic vocabulary these remain unknown.

4.11 *Explaining the Patterns of Older Borrowings in Basic Vocabulary*

Basic vocabulary is powerful at retaining the most archaic loanword layers: our results reveal that they are mostly overrepresented, and in some cases congruent in size when compared to whole vocabulary. A challenge for explaining this representation is that the mechanisms behind contact-induced changes in (pre-)historical linguistic contacts cannot always be ascertained with confidence. In many historical cases, the details of the sociolinguistic settings (e.g., attitudes towards languages) for contact-induced change are unknown. However, new studies discuss how the presence of other languages in philological material can be interpreted from the perspective of modern contact linguistics (Johanson, 2013; Andrason and Vita, 2016). Each historical contact situation has to be assessed individually, and the text attestations likely reflect the specific sociolinguistic situation of the writer and literary traditions. Nonetheless, these studies indicate that also in the past lexical and structural contact influence was acquired during long-standing contacts between geographically neighboring languages, bilingualism and status of the languages being the most important factors (Johanson, 2013: 342; Andrason and Vita, 2016: 329). These are the same factors shaping contact between languages today. The observed mechanisms of contact-induced change in synchronic research can offer a framework for studying prehistoric contact settings where there are no texts acting as a window to the past.

It seems likely that the older loanword layers present in the basic vocabulary of our languages are also the outcomes of intense contact among bilingual

communities, involving different social statuses of the languages. The Proto-Indo-European and North-West Indo-European loanwords are overrepresented in basic vocabulary, but due to the age and heterogeneous nature of the layers not much can be said about the sociolinguistic conditions during these contacts. Doubtlessly their semantic neutrality and frequency of use have helped to preserve these older loanwords despite multiple changes in the sociolinguistic situations of the focal languages. Despite the strong influence younger contact languages such as Russian have had on the focal languages, this has generally not yet led to the replacement of the oldest borrowings in basic vocabulary.

It is generally assumed that the Indo-Iranian and later Iranian borrowings reflect a situation of cultural exchange where Indo-Iranian speakers were of a higher status and provided new expertise in the area of animal domestication (Korenchy, 1988: 679; Holopainen, 2019: 345). In addition, the borrowing of e.g., kinship terms point to more intimate contacts. An underresearched area is possible Uralic influence on Indo-Iranian which would offer valuable information on the nature of these contacts (Holopainen, 2019: 346).

For some older instances it is uncontroversial to postulate an intimate and long-standing nature for these contacts. The intense contacts between Finnic and Germanic as well as between Finnic and Baltic are evinced by hundreds of loanwords which can be found in high numbers in nearly all semantic domains (Table S2). In addition, some Finnic loanwords have also been detected in Germanic, indicating significant mutual relationships (Hofstra, 1997: 132). The development of the Volga Bulgar state around 500 AD is connected to the dominance of an Oghur Turkic-speaking elite in the Volga area (Agyagási, 2012: 21). Oghur Turkic layers in the basic vocabularies of Erzya, Mari and Komi further confirm that the Turkic presence led to strong cultural pressure on the Uralic-speaking peoples. This includes Hungarian, spoken in the east before the conquest of the Carpathian Basin (Róna-Tas, 1988: 752).

4.12   *Explaining the Patterns of More Recent Borrowings in Basic Vocabulary*

The representation of younger loanword layers in the borrowing profiles of our six languages shows three general tendencies: 1) The layers are either congruent in size between whole and basic vocabulary, 2) they are underrepresented in the basic vocabulary, or 3) the layers are missing from the basic vocabulary altogether. Well-documented sociolinguistic conditions and changes in them explain these tendencies.

The large congruent loanword categories mostly occur in those languages that are currently spoken as minority languages. The increasing pressure from dominant languages is a major driver of change in the domains of language

use, leading to major shifts in the vocabulary, and in turn to large new loanword layers even in basic vocabulary. This is the case with North Saami, Erzya and Komi, all endangered minority languages strongly influenced by the dominant majority languages. Mari is a minority language as well, but its close and longstanding contact with Oghur Turkic, mainly Chuvash, has led to a different result: it seems that the contacts with Oghur Turkic, starting already in Volga Bulgar times and still ongoing at the periphery of the Mari-speaking area, have slowed the adoption of Russian loanwords into the basic vocabulary of Mari, causing their underrepresentation.

The North Saami borrowing profile is characterized by North Germanic, Modern Scandinavian and family-internal influence from Finnic; these loanword layers are the largest in whole and basic vocabulary. Since the 19th century assimilation and intensive contacts with modern majority languages such as Norwegian, Swedish and Finnish has greatly increased (Saarikivi, 2011: 81). Together with demographic changes, changes in state borders, and harsh assimilation policies have changed and split the traditional North Saami speech communities and reduced the number of speakers (Aikio et al., 2015: 244; Marjomaa, 2014: 57–58). Virtually all North Saami speakers are bilingual in Saami and a state language (Norwegian, Swedish or Finnish) (Aikio et al., 2015: 244). It must be noted that the UraLex data is an approximate representation of North Saami and a dialect dataset is needed in order to discern micro-scale patterns.

Erzya Mordvin and Komi-Zyrian are both spoken as minority languages in Russia and are heavily influenced by Russian, as shown by the large and congruent representation in the basic vocabulary. The speakers of Erzya and Komi are probably all fluent bilinguals (Riese, 1998: 252; Janurik, 2017: 27). The speakers of Erzya and Komi have been affected by the fragmentation of the traditional speaker areas, urbanization, as well as by official policies reducing the areas of language use (Mosin, 2002: 161, 2002; Riese, 1998: 252; Rueter, 2013: 5–6). Duration of contact might have affected the volume of Russian borrowings in the basic vocabulary. Contacts between Mordvin and Russian started a few centuries earlier than between Russian and Mari or Russian and Komi (Décsy, 1988: 632). This could contribute to the strong Russian presence in Erzya; the position of Erzya is also weaker in comparison to Mari and Komi as the majority of Erzya speakers live outside the Republic of Mordovia (Bartens, 1999: 10).

Considering these general developments, the strong Finnic, North Germanic and Russian presence in the basic vocabulary of the focal languages is likely caused by the rapid change in the sociolinguistic situation, leading to a strong command of two linguistic repertoires, i.e., total bilingualism.

In addition to relatively sudden and recent changes in power dynamics between speakers leading to change in domains of language use, another con-

tributing aspect of the congruent representation of new loanword layers can be that these loanwords are more confidently detectable because of their transparency. It is also possible that some older borrowings have been replaced by these newer loanwords, obscuring evidence of older contacts. Replacement might similarly play a role in the representation of the small and congruent layers since their emergence in basic vocabulary hints at more influential contacts, even though the number of detected borrowings in the whole language is smaller. Alternatively, a low number of loanwords could also point to a shorter duration of contact, without the concurrent development of extensive bilingualism.

The underrepresented or missing loanword strata in UraLex contain important semantic fields of so-called cultural vocabulary such as words related to agriculture (see listing of semantic fields in the Supplementary material). Since the goal of basic vocabulary lists is to filter out cultural meanings, loanwords referring to change in societal order and certain lifestyles are not visible in basic vocabulary material. While the exact number of borrowings in these semantic fields is not clear, it is likely that the underrepresented or missing loanword layers are the result of the most common type of contact-induced change in vocabulary, namely the acquisition of cultural vocabulary, and that borrowing has mostly affected only a certain part of the vocabulary (Thomason and Kaufman, 1988: 77; Matras, 2009: 156).

The missing or underrepresented loanword categories in basic vocabulary tend to be newer and acquired during the independent developments of the focal languages:

In Estonian, Low German influence started in the 13th century; its status as the language commonly used in administrative contexts lasted until the late 16th century, though it was probably still spoken as late as the early 19th century (Soosaar, 2013: 288). High German influence started around the 16th century and was strong up until the 20th century (Soosaar, 2013: 290). The Russian influence in Estonian is concentrated in cultural vocabulary and is considered less intense (Blokland, 2009: 393). The Finnish layer is also restricted to cultural domains.

Regarding the missing layers in Komi, contacts with Ob-Ugric and Samoyedic languages are relatively late; loanwords from these languages are mostly found in northern Komi varieties, referring mostly to reindeer herding and life in the tundra (Rédei, 1963; 1964). For Ob-Ugric, Komi is also a strong donor language, indicating that the Komi imposed cultural dominance over Ob-Ugric speakers (Rédei, 1963; 1964; 1970).

For Hungarian, the layers missing from the basic vocabulary contain large numbers of words covering various semantic fields belonging to different cul-

tural domains and practices, e.g., societal order, education and trade (see Gerstner, 2006: 310–318). The underrepresented borrowings of Slavic origin in Hungarian are similar: the multi-layered Slavic influence is strong in the cultural vocabulary but weak in the basic vocabulary (Table S10), even though Hungarian is estimated to have around 2000 Slavic borrowings (Décsy, 1988: 621). The loanword strata present in the Hungarian UraLex list mostly involve older borrowings acquired before the conquest of the Carpathian Basin in the 9th century.

The basic vocabulary of the focal languages (Erzya, Mari and Hungarian) which have acquired large numbers of loanwords from Turkic show stronger influence from the Oghur Turkic branch, whereas the Common Turkic layers tend to be underrepresented. The (Common Turkic) Tatar words in Erzya and Mari have especially affected several important cultural domains such as agriculture and housekeeping (Tables S5, S9). In Hungarian, the Common Turkic layers are small and also limited to cultural vocabulary (Table S10). In Mari, the effect of Tatar influence, starting in the 13th century (Isanbaev, 1989: 28), varies considerably dialectally, an indication of a more regional and later relationship in contrast to the all-encompassing contact with Oghur Turkic (Hesselbäck, 2005: 167).

5   Conclusion

The historical development of languages results in a mesh of inherited, contact-induced and borrowed material. In this paper we have demonstrated that it is possible to extract a borrowing profile representing the contact history of a language from basic vocabulary data. The borrowing profiles provide a coarser, zoomed-out overview of contact history, but if necessary a more detailed picture can usually be obtained by using the existing information about the contact situations, supplemented with other, non-basic semantic domains, and taking into account the wider context of historical events.

The historical development of languages results in a mesh of inherited, contact-induced and borrowed material. In this paper we have demonstrated that it is possible to extract a borrowing profile representing the contact history of a language from basic vocabulary data. The borrowing profiles provide a coarser, zoomed-out overview of contact history, but if necessary a more detailed picture can usually be obtained by using the existing information about the contact situations, supplemented with other, non-basic semantic domains, and taking into account the wider context of historical events. In all, studies on linguistic contacts with vocabulary can be considered a start-

ing point, and, in order to gain a fuller picture on contact history, more studies are needed on borrowing in the domains of morphology and syntax (Norvik et al., 2022).

New systematic research in Uralic historical lexicology would further increase the quantity and quality of loanword material available for establishing the borrowing profiles; this material is currently biased towards western Uralic branches, especially Finnic and Saamic. However, this research underlines the need for more input on studies identifying the direct loanwords in the Saami basic vocabulary from those mediated via Finnic. The earliest archaic Indo-European borrowings also need further evaluation. Moreover, the effect of semantic change and the role of descriptive word formation especially in basic vocabulary is not yet well understood.

The context of borrowing could shed light on changes in prehistoric lifestyle. For example, some of the earliest loanwords have a specialized meaning in Uralic but a broad meaning in Indo-European. Currently, besides loanwords and hypotheses about possible language shifts, the effects of language contact are an understudied area in Uralic historical linguistics (Laakso, 2014: 2). From this would follow a new linguistic research prospect (see Section S2 for other suggestions): if we can establish an equally strong base unit for language shift (as loanwords are for borrowing profiles), we could capture a shift profile of a language, which in turn could give insights into the changing sociolinguistic environment.

The idea of language being a proxy for a speaker population has already been used in interdisciplinary research in attempts to shed light on the human past. Still, the evidence of contacts visible in linguistic material seems to be an underused resource in inter- and multidisciplinary work aiming towards a holistic understanding of human history. Previous studies mostly aim at finding the origins of linguistic groups by linking reconstructed languages and cultures together to a specific space and time (e.g., Frog and Saarikivi, 2015). There are many detailed syntheses on the origins of language families, Indo-European being a well-discussed example (Renfrew, 1987; Mallory, 1989; Anthony, 2007). The fit of language spread to archaeological dispersal models has been discussed for e.g., the Austronesian language family (Jordan and Gray, 2000; Green, 1999) and for Bantu (Holden, 2002). Often data availability shifts the attention to the micro-level, which is the case with correlation studies conducted using e.g., toponymic material (Saarikivi and Lavento, 2012: 82–83). It is less common that the perspective of contact between language families is emphasized (e.g., Carpelan and Parpola, 2001; Parpola, 2012). In addition to archaeology, other fields such as genetics provide new types of data which have been contextualized with linguistic information (e.g., Pakendorf, 2014; Tambets

et al., 2018). Furthermore, the multidisciplinary field of archaeogenetics has developed due to the availability of methods to analyze ancient DNA. This collaborative line of research has yielded valuable studies on the characteristics of past human populations (e.g., Haak et al., 2015; Peltola et al., 2023; Översti et al., 2019).

In the future, contact information could strongly contribute to tracking e.g., migration routes and locations for prehistoric encounters if corresponding "contact profiles" aligning with the borrowing profiles could be established in other fields. Correlating different results from different fields is by no means straightforward or always uncontroversial, but refining the material and increasing data accessibility for such analyses is a step towards a more complete interdisciplinary synthesis.

Author Contributions

MdH and OV planned the paper. MdH, OV and MD contributed to the technical description in Section 2 and practical execution of the data visualization for the quantitative comparisons of large word stocks. MdH compiled, updated and systematized the loanword data. All authors participated in the interpretation of the results. All authors took part in writing the introduction and conclusion and in commenting on all sections of the manuscript.

Acknowledgments

References

Agyagási, Klára. 2012. Language Contact in the Volga-Kama Area. *Studia Uralo-Altaica* 49: 21–37. Szeged: University of Szeged, Department of Altaic Studies and Department of Finno-Ugrian Philology.

Aikio, Ante. 2006. On Germanic-Saami Contacts and Saami Prehistory. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 91: 9–55. Helsinki: Suomalais-Ugrilainen Seura.

Aikio, Ante. 2004. An Essay on Saami Ethnolinguistic Prehistory. In Irma Hyvärinen, Petri Kallio, and Jarmo Korhonen (eds.), *Etymologie, Entlehnungen Und Entwicklungen: Festschrift Für Jorma Koivulehto Zum 70. Geburtstag*. Mémoires de La Société Néophilologique de Helsinki 6: 5–34. Helsinki: Modern Language Society.

Aikio, Ante. 2007. Etymological Nativization of Loanwords: A Case Study of Saami and Finnish. In I. Toivonen and D. Nelson *Saami Linguistics*, 17–52. Amsterdam and Philadelphia: John Benjamins.

Aikio, Ante. 2012. An Essay on Saami Ethnolinguistic Prehistory. In Riho Grünthal and Petri Kallio (eds.), *A Linguistic Map of Prehistoric Northern Europe*. Suomalais-ugrilaisen seuran toimituksia 266: 63–117. Helsinki: Suomalais-Ugrilainen Seura.

Aikio, Ante, Laura Arola, and Niina Kunnas. 2015. Variation on North Saami. In Dick Smakman and Patrick Heinrich (eds.), *Globalising Sociolinguistics: Challenging and Expanding Theory*, 5254. London: Routledge.

Alpher, Barry and David Nash. 1999. Lexical replacement and cognate equilibrium in Australia. *Australian Journal of Linguistics* 19:1, 5–56

Andrason, Alexander, and Juan-Pablo Vita. 2016. Contact Languages of the Ancient Near East – Three More Case Studies (Ugaritic-Hurrian, Hurro-Akkadian and Canaano-Akkadian). *Journal of Language Contact* 9 (2): 293–334. DOI: https://doi.org/10.1163/19552629-00902004.

Anthony, David. W. 2007. *The Horse, the Wheel and Language. How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World*. Princeton: Princeton University Press.

Bartens, Raija. 1999. *Mordvalaiskielten rakenne ja kehitys*. Suomalais-ugrilaisen seuran toimituksia 232. Helsinki: Suomalais-Ugrilainen Seura.

Bartens, Raija. 2000. *Permiläisten kielten rakenne ja kehitys*. Suomalais-ugrilaisen seuran toimituksia 238. Helsinki: Suomalais-Ugrilainen Seura.

Bereczki, Gábor. 1994. *Grundzüge Der Tscheremissischen Sprachgeschichte: 1*. Studia Uralo-Altaica 35. Szeged: University of Szeged, Department of Altaic Studies and Department of Finno-Ugrian Philology.

Benzing, Johannes and Karl Heinz Menges. 1959. Classification of the Turkic Languages. In Deny, Jean et al. (eds.), *Philogiae Turcicae Fundamenta. Band 1*, 1–10. Wiesbaden: Steiner.

Björklöf, Sofia. 2019. Mutual contacts and lexical relations among the Finnic varieties

of western Ingria and northeastern Estonia. In Sofia Björklöf and Santra Jantunen (eds.), *Multilingual Finnic: Language contact and change*. Uralica Helsingiensia: 14. 89–153. Helsinki: Suomalais-Ugrilainen Seura. DOI: https://doi.org/10.33341/uh .85034

Blokland, Rogier. 2009. *The Russian Loanwords in Literary Estonian*. Veröffentlichungen der Societas Uralo-Altaica 78. Wiesbaden: Harrassowitz.

Butylov, Nikolaj V. 2007. *Tjurkskie zaimstvovanija v mordovskich jazykach*. Saransk: Respublikanskaja Tip. Krasnyj Oktjabr'.

Calude, Andreea. S. and Mark Pagel. 2011. How Do We Use Language? Shared Patterns in the Frequency of Word Use across 17 World Languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1567): 1101–1107. https://doi.org/10.1098/ rstb.2010.0315

Campbell, Lyle, and Mauricio J. Mixco. 2007. *A glossary of historical linguistics*. Glossaries in Linguistics. Edinburgh: Edinburgh University Press.

Carpelan, Christian, and Asko Parpola. 2001. Emergence, Contacts and Dispersal of Proto-Indo-European, Proto-Uralic and Proto-Aryan in Archaeological Perspective. In Christian Carpelan, Asko Parpola, and Petteri Koskikallio (eds.), *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations*. Suomalais-Ugrilaisen Seuran Toimituksia 242: 55–150. Helsinki: Suomalais-ugrilainen seura.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis. *Language* 91 (1): 194–244.

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. Longman Linguistics Library. Harlow: Longman.

Cronhamn, Sandra. 2018. *Quantifying Loanwords: A Study of Borrowability in the Finnish Vocabulary*. Master thesis, Lund University.

Curnow, Timothy Jowan. 2001. What Language Features Can Be 'Borrowed'? In Alexandra Aikhenvald and Robert M.W. Dixon (eds.), *Areal Diffusion and Genetic Inheritance. Problems in Comparative Linguistics*, 41–36. Oxford: Oxford University Press.

Cygankin, D.V. and Mosin, M.V.M. 2015. *Erzjan' kelen' nur'kine etimologičeskoj slovar'*. Saransk: Mordovskoj knižnoj izd-vas.

Décsy, Gyula. 1988. Slawischer Einfluss Auf Die Uralischen Sprachen. In Denis Sinor (ed.), *The Uralic Languages: Description, History, and Foreign Influences*. Handbuch Der Orientalistik 8: 616–635. Leiden: Brill.

De Heer, Mervi, Mikko Heikkilä, Kaj Syrjänen, Jyri Lehtinen, Outi Vesakoski, Toni Suutari, Michael Dunn, Urho Määttä and Unni-Päivä Leino. 2021. *Uralic basic vocabulary with cognate and loanword information* (Version v2.0) [Data set]. Zenodo. DOI: http://doi.org/10.5281/zenodo.4777568

De Heer, Mervi, Blokland, Rogier, Dunn, Michael, and Vesakoski, Outi. (2023). *Loanwords in basic vocabulary as an indicator of borrowing profiles supplementary materials*. Zenodo. DOI: https://doi.org/10.5281/zenodo.7716522

Dench, Alan. 2001. Descent and Diffusion: The Complexity of the Pilbara Situation. In Alexandra Aikhenvald and Dixon, R.M.W. (eds), *Areal diffusion and genetic inheritance. Problems in comparative linguistics*, 105–133. Oxford: OUP.

Dimmendaal, Gerrit. 2006: Areal Diffusion versus Genetic Inheritance: An African Perspective. In Alexandra Aikhenvald and Dixon, R.M.W. (eds), *Areal diffusion and genetic inheritance. Problems in comparative linguistics*, 358–392. Oxford: OUP.

Dolgopolsky, Aharon. 1964. A Probabilistic Hypothesis Concerning the Oldest Relationships Among the Language Families in Northern Eurasia. In Vitalij V. Shevoroshkin and Thomas L. Markey (eds.), *Typology, Relationship and Time. A collection of papers on language change and relationship by Soviet linguists*, 27–50. Ann Arbor: Karoma Publishers.

EES = Metsämägi, Iiris, Meeli Sedrik, and Sven-Erik Soosaar. 2012. *Eesti Etümoloogiasõnaraamat*. Tallinn: Eesti Keele Sihtasutus.

Evans, Nicholas. 1998. Iwaidja mutation and its origins. In Anna Siewierska, and Jae Jung Song (eds.), Case, *Typology and grammar: In honor of Barry J. Blake*, 115–149. Amsterdam: John Benjamins.

EWUng = Benkő, Loránd, and Béla Büky. 1993. *Etymologisches Wörterbuch Des Ungarischen 1–2*. Budapest: Akadémiai Kiadó.

Frog, and Saarikivi Janne. 2015. De Situ Linguarum Fennicarum Aetatis Ferreae, Pars I. *RMN Newsletter* 9: 64–115.

Gardiner, S.C. 1983. Loan adaptation and the discovery of the genetic relationships of languages. The Slavonic and East European Review 61(4): 512–517.

Gerstner, Károly. 2006. Az idegen eredetű szókészlet. In Péter Siptár (ed.), *A Magyar Nyelv*, 437–480. Budapest: Akadémiai Kiadó.

Gray, Russell D., and Fiona M. Jordan. 2000. Language Trees Support the Express-Train Sequence of Austronesian Expansion. *Nature* 405 (6790): 1052–1055. DOI: https://doi.org/10.1038/35016575.

Green, Roger C. 1999. Integrating Historical Linguistics with Archaeology: Insights from Research in Remote Oceania. *Bulletin of the Indo-Pacific Prehistory Association* 18: 3–16. DOI: https://doi.org/10.7152/bippa.v18i0.11694.

Greenberg, Joseph. 1957. *Essays in Linguistics*. Chicago: University of Chicago Press.

Grünthal, Riho. 2012. Baltic Loanwords in Mordvin. In Riho Grünthal and Petri Kallio (eds.), *A Linguistic Map of Prehistoric Northern Europe*. Suomalais-ugrilaisen seuran toimituksia 266: 297–343. Helsinki: Suomalais-Ugrilainen Seura.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, Qiaomei Fu, Alissa Mittnik, Eszter Bánffy, Christos Economou, Michael

Francken, Susanne Friederich, Rafael Garrido Pena, Fredrik Hallgren, Valery Khartanovich, Aleksandr Khokhlov, Michael Kunst, Pavel Kuznetsov, Harald Meller, Oleg Mochalov, Vayacheslav Moiseyev, Nicole Nicklisch, Sandra L. Pichler, Roberto Risch, Manuel A. Rojo Guerra, Christina Roth, Anna Szécsényi-Nagy, Joachim Wahl, Matthias Meyer, Johannes Krause, Dorcas Brown, David Anthony, Alan Cooper, Kurt Werner Alt and David Reich. 2015. Massive Migration from the Steppe Is a Source for Indo-European Languages in Europe. *Nature* 522: 207–211. DOI: https://doi.org/10.1101/013433.

Häkkinen, Jaakko. 2009. Kantauralin ajoitus ja paikannus: perustelut puntarissa. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 92, 10–56. Helsinki: Suomalais-Ugrilainen Seura.

Häkkinen, Jaakko. 2012. Uralic Evidence for the Indo-European Homeland. Unpublished manuscript. (Accessed 29 October 2020). http://www.elisanet.fi/alkupera/UralicEvidence.pdf.

Haspelmath, Martin, and Uri Tadmor (eds.), 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: de Gruyter Mouton.

Haugen, Einar. 1950. The Analysis of Linguistic Borrowing. *Language* 26 (2): 210–231. https://doi.org/10.2307/410058.

Hesselbäck, André. 2005. *Tatar and Chuvash Code-Copies in Mari*. Studia Uralica Upsaliensia 35. Uppsala: Acta Universitatis Upsaliensis.

Hofstra, Tette. 1997. Zu mutmasslichen frühen ostseefinnischen Lehnwörtern in Nordgermanischen. *Finnisch-ugrische Sprachen in Kontakt*, 127–134.

Holden, C.J. 2002. Bantu Language Trees Reflect the Spread of Farming across Sub-Saharan Africa: A Maximum-Parsimony Analysis. *Proceedings of the Royal Society B: Biological Sciences* 269 (1493): 793–799. https://doi.org/10.1098/rspb.2002.1955.

Holopainen, Sampsa. 2018. Indo-Iranian Loans Confined to Saami? In Sampsa Holopainen and Janne Saarikivi (eds.), *Peri Orthotētos Etymōn. Uusiutuva Uralilainen Etymologia*. Uralica Helsingiensia 11, 135–179. Helsinki: Suomalais-Ugrilainen Seura.

Holopainen, Sampsa. 2019. *Indo-Iranian Borrowings in Uralic: Critical Overview of Sound Substitutions and Distribution Criterion*. PhD dissertation, Helsinki: University of Helsinki.

Holopainen, Sampsa, Petri Kallio, and Janne Saarikivi (eds.), 2016. *Verba vagantur: Jorma Koivulehto in memoriam*. Suomalais-ugrilaisen seuran toimituksia 274. Helsinki: Suomalais-ugrilainen seura.

Honkola, Terhi, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski, Kaj Syrjänen, and Niklas Wahlberg. 2014. Behind Family Trees: Secondary Connections in Uralic Language Networks. *Language Dynamics and Change* 4 (2): 189–221. DOI: https://doi.org/10.1163/22105832-00402007.

Isanbaev, Nikolaj Isanbaevič. 1989. *Marijsko-tjurkskie jazykovye kontakty: Č. 1, (Tatarskie i Baškirskie zaimstvovanija*). Joškar-Ola: Marijskoe knižnoe izd-vo.

Isanbaev, Nikolaj Isanbaevič. 1994. *Marijsko-tjurkskie jazykovye kontakty. 2, Slovar' tatarskih i baškirskih zaimstvovanij*. Joškar-Ola: Naučnyj centr Finno-Ugrovedenija.

Itkonen, Erkki. 1961. *Suomalais-ugrilaisen kielen- ja historiantutkimuksen alalta*. Tietolipas 20. Helsinki: Suomalaisen kirjallisuuden seura.

Johanson, Lars. 2002. Contact-Induced Change in a Code-Copying Framework. In Mari C. Jones and Edith Esch (eds.), *Language Change*, 285–310. Berlin – New York: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110892598.285.

Johanson, Lars. 2013. Written Language Intertwining. In Peter Bakker and Yaron Matras (eds.), *Contact Languages: A Comprehensive Guide*, 273–351. Berlin – Boston: De Gruyter. DOI: https://doi.org/10.1515/9781614513711.273.

Joki, Aulis Johannes. 1973. *Uralier Und Indogermanen: Die Älteren Berührungen Zwischen Den Uralischen Und Indogermanischen Sprachen*. Suomalais-ugrilaisen seuran toimituksia 151. Helsinki: Suomalais-ugrilainen seura.

Junttila, Santeri. 2012. The Prehistoric Context of the Oldest Contacts between Baltic and Finnic Languages. In Riho Grünthal and Petri Kallio (eds.), *A Linguistic Map of Prehistoric Northern Europe*. Suomalais-ugrilaisen seuran toimituksia 266: 261–296. Helsinki: Suomalais-Ugrilainen Seura.

Junttila, Santeri. 2015. *Tiedon kumuloituminen ja trendit lainasanatutkimuksessa: kantasuomen balttilaislainojen tutkimushistoria*. PhD dissertation, Helsinki: University of Helsinki.

Junttila, Santeri. 2018. Altlettgallische Lehnwörter in den mordwinischen Sprachen? *Finnisch-Ugrische Forschungen* 64: 72–91.

Kallio, Petri. 2006. Suomen Kantakielten Absoluuttista Kronologiaa. *Virittäjä* 110: 2–25.

Kallio, Petri. 2012. The Prehistoric Germanic Loanword Strata in Finnic. In Riho Grünthal and Petri Kallio (eds.), *A Linguistic Map of Prehistoric Northern Europe*. Suomalais-ugrilaisen seuran toimituksia 266: 225–238. Helsinki: Suomalais-Ugrilainen Seura.

Keresztes, László. 1998. Etymologisches Wörterbuch Des Ungarischen. Band II. (Kor-Zs). In Loránd Benkö, Red. Károly Gerstner, Antónia S. Hámori, Gábor Zaicz (eds.), *Acta Linguistica Hungarica* 45 (3): 383–388. DOI: https://doi.org/10.1023/A:100967131 8540.

KESK = Lytkin, Vasilij. and F.S. Guljaev. 1970. *Kratkii etimologitšeskii slovar komi jazyka*. Institut jazykoznanija An SSSR: Komi fifial. Moskva: Nauka.

Koivulehto, Jorma. 2001. The Earliest Contacts between Indo-European and Uralic Speakers in the Light of Lexical Loans. In Christian Carpelan, Asko Parpola, and Petteri Koskikallio (eds.), *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations*. Suomalais-Ugrilaisen Seuran Toimituksia 242: 235–264. Helsinki: Suomalais-ugrilainen seura.

Koivulehto, Jorma. 1999. Varhaiset indoeurooppalaiskontaktit: aika ja paikka lainasanojen valossa. In Paul Fogelberg (ed.), *Pohjan poluilla: suomalaisten juuret nykytutki-*

muksen mukaan. Bidrag till kännedom av Finlands natur och folk 153: 207–236. Helsinki: Tiedekirja, Suomen tiedeseura.

Korenchy, Éva. 1988. Iranischer Einfluß in Den Finnisch-Ugrischen Sprachen. In Denis Sinor (ed.), *The Uralic Languages: Description, History, and Foreign Influences.* Handbuch Der Orientalistik 8: 665–681. Leiden: Brill.

Kümmel, Martin. 2020. Substrata of Indo-Iranic and related questions. In Romain Garnier (ed.), *Loanwords and substrata: proceedings of the colloquium held in Limoges (5th–7th June, 2018)*: 237–277. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.

Labov, William. 1963. The Social Motivation of a Sound Change. *Word* 19 (3): 273–309. DOI: https://doi.org/10.1080/00437956.1963.11659799.

LÄGLOS = Hahmo, Sirkka-Liisa, Tette Hofstra, Osmo Nikkilä, and Andries Dirk Kylstra. 1991. *Lexikon Der Älteren Germanischen Lehnwörter in Den Ostseefinnischen Sprachen.* Vol. 1–3. Amsterdam: Rodopi.

Laakso, Johanna. 2014. The Prehistoric Multilingual Speaker: What Can We Know about the Multilingualism of Proto-Uralic Speakers? *Finnisch- Ugrische Mitteilungen* 38: 93–114.

Leer, Jeff. 1990. Tlingit: a portmanteau family? In Baldi, Philip (ed). *Linguistic change and reconstruction methodology. Trends in Linguistics. Studies and monographs 45*: 73–98. Berlin/ New York: Mouton de Gruyter.

Lytkin, Vasilij and F.S Guljaev. 1975. *Dopolnenija k Kratkomu Slovarju Komi Jazyka* 18. Komi Filologija. Syktuvkar: Trudy instituta jazyka, literatury i istorii.

Mallory, James. P. 1989. *In Search of the Indo-Europeans: Language, Archaeology and Myth.* London: Thames and Hudson.

Marjomaa, Marko. 2014. North Sámi in Norway: ELDIA Case-Specific Report. In Johanna Laakso (ed.), *Working Papers in European Language Diversity.* Research consortium ELDIA. Mainz: Research consortium ELDIA.

Matras, Yaron. 2009. *Language Contact.* Cambridge–New York: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511809873.

Metsäranta, Niklas. 2020. *Periytyminen ja lainautuminen: Marin ja permiläisten kielten sanastontutkimusta.* PhD dissertation, Helsinki: University of Helsinki.

Milroy, James, and Lesley Milroy. 1985. Linguistic Change, Social Network and Speaker Innovation1. *Journal of Linguistics* 21 (2): 339–384. DOI: https://doi.org/10.1017/S0022226700010306.

Mosin, Mihail. 2002. Об Особенностях Мордовско-Русского Двухязычия у Мордвы-Эрзи в Республике Мордовия [On the Characteristics of the Mordvin–Russian Bilingualism of Erzya-Mordvins in the Republic of Mordovia]. In Jorma Luutonen (ed.), *Volgan Alueen Kielikontaktit*: 160–166. Turku: University of Turku.

Muysken, Pieter. 2013. Language Contact Outcomes as the Result of Bilingual Optimization Strategies. *Bilingualism: Language and Cognition* 16 (4): 709–730. DOI: https://doi.org/10.1017/S1366728912000727.

Norvik Miina, Jing Yinqi, Dunn Michael, Forkel Robert, Honkola Terhi, Klumpp Gerson, Kowalik Richard, Metslang Helle, Pajusalu, Karl, Piha Minerva, Saar Eva, Saarinen Sirkka, Vesakoski Outi. (2022). Uralic typology in the light of a new comprehensive dataset. *Journal of Uralic Linguistics*, 1(1), 4–42. DOI: https://doi.org/10.1075/jul.00002.nor.

Översti, Sanni, Kerttu Majander, Elina Salmela, Kati Salo, Laura Arppe, Stanislav Belskiy, Heli Etu-Sihvola, Ville Laakso, Esa Mikkola, Saskia Pfrengle, Mikko Putkonen, Jussi-Pekka Taavitsainen, Katja Vuoristo, Anna Wessman, Antti Sajantila, Markku Oinonen, Wolfgang Haak, Verena J. Schuenemann, Johannes Krause, Jukka U. Palo and Päivi Onkamo. 2019. Human Mitochondrial DNA Lineages in Iron-Age Fennoscandia Suggest Incipient Admixture and Eastern Introduction of Farming-Related Maternal Ancestry. *Scientific Reports* 9 (1): 1–14. DOI: https://doi.org/10.1038/s41598-019-51045-8.

Paasonen, Heikki. 1897. *Die türkischen Lehnwörter im Mordwinischen*. Suomalais-ugrilaisen Seuran aikakauskirja 15. Helsinki: Suomalais-Ugrilainen Seura.

Pakendorf, Brigitte. 2014. Coevolution of Languages and Genes. *Current Opinion in Genetics and Development*, Genetics of human evolution, 29: 39–44. DOI: https://doi.org/10.1016/j.gde.2014.07.006.

Van Pareren, Remco. 2009. Die direkten baltischen Lehnwörter im Mordwinischen. *Finnisch-Ugrische Mitteilungen* 30/31: 69–147.

Parpola, Asko. 2012. Formation of the Indo-European and Uralic Language Families in the Light of Archaeology: Revised and integrated 'total' correlations. In Riho Grünthal and Petri Kallio (eds.), *A Linguistic Map of Prehistoric Northern Europe*. Suomalais-ugrilaisen seuran toimituksia 266: 119–184. Helsinki: Suomalais-Ugrilainen Seura.

Peltola, Sanni, Majander, Kerttu, Makarov, Nikolaj, Dobrovolskaya, Maria, Nordqvist, Kerkko, Salmela, Elina, and Onkamo, Päivi. 2023. Genetic admixture and language shift in the medieval Volga-Oka interfluve. *Current biology*. 33 (1): 174–182. DOI: https://doi.org/10.1016/j.cub.2022.11.036.

Rantanen, Timo, Tolvanen, Harri, Roose, Meeli, Ylikoski, Jussi, and Vesakoski, Outi. 2022. Best practices for spatial language data harmonization, sharing and map creation – A case study of Uralic. *PLoS ONE* 17(6). DOI: https://doi.org/10.1371/journal.pone.0269648.

Rédei, Károly. 1963. Juraksamojediche Lehnwörter in Der Syrjänischen Sprache. *Acta Linguistica Academiae Scientiarum Hungaricae* 13: 275–310.

Rédei, Károly. 1964. Ob-Ugor Jövevényszavak a Zürjén Nyelvben. *Nyelvtudományi Közlemények* 66: 3–15.

Rédei, Károly. 1970. *Die syrjänischen Lehnwörter im Wogulischen*. Budapest: Akadémiai Kiadó.

Rédei, Károly. 1986. *Zu Den Indogermanisch-Uralischen Sprachkontakten*. Veröffentli-

chungen Der Kommission Für Linguistik Und Kommunikationsforschung 16. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Rédei, Károly, and András Róna-Tas. 1972. A permi nyelvek őspermi kori bolgártörök jövevényszavai. *Nyelvtudományi Közlemények* 74: 281–298.

Renfrew, Colin. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins*. New York: Cambridge University Press.

Riese, Timothy. 1998. Permian. In Daniel Abondolo (ed.), *The Uralic Languages*, 249–275. London: Routledge.

Rießler, Michael. 2009. Kildin Saami Vocabulary. In Haspelmath, Martin, and Uri Tadmor (eds.), 2009. *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wold.clld.org/.

Róna-Tas, András. 1988. Turkic Influence on the Uralic Languages. In Denis Sinor (ed.), *The Uralic Languages: Description, History, and Foreign Influences*. Handbuch Der Orientalistik 8: 742–780. Leiden: Brill.

Róna-Tas, András, and Berta, Árpád. 2011. *West Old Turkic. Turkic Loanwords in Hungarian*. Turcologica 84, 1–2. Wiesbaden: Harrassowitz.

Ross, Malcolm. 2007. Calquing and Metatypy. *Journal of Language Contact* 1 (1): 116–143. DOI: https://doi.org/10.1163/000000007792548341.

Rozhanskiy, Fedor, and Zhivlov, Mikhail. 2019. Votic and Ingrian Core Vocabulary in the Finnic Context: Swadesh Lists of Five Related Varieties. *Linguistica Uralica* 55 (2). DOI: https://doi.org/10.3176/lu.2019.2.01.

Rueter, Jack. 2013. The Erzya Language. Where Is It Spoken? *Études Finno-Ougriennes* 45. DOI: https://doi.org/10.4000/efo.1829.

Saarikivi, Janne, and Lavento, Mika. 2012. Linguistics and Archaeology: A Critical View of an Interdisciplinary Approach with Reference to the Prehistory of Northern Scandinavia. In Charlotte Damm and Janne Saarikivi (eds.), *Networks, Interaction and Emerging Identities in Fennoscandia and Beyond*. Suomalais-Ugrilaisen Seuran Toimituksia 265, 177–216. Helsinki: Suomalais-Ugrilainen Seura.

Saarikivi, Janne. 2011. Saamelaiskielet – nykypäivää ja historiaa. In Irja Seurujärvi-Kari, Petri Halinen and Risto Pulkkinen (eds.), *Saamentutkimus tänään*. Tietolipas 234: 177–219. Helsinki: Suomalaisen Kirjallisuuden Seura.

Saarikivi, Janne. 2018. Finnic and Other Western Uralic Borrowings in Permian In Sampsa Holopainen and Janne Saarikivi (eds.), *Peri Orthotētos Etymōn. Uusiutuva Uralilainen Etymologia*. Uralica Helsingiensia 11, 269–355. Helsinki: Suomalais-Ugrilainen Seura.

Saarinen, Sirkka. 2010. Marin Sanaston Alkuperästä. In Sirkka Saarinen, Kirsti Siitonen, and Tanja Vaittinen (eds.), *Sanoista Kirjakieliin: Juhlakirja Kaisa Häkkiselle 17. Marraskuuta 2010*. Suomalais-Ugrilaisen Seuran Toimituksia 259: 335–341. Helsinki: Suomalais-Ugrilainen Seura.

Sammallahti, Pekka. 2001. The Indo-European Loanwords in Saami. In Christian

Carpelan, Asko Parpola, and Petteri Koskikallio (eds.), *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations*. Suomalais-Ugrilaisen Seuran Toimituksia 242: 397–415. Helsinki: Suomalais-ugrilainen seura.

Simon, Zsolt. 2020. Urindogermanische Lehnwörter in Den Uralischen Und Finno-Ugrischen Grundsprachen. *Indogermanische Forschungen* 125 (1): 239–266.

Soosaar, Sven-Erik. 2013. The Origins of Stems of Standard Estonian – a Statistical Overview. *Trames: A Journal of the Humanities and Social Sciences; Tallinn* 17 (3): 273–300.

Swadesh, Morris. 1955. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics* 21 (2): 121–137. DOI: https://doi.org/10.1086/464321.

Swadesh, Morris. 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.

Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski and Niklas Wahlberg. 2013. Shedding More Light on Language Classification Using Basic Vocabularies and Phylogenetic Methods: A Case Study of Uralic. *Diachronica* 30 (3): 323–352. DOI: https://doi.org/10.1075/dia.30.3.02syr.

Syrjänen, Kaj, Jyri Lehtinen, Outi Vesakoski, Mervi de Heer, Toni Suutari, Michael Dunn, Urho Määttä, and Unni-Päivä Leino. 2018. *Lexibank/UraLex: UraLex Basic Vocabulary Dataset*. (Version v1.0) [Data set]. Zenodo. DOI: http://doi.org/10.5281/zenodo.1459402

Tambets, Kristiina, Bayazit Yunusbayev, Georgi Hudjashov, Anne-Mai Ilumäe, Siiri Rootsi, Terhi Honkola, Outi Vesakoski, Quentin Atkinson, Pontus Skoglund, Alena Kushniarevich, Sergey Litvinov, Maere Reidla, Ene Metspalu, Lehti Saag, Timo Rantanen, Monika Karmin, Jüri Parik, Sergey I. Zhadanov, Marina Gubina, Larisa D. Damba, Marina Bermisheva, Tuuli Reisberg, Khadizhat Dibirova, Irina Evseeva, Mari Nelis, Janis Klovins, Andres Metspalu, Tõnu Esko, Oleg Balanovsky, Elena Balanovska, Elza K. Khusnutdinova, Ludmila P. Osipova, Mikhail Voevoda, Richard Villems, Toomas Kivisild, and Mait Metspalu. 2018. Genes Reveal Traces of Common Recent Demographic History for Most of the Uralic-Speaking Populations. *Genome Biology* 19. DOI: https://doi.org/10.1186/s13059-018-1522-1.

Thomason, Sarah. 2008. Social and Linguistic Factors as Predictors of Contact-Induced Change. *Journal of Language Contact* 2 (1): 42: 56. DOI: https://doi.org/10.1163/000000008792525381.

Thomason, Sarah Grey, and Terrence Kaufman. 1988. *Language Contact, Creolization and Genetic Linguistics*. Berkeley: University of California Press.

Trask, Larry and Robert McColl Millar (ed). 2015. *Trask's historical linguistics* (*3rd ed.*). London: Routledge.

TschWB = Moisio, Arto, and Sirkka Saarinen. 2008. *Tscheremissisches Wörterbuch*. Lexica Societatis Fenno-Ugricae 32. Helsinki: Suomalais-Ugrilainen Seura.

Van Hout, Roeland, and Pieter Muysken. 1994. Modeling Lexical Borrowability. *Language Variation and Change* 6 (1): 39–62. DOI: https://doi.org/10.1017/S095439450000 1575.

Veršinin, V.I.B. 2004. *Etimologičeskij slovar' mordovskih* (*erzjanskogo i mokšanskogo jazykov*). Joškar-Ola.

Wichmann, Søren, Cecil H. Brown, and Eric W. Holman. 2020. *The ASJP Database*. Jena: Max Planck Institute for the Science of Human History. https://asjp.clld.org/.

Wickman, Bo. 1988. The History of Uralic Languages. In Denis Sinor (ed.), *The Uralic Languages: Description, History, and Foreign Influences*. Handbuch Der Orientalistik 8: 792–818. Leiden: Brill.

WOLD = Haspelmath, Martin, and Uri Tadmor (eds). 2009. *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wold.clld.org/.

Zaicz, Gábor. 1998. Mordva. In Daniel Abondolo (ed.), *The Uralic Languages*, 184–218. London: Routledge.