LUND
UNIVERSITY

Quantifying loanwords

A study of borrowability in the Finnish lexicon

Sandra Cronhamn

Supervisor: Gerd Carling

Co-supervisor: Niklas Johansson

Centre for Language and Literature, Lund University
MA in Language and Linguistics, General Linguistics
SPVR01 Language and Linguistics: Degree Project – Master's (Two Years) Thesis, 30 credits

January 2018

Abstract

The current study set out to investigate patterns of loanwords in a sample of 1,460 lexical meanings in the Finnish lexicon by means of quantitative methods. The methodology used was borrowed from the Loanword Typology project (Haspelmath & Tadmor 2009a), and consisted of a template including various fields, where information about each lexical item was coded. The fields included measures such as *Borrowed status*, *Age* and *Donor language*, and the data was collected from etymological dictionaries. The values coded for the lexical meanings were analysed to answer the research questions, which had to do with e.g. loanword patterns in relation to semantic domains, immediate donor languages and loanword age. The loanword patterns found in Finnish were also compared to the cross-linguistic averages found by the Loanword Typology project. It was found that, in general, Finnish is a fairly typical language from a loanword typological point of view. It was also corroborated that the overwhelming majority of loanwords in Finnish come from Indo-European, especially from Germanic languages. Support was also found for correlations between loanword age and donor language branch, in that the loanwords from different language branches layered themselves timewise. Although the findings of this study are largely in line with the previous research on loanwords in Finnish, the most important contribution of this thesis is the restructuring of the previous research into a format which makes it comparable to corresponding data in a relatively large sample of languages cross-linguistically.

Keywords: loanwords, Finnish, Uralic languages, language contact, borrowing

Abstrakti

Käsillä olevan tutkimuksen tavoitteena on tutkia suomen kielen sanastossa esiintyviä lainasanoja. Tutkimus on toteutettu kvantifioimalla 1460 leksikaalisen merkityksen etymologiaa lainaamalla projektissa Loanword Typology project (Haspelmath & Tadmor 2009a) käytettyä metodia, jossa sovelletun mallin mukaan etymologista tietoa jokaisesta lekseemistä on kerätty etymologisista sanakirjoista ja kvantifioitu. Analyysi keskittyy löytämään vastauksia kysymyksiin esimerkiksi lekseemien lainautumistilasta, iästä, sekä lainanantajakielestä ja -kieliperheestä. Kerättyä aineistoa analysoimalla tämä tutkimus pyrkii vastaamaan tutkimuskysymyksiin, joiden aiheena on muun muassa tutkia lainasanojen suhteita esimerkiksi semanttisiin luokkiin, lainanantajakieliin sekä lainasanojen ikään. Lainasanatutkimuksen tuloksia verrataan myös vastaaviin, kielirajat ylittäviin tuloksiin, jotka löytyivät edellisessä tutkimuksessa Loanword Typology project. Tulokset osoittavat suomen kielen lainasanojen seuranneen pääsääntöisesti typologisesta perspektiivistä melko tyypillisiä taipumuksia. Tulokset vahvistavat myös valtaosan suomen kielen lainasanoista olevan indoeurooppalaisperäisiä, joista puolestaan valtaosa on germaanisperäisiä lainoja. Tutkimustulokset vahvistavat myös lainanantajakieliryhmien välistä korrelaatiota siten, että indoeurooppalaisista kielihaaroista peräisin olevat lainasanat ryhmittyvät selkeästi toisistaan erottuviin ikäkerrostumiin. Vaikka tutkimuksen tulokset ovatkin pääasiassa odotuksenmukaisia edellisen tutkimuksen valossa, tämän tutkimuksen tärkein myötävaikutus onkin edellisen etymologisen tutkimuksen uudelleenjärjestely sellaiseen muotoon, että tuloksia voi helposti verrata muiden kielten osalta tehtyjen, samankaltaisten tutkimusten tuloksiin.

Acknowledgements

First of all, I would like to thank my two supervisors, Gerd Carling and Niklas Johansson, for their time, help, support and advice. I would also like to thank my consultant and mother, Taina Cronhamn, who spent many hours helping me identify the Finnish lexemes to the word list. *Sydämelliset kiitokset* as well to my aunts Leena Töyry and Anna-Liisa Kiesi and my godmother Eeva Saukkonen, who all served as over-the-phone advisors in the collection process. I would also like to thank Juhan Luomala and Lari-Valtteri Suhonen for helping me with proof-reading, Olof Lundgren for his comments, Uri Tadmor and Brad Taylor for providing me with the Loanword Typology project's original data collection template, and Mechtild Tronnier for reminding me again and again to focus on my studies and finalise this thesis. Finally, I would like to thank Anneliese Kelterer and Arthur Holmer, the opponent and examiner at the defence of this thesis, for their valuable comments and suggested improvements that played an important role in finalising the published version.

Table of contents

List of figures

List of tables

Abbreviations

2SG	2nd person singular
3SG	3rd person singular
BS	Balto-Slavic
C	consonant
CONNEG	connegative
Eng.	English
F	Finnic
Fi.	Finnish
FP	Finno-Permic
FS	Finno-Saamic
FU	Finno-Ugric
FV	Finno-Volgaic
Germ.	Germanic
IE	Indo-European
II	Indo-Iranian
IMP	imperative
LG	Low German
LWT	Loanword Typology (project)
Neo-Cl.	Neo-Classical
NES	Nykysuomen etymologinen sanakirja (dictionary)
NF	Northern Finnic
OES	Old East Slavic
ON	Old Norse
OS	Old Swedish
PB	Proto-Baltic
PBS	Proto-Balto-Slavic
PF	Proto-Finnic
PG	Proto-Germanic
PIE	Proto-Indo-European
PII	Proto-Indo-Iranian
PI	Proto-Iranian
POT	potential
PS	Proto-Slavic
PU	Proto-Uralic
RQ	research question
Ru.	Russian
S	segment
Swe.	Swedish
SG	singular
SSA	Suomen sanojen alkuperä (dictionary)
U	Uralic
UEW	Uralisches etymologisches Wörterbuch (dictionary)
V	vowel

1 Introduction

Loanwords consist of lexical material that has been transferred from one linguistic variety to another through contact between their speakers. Studying them can give us access to a lot of interesting information about language contact, not least from the past – once a loanword is incorporated into the recipient language, it can be passed down to coming generations just like native vocabulary, and constitute evidence of language contact that may have happened thousands of years ago.

In order to establish a loan etymology, one has to go into great detail when it comes to historical/societal context, dating of the borrowing, semantic matching, rules of sound substitution etc., so many loanword studies have necessarily been focused on individual lexical items or groups thereof. However, with these in place, it is possible to start conducting studies with a broader aim, focusing on the bigger picture. One such study, which had a cross-linguistic scope, is the Loanword Typology project (Haspelmath & Tadmor 2009a). The findings of this project serve as material of comparison for the study at hand, and therefore the methodology has also been borrowed from the project (more on this under 3.1).

Finnish is a language that, through its various language stages, has borrowed a large amount of linguistic material from neighbouring languages (cf. e.g. Carpelan, Parpola & Koskikallio 2001). There is a long tradition of loanword studies within Finnish linguistics (cf. section 1.1 below), and loanwords in Uralic languages have also played an important part in Indo-European linguistics. To the best of my knowledge, however, no study so far has aimed to quantify the borrowability of the Finnish lexicon and compare the patterns to that of a large number of other languages. This is the aim of the current study, which is based on data from etymological dictionaries (primarily Häkkinen 2013) and performed by means of the methodology from the Loanword Typology project (Haspelmath & Tadmor 2009a; see more under 3.1 below).

The well-described state of Finnish makes it a suitable candidate for the study - the better the existing etymologies, the more reliable the statistics. In addition, for future purposes, the results could perhaps contribute to a deeper understanding of how the sociolinguistic history of a language can be reflected in its etymological structures, which in turn could help us interpret such structures in lesser-known languages.

1.1 Aims, scope, research questions and hypotheses

The aims of this thesis are to identify large-scale patterns in Finnish loanwords, and to focus on the big picture. The subject of study is the *direct* contact situations Finnish has been involved in. Therefore, it does *not* lie within the scope of this thesis to e.g. trace the oldest discoverable origin of any individual word. It is also outside the scope to, in any way, focus on any specific lexical item. The study is focused particularly on Finnish – albeit including earlier language stages shared with other Uralic languages – not on Uralic languages in general.

More specifically, the topics I am interested in investigating can be divided into two groups: on the one hand, the big, overarching loanword patterns (as well as how they compare to those of other languages), and on the other hand, the relationship of these patterns to other factors, such as loanword age, semantic domains and donor languages. The research questions that the project revolves around have been formulated as follows:

1. How much of the Finnish vocabulary is borrowed?
2. Which semantic domains have the highest vs. the lowest amount of loanwords?
3. Does the Finnish sample differ in any noticeable way from the cross-linguistic tendencies found in the Loanword Typology project?
4. Is there a correlation between semantic domain and donor language (family)?
5. Is there a correlation between time period and donor language (family)?

In order to answer these questions, I have chosen a suitable method by which the relevant data is quantified and coded in a format that makes it comparable to other languages. The methodology is explained in detail in section 3.

Based on the previous literature within the subject and in adjacent fields of study, four hypotheses have been developed:

A. There is a salient difference in the percentage of loanwords between different semantic domains.
B. Finnish has a higher percentage of loanwords in general compared to other languages.
C. There is a correlation between donor language and semantic domain.
D. The donor languages will cluster in (possibly partly overlapping) layers timewise, roughly indicating when the language contact took place.

Hypothesis A is based on earlier research about lexical stability vs. contact-sensitivity, which clearly points to a variation in borrowability between different semantic domains, even though the more precise circumstances still need further investigation (cf. e.g. Swadesh 1950, Haspelmath & Tadmor 2009a, Hock & Joseph 1996:257-258; more under 2.2.2 and 3.1).

Hypotheses B, C, and D were developed in response to theories on the prehistory of Uralic languages and their contact with the Indo-European language family (cf. e.g. Häkkinen 1998, 2001, Kallio 2002, 2015a, 2017; further discussed under 2.3). Hypothesis B was also influenced by the more recent socio-political history of the Finnish-speaking territories (more under 2.1.2).

1.2 Outline

Chapter 2 presents the theoretical framework that have served as the background for this thesis, such as an overview of important theoretical concepts (the comparative method, loanword studies and linguistic palaeontology), as well as an introduction to Finnish and to the Uralic language family. Earlier research on loanwords in Uralic languages is also presented in this chapter.

Chapter 3 introduces the methodology, beginning with a presentation of the Loanword Typology project, and thereafter a detailed description of the working process of this study. Methodological problems are reflected on and exemplified, and some criticism towards the method is brought up.

Chapter 4 presents the analyses and results of the study, research question by research question. The results are illustrated by tables and diagrams, and the most important points are explained in text.

Chapter 5 contains a more detailed discussion of the results presented in chapter 4, including implications beyond answering the research questions. More aspects of the methodology are being discussed here, in the light of the results it produced. The study is being put in a larger context in the discussion of future research.

Lastly, chapter 6 provides a summary of the thesis as well as some concluding remarks.

2 Background

2.1 Finnish

Finnish is spoken by around 5 million people and belongs to the Finnic branch of the Uralic language family. The Finnic languages are all spoken around the Baltic Sea area. Just like its relatives, Finnish is characterised by a synthetic, agglutinating morphology (with some fusional traits). Word formation is mainly achieved by suffixation (Karlsson 1999). Below, a brief introduction to some phonological phenomena relevant for the current study is provided.

2.1.1 Phonological structure

The Finnish vowel inventory consists of eight vowels, graphemically represented as *i, e, ä, y, ö, u, o* and *a.* Their approximate locations in the vowel space are shown in Figure 1 below (adapted from Suomi et al. 2008:21; *ä* and *ö* are represented by *æ* and *ø*, respectively). All vowels have both short and long forms, which are phonemically contrasted and differ only in quantity, not in quality. In addition, Finnish has 16 common diphthongs: *ei, äi, ui, ai, oi, öi, yi, au, ou, eu, iu, äy, öy, ie, yö,* and *uo* (Suomi et al. 2008:20-23, Karlsson 1999:10-14).

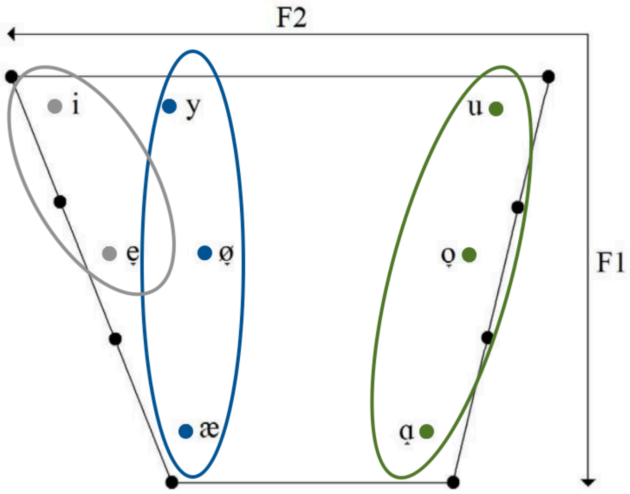

Figure 1. The Finnish vowel inventory.

Finnish also has *vowel harmony*, which means that its vowels are divided into groups according to which other vowels they can co-occur with (see colour coding in Figure 1: blue (front harmonic), green (back harmonic) and grey (neutral)). Vowel harmony operates on the word level, making it a suprasegmental feature, where the root governs what vowels are allowed in the affixes. Vowels from the back harmonic and the front harmonic sets can never co-occur; the two neutral vowels, however, may occur with either set. Because of this, many suffixes have two allomorphs (Karlsson 1999:16-17). All native vocabulary (as well as most loanwords) obeys these rules – however, compounds, which consist of two lexical roots, are excepted from this system and do not count as violations. There are some examples of loanwords which breach the vowel harmony, e.g. *amatööri* 'amateur' (Duncan 2008). Such words usually consist of several syllables, and according to Duncan and others, they may be syllabically analysed as compounds. In other words; while not all loanwords obey this rule in a strict sense, the system is still productive enough to force an alternative, harmonic interpretation.

Unlike the vowels, the consonant inventory differs in size between different varieties of Finnish – the minimum is 11 phonemes and the maximum is 17 phonemes (Suomi et al. 2008:24-25). The consonants can be plotted out as in the IPA chart in Table 1 below, a slightly simplified adaptation from Suomi et al. (2008:38), with the most unusual phonemes in parentheses.

Table 1. The Finnish consonant inventory.

	Bilabial	**Labio-dental**	**(Denti-) alveolar**	**Palato-alveolar**	**Palatal**	**Velar**	**Glottal**
Plosive	p (b)		t			k (g)	
Semiplosive			d				
Fricative		f	s	(ʃ)			
Glottal continuant							h
Nasal	m		n			ŋ	
Trill			r				
Lateral approximant			l				
Central approximant		ʋ			j		

Just as with the vowels, all consonants have a two-way phonemic contrast in length (although some long phonemes are rarer than others). Most consonants also have several allophones – see Suomi et al. (2008:23-38) for a thorough survey.

Both long and short sounds (vowels as well as consonants) are permitted in almost every position in a word. This leads to many possible combinations, and the same phoneme sequences, differing only in quantity, can convey different meanings: e.g. *tule* 'come.IMP.2SG', *tuule* 'blow.CONNEG (of wind)', *tuulle* 'blow.CONNEG.POT (of wind)', *tuulee* 'blow.3SG (of wind)', *tulee* 'come.3SG', *tullee* 'come.3SG.POT', *tuullee* 'blow.3SG.POT (of wind)' (Karlsson 1999:12).

Another important phonological phenomenon in Finnish is *consonant gradation*, i.e. the alternating grades of *p, t* and *k* depending on their environment. For example, *pp* alternates with *p* (*kaappi* 'cupboard' - *kaapi/ssa* 'in the cupboard'), *p* alternates with *v* (*tupa* 'hut' - *tuva/ssa* 'in the hut'), and *mp* alternates with *mm* (*ampu-* 'shoot' - *ammu/mme* 'we shoot'). For more examples, see Karlsson (1999:28-29).

Finnish has a quite simple syllabic structure, which (for fully native words) "can be described by the template (C)V(S)(C) in which "S" refers to a segment, either V or C, and in which each segment is a phoneme, given the syntagmatic interpretation of quantity" (Suomi et al. 2008:65). A few other types of syllables (introduced through contact) exist, including syllables with initial consonant clusters, although they are much less frequent. Finnish has a strict rule of word-initial stress (Karlsson 1999:15).

Patterns such as vowel harmony and consonant gradation put strong constraints on Finnish phonology, and therefore also have a strong influence on the phonological integration of loanwords. We will return to loanword integration under 2.2.2 below.

2.1.2 Socio-political history

(Pre-)Saami people have lived in the northern areas of Scandinavia for thousands of years. It is possible that also (pre-)Finnish-speakers have resided in present-day Finland for long, but the first certain records of such a population date to the 12th century BCE (Latomaa & Nuolijärvi 2002:100). During the same century, people from central Sweden started settling along the Gulf of Bothnia. Swedes kept migrating to the coastal areas in what we today call Finland until the end of the 14th century, and have since then lived side by side with Finnish-speakers. (Latomaa & Nuolijärvi 2002:104). However, there are strong indications of several contact situations between the linguistic ancestors of these groups much earlier in prehistory (which, by definition, is undocumented), cf. e.g. Carpelan, Parpola & Koskikallio 2001.

In 1323, the peace treaty of Nöteborg divided the areas constituting present-day Finland in two: the eastern parts were assigned to Novgorod while the western parts were assigned to

Sweden. “Finland” – or most of it, at least – then belonged to Sweden for the next five centuries. During this period, Swedish was the language used in all higher functions in society. Latomaa & Nuolijärvi (2002:96-97) state that this “implanted not only the Swedish language but also the legal and social structures of Sweden deeply into the Finnish soil”, and that “[w]hile under Swedish rule, Finland was legally a set of provinces governed from Stockholm rather than a national entity”.

In the 16th century, the religious Reformation reached Sweden and Finland. Mikael Agricola, the bishop of Turku considered “the father of written Finnish” (Karlsson 1999:3), wrote the first book published in Finnish in 1543 and translated the New Testament into Finnish in 1548. These events initiated the rise of Finnish as a written language (Latomaa & Nuolijärvi 2002:97, 101). The written tradition of Finnish is thus remarkably younger than that of many Indo-European languages.

In 1809, Sweden lost a war to Russia. This led to the surrendering of the Finnish provinces, and Finland became an autonomous Grand Duchy of Russia. Despite the Russian rule, Swedish remained the official language of the country. During the following century, however, nationalism gained foothold in Finland as in many other countries, particularly through Elias Lönnrot’s publication of Kalevala, the Finnish national epic, in 1835 (Latomaa & Nuolijärvi 2002:97). This movement led to the strengthening of the national spirit, and thereby also of the Finnish language. In 1863, Tsar Alexander II signed the Language Decree, which “gave Finnish an equal status with Swedish in official matters concerning the Finnish-speaking population” (Halonen, Ihalainen & Saarinen 2014:154) and led to Swedish gradually losing its dominant position.

In 1917, the Finnish Parliament approved the declaration of independence, and in 1919 Finland became a republic (Latomaa & Nuolijärvi 2002:97). Although the Swedish-speaking population is decreasing by the year and currently only makes up around 5% of the population (Suomen virallinen tilasto 2017), Finland has remained officially bilingual to this day. The Constitution states that the two languages have equal status, and recognises (apart from the two national languages) three official minority languages: Saami, Romani & Finnish Sign language (Constitution Act of Finland 1919, §17).

2.2 Theoretical framework

2.2.1 The comparative method and language classification

Languages that belong to the same language family are said to be *genetically related*. This means that they descend from a common ancestor, i.e. an earlier language stage, called the *proto-language*. The comparative method is a way of reconstructing this earlier language stage by comparing the descendants, and forming theories on how changes from the earlier stage (i.e. *divergence*) can be accounted for (Campbell 2013:107).

The comparative method constitutes the basis for the classification of languages into language families, as well as for the sub-classification of the languages into branches within the family tree. In simplified terms, the method is applied to sets of *cognates*, i.e. related (or presumably related) words in related (or presumably related) languages, all (presumably) inherited from the proto-language, to determine if any regular sound correspondences can be found in the material. These sound correspondences can then serve as the basis for setting up rules about how the sounds must have evolved. The underlying theory is that sounds change in ways that are regular and therefore traceable[1] (Campbell 2013:14ff). Thus, if we find regularity in our cognate set, we may postulate reconstructed proto-forms[2], sound changes, and a relative chronology of the sound changes. If some of the related languages share the same sound changes, we may set up theories on intermediate proto-languages, which in turn serve as the basis for the nodes in the family tree. In Figure 2, which is an illustration of the basic difference between inherited and borrowed vocabulary, genetic inheritance is represented by the blue, vertical arrows.

Naturally, the ability to reconstruct any aspect of a proto-language depends on the availability of the material in accessible sources. This means, for example, that a proto-language may have contained words which, for some reason, have been lost in all its descendants that are available to us (i.e. either living descendants or documented dead descendants). Such words can never be reconstructed, as there is simply no evidence of them ever having existed.

A more in-depth description of the comparative method and its application can be found in Campbell (2013), chapter 5. Sound change is treated at length in Campbell (2013), chapter 2.

[1] There are exceptions to the regularity of sound change – for examples, see Campbell (2013, chapter 4).

[2] Reconstructed proto-forms are preceded by an asterisk (*).

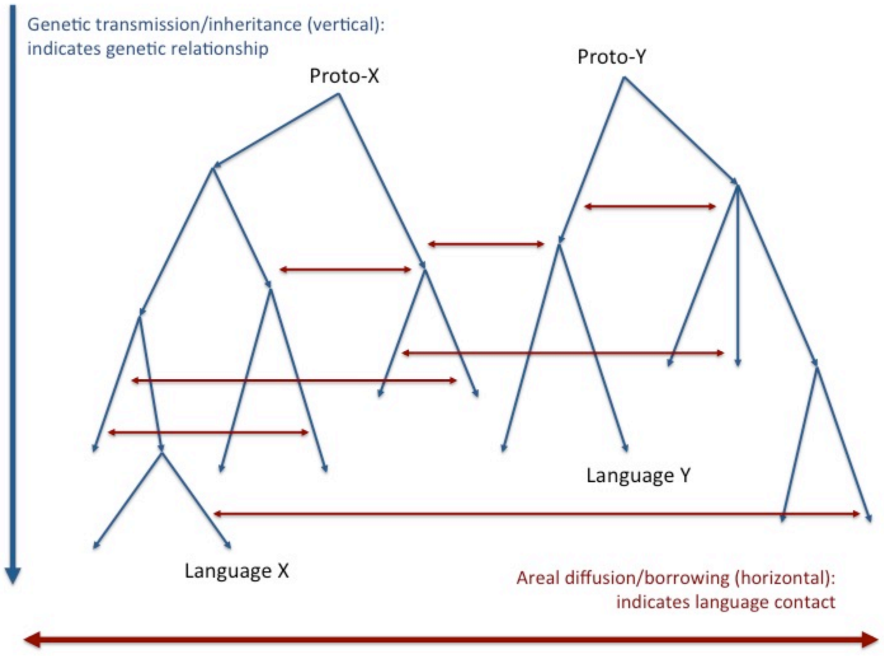
Genetic transmission/inheritance (vertical):
indicates genetic relationship
Proto-X
Proto-Y
Language Y
Language X
Areal diffusion/borrowing (horizontal):
indicates language contact

Figure 2. Illustration of inheritance vs. borrowing.

2.2.2 Loanword studies

A *loanword* is here defined as a lexical item that has been transferred from one language variety into another (Haspelmath 2009:36-38, Haspelmath & Tadmor 2009b:13) by means of contact instead of inheritance. Usually, borrowing entails a certain amount of bilingualism. In Figure 2, borrowing is illustrated with red, horizontal arrows. In contrast to genetic inheritance, which leads to *divergence*, borrowing leads to *convergence*. As Campbell (2013:68) notes, "virtually any aspect of language can be borrowed" (e.g. phonological, morphological or syntactic features), but with this definition, we narrow the object of study down to *lexemes* (i.e. words), thus excluding all other types of borrowing. The definition also excludes so called *calques* or *loan translations* (cf. e.g. Fi. *rautatie* 'railroad' and Swe. *järnväg* 'id.', both lit. 'road of iron' (Campbell 2013:71)), since these do not involve the transfer of any lexical material – only meaning or function. Words that are derived from borrowed material also do not constitute loanwords in this sense, as the derivations are separate words that have been formed in the language in question.

The loanword originates in a *donor language* and ends up in a *recipient language*. These are the two roles necessarily involved in any *borrowing event*. The donor language, therefore, does not automatically equal the earliest traceable source of a certain etymology, but the language that (in this case) Finnish has been in direct contact with.

Languages normally borrow words either out of *need* or out of *prestige*. When a new concept or item is acquired by contact with another group, the need for a word to go along with it arises, and often the word is borrowed along with the concept, which is why many languages have similar words for e.g. *coffee* and *tobacco* (Campbell 2013:58). In other cases, the donor language may be associated with a higher status, which can result in borrowing despite the lack of a "need" for it. Hock & Joseph (1996:274) outline three different types of relative social status of the participants in a borrowing event: *adstratum, superstratum* and *substratum*. Languages of roughly equal social status that come into contact with one another are referred to as adstrata, whereas a more socially imbalanced contact relationship contains a superstratum and a substratum (a high prestige and a low prestige language, respectively). Adstratal relationships are, according to Hock & Joseph (1996:274), the one's most likely to give rise to borrowing of "everyday-life vocabulary, even basic vocabulary". When a superstratum serves as the donor language, the loanwords tend to belong to the more prestigious domains of the lexicon, and their connotations tend to be equally highly esteemed – a famous example being the Norman French loanwords for animal meat borrowed into Middle English, which to this day exist in parallel with the inherited words for the animals themselves (e.g. *mutton*, *poultry* and *pork* vs. *sheep*, *hen* and *pig*; cf. Epps 2014:585). Borrowing from a substratum, however, is usually limited to need borrowings, often with derogatory connotations.

Looking at it from a different angle, prestige borrowings almost always imply an imbalanced relationship between the donor and the recipient language, where a superstratum serves as the donor language. Need borrowings are less uniform in this respect, since they merely imply that the speakers of the recipient language are becoming familiar with a new concept of some kind, and can thus involve both an adstratal or a super- vs. substratal relationship. Epps (2014:580) points out that "the source of the loan is likely to represent the source of the concept", and that "where loans have replaced pre-existing terms, they are likely to indicate the social importance of the corresponding concept in the interaction".

When words are borrowed, they often undergo *adaptation*, i.e. substitution of non-native phonemes to fit the recipient language's sound structure. Loanwords can also undergo *accommodation*, where phonological patterns are modified according to the phonological rules of the recipient language. Both of these processes are frequently represented in Finnish loanwords, as can be seen in Fi. *peti* 'bed' < Swe. *bädd* 'id.', where the foreign sounds *b* and *d* have been adapted to the native *p* and *t* (Häkkinen 2013), and in Fi. *ruuvi* 'screw' < Swe.

skruv 'id.', where the initial consonant cluster formerly unpermitted in Finnish has been simplified into a single consonant, and thus accommodated into the native phonological structure (Campbell 2013:60). Often, typical substitution patterns like these can be found, but they should not be confused with the regularity of sound change in inherited words. A few of the factors that may have an effect on the outcome of the substitution are the location in time of the borrowing event (due to the changing nature of languages' phonology) and the extent to which the speakers of the recipient language are familiar with the donor language (Campbell 2013:59-61).

In order to establish reliable loanword etymologies, it is important that we are able to determine which language is the donor and which is the recipient. This information is also important for understanding the social relations between the language groups. This can usually be assessed on the basis of clues relating to morphological complexity, phonological information, cognates in related languages, or geographical, ecological and cultural clues (Campbell 2013:61-66). However, sometimes the directionality can be difficult to establish, particularly in the case of *Wanderwörter* ('wandering words'). Campbell & Mixco (2007:220) define Wanderwörter as "borrowed word[s] diffused across numerous language[s], usually with a wide geographical distribution", and state that "typically it is impossible to determine the original donor language from which the loanword in other languages originated". If we are dealing with a Wanderwort, as Epps (2014:586) notes, the immediate contact situations between languages can be obscured, since "a loan may be passed along several languages via a borrowing chain, and therefore cannot be taken as evidence of direct contact among all the groups concerned."

Loanwords can be dated using various methods. First of all, there may be records of early written attestations of a lexical item, which can help us set at least a lower boundary for the borrowing. If this information is not available, we can check related languages for clues that help us narrow down the time of borrowing to a certain language stage. If the loanword is found in a number of related languages, following regular patterns of change (and if no other information is available), the most plausible proposal is that the borrowing took place in the most recent common ancestral state of these languages, rather than having been borrowed into each language individually (which, on the other hand, would be the most plausible explanation if the words did *not* follow regular patterns of change). By locating the loanword to a certain language stage, one can also obtain an approximate time period for the borrowing – i.e. the period during which this particular language stage is believed to have been spoken.

Some general assumptions have been made (on more or less firm evidence) regarding the likelihood for different groups of words to get borrowed. First of all, there is the idea that "basic" vocabulary, i.e. concepts common to all human cultures (e.g. body parts), is more often inherited than "cultural" (or culture-specific) vocabulary (e.g. agricultural terminology), which is more likely to be borrowed (e.g. Swadesh 1950, Haspelmath & Tadmor 2009a, Hock & Joseph 1996:257-258). This fits in well with the two main borrowing reasons mentioned above (i.e. *need* and *prestige*), which both seem more geared towards cultural than basic vocabulary, and the idea has had a far-reaching influence on the choice of e.g. methodology and data within both language classification and loanword or contact studies – sometimes perhaps applied too incautiously. The most evident example of this is the *Swadesh list*, a set of supposedly "stable" lexical items which is still widely in use today, despite the fact that it does not rest on a solid ground of linguistic research. Swadesh lists are dealt with in more detail under 3.1 below. Another general assumption concerns the frequency with which a word occurs in a language. As Haspelmath & Tadmor (2009b:15) point out, "it is generally assumed that lexical stability increases (and therefore borrowability decreases) with frequency." Claims have also been made about varying borrowability depending on word class, e.g. that verbs should be more difficult to borrow than nouns due to an increased requirement of grammatical adaptation (Haspelmath 2009:35).

The "inherited vs. borrowed" dichotomy is an important basic distinction in the lexicon of a language, but unfortunately reality does not always present us with cases that can be neatly divided into either of these groups. For instance, there are examples of words belonging to categories traditionally thought of as the most "stable" ones (i.e. – allegedly – immune to borrowing) which can be shown to have been borrowed, e.g. Fi. *vesi* 'water' < PU **wete* 'id.' < PIE **wed*- 'id.' (denoting a very basic concept which must have been known to the speakers of Proto-Uralic, or to any human population for that matter) (Häkkinen 2013). Also, there are plenty of linguistic phenomena that may obscure etymologies or mislead us in our analyses. One such thing is *sound symbolism*, defined by Campbell & Mixco (2007:187) as "[a] direct association in a language between sounds and meaning, where the meaning typically involves the semantic traits of 'size' or 'shape'". Sound symbolism can "interfere" with the regularity of sound change by driving sounds in certain directions to represent the semantics of the word – thus obscuring the lineage. Another important notion is that of *analogy*, i.e. change caused by a relation of similarity of some kind (Campbell 2013:91-92), such as a word's phonological shape being changed under the influence of a somehow reminiscent word.

Analogy is sometimes described as “internal borrowing”. Finally, in populations where bilingualism in two genetically related languages is widespread, we may be dealing with the concept of *etymological nativisation* (Hock & Joseph 1996:261-262, Aikio 2007), which means that loanwords can be “impersonating” native words due to the speakers’ knowledge and intuition about regular sound correspondences, which can have the effect that these sound correspondences (partially or entirely) get applied to words that are borrowed between the languages. An example cited in Aikio (1997:28) is Northern Saami *haddi* ‘price’ < Fi. *hinta* ‘id.’, where both Finnish vowels are replaced by their regular Northern Saami reflexes (*a* < *i* (stressed first syllable) and *i* < *a* (unstressed second-syllable), respectively).

Another important concept is that of *semantic change*, the well-attested phenomenon of linguistic forms taking on a slightly different meaning. For example, English *deer* descends from Old English **dēor*, which used to be the general term for ‘animal’ (cf. Swedish *djur*, German *Tier*) (Campbell 2013:224). We know this from documentation and from cognates in related languages, but for some reconstructed terms it may be harder to establish an unambiguous semantic referent. The research on semantic change has not yet led to any overarching models of how this mechanism operates (Campbell 2013:232ff., Epps 2014), but the cases can be divided into different types, e.g. widening or narrowing (of the meaning), metaphor, metonymy, etc. (Campbell 2013, chapter 9). A related topic is that of *taboo replacement*, which involves the lexical substitution of words that are considered taboo or obscene with more neutral terms (Campbell 2013:229-230). This often affects words with multiple meanings, where some are obscene, e.g. English *ass* and *cock* (when denoting animals) having largely been replaced by *donkey* and *rooster*, respectively (Campbell 2013:230), but it has also frequently been applied to names of predator animals, due to the superstitious fear of “summoning” them by pronouncing their actual names – a well-known example being Russian *medved’*, lit. ‘honey-eater’, which has replaced the inherited Indo-European word for ‘bear’ (Epps 2014:585).

Finally, it has to be pointed out that negative evidence, i.e. the lack of a loan etymology for a word, does not mean that we can prove that it has not been borrowed. Indeed, we can only prove borrowing – not “non-borrowing”. As Haspelmath & Tadmor (2009b:13) write, “any word could have been borrowed at some prehistoric time, so we can never be sure what is *not* an old loanword”. Words can also have been borrowed at an earlier language stage, and then be inherited in the descendants of that ancestor. We return to this topic in Borrowed and Created on loan basis under 3.2.3 below.

2.2.3 Linguistic palaeontology

Linguistic palaeontology (or *linguistic prehistory, linguistic archaeology*, etc.) is a sub-discipline that "uses historical linguistic findings for cultural and historical inferences" (Campbell 2013: 405). For instance, the reconstructed vocabulary of a proto-language – when correlated with loanword studies, linguistic homeland and migration theory, language classification, etc. – may be able to provide us with a glimpse into various aspects of the culture of its speakers.

A closely related topic is *Wörter und Sachen* ('words and things'), a 19th-20th century movement that builds on the idea of a close relationship between words and their referents, and has to do with the historical cultural inferences we can make from studying the lexicon (Campbell 2013:434-435, Epps 2014:580ff.). One of the assumptions of this method is that analysable words tend to be more recent than unanalysable ones (Campbell 2013:434-435). This also entails that the referents behind unanalysable terms have long been known to the speakers, while those behind analysable terms have been acquired more recently. Another assumption holds that words inherited from a proto-language represent meanings associated with a certain cultural salience (Epps 2014:580). However, as (Epps 2014:580-581) states, "while inheritability implies salience, the converse is not necessarily true; words for which cultural relevance is linked to interaction with other groups may be particularly prone to borrowing, and terms associated with taboo topics tend to undergo rapid replacement".

Linguistic palaeontology is often applied in attempts to locate the *Urheimat*, or *homeland*, of a language family, sometimes by matching the proto-language's reconstructible words for flora and fauna to a geographical location. A well-known example is the role tree names have played in the various attempt to locate the Proto-Indo-European homeland (cf. e.g. Campbell 2013:432). While this kind of information can certainly be of use, it is important to remember that a reconstructed lexical item does not only consist of a phonological shape, but also a semantic meaning. The semantics is often more difficult to reconstruct, as it does not follow the same kind of regular patterns as sound change (Epps 2014:583). In other words, if we cannot know for sure what meaning a reconstructed word had in the proto-language, it does not help us much in the location of the homeland – something that has become evident also in the case of the Indo-European tree names mentioned above. The meaning of a reconstructed item can be further obscured by mechanisms such as *semantic extension* and *markedness reversal* (Epps 2014:583). Both of these terms refer to the process of lexical items taking on a new, secondary meaning (while keeping the original one) – and in the case of markedness reversal, what originated as a secondary meaning eventually shifts to the primary one.

Kallio (2002:34) comments that “linguistic paleontology seems to be more able to date than to locate proto-languages. Even then, one should rather concentrate on semantic categories instead of semantic units”. Epps (2014:586) makes a similar claim about the increased reliability of whole semantic domains over single lexical items.

Another homeland location method is *linguistic migration theory* (or *center of gravity model*), a method which looks at the language classification within the family and relies on a “model of maximum diversity and minimal moves” (Campbell 2013:432-433). The hypothesised homeland is thus placed in the area of greatest linguistic (in-family) diversity, i.e. the area which contains the highest number of primary family tree branches (or nodes). The technique generally produces fairly good results, and Campbell (2013:431) states that “migration theory has a stronger probability of being correct than any random guess we might make which is not based on these principles”. There are, however, reasons to be cautious, as it is easy to imagine situations where linguistic migration theory would be of little use: e.g. if the speakers for some reason have been driven away or drawn to another place, or if the original area’s population and languages have been lost without a trace.

2.3 The Uralic language family

The Uralic language family consists of some 30-40 languages (slightly more than the average language family), is relatively small by number of speakers, but is one of the largest ones when it comes to geographical extension (Janhunen 2009:59). Figure 3 shows focal points representing the current geographical locations of the Uralic languages, colour coded according to subgroup (map from Hammarström, Bank, Forkel & Haspelmath (2018)). According to Salminen (2002), the Uralic family has “a number of nodes representing branches that are so transparent, closely-knit and well-established that they can be immediately and beyond doubt recognised as historical linguistic entities, each deriving from a highly distinct proto-language”.

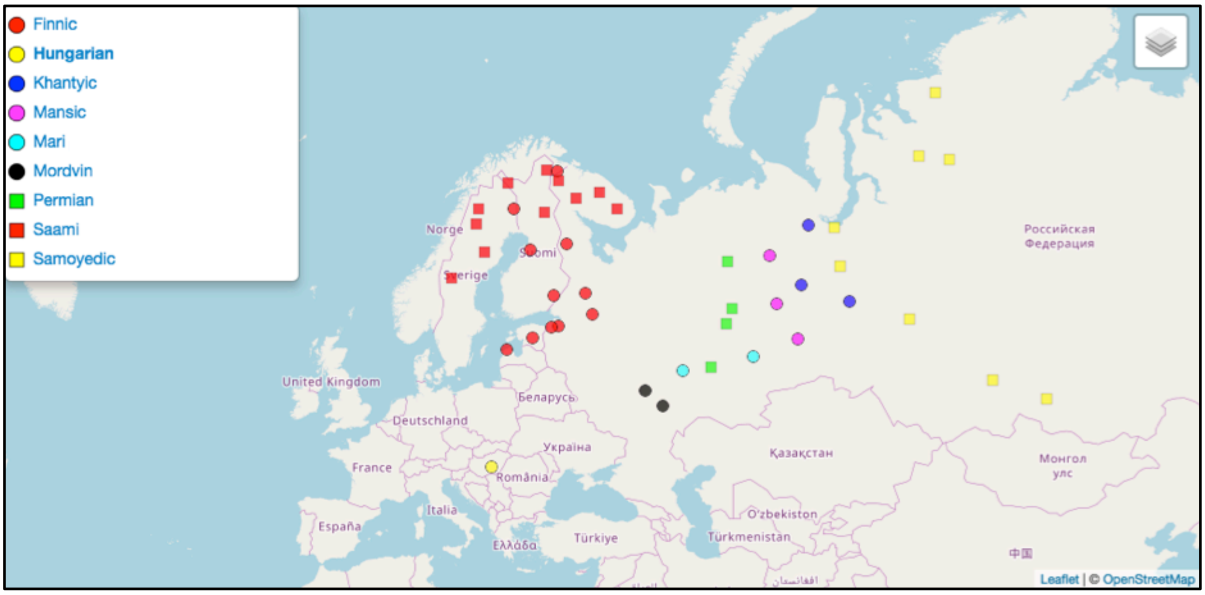

Figure 3. Map showing the spread of the Uralic languages.

The nine Uralic branches and a sample of their member languages are listed below:

Saami	South, Ume, Pite, Lule, North, Inari, Skolt, Kildin, Ter
Finnic	Finnish, Estonian, Livonian, Votic, Karelian, Ingrian, Ludic, Veps
Mordvin	Erzya, Moksha
Mari	Mari
Permic	Komi, Udmurt
Hungarian	Hungarian
Mansi	Mansi
Khanty	Khanty
Samoyed	Nganasan, Enets, Nenets, Selkup

2.3.1 Uralic family tree & homeland hypotheses

It is considered uncontroversial that the languages listed above should derive from a common ancestor (Proto-Uralic); however, the relationships between the branches are not entirely clear (Laakso 2001:203). Jaakko Häkkinen (2009) identifies three different stages reflecting in the emergence of different theories about the Uralic family tree: *the heyday* (end of the 19th and most of the 20th century), *the cutting stage* (the 1980's and -90's) and *the recovery stage* (21st century).

In 1879, the Finnish linguist Otto Donner was the first scholar to propose what we might call a "traditional" Uralic language tree[5] (Häkkinen 2009), representing Jaakko Häkkinen's 1st stage. Figure 4 shows a representation of this traditional family tree (adapted from Campbell 2013:178), where the nodes marked with thicker frames represent the undisputed groupings

[5] However, the Samoyed branch was not included, as it was not yet recognised as belonging to the Uralic family.

we saw in section 2.3. The basic structure of this family tree remained the standard model for more than a century. Most versions of this tree are characterised by their branching being almost exclusively binary.

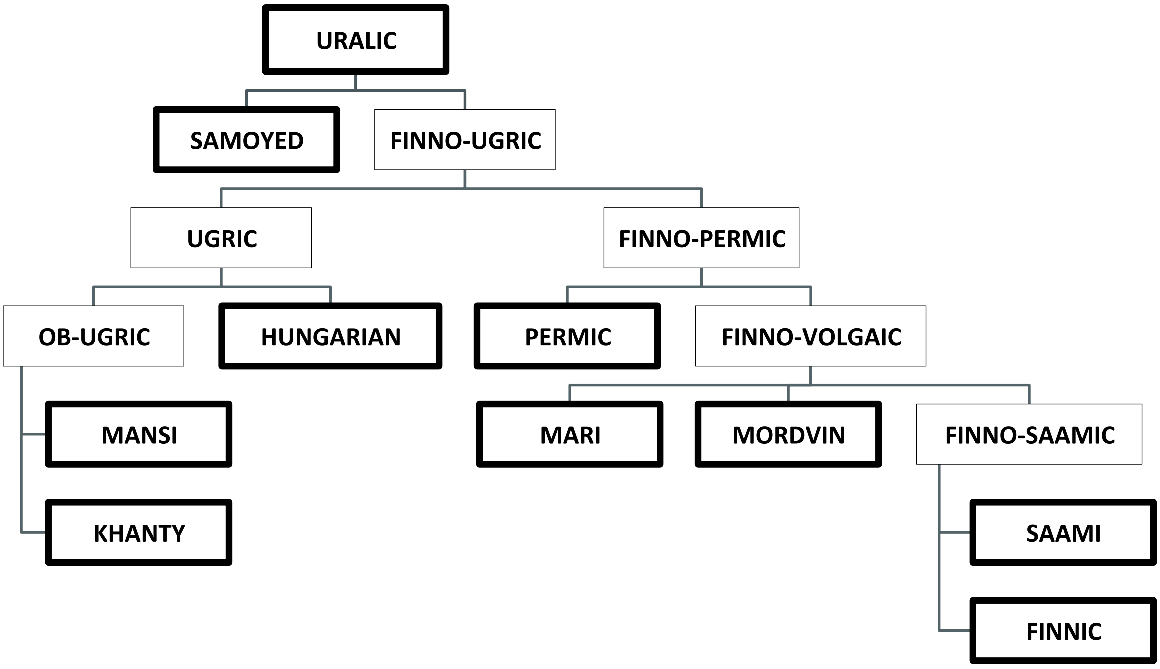

Figure 4. A "traditional" Uralic family tree.

The first scholar to represent Jaakko Häkkinen's 2nd stage was Kaisa Häkkinen (1984), who proposed her so called "family bush" (the name referring to its non-tree-like structure), with Ugric as the only intermediate node between Uralic and the 9 individual branches. Salminen 2002) notes that

> "it must be carefully examined whether all of the traditionally assumed proto-languages qualify as distinct genetic units, or whether they are either based on very few diagnostic features that do not make them notably different from their parent languages, or whether the features attributed to them are actually better explained by areal influences".

Salminen (e.g. 1999, 2001, 2002) also argues for the abandonment of higher nodes in the Uralic tree, i.e. the groupings of the nine undisputed Uralic branches. His suggested family tree is illustrated in Figure 5. He is of the view that some of these traditional groupings, e.g. *Ugric* (Hungarian, Khanty & Mansi), *Finno-Volgaic* (Finnic, Saami, Mari & Mordvin) and even *Finno-Ugric* (all branches except Samoyed), are better explained as so called *areal genetic units*, and thus argues that a tree "involving uncontroversial branches only, reflects the structure of Uralic more accurately, especially when supplemented with information on the areal contacts between the branches" (Salminen 2002). Such a situation could also be argued to be supported by Aikio's (2007) research on the *etymological nativisation* (see 2.2.2 above) of loanwords between Finnish and Saami, which, of course, blurs the line between shared vocabulary and intra-family loanwords, in turn making it harder to render an accurate family tree.

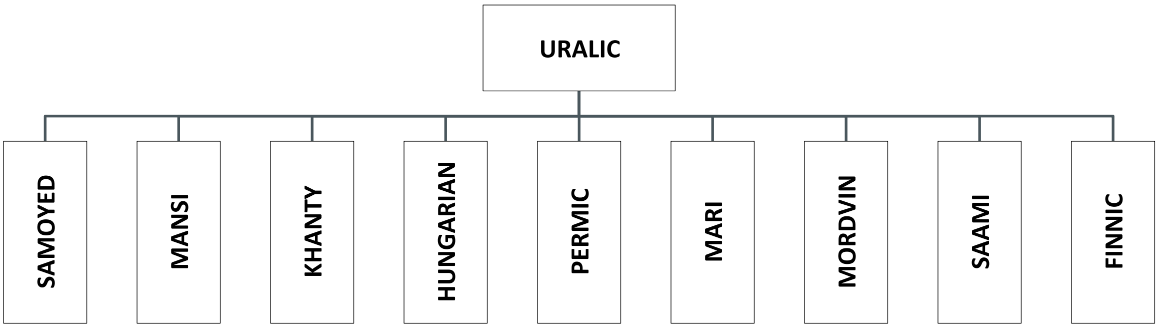

Figure 5. A Uralic family tree without intermediate nodes.

Salminen's hypothesis has been criticised, e.g. by Janhunen (2009:67), among other things for assuming that the small amount of lexical material that the Samoyed languages share with the rest of the Uralic branches is due to "relexification" in Samoyedic (a process that, according to Janhunen, is always accompanied by considerable grammatical restructuring or simplification – which the Samoyedic languages apparently have not undergone), rather than Samoyed/Finno-Ugric representing the primary split of the Uralic tree.

Jaakko Häkkinen himself is an advocate of what he considers the 3rd stage, where the Uralic tree once again contains intermediate nodes, but this time with the important difference of a "Samoyed-Ugric" branch. Kallio, in his more recent works (2012, 2015a), also appears to be in support of a similar model.

Note that while Jaakko Häkkinen's 3-stage model accurately describes the consecutive *emergence* of the different generalised family tree theories, the onset for a new stage does not necessarily imply the immediate decline of another. Indeed, to some extent, all three theories (or some variant thereof) have their advocates to this day.

Regarding the Finnic and post-Finnic developments, Laakso (2001:204-207) describes the Finnic languages as a dialect continuum rather than a uniform proto-language with neatly divided groups. Variation seems to have spread between the dialects, creating various overlapping isoglosses which cannot straightforwardly be explained with inheritance only, or do not constitute large enough sets of sound changes. Contemporary Standard Finnish (which is the object of study here) can also be described as a 19th century construction, born out of a rather "artificial" admixture of Western and Eastern Finnish dialects.

In sum, capturing the structure of the Uralic language family in an undisputed tree shape is a difficult task. For the study at hand, a model had to be chosen to serve as the basis for the *Age* coding of the lexemes (more in the subsection Age under 3.2.3 below). Kaisa Häkkinen (2001:171) mentions that the traditional Uralic family tree is indirectly implied in the reconstructions of some etymological dictionaries (e.g. *Uralisches etymologisches Wörterbuch*, Rédei (1986-1991); henceforth *UEW*), despite the fact that the tree model is far from universally agreed upon. However, she also argues that this presupposition does not cause any serious problems from the perspective of the Finnic branch, as the languages that (according to the traditional tree model) are classified as being distantly related to Finnic also happen to be geographically distant, which significantly reduces the risk of subsequent contamination by means of language contact[6]. Figure 6 contains a map showing the current geographical locations of the Uralic languages, with rings demonstrating (roughly) how, accepting a "traditional" Uralic family tree, the genetic and geographical distance from the Finnic branch is correlated, increasing with the numbers: 1. Finnic, 2. Finno-Saamic, 3. Finno-Volgaic, 4. Finno-Permic, 5. Finno-Ugric, 6. Uralic (map from Glottolog).

Thus, for a study from a Finnish perspective, the available materials on Uralic etymologies can be employed irrespective of their authors' views on Uralic subbranching. Since the traditional family tree serves as the basis for most etymological dictionaries, and since I will be relying on these existing etymological dictionaries for my data, accepting it provides the most reliable way of quantifying the data with respect to age.

[6] It is important to point out, however, that this is a lucky coincidence and not the case for all languages. Häkkinen gives the example of Komi and Mansi, which are as distantly related as Finnish and Mansi, but spoken in the same area. In such cases, "the genetic model is of dubious validity" (Häkkinen 2001:171), as it can be difficult to determine what lexical material is due to genetic inheritance and areal diffusion, respectively. Words present in both Komi and Mansi have, according to Häkkinen possibly incorrectly, been attributed to Proto-Finno-Ugric, where they could perhaps have been explained by later instances of contact.

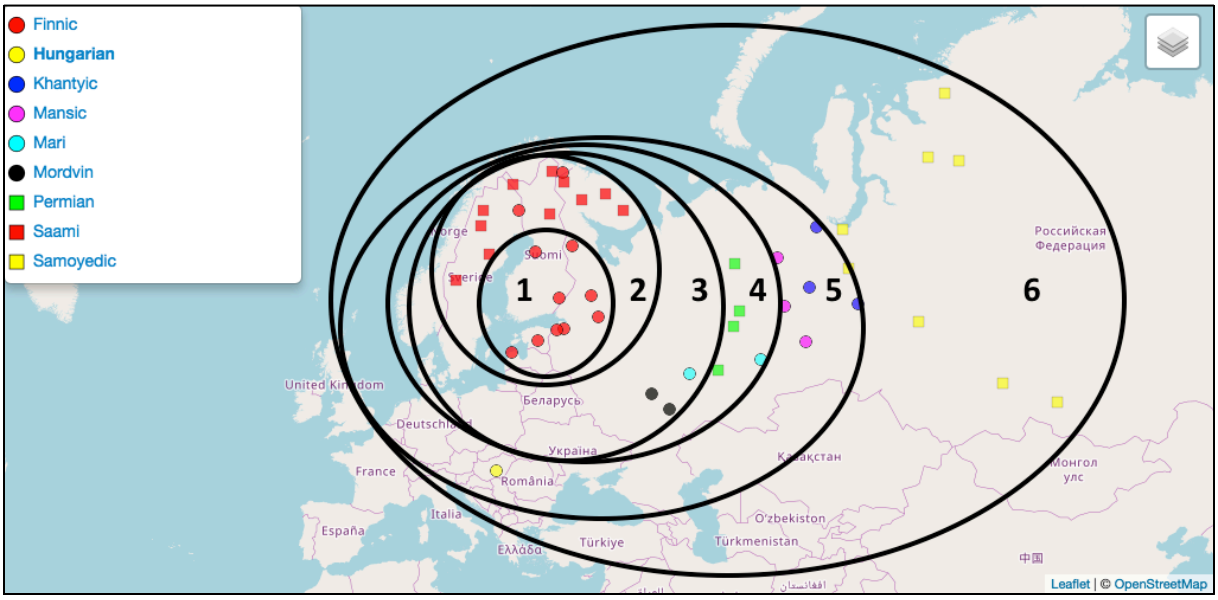

Figure 6. Correlations between the genetic and geographical distance from the Finnic branch.

Another question which is closely linked to the theories on the branching of the Uralic family tree is the dating of Proto-Uralic and the location of its homeland. This is still the subject of some debate. As for the dating, most scholars agree that the primary split of Proto-Uralic must have happened around 4000-3500 BC (e.g. Décsy 1965, Hajdú 1975, Korhonen 1981, Carpelan & Parpola 2001, Laakso 2011) – although absolute dating of any language stages will not be explored further in this study. As for the homeland location, Campbell (2013:428) notes that the various homeland suggestions differ from each other in size more than in location[7], and summarises the most popular hypotheses in the six groups listed below. The approximate locations of the suggested homelands are plotted out in Figure 7, with the corresponding numbers (map from Google Maps).

1. The region of the middle course of the Volga River and its tributaries
2. The region of the northern Urals on both sides of the mountains
3. The central and southern Urals on both sides
4. Rather eastward on the Asian side of the Urals
5. Rather westward on the European side
6. The broad area between the Urals and the Baltic Sea

Most scholars would thus seem to agree on the fact that the Uralic homeland is located farther east than the areas where Finnic languages are spoken today, implying a westward spread of this language branch.

[7] It should be mentioned that homeland locations wildly divergent from these have also been suggested in the past – however, these have very few supporters today.

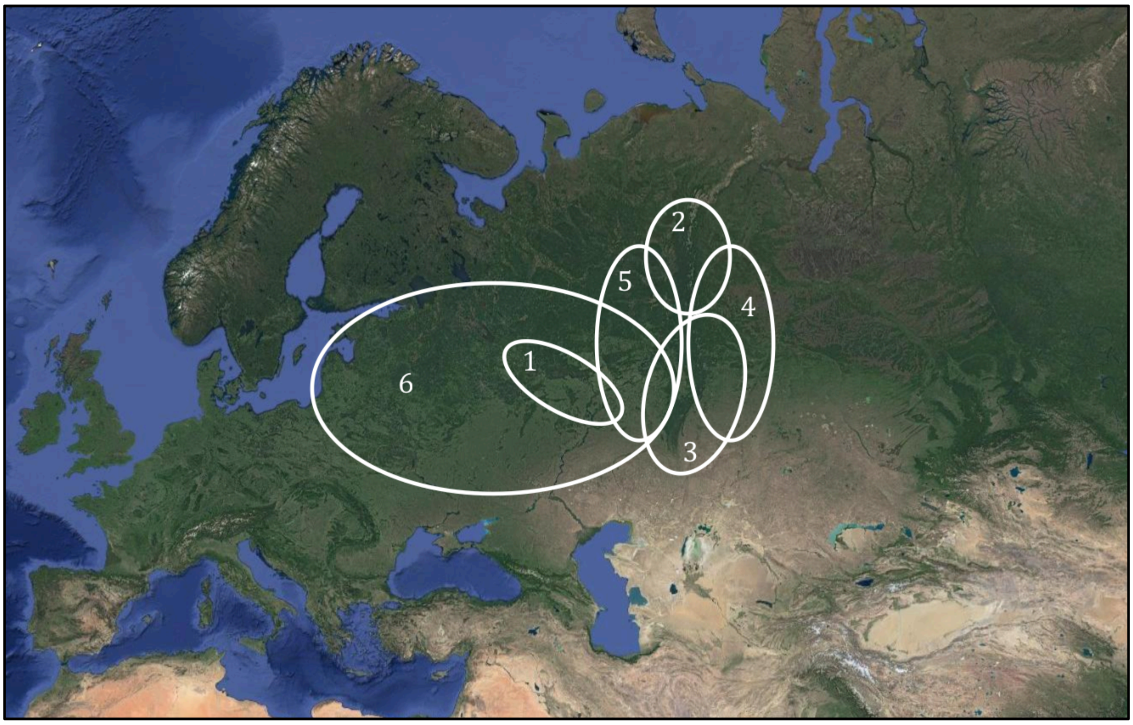

Figure 7. Map of hypothesised homeland locations for Proto-Uralic.

Many see the existence of Proto-Indo-European loanwords in Proto-Uralic as evidence for locating the two homelands close to each other, usually somewhere north of the Black Sea (e.g. Kallio 1997, Häkkinen 1998). Those who do not support this hypothesis argue that the proposed Proto-Indo-European loanwords are instead borrowed from some early Indo-Iranian language, and see evidence for locating the Proto-Uralic homeland elsewhere, e.g. Janhunen (2009), who is in favour of an Eastern homeland location in the Minusinsk basin and a gradually westward expansion reflected by a binarily split tree.

Campbell (2013:428) concludes the homeland discussions as follows:

> "In any event, most scholars assume that the relative homogeneity of the family was broken up by the introduction of Neolithic techniques and agriculture from areas south of the Proto-Uralic and Proto-Finno-Ugric homeland, and that the onset of farming and cattle herding – factors contributing to sedentarism – probably contributed to diversification of the family."

All three aspects of linguistic prehistory brought up under this section are tightly interconnected: the relative chronology of the branching, the dating of the proto-language and the location of the homeland. The branching theories build on e.g. linguistic reconstruction and loanword studies, the dating theories build on e.g. theories on lexical replacement, and the homeland location theories build on e.g. linguistic migration theory and linguistic palaeontology. Findings from other disciplines, such as archaeology, are also often taken into account. While the combination of results from different disciplines can be a powerful resource, Epps (2014:592) reminds us about the importance of the independence of each discipline's contributions: "[i]f historical linguists take archaeological findings into account in interpreting linguistic data, for example, then these interpretations cannot be argued later to corroborate the archaeological evidence".

It is also crucial not to take any theory as the truth. As Epps (2014:579) also notes, "our inferences about the past are only as good as our reconstructions, which are necessarily hypotheses."

2.3.2 Previous loanword studies in Uralic languages

The relationship between the Uralic and Indo-European language families has been the subject of countless studies. There is more or less universal agreement on the fact that these speech communities have a lot in common. Some have postulated a theory of deep genetic relationship between the families (often called the *Nostratic* language family; cf. Campbell 2013:360-361), but this theory has rather limited support. Instead, most scholars believe that the similarities are due to early, long-standing and intensive language contact. Most agree that there are very old Indo-European loanwords in Uralic languages. Some argue that some of these may be as old as to come from PIE itself, whereas others postulate the oldest Indo-European loanword layer to come from some early Indo-Iranian language variety. Kallio (2002:35) sees no reason to consider these early loanwords as anything else than PIE proper unless they have some distinctively Indo-Iranian feature, and also points out that the advocates for this theory often happen to belong to those who favour an Indo-European homeland far away from the Uralic speech area. In any case, despite the discrepancies in the Proto-Uralic homeland theories of different scholars presented in 2.3.1 above, the majority of scholars seem to believe that Proto-Uralic (or Proto-Finno-Ugric, depending on the classification they favour) was spoken in the vicinity of Proto-Indo-European (Campbell 2013:430). Studies on the Uralic language family – especially those on Indo-European loanwords – have also had a great influence in Indo-European linguistics, e.g. on theories about the Proto-Indo-European homeland, as well as on the reconstruction of e.g. Proto-Germanic (cf. e.g. Campbell 2013:66, Mallory & Adams 2006:81-82, Mallory & Adams 1997:290ff).

Kaisa Häkkinen (1998, 2001) has conducted a well-known and influential study about the Uralic lexicon, making a rough semantic division of the reconstructed vocabulary for the earliest Uralic language stages (i.e. Proto-Uralic and Proto-Finno-Ugric) in *UEW*[8]. Reconstructed items have only been included based on stable criteria regarding their certainty (see Häkkinen 2001:172-173).

All identified reconstructed lexemes for these two strata were included, i.e. no distinction was made between native Uralic words and (possible or certain) loanwords, as the purpose of the study was merely to investigate the kind of culture that could be associated with the early speakers of Uralic – not the contact with other linguistic groups (Häkkinen 1998:190). The counts for the identified semantic domains are given in Table 2 (Häkkinen 2001:173-174): the number of lexemes reconstructible for Uralic (U), the number of lexemes reconstructible for Finno-Ugric (FU), and the total number of reconstructible lexemes for both stages.

The study has been particularly influential when it comes to theories about the subsistence system of the early Uralic speakers, since it points to a hunter-gatherer rather than a agriculturalist or pastoralist lifestyle. Häkkinen (2001:169) states that

> "[…] there is no Uralic or Finno-Ugric vocabulary whatsoever which unambiguously refers to the cultivation of crops, and only a few lexical items which putatively refer to the keeping of domestic animals. The terms referring to agriculture typically have a narrow regional distribution, with cognates traceable only in those related languages spoken in geographically adjacent areas (Häkkinen & Lempiäinen 1996). In strikingly many cases, the agricultural vocabulary can be shown to consist of loanwords. The linguistic ancestors of Finnish, for example, appear to have begun to practise agriculture in the region surrounding the Baltic Sea, and to have acquired this activity through the mediation of their Indo-European-speaking neighbours."

Other underrepresented categories include words for social organisation. Given the rather ample category of kinship terminology, Häkkinen (1998:192) hypothesises that family relations were the main basis for the societal structure.

Among the overrepresented categories we find e.g. parts of the body, names of animals, work- and activity-related vocabulary, as well as a surprisingly large amount of words for quality (Häkkinen 1998:191).

[8] It should be mentioned that not everybody agrees with these reconstructions. A similar analysis of another scholar's list of reconstructible items would thus, in all likelihood, yield a different result. To the best of my knowledge, no such study has been made to date.

Table 2. The results of Häkkinen's study on semantic domains in the reconstructible Uralic lexicon.

Domain	**Total**	**U**	**FU**
time	12	4	8
sensations	14	4	10
fauna			
a) names of animals	60	24	36
b) terminology relating to animals	15	7	8
human society	2	1	1
flora			
a) names of plants	27	13	14
b) parts of plants, etc.	22	10	12
trade	2	1	1
transport, traffic, motion	32	17	15
quality	49	13	36
quantity, measurement, value	13	1	12
nature			
a) land and landscape	27	10	17
b) water and water systems	10	5	5
c) materials, surface	19	8	11
d) atmosphere, sky	14	3	11
hunting & fishing	18	9	9
form, posture	8	2	6
pronouns	16	11	5
processes and states			
a) life & health	13	5	8
b) emotions & perceptions	5	1	4
c) miscellaneous states & changes	27	6	21
buildings, constructions, equipment	14	6	8
construction processes, materials, pieces	10	5	5
nourishment			
a) eating & drinking	8	5	3
b) foodstuffs	10	2	8
c) dishes, preparation of food	12	4	8
the body			
a) parts of the body	77	35	42
b) bodily functions, etc.	12	5	7
speech, thought	6	1	5
family & personal relationships	27	16	11
relations in space & time	21	10	11
activities & processes	60	12	48
fire, the handling of fire	3	2	1
work, tools, working materials	59	22	37
religion, beliefs	7	-	7
clothing	7	2	5
miscellaneous other items	5	2	3
Total	**743**	**284**	**459**

Häkkinen (2001:169) finds that the study "yields abundant evidence from the earliest lexical strata of hunting cultures, e.g. terms for hunting and fishing equipment and for game animals". While this claim about evidence for a hunting culture is entirely legitimate, and while Häkkinen has indeed focused on semantic domains rather than single lexical items (as per the discussion under 2.2.3 above), it should be pointed out that the claims about agriculture rest heavily on the *lack* of reconstructible vocabulary, i.e. on *negative evidence*. As Epps (2014:584) states, “a crucial caveat in using historical linguistics to draw inferences about the past is that our inability to reconstruct a word to the proto-language does not entail its absence in that language, or the absence of its referent in the lives of its speakers” (Epps 2014:584). In other words, what Häkkinen found in her study is that no agricultural lexicon whatsoever is (to the current knowledge) reconstructible for Proto-Uralic or Proto-Finno-Ugric, whereas reconstructed terms for hunting and fishing are abundant. Based on this, a proposed hunter-gatherer lifestyle for the early Uralic speakers is clearly a valid hypothesis (as the study certainly does not point to any other subsistence system), but it is necessary to remember that this is a hypothesis and nothing else.

A final important takeaway from Häkkinen’s study has to do with loanwords and the relationship to Indo-European: a fourth of the most certain etymologies have been explained as Indo-European loanwords (Häkkinen 1998:193).

Koivulehto (e.g. 1976, 2006, 2007, etc.) has provided Indo-European loan etymologies for many previously unexplained words. Kallio (e.g. 1995, 1998, 1999, 2002, 2006, 2008, 2012, 2015a, 2015b, 2017, etc.) has written a number of articles on e.g. the layers of loanwords coming from different Indo-European sources.

A lot has also been written about the phonological adaptation of loanwords into Finnish. Shifts in sound substitution patterns (reflecting changes in the phonological profile of Finnish) give us clues about (roughly) when a particular borrowing event may have taken place. For example, Finnish used to have a constraint against initial consonant clusters (Suomi et al. 2008:55), which forced the Swedish loanword *strand* ‘shore’ to be integrated as *ranta*. As this constraint is no longer in force (cf. the more recent loanwords *krokotiili* ‘crocodile’ and *presidentti* ‘president’, both from Swedish (Campbell 2013:60, Häkkinen 2013)), we can temporally locate the borrowing event to a period when it still was. Another example of a shift in sound substitution can be seen in loanwords containing a *d*: as can be seen in the *ranta* example above, *d* used to be replaced with a native *t* earlier, but as *d* has gained phonemic status (at least in some varieties, including Standard Finnish, cf. Table 1), it is normally retained in more recent loanwords: *demokratia, indeksi* (Suomi et al. 2008:34).

3 Methodology

Using the definitions outlined in 2.2.2 above, this study aims to examine Finnish as a recipient language, as well as its relationship with the donor languages.

The study combines qualitative and quantitative research. As for the actual analysis, i.e. the contribution of this thesis, quantitative methods were used. The data, however, was retrieved from etymological dictionaries, primarily *Nykysuomen etymologinen sanakirja* [Etymological dictionary of Contemporary Finnish] (Häkkinen 2013; henceforth *NES*), which means that the time-consuming qualitative work had already been done by much more competent Uralic etymologists – the data only needed to be organised into a suitable format for the quantitative analysis. As Angouri (2010:33) puts it: "while quantitative research is useful towards generalising research findings […], qualitative approaches are particularly valuable in providing in-depth, rich data", and "[w]hether combining or integrating qualitative/quantitative elements, mixed methods designs arguably contribute to a better understanding of the various phenomena under investigation".

In this chapter I will present the Loanword Typology project (whose method I have borrowed and adapted), explain how I have modified the method to suit my study, go through the different stages in the working process, as well as comment on some problems that were discovered during the process and express some criticism towards the method.

3.1 The Loanword Typology project

The *Loanword Typology project* (henceforth LWT project) and the accompanying *World Loanword Database* was a research project run by Martin Haspelmath and Uri Tadmor in 2004-2009. The aim of the LWT project was to answer questions about borrowability from a cross-linguistic perspective. This was done by making qualitative studies on a cross-section of the basic vocabulary (1,460 lexical meanings[9]) in 41 languages. They were striving for a genetically and areally balanced sample – however, this is of course difficult to achieve with such a small number of languages. In addition, some of the languages are more unusual in their nature – e.g. Seychelles Creole and Saramaccan, being creole languages, and Old High

[9] "By asking the contributors to provide the counterparts of these meanings, we aimed to obtain comparable lexical samples from all project languages. Note that the list is a 'meaning list', not a 'word list'. The items on the list are meanings that could be relevant in any language, not words of a particular language (in particular, they are not words of our working language English […])." (Haspelmath & Tadmor 2009:5).

German, being a dead language – and their place in the language sample could perhaps be questioned (or at least be argued to have been given more space than can be considered representable).

Once the quantification was completed, cross-linguistic generalisations could be made as to which lexical items actually get borrowed the most (and which do not). There is a historical background to why this is an important study. As explained in section 2.2.2 above, the idea that there is a distinction between *stable* (i.e. resistant to borrowing) and *contact-sensitive* (i.e. easily borrowed) vocabulary has existed at least since the mid-20th century, most prominently put forward by and associated with Morris Swadesh (e.g. 1950) who developed the so-called *Swadesh list*, a list of the 100 most "stable" lexemes[10] – a tool for establishing genetic linguistic relationship still widely in use today. The purpose of the Swadesh list is to separate the two categories in order to separate borrowed and inherited material, since only inherited material can tell us something about genetic affinities between languages. The only problem is that nobody ever actually *investigated* which lexical items were stable and which were contact-sensitive, and the Swadesh list is thus based on the anecdotal intuition of Morris Swadesh and on circular argumentation. The LWT project was the first research project that set out to verify statistically which lexical meanings are the most stable, and it did so by using a relatively large and balanced language sample. The project resulted in the *Leipzig-Jakarta list* (see Table 3), an alternative to the Swadesh list including the 100 least borrowed lexical meanings as shown by the generalisations of the sample (Tadmor 2009:68ff). The two lists overlap to 62% (marked in black in Table 3) – a 62% that Swadesh and his intuition should be given credit for. However, the remaining 38% that do *not* overlap, i.e. that are present in the Leipzig-Jakarta but not in the Swadesh list (marked in grey and strikethrough in Table 3), imply a quite substantial difference which "can lead to rather different lexicostatistical and other results" (Tadmor 2009:73).

[10] Swadesh created several different versions of the Swadesh list (including some with around 200 words), but the 100-item list is the one that is most widely used.

Table 3. The Leipzig-Jakarta list and its overlap with the Swadesh 100 list.

1. fire
2. nose
~~3. to go~~
4. water
5. mouth
6. tongue
7. blood
8. bone
9. 2sg pronoun
10. root
11. to come
12. breast
13. rain
14. 1sg pronoun
15. name
16. louse
~~17. wing~~
18. flesh/meat
19. arm/hand
~~20. fly~~
21. night
22. ear
23. neck
~~24. far~~
~~25. to do/make~~
~~26. house~~
27. stone/rock
~~28. bitter~~
29. to say
30. tooth
31. hair
32. big
33. one
~~34. who?~~
~~35. 3sg pronoun~~
~~36. to hit/beat~~
37. leg/foot
38. horn
39. this
40. fish
~~41. yesterday~~
42. to drink
43. black
~~44. navel~~
45. to stand
46. to bite
~~47. back~~
~~48. wind~~
49. smoke
50. what?
~~51. child (kin term)~~
52. egg
53. to give
54. new
55. to burn (intr.)
56. not
57. good
58. to know
59. knee
60. sand
~~61. to laugh~~
62. to hear
63. soil
64. leaf
65. red
66. liver
~~67. to hide~~
68. skin/hide
~~69. to suck~~
~~70. to carry~~
~~71. ant~~
~~72. heavy~~
~~73. to take~~
~~74. old~~
75. to eat
~~76. thigh~~
~~77. thick~~
78. long
~~79. to blow~~
~~80. wood~~
~~81. to run~~
~~82. to fall~~
83. eye
84. ash
85. tail
86. dog
~~87. to cry/weep~~
~~88. to tie~~
89. to see
~~90. sweet~~
~~91. rope~~
~~92. shade/shadow~~
93. bird
~~94. salt~~
95. small
~~96. wide~~
97. star
~~98. in~~
~~99. hard~~
~~100. to crush/grind~~

3.1.1 Meanings & semantic domains

As mentioned above, the LWT list consists of 1,460 lexical meanings. The list is based on the 1,310 meanings in the IDS (Intercontinental Dictionary Series; Key & Comrie 2015), which in turn is based on the roughly 1,200 meanings in Buck (1949). A few meanings were added to the LWT list for various reasons: to cover the entire Swadesh 207 list, to include some modern phenomena missing in Buck (1949) and the IDS, or to compensate for the geographical and cultural biases in these previous versions (Haspelmath & Tadmor 2009b:6). The meanings are divided into 24 semantic domains (see Table 4), of which the first 22 were retained from the IDS and Buck (1949). The two last ones – strictly speaking no "semantic" domains – were added to the original 22.

Table 4. Semantic domains of the Loanword Typology project.

	Semantic domain label	**Number of meanings**
1.	The physical world	75
2.	Kinship	85
3.	Animals	116
4.	The body	159
5.	Food and drink	81
6.	Clothing and grooming	59
7.	The house	47
8.	Agriculture and vegetation	74
9.	Basic actions and technology	78
10.	Motion	82
11.	Possession	46
12.	Spatial relations	75
13.	Quantity	38
14.	Time	57
15.	Sense perception	49
16.	Emotions and values	48
17.	Cognition	51
18.	Speech and language	41
19.	Social and political relations	36
20.	Warfare and hunting	40
21.	Law	26
22.	Religion and belief	26
23.	Modern world	57
24.	Function words	14
	Total	**1,460**

Indeed, the categorisation is not as obvious for all meanings – Haspelmath & Tadmor (2009b:6) acknowledge this with the example 'wheel' – a meaning which is placed in domain 10 (*Motion*), but could just as well have been in e.g. domain 9 (*Basic actions and technology*). In any case, for comparability reasons, the meaning list as well as the semantic domain classification in this thesis had to remain exactly the same as that of the LWT project, so this was not altered at all.

An unlimited number of lexemes are allowed per meaning, and each lexeme can also be linked to several meanings. This will be discussed under 3.2.1.

3.1.2 Application of the method to the current study

Whereas the aim of the LWT project was to make cross-linguistic generalisations on the kind of lexical meanings that are most likely to be borrowed, the aim of this thesis was slightly different: the same methodology and the corresponding data was used, but the focus of this study was instead to see how well Finnish fits into the cross-linguistic patterns; where it behaves like other languages, and where it differs from these general tendencies. Using the same methodology as the LWT project ensured the existence of a large, matching cross-linguistic study to which the results of this study could be compared, required for answering research question 3 (cf. section 1.1). While there are certainly other conceivable methods for collecting and quantifying the data for a study of this kind (e.g. different word lists, different semantic domains, more or less fine-grained variables), and while there are aspects of the LWT project which can be criticised, it is only by using the *exact* same methodology that the results can be compared to an existing, corresponding cross-linguistic sample, i.e. the findings of the LWT project.

The quantitative analysis was carried out by coding various values for each lexical item, relying on the information found in the dictionaries. These values include the borrowing status (according to the reliability scale mentioned earlier), donor language, time of borrowing (see 2.2.2 above for more information about the dating of lexemes), etc. Once the database feeding was completed, statistical analyses were performed on the values to obtain the final results. In section 3.2 below, the working process will be described in detail.

3.2 Working process

3.2.1 Identifying the meanings in Finnish

In order to carry out the data collection (i.e. the coding of the etymological data for the lexemes in question), some initial preparations had to be done. First of all, this meant identifying the Finnish term(s) corresponding to each LWT meaning. Standard Finnish was chosen as the subject of study, based on the availability of the material and its broader lexical coverage compared to more regional dialects. All the English LWT labels were looked up in an English-Finnish dictionary (Hurme et al. 2003) and were supplemented with terms by native intuition, as I am a native speaker of Finnish. Initially, *all* terms judged to correspond to each meaning were collected, which resulted in the expansion of the 1,460 LWT meanings to 1,896 Finnish terms. The next step in the process was to go through the words again, making decisions about which ones to keep and which ones to dispose of, with the help of another native speaker. This process was carried out by going through the LWT meanings one by one, explaining them to the native speaker (in Finnish, to avoid translational issues) using other words (carefully avoiding the terms in the list), and asking them to provide all words they knew that denoted each concept. Based on this, some terms collected in the initial step with the bilingual dictionary were deleted from the list, while others were added. The final version of the list ended up containing 1,825 Finnish lexemes, and can be found in the Appendix along with the rest of the coded data.

In some cases, delimiting the data turned out to be a difficult task. For sure, if any other person would set out to do the same thing, the list would almost certainly look different. One of the main difficulties was to assess guidelines for when it was reasonable to add more than one word for a meaning, and when it was not. There were several reasons to why this was problematic. For example, some meanings did not have an exact equivalent provided by a single Finnish word. The difficulty here had to do with expanding the scope of that meaning enough to find an equivalent (or more than one), but at the same time not stray too far from the core. There were also cases where there was a large number of close synonyms for a meaning. The problem here was the difficulty to motivate the inclusion of one lexeme and not another, while on the other hand this could potentially have yielded up to 10 lexemes for a single meaning. This was the case e.g. for the meanings 4.464 ‘the buttocks’, 4.67 ‘to have sex’, 4.65 ‘to piss’ and 4.66 ‘to shit’ – meanings that are often subject to taboo or euphemism replacement, “baby talk” and slang, and there are potentially endless variants that could have been included in the list. In the end, the meanings that were subjectively identified by my consultant as being part of “standard language” (and/or found in Hurme et al. 2003 without special usage remarks) were included.

Another problematic situation concerned the meanings that had two (or more) lexical equivalents with the same degree of neutrality and (more or less) the exact same connotations, but had so similar origins it would misrepresent the data to include them twice (or more) under the same meaning. An example is 1.26 ‘the mainland’, which in Finnish has (at least) three possible counterparts, *manner, mannermaa* and *mantere*. They are all compounds including or parallel forms of the same lexical root. It would be very misrepresentative of the data to include all three variants as separate entries under the same lexical meaning, and therefore this was avoided as far as possible in situations similar to this one. Thus, only *manner* ended up in the list.

3.2.2 Collection of etymological data

Once the list of Finnish lexical items was set, etymological data for each item was collected primarily from *NES* (Häkkinen 2013). Whenever a lexeme was unavailable here, it was instead looked up in *Suomen sanojen alkuperä* [The origin of Finnish words] (Itkonen 1992-2000; henceforth *SSA*). If a lexeme was unavailable in both of these sources (or needed supplementation), a number of other sources were consulted in a less strict order, depending on the nature of the data needed. All sources used in the data collection are listed in the References section, under Sources to etymological data. Completed items were not looked up in more than one source, unless there was some issue or detail that needed clarification.

The sources were selected on the basis of their comprehensiveness and their recency. *NES* and *SSA* are the only reasonably modern etymological dictionaries that focus specifically on contemporary Finnish and cover a large part the lexicon. Of these two, *NES* is more recent and was therefore chosen as the primary source.

3.2.3 Coding

The LWT template[11] was modified to only include those fields that would help answer the research questions of this study. Table 5 below shows which of the fields (horizontal axis) were used for answering each research question (vertical axis), and the symbols show the applicability: ✔ = applicable, (✔) = partly or indirectly applicable, − = not applicable. The research questions can be found under 1.1 above. The entire data set can be found in the Appendix.

Table 5. Fields from the LWT template used for the research questions of the current study.

	For all words				**Only for loanwords**			
RQ ↓	Finnish word form	Borrowed	Age	Created on loan basis	Immediate source word:			
					Word form	Donor language	Meaning	Language branch
1	✔	✔	−	(✔)	−	−	−	−
2	✔	✔	−	(✔)	−	−	−	−
3	✔	✔	−	(✔)	−	−	−	−
4	✔	✔	−	(✔)	(✔)	✔	(✔)	✔
5	✔	✔	✔	(✔)	(✔)	✔	(✔)	✔

In addition to the fields in Table 5, three pre-set Meaning fields with fixed values were also kept and used in the analysis: *LWT code*, *Semantic field* and *LWT label*. Below follows a more detailed description of each field.

LWT code, Semantic field and LWT label

Every meaning in the list has a unique identifier (the LWT code), which allows it to be linked to words denoting the same meaning in other languages. Each LWT code also contains information

[11] For an elaborate description of the original version, see Haspelmath & Tadmor (2009:1-34).

about the semantic domain it belongs to. Each meaning is connected to exactly one semantic domain, as assigned by the LWT project (read more under 3.1.1). The LWT label field contains the closest English approximation to the LWT code for each meaning in the list. This field really only fills a purpose for the coder, as a semantic meaning is most easily conveyed to humans by translation into a common language. Crucially for the comparison with the original study, none of these three pre-set fields have been altered.

Finnish word form

This is where the word form corresponding to each meaning was given (the meaning identification process is described under 3.2.1). An unlimited number of word forms was allowed per meaning, and if a meaning lacked an equivalent, the field could be left blank too. Naturally, the lack of an equivalent for a particular meaning also meant that the rest of the fields were left blank for that meaning.

In a number of cases, identical forms were given for more than one meaning. A few of these were *homonyms,* i.e. "unrelated senses of the same phonological word" (Saeed 2009:63), e.g. *kuusi* 'six' and *kuusi* 'fir'. Homonyms were simply marked with indices in parentheses (*kuusi (1), kuusi (2)*, etc.), and caused no further problem. On the other hand, *polysemous* words, i.e. phonological words which have multiple senses that – contrary to homonyms – are judged to be related (Saeed 2009:64), e.g. *kynsi* 'fingernail; claw', needed further action. To treat these polysemous entries as several different words would be a misrepresentation of their very character, as in reality a polysemous word is merely *one* word covering a relatively large semantic "space". The problem is thus one of considerable mismatch in "semantic size" between items in the LWT meaning list and the meanings connected to certain Finnish lexemes. Following the LWT methodology, this was dealt with by inserting a separate field, where, if applicable, the total number of LWT meanings per polysemous entry (this varied between 2 and 3) was coded. Based on this, the polysemous word forms were then attributed a weight of either 0.5 (if they occurred twice) or 0.33 (if they occurred three times) – all non-polysemous entries receiving a default weight of 1 (Haspelmath & Tadmor 2009b:9, 20-21). This weighting was used to calculate the *borrowed score* (see Borrowed and Created on loan basis below).

The most difficult cases of polysemy were those including a kind of taxonomic hierarchy. For example, when two hierarchically independent meanings in the list were covered by the same lexical item (e.g. *kynsi* 'fingernail; claw' mentioned above), neither meaning can be said to constitute a subcategory of the other, and therefore *kynsi* had to be occupy two slots in the list – under both 'nail' and 'claw' (cf. Figure 8). Rather, they are both subcategories of a common hierarchical parent. The problem appears when there is a hierarchical relationship between two meanings in the list that are both covered by the same lexical item in Finnish, e.g. in the case of terms for many family relations. A good example is the relationship between the meanings 2.52 'the aunt', 2.521 'the mother's sister' and 2.522 'the father's sister', where the former constitutes a hierarchical parent to the latter two (i.e., 'the mother's sister' and 'the father's sister' are different kinds of aunts). In Finnish – mirroring the English word *aunt* – the term *täti* covers all of these meanings, which led me to fill in *täti* under 'the aunt' and leave both 'the mother's sister' and 'the father's sister' empty (cf. Figure 8). The opposite situation was also encountered, best illustrated by the analogous meanings 2.51 'the uncle', 2.511 'the mother's brother' and 2.512 'the father's brother'. Here, the Finnish kinship system is asymmetrical, and instead of having a general meaning covering both parent's brothers (as is the case for 'aunt'), there is a term *eno* for 'the mother's brother' and another term *setä* for 'the father's brother'. No general term covering both meanings (like English *uncle*) exists, instead the speaker has to specify what kind of uncle they're talking about. Thus, in contrast to the 'aunt' situation described above, the meanings 'the mother's brother' and 'the father's brother' were filled in with *eno* and *setä* respectively, whereas 'the uncle' was left empty (cf. Figure 8). It is worth pointing out, however, that had 'the mother's brother' and 'the father's brother' not been present in the meaning list, they would both have ended up in separate entries under 'the uncle'.

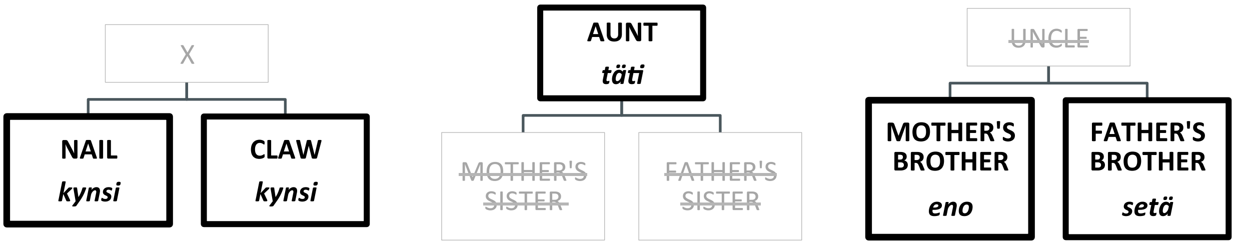

Figure 8. Diagrams illustrating three kinds of taxonomic hierarchies.

Borrowed and Created on loan basis

The *Borrowed* field contains information on the "borrowing status" of each lexeme in the meaning list (the data retrieved from etymological dictionaries), and can thus be considered the most central field in the entire study. This field was filled out with one of five labels, representing a *reliability scale*[12] which was used as another weighting basis in the analysis (Haspelmath & Tadmor 2009b:12-13). The weighting was achieved by converting the label in the *Borrowed* field to a *borrowed score*, ranging from 0 to 1. The labels and their corresponding borrowed scores can be seen in the list below. The borrowed score was then itself weighted by the polysemy score described under Finnish word form above. This score could then be used for calculations, as it included information also on how much weight should be attributed to a certain data point.

	Borrowed label	**Borrowed score**
4.	Clearly borrowed	1.00
3.	Probably borrowed	0.75
2.	Perhaps borrowed	0.50
1.	Very little evidence for borrowing	0.25
0.	No evidence for borrowing	0.00

The quantification of a certainty measure such as this one necessarily entails a certain degree of subjectivity. I have tried to represent, as closely as possible, the etymological certainty expressed in the source for each item, but as this is a question of interpretation, the result would undoubtedly differ at least slightly if the coding was made by someone else.

The *Created on loan basis* field was used to code whether a lexeme contained borrowed material without being a loanword (for definitions, see 2.2.2). This was indicated using the same reliability scale as for the *Borrowed* field, but with slightly renamed labels: *4. Clearly created on loan basis*, *3. Probably created on loan basis*, *2. Perhaps created on loan basis*, *1. Very little evidence for loan basis* and *0. No evidence for loan basis*. This field was not really used in the

[12] Note that the labels in the lower spectrum of this scale do not imply that the word in question has not been borrowed – only that we have no evidence for it (cf. the discussion under 2.2.2).

analysis per se, but served more as background or additional information – the data is presented as additional results under 4.1.

If the *Borrowed* label was either '3. Probably borrowed' or '4. Clearly borrowed', the fields under 'Only for loanwords' were also required. For reliability degrees lower than that, these fields did not require values (Haspelmath & Tadmor 2009b:15). Also, in the basic percentage counts (i.e. analyses not involving the borrowed score) in the analyses presented in chapter 4 below, only 3's and 4's count as loanwords.

In order to categorise a word list item as a loanword (whether certain, probable, possible etc.), it must have been borrowed directly. For example, *sokeri* 'sugar' is an undisputable loanword, borrowed from Swe. *socker* (Häkkinen 2013:1175). *Sokeriruoko* 'sugar cane', on the other hand, does *not* count as a loanword according to this method of analysis, even though it is a compound created on the basis of two borrowed components: *sokeri* and *ruoko* 'cane' (of Germanic or Baltic origin (Häkkinen 2013:1070)). The key part here is whether or not the word list item has been borrowed *as such* – if not, it gets the value '0. No evidence for borrowing' in the Borrowed field, and its borrowed origins are instead coded in the field Created on loan basis. This is the case for *sokeriruoko*, which must have been created in Finnish from "Finnish" material – albeit of loan origin, and probably under the influence of similar forms in other languages, e.g. Eng. *sugar cane* and Swe. *sockerrör* 'sugar cane' (lit. 'sugar cane/tube') – but, as outlined in 2.2.2 above, *calques* like this one do not qualify as loanwords, as they do not involve the direct borrowing of any lexical material.

The same criteria of directness also apply to derivations. However, an important distinction has to be made between *inflectional morphology* on the one hand and *derivational morphology* on the other[13]. An inflection of a borrowed word, e.g. *tikkaat* 'ladder' (lit. the plural form of *tikas* 'ladder', which in its singular form is used only in compounds) still counts as an instance of the same word, and is thus coded as a loanword (*tikas* < OS *stighi, stige* 'ladder' (Häkkinen

[13] It should be mentioned that there are cases which do not easily fall into either category in this dichotomy – cf. e.g. chapter 4 in Haspelmath (2002) for a thorough discussion on inflectional vs. derivational morphology.

2013:1309-10)). A derived term, however, no longer constitutes an instance of the word it was derived from. Thus, inflectional morphology does not interfere with a loanword's borrowed status, whereas derivational morphology does. Unless, of course, an inflected form has fossilised as a separate word, as is the case with e.g. the adverb *ääressä* 'at', which morphologically is analysable as the inessive singular form of the noun *ääri* 'end, verge', but which can be categorised as a (null) derivation due to its shifted grammatical function as well as belonging to another word class.

The etymology of a word can be thought of as a lineage where various mechanisms may have played a role. For example, the origin of a word can sometimes (to the best of our current knowledge, at least) be rather uncomplicated, the modern form simply being genetically inherited from an ancestral stage. An example of this would be Fi. *tuli*, which is inherited from Proto-Uralic **tule* (Häkkinen 2013), as illustrated in "A" in Figure 9. As the figure shows, no language boundary is being crossed horizontally (to reuse the metaphors from Figure 2). A situation like this one would, in the present study, generate a simple "0. No evidence for borrowing" in the *Borrowed* field.

As soon as a form (or its direct ancestor at any given stage) can be identified as a loanword, this would be coded in the *Borrowed* field (the value depends on the certainty of the evidence). An example is Fi. *nappi*, borrowed from Swe. *knapp* (Häkkinen 2013). As illustrated in "B" in Figure 9, a language boundary has been crossed horizontally in this case. Regardless of whether the form was borrowed into Finnish quite recently or into e.g. Proto-Uralic thousands of years ago, the coding would be the same.

A third kind of situation arises when a word is at some point in the language development derived from or part of a compound including a word that has been borrowed at some language stage. Derivation is represented by the *diagonal* arrow in "C" in Figure 9. The example used here is Fi. *lainata* 'to lend, to borrow', which is a verb derived from Proto-Finnic **laihna* 'loan', in turn borrowed from Proto-Germanic (Häkkinen 2013, Itkonen 1992-2000). Even though a language boundary has been crossed horizontally, a derivation has taken place since, which "cancels" the loan etymology from a coding point of view. A word that has undergone derivation or compounding is no longer regarded as a representation of the "original" form, and was therefore not coded as a loanword of any kind. The foreign origin was, however, coded in the *Created on loan basis* field.

Thus, the three situations described above show that a word was coded as a loanword *if and only if* there is evidence of *horizontal* transmission which is not followed by any *diagonal* developments. The age itself of the borrowing was not relevant for the coding.

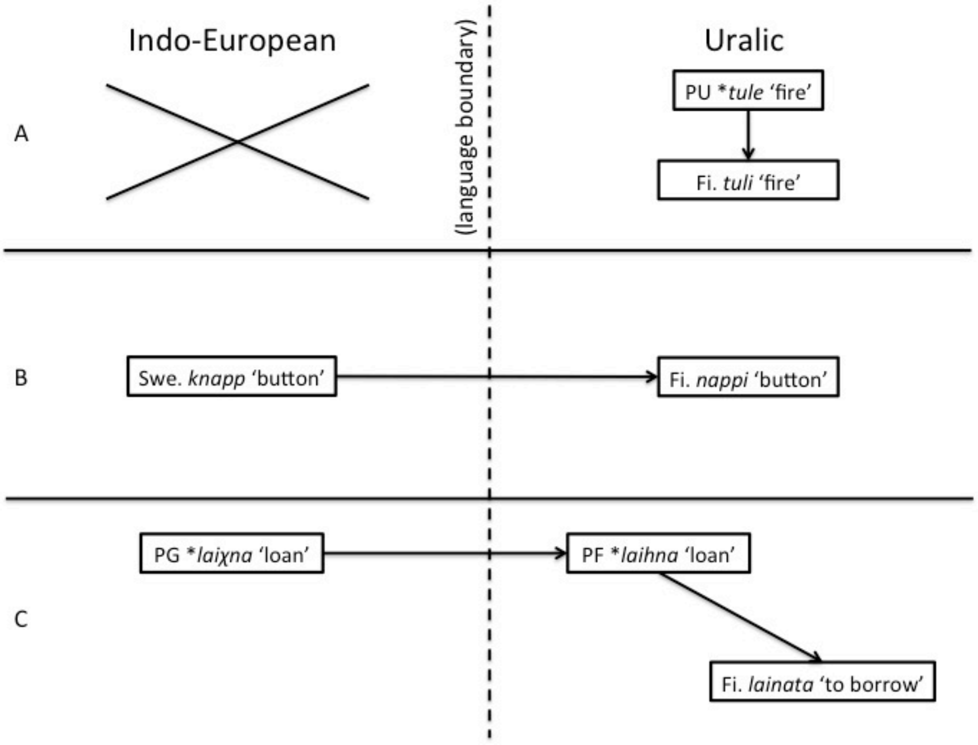

Figure 9. Illustration of three kinds of etymological lineages.

On a final note, a word's origin can sometimes be misrepresented by the system of categorisation in this template (as, indeed, would be expected for any study where qualitative data has to be forced into quantifiable units). For example, *mitata* 'to measure' may according to Häkkinen (2013:718-719) either be a loanword itself or a derivation of *mitta*[14] 'measure', which in turn is

[14] This is just an example – *mitata* is not part of the word list included in this thesis.

either a loanword itself or a derivation of *mitata*. In other words, it is clear that both words stem from the same Germanic root, but as the forms make it impossible to tell which was first borrowed and which is a derivation from the original loanword (unless they are two separate borrowings from the same root, which is also a possibility), the method forces us to categorise them both as '2. Perhaps borrowed' and '2. Perhaps created on loan basis' in the fields *Borrowed* and *Created on loan basis*, respectively. Thus, even when the etymological origin of a word can be pinpointed with quite a fair amount of both certainty and accuracy, that word can be forced into the categorisation in a way that does not mirror the actual situation – in this case that a word with undisputed loan origins only gets categorised as "perhaps" borrowed/created on loan basis. However, as this analysis is not meant to focus on specific words but rather on the big picture, the overall representation should not suffer from this.

Age

As explained and motivated in section 2.3.1 above, the family tree that was chosen to serve as the basis for the quantification in the *Age* field is the traditional, binarily branching one. A model of this tree can be seen in Figure 10, where the nodes with thicker frames represent the values that were coded in the *Age* field. In this field, for every word list item, the earliest attested/reconstructible language stage is given: *Uralic* (U), *Finno-Ugric* (FU), *Finno-Permic* (FP), *Finno-Volgaic* (FV), *Finno-Saamic* (FS), *Finnic* (F), *Northern Finnic* (NF) or *Finnish* (Fi.). This has been assessed by accounts of cognates in related languages, as given in the etymological dictionaries. This study has not been concerned with any absolute dating of the language stages.

The general principle in assigning age has been to aim low rather than high, so the language stages here should be read as reflecting the oldest *known* age of any particular word, not necessarily its *actual* age. *Finnish* was used in cases when it was reported that no cognates were known, but it was also used as a default value when no information was available. Most phrasals and many compounds have been assigned this value, simply for the reason that they usually do not constitute entries in etymological dictionaries, and therefore lack accounts of cognates.

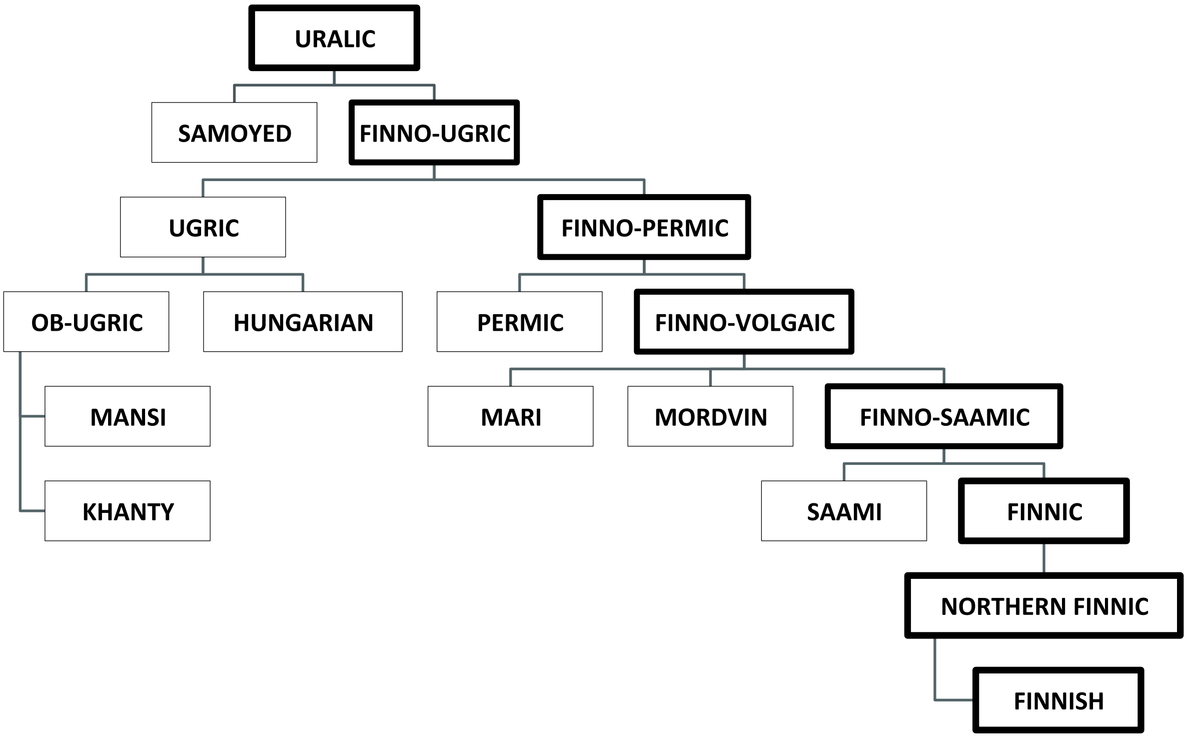

Figure 10. The Uralic family tree model and nodes settled on for this study.

The model does *not* take into account e.g. derivations or compounds, i.e. if a Finnish word is derived (at the Finnish language stage) from an older, inherited word that might go as far back as Proto-Uralic, the *Age* value of that particular word would remain at 'Finnish' – representing the oldest known age of that particular derivation.

See section 2.2.2 for more on loanword dating, and 2.3.1 for more on (the disagreements regarding) the Uralic family tree and for a motivation for the choice of the included language stages for this study.

Immediate source word: Word form, Donor language, Meaning and Language branch

As stated in Borrowed and Created on loan basis under 3.2.3, these fields were only required if the word form was given the label '3. Probably borrowed' or '4. Clearly borrowed' in the *Borrowed* field. These four fields were populated with information about the immediate source word to each loanword: word form, meaning, donor language and/or donor language branch. If all or some of this information was for some reason unavailable, or if it was not unambiguous, the corresponding field(s) was/were left empty.

This thesis focuses on the languages Finnish has been in *direct* contact with, and thus what is being coded and quantified are the immediate donor languages and the immediate source words. Coding for the original donor language (or, more realistically: the oldest *known* donor language) would tell us nothing about the situation when the borrowing into Finnish took place, which is what we're interested in. For example, Fi. *tupakka* 'tobacco' is borrowed from Swe. *tobak* 'tobacco', although Swedish is obviously not the oldest traceable source to this widespread root.

16 immediate donor languages were found: *English, Low German, Neo-Classical, Old East Slavic, Old Norse, Old Swedish, Proto-Baltic, Proto-Balto-Slavic, Proto-Germanic, Proto-Indo-European, Proto-Indo-Iranian, Proto-Iranian, Proto-Slavic, Russian, Saami (undefined)* and *Swedish.* In addition to these, there was also an option for when the information was unavailable: a 17th value *Unknown*.

Some simplifications were made: for instance, many different stages of Swedish were identified in the dictionaries, but for the purposes of this thesis (which focuses on large-scale patterns, cf. 1.1), they were grouped under *Old Swedish* and *Swedish*, respectively[15]. Similarly, Old Low German and Middle Low German were collapsed into *Low German.* Proto-Norse and Old Norse were also collapsed into *Old Norse*. The "language" *Neo-Classical* is a simplification in its very essence, and collects cases of hard-to-trace modern loanwords that all occur in a number of (Indo-)European languages[16] and contain neologisms based on Latin and/or Greek material. *Proto-Baltic* should also be seen as an approximation, denoting early borrowings of (unspecified) Baltic origin[17]. Likewise, *Saami (undefined)* is a generalisation that had to be made due to the lack of information about more specific Saami origins.

[15] *Early Old Swedish, Late Old Swedish, Old Swedish > Old Swedish*; *Early Modern Swedish, Late Modern Swedish, Modern Swedish, Contemporary Swedish > Swedish.*

[16] According to the definition under 2.2.2 above, these words could perhaps be described as Wanderwörter.

[17] Many scholars do not consider it justified to reconstruct a common Baltic stage – see e.g. Derksen (2010).

Table 6 below shows the grouping of the donor languages into their respective donor language families. Note that the grouping called "Indo-European" does not only comprise loanwords from PIE, but also loanwords of clearly Indo-European origin (including the Neo-Classicisms) that cannot be identified as belonging to any of the specified Indo-European branches in the list (*Balto-Slavic, Germanic* or *Indo-Iranian*). Also note that a loanword's precise donor language could be *Unknown*, while the donor language branch could be identified. This is why *(Unknown)* occurs under all groups in the table (except Saami).

Table 6. The identified donor languages, grouped according to language branch.

Donor language branch	**Donor languages**
Balto-Slavic	*Old East Slavic, Proto-Baltic, Proto-Balto-Slavic, Proto-Slavic, Russian, (Unknown)*
Germanic	*English, Low German, Old Norse, Old Swedish, Proto-Germanic, Swedish, (Unknown)*
Indo-European	*Neo-Classical, Proto-Indo-European, (Unknown)*
Indo-Iranian	*Proto-Indo-Iranian, Proto-Iranian, (Unknown)*
Saami	*Saami (undefined)*
(Unknown)	*(Unknown)*

The entire data set used for the quantification can be found in the Appendix.

4 Analysis and Results

In this section, the results corresponding to each research question will be presented and analysed individually. The various analyses that had to be conducted in order to answer the research questions are summarised in steps in the schema in Figure 11 below.

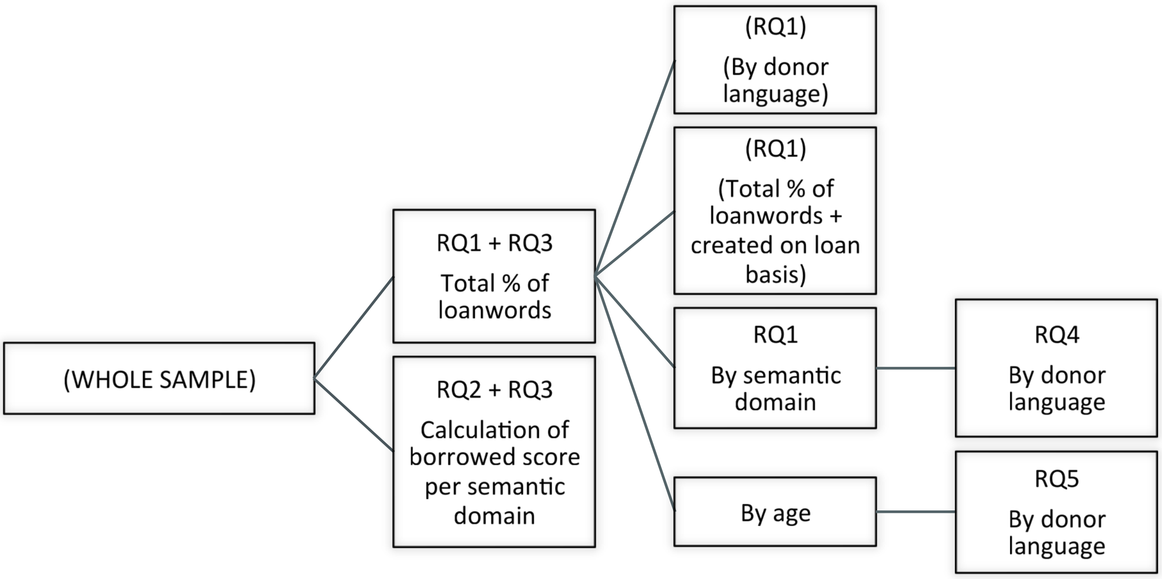

Figure 11. The steps involved in the analyses.

First, the total percentage of loanwords in the entire sample was calculated. Calculations were then performed with the loanwords grouped according to different codings (donor language; semantic domain; age), and thereafter (for the latter two) by donor language. In addition, as the borrowability per sematic domain in the results of the LWT project is presented as borrowed scores (and not basic percentages), the borrowed scores per semantic domain had to be calculated as well, to ensure comparability.

The analyses only involved basic statistics, and were therefore all performed using Excel formulae. The research questions will be answered in a subsection each below, where the analyses as well as the results will be presented and described.

It is worth pointing out that the Finnish vocabulary will be represented by the 1,825 words in the study, and that the "cross-linguistic tendencies" the Finnish sample is being compared to is represented by the average scores in a quite small number of languages (i.e. the findings of the Loanword Typology project, cf. Tadmor 2009). Thus, henceforth, when I refer to things like "the Finnish vocabulary" or "the cross-linguistic average", bear in mind that these are represented by these samples.

4.1 Overall percentage of loanwords

RQ 1: How much of the Finnish vocabulary is borrowed?

The Finnish language sample ended up containing 1,825 words. Of these, 480 were borrowed, i.e. were coded as either 3 or 4 on the reliability scale (see 3.2.3 (subsection Borrowed and Created on loan basis) above; Haspelmath & Tadmor 2009b:13). The Finnish vocabulary, as represented by this word list, thus contains 26.3% loanwords.

If we – for the sake of loan origins, not just loanwords – were to calculate the percentage not just of words which were "clearly" or "probably" borrowed, but also those that were "clearly" or "probably" created on loan basis (see Borrowed and Created on loan basis under 3.2.3 above), the total number of words would be 716, which corresponds to 39.23% of the sample.

In close connection to this research question (and since I had the data at hand), I also chose to analyse how much of the loanwords came from each immediate donor language. Figure 12 shows the distribution of donor languages to the loanwords found in the sample, colour coded by language branch and ordered from the highest to the lowest amount of contributed loanwords. Figure 13 shows how much each language branch has contributed.

As shown by the pie chart in Figure 13, Germanic languages (70.63%) have been the immediate source of an overwhelming majority of the borrowings. Swedish is in first place (30%) and Proto-Germanic in second (24.58%). Old Swedish also makes up a considerable portion (10%), as does Proto-Baltic (9.79%). The rest of the individual donor languages make up less than 5% each. 7.92% of the loanwords could not be attributed to a specific language, and were therefore categorised as *Unknown*.

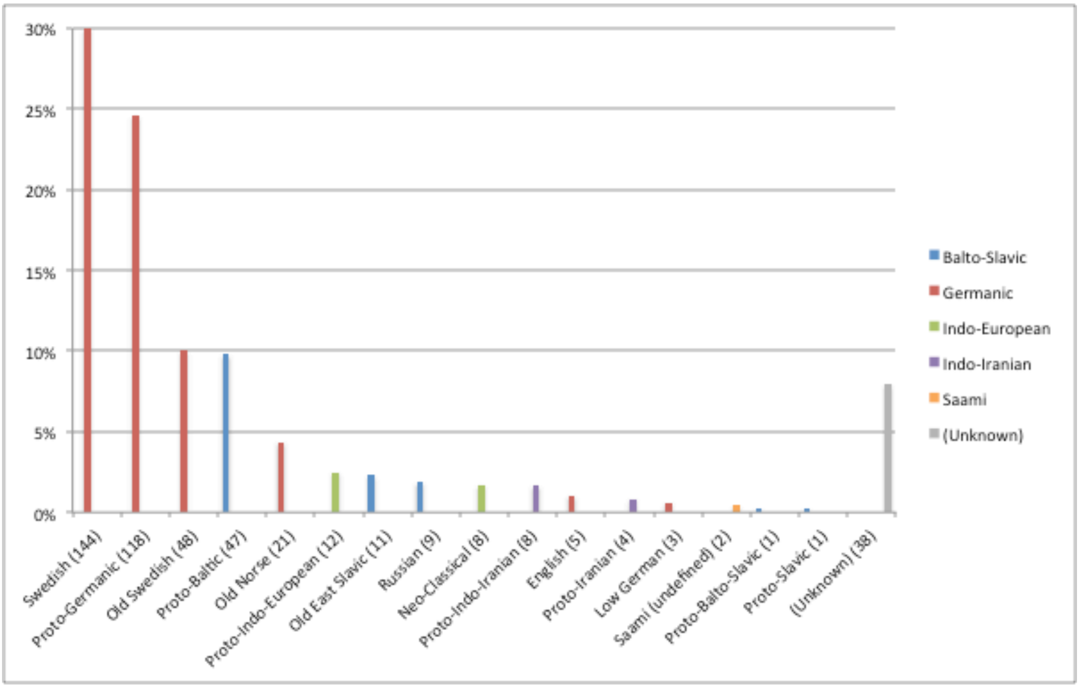

Figure 12. Percentages of loanwords from each donor language.

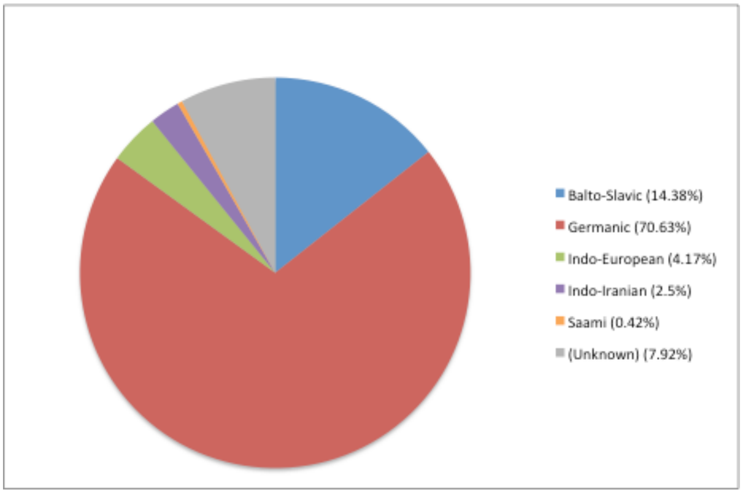

Figure 13. Percentages of loanwords from each donor language branch.

It should also be pointed out that all language branches except Saami "Unknown" are in fact Indo-European, which means that Indo-European languages as a whole are the source of (at least) 91.67% of all loanwords in this study. Further analyses relating to donor languages are presented under 4.4 and 4.5 below.

4.2 Semantic domains

RQ 2: Which semantic domains have the highest vs. the lowest amount of loanwords?

To answer this question in a way compliant with the LWT project, the loanword percentages for each of the 24 semantic domains were converted to *borrowed scores* (see Borrowed and Created on loan basis under 3.2.3 above), weighted according to the polysemous entries. The average borrowed score for each semantic domain was then calculated on the basis of the total number of words per domain. These are shown in Table 7, ordered from the highest to the lowest borrowed score. (The loanword percentages are also given for reference, as are the total number of words and the total number of loanwords, but the borrowed scores are the important numbers for this analysis.)

Table 7. Percentage of loanwords and borrowed score for each semantic domain.

Semantic domain	**Words, total**	**Loanwords**	**Loanwords, %**	**Borrowed score**
The house	63	32	50.79%	0.51
Clothing and grooming	79	40	50.63%	0.50
Modern world	66	32	48.48%	0.48
Animals	136	55	40.44%	0.43
Agriculture and vegetation	86	34	39.53%	0.40
Religion and belief	35	14	40.00%	0.40
Food and drink	108	40	37.04%	0.35
Social and political relations	49	17	34.69%	0.34
Basic actions and technology	104	30	28.85%	0.31
Warfare and hunting	52	14	26.92%	0.30
Possession	55	15	27.27%	0.30
The physical world	93	19	20.43%	0.24
Quantity	42	9	21.43%	0.23
Law	28	7	25.00%	0.23
Time	71	12	16.90%	0.23
Emotions and values	65	14	21.54%	0.23
The body	199	35	17.59%	0.21
Motion	104	19	18.27%	0.20
Spatial relations	95	11	11.58%	0.16
Speech and language	49	5	10.20%	0.15
Cognition	71	10	14.08%	0.15
Sense perception	54	5	9.26%	0.12
Kinship	100	10	10.00%	0.11
Function words	21	1	4.76%	0.07

Two semantic domains – *The house* and *Clothing and grooming* – scored 0.5 or higher (the maximum possible borrowed score being 1.0). *Modern world* arrives in third place not far behind. Continuing downwards, many of the following semantic domains – e.g. *Agriculture and vegetation, Religion and belief, Warfare and hunting* – also contain typically cultural vocabulary, i.e. vocabulary that is traditionally thought of as contact-sensitive (cf. 2.2.2 and 3.1). Many semantic domains that would be classified as "basic vocabulary" can be found at the bottom of the list, e.g. *Function words, Kinship, Sense perception*. Generally speaking, the order in this list is quite in line with the expectations for any language. It will be compared to the corresponding average borrowed scores from the LWT sample in 4.3 below, and we will get back to semantic domains again in both 4.4 and 4.5.

4.3 Comparison to the LWT results

RQ 3: Does the Finnish sample differ in any noticeable way from the cross-linguistic tendencies found in the Loanword Typology project?

The languages in the LWT project were divided into four groups, based on their overall percentage of loanwords: *very high borrowers* (> 50%), *high borrowers* (25-50%), *average borrowers* (10-25%) and *low borrowers* (< 10%) (Tadmor 2009:56-57). As presented in 4.1 above, the Finnish sample contains 26.3% loanwords, which places it in the lower realm of the high borrowers.

Table 8 (adapted from Tadmor 2009:56-57) shows the basic results of the entire LWT sample, with Finnish inserted in the right spot and marked in bold.

Table 8. Loanword percentages of each individual language in the Loanword Typology project.

Borrowing type	Languages	Words, total	Loanwords	Loanwords, %
Very high borrowers	Selice Romani	1,431	898	62.7%
	Tarifiyt Berber	1,526	789	51.7%
High borrowers	Gurindji	842	384	45.6%
	Romanian	2,137	894	41.8%
	English	1,504	617	41.0%
	Saramaccan	1,089	417	38.3%
	Ceq Wong	862	319	37.0%
	Japanese	1,975	689	34.9%
	Indonesian	1,942	660	34.0%
	Bezhta	1,344	427	31.8%
	Kildin Saami	1,336	408	30.5%
	Imbabura Quechua	1,158	350	30.2%
	Archi	1,112	328	29.5%
	Sakha	1,411	409	29.0%
	Vietnamese	1,477	415	28.1%
	Swahili	1,610	447	27.8%
	Yaqui	1,379	366	26.5%
	Finnish	**1,825**	**480**	**26.3%**
	Thai	2,063	539	26.1%
	Takia	1,123	291	25.9%
Average borrowers	Lower Sorbian	1,671	374	22.4%
	Hausa	1,452	323	22.2%
	Mapudungun	1,236	274	22.2%
	White Hmong	1,290	273	21.2%
	Kanuri	1,427	283	19.8%
	Dutch	1,513	289	19.1%
	Malagasy	1,526	267	17.5%
	Zinacantán Tzotzil	1,217	195	16.0%
	Wichí	1,187	188	15.8%
	Q’eqchi’	1,774	266	15.0%
	Iraqw	1,117	162	14.5%
	Kali’na	1,110	156	14.0%
	Hawaiian	1,245	169	13.6%
	Oroqen	1,138	137	12.0%
	Hup	993	114	11.5%
	Gawwada	982	111	11.3%
	Seychelles Creole	1,879	201	10.7%
	Otomi	2,158	231	10.7%
Low borrowers	Ket	1,030	100	9.7%
	Manange	1,009	84	8.3%
	Old High German	1,203	70	5.8%
	Mandarin Chinese	2,042	25	1.2%

If we exclude Finnish from the list above, and divide the total number of words in all languages (57,520) with the total number of loanwords in all languages (13,939), this gives us an average loanword percentage of 24.23% based on the whole sample. On the whole, then, Finnish only has a slightly higher percentage of loanwords than the average language (as represented by the LWT project). The median percentage of the languages above (again, Finnish excluded) is 22.2% (Hausa). This calculation method also supports the position of Finnish as just above average. Hypothesis B – the one about Finnish having a higher amount of loanwords than the cross-linguistic average – is thus hardly supported by the results of this study, but see chapter 5 below for a discussion on some methodological shortcomings which may have affected these statistics.

Table 9. Average borrowed score for each semantic domain in the Loanword Typology project.

Semantic domain	**Borrowed score**
Modern world	0.64
Social and political relations	0.64
Religion and belief	0.49
Agriculture and vegetation	0.45
Function words	0.44
The house	0.40
Clothing and grooming	0.40
Food and drink	0.37
Speech and language	0.36
Law	0.36
Possession	0.34
Warfare and hunting	0.34
Basic actions and technology	0.33
Quantity	0.33
Animals	0.32
Emotions and values	0.30
Cognition	0.29
Time	0.28
Kinship	0.23
The physical world	0.21
Motion	0.21
Sense perception	0.19
The body	0.17
Spatial relations	0.15

Turning back to semantic domains, Table 9 shows the average borrowed score for each semantic domain, levelled out across all languages in the LWT sample. The corresponding scores for Finnish were presented in Table 7 above. The two sets of scores are compared in Figure 14, the semantic domains ordered by their individual borrowed score in Finnish (from highest to lowest).

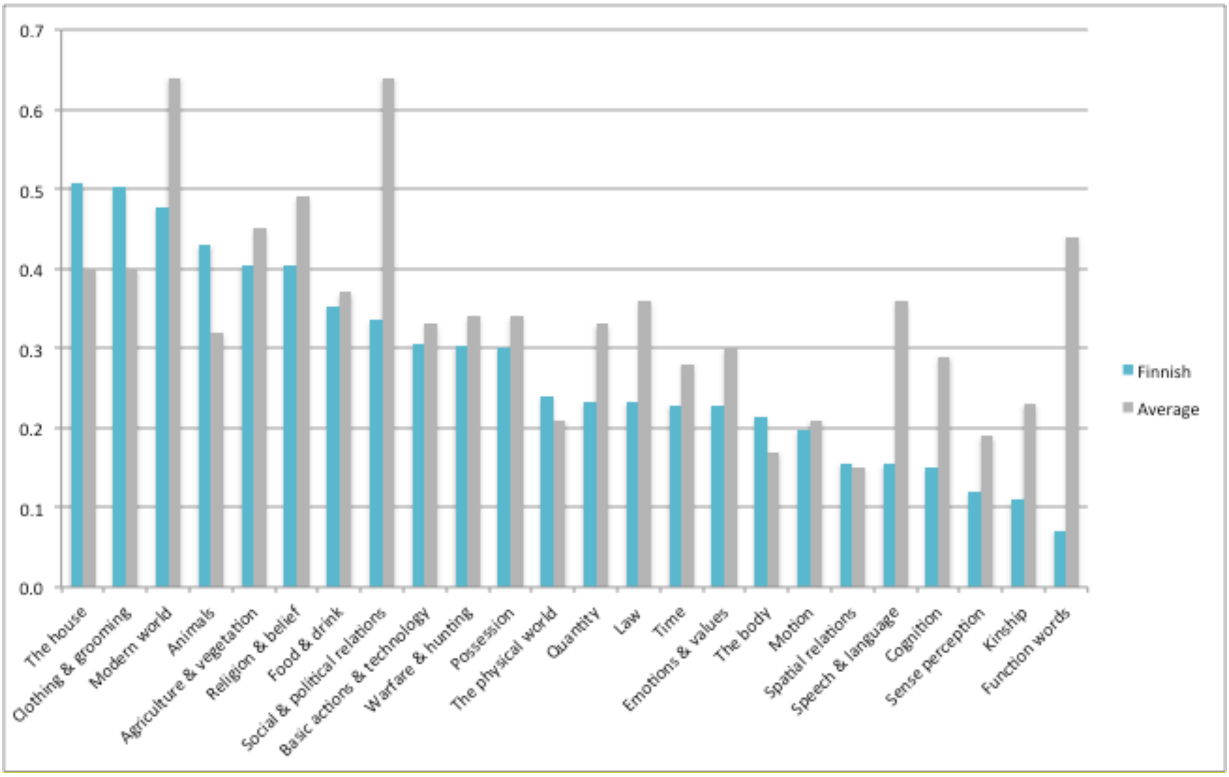

Figure 14. Borrowed score per semantic domain in Finnish vs. the LWT average.

The results that stand out immediately in Figure 14 are those where Finnish has a remarkably lower amount of loanwords than the average (a difference of more that 0.10 in the borrowed score), e.g. *Function words, Social and political relations, Speech and language, Modern world, Cognition, Law, Kinship* and *Quantity*.

Only three domains have a considerably higher amount of loanwords in Finnish than in the average language sample (again, a difference of more that 0.10 in the borrowed score): *Animals, The house* and *Clothing and grooming*.

The rest of the domains, namely *Religion and belief, Emotions and values, Sense perception, Time, Agriculture and vegetation, Possession, Warfare and hunting, Basic actions and technology, Food and drink, Motion*, *Spatial relations*, *The body* and *The physical world,* have a borrowed score in Finnish which differs less than 0.10 units from the corresponding score in the LWT results. These domains can thus be considered relatively compliant with the cross-linguistic patterns identified by the LWT project.

4.4 Semantic domain vs. donor language

RQ 4: Is there a correlation between semantic domain and donor language (family)?

These scores were achieved by first grouping the loanwords by semantic domain, and then calculating the amount of loanwords per donor language branch in each domain (percentually). The results of the individual languages were too fine-grained to yield any discernible patterns (due to their large number, cf. Figure 12 above), so these were abandoned in favour of the language branches, which generated more legible results. Figure 15 shows the results, the semantic domains ordered by their amount of loanwords in real numbers (from highest to lowest). Each number can be seen in parentheses after the name of the semantic domain. As the number of loanwords within a domain varies between 55 and 1, the percentages for the semantic domains farther to the left are generally more reliable, as they are based on a larger loanword sample.

As we have already established (see 4.1 above), Germanic stands for 70.63% – almost three quarters – of all loanwords, so it is not very surprising that Germanic is the language branch that clearly distinguishes itself from the rest in Figure 15. Only in one semantic domain, *Quantity*, Balto-Slavic can meet the challenge, and the two language branches have contributed an equal amount of loanwords (33.33% each) – although this domain only contains 9 loanwords. In the domain *Sense perception*, the loanwords of Germanic origin “only” amount to 40%. In all of the remaining 22 semantic domains, at least 50% of the loanwords come from Germanic, and in two cases (*Law* and *Function words*), Germanic stands for 100% of the loanwords – however, these domains, in all, only contain 7 and 1 loanwords, respectively.

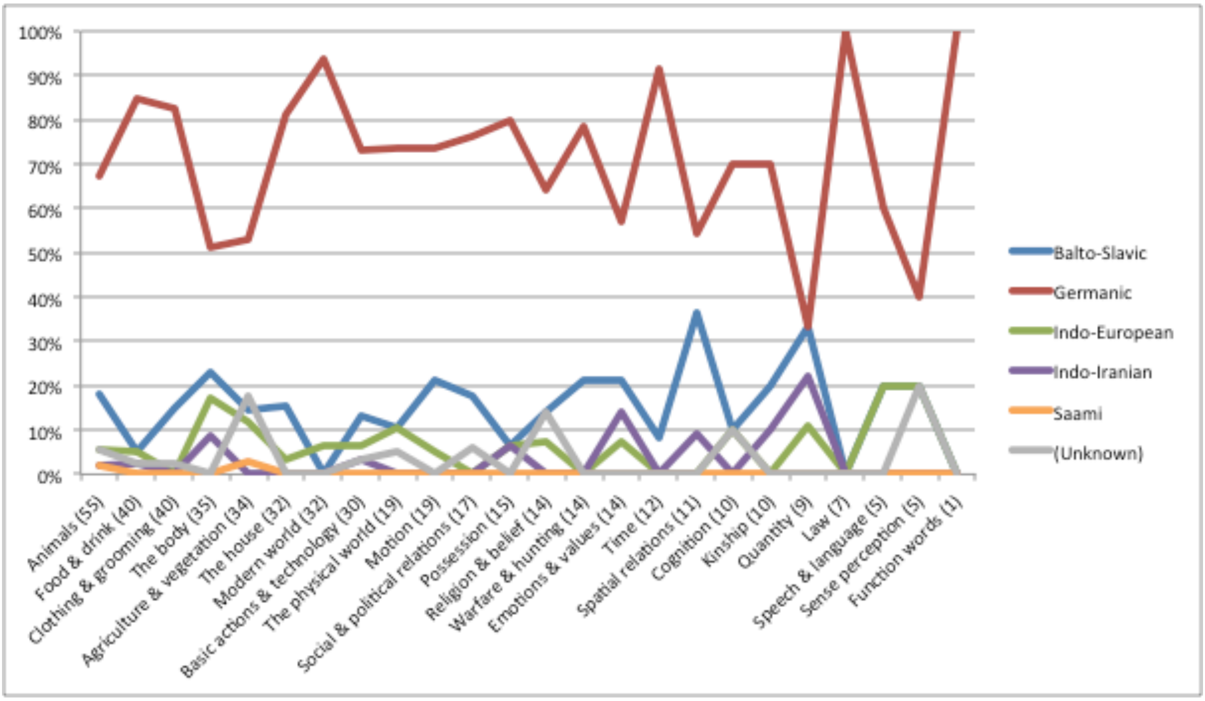

Figure 15. Percentage of loanwords per donor language branch and semantic domain.

As for the Balto-Slavic loanwords, two semantic domains stand out from the rest: *Spatial relations* (36.36%) and *Quantity* (33.33%). *Quantity* (22.22%) is also the domain which distinguishes itself from the rest when we look at the Indo-Iranian loanwords, alongside *Emotions and values* (14.29%). The Indo-European grouping is perhaps not as interesting, as it is a bit of a dustbin category (containing – in addition to the Proto-Indo-European loanwords – a lot of uncategorisable but clearly Indo-European loans). In any case, *Speech and language* (20%), *Sense perception* (20%) as well as *The body* (17.14%) clearly stand out here. To give some additional information, Proto-Indo-European stands for 20%, 20% and 2.86% of the loanwords in these domains, respectively. (Keep in mind, however, that all semantic domains discussed in this paragraph (except *The body*) contain very few loanwords.)

As for Saami, there is little meaning in drawing any generalised conclusions based on the two loanwords that were found in the entire sample.

4.5 Time period vs. donor language

RQ 5: Is there a correlation between time period and donor language (family)?

For this analysis, the loanwords were first divided by their age in the Finnish lineage (as per the language stages identified in the 'Age' field), and then the amount of loanwords per donor language branch was calculated (percentually) for each language stage. In this analysis, as in 4.4 above, the individual languages were abandoned for the benefit of the language branches, since the former yielded too fine-grained and hard-to-interpret results.

It should be pointed out that the division into time periods is based purely on cognate data from related Uralic languages (see the description in subsection Age under 3.2.3, and the motivation for the included language stages in 2.3.1). This entails that the age assessment is better described as a measurement of the earliest known language stage to which we can reconstruct the word. In other words, nothing is in the way of a word being older than the stage as which it has been classified, but – assuming our cognate assessment is correct – the word should not be younger. No data relating to the contact situation or the donor language has explicitly been taken into account, although such information may have been influential in the etymological work behind the dictionary entries.

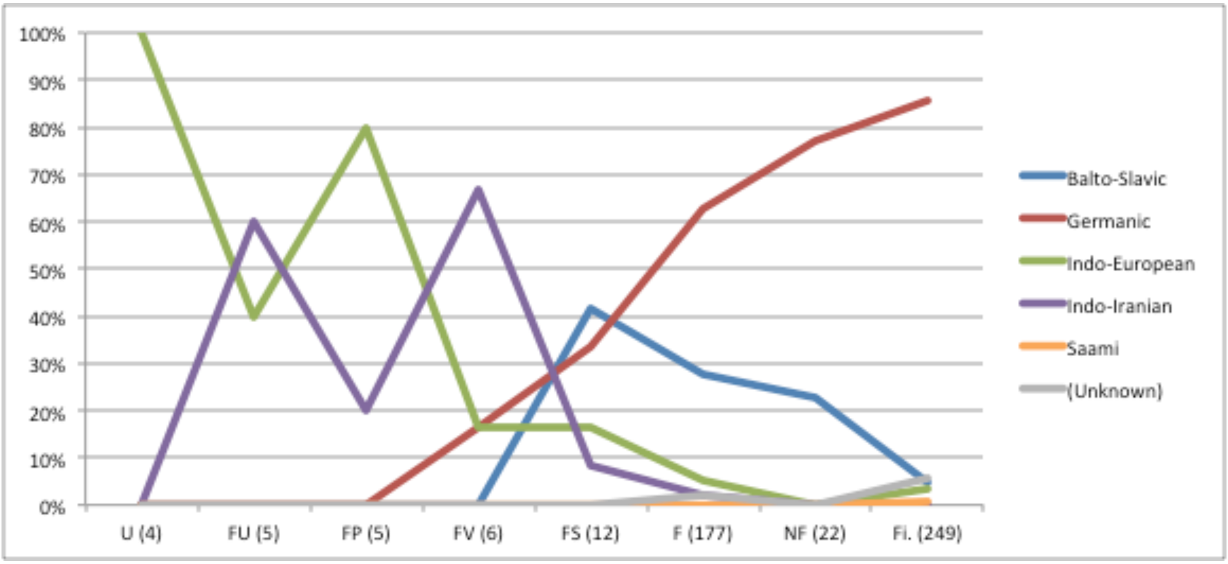

Figure 16. Percentage of loanwords per donor language branch and time period.

Figure 16 shows the percentual distribution of each donor language branch with respect to the most recent age the loanword can be reconstructed to. The loanword stages are ordered according to their age, from the oldest (Uralic) to the most recent (Finnish). The numbers in parentheses after each abbreviated language stage indicate the total number of loanwords reconstructible for that stage. Before moving on to the results, the large differences in the number of loanwords (ranging from 4 to 249) should be mentioned.

As mentioned in the Age section under 3.2.3 above, the principle in the age assignment has been to aim low rather than high, with the inevitable consequence that there is a considerable overweight of words in the more recent language stages. These stages may thus contain some loanwords which are in fact older, but which can't be proved to be older. In addition, the words lacking information on cognates have simply been placed at the stage *Finnish*. Last but not least, despite the large differences, it is unsurprising that the older stages contain less loanwords than the more recent ones, as the farther back in time we go, the more obscured our linguistic knowledge is – due to e.g. lexical replacement, the data from the earliest language strata is of course not as well preserved as that of the younger stages.

It should also be borne in mind that at the time of Proto-Uralic, there were (in all likelihood) not yet any distinct Indo-European branches, so it is unsurprising that Indo-European stands for 100% of the loanwords at this stage (although the number of loanwords reconstructible to this stage only amount to 4). (Again – "Indo-European" may be a bit misleading as a category, as it incorporates loanwords from Proto-Indo-European, but also those loanwords that can be identified as Indo-European but not ascribed to any certain branch.) Similarly, the Saami and Finnic branches only part after the Finno-Saamic stage[18], so it is unsurprising that these loanwords only turn up in the later stages.

These caveats aside, we can still see some interesting patterns. We note an Indo-European peak back when the family had not yet been divided into its individual branches, as well as a decline

[18] At least according to the family tree model selected for this particular study, cf. sections 2.3.1 and 3.2.3 (subsection Age).

in favour of more specific Indo-European origins. Two relatively early Indo-Iranian peaks can be distinguished, with a decline probably due to the increasing geographical distance between the speech communities (cf. the most widely accepted theories on the Uralic homeland under section 2.3.1). The relatively large fluctuations of these two language groups during the four earliest language stages should, however, not be paid much attention to, as the stages consist of so few loanwords each. The Balto-Slavic contacts set in roughly at the time of decline of the Indo-Iranian contacts, probably reflecting the onset of their contact as Uralic languages began to be spoken in the area around the Baltic Sea. The Balto-Slavic peak at the Finno-Saamic language stage should, due to the small number of loanwords at this stage (12, whereof 5 from BS), perhaps be overlooked in favour of the Balto-Slavic loanwords at the Finnic stage, which has a considerably higher amount of loanwords (177, whereof 49 from BS). Germanic loanwords are beginning to be introduced at around the same time, and increase with every language stage up until the current one (i.e. *Finnish* as a separate language variety), when almost all loanwords are of Germanic origin – in contrast to the Balto-Slavic curve, which has an early peak followed by a gradual decline. The Saami loanwords are, again, too few to serve as a ground for any generalisations.

5 Discussion

First of all, when compared to the languages in the Loanword Typology project, Finnish is a fairly typical language from a loanword typological viewpoint. According to the present results, 26.3% of the investigated Finnish lexicon consists of loanwords, which is just slightly above the average of 24.23%. While it – by a narrow margin – places Finnish in Haspelmath's & Tadmor's 'high borrowers' category, we can hardly say that the study lends any strong support for hypothesis B: *Finnish has a higher percentage of loanwords in general compared to other languages.*

However, while the LWT project aimed to design a study which would allow for comparison of different languages from a loanword typological perspective, there is reason to question the comparability of the languages with respect to the coding criteria set up by the project. For instance, the definition of a loanword is narrowed down to only include words borrowed as entities (cf. sections 2.2.2 and 3.2.3) – subsequent derivations from or compounds including borrowed material are excluded by this definition. A consequence of this principle for highly synthetic languages such as Finnish, where word-formation is to a large extent achieved by derivation and compounding, is that the addition of a morpheme to a borrowed word – be the morpheme native or borrowed itself – completely obscures the foreign origin of that word in the statistics. For some of the LWT languages with a higher percentage of loanwords than Finnish, this can hardly be as problematic, as they have a more analytic profile – e.g. Vietnamese, but also English and Indonesian. The typological variation on the synthetic-analytic scale is thus an inherent problem of cross-linguistic lexical comparison.

Another point worth mentioning is the variation in linguistic description between many of the languages compared. The Uralic languages are perhaps the most well studied ones after the Indo-European languages, not least when it comes to etymology, meaning that the Finnish loanword etymologies have been under close scrutiny with strict demands of explanations for e.g. sound substitutions. According to Kallio (2015b:23), this has not been the case for "most contributions" to the Loanword Typology project. As the borrowed status of each loanword is weighted by a reliability scale (cf. Borrowed and Created on loan basis under 3.2.3), it is possible that more well-studied languages have been under more rigid constraints regarding the reliability of the etymologies, due to the increased demand of detailed explanation that goes along with the well-studied state of a language.

5.1 Donor languages

The overwhelming majority of all loanwords in Finnish come from Indo-European languages (91.67%), which, based on earlier research, is expected. Furthermore, almost three quarters of all loanwords are of Germanic origin (70.63%). These Germanic loanwords constitute a steady and ever-increasing stream that stretches all the way from Proto-Germanic to modern-day Swedish and English, indicative of a very intensive and long-standing contact situation. 14.38% of the loanwords come from Balto-Slavic languages; the rest of the language groups have contributed less than 5% each.

Sadly, the amount of words reconstructible to each Uralic stage varies a lot (generally, the older the language stage, the less words) – which, at least in part, may be an artefact of the method. In any case, there is good support for hypothesis D: *The donor languages will cluster in (possibly partly overlapping) layers timewise, roughly indicating when the language contact took place*, as shown in 4.5 (cf. Figure 16). Indo-European shows an early peak, before the descendants of Proto-Indo-European start diverging into their own language branches. Indo-Iranian loanwords start showing up a little later, and disappear completely from the graph towards the more recent language stages. The onset of Germanic and Balto-Slavic loanwords then takes over; Balto-Slavic has an early peak followed by a gradual decline, whereas Germanic increases steadily up until the present day. All of these relationships are unsurprising, and in line with the most mainstream theories on the Uralic prehistory – cf. section 2.3.1.

As for hypothesis C, *There is a correlation between donor language and semantic domain*, the most obvious trend is that Germanic has provided between 50% and 100% of the loanwords in nearly every domain, while the other language branches stand for 20% or less of the loanwords in more or less all domains (with a few Balto-Slavic exceptions). With Germanic being so overwhelmingly overrepresented with respect to the other language branches, it is difficult to find any other patterns. This was further obstructed by the fact that many of the semantic domains contained very few loanwords in total. In any case, the two domains where Balto-Slavic stands out – *Spatial relations* (36.36%) and *Quantity* (33.33%) – could perhaps be said to share some kind of similar properties. The domains of *Speech and language* and *Sense perception* are the ones that stand out when it comes to Proto-Indo-European loanwords, and they clearly seem to have properties in common.

The Saami loanwords are too few (two in all) to serve as the basis for any generalisations time- or domain-wise – the only pattern that really emerges is how few they are[19]. Loanwords from Finnish (and its previous Balto-Finnic stage), on the other hand, are much more common in Saami, at least in those Saami languages that are spoken within the national borders of Finland (Aikio 2007:24-25, Rießler 2009). The two loanwords consist of an animal name (Fi. *norsu* 'elephant'), originally denoting an animal native to Saami-speaking regions (Saami **moršę* 'walrus'), as well as a tool used in reindeer herding, traditionally practised by Saami-speakers (Fi. *suopunki* 'lasso' < Saami **suoppęnję* 'lasso'). Without analysing the Finnish loanwords in Saami any further, the two loanwords in the opposite direction would seem to present us with a rather classic case of a socially imbalanced borrowing situation, with Finnish as a superstratum and Saami as a substratum (cf. 2.2.2). There are parallels to be drawn here to the (former) relationship between Finnish and Swedish (cf. 2.1.2), where Finnish-speaking areas were incorporated under a nation-state with a different language for its official functions, leading to countless Swedish loanwords in Finnish (30% of all loanwords in this study, cf. 4.1), while Standard Swedish only contains a handful of loanwords from Finnish[20]. This pattern would seem

[19] In Standard Finnish, that is – in the northern Finnish dialects (i.e. those that are spoken in the vicinity of Saami languages), Saami loanwords are much more frequent (Aikio 2009:40ff.).

[20] Having participated in a more intensive and long-standing language contact with Finnish, Finland-Swedish dialects contain a much larger share of Finnish loanwords than does Standard Swedish (spoken in present-day Sweden).

to paint the picture of Finnish as a substratum language, with Swedish as a superstratum, which is unsurprising given the socio-political history described under 2.1.2 above.

5.2 Semantic domains

As for hypothesis A: *There is a salient difference in the percentage of loanwords between different semantic domains*, the loanword percentages per semantic domain varied from 4.76% (*Function words*) to 50.79% (*The house*)[21], which means that the hypothesis is definitely supported. In relation to this, however, it should be mentioned – again – that the LWT categorisation of the lexemes into semantic domains is somewhat arbitrary: as Häkkinen (1998:190; my translation) states, "a universal and all-encompassing semantic classification model does not exist, and no commonly used categories are mutually exclusive". For example, the item 4.97 'blind' is categorised under *The body*, but could easily have been under *Sense perception*. In addition, the selection of semantic domains could also have been different: for example, the domain *Animals* could have been divided into *Wild animals* and *Domestic animals*, respectively. This being said, the percentual differences between the semantic domains at hand are still large enough to warrant some legitimacy to the claim that there are semantic domains containing meanings which are more easily borrowed than others – both language-specifically (as per the results in this study, cf. 4.2) and cross-linguistically (as per the findings of the LWT project, cf. Tadmor 2009:64-65).

While on the topic of semantic domains, I would like to make a few comparisons between the findings of this study and those of Häkkinen's (1998) study on the reconstructed Uralic lexicon referred to in length under 1.1. Häkkinen's most important results concern the subsistence system of the early Uralic speakers, and she found indications of them having had a hunter-gatherer lifestyle. As this study focuses on general patterns alone, and explicitly refrains from focusing on single lexical items, the central claims of Häkkinen's study are hard to either corroborate or disprove, as the semantic domains in the present study which contain vocabulary relating to agriculture also contain vocabulary which has nothing to do with agriculture: *Animals*

[21] Their respective borrowed score being 0.07 and 0.51.

contains both wild and domesticated fauna, and *Agriculture and vegetation* contains both wild and cultivated flora (Häkkinen's semantic domain classification for her own study was more custom-made and to-the-point). As for social organisation, Häkkinen finds that kinship terminology is overrepresented in the reconstructible Uralic lexicon, whereas terms in the category *Human society* are underrepresented, leading her to hypothesise that family relations were the main basis for the societal structure. As the low reconstructibility of a semantic domain must be related to some form of lexical replacement, it is not unlikely that this replacement has taken place by means of borrowing. If we accept this hypothesis, then Häkkinen's conclusions regarding societal structure are supported by the findings of the present study, as the semantic domain *Social and political relations* contains 34.69% loanwords, while *Kinship* only contains 10.00%[22]. The underlying assumption here has to do with the relation between cultural salience and inheritability (cf. 2.2.3; Epps 2014).

5.3 Where Finnish differs from the average

As for the comparison to the results of the Loanword Typology project (cf. 4.3), roughly half (13) of the 24 semantic domains differ with less than 0.10 in their borrowed score. These domains can thus be said to behave more or less according to the expectations. Among the results that have shown a larger difference, the three domains *Animals*, *The house* and *Clothing and grooming* contain a higher share of loanwords in Finnish than in the cross-linguistic average. Indeed, *Animals* contains both domesticated and wild animals, but at least to some extent, all three categories are related to sedentarism. As Campbell (2013:428) notes (see full citation under 2.3.1 above), most scholars believe that Neolithic agricultural techniques reached Proto-Uralic by contact with groups from the south and had a large influence on the diversification as well as the lifestyle of the early Uralic speakers, contributing among other things to sedentarism. It is

[22] It should be mentioned that kinship terminology is one of the semantic domains generally considered resistant to borrowing. This assumption seems to be supported by the LWT results, too: *Kinship*, with a borrowed score of 0.23, has the 6th lowest borrowed score out of the 24 semantic domains. Despite its low score in the LWT results, the *Kinship* terms in Finnish scored considerably lower, namely 0.11 (2nd lowest borrowed score). The relatively large difference in borrowed score between Finnish and the average LWT results for *Kinship* (compared to the other semantic domains) is the reason it is brought up here.

possible that the language contact leading to this subsistence revolution in the Uralic speech community is reflected in these results.

However, it is important to point out that the *Animals* domain also contains a lot of what Bowern et al (2014:224) call "zoo animals" – indeed, there is no language whose speakers would be familiar with all animals in the LWT list. As Finnish is an official majority language serving all kinds of functions, there are zoology textbooks written in Finnish, which means that there *are* Finnish names for many animals, even though the speakers very rarely have reason to talk about them in their everyday lives. This may not be the case for all languages: smaller minority languages used in more limited situations simply may not have words for animals other than those that are relevant to the speakers. A different approach is used in Bowern et al. (2014), where the list of lexemes is adapted to the geographic areas of the languages studied, thus avoiding the problems mentioned above.

The semantic domains where Finnish has a lower share of loanwords than the average include *Function words, Social and political relations, Speech and language, Modern world, Cognition, Law, Kinship* and *Quantity*. It is harder to find a single, common denominator for these semantic domains. In fact, this is not surprising – if we combine the LWT comparison with the list of borrowed score per semantic domain in Finnish, as in Table 10 below, we notice the following: the three semantic domains that contain a *higher* amount of loanwords in Finnish than the LWT average (marked in bold) are all in the top four on the list over borrowed score per languloanwordsage (making it a rather uniform category), while the eight domains containing a *lower* amount of vocabulary than the LWT average (marked in italics) can be found all over the list, from third place to the very bottom (making this category less homogenous and generalizable). The semantic domains that differ less than 0.1 in borrowed score are marked in grey and strikethrough.

Table 10. The borrowed score of the semantic domains in relation to the LWT comparison.

Semantic domain	Words, total	Loanwords	Loanwords, %	Borrowed score
The house	**63**	**32**	**50.79%**	**0.51**
Clothing and grooming	**79**	**40**	**50.63%**	**0.50**
Modern world	*66*	*32*	*48.48%*	*0.48*
Animals	**136**	**55**	**40.44%**	**0.43**
~~Agriculture and vegetation~~	~~86~~	~~34~~	~~39.53%~~	~~0.40~~
~~Religion and belief~~	~~35~~	~~14~~	~~40.00%~~	~~0.40~~
~~Food and drink~~	~~108~~	~~40~~	~~37.04%~~	~~0.35~~
Social and political relations	*49*	*17*	*34.69%*	*0.34*
~~Basic actions and technology~~	~~104~~	~~30~~	~~28.85%~~	~~0.31~~
~~Warfare and hunting~~	~~52~~	~~14~~	~~26.92%~~	~~0.30~~
~~Possession~~	~~55~~	~~15~~	~~27.27%~~	~~0.30~~
~~The physical world~~	~~93~~	~~19~~	~~20.43%~~	~~0.24~~
Quantity	*42*	*9*	*21.43%*	*0.23*
Law	*28*	*7*	*25.00%*	*0.23*
~~Time~~	~~71~~	~~12~~	~~16.90%~~	~~0.23~~
~~Emotions and values~~	~~65~~	~~14~~	~~21.54%~~	~~0.23~~
~~The body~~	~~199~~	~~35~~	~~17.59%~~	~~0.21~~
~~Motion~~	~~104~~	~~19~~	~~18.27%~~	~~0.20~~
~~Spatial relations~~	~~95~~	~~11~~	~~11.58%~~	~~0.16~~
Speech and language	*49*	*5*	*10.20%*	*0.15*
Cognition	*71*	*10*	*14.08%*	*0.15*
~~Sense perception~~	~~54~~	~~5~~	~~9.26%~~	~~0.12~~
Kinship	*100*	*10*	*10.00%*	*0.11*
Function words	*21*	*1*	*4.76%*	*0.07*

The domain *Function words* contains a great deal of meanings that are often realised as adpositions or case affixes, and perhaps the divergent morphological structure of Finnish (compared to the Indo-European donor languages) could explain its considerably lower borrowability. After all, borrowed items have to be incorporated into the recipient language, and the incorporation of items with more grammatical functions is naturally easier if the languages in contact share a similar morphological structure.

As for *Modern world*, this domain achieved a borrowed score of 0.48 in Finnish, compared to 0.64 in the LWT average. While this is a salient difference, it does not mean that the domain is particularly short of loanwords in Finnish – on the contrary, it is the semantic domain with the third highest amount of loanwords. Despite this, the relatively low borrowability of this domain may have at least part of its explanation in the fact that it contains several Finnish neologisms (both derivations and compounds), such as the following (Häkkinen 2013, Ikola 1985):

Word	**Meaning**	**Origin**	**Year**
elokuva	‘film, movie’	< *elo* ‘life’ + *kuva* ‘picture’	1927
kirje	‘letter’	< *kirja* ‘book’	1844
osoite	‘address’	< *osoittaa* ‘to show, to point’	1836
polkupyörä	‘bicycle’	< *polkea* ‘to pedal’ + *pyörä* ‘wheel’	1880’s-1900’s
puhelin	‘phone’	< *puhua* ‘to talk’	1897
sairaala	‘hospital’	< *sairas* ‘sick’	1860
savuke	‘cigarette’	< *savu* ‘smoke’	1910’s
sähkö	‘electricity’	< *sähähtää* ‘to sizzle’, *säpenöidä* ‘to sparkle’	1845

These neologisms are all “artificial” coinages from the 19th and early 20th centuries, coinciding with the rise of the Finnish nationalistic movement (cf. section 2.1.2 above). McRae (1997:117) writes that during this period, a cardinal principle in the development of Finnish was linguistic purism: “Every effort was made to build neologisms on native Finnish roots”. This was partly motivated as a “democratisation” of the language, making this higher function terminology more transparent to less educated Finnish-speakers (Ikola 1985), but undoubtedly, these puristic endeavours also went hand in hand with the prevailing nationalistic ideals.

Another motivation for the neologisms can be mentioned: as described in section 2.1.2 above, Swedish was the language used in all higher functions in society (except, to a certain extent, in the church and the judicial system) until 1863, when Finnish was promoted to an equal status (something that was achieved gradually over the course of the subsequent 20 years). At this time, however, the entire educational system was still operating in Swedish; the first Finnish-language school was founded in 1858, meaning that the Finnish vocabulary within the fields of education, administration and science was simply non-existent at the time (Ikola 1985). The development of this new vocabulary was thus also driven by an increasing need for new terminology.

A lot of the neologisms were built on dialectal Finnish words with a modified or specified meaning. Ikola (1985) argues that the creation of new words was facilitated by the structural nature of Finnish: on the one hand, its rich derivational morphology and its ability to create new words by compounding allows for the formation of new words; on the other hand, the Finnish phonological system (cf. section 2.1.1) is described as an obstacle for the incorporation of loanwords[24], which would then also serve as motivation for the coining of neologisms. In any case – it is possible that these mechanisms of word formation have been used in Finnish for meanings where other languages, to a greater extent, have actual loanwords.

5.4 Future research

Regarding the contributions of this study to future research, the addition of more languages with comparable material to those already in the Loanword Typology project can contribute to the general understanding of loanword typology, as well as to the fine-tuning of the outcomes of this research project. The project has already shed a great deal of light on aspects of borrowability that were previously founded on a non-scientific ground, and the Leipzig-Jakarta list is a linguistic tool that can help making this area of research more empirically grounded. However, as only a fraction of the world's languages have been covered using this methodology, the addition of more languages would doubtlessly lead to a more complete picture of borrowability and further help us understand this phenomenon, so that we can set up more realistic hypotheses in the future.

From a Finnish-language point of view, this study has also contributed with the restructuring of earlier etymological work into a format that makes it comparable to similar results from other languages.

With regards to lesser-known languages, perhaps the findings of the Loanword Typology project (potentially with the addition of more languages) can help us construct hypothesis-driven studies where our expectations are more in line with reality. The contribution of well-studied languages of which we have a relatively high etymological understanding (such as Finnish) would then be of importance in the revelation of patterns which may hold cross-linguistically, and thus be of help in the investigation of the linguistic history of lesser-known languages. Of course, the findings in one language do not necessarily hold in the case of another, but if the alternative is to have little or nothing to base one's hypotheses on, basing them on attested patterns from another natural language is the better option. For instance, we may assume that the agricultural lexicon of

[24] We do know, however, that despite its phonological structure, Finnish is hardly devoid of loanwords!

languages whose speakers have been prominent in the prehistoric spread of agriculture in a certain geographical area will behave similarly to the agricultural lexicon of other languages who have had a similar function in another geographical area.

Furthermore, as has been discussed throughout this thesis, the classification of meanings into semantic domains, at its current state, leaves a lot to be desired. The semantic domain division used in the Loanword Typology project (and, by extension, in this study) is largely inherited from the IDS, which in turn has inherited it largely from Buck (1949), and this classification is rather coarse and borderline misleading. Admittedly, semantics is inherently tricky to boil down and categorise, but future studies focusing on semantic domain classification would be of great importance to the scientific community and to future lexical research.

6 Conclusion

This quantitative study has investigated various aspects of loanwords in the Finnish lexicon, as well as compared them to earlier findings based on corresponding data from other languages.

Five research questions were formulated:

1. How much of the Finnish vocabulary is borrowed?
2. Which semantic domains have the highest vs. the lowest amount of loanwords?
3. Does the Finnish sample differ in any noticeable way from the cross-linguistic tendencies found in the Loanword Typology project?
4. Is there a correlation between semantic domain and donor language (family)?
5. Is there a correlation between time period and donor language (family)?

Four hypotheses were set up based on earlier research:

A. There is a salient difference in the percentage of loanwords between different semantic domains.
B. Finnish has a higher percentage of loanwords in general compared to other languages.
C. There is a correlation between donor language and semantic domain.
D. The donor languages will cluster in (possibly partly overlapping) layers timewise, roughly indicating when the language contact took place.

The data used to answer these questions was retrieved from etymological dictionaries and quantified according to parameters set by the Loanword Typology project (Haspelmath & Tadmor 2009a). The most central feature of the quantification model was the *Borrowed* field, where information on the borrowing status of each lexeme was coded, as well as an assessment of how certain this information was. Other information that was quantified include the age of each lexeme (represented by its earliest reconstructible stage in Uralic), and (for loanwords) information about the source word and donor language.

A series of quantitative analyses where then performed to answer each of the research questions, the results of which where presented in tables and diagrams. Hypotheses A and D were corroborated by the results; C to a lesser extent, and B was refuted. Some findings of the study worth noting are the following:

- Generally speaking, Finnish is a fairly typical language from a loanword typological point of view according to the results of this study, albeit with a few caveats regarding the comparability in the language sample.
- The overwhelming majority of loanwords in Finnish come from Indo-European, especially from Germanic languages.
- Clearly visible tendencies could be found in support of timewise layers of loanwords from different language branches. These findings are in line with mainstream hypotheses on the prehistory of Uralic-speakers.
- The semantic domains where Finnish deviates the most from the general cross-linguistic patterns of borrowability (as represented by the averages of the Loanword Typology project) in that it has a higher share of loanwords include parts of the lexicon related to sedentarism, albeit with a few caveats regarding the semantic domain classification.

Although the findings of this study are largely in line with the previous research on loanwords in Finnish, the most important contributions of this thesis are related to its data reorganisation and typological comparability: the restructuring of the previous research into a format which makes it comparable to corresponding data in a relatively large sample of languages cross-linguistically, as well as the implications that the results could have for hypothesis formulation in loanword studies on lesser-known languages with a similar political or sociolinguistic history.

References

Aikio, A. (2007). Etymological nativization of loanwords. A case study of Saami and Finnish. In Toivonen, I. & Nelson, D. (Eds.), *Saami Linguistics* (pp. 17-52). Amsterdam/Philadelphia: John Benjamins.

Aikio, A. (2009). *The Saami loanwords in Finnish and Karelian.* (Doctoral dissertation, University of Oulu, Faculty of Humanities).

Angouri, J. (2010). Quantitative, Qualitative or Both? Combining Methods in Linguistic Research. In Litosseliti, L. (Ed.), *Research Methods in Linguistics* (pp. 29-45). London: Continuum.

Bowern, C., Haynie, H., Sheard, C., Alpher, B., Epps, P., Hill, J. & McConvell, P. (2014). Loan and Inheritance Patterns in Hunter-Gatherer Ethnobiological Systems. In *Journal of Ethnobiology*, 34 (2) (pp. 195-227).

Buck, C.D. (1949). *A dictionary of selected synonyms in the principal Indo-European languages: a contribution to the history of ideas*. Chicago: Univ. of Chicago Press.

Campbell, L. (2013). *Historical linguistics: an introduction*. (3rd ed.). Edinburgh: Edinburgh University Press.

Campbell, L. & Mixco, M.J. (2007). *A glossary of historical linguistics*. Edinburgh: Edinburgh University Press.

Carpelan, C. & Parpola, A. (2001). Emergence, contacts and dispersal of Proto-Indo-European, Proto-Uralic and Proto-Aryan in archaeological perspective. In Carpelan, C., Parpola, A. and Koskikallio, P. (Eds.), *Early contacts between Uralic and Indo-European: linguistic and archaeological considerations*. Papers presented at an international symposium held at the Tvärminne Research Station of the University of Helsinki, 8-10 January, 1999 (pp. 55-150). Helsinki: Suomalais-Ugrilaisen Seuran Toimituksia 242.

Carpelan, C., Parpola, A. and Koskikallio, P. (Eds.) (2001). *Early contacts between Uralic and Indo-European: linguistic and archaeological considerations.* Papers presented at an international symposium held at the Tvärminne Research Station of the University of Helsinki, 8-10 January, 1999. Helsinki: Suomalais-Ugrilaisen Seuran Toimituksia 242.

Constitution Act of Finland. (1919, July 17). Consulted online January 2018: <https://www.finlex.fi/en/laki/kaannokset/1999/en19990731.pdf>

Décsy, G. (1965). *Einführung in die finnisch-ugrische Sprachwissenschaft*. Wiesbaden: Harrassowitz.

Derksen, R. (2010). *Etymological Dictionary of the Baltic Inherited Lexicon: Introduction*. Consulted online October 2017: < http://dictionaries.brillonline.com/pdfdocument/baltic/introduction>

Duncan, L. (2008). *Vowel Quality in Finnish Loanwords of Swedish Origin: An Acoustic Study.* University of Toronto, Department of Linguistics.

Epps, P. (2014). Historical linguistics and socio-cultural reconstruction. In Bowern, C. & Evans, B. (Eds.), *The Routledge Handbook of Historical Linguistics* (pp. 579-597). New York: Routledge.

Hajdú, P. (1975). *Finno-Ugrian languages and peoples*. London: Deutsch.

Halonen, M., Ihalainen, P. & Saarinen, T. (Eds.) (2014). *Language policies In Finland and Sweden: interdisciplinary and multi-sited comparisons.* Bristol: Multilingual Matters.

Hammarström, H., Bank, S., Forkel, R. & Haspelmath, M. (2018). *Glottolog 3.2*. Jena: Max Planck Institute for the Science of Human History. Consulted online October 2017: <http://glottolog.org/resource/languoid/id/ural1272.bigmap.html#3/62.80/58.28>

Haspelmath, M. (2002). *Understanding morphology*. London: Arnold.

Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. In Haspelmath, M. & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook* (pp. 35-54). Berlin: De Gruyter Mouton.

Haspelmath, M. & Tadmor, U. (Eds.) (2009a). *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: De Gruyter Mouton.

Haspelmath, M. & Tadmor, U. (2009b). The Loanword Typology project and the World Loanword Database. In Haspelmath, M. & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook* (pp. 1-34). Berlin: De Gruyter Mouton.

Haspelmath, M. & Tadmor, U. (2009c) *The World Loanword Database*. Retrieved from: <http://wold.clld.org/>

Hock, H.H. & Joseph, B.D. (1996). *Language History, Language Change, and Language Relationship*. Berlin: Mouton de Gruyter.

Hurme, R., Pesonen, R. & Syväoja, O. (2003). *Englanti-suomi-suursanakirja* [English-Finnish general dictionar*y*]. Helsinki: WSOY.

Häkkinen, J. (2009). *Uralilaisen sukupuun kehitys* [The development of the Uralic family tree]. University of Helsinki. Retrieved from: < http://www.elisanet.fi/alkupera/Sukupuu.pdf>

Häkkinen, K. (1984). Wäre es schon an der Zeit, den Stammbaum zu fällen? In *Ural-Altaische Jahrbücher, Neue Folge 4* (pp. 1–24). (In Häkkinen, J. (2009). *Uralilaisen sukupuun kehitys* [The development of the Uralic family tree]. University of Helsinki. Retrieved from: < http://www.elisanet.fi/alkupera/Sukupuu.pdf>)

Häkkinen, K. (1998). Uralilainen muinaiskulttuuri sanahistorian valossa [Prehistoric Uralic culture in the light of lexical history]. In Grünthal, R. & Laakso, J. (Eds.), *Oekeeta asijoo. Commentationes Fenno-Ugricae in honorem Seppo Suhonen sexagenarii* (pp. 188-194). Helsinki: Mémoires de la Société Finno-Ougrienne 228.

Häkkinen, K. (2001). Prehistoric Finno-Ugric culture in the light of historical lexicology. In Carpelan, C., Parpola, A. and Koskikallio, P. (Eds.), *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations.* Papers presented at an international symposium held at the Tvärminne Research Station of the University of Helsinki, 8-10 January, 1999 (pp. 169-186). Helsinki: Suomalais-Ugrilaisen Seuran Toimituksia 242.

Häkkinen, K. (2013). *Nykysuomen etymologinen sanakirja* [Etymological dictionary of Contemporary Finnish]. (6th ed.). Helsinki: WSOY.

Häkkinen, K., & Lempiäinen, T. (1996). Die ältesten Getreidepflanzen der Finnen und ihre Namen. In *Finnisch-ugrische Forschungen: Zeitschrift für finnisch-ugrische Sprach-und Volkskunde,* Vol. 53 (pp. 115-182).

Ikola, O. (1985). Ordbildning på inhemsk bas i finskan. In *Språk i Norden* (pp. 13-19).Itkonen, E. (Ed.) (1992-2000). *Suomen sanojen alkuperä: etymologinen sanakirja* [The origin of Finnish words: an etymological dictionary]. Helsinki: Suomalaisen kirjallisuuden seura.

Janhunen, J. (2009). Proto-Uralic – what, where, and when? In Ylikoski, J. (Ed.), *The quasquicentennial of the Finno-Ugrian Society* (pp. 57-78). Helsinki: Société Finno-Ougrienne.

Kallio, P. (1995). Suomen kielen kivikautisista lainasanakerrostumista [On the Stone Age Loanword Strata in Finnish]. In *Virittäjä,* Vol. 99, Iss. 3 (pp. 380-389). Helsinki: Kotikielen Seura.

Kallio, P. (1997). Uralilaisten alkuperä indoeuropeistisesta näkökulmasta [The origin of the Uralics from an Indo-European poit of view]. In *Virittäjä,* Vol. 101, Iss. 1 (pp. 74-78). Helsinki: Kotikielen Seura.

Kallio, P. (1998). Vanhojen balttilaisten lainasanojen ajoittamisesta [On the dating of old Baltic loanwords]. In Grünthal, R. & Laakso, J. (Eds.), *Oekeeta asijoo: Commentationes Fenno-Ugricae in honorem Seppo Suhonen sexagenarii 16.V.1998* (pp. 209-217). Helsinki: Mémoires de la Société Finno-Ougrienne 228.

Kallio, P. (1999). Varhaiset indoeurooppalaiskontaktit [Early Indo-European contacts]. In Fogelberg, P. (Ed.), *Pohjan poluilla: Suomalaisten juuret nykytutkimuksen mukaan* [On the paths of the North: The roots of the Finns in the light of present day research] (pp. 237-239). Helsinki: Bidrag till kännedom av Finlands natur och folk 153.

Kallio, P. (2002). Prehistoric Contacts between Indo-European and Uralic. In Jones-Bley, K., Huld, M.E., Della Volpe, A., & Robbins Dexter, M. (Eds.), *Proceedings of the Thirteenth Annual UCLA Indo-European Conference* (pp. 29-44). Washington, DC: The Journal of Indo-European Studies Monograph Series 44.

Kallio, P. (2006). On the Earliest Slavic Loanwords in Finnic. In Nuorluoto, J. (Ed.), *The Slavicization of the Russian North: Mechanisms and Chronology* (pp. 154-166). Helsinki: Slavica Helsingiensia 27.

Kallio, P. (2008). On the 'Early Baltic' Loanwords in Common Finnic. In Lubotsky, A., Schaeken, J. & Wiedenhof, J. (Eds.), *Evidence and Counter-Evidence: Essays in Honour of Frederik Kortlandt. Vol. 1: Balto-Slavic and Indo-European Linguistics* (pp. 265-277). Amsterdam/New York: Studies in Slavic and General Linguistics 32.

Kallio, P. (2012). The Prehistoric Germanic Loanword Strata in Finnic. In Grünthal, R. & Kallio, P. (Eds.), *A Linguistic Map of Prehistoric Northern Europe* (pp. 225-238). Helsinki: Mémoires de la Société Finno-Ougrienne 266.

Kallio, P. (2015a). The Language Contact Situation in Prehistoric Northeastern Europe. In Mailhammer, R., Vennemann, T. & Olsen, B.A. (Eds.), *The Linguistic Roots of Europe: Origin and Development of European Languages* (pp. 77-102). Copenhagen: Copenhagen Studies in Indo-European 6.

Kallio, P. (2015b). The Stratigraphy of the Germanic Loanwords in Finnic. In Askedal, J.O. & Nielsen, H.F. (Eds.), *Early Germanic Languages in Contact* (pp. 23-38). Amsterdam/Philadelphia: North-Western European Language Evolution Supplement Series 27.

Kallio, P. (2017). The Indo-Europeans and the Non-Indo-Europeans in Prehistoric Northern Europe. In Hyllested, A., Nielsen Whitehead, B., Olander, T. & Olsen, B.A. (Eds.), *Language and Prehistory of the Indo-European Peoples: A Cross-Disciplinary Perspective* (pp. 187-203). Copenhagen: Copenhagen Studies in Indo-European 7.

Karlsson, F. (1999). *Finnish: An essential grammar*. New York: Routledge.

Key, M.R. & Comrie, B. (Eds.) (2015). *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Consulted online January 2018: <http://ids.clld.org>

Koivulehto, J. (1976). Vanhimmista germaanisista lainakosketuksista ja niiden ikäämisestä [On the oldest Germanic loan contacts and their dating]. In *Virittäjä,* Vol. 80, Iss. 1 (pp. 33-47). Helsinki: Kotikielen Seura.

Koivulehto, J. (2006). Wie alt sind die Kontakte zwischen Finnisch-Ugrisch und Balto-Slavisch? In Nuorluoto, J. (Ed.), *The Slavicization of the Russian North: Mechanisms and Chronology* (pp. 179-196). Helsinki: Slavica Helsingiensia 27.

Koivulehto, J. (2007). Auf der Suche nach germanischen Elementen im heutigen Wortschatz des Ostseefinnischen und Saamischen: Veröffentlicht anlässlich des 120-jährigen Bestehens des Neuphilologischen Vereins. In *Neuphilologische Mitteilungen* (pp. 577-589).

Korhonen, M. (1981). *Johdatus lapin kielen historiaan* [A guide to the history of the Lapp language]. SKS:n toimituksia 370. Helsinki: Suomalaisen Kirjallisuuden Seura.

Laakso, J. (2001). The Finnic languages. In Dahl, Ö. & Koptjevskaja-Tamm, M. (Eds.), *The Circum-Baltic languages: typology and contact. Vol. 1: Past and present* (pp. 179-212). Amsterdam/Philadelphia: John Benjamins.

Laakso, J. (2011). The Uralic languages. In Kortmann, B., & van der Auwera, J. (Eds.), *The Languages and Linguistics of Europe: A Comprehensive Guide* (pp. 179-197). Berlin: de Gruyter Mouton.

Latomaa, S. & Nuolijärvi, P. (2002). The Language Situation in Finland. In *Current Issues in Language Planning,* Vol. 3, Iss. 2 (pp. 95-202).

Mallory, J.P. & Adams, D.Q. (Eds.) (1997). *Encyclopedia of Indo-European culture*. London: Fitzroy Dearborn.

Mallory, J.P. & Adams, D.Q. (2006). *The Oxford introduction to Proto-Indo-European and The Proto-Indo-European world*. Oxford: Oxford University Press.

McRae, K.D. (1997). *Conflict and compromise in multilingual societies.* Vol. 3: Finland. Waterloo, Ontario: Wilfrid Laurier University Press.

Rédei, K. (1986-1991). *Uralisches etymologisches Wörterbuch*. Wiesbaden: Harrassowitz.

Rießler, M. (2009). Loanwords in Kildin Saami, a Uralic language of northern Europe. In Haspelmath, M. & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook* (pp. 384-416). Berlin: De Gruyter Mouton.

Saeed, J.I. (2009). *Semantics*. (3rd ed.). Malden, Massachusetts: Wiley-Blackwell.

Salminen, T. (1999). Euroopan kielet muinoin ja nykyisin [The languages of Europe in the past and today]. In Fogelberg, P. (Ed.), *Pohjan poluilla: Suomalaisten juuret nykytutkimuksen mukaan* [On the paths of the North: The roots of the Finns in the light of present day research] (pp. 13-26). Helsinki: Bidrag till kännedom av Finlands natur och folk 153.

Salminen, T. (2001). The rise of the Finno-Ugric language family. In Carpelan, C., Parpola, A. and Koskikallio, P. (Eds.), *Early contacts between Uralic and Indo-European: linguistic and archaeological considerations* (pp. 385-396). Papers presented at an international symposium held at the Tvärminne Research Station of the University of Helsinki, 8-10 January, 1999. Helsinki: Suomalais-Ugrilaisen Seuran Toimituksia 242.

Salminen, T. (2002). Problems in the taxonomy of the Uralic languages in the light of modern comparative studies. In *Лингвистический беспредел: сборник статей к 70-летию А. И. Кузнецовой* [Linguistic lawlessness: a collection of articles dedicated to the 70th birthday of A.I. Kuznetsova] (pp. 44–55). Moscow: Moscow University Publishing house. Retrieved from: <http://www.helsinki.fi/~tasalmin/kuzn.html>

Suomen virallinen tilasto [Official statistics of Finland] (2017, September 22). Väestörakenne [Population structure] (online publication). In *Vuosikatsaus 2016* [Annual report of 2016]. Helsinki: Tilastokeskus. Consulted online January 2018: <http://www.stat.fi/til/vaerak/2016/01/vaerak_2016_01_2017-09-22_tie_001_fi.html>

Suomi, K., Toivanen, J. & Ylitalo, R. (2008). *Finnish sound structure: Phonetics, Phonology, Phonotactics and Prosody*. Oulu: University of Oulu.

Swadesh, M. (1950). Salish Internal Relationships. In *International Journal of American Linguistics,* Vol. 16, No. 4 (pp. 157-167). The University of Chicago Press.

Tadmor, U. (2009). Loanwords in the World's Languages: Findings and results. In Haspelmath, M. & Tadmor, U. (Eds.), *Loanwords in the World's Languages: A Comparative Handbook* (pp. 55-75). Berlin: De Gruyter Mouton.

Sources to etymological data

Aikio, A. (2009). *The Saami loanwords in Finnish and Karelian.* (Doctoral dissertation, University of Oulu, Faculty of Humanities).

Bartholomae, C. (1979 [1904]). *Altiranisches Wörterbuch: zusammen mit den Nacharbeiten und Vorarbeiten*. (2. Nachdr.) Berlin: Gruyter.

Beekes, R.S.P. (2010). *Etymological Dictionary of Greek*. Consulted online October 2017: <http://dictionaries.brillonline.com/search#dictionary=greek&id=gr1986>

Chan, E. (2006-2015). *Numeral Systems of the World's Languages*. Consulted online October 2017: <https://mpi-lingweb.shh.mpg.de/numeral/>

Cheung, J. (2010). *Etymological Dictionary of the Iranian Verb*. Consulted online October 2017: <http://dictionaries.brillonline.com/search#dictionary=iranian&id=iv0022>

Derksen, R. (2010a). *Etymological Dictionary of the Baltic Inherited Lexicon*. Consulted online October 2017: <http://dictionaries.brillonline.com/search#dictionary=baltic&id=blt0001>

Derksen, R. (2010b). *Etymological Dictionary of the Slavic Inherited Lexicon*. Consulted online October 2017: <http://dictionaries.brillonline.com/search#dictionary=slavic&id=ps0001>

Det Danske Sprog- og Litteraturselskab: *Den Danske Ordbog*. Retrieved from: <http://ordnet.dk/ddo>

Hakulinen, L. (1968). *Suomen kielen rakenne ja kehitys* [The structure and development of the Finnish language]. Helsinki: Otava.

Hellquist, E. (1922). *Svensk etymologisk ordbok*. Lund: Gleerup.

Häkkinen, K. (2013). *Nykysuomen etymologinen sanakirja* [Etymological dictionary of Contemporary Finnish]. (6th ed.). Helsinki: WSOY.

Itkonen, E. (Ed.) (1992-2000). *Suomen sanojen alkuperä: etymologinen sanakirja* [The origin of Finnish words: an etymological dictionary]. Helsinki: Suomalaisen kirjallisuuden seura.

Kallio, P. (2006). On the Earliest Slavic Loanwords in Finnic. In Nuorluoto, J. (Ed.), *The Slavicization of the Russian North: Mechanisms and Chronology* (pp. 154-166). Helsinki: Slavica Helsingiensia 27.

Korhonen, M. (1981). *Johdatus lapin kielen historiaan* [A guide to the history of the Lapp language]. SKS:n toimituksia 370. Helsinki: Suomalaisen Kirjallisuuden Seura.

Koukkunen, K. (1990). *Nykysuomen sanakirja 8: Vierassanojen etymologinen sanakirja* [Dictionary of Contemporary Finnish 8: Etymological dictionary of foreign words]. (2nd ed.). Porvoo: WSOY:n graafiset laitokset.

Kroonen, G. (2010). *Etymological Dictionary of Proto-Germanic*. Consulted online October 2017: <http://dictionaries.brillonline.com/search#dictionary=proto_germanic&id=pg0001>

Laakso, J. (2001). The Finnic languages. In Dahl, Ö. & Koptjevskaja-Tamm, M. (Eds.), *The Circum-Baltic languages: typology and contact*. Vol. 1: Past and present (pp. 179-212). Amsterdam/Philadelphia: John Benjamins.

Mallory, J.P. & Adams, D.Q. (Eds.) (1997). *Encyclopedia of Indo-European culture*. London: Fitzroy Dearborn.

Mayrhofer, M. (1986-2001). *Etymologisches Wörterbuch des Altindoarischen*. Heidelberg: Winter.

Sadeniemi, M. (Ed.) (1979). *Nykysuomen sanakirja 4: Vierasperäiset sanat* [Dictionary of Contemporary Finnish 4: Words of foreign origin]. (5th ed.). Porvoo: WSOY:n graafiset laitokset.

Plöger, A. (1973). *Die russischen Lehnwörter der finnischen Schriftsprache*. Wiesbaden: Veröffentlichungen der Societas Uralo-Altaica 8.

Rédei, K. (1986-1991). *Uralisches etymologisches Wörterbuch*. Wiesbaden: Harrassowitz.

Ringe, D.A. (2006). *From Proto-Indo-European to Proto-Germanic*. Oxford: Oxford University Press.

Svenska Akademien (1893–). *Ordbok över svenska språket*. Lund. Consulted online October 2017: <www.saob.se>