SEMI-AUTOMATED METHODS OF DIRECT ANGLICISM IDENTIFICATION IN FINNISH CORPORA

Nina Mikušová
Master's Thesis
Language Technology
Department of Digital Humanities
Faculty of Arts
University of Helsinki
November 2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty
Faculty of Arts/Department of Digital Humanities

Tekijä – Författare – Author
Nina Mikušová

Työn nimi – Arbetets titel – Title
Semi-Automated Methods of Direct Anglicisms Identification in Finnish Corpora

Oppiaine – Läroämne – Subject
Language Technology

Työn laji – Arbetets art – Level Master's Thesis	Aika – Datum – Month and year November 2020	Sivumäärä– Sidoantal – Number of pages 60

Tiivistelmä – Referat – Abstract

The goal of this thesis is to investigate methods that could help with harvesting neologisms and more specifically anglicisms (i.e. English-sourced borrowings) in Finnish language. The work is partially motivated by the Global Anglicism Database project to gather anglicisms from various languages, which can serve both as an anglicism dictionary and researchers as a source of information for studying language contact and borrowing either in depth for a specific language or cross-linguistically.

A systematic way of harvesting anglicisms in current Finnish language from a suitable corpus is devised. The research examines what kinds of data sources suitable for this goal are available, and what would be the criteria for a useful data source; how to use a data source like that to prepare a good list of anglicisms candidates so that there would be as little irrelevant material as possible but so that no anglicisms would not be lost in the process, and how could the candidates be scored so that the more probable anglicisms would appear closer to the top of a candidate list.

Several of Language Bank's Finnish language monolingual corpora are considered. The most important criteria are identified to be the size and genre of the corpus and its annotation. The criteria are explored from the description of corpora on Language Bank's website and available literature and by hands-on examination of the data. Other important measures of corpus suitability are the amount of unannotated foreign language material, amount of noise, and potential anglicism proportion in the corpora. This information is gained via meticulous exploration of random samples of the corpora neologism candidate lists and evaluation on previously gained anglicism set. A combination of two corpora with good coverage of known anglicisms and relatively low amount of noise is chosen as the dataset for the next phase of the anglicism identification process.

Anglicism candidate lists are prepared by a process of removing tokens irrelevant for anglicism harvesting. That includes an identifiable part of foreign language material in the corpus, formally recognizable noise, known lemmas of the words that were present in Finnish language around the time just before the major influx of English borrowings to Finnish language started, and their inflected forms.

Several methods of scoring candidates are devised that would assign better scores to tokens with higher probability to be an anglicism. The score is based on tokens' frequency in the corpus and relative frequency of the character-level n-grams made out of tokens in representative purely English and purely Finnish corpora. The tokens in the candidate list are scored and ordered, and the resulting list is evaluated based on the ranking of a set of previously identified anglicisms. The method is proved to be somewhat effective; the resulting average ranking of known anglicisms is better than it would be in a randomly sorted candidate list.

Avainsanat – Nyckelord – Keywords
anglicism, neologism, corpus, direct anglicism, anglicism harvesting, internationalism, character level n-gram, Language Bank of Finland, GLAD network

Säilytyspaikka – Förvaringställe – Where deposited
City Centre Campus Library

Muita tietoja – Övriga uppgifter – Additional information

INDEX

1. INTRODUCTION

The goal of this thesis is to investigate methods that could help with harvesting neologisms and more specifically anglicisms (i.e. English-sourced borrowings) in contemporary Finnish language. I became interested in the topic while working on a pilot project for the Global Anglicism Database (GLAD) network. It is a network of scholars and researchers that aims at studying anglicization of world languages and, among other goals, creating an online global database of anglicisms. The database contains anglicisms from various languages, organized by their respective English etymons[1]. The idea is that the database can serve both end users as an anglicism dictionary and researchers as a source of information for studying language contact and borrowing either in depth for a specific language or cross-linguistically. (Gottlieb et al. 2018: 5, 16-18)

A database like this and the identification of anglicisms in general are relevant for the field of linguistics in many ways. The information can be important for language authorities and enhance lexicographical work by giving researchers a direct access to words that may be relevant for inclusion in dictionaries, as well as knowledge about their origin. It can give linguists studying language contact information that can contribute to quantifying the impact of English, a global language par excellence, on the lexicon of a small language like Finnish. In language technology there is use for it for example in text-to-speech synthesis, because a word's origin can affect its pronunciation, or for term extraction in the identification of new terminology in different domains (Andersen 2005, ch. 3.1).

The GLAD database is publicly accessible[2], but it is still in early stages. As of the end of 2019 it included contributions from 16 languages, varying between 30 and 6000 anglicisms per language. The project started with a pilot phase in which a contributor for each language was supposed to provide the list of anglicisms in their language the etymon of which starts with the letter O, but many contributors have since progressed to further phases and provided more anglicisms to the database. (Gottlieb 2019)

I was hired as a contributor for the pilot phase of the project. The task consisted of finding as many Finnish anglicisms in scope of the project as possible, categorizing them and adding information such as first attestation, frequency or usage example about each anglicism (Gottlieb et al. 2018: 14).

[1] etymon: a word that is the source of a particular loanword

[2] available at https://www.nhh.no/en/research-centres/global-anglicism-database-network/resources/

During my work for the GLAD project I came to notice that for the Finnish language there was no ready good source to go for anglicisms and there appears to be no systematic effort in this direction from the Institute for the Languages of Finland at this time. In search for anglicisms I first used some obvious sources (Kotimaisten kielten tutkimuskeskus 1979; Görlach (ed.) 2001; Eronen 2007; relevant articles in Kielikello) of potential candidates for the database. The problem was that while there were good and comprehensive sources like Uudissanasto 1980 and Dictionary of English Anglicisms, they have become outdated, and the other, newer sources were not comprehensive and contained rather just examples or words topical at the moment of publishing. A systematic way of harvesting anglicisms from large corpora of text seemed like a good option, as compared to less systematic and more manual work requiring methods. For this purpose I devised a simple method with the idea that it could sift through data in search of neologisms quite effectively, especially if it used a well suited corpus as its data source.

This method turned out to be not at all as successful as I had imagined (the method and its shortcomings are described in the chapter 4.1.). The pilot phase of the project had to be finished nevertheless, so I went through the data, pre-filtered by the method, manually. That allowed me to gather a number of additional O-starting anglicisms that I categorized, and which were then submitted to the database. I realized that those anglicisms could be used as evaluation data for potential more advanced automatic methods that I could come up with were I to continue with the work. I am not employed by the project anymore, but out of interest I decided to make the effort of improving the efficiency of search for anglicisms my master's thesis project.

In this thesis I try to find a systematic way of harvesting anglicisms in current Finnish language from a suitable corpus. The expected result is a candidate list of potential anglicisms created from the word forms contained in the corpus, which would be narrowed down enough so that going through them manually is not too demanding. The thesis aims to answer the following questions:

- What kinds of data sources suitable for this goal are available, and what would be the criteria for a useful data source?
- How to use a data source like this to prepare a good list of anglicisms candidates so that there would be as little irrelevant material as possible but so that no anglicisms would not be lost in the process?
- How could the candidates be scored so that the more probable anglicisms would appear closer to the top of the list?

1.1. Structure of this work

The next chapter focuses on explaining some key concepts, looks at previous research and takes up other elements that are important to take into consideration when thinking about the scope of this work. Chapter 3 looks at the available data sources and describes their specifics. Chapter 4 explains the idea behind the original method and the process of creating neologism candidate lists from available datasets up to the point of choosing the most suitable source. Chapter 5 then shows a method of scoring the candidates by their probability to be an anglicism. Chapter 6 contains the discussion of the results, suggestion how to work with them effectively, and a reflection on how the results could be improved with additional resources.

2. Background

2.1. Terminology and historical background

In this chapter I present the definitions and limitations of the terms anglicism and neologism and some other related concepts relevant to this work. Even though anglicisms are the core of this work, neologisms cannot be left out. Anglicisms are basically just a special case of neologisms, so they have many characteristics in common, and in the last couple of decades most borrowed neologisms in Finnish have been coming from English (Battarbee 2002: ch. 14.1.1; Eronen 2007: 28). Some of the methods used in this work use to their benefit English graphotactics and thus will be more effective in search for anglicisms, but others are unspecific regarding the origin of the neologisms. That is not a problem per se, as anglicisms can be then identified from among other neologisms manually.

It would seem obvious that when explaining terms, one should start from the more general case, or the hypernym, and make their way towards the narrower term. But in this case, where to start is not completely obvious; if anglicisms are a special case of neologism, all anglicisms should also be neologisms in the given language. But let's consider the English word *to realize* (originally a borrowing from French) that has meanings roughly in the sense "to make something happen" and "to become aware of". In French, the verb *réaliser* used to have only the former meaning, but under the influence of English it also acquired the latter (Onysko 2007: 19). That would mean that the French word *réaliser* with its latter meaning is an anglicism, but is it also a neologism in French? It appears it all comes down to the definitions we decide to work with.

2.1.1. Neologisms

The term "neologism" in its most narrow meaning, also called "pure neologism", means a word made up from scratch, like for example *quark* (meaning the elementary particle) (Geeraerts 2015). But this type of neologisms is rare and for this work a wider definition will be applied.

In his essay, Alain Rey (2005) works out his definition of neologism through exploring the limits of the term. He starts from the widest definition and then explores all three components of it – linguistic unit, language, and newness. The "word" or more exactly the linguistic unit in question is better specified as follows: *"It is always a sign, not a phoneme, […], it cannot be a morpheme as such and certainly not agrammatical word; one may provisionally admit that it is a phrase, however complex, but certainly not a proposition or a sentence."*

Regarding language, Rey (2005) specifies that neologism can belong to language in general or to a more specific area of usage, be it a dialect or some other language variety. Finally, regarding novelty, he brings up Saussure's concept of "signified" and "signifier", and emphasizes that it is not just the "signifier" that has to be new in the lexicon of a given language, but precisely the "signifier"-"signified" combination (therefore the above mentioned example of *réaliser* is a neologism by Rey's definition). Rey also says that novelty depends on the general feeling of the speakers of the language or its variant and must be shared by majority of its speakers.

Rey's definition helps point out what is not a neologism: for example nonce words or children's agrammatical words (limited away by the condition of collective feeling about the neologism), code-switching (which will be mentioned later in chapter 2.2.2; it is usually filtered out by the definition of a linguistic unit), words formed by morphological processes (derivation, composition) from existing words (also filtered out by the relevant linguistic unit definition).

From the diachronic point of view the concept of a neologism is irrelevant, because every linguistic unit in the language is a neologism at one point (Rey 2005: 71-74). But a chronological aspect is needed here. Every linguistic unit has its own life in which it is introduced to a language, it is accepted and expanding into more language varieties or unused and forgotten, and at some point might or might not be integrated "officially", which in our case would probably mean that it is included to official materials published by a language authority[3], e.g. a codifying dictionary.

[3] In Finland and at this time, that would probably mean inclusion in Kielitoimiston sanakirja https://www.kielitoimistonsanakirja.fi/ managed and updated by The Institute for the Languages of Finland.

According to Rey's (2005) definition, a linguistic unit starts being a neologism when it is generally accepted by the majority of the language's speakers, but when does it stop being one?

There does not seem to be a consensus about the answer to that question; e.g. Rey (2005) does not even mention this aspect in his otherwise very detailed essay. One could argue that when a word is fully integrated into a language, it ceases being a neologism. On the other hand, even after full integration words can have a "novelty" feel about them, so maybe a neologism stays a neologism for as long as a majority of speakers can identify it as one. Further research in this direction would be interesting.

Neologisms can be created in many different ways; e.g. the Oxford Companion to the English Language (McArthur et al. 2018) in its encyclopaedical entry on neologism cites the following categories: compounding, derivation, abbreviation, back-formation, blending, shifting meaning, extension in grammatical function, borrowing and coinage (this article does not mention words with unknown origin, although they also deserve their own category). Most of these processes work with language-internal material: neologisms created by modifying or connecting existing words in different ways (compounding, derivation, abbreviation, back-formation and blending) or neologisms where the form is left untouched but their function is extended (shifting meaning and extension in grammatical function). One category, coinage, is equal to the "pure neologism" mentioned above. Just one category, borrowing, is dedicated to neologisms coming from language-external material.

How big part of neologisms are borrowings will differ by language and time of measuring; for example Algeo (1991) reports that only 3 % of neologisms in American English were borrowings at the time – which makes sense when we know that English has been a supplier of borrowing, rather than a receptor, during this and the last century, while (and these is a relatively random reference, but this was somehow a difficult fact to find about any language) e.g. Sijens (2004: 274, referred in Sijens & Van de Velde 2020) reports that around 20 % of Dutch neologisms in the beginning of this millennium were borrowings. Regarding Finnish, Haarala & Nissinen (1994) report about 25 % of neologism as words with foreign elements. I'd argue that it is logical that the number of borrowings in Finnish has grown since 1990s and also that the proportion of borrowings among neologism would be even larger if it were not for the high number of compounds in Finnish language that take up a substantial part of new words, compared to other, less synthetic languages.

2.1.2. Anglicisms

Anglicisms are the neologisms that in some way originate from English language. Formal definitions differ according to the model that their authors use for their research. Görlach (2003: 1)'s definition is handy for lexicographic purposes as it defines anglicism as a word or idiom that is in some way recognizably English but is accepted in the receptor language. Then again Pulcini et al. (2012: 5) points out that the "Englishness" in anglicisms might be obscured, e.g. in calques. A broader definition is offered e.g. by Onysko (2007), who draws on Frans van Coetsem's theory of borrowing in contact between source and receptor languages and defines anglicism as an "umbrella term that covers any instance of transmission from English to a receptor language" (2007: 89). Pulcini et al. (2012: 5) also say that the definition of anglicism is flexible and can be tailored to a specific research.

2.1.3. Anglicisms classification

Anglicisms can be classified in many ways according to different scopes and interests, but I will directly reproduce here (with Finnish examples drawn from Sajavaara (1989)) the typology from Pulcini et al. (2012), because it encompasses a wide range of anglicism types and includes (for this work) important division to direct and indirect borrowings. In the next subchapter I will then dedicate attention to classification of anglicisms by their degree of integration to the receptor language.

- phrasal borrowings: usually multi-word units or whole phrases
    - for example collocations, idioms, catch phrases, routine formulas, proverbs
- lexical borrowings: words or multi-word units
    - direct: formal evidence of the source language is detectable
        - loanword: borrowed from source language; meaning in receptor language is close to meaning in source language
            - non-adapted: none or minimal semantic integration; phonological integration is possible; e.g. *show*
            - adapted: orthographic and/or morphosyntactic integration into the receptor language structures; e.g. *buumi* (Eng. *boom*)
        - false anglicism or pseudo-loan: made up from source language elements, but unknown or with different meaning than it had it the source language; e.g. *sitteri* (from Eng. *babysitter* but meaning *baby bouncer*; this example is from Käenmäki 2019: 26)
        - hybrid: combination of source and receptor language elements; e.g. *cocktailkutsut* (Eng. *cocktail party*)
    - indirect: the source language model is reproduced in the receptor language through native elements
        - calque
            - loan translation: translation of source language item into receptor language; e.g. *kehonrakennus* (Eng. *bodybuilding*)

    - loan rendition: compound or multi-word unit, one part of which is translated from source language and the other is a loose equivalent of the source language part; e.g. *vesipatja* (Eng. *waterbed*)
    - loan creation: receptor language freely renders the source language equivalent; e.g. *turvatyyny* (Eng. *airbag*)
  - semantic loan: an already existing item in the receptor language takes a new meaning after a source language one; e.g. *hiiri* (Eng. *mouse*, meaning mouse as a computer device)

(classification reproduced from Pulcini et al. (2012: 6-8, 13), with examples from Sajavaara (1989: 85-6))

Anglicisms are quite a heterogenous group, but they can be divided into three more homogenous categories by their "harvestability". The group of anglicisms that this thesis concentrates on consists of direct lexical borrowings (including pseudo-loans and hybrids). This group of anglicisms by definition keeps formal evidence of its origin, albeit in different quantity (completely in direct unadapted borrowings, less or little in adapted). This formal evidence is the feature that can be used to identify such anglicisms, especially in languages like Finnish which is not related to English and up until the 20th century had not been in close contact with it, so any shared lexicon except for very old common loans and internationalisms (see chapter 2.2.1.) has to come from relatively recent borrowing processes.

Phrasal anglicisms are ready-made phraseological units that common speaker will not recognize as foreign because they are either translated or reproduced with language-inherent material (Pulcini 2012: 13-14). They are multi-word expressions, a fact that can be used for their identification. For example Lyse and Andersen (2012) use an approach that employs collocation strength of phraseological neologisms, and uses what they call association measures based on frequency and statistical measure of association between tokens of n-grams in corpora to find suitable candidates for phrasal anglicisms.

Indirect lexical borrowings (calques and semantic loans) are the most difficult group of anglicisms to be harvested. They do not have any formal evidence of their origin and they are also single words and not multi-word expression, so that the above-mentioned methods will not help with their identification. They can be found either by manually going through historical/etymological information, or by searching in corpora which contain suitable metadata for checking the use and frequency of such anglicisms in different historical periods or in different text types (Pulcini 2012: 14). However I suppose that such corpus would have to be very large and relatively noise-free for such methods to even have a slight chance to work.

2.1.4. Integration categorization of neologisms and phonetic matching

Direct loanwords, i.e. the borrowings whose formal evidence of the source language is detectable, are integrated into the receptor language in various degrees and on different levels. Loanwords can undergo phonetic, graphemic and grammatical integration to the receptor language.

In different languages various tendencies to integrate loanwords can be found. When the source and receptor languages have different writing systems, there will be pressure for orthographical adaptation; big differences in phonological systems between the languages will require the loanwords to be subjected to phonological adaptation. Good knowledge of the source language among the speakers and positive attitude towards it on the other hand can reduce the pressure to adapt. (Onysko 2007: 39, 51)

Sajavaara (1989: 70-72, 96-97) divides direct loanwords in Finnish into three categories, mostly according to their phonemic adaptation, although the degree of adaptation is presented in relation to their use:

1) unadapted loanwords or *sitaattilainat* stay unmodified (e.g. *sarong*, *status quo*) - although in principle their spoken form can be modified to fit Finnish sound structure. They are often not known or understood by most of the population and some of them are just in temporary use. Foreign proper names are also included in this group. In the light of the definitions mentioned in the previous subchapters we could ask here whether these words should be even considered loanwords. On the other hand, there are also unadaptet loanwords that have not changed their form at all and are still well-known and often used, like e.g. *copyright*.
2) partially adapted loanwords or *erikoislainat* (e.g. *kognitiivinen*) have been somewhat adapted to the receptor language structures but can still contain phonemes or structures that are unusual in Finnish (e.g. *synonyymi* is partially adapted but has not succumbed to the vowel harmony). These words' meanings often are not completely clear to all language speakers.
3) adapted loanwords or *yleislainat* (e.g. *kahvi*, *tulppaani*). They are fully adapted orthographically and phonologically. They are familiar to all language speakers.

When a foreign word is used in Finnish, it must be integrated to the Finnish morphosyntactic system up to some degree in order to fulfill syntactic functions. For nouns that end in a consonant that generally means that the ending *-i* is added, which both helps with pronounceability and enables the noun to be integrated to the declension system (*box* - *boksi*), or some suffixes can be used for the same purpose (*active* – *aktiivinen*), which is a process that is even more common in adjectives.

Additionally some suffixes in the loanword might be considered superfluous and left out or replaced (*sexuality – seksuaalisuus*). In case of verbs, the most often used endings are *-ata* and *-oida* (*to observe – observoida*, *to check – tsekata*). (Sajavaara 1989: 97-104)

The result of the adaptation process follows some general rules, but there is a lot of variability. In this research one of the approaches I considered was to use phonetic matching to try to identify neologisms, but after some initial experiments I reconsidered. The idea was to take a list of English words, generate their potential equivalents (with massive overgeneration of course, but that would not necessarily be an issue) and compare them with the cleaned list of potential neologism candidates. All matches found would have to be explored individually to see whether they are actually attested in use or whether the match is coincidental, but the process could help narrow down the amount of candidates.

There were multiple problems; to begin with, while some loanwords' form come from the how the etymon is pronounced in the source language (*sound – saundi*), sometimes it stays true to the written form (*chat – chatti, pub – pubi*), or some mix of the two (*scene – skene*, *ketchup – ketsuppi*, *jazz – jatsi*). Both graphemes and phonemes can have multiple equivalents (e.g. *z* or */z/* -> *z*/*s*/*ts*). Suffixes in etymons can be kept, adapted, or replaced (*doping – doping*, *feeling – fiilinki*, *dumping – dumppaus*). Consonant groups that are difficult to be pronounced can be kept or adapted (*stress – stressi*, *training – (t)reeni* (col.)). There are additional, not completely regular changes that can happen in the process of adaptation: for example consonant or vowel gemination (*tape – teippi*, *banal – banaali*). (Sajavaara 1989: 97-104)

Even though it would be possible to take all these ambiguities in consideration, the overgeneration would be very large. It is probable that in addition to real anglicisms, a very large part of the matches between the generated list and the candidate list would consist of typos, errors induced by OCR (optical character recognition) and other non-words, so the question arises whether the manual work time saved by somewhat narrowing down the candidate list would not be nevertheless smaller than the time dedicated to designing the rules for generation. And the list of the problems to solve is not even complete yet: e.g. would there have to be two sets of rules, one for matching phonemes in an English pronunciation dictionary (e.g. the CMU pronouncing dictionary[4]) and another for matching graphemes in an English wordlist? Which English wordlist would be suitable and would it even include all English neologisms that can become anglicisms in Finnish? Further, if it happens that an

[4] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

anglicism appears in a corpus only in an inflected form, it would not be found this way unless we the whole declension paradigm is generated for each already generated form, and that would result in even more massive overgeneration; etc. The issues were so plentiful that I decided to leave this approach completely.

2.1.5. English in Finland

According to Battarbee (2002)'s summary of history of English-Finnish language contact, there was no significant English influence on Finnish before the industrialization era, and the little that was there came through Swedish. During the industrialization era, there was a strong nationalist movement that put emphasis on using Finnish lexis for all the new terminology, so language borrowing was rare then too. That means that the first influx of English loanwords appeared in the era between the two world wars, and it was especially concentrated in the part of the lexis that has to do with culture and entertainment. Sajavaara (1989: 84) on the other hand places the era of dominance of English in this field to only after the World War II and says that between wars the strongest position belonged to the German language.

The position of English as a carrier of innovation in science, as the language of business, academia and entertainment, with its strong position in the global media, has enormous effect on the lexis of all languages in the world (Pulcini et al. 2012: 2), and Finnish is no exception. In the 1980s and 1990s Finland has changed very fast into a country where services, high technology and especially the IT-sector have important economic roles. A large amount of vocabulary to describe novel things in these fields has been needed. It is interesting that it seems that Finnish has a preference for using indigenously coined vocabulary even there where many other European languages use anglicisms, like in the field of computing, but the amount of anglicisms of Finnish has been growing in this and basically all other fields anyway. (Battarbee 2002: ch. 14.1.1)

2.2. Related challenges

2.2.1. Internationalisms

Internationalism is a term that depending on definition either denotes lexemes or smaller linguistic units (affixes like *auto-*, *hyper-*, *-logy*, etc.) that are derived from neo-classical roots or are borrowed from Greek or Latin directly (Onysko 2007: 77), or in a somewhat wider definition it can mean all cognates that are shared across most European languages (which can be called Europeanisms) or even wider due to common Indo-European descent (Görlach 2003: 15-16). They constitute a problem for this work, as because of their ubiquity it is often difficult to establish when and how they are borrowed to a new language. This is further obscured by the fact that a big part of English language vocabulary comes originally from Latin and Greek, even though those lexemes have undergone extensive adaptation, including semantic (e.g. extension, narrowing or shift of meaning) (Sajavaara 1989: 82-83).

Internationalisms came to Finnish language for a long time through Swedish, which is logical due to the long-lasting exceptional position of Swedish language for Finns, and for which an occasional clear evidence can be seen in the adapted form of the Finnish word (e.g. Lat. *Caesar*, Ger. *Kaiser*, Swed. *kejsare*, Fin. *keisari*). Nowadays internationalisms are borrowed almost exclusively through English, even though they are hardly ever anymore real words from old Latin or Greek and more often neologisms that are put together from elements that proceed from those languages to denote new phenomena (e.g. Greek *tele* + Latin *visio* -> Eng. *television* and Fin. *televisio*). (Sajavaara 1989: 66-67, 74).

GLAD project's guidelines say to not to include internationalisms among the anglicisms, or more specifically not those whose English origin is not possible to determine with certainty (Gottlieb et al. 2018: 8). But in the scope of this research it is not possible to explore etymology, and the methods used here will very probably find and group internationalisms together with anglicisms. On the other hand, a source that is available, Nykysuomen sanakirja (see chapter 3.4.), reflects the Finnish language of approximately late 1940s, and even though its coverage is not 100 per cent, it can serve as a simple way to place the first attestation of a word in Finnish approximately before or after the end of the WWII.

Therefore if I am unsure whether to categorize a word as internationalism or as an anglicism, as a rule of thumb I mark them as internationalisms if they appear in Nykysuomen sanakirja unless their English provenance is clear (e.g. in *outsideri*) or unless I have their date of first attestation which precedes WWII from a different source. If the word is not in Nykysuomen sanakirja, I categorize them as anglicisms because I suppose that they came to Finnish language later, through English. A certain amount of mistakes in this case cannot be avoided, but those can be sorted out later by closer manual inspection.

2.2.2. Language identification

One of the issues that has to be dealt with while trying to harvest anglicisms is a certain amount of foreign language text that appears in most large corpora. Foreign texts can appear in different scopes, from large texts that are either a part of a larger Finnish text that belongs to the corpus through longer quotes to shorter code-switches.

Code-switching happens when a speaker switches between two language varieties, and it can appear also in written text. There is not an exact consensus on the boundary between code-switching and borrowing; they are said to form a continuum. However, borrowing tends to encompass primarily single lexical items while code-switching usually consists of multi-element syntactic units which are not adapted on any level to the language around them, while the term borrowing is generally used to denote a higher degree of acceptance into the receptor language, which is often reflected by adaptation on morphosyntactic and/or phonological level. (Onysko 2007: 36-38)

Corpora can contain annotation of foreign-language material, but especially in case of monolingual corpora there can be an assumption that because the source of the data is monolingual overall, there will not be any significant amount of foreign-language material. That is not necessarily true. As shown in the chapter 4.2.5., that amount can be quite high after all, up to 12 % of unrecognized unique tokens. That can be problematic for further work with the corpus, especially in case of corpora that are distributed in a form of frequency lists (like FNC1, see chapter 3.3.) that do not show tokens' context and when the foreign language material is not annotated, as it is in case of this work. The goal of this work is to find anglicisms, i.e. English-sourced borrowings. The only way to recognize unadapted borrowings from foreign language material is from context, and without either context or reliable annotation, that goal is very difficult to accomplish.

The corpora in Language Bank which are pre-lemmatized do carry word-level annotation with a tag "Foreign" that is created by the parser, but this annotation is not very reliable. The parser used is a tool based on Turku Dependency Treebank statistical dependency parser (Haverinen et al. 2013) and it seems to assign the "Foreign" tag to all the words that it does not recognize and that it does not identify as symbols, so the tag lumps together foreign language material with all kinds of other tokens.

In theory it should be possible to annotate the corpus material for languages by myself. Language identification as a field is well-developed and has very high precision in general, but as the length of the foreign language insertions in the corpus can be as short as just one word, language identification would have to be done at that level. Language identification of very short texts is more problematic.

The works in that subfield often use character-based n-gram models (Vatanen, Väyrynen & Virpioja 2010: 1) in connection with some machine learning method, e.g. conditional random fields (Jauhiainen et al. 2019: ch. 10.7), but also with different ranking methods, e.g. some kind of a distance measure (Cavnar and Trenkle 1994: ch. 3.2). The basic idea is that n-grams and their relative frequencies give a generalized picture about the language (especially the shorter and the very frequent longer n-grams). Unfortunately I was not able to find any freely available and easy-to-use enough tool that I could use for language identification of the corpora. In lack of better tools, I decided to use the "Foreign" annotation in a heuristic manner described in the chapter 4.2.1. to get rid of at least some percentage of the foreign language material.

If such a tool happened to be available, I would probably try to identify the language of each corpus token and its immediate context (two or three words in both directions). Then I would remove from data clusters where all the tokens show high probability of not being Finnish. It's important to note that, on the other hand, places where most of the context appears to be Finnish but the token in question is not, are exactly the places that should be kept or even separately marked for closer inspection, as those tokens will have high probability to be borrowings.

2.3. Related research

On the topic of gathering anglicisms it is impossible to leave out **Manfred Görlach**. His **Dictionary of European anglicisms** was published in 2001 and it contains manually collected anglicisms from 16 European languages, including Finnish. Gathering the anglicisms must have been quite an effort. Keith Battarbee, the contributor for the Finnish part of the volume put the Finnish list together with three students. He does not explain their collecting method more closely, only notes that they used for that purpose all available dictionaries, including those for particular fields (Battarbee 2000: ch. 14.5). It's unknown whether they undertook any kind of systematic search for anglicisms that are not present in the dictionaries. The amount of anglicisms they found is difficult to find out (I would have to do it by hand because search in a document in a PDF form is making it very complicated), but for illustration, for O-starting etymons I counted 25 anglicisms, with at least a third of them being calques (which among other things mean that they most probably start with some other letter than O in Finnish). Most of them are included also in my list of O-starting anglicisms (see chapter 4.1.6.); some were left out as they in the meantime fell out of use.

Görlach's work was the first to collect anglicisms for cross-linguistic comparison in this scale. But due to the time it was made at the collectors could not use the technical means available now, which means that their effort was not necessarily systematic.

Andersen (2005) describes an approach to automatic anglicism extraction in Norwegian that was a direct inspiration for my work. Andersen works with the Norwegian Newspaper Corpus[5], a self-expanding corpus of newspaper texts that currently contains cca 1,7 billion words. Each day about

[5] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-4/

1500 previously unseen word forms are added to the corpus, most of them compounds but also names, spelling errors, and also anglicisms. The paper describes a part of the effort to classify these new word forms automatically, i.e. to retrieve anglicisms.

Andersen (2005)'s process uses single tokens as inputs, and it outputs a value that indicates whether the word is an anglicism or not. As evaluation data a set of randomly chosen 10 000 new words from the corpus was chosen, and among them were manually identified 563 anglicisms. Andersen compares several methods of anglicism retrieval and their combinations: lookup in a BNC[6]-based lexicon with fuzzy matches and restriction of very short words (precision 45,78%, recall 6,75%), similar lookup in the BNC lexicon with lexemes that were also found in Norwegian lexicon removed (58,7% precision, 6,57% recall), comparison with a list of character-based n-grams (4-6 -grams from BNC with frequency > 100 which do not exist in Norwegian lexicon; recall 36,23%, precision 74,73%), lookup using regular expression constructed intuitively based on the knowledge of English orthography (recall 39%, precision 60,6%), and combination of the last two methods (recall 48,8%, precision 75,76%).

Andersen (2005) concentrates only on unadapted anglicisms, which is where his work differs from this one. It is possible that based on the affinity between English and Norwegian languages, fully adapted loanwords from English that are formally indistinguishable from native Norwegian words are more common. I believe that Finnish direct loanwords are under bigger pressure to be adapted due to the differences between English and Finnish phonology, partly because some Finnish speakers have troubles pronouncing them and make no contrast between e.g. /b/, /g/, or /z/ and their voiceless counterparts (Suomi, Toivanen & Ylitalo 2008: 35-37; Sajavaara 1989: 106-108), and partly because there is pressure on adapting loanwords to the needs of the Finnish inflection system (Sajavaara 1989: 98). But it also seems that adapted loanwords in Finnish often keep features that distinguish them formally from fully native Finnish words. In support of this claim I can say that out of my own evaluation data, the list of O-starting anglicisms (see chapter 4.1.6.), only 42 % fit all the rules of fully native Finnish graphotax (see chapter 5.2.2), while 58 % do not.

Andersen (2005) apparently does not need to worry about language mixing in his data, at least there is no mention of that in his paper. My assumption is that the Norwegian Newspaper Corpus is carefully curated in that respect. However language mixing is a major problem in my work, because (as can be seen from the chapter 4.2.5. on data cleaning) available corpora of all genres contain a

[6] British National Corpus, http://www.natcorp.ox.ac.uk/

surprisingly high amount of foreign language material. Andersen's approach can be simplified as retrieving the most English-looking tokens in his data (and this is augmented by his decision to concentrate on non-adapted anglicisms only), but if I apply the same approach, I will most probably end up with English words from English context that are not necessarily used at all in Finnish context and thus are not anglicisms.

Losnegaard & Lyse (2012) expand on Andersen (2005)'s experiments with character level n-gram matching by involving machine learning. They use the Tilburg Memory-Based Learner software package to calculate similarity between words' character level n-grams, and for computational purposes restrict the n-grams' order to trigrams. Their method can identify anglicisms with 70% precision in some settings, but in practice their method does not reach significantly better results than the original method unaided by machine learning.

Andersen (2012) then summarizes his and Losnegaard and Lyse (2012)'s effort and explores some patterns common to the anglicisms that evade detection. The anglicisms that stay unrecognized are often compounds with an anglicism component, compounds with some specific components with high frequency, and words than contain Norwegian-specific characters (those are usually compounds with an English component).

Alex (2008) identifies English inclusions in German with an unsupervised method and then extends the same system to French in demonstration of her system's adaptability. Similar to Andersen (2005), the method does not need to take the language of tokens' context into consideration, but for a different reason; the term "English inclusion" encompasses English words in English context, English words in German context, anglicisms, and even English proper nouns. Her process is based on lexicon lookup, and where that fails, on search engine lookup and comparison of a token's frequency on English and German webpages. The method is quite dependent on the internal language identification method of the search engine (Alex uses Yahoo because it allows more automatic queries than Google), the precision of which is not evaluated in the work. Alex also experiments with machine learning, more precisely with a conditional Markov model tagger, but even though her intra-domain results look promising, cross-domain results are not very good (F-score between 22 and 55 %).

Säily, Mäkelä and Hämäläinen (2018) search for neologisms in the Corpora of Early English Correspondence (CEEC). Their working definition of a neologism is slightly different from the one used in this work; a neologism is understood as a word that appears in the corpus no longer than 100 years after its first attestation in the Oxford English Dictionary. The 100 years make sense in the era in question (CEEC consists of texts from the 15th until the end of the 18th century), when a new word could take very long to be established as a part of general vocabulary. The goal of their research is also quite different as they use the neologisms to analyze their sociolinguistic context.

CEEC should not have any errors or mistakes introduced by OCR because the whole corpus was proofread several times (Mäkelä et al. 2020: 90), but there is a problem of large variety in spelling in Early English. The authors devised a graphical interface that presents the list of tokens grouped by keys created by a spelling normalization algorithm commonly used in that field of research, and also context from the corpus for the active token and it's dictionary entry from the Oxford English Dictionary, which contains the year of the first attestation of the word. The interface certainly makes manual work faster, but the neologism identification is only made possible by focusing the research to words with specific derivational suffixes, like *-er/or* or *-ity*. (Säily, Mäkelä and Hämäläinen 2018: chapter 3.2)

Säily, Mäkelä and Hämäläinen (2018)'s work has the advantage that they have access to a dictionary that carries the dates of first attestation for each word and also, because their language of interest is English, access to all kinds of NLP tools and useful datasets. In the discussion part of the article they suggest some methods that could improve the problem which our researches share, a large number of tokens that cannot be parsed automatically so they could be excluded from the pool of potential neologisms. Unfortunately, these methods, like using edit distance and distributional semantics or then a statistical character level machine translation model to map such tokens to the dictionary words, apparently have not been helpful enough so far, because they end up producing too many incorrect results. (Säily, Mäkelä and Hämäläinen 2018: chapter 5, paragraph 5)

In general, even though research on **OCR post-correction** and my work share some aspects, ultimately their goals go against each other. When an unknown string is met in OCR post-correction, it is supposed that it is actually a known word form that is somehow obscured (by imperfect OCR or, e.g. in historical document retrieval, also by historical or non-standard spelling) and the unknown string needs to be mapped onto this known word form. There is a non-negligible possibility that trying to clean the corpus from OCR errors would actually omit a part of neologisms present in it by mapping them onto existing, graphemically similar words.

2.4. Scope

2.4.1. Anglicisms in the Global Anglicism Database

For the purpose of the GLAD project, an anglicism is defined as "any individual or systemic language feature adapted or adopted from English, or inspired (…) by English models, used in intralingual communication in a language other than English" (Gottlieb 2005: 163, as cited in Gottlieb et al. 2018: 7). In addition, some specific limitations are mentioned to define the scope. The anglicisms the project is interested in should be attested in language of the 20th and 21st centuries (but can be borrowed before that). They can include common slang but should exclude in-crowd jargon and vocabulary only known to specialists. They should be attested in a variety of sources and not to be just a part on one person's idiolect, but there is not a hard-set lower limit of citations necessary for inclusion in the database. In addition, the anglicisms should not be proper names and proper-name based adjectives nor internationalisms whose English origin is impossible to determine. (Gottlieb et al. 2018: 6-13) The anglicisms in the database are categorized in a way that more or less corresponds with Pulcini et al. (2012)'s classification described in the previous subchapter.

2.4.2. Etymological information

In chapter 2.1.1. it was said that it could be argued that a neologism stops being a neologism when it is formally accepted into official language. Nevertheless, for the purpose of this work an anglicism is understood as any linguistic unit whose English origin is still in some way recognizable. This does in principle include anglicisms that are fully integrated and not salient in any way in the eye of an average native language speaker, and whose origin is only given away by its etymological information, but in practice such anglicisms most probably cannot be identified by any method described in this work.

As already implied in the previous paragraph, an important limitation of the scope of this work is that it cannot take into consideration etymology as a parameter for harvesting anglicisms. In practice, that means that if an average speaker of Finnish cannot tell that a specific linguistic unit is a neologism (not necessary that it is an anglicism), then most of the methods mentioned in this work cannot recognize it either. This includes most calque-like neologisms, which take only their meaning from the source language but are realized by receptor language means. Those could be identified by exploring their frequency and collocations, but even then they can be at maximum identified as neologisms, not as anglicisms, without knowing anything about their etymology.

That is unpleasant because it means part[7] of anglicisms is lost by definition, but it is a necessary limit, because the information simply is not available. And even if it was, maybe it would not be smart to use it. Even though his work on anglicisms has nothing to do with harvesting anglicisms by automatic

[7] Unfortunately, it's not known how big part that is; this is very language-specific information, and there is no research about Finnish language on this topic.

means, Onysko (2007: 61-62) gathers arguments against the use of etymologic knowledge in any larger-scope work on anglicisms, and in support of rather exploiting word form. He states that in many cases, the etymological information is conflictive and unreliable. Various references can provide varying interpretations of the origin of a borrowing.

It is difficult to prove where a neologism exactly came from and when. Nowadays the mass media, especially the internet, allow the first appearance of a word to be tracked more easily, but that is a question of the last couple of decades. The lag between the coinage of a term and its appearance in searchable data can be quite short, but before the internet times the first record of a neologism in written form could be preceded by years or longer of its spoken or even written use without it leaving any mark (Onysko 2007: 61).

In short; if we have reliable etymologic background for a word, we have probably already enough information to decide whether it is an anglicism or not, and there's no need to go searching for it. In this work, the information that is used for identifying anglicisms is most importantly the form of a linguistic unit and the information contained within, especially the phonotactic (or actually in our case graphotactic) information. For confirmation of a status of an anglicism in contemporary Finnish also information gained from context will be used, such as the language of the linguistic unit's surroundings. Some statistical information can also be used for that purpose, for example frequency of appearance of the linguistic unit in a corpus and any additional information, such as its frequency and distribution in corpora of different genres or from different time periods.

2.4.3. Scope of this work

To sum up the scope: in this work I try to identify direct (i.e. those that keep some formal evidence of the source language) Finnish anglicisms, both adapted and unadapted. I concentrate on their form and disregard potential etymological information. Regarding the approximate time of borrowing, I posit that anglicisms are for the most part just those borrowings with English elements that appeared in Finnish only after the WWII. Earlier borrowings that formally resemble anglicisms are most probably internationalisms, which don't belong to GLAD project's scope, but because the boundary between the two might be unclear, some internationalisms can be lumped together with anglicisms. As a simplified method of finding out which is which I'll use lexeme's presence or absence in Nykysuomen sanakirja. Information that can be gained from context and statistical information in corpora is used to ensure that the borrowing is actually established and in use in contemporary Finnish language.

3. DATA

Even though humankind nowadays produces more texts than ever, it is not easy to come across good datasets that would be available, high quality and in a suitable format. This is even more true of linguistic resources for languages with a relatively small number of speakers. Luckily, Finnish Language Bank manages an amount of text corpora, some of which will be suitable for this project. Kotus (The Institute for the Languages of Finland) can also provide researchers with some wordlists that can be helpful for projects of this type.

The characteristics of a corpus are essential for the quality of the result of this work. As mentioned earlier, in an ideal case and with ideal tools, non-calque/non-phraseological neologism harvesting would be easy, because all known wordforms would be correctly lemmatized, recognized and discarded, all foreign words would be marked as foreign and discarded, all proper nouns would be recognized from their form and discarded too, and all that would be left would be either neologisms or non-established neologism. However, the reality is far from this ideal scenario. The amount of noise (typos, errors, unmarked non-Finnish text) add up to other problems, like parser errors. Issues are more difficult to deal with when they're combined. In addition to that, the documentation of the available corpora is often not very detailed, or it is not easily available. For this reason, the following section describes each dataset quite extensively so that their characteristics can serve for choosing a suitable corpus as a source for the neologism candidate list creation.

3.1. PROJECT'S REQUIREMENTS FOR CORPORA

Suitable corpus data for this research would cover a broad spectrum of genres; ideally it would be representative, meaning that it would contain all types of text. This would make it a true reflection of the whole of Finnish language even though it does not need to be balanced, i.e. contain those types of text in correct proportions. (McEnery & Hardie 2011: 250) The data should include relatively recent texts in order to catch also newer anglicisms. It should use relatively standard language to avoid overrepresenting personal eccentricities of expression. The data should be as clean as possible from typos and spelling mistakes.

In addition, an ideal corpus would be either monolingual (in this case Finnish) or annotated for language, it would be lemmatized, and it would also carry the information about the origin of the texts with chronological information. Such information would be very handy for identifying first appearances of neologisms in the dataset. It would be practical if the corpus covered a longer time period, as that would enable comparing frequencies of appearance of a word at different time points. And the corpus should be quite large in order to be able to cover as large vocabulary as possible and include even rarer words in high enough quantity to give sufficient information about their use.

One type of corpus that is both large enough and diverse enough to fill the criteria at least partially is newspaper corpora. Newspapers, especially large and web-based ones, are a good source for an adequately representative corpus not only for their convenient -already digitalized- form, but also because of the regular and large production of text, diversity of the topics they cover and genres and styles they use, and because of a certain level of quality brought by the professional writing and editing (Andersen & Hofland 2012: 3).

Another type of corpus that might be suitable is text of Wikipedia articles. The articles are crowdsourced - written and edited by many different authors, and hopefully that could prevent unestablished words from appearing in the text of articles. It is also reasonable to expect that any typos and other ungrammatical expressions would get quickly spotted and fixed. Regarding corpus representativeness, Wikipedia does not cover diverse genres; its language is in general formal, even almost as formal as traditional print encyclopedia articles, neutral, and relatively homogenous (Hiltunen (2014) says that about English Wikipedia, but I'd argue this is generalizable). This corpus will not be able to help with neologisms in slang and colloquial Finnish.

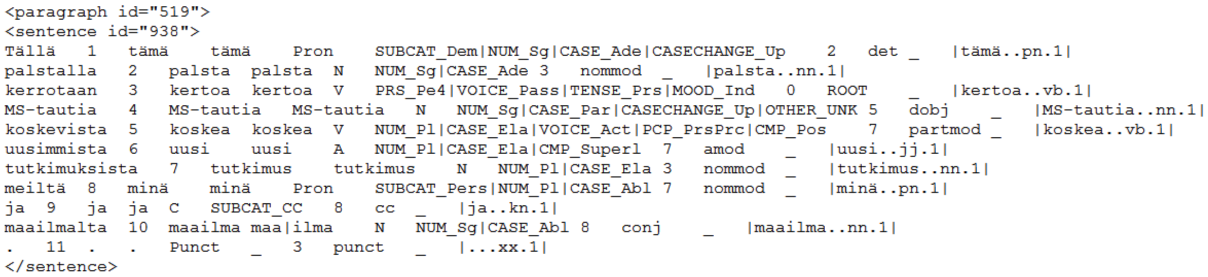
In addition to the previous types of corpora, some kind of online discussion board text should be included to cover more colloquial language; in theory spoken language corpora could be even better source for this purpose, but they tend to be significantly smaller because the material is more difficult to obtain and process[8]. An online discussion board corpus contains more colloquial text than the two above mentioned types of corpora, which of course also brings potential risks. Uncurated text in an internet forum will have more typos than e.g. newspaper, and personal idiosyncrasies might appear with higher frequency. On the other hand, if only texts like newspapers or Wikipedia are explored, colloquial language and slang will be unavoidably left completely uncovered.

[8] E.g. Language Bank has the Longitudinal Corpus of Finnish Spoken in Helsinki, the exact size of which I have not been able to find out, but it appears that the size of the transcribed corpus is 40 MB as noted on http://urn.fi/urn:nbn:fi:lb-2017030103 – in comparison with gigabyte sizes of the text corpora used in this research!

3.2. Language Bank's corpora

The corpora listed in the next subchapter are all available in the Language Bank of Finland and have some characteristics that make them useful for this project, as per the description of an ideal corpus for neologism harvesting above. The Language Bank does have some basic information about the corpora on their pages, but surprisingly often the information is very brief and incomplete. It was necessary to download the available corpora and explore them in order to find out the details. For that reason, I decided to describe the corpora in detail here. It's an interesting question whether researchers would use the corpora more if more detailed description were easily available, maybe even with an example of what the data actually looks like (where the conditions of use do allow it).

Most of the corpora are stored in the VRT format, apparently because VRT format is the data format that the Korp corpus search interface uses. VRT stands for VeRticalized Text, which consists of XML-style tag elements and columns with the corpus tokens and their annotation.



```
<paragraph id="519">
<sentence id="938">
Tällä   1   tämä    tämä    Pron    SUBCAT_Dem|NUM_Sg|CASE_Ade|CASECHANGE_Up    2   det _   |tämä..pn.1|
palstalla   2   palsta  palsta  N   NUM_Sg|CASE_Ade 3   nommod  _   |palsta..nn.1|
kerrotaan   3   kertoa  kertoa  V   PRS_Pe4|VOICE_Pass|TENSE_Prs|MOOD_Ind   0   ROOT    _   |kertoa..vb.1|
MS-tautia   4   MS-tautia   MS-tautia   N   NUM_Sg|CASE_Par|CASECHANGE_Up|OTHER_UNK 5   dobj    _   |MS-tautia..nn.1|
koskevista  5   koskea  koskea  V   NUM_Pl|CASE_Ela|VOICE_Act|PCP_PrsPrc|CMP_Pos    7   partmod _   |koskea..vb.1|
uusimmista  6   uusi    uusi    A   NUM_Pl|CASE_Ela|CMP_Superl  7   amod    _   |uusi..jj.1|
tutkimuksista   7   tutkimus    tutkimus    N   NUM_Pl|CASE_Ela 3   nommod  _   |tutkimus..nn.1|
meiltä  8   minä    minä    Pron    SUBCAT_Pers|NUM_Pl|CASE_Abl 7   nommod  _   |minä..pn.1|
ja  9   ja  ja  C   SUBCAT_CC   8   cc  _   |ja..kn.1|
maailmalta  10  maailma maa|ilma    N   NUM_Sg|CASE_Abl 8   conj    _   |maailma..nn.1|
.   11  .   .   Punct   _   3   punct   _   |...xx.1|
</sentence>
```

Table 1: Example of VRT format from the Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s.

The tags carry information about the structure and any additional information, e.g. about the origin of the text. The tokens and their annotation are organized in columns; one column for the tokens themselves and another set of columns is generated automatically by the TDT (Turku Dependency Treebank) parser. The annotation contains one analysis for each word, disambiguated with a hybrid approach of machine learning and a small set of rules (Haverinen et al. 2014: 514-516). The analysis contains the token's lemma with and without compound boundary marked, POS annotation, more detailed morphological analysis and dependence relation information. (The Korp corpus input format (n.d.); Korpin korpusannotaatio: TDT (n. d.))

The annotation made by the TDT parser is not completely identical for all the corpora. Potentially useful for this research is that there is an annotation option "foreign" that in some corpora appears among POS annotation or elsewhere in the morphological analysis column. The difference seems to be unimportant; in both cases tokens that are marked as "foreign" can be really a non-Finnish word, but it can just as well be a typo or other (from the point of view of a parser) difficult or atypical token, e.g. a website address. This annotation was used for removing some parts of corpora that could make the harvesting of neologisms more difficult (see chapter 4.2.1.).

3.3. Suitable Language Bank's corpora

The Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland (FNC1). The corpus consists of 1-, 2- and 3-grams created from digitalized (via OCR) Finnish newspapers, magazines, and periodicals published between 1820 and 2000, divided by decades. The focus of this corpus is on the older period (out of the 6 gigabytes of 1-gram lists, the lists from the 1940-2000 decades consist of only 390 MB), but it is nevertheless a large corpus. Its source, The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland (KLK-fi), contains 5.2 billion tokens; the part from the period of interest for this project (in this case 1950-2010) contains around 5% from that, 260 million tokens.

Most of the Language Bank's corpora come in the VRT format with the XML-like tags that, apart from carrying extra information, also structure the text into sections and sentences, which means that the information about the tokens' context is preserved. This corpus comes in frequency lists of n-grams, which means that it is already pre-processed. Even though a lot of information about the context and origin is lost in this processing, a researcher is spared from manipulating gigabytes of data.

A problem is that due to the OCR in the digitalization process, this corpus is full of erroneous word forms and non-word strings of characters. The division by decades is on the other hand a plus, because it can provide chronological points of comparison of Finnish language in each decade.

Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, version 2 in VRT format (lehdet90ff-vrt-v2) consists of 247 million tokens (19 GB of data). This dataset could be very useful, but it also brings a potential problem. It includes many scientific magazines, so it is possible that the amount of scientific terms, often unknown to average speaker, will end up masking the neologisms we're trying to find. Fortunately, the magazines are divided into two parts, "magazines" and "scientific magazines", which could help with this issue. On the other hand, the corpus was apparently made by collecting the texts of the magazines in various formats (text, PDF, maybe also HTML?), so the quality is uneven.

Finnish Wikipedia is available for download at any point from Wiki dumps, but for convenience, Language Bank also has **Finnish Wikipedia corpus (Wikipedia-fi-2017-src)** with all Finnish Wikipedia articles available at 1. 1. 2018, in VRT format, tokenized and annotated with the TDT parser. It contains 83 million tokens. It's difficult to tell beforehand how big potential this corpus might have for harvesting neologisms that are not specific to scientific jargon, but on the other hand this is one of the corpora that are digital from the beginning, not digitalized, and as such should not contain too many errors and other noise.

Yle Finnish News Archive 2011-2018 (ylenews-fi-2011-2018-src) contains over 700 000 articles from YLE webpage with more than 200 million tokens. It could be very useful because unlike other corpora with digitalized newspaper, these are articles that were published online and as such could be close to error-free. This corpus is available for download in JSON format and is without lemmatization or any other linguistic analysis. Analyzing the Finnish text extracted from the corpus with the *finnish-parse* parser on Language Bank's supercomputer Puhti would, according to my calculations, take approximately a week of non-stop processing, so I will only use this corpus as a source of 1-grams and only do the morphological analysis on word forms in frequency list, without context.

Regarding online discussion board corpora, the Language Bank of Finland has two of them, Suomi24 and Ylilauta. **The Suomi24 Corpus 2001-2017 (suomi24-2001-2017-vrt-v1-1)** contains texts from Suomi24 online discussion board from the years 2001-2017. It is massive, with more than 4 billion tokens (71 gigabytes). The downloadable version of the **Ylilauta corpus (ylilauta-dl)** is the text from the Ylilauta discussion board from between years 2012 and 2014, contains 26 million tokens, is lemmatized and contains also morphosyntactic and semantic annotation. The size of Ylilauta is much more manageable, which is why I decided to use it. It also comes in the VRT format.

I also considered the following corpora, but the access to them is more restricted than to those above, and they would only cover the styles and topics already covered by the corpora above. I mention them for completeness. **Finnish Text Collection (FTC)** contains a selection of electronic texts (newspapers, journals and also books) from the years 1987-2000. It would be quite practical for this project, but the access is restricted, and the texts are at this point 20-33 years old. Unsure if this is originally digital or digitalized resource. The size of the corpus is 180 million tokens. **The HS.fi News and Comments (HS.fi)** Corpus contains domestic news and their comments from the Helsingin Sanomat website from one year's time 2011/2012. Its size is 8 million tokens.

Finally, also subtitles could provide some interesting data to explore regarding neologisms; they could bring in more colloquial language, like the discussion boards, but maybe they would be less riddled by errors. On the other hand, many such subtitles are direct translations from English, often made by amateurs, which makes them more prone to contain anglicisms that are not actually in use elsewhere but in bad translations. It could be interesting to try out whether such corpus contains anglicisms not seen in other corpora. **Finnish OpenSubtitles 2017 (opensub-fi-2017-src)** is a freely available corpus of Finnish subtitles for movies and series and contains more than 267 million tokens.

It consists mostly of user uploads which generally contain lots of noise and errors and they can come in many different encodings, but the dataset has gone through ample preprocessing and correction of spelling errors and errors induced by OCR extraction (Lison & Tiedemann 2016: 924-926).

Corpus	**Language Bank's code**	**size**	**format**
Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, (only the non-scientific magazines)	lehdet90ff-vrt-v2	cca 166 million	VRT format, pre-lemmatized
Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland, just 1950+	FNC1	260 million	only 1-, 2- and 3-gram lists
Finnish OpenSubtitles 2017	opensub-fi-2017-src	267 million	VRT format, pre-lemmatized
Finnish Wikipedia corpus	wikipedia-fi-2017-src	83 million	VRT format, pre-lemmatized
Yle Finnish News Archive 2011-2018	ylenews-fi-2011-2018-src	200 million	JSON format
Ylilauta corpus	ylilauta-dl	26 million	VRT format, pre-lemmatized

Table 2: List of Language Bank's corpora used in this thesis.

3.4. Wordlist

Another readily made dataset needed for this work is a wordlist. Its function is to serve as a list that contain as many correct Finnish words as possible and at the same time as few neologisms in the scope of this work as possible. An available wordlist that fulfills this function the best comes from the first codifying dictionary of Finnish language, **Nykysuomen sanakirja**. This six-part descriptive dictionary was published between years 1951 and 1961, but the work on it had started much earlier and it mostly reflects the codified Finnish of the 1940s (Räikkälä 1995). This is ideal, because as mentioned in chapter 2.1.5., until the end of the WWII there had not been many borrowing from English language in Finnish.

The wordlist from Nykysuomen sanakirja is not publicly available, but Kotus makes it available to researchers upon request. The wordlist contains some OCR errors and there might be some missing keys, which is why Kotus does not really advertise it (R. Widenius, Kotus, personal communication, 21. 2. 2018). It is nevertheless in good enough condition to be used in this research. The list itself is separated into compounds and non-compounds, and compounds have the lexeme boundary marked.

It contains 206723 lemmas (135181 compounds and 71542 non-compounds), but because some lemmas have more than one meaning in the dictionary, there are duplicates. After removing them the list contains 201366 lemmas.

3.5. Data for evaluation

When I started working on this project, I soon realized that there are no data that I could use for evaluating any of the potential semi-automatic methods that I was hoping to contribute with. Luckily, the GLAD project started with a pilot phase; each contributor was supposed to gather a list of anglicisms in the given language whose etymon starts with the letter O[9]. I decided to gather them by any methods I'd have at my disposal in the time given, including also time-consuming handpicking, with the goal of not only finishing the first part of the work fast, but also as a way to gather data that could be later used as data to evaluate any potential automated methods on. The next chapter explains the process I decided for.

4. Narrowing down the list: cleaning and comparing corpora

4.1. Exploratory run

At first, my idea about how the work could be done was pretty straightforward. Without going into more details, it could be summarized as sorting a written corpus of Finnish into what is basically a frequency list, removing from the list all word form whose lemmas can be found in a dictionary and proper nouns, and in the ideal case all that would be left after this process is finished would be neologisms. Then it would only be necessary to devise a way to identify those of the neologisms that are borrowed from English, perhaps taking advantage of their graphotactic rules (i.e. rules for acceptable letter sequences in a language). Here I refer to the fact that borrowings often contain

[9] In practice, though, the first gathered subgroup would consist of anglicisms that themselves start with O, not anglicisms whose etymon starts with O. It is unnecessarily difficult to search Finnish text for anglicisms according to the first letter of their English etymon. In some instances they do not correspond 1-on-1 on an overall level (*c-* in English to *c-*, *k-* or *s-* in Finnish), sometimes because of pronunciation in adapted borrowings (*alright – oolrait*), and then there is a whole category of calque-like neologisms (*credit card - luottokortti*).
But that only matters in case where we are actually separating the groups by their first letter. As the point of the GLAD project was to gather as complete list of anglicisms as possible, it's easy to gather the words by the starting letter of the anglicism itself, and in the end re-sort them by the etymon's starting letter after they are analyzed by a linguist.
It makes sense that the project required ordering by etymon's first letter though; it was important to have a possibility to compare the results from different languages.

phoneme sequences that are either not allowed (violate phonotactic constraints) in the recipient language or are scarce in it (perhaps only appear in words that were originally borrowed from other languages and were adapted a long time ago). In a phonetic language like Finnish this should be easy enough to translate from phonotactic rules into graphotactic ones. English language's phonology and graphotax differ quite a lot from Finnish, so it is reasonable to suppose that borrowings from English to Finnish will often differ from non-borrowed Finnish words.

I was aware that this method would have to be much more fine-tuned in order to actually work, and it has many obvious (and also a lot of less obvious) pitfalls. Some of them I was able to deal with by including extra steps in the process, and some of them I was not able to deal with at all. As a result, parts of the process had to be replaced by unautomated, manual work. I will now describe in detail the process I used and discuss conditions that would have to be fulfilled for the process to work well enough.

4.1.1. Data

For the purpose of the exploratory run I chose to use both the FNC1 corpus and Ylilauta corpus. The FNC1 corpus comes in frequency lists of n-grams, which was very practical as it did not require handling gigabytes of data and the exploratory run method did not require any context anyway. In addition, the division of the frequency lists according to decades made it possible to concentrate just on the period in question (1950s-2000s). Including both newspaper and internet forum language meant the data was more diverse and covers both standard language and more colloquial text. In addition, in the exploratory run I used a wordlist from Nykysuomen sanakirja, which reflects the codified Finnish language from around 1940s.

4.1.2. Creating lists for comparison

I began by creating one frequency list of tokens. For FNC1, I took the 1-grams from the year 1950 until 2010 and then employed various methods to get rid of some obvious typos and other wordforms that cannot be of any use in the process. Some of the categories of the forms that were removed are listed just below, along with a short reasoning for their removal and brief pondering if their removal also brought along some problems. All of the categories are then listed in the table 3 a little bit further on.

- all tokens that contain uppercase; they include mostly names, but also acronyms. Unfortunately first words from sentences fall under this category, but even though their removal lowers the frequency for some words, the risk of removing a potential neologism completely in this way is low and worth it for being able to not have to deal with proper nouns.

- all tokens that contain other characters than those used either in Finnish or English orthography, expressed by the following regular expression: [^\-'abcdefghijklmnopqrstuvwxyzäö]. Hyphen can appear in compounds; apostrophe is used in some places, e.g. to divide two same vowels that belong to different syllables or in declension of some borrowings (Kotimaisten kielten keskus, n.d.). Colon is also used in Finnish declension, but only in the case of abbreviations or symbols (Kotimaisten kielten keskus 2006) that should be discarded anyway, so colon is not included here.

  Such tokens are most often typos, mistakes, and all kind of non-words. In addition, Finnish orthography uses also other graphemes (š, ž, å), but those and any other characters than the ones listed above point towards borrowings from other languages than English.
- all tokens with frequency under 3. This was a heuristic cut-off. Words with such low frequency in a large corpus are most often typos or mistakes, and in a case of a potential neologism, a frequency this low would also hint at the fact that a neologism is not established, that it's just a part of someone's idiolect. But this step does include a risk that there is a relatively frequent lemma that gets discarded because it appears in several inflected forms, each of which only appears less than 3 times in the corpus. For this reason, in the later processes I only discard words after they're lemmatized (if possible). Unfortunately, if the word in question is a neologism and the parser does not recognize it, it does not try to guess the lemma, the word stays unlemmatized and will be discarded anyway. But to lower this risk, in all subsequent cases I only discard tokens with frequency of 1.
- too short tokens. Tokens with only one or two characters are functional or closed-class words that are unlikely to be borrowed (Pulcini et al. 2012: 12), or very often also OCR errors.

The Ylilauta corpus is already lemmatized, but other than that, I employed similar criteria for it.

4.1.3. Lemmatization and comparison of lists

With the wordlist acquired through previous steps, the next phase should eliminate all words already established in language and listed in codified dictionaries. This is a banal step for languages without complex morphology where one could basically subtract the vocabulary list from the corpus list and have their result. In Finnish the morphology makes this task more challenging, but it only requires a suitable tool. I used *finnish-parse* – Dependency Parser for Finnish based on the Turku Dependency Treebank statistical dependency parser (Haverinen et al. 2013), available at the CSC's supercomputer Puhti.

The Dependency Parser for Finnish attempts to do morphological analysis for every token and outputs the most probable result if it recognizes the token. The part of the output relevant for this work's purpose is the lemma (in case of compounds, the parser also marks the boundary between the compound's components. In case of unknown word, the parser will output that very same wordform as lemma. That means that with the input in the form of the list created from the corpus, the output gained in this phase will be a list of analyzed known lemmas and of unknown wordforms. As the Ylilauta corpus already includes lemmas for each token, this step can be skipped in its case.

At this point, I joined the lists of tokens from the two corpora together. This list can be easily compared with the dictionary list (Nykysuomen sanakirja described in the chapter 3.4.) and the lemmas present in the dictionary are removed. The result is a list of potential neologisms. Its quality, i.e. its precision (because only a fraction of it really will be actual neologisms) will depend on many factors discussed further on.

4.1.4. Numbers

The table 3 below illustrates the process described above on an example of the data from the decade of 1970s in FNC1. The steps described above in the chapter 4.1.2 are sometimes divided into smaller separate steps. All in all, it's clear that out of almost 3 million different word forms, almost 77% were removed because they only appeared once or twice; more than 7% others contained uppercase letters, around 2% others contained characters from outside of Finnish and English orthography, and more than 10% others were wordforms of lemmas that were already in the same list (this includes also separate parts from divided compounds). Approximately 68 000 lemmas plus other wordforms unrecognized by the parser (2,3% of the original token list) are left. Out of those, about one third can be found in Nykysuomen sanakirja wordlist. The rest, 43381 wordforms, 1,5% of the original token list, constitute the result of this process - the list of potential neologisms to go through.

In my exploratory run, when I employed the same process, I started with combined frequency lists of the FNC corpus from 1960s to 2010's and Ylilauta corpus, which contained tens of millions of tokens. After the processing and removing the words from Nykysuomen sanakirja, I ended up with a list of a little over 100 000 lemmas and unrecognized wordforms. The O-starting ones constituted about 2900 of those.

1-grams from 1970s, 2996335 tokens	**tokens left after step**	**removed in step, % of whole**
delete:		
all under frequency of 3	696519	**76,75%**
everything that contains words "superscript" and "subscript" (that’s a FNC corpus specific)	692153	0,15%
words that start with nonalphabetical characters or numbers	646395	**1,53%**
words starting with capital letters	432556	**7,14%**
all that contains any characters outside of the allowed	427885	0,16%
all that end with '-:	411692	**0,54%**
1- and 2-character words	411006	0,02%
words with same character more than 2times next to each other	410234	0,03%
words with many -':	410191	0,00%
short words with -':	409881	0,01%
words with ' or : among first three characters (FNC1 specifics? OCR issue probably)	409619	
lemmatize: *finnish-parse*	**409619 lemmas and unrecognized wordforms**	
delete:		
duplicates after lemmatization	131033	**9,30%**
words that have starting capital letter after lemmatization	130838	0,01%
separate recognized compounds into lemmas	**68440**	**2,08%**
delete words found in the Nykysuomen sanakirja wordlist	**43381**	**0,84%**

Table 3: Example of percentage of word forms removed in every processing step.

4.1.5.Evaluation

I went through the O-starting list manually and found 140 potential anglicisms that would fit the scope of the GLAD project's definitions. I also added some anglicisms found in other sources (e.g. Kotimaisten kielten tutkimuskeskus 1979; Eronen 2007). After closer inspection, I deemed some of the candidates too field-specific to fit the scope (*objektivismi*), other originated from other languages than English (*otsoni*), and more than 40 were internationalisms that with very high probability came to Finnish through Swedish because they were established well enough before 1950 that they were included in Nykysuomen sanakirja (*oodi*, *orkesteri*). The rest, 82 linguistic units, are most probably anglicisms that belong in scope of the GLAD project. That means less than 3% of the list I went through manually (or around 4,2% if I count internationalisms) were what I was looking for, even though in reality the percentage is a bit higher, because many of the found anglicisms/internationalisms appear also in other forms than just as a lemma. In any case, the proportion of the desired words is very low, and it definitely calls for some computer-aided process that would make the search more effective.

Examples of found anglicisms can be found in table 4 in the next subchapter. Here are some types of O-starting noise with some higher-frequency examples:

- Finnish words, the lemma of which is in Nykysuomen sanakirja, but the parser for some reason does not recognize them: *onneksi*, *oikeassa*, *oikeutettu*, *omillaan*, *oikotietä*, …
- Finnish words that either were not in use yet in the 1940s or were for some reason left out from Nykysuomen sanakirja; also compounds: *osasyy*, osa-*aikainen*, *omistusasunto*, *omakotirakentaja*, opinto-*ohjaaja*, *oikeanlainen*, *ostoskori*, *ohjeistus*…
- names that did not have capitalized first letter: *oulu*, *olavi*, *oslo*…
- typos or OCR errors: *oik*, *oii*, *oiisi*, *opetaja*, *ostajiem*, *olev*, *ort*, *ork*, *ohjeh*, *onen*, *oni*, *oila*, *osk*, *olia*, *oiiut, olekkaan*…
- words that were probably divided with at the end of a line and OCR did not connect them back: *opisk*, *opett*, *ollisuuskatu*, *orstaina*…
- foreign words that appear in the corpus in foreign context: *other*, *oder*, *officials*, *offentligrättsliga*, *officio*, *ohne*, *omarbetade*, *ocksä*…
- borrowings from other languages: *otsonikerros*, *oregano*

4.1.6.Data for future evaluation

From the point of view of evaluation for the next phase of the project, the list of O-starting anglicisms has to be looked at with a different set of requirements. The most important difference is that in the result list for the GLAD project, all words that belong under the same English etymon are understood as one item. For future evaluation all the potential orthographical or other variations (e.g. *off topic* in English -> *off topic*/*off-topic*/*offtopic*/*offari* in Finnish) have to be listed separately for automatic processing.

Another point is that if an anglicism was established enough in Finnish language before the 1950 or so, it fits the GLAD's scope (e.g. *outsider*), but it will not be found by any method used in this thesis, because to sift through the giant amount of tokens in some systematic way, I have to use Nykysuomen sanakirja to filter out "known" words. Nevertheless, it may still be present in compounds and derivations from the original lemma.

And in addition, any multi-word expressions among these anglicisms cannot be used to evaluate the results of methods used on tokens without context (*off the record*; *yhden miehen show 'one-man-show'*).

After updating the list according to the above-mentioned requirements, I ended up with a list of 62 anglicisms, out of which 7 are in Nykysuomen sanakirja, and 42 internationalisms, out of which 30 are in Nykysuomen sanakirja; altogether 104 lemmas. Some of the items have (very close) spelling variants which are counted all as one item in the list. This list will be used for evaluation. The words that are contained in Nykysuomen sanakirja are kept in the list, because compounds or words derived from them might still appear among unrecognized word forms from the corpora; their word forms can also still appear among to-be-evaluated tokens if the parser cannot parse them correctly for some reason.

O-starting anglicisms used for evaluating the precision and "recall" in corpora (altogether 104 lemmas)			
anglicisms	not in Nykysuomen sanakirja	*online*, *offari*, *oolrait*, *outletti*, *orkku*...	55
	in Nykysuomen sanakirja	*outsider*, *ohjelmoida*, *observoida*...	7
internationalisms	not in Nykysuomen sanakirja	*optio*, *orgiat*, *osteopatia*...	12
	in Nykysuomen sanakirja	*operaatio*, *optimisti*, *originaali*, *optimaalinen*..	30

Table 4: O-starting anglicisms used for evaluating the precision and "recall" in corpora.

4.1.7. Assessment of the exploratory run

The biggest issue of the exploratory run ended up being the quality of the data; there is an immense amount of noise. The FNC1 corpus is filled with OCR errors, while Ylilauta contains large amounts of typos and mistakes. On the other hand, no ideal corpus is available. But this realization led me to the decision to explore the available corpora more closely, because in some of them the ratio of neologisms to noise might be more favorable.

Another problem is the lack of context in 1-grams. Many actual English words appeared in the list of O-starting neologism candidates which must have come from quotes and code-switching in the corpora. Without context it is impossible to recognize code-switching from borrowing, so I had to explore each such word separately to find out if it appears in Finnish context.

Lack of context also means that to be able to decide about the status of the anglicism (to see whether it is actually established) whenever my speaker's intuition was not up to it, I either had to search the gigabytes of text in the corpora for the concordances myself to see the term in question in its context, or to use Korp (the web-based concordance search tool used by Language Bank of Finland; Borin et al. 2012). Korp is also slow as it has to go through the same gigabytes of text, but in addition for some corpora it is unable to show diachronic statistic. Thus identifying the first attestation of a word in the available corpora is problematic, let alone finding out about its frequency of use presented diachronically.

The context is the more complicated of the two problems to work with. So first I decided to explore the available corpora to see if better input would make better output; is there a corpus available that would make a better dataset to work with?

4.2. Context-less pipeline: comparing the corpora

The next step was to subject all the available and suitable corpora to the same procedure and compare the results. This procedure is still designed to only deal with frequency lists of tokens, without taking context into consideration, so it is basically an enhanced version of the exploratory run. Its purpose is both to produce a narrowed-down list of neologism candidates and to be a process applicable on any corpus, so that the result of the process can be compared, and the most suitable corpus can be chosen as the basis for any other further processing.

There are six corpora in question and they're the ones described in detail in the chapter 3.3.: the Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, version 2 in VRT format (lehdet90ff-vrt-v2), Finnish Wikipedia corpus (Wikipedia-fi-2017-src), Yle Finnish News Archive 2011-2018 (ylenews-fi-2011-2018-src), Ylilauta corpus (ylilauta-dl), Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland (FNC1), and Finnish OpenSubtitles 2017 (opensub-fi-2017-src). The Lehdet90ff corpus is divided in two parts, scientific magazines and other magazines, and only the latter was used in this work because there's a high probability that the scientific magazines contain plenty of very specialized terms not used in common language that would just add to the unwanted noise in the candidate lists. Of FNC1, only the data from 1950s onwards was used.

The tools I used in this process are the following: shell scripts; Python scripts; Excel in cases where the data needs to be presented in a structured and easy-to-evaluate-for-humans form; *finnish-parse* parser tool at CSC's Puhti supercomputer; NLTK's Snowball stemmer for Finnish as the last resort to automatically connect lemmas to inflected word forms that the parser does not recognize.

4.2.1. Foreign cluster cleaning

Most of the corpora are stored in multiple .VRT files which preserve the structure of the original text with tags on the level of text, paragraph and sentence. They also contain the lemma and a more detailed morphological analysis of each token (if the parser managed to determine one). In addition to other analyses, the parser attempts to determine which words are not Finnish and tags them as "foreign" (based on my observation I would estimate that the precision of this analysis is pretty high, but recall is not; I have not seen almost any Finnish words tagged as foreign, but many foreign words without the "foreign" tag).

After some experimenting, I decided to try to get rid of at least a part of the foreign words because they do get mixed up with unadapted borrowings in the manual phase of neologism detection. But the algorithm to remove them should not be too aggressive, because it's exactly borrowings that are in high risk of being tagged erroneously as "foreign". In the end I designed the algorithm to remove whole sentences in case that more than half of the tokens of a sentence are marked as foreign, and then also any clusters of 4 or more tokens tagged as "foreign" in a row were also removed. Such clusters were replaced by an end/start of sentence tag, so that if the data is later used to explore the context of the tokens, there is not new, falsely created context around them.

4.2.2.Format change and cleaning

The VRT files were transformed into lemma frequency lists (which apart from lemmas also contain unrecognized word forms) for the whole corpus. Lemmas with the frequency of 1 were discarded (this was a very careful heuristic cut-off). The frequencies of the tokens were kept throughout the whole process as an important characteristic that is helpful in the manual phase.

The cleaning phase pretty much copied the one from the exploratory run described in 4.1.2 (see the chapter for details and discussion). The following were discarded: all tokens with non-FinnoEnglish characters, including uppercase (to take care of some of borrowings and foreign words from other languages, proper nouns, abbreviations); 1- and 2-character long tokens (functional words, in some corpora OCR errors); tokens with some character clusters or patterns that point towards problematic or erroneous words (same character more than 2 times in a row; 2 or more special characters [:-'] in a row; token starting or ending with a special character).

4.2.3.Comparing with the dictionary; compounds

At this point, the lemma/unrecognized word forms list was compared with the wordlist from Nykysuomen sanakirja and all the recognized tokens were discarded.

Finnish has plenty of compounds; e.g. in Nykysuomen sanakirja, the largest codified dictionary, almost two thirds of entries are compounds. In the corpus, the compounds recognized by the parser were marked by a hashtag (#) in place of the lexeme boundary. But the ones not recognized by the parser still might be made of parts that are in Nykysuomen sanakirja, or at least a part of them could be. That is why all the parts were also checked against Nykysuomen sanakirja and the compounds of which all parts were recognized were also discarded.

This phase ended with a cleaned list of neologism candidates that still carry the frequency information.

4.2.4. Comparison of corpora

At this point, the result of the previous steps -the cleaned lists of neologism candidates from each corpus- still contained tens of thousands of tokens. That was too much to go through manually. But before proceeding to the next step of anglicism identification, the best data source had to be chosen as a basis for it. For that purpose it was a good idea to learn what the lists of candidates were like and how they differed from each other.

To gain some idea about the candidate lists, I proceeded to make simple random samples from each list. The exact size of the neologism candidate list depends on the corpus, but it is usually around 100 000 items. With the population of that size, if a simple random sample of 100 items is chosen, it can be said with 95% confidence that the true population proportions are within a margin of error of 10 percentage points from the sample proportions (counted with the simplified, approximate formula for counting the margin of error $E \approx \frac{1}{\sqrt{n}}$ (Bennett et al. 2014: 326)). That means that even from a relatively small random sample it is possible to gain some useful insight into what the candidate list contains and how many identifiable neologisms there might be.

The evaluation data described in 4.1.5 was used to have another way of finding out how useful a neologism candidate list from a specific corpus might be. The candidates that start with "o" were gathered in order to find out how many of the candidates were also in evaluation data to count the precision, and how many entries from the evaluation data were found among out candidates to find out about recall. These numbers were not very high, but they did help with comparison between the corpora.

The comparison of the two lists can be problematic, because the tokens in the neologism candidate list can be word forms that the parser does not recognize and will not try to guess their lemma. For this purpose I used the Snowball stemmer for Finnish from NLTK (based on Porter 1980), with a slight modification for very short tokens (because especially unadapted anglicisms can be quite short). All neologism candidates were stemmed and compared with stemmed anglicisms in the list. This method is not bulletproof and it can make mistakes, especially when applied to very short tokens. But e.g. in the case of the Wikipedia corpus, it found 23 additional word forms of neologisms (i.e. not found by direct comparison of a candidate or its recognized compound components with the evaluation list; e.g. *outroon*, *ookoolta*, *outletin*) and made only 1 mistake that was actually a typo in the corpus (*onsitten*, with a frequency of 10, which is supposed to be two words *on sitten*, was stemmed as *onsit* and paired with an anglicism *onsite*, a spelling variant of *on-site*).

4.2.5. Results for the available corpora

The differences between corpora are shown primarily by their genre and type; there are news articles and magazines (probably professionally edited), crowdsourced encyclopedia, both professional and fan-made TV program subtitles, and often highly colloquial discussions from an internet forum. How the text for a corpus was gathered is also relevant; some corpora were mainly OCR:d while some were written directly in digital form.

There were some interesting facts to notice about the corpora (see table 5 below). A high amount of unique tokens in one corpus can point to either rich vocabulary or many non-words like typos or OCR errors (or both, in the case of the Lehdet90ff corpus). The amount of unique lemmas in a corpus is uncertain because of the confounding noise, but the amount of lemmas in a corpus that appear also in Nykysuomen sanakirja is easy to find out. It looks like the richest and most diverse language is used in Lehdet90ff and in the Wikipedia corpus. From the sample evidence, I suppose that at least in case of the Wikipedia corpus this is mostly due to a word category that I ended up naming "special domain" when going through the samples. It consists of nouns that do not need to be translated or borrowed to Finnish because most of the speakers have never heard of them, but they are often formally adapted orthographically and for inflection purposes. They are most often scientific terms, demonyms[10], and zoological and botanical generic names, both real and fictional (e.g. *pikti*, *divetysulfidi*, *cardinalis, kaguaani*, *bilbaolainen*, *gammagong*). These words make almost 30 % of the sample from the Wikipedia corpus candidate list, but cannot be considered anglicisms and probably even internationalisms, for their lack of use among the speakers.

Other interesting observations include for example a confirmation of the premise that OCR brings among a huge amount of noise (in the samples, the two corpora produced via OCR of texts show 80 % of noise in the candidate list sample, as opposite to 28-34 % in other corpora), or on the other hand an observation that even professionally edited or crowdsourced text is not at all free from typos (that specific type of noise comprised 19 % of the sample of the Yle News corpus neologism candidate list, and 14% at the one made from the Wikipedia corpus).

The amount of non-Finnish words that are not anglicisms but foreign words in foreign context appears to be as high as 12 % even in some monolingual corpora. Yle News, which had 18 % of

[10] a word that identifies a group of people (inhabitants, residents, natives) in relation to a particular place

foreign language material in the sample, is not completely monolingual, as it contains also Swedish and English texts. When processing this corpus stored in JSON format, I extracted only the elements that were storing Finnish language text, but the process was not necessarily absolutely perfect in that regard.

The OpenSubtitles corpus has some interesting examples of what would have to be called bad anglicisms, calque translations that are not used at all among speakers, because there already is a well-used word for it, anglicism or not. It probably happens easily, especially in fan-made amateur translations of subtitles that certainly make up at least a part of that corpus, but it can also happen when a translator has to come up with a translation that matches what happens in the movie (the example from the sample: *ideahousut*, in context from the corpus *minulla on ideahousut jalassa*). These translations could play a role in creating an impression of established use of certain anglicism that are not actually in use, but in the end there probably are not as many cases present in the corpus that it would really make a difference.

The last interesting observation is regarding the Ylilauta corpus. There appear to be many duplicate texts in it, the reason for which I can only guess. The duplicate (or sometimes even more than two copies) significantly distort the frequencies of the words, especially the rarest ones.

Anglicism-wise, the sample analysis shows that the estimated proportion of anglicisms from all the candidates in the list is highest in Ylilauta corpus (18 %), second place is shared by Wikipedia and OpenSubtitles corpora (11-12 %) and the lowest proportion is in Lehdet90ff corpus (3 %).

This result is at least partly corroborated by the precision measure among O-starting neologism candidates. Out of all O-starting tokens in the corpora candidate lists, the highest proportion of those that were either anglicisms or compounds whose part is an anglicism was found in Wikipedia corpus (12,5 %), then Yle News (9,7 %) and OpenSubtitles and Ylilauta (both 6 %). The difference of results that Ylilauta and OpenSubtitles corpora show is probably due to those compounds; the news corpora clearly contain larger amount of compounds among their tokens. For comparison see e.g. Yle News and Ylilauta corpora, which have almost identical size of their anglicism candidate lists. The lemmas found in Nykysuomen sanakirja from Yle News have a ratio of 1,2:1 between compounds and simple lexemes; in Ylilauta the ratio is 1:1,4 with simple lexemes prevailing. This can be shown also more tangibly: among Yle News corpus O-starting candidates we'd find 36 different compounds that start with *online-*; in Ylilauta corpus there are only 12.

The analysis of the evaluation data on the other hand shows that the Lehdet90ff corpus has the highest coverage: out of the 104 O-starting anglicisms in the evaluation list, 64 were found in it. In this measure, Yle news and Wikipedia corpora share the second place with 54 covered anglicisms, and the worst is the OpenSubtitles corpus with 44 covered anglicisms.

	FNC1 (1950+)	**lehdet90ff (not scientific)**	**Yle News**	**Finnish Wikipedia**	**Open Subtitles**	**Ylilauta**
unique tokens before cleaning	10239948	8218373	5526072	3898553	3232734	1485378
lemmas[11] when foreign clusters removed	not prelemmatized	5400461	not prelemmatized	2353598	1557194	932423
lemmas[11] after cleaning	606377	846447	449175	393126	394974	295060
lemmas found in Nykysuomen sanakirja	82393	69270	67871	44624	76577	82887
not in NS (unique tokens in neologism candidate list)	329642	461805	189561	129632	147818	178665
estimate of anglicism proportion	4 %	3 %	9 %	11 %	12 %	18 %
estimate of noise proportion	80 %	79 %	44 %	28 %	36 %	43 %
estimate of unwanted foreign words proportion	12 %	4 %	18 %	9 %	5 %	2 %
O-starting anglicisms, precision	4,3 %	2,3 %	9,7 %	12,5 %	6,0 %	6,0 %
O-starting anglicisms, recall	43,3 %	61,5 %	52 %	52 %	42 %	47 %

Table 5: Corpora compared (darker is better)

Regarding the evaluation data, 6 of the anglicisms/internationalisms in the list were not found in any of the corpora (e.g. *outsourcata*, *outata*, *op-taide*). All of them are attested in some other Language Bank's corpora, even though with very low frequency, and are marked as "rare" in the GLAD database.

4.2.6. Picking up the corpus/corpora for the n-grams comparison

The criteria for the best data source for the next step in the process were good coverage of known anglicisms and relatively low amount of noise and foreign words that appear in foreign contexts.

The analysis in the previous subchapter implies that to find the highest amount of anglicisms no matter the amount of work, Lehdet90ff is the right corpus. But the manual work would be highly

[11] Lemmas here mean both lemmas of words recognized by the parser and unique unlemmatized tokens which the parser did not recognize.

ineffective, because that corpus has the largest candidate list and around 80% of it is noise. Out of the other corpora that have decent coverage (Wikipedia and Yle news, 51,9%), Wikipedia was the one with lower amount of noise and also lower amount of foreign words in foreign context, so it seemed that to make the manual work as effective as possible, Wikipedia corpus would be good dataset to go with.

Because having a good recall of anglicisms is important for this work, I decided to combining two corpora together. In the selection of available corpora there very different genres: newspaper language, discussion forum, encyclopedia language, subtitles… combining them could lead to better coverage without necessarily adding too much noise to the equation. As is visible from the table 6 below, the best combination regarding coverage includes almost always the single-corpus coverage winner, the Lehdet90ff corpus. But as that one brings along a lot of noise (similar to the FNC corpus), the ideal combination would be Wikipedia+Ylilauta. For comparison, the coverage of all six corpora combined would be 94 % (see the subchapter 4.2.5. above).

	lehdet90	**FNC**	**yle**	**ylilauta**	**subs**	**wiki**
lehdet90	64	71	70	74	69	72
FNC		45	63	69	63	68
yle			54	67	62	68
ylilauta				49	60	69
subs					44	62
wiki						54

Table 6: Anglicism coverage in combined corpora

Ylilauta also has the lowest estimated amount of foreign words in foreign context from all the corpora (2%) and the amount of noise is around the average for all considered corpora (43%).

5. Ordering the list: N-grams comparison approach

After deciding for a specific corpus or corpora, cleaning the data and extracting the neologism candidate list, the final task was to sort the candidate list in a way that would put more probable candidates for anglicisms higher in the list.

5.1. Method description

Candidates were each assigned a score that defines the order in which they are presented to the human evaluator, so that the candidates considered as the ones with the higher probability to be an established anglicism come earlier.

The idea was to assign each candidate a score based on:

- its frequency in the source corpus, which reflects how much is the word used and necessarily correlates up to some degree with how established the word, if not in general, then at least in the genre of the corpus
- the relative frequencies of the candidate's character level n-grams in a Finnish n-gram list (a number between 0 and 1), which should reflect how typically Finnish the word is – with highest scores for fully Finnish words (most of which should not be in the candidate list because they've been removed in the step of the process that excluded word forms from Nykysuomen sanakirja) and then borrowings adapted to Finnish phonologic/graphemic system
- the relative frequencies of the candidate's character level n-grams in an English n-gram list (a number between 0 and 1), which should reflect how typically English the word is – with highest scores for unadapted borrowings from English (or, unfortunately, actual English words that were in the corpus in English context) and then adapted borrowings, the form of which still points towards their English origin.

As an input, the following datasets were needed:

- an anglicism candidate list, created by merging the candidate lists from the Wikipedia ad Ylilauta corpora; it contains wordforms along with their frequencies in the said corpora
- list of character level n-grams extracted from a list of fully native-like Finnish words with their relative frequencies
- list of character level n-grams extracted from a list English words with relative frequencies

The candidate list with frequencies was ready, a result of the process described in the previous chapter.

5.2. Finnish and English character level n-gram list

The list of character level n-grams extracted from Finnish words needed to be prepared from words that are certainly Finnish, without any formal features that would suggest that they are borrowings. First I was thinking about just using Nykysuomen sanakirja for this purpose, but I decided against it: Nykysuomen sanakirja also contains borrowings (mostly borrowings from Swedish and internationalisms, but nevertheless), and more importantly, it only contains lemmas, which means that the n-gram list learned from it would not contain any n-grams that reflect Finnish inflection. It also does not contain any information about the frequency of use.

I decided to pick a dataset with relatively low content of foreign words in foreign context, without too much noise (and especially without OCR-induced noise), which is presented in the VRT format (i.e. known words are pre-lemmatized); the Wikipedia corpus. The idea was to choose from the corpus all such tokens that are either lemmas or know inflections of the lemmas that are in Nykysuomen sanakirja (i.e. it can be said with certainty they are not noise), and then remove those which, judged from their graphemic form, cannot be direct borrowings. The rest was then divided into character level n-grams of different order and their frequencies are counted.

To create this uniquely Finnish word list a set of regular expression rules was needed that would serve for identifying clearly non-fully native-like Finnish words by their graphemic form. The following subchapter describes how it was prepared.

5.2.1. Finnish and English graphotactic information

English and Finnish are phonologically speaking very different languages, which can be used to our advantage in the hunt for anglicisms. Because we're working with text, we will translate the phonotactic rules into graphotactics. Thanks to the fact that Finnish orthography is based on the alphabetical principle, which means that its phonemes correlate significantly with its graphemes (Suomi, Toivanen & Ylitalo 2008: 141), it is not too complicated a process.

It's important to note that the graphotactic information I attempt to describe here belongs to a partially idealized, conservative picture of Finnish language without external influences. This does not correspond with what we know about language; languages evolve constantly and unless their speakers live in total isolation, they keep accepting new influences. On the other hand, standard Finnish has kind of helpful history in this regard; in the 19th century the nationalist movement attempted to "finnicize" Finnish by disposing of Swedish loanwords and grammatical structures (Karlsson 1999: 3). In addition, many external influences were in the course of time adapted to the constraints of the graphotax as I present it here (Suomi, Toivanen & Ylitalo 2008: 53). That makes it easier for this work to choose this snapshot of Finnish in time as the -albeit idealized- starting point and categorize any later influences as neologisms.

I identify the features of traditional (in the meaning explained by the previous paragraph) Finnish graphotax that can be used for checking whether a word can potentially be a "fully native" (term used in Suomi, Toivanen & Ylitalo 2008) word. Those features are then used to help create a Finnish wordlist with formally distinguishable borrowings removed. There will certainly be borrowings in the list that are not formally distinguishable by these rules (e.g. *enkeli*, *meijeri*, *tentti*…), but the rule is absolutely valid in the opposite direction, i.e. a word that fits at least one of the rules is certainly a borrowing, a non-fully native Finnish word.

All the information in the next subchapter is based on Suomi, Toivanen & Ylitalo (2008: 20-38, 49-64) about Finnish phonological system and adapted to serve as information about graphotax. Each applicable rule is followed by a regular expression that serves for finding tokens that break the allowed pattern.

5.2.2. Non-fully native graphotax

All vowels used in English graphotax can be also used in Finnish graphotax. Finnish has two extra vowel graphemes, ä and ö, that can be nevertheless used in borrowings when the phoneme corresponds to the one used in the pronunciation of the foreign etymon (e.g. *laptop* -> *läppäri*). Only *å*, which appears exclusively in unadapted borrowings from Swedish, can serve as an indicator of a non-native Finnish word. Diphthongs are similarly unhelpful, because even though there are some vowel combinations that do not appear in native Finnish roots (e.g. *eö*), they can well enough stand next to each other in a compound (e.g. *koneöljy*). Also sequences of many vowels are allowed in Finnish (case in point: *hääyöaie*) and will not be helpful in this regard.

Unfortunately, vowel harmony cannot be used for this purpose either; there are not many words in Finnish that would break vowel harmony (and those certainly are borrowings), but there is an abundance of compounds where lexemes with front vowels and lexemes with back vowels are freely combined. For that reason, detecting tokens that break vowel harmony is not useful. A slightly more complex algorithm could check if a token seems to have a case ending and whether such case ending observes the rules of vowel harmony – but breaking this rule points more towards a spelling or OCR error, not towards a borrowing.

Some consonants that are common in contemporary Finnish were not allowed at all in the native Finnish. Their respective graphemes are *b*, *f*, *š*. Also graphemes *c*, *q*, *w*, *x*, *z* point towards a borrowing.

```
^.*[bfšcqwxzå].*$
```

Some consonants' distribution is restricted; e.g. *d* can only appear in an intervocalic position or between *h* and a vowel.

```
^.*[^aeiouyäöh]d.*$
```

```
^.*d[^aeiouyäö].*$
```

Similarly, *g* can only appear after *n*.

```
^.*(?<!n)g.*$
```

Consonants *v* and *j* can only be followed by a vowel, never a consonant.

```
^.*[jv][^aeiouyäö].*$
```

There are also some not-allowed consonant clusters independent on their position in the word: *pm, pn, km, kn, ph, kh, pv, pj, kv, kj, mh, mj, ms, pt, pk, kp, kt*. There would be more, but some of clusters that would otherwise belong to this group can, again, appear on the boundary between compound components even in fully native words, e.g. *tn* (*lyhytnäköinen*), *sr* (*ihmisranvinto*), *lr* (*nivelrikko*), etc.

```
^.*(kh|kj|km|kn|kp|kt|kv|mh|mj|ms|ph|pj|pk|pm|pn|pt|pv).*$
```

All CCCC clusters are prohibited.

```
^.*[dghjklmnprstv][dghjklmnprstv][dghjklmnprstv][dghjklmnprstv].*$
```

CCC clusters are allowed if the first consonant is a liquid or a nasal (*l, m, n, r*) and the two last consonants are obstruents (*k, p, s, t*).

```
^.*[dghjklmnprstv][dghjklmnprstv][dghjlmnrv].*$
```

```
^.*[dghjklmnprstv][dghjlmnrv][dghjklmnprstv].*$
```

```
^.*[dghjkpstv][dghjklmnprstv][dghjklmnprstv].*$
```

In the fully native Finnish graphotax word-initial consonant clusters are not allowed. Borrowings that precede our idealized snapshot in time whose etymon starts with a consonant cluster were adapted (e.g. Swed. *spel* -> Fin. *peli*); in the newer borrowings that we'd like to identify this adaptation has not happened (e.g. *sponsori*).

```
^[dghjklmnprstv][dghjklmnprstv].+$
```

In the word-final position --VC the only consonants allowed are *t,s, n, l, r*.

```
^.+[dghjkmpv]$
```

Word-final consonant clusters CC appear in some onomatopoetic interjections, and it has to be a plosive (*p, t, k*) followed by *s*.

```
^.+[dghjklmnrv]s$
```

And finally, in fully native words word-initial positions *je-* and *ji-* appear almost exclusively in borrowings; *vu-* can only appear if followed by *o*.

```
^(je|ji|vu[^o]).*$
```

5.2.3. Finalizing Finnish character level n-gram list

A script was made that chose from the corpus all the word forms the lemmas of which are in Nykysuomen sanakirja and which pass the fully native-like graphotax test made with the rules described in the previous subchapter. Then it divided each word form to character level n-grams with padding on both sides (padding is useful to distinguish between n-grams from the middle of the word and from the beginning or end of the word) and counted the n-grams' relative frequency in the cleaned version of the corpus. Regarding the character level n-gram order, I decided to go with 2- to 4-grams and not more for computational reasons (processing time -wise). The result of this process was a list of 2-, 3- and 4-grams made of fully native-like Finnish word forms with relative frequencies which preserved the information about their frequency in the original corpus.

5.2.4. English character level n-gram list

The requirements for the English n-gram list were similar to those for the Finnish one, although here there was no reason to make the source list in any way innately English. Here I did not need to prepare my own frequency wordlist as I as able to choose a source that already covers word forms and not only lemmas, and it also carries information about frequency of use. The source I used was the British National Corpus (BNC) frequency wordlist[12].

This wordlist was prepared from the full BNC and published by Leech, Rayson & Wilson in 2001. In its full form it contains more than 750 000 different word forms. I removed all tokens with frequency below 50 and dispersion (i.e. the measure of how evenly a word is distributed in the corpus) below 0,3 to weed out too rare and too specialized word forms. I also removed all tokens containing uppercase characters, and in general characters that do not appear in native English words. This way the final wordlist contained only approximately 32 000 wordforms (which still cover 86% of all the corpus tokens) with their respective frequencies. Then I followed the same process like with the Finnish list and created a list of 2-, 3- and 4-grams with their relative frequencies.

5.3. Scoring and evaluation method

All words in the candidate list were transformed into a list of character level 2-, 3- and 4-grams and then assigned a score based on the relative frequencies of the n-grams in the English and/or Finnish n-gram lists, and on the candidate's frequency in their source corpus. I experimented with various

[12] available at http://ucrel.lancs.ac.uk/bncfreq/flists.html. This wordlist is almost 20 years old, but a cursory search showed that similarly large frequency wordlists for English newer than this one are subject to a charge

methods of scoring based for a large part on Andersen (2005)'s article and on several intuitions about the relationship between anglicisms and the English and Finnish n-gram lists.

To evaluate different scoring methods, I again turned to the list of known O-starting anglicisms. If the candidate list is sorted fully randomly and then a ranking is assigned to each candidate according to the order in the list, the average rank of the known O-starting anglicisms should lie close to the average rank of all the candidates. The candidate list created from the combined Wikipedia and Ylilauta corpora contains 287 000 unique candidates, so in a randomly sorted list the average rank of the known O-starting anglicisms would be close to 143 500. If around 15 % of the candidate list consists of anglicisms (average for the combination of Wikipedia and Ylilauta corpora, as suggested by the samples' analysis for each corpora, see chapter 4.2.5), then if the list was ordered absolutely ideally with all the anglicism word forms on top, the rank of all known O-starting anglicism would be close to the average rank of the first 15% of the candidate list (i.e. first 43 050 candidates, so the average rank would lie close to 21525).

evaluation by O-starting known anglicisms (O-test)	result
ideal order of candidates, better than random order of O-starting anglicisms	< 21525
ideal order of candidates, random order of O-starting anglicisms	21525
better than random order of candidates	21525 - 143500
random order of candidates	143500
worse than random order of candidates	> 143500

Table 7: Possible evaluation results, O-test

On the other hand, it would be important that more frequently appearing candidates appear higher in the resulting list. To have an idea how the ordering of the lists affects them, I went through the top 150 tokens in the candidate list and picked the 64 highly probable anglicisms (only highly probable, not certain, because for some of them are unadapted and so it is possible that they appear in the corpus only in foreign language context). Their average rank in frequency-sorted candidate list is approximately 75,5 (average of 1…150); in an ideally sorted candidate list their average rank would be close to 32,5 (average of 1…64). In a randomly sorted candidate list their average frequency would be, like in the O-starting anglicism measure above, around 143 500.

evaluation by most frequent anglicisms (Freq-test)	result
ideal order of candidates	32,5
frequency order as in the list of candidates	75,5
better than random order of candidates	75,5 - 143500
random order of candidates	143500
worse than random order of candidates	> 143500

Table 8: Possible evaluation results, Freq-test

The intuition behind the scoring method is that anglicisms would get a high score (translated to low rank in the resulting list) when scored with the character level n-grams from the English list; especially the unadapted ones, as their written form matches the English etymon (which may or may not be present in the BNC corpus from which the English n-gram list was made). But also the adapted anglicisms could get high score, as many of them still formally resemble their etymons, together with compounds which consist of mixed origin lexemes. By this logic, adapted anglicisms and mixed compounds should get high score also when compared with Finnish n-grams. Then again, typos, OCR-errors and non-English foreign words or borrowings should get lower score in both cases, as they would often contain n-grams that are not present in either list. And finally, higher frequency in the candidate list should translate into a better (lower) rank, but I experimented with how much the frequency would affect the scores.

I also tried out several modifications of the above described scoring procedure that are based on Andersen (2005)'s observations on what worked for Norwegian anglicisms. Those modifications included the following:

- using only unique lists instead of full n-gram lists (not using n-grams that appear in both English and Finnish lists)
- "cutting the tail" of the full n-gram lists (not using low frequency n-grams in the lists)
- using only higher order n-grams, mostly 4-grams
- using only English n-gram list

5.4. Results

The intuition regarding the scoring method described in the previous part did not work as well as I had hoped.

I first tried some experiments to see the effects of the scoring method components separately. Scoring the list purely on the Finnish n-grams and without frequency affecting the results led to clearly worse-than-random results in both evaluation methods (that was expected), and it put to the top of the list mostly very short tokens that consist of only the most frequent Finnish n-grams. Those tokens often belonged into the noise part of the list: frequent misspellings or second parts of divided words (examples: *sista*, *kaan*, *mista*, *tasta*, *sita*, *tinen*, *kaisten*…). When using purely the English n-grams, the top results were short and frequent English words (*the*, *and*, *that*, *there*, *this*…). This scoring method version would have worked better if it were not for English-language texts in the corpus. The evaluation results were better and there were definitely also unadapted anglicisms among the top of the list (*single*, *online*, *software*, *blues*…), but they were mixed among the frequent English words that appear among English language texts.

Using the unique n-grams' lists had moderately positive effect on the evaluation scores, but it lead to a situation where tokens are scored based only on the basis of a couple of n-grams (example: *biomassa* contains only one n-gram that is in the unique list of Finnish n-grams, "ssa$", and would get score equal to the score of that one n-gram). In practice that pushed to the top of the list the tokens that contain only one unique, frequent and highly scored n-gram, because all other n-grams would be ignored as they are not in the unique Finnish n-gram list. For example, when scored only with unique Finnish n-gram list and moderate frequency influence, the beginning of the list looked like this: *ssa*, *biomassa*, *schoolissa*, *erossa*, *steamissa*, *selaimessa*, *somessa*, *darrassa*.

Cutting the tail, that is to use only the more frequent n-grams for scoring, did not seem to affect the results much. That's probably because the scoring method already gived substantially more weight to the more frequent n-grams. Using only higher order n-grams was similar, it did not affect the results a lot, even if it had a slight positive effect on the evaluation results.

The balancing of the grade of influence of the candidates' frequency in the corpus was difficult. It seemed to always play either too big or too small role and I did not manage to tune it to get desired behavior. Here I suppose that more research regarding possible scoring methods or potentially more mathematical background would have helped.

The most effective method out of all those tried out with this scoring system was the one that used only unique English 4-grams, which agrees with Andersen (2005)'s findings for Norwegian anglicisms. Combining the Finnish and English n-gram scoring did not bring better results than using exclusively English n-gram scoring, especially if evaluated on O-starting known anglicisms.

After trying out the above-mentioned scoring methods, the results did not seem very good, both the O-test/Freq-test results and what the content of the upper part of the actual ranked list of candidates looked like. Either the intuition on which the scoring was based must have been wrong, or the scoring system itself should have been designed more carefully to actually reflect the intuition. Using the relative frequency of the tokens in the corpus did not seem to be very balanced, especially when the effects of Finnish and English n-gram scores were combined.

If the problem was the balance in combination of the two scores, it would have probably been better to use ranking in the list rather than directly the result of the n-gram score directly. To try this idea, I identified the experiments that had best results for English and Finnish n-gram scoring method separately, took the ranked lists obtained as the results, and combined them. This way the unbalanced scoring method between the two languages did not affect the result because its combined result depended purely on the rank of the candidate in the two lists. Evaluation scores show that this direction had somewhat better results, and that the combined ranking had an enormous positive effect on the ranks of the more frequent anglicisms, as can be seen in the table 9 with the evaluation test results below.

no.	method	O-test result	Freq-test result
1	Finnish n-grams, without frequency of candidates	195689	170050
2	Finnish unique n-grams, without frequency	188212	167366
3	English n-grams, without frequency	92651	100177
4	English unique n-grams, without frequency	74826	76100
5	English unique 4-grams, with frequency	70518	68725
6	Finnish & English unique n-grams, with frequency	109632	76135
7	methods 2 and 5, combination of ranked results	70400	567

Table 9: Evaluation results

6. Discussion and conclusion

6.1. Discussion of the results

The goal of the process described in chapter 5 was to reorganize the neologism candidate in such way that candidates with the most potential to be anglicisms would be positioned towards the beginning of the list. This was proven quite difficult to do; the best result obtained in the process was significantly better than random (or for example alphabetical) order of candidates, but it was not very close to the ideal. One of the reasons for that was already discussed above; the method of scoring. There should have been more time dedicated to exploring suitable scoring methods, and if a more suitable would not have been identified, then the method by which the best result was obtained, which combined ranks instead of n-gram scoring method scores, should have been explored more.

In retrospect it would have been also smart to get rid of the lowest frequency items (freq = 2 or 3) from the candidate list before starting and worry less about losing potential candidates. The amount of noise is so large that getting rid of a substantial portion of it would be worth losing a couple of low-frequency (i.e. not very well established) anglicisms.

Another potential approach to this problem could be to first group the candidates into clusters, using an algorithm that could include a stemmer or a parser that unlike *finnish-parse* also tries to guess lemmas of unknown tokens. This grouping could cluster neologism candidates related to each other; those could be for example inflected forms of the same lemma, or not completely settled orthographic variants of one borrowing, or even compounds with the same first component (there are e.g. tens of compounds that start with *online*-, and around ten compounds that start with *optio*-, *organisaatio*-, or *optimointi*- in just the candidate list created from the Wikipedia corpus). This could be either done already before scoring and just e.g. the most frequent variant could be scored, or then it could be done after and the all variants' scores could be combined.

The approach that I believe would have helped the most is better cleaning of the foreign language parts from the original corpus (see chapter 2.2.2.), that is, when context is still available for each token. It is probable that the English n-gram scoring results would be much better if it was not for all the English words contained in the corpora that also end up in high ranks. In general, it can be said that any approach that manages to lower the amount of noise in the corpus to begin with would be certainly rewarded with significantly better results regarding anglicisms.

6.2. Ideas for other approaches

Here I discuss a few approaches that could improve the results of the method examined in this work with more time or more resources. For example, if more already found anglicisms were available, they could be used as training data for some kind of machine learning approach. There could be used for machine learning also in the scoring of the neologism candidates.

Other languages' anglicism lists could be used to help search for their Finnish counterparts in several ways. As mentioned in the chapter 2.3. on related research, in the part dedicated to OCR recognition, e.g. in information retrieval there's need for approximate string matching. I considered that method for getting rid of noise in the corpora and rejected it, because I believe it would also remove a part of anglicisms. On the other hand, if we turn the point of view, we could say that direct adapted anglicisms could be seen as approximate matches of their English etymons. Järvelin et al. 2016 describes a method called s-gram matching, which would be interesting to try out for this purpose. It is basically a variant of n-gram matching, but while n-grams are created of adjacent characters, s-grams are allowed to skip a certain number of characters. Thus when we create e.g. 2-grams with padding from *hello*, we get {h, he, el, ll, lo, o}, while if we created s-grams of the same length and with allowed skip length equal to 0 and 1 from the same word, we'd get {$h, e, he, hl, el, ll, lo, l, o$}. I think s-grams could be useful for some specific methods of anglicism harvesting where we know what we search for. We could e.g. use the GLAD database at its current state, make a list of all the etymons that are there (i.e. list of all words that were borrowed into at least one of the 16 languages covered in the database) and find all tokens in our neologism candidate list that would match them.

Finding a new, less noisy corpora would be an obvious way to improving the results. It is possible that even small systematic improvements in the corpora used in this work could also help. According to the corpus sample analysis, for example in Lehdet90ff corpus, which has a great coverage of the O-starting anglicisms and would be a great anglicisms source if it were not so noisy, 65% of unknown words come from words divided on the end of a row that the OCR tool understood as two separate items. In general, OCR post-correction is probably a counterproductive measure for finding neologism as it can mask them by considering them a badly OCR:d known word (see chapter 2.3.), but this kind of fix would probably be improvement even for neologism harvesting.

Another approach could be to add to the corpora selection some freely available, albeit very noisy sources of anglicisms. A Finnish equivalent to English language Urban Dictionary[13], Urbaani Sanakirja[14] is an online crowdsourced dictionary where anyone can add a slang word or phrase, its definition and examples of use. It contains more than 50 000 terms as of this moment and even though most of it will be either outright nonsense or expressions that only belong to random idiolects, it also contains neologisms (mostly anglicisms) in different phases of being established.
A script that would go through the list of words on Urbaani Sanakirja, find their frequencies in a Finnish corpus with reasonably recent content and gather those with non-zero frequency in the corpus could provide a list of interesting anglicisms that one will not find anywhere else in one place.

6.3. Speeding up manual work

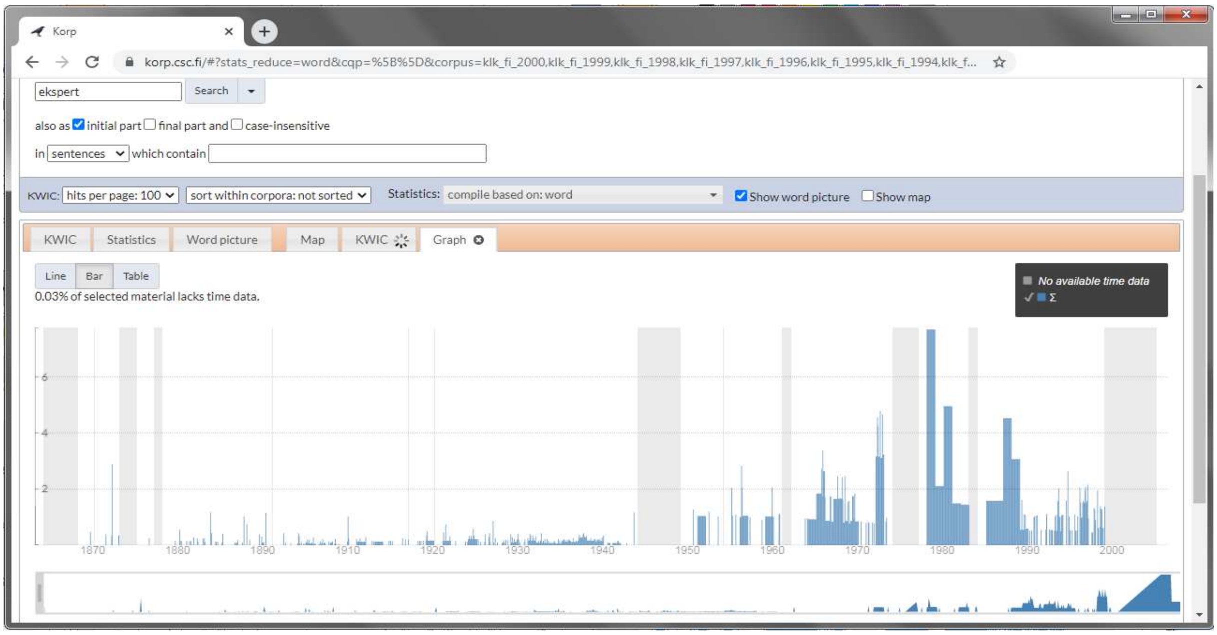
One way to make manual work more effective would certainly be to create an interface that gathers on one screen all the available information about the word that is being processed, enables easy moving between words in the queue to be processed, and after the decision about the status of the word is made, makes it easy to assign it to its respective group. Such graphical interface was used for example in Säily, Mäkelä and Hämäläinen (2018), where it shows on one screen the active word in a queue, a dictionary entry on the word with the date of its first attestation, and its examples from the

[13] https://www.urbandictionary.com

[14] https://urbaanisanakirja.com/

corpus. In addition, the word is grouped together with other words with the same key created by a spelling normalization algorithm commonly used in that field of research.

In this work, some kind of a graphical interface like that would also certainly speed up both the processing of the queue of words and the deciding about the status of the word. Context shows how exactly is the word used in the relevant corpus; in addition, if there happens to be a lot of other language material mixed up with Finnish in the corpus (and no reliable way of getting rid of it), checking context is a fast way to decide if the word appears as a borrowing, and not only in a code-switching situation. If there are many appearances of that word in the corpus and the other-language material abundance is a problem, I suggest that some simple rating of the word's context to measure its "Finnishness" could be used to present the "most Finnish" context first.



korp.csc.fi/#?stats_reduce=word&cqp=%5B%5D&corpus=klk_fi_2000,klk_fi_1999,klk_fi_1998,klk_fi_1997,klk_fi_1996,klk_fi_1995,klk_fi_1994,klk_f...

ekspert

also as ☑ initial part ☐ final part and ☐ case-insensitive

in sentences which contain

KWIC: hits per page: 100 sort within corpora: not sorted Statistics: compile based on: word ☑ Show word picture ☐ Show map

0.03% of selected material lacks time data.

Table 10: Trend diagram of words that start with 'ekspert' in the KLK-fi corpus.

Finding first appearance of the neologism candidate in the available corpora would also be helpful and its availability can speed up manual work significantly. This could include searching for the candidate's context in the corpora used in this work, but it could be useful to have a look at other corpora, e.g. the full Newspaper and Periodical Corpus of the National Library of Finland (KLK-fi) because it contains newspaper texts starting from 1820 (in this research we only used the n-gram list version of this enormous corpus, FNC1). Korp corpus search interface (https://korp.csc.fi/) provides a trend diagram, a graph of a word's use in time which would be greatly helpful, but unfortunately its generation from a corpus large enough to have enough instances takes a long time. Maybe it would be possible to use Korp web API to preload such graphs so that they would be available during the manual processing of the queue, but the context also needs to be checked, because very often the putative first example can be just a typo or another badly OCRed word.

With or without a graphical interface, grouping of tokens could certainly help speed up the processing. In Säily, Mäkelä and Hämäläinen (2018), the words are grouped together with the help of a spelling normalization algorithm used specifically for Early English in order to get all the spelling and dialectal variants together. This research could benefit from grouping tokens several parts of the process, for example it could be used in a corpus token list to try to find all typos of a certain word (and then keep just the version with the highest frequency as the correct one). To try this out, I experimented with a combination of a stemmer that would help group together word forms that the parser does not try to parse and a minimal edit distance measurement that could group together typos with the intended word. My experiments with a self-written algorithm that used NLTK's Snowball stemmer and a basic Levenshtein distance of 1 had promising results. The problem was that my algorithm was not scalable because it had order of n^2 time complexity. For that reason I only applied it on relatively short lists (max. thousands of tokens). When applied on A-starting neologism candidates from the OpenSubtitles corpus, it grouped 7200 items to 4700 groups (3700 of them were still groups of one, 1000 groups had between 2 and 24 items). The algorithm worked well for tokens with 5+ characters; it would need further improvements for handling the 3-4 characters long ones. Other approaches to grouping (more efficient ones) could be also tried out, e.g. some kind of machine learning method for cluster analysis, but without any experience with that it is difficult to ascertain how precise such methods would be, especially regarding grouping together inflected word forms.

6.4. Conclusion

In this work, I attempted to identify a way to partially automatize anglicism harvesting in Finnish corpora and thus make it more effective. In response to the research questions:

What kinds of data sources suitable for this goal are available, and what would be the criteria for a useful data source? Several of Language Bank's Finnish language monolingual corpora were considered and their pros and cons explored. The most important criteria were identified to be the size and genre of the corpus and its available pre-made annotation, which were explored from the description of corpora on Language Bank's website and available literature, and because those were often brief and insufficient, also by hands-on examination of the data. Other important measures were the amount of unannotated foreign language material, amount of other noise (for example errors introduced by OCR and typos), and potential anglicism proportion in the corpora. This information was gained via meticulous exploration of random samples of the corpora neologism candidate lists and evaluation on previously gained anglicism set. A combination of two corpora with good coverage of known anglicisms and relatively low amount of noise was chosen as the dataset for the next phase of the anglicism identification process.

How to use a data source like this to prepare a good list of anglicisms candidates so that there would be as little irrelevant material as possible but so that no anglicisms would not be lost in the process? Candidate lists were prepared by a process of removing tokens irrelevant for anglicism harvesting. That includes an identifiable part of foreign language material in the corpus, formally recognizable noise, known lemmas of the words that were present in Finnish language around the time just before the major influx of English borrowings to Finnish language started, and their inflected forms.

How could the candidates be scored so that the more probable anglicisms would appear closer to the top of the list? Several methods of scoring candidates were devised that would hopefully assign better score to tokens with higher probability to be an anglicism. The score is based on tokens' frequency in the corpus and relative frequency of the character level n-grams made out of those tokens in representative purely English and purely Finnish corpora. The tokens in the candidate list were scored and ordered, and the resulting list was evaluated based on the ranking of a set of previously identified anglicisms. The method was proved to be somewhat effective; the resulting average ranking of known anglicisms was better than it would be in a randomly sorted candidate list, but it was also worse than it could be in ideal case.

Based on the observation of the data and the obtained results, it is my opinion that better results could be gained particularly by improving two elements of the process; one of them is a better method of language identification of short text, which could be used for both cleaning the corpus from foreign language material and for pinpointing tokens that have a higher chance to be a borrowing by assessing their and their context's language. The other would be a more careful design of the scoring method.

Hopefully this work will be beneficial to the GLAD project. The neologism candidate lists will be made available for further manual processing to anyone who shall continue with harvesting Finnish anglicisms for the project. Moreover, the process of corpus cleaning and neologism candidate list scoring is replicable and easy enough to do, and should a more suitable Finnish language become available or should any of the corpora used in this work be improved regarding the amount of noise, for example by better OCR post-correction method, it can be run again with results that can be only improved. The previous subchapter regarding speeding up manual work can also provide some ideas that could make work on anglicism harvesting more efficient.

7. LIST OF REFERENCES

- Alex, Beatrice 2008*. Automatic detection of English inclusions in mixed-lingual data with an application to parsing*. Doctoral dissertation, University of Edinburgh.
- Algeo, John 1991. Fifty Years among the New Words: A Dictionary of Neologisms 1941-1991 *Centennial Series of the American Dialect Society*. New York: Cambridge University Press.
- Haarala, Risto, and Irma Nissinen 1994. *Perussanakirjan uudissanat.* Kielikello 3. Available at https://www.kielikello.fi/-/perussanakirjan-uudissanat
- Andersen, Gisle 2005. *Assessing algorithms for automatic extraction of anglicisms in Norwegian texts*. Proceedings of Corpus Linguistics 2005. Birmingham: University of Birmingham.
- Andersen, Gisle 2012. Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Furiassi, Cristiano, Virginia Pulcini and Félix Rodríguez González (eds.) 2012. *The anglicization of European lexis*. John Benjamins Publishing.
- Andersen, Gisle & Knut Hofland 2012. Building a large corpus based on newspapers from the web. In Andersen, Gisle (ed.) 2012. *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian.* Vol. 49. John Benjamins Publishing. (Andersen & Hofland 2012: X)
- Battarbee, Keith 2002. Finnish. In Manfred Görlach (ed.): *English in Europe*. Oxford University Press. Retrieved on 16 Aug 2020 from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199273102.001.0001/acprof-9780199273102-chapter-14
- Bennett, Jeffrey O., William L. Briggs and Mario F. Triola 2014. *Statistical reasoning for everyday life*, 4th edition. Pearson.
- Borin, Lars, Markus Forsberg and Johan Roxendal 2012. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA, pages 474–478.
- Cavnar, William B., and John M. Trenkle 1994. "N-gram-based text categorization." *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175.
- Eronen, Riitta 2007. *Uudissanat rötösherrasta salarakkaaseen*. Otava.
- Geeraerts, Dirk 2015. How Words and Vocabularies Change. In (Ed.), *The Oxford Handbook of the Word: Oxford University Press*. Retrieved on 11 Aug. 2018 from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199641604.001.0001/oxfordhb-9780199641604-e-026
- Gottlieb, Henrik 2019. *GLAD newsletter # 6 – fall 2019*. Retrieved from https://www.nhh.no/globalassets/centres/glad/glad-newsletter-6-december-2019.pdf
- Gottlieb, Henrik, Gisle Andersen, Ulrik Busse, Elżbieta Mańczak-Wohlfeld, Elizabeth Peterson and Virginia Pulcini 2018. Introducing and developing GLAD : The Global Anglicism Database Network. *The ESSE Messenger* 27(2): 4-19.
- Görlach, Manfred 2003. *English words abroad.* Vol. 7. John Benjamins Publishing.
- Görlach, Manfred (ed.) 2001. *A dictionary of European anglicisms: A usage dictionary of anglicisms in sixteen European languages.* Oxford University Press on Demand.
- Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski & Filip Ginter 2014. Building the essential resources for

Finnish: the Turku Dependency Treebank. *Lang Resources & Evaluation* 48, no. 3, 493–531. https://doi.org/10.1007/s10579-013-9244-1

- Hiltunen, Turo 2014. Choice of national variety in the English-language Wikipedia. *Studies in Variation, Contacts and Change in English*, vol. 15.
- Jauhiainen, Tommi, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research* 65: 675-782. Available at https://arxiv.org/pdf/1804.08186.pdf
- Järvelin, Anni, Heikki Keskustalo, Eero Sormunen, Miamaria Saastamoinen & Kimmo Kettunen 2015. Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach*. Journal of the Association for Information Science and Technology.*
- Karlsson, Fred 1999. *Finnish – an essential grammar*. Routledge.
- *Korp corpus input format.* (n.d.) The Language Bank of Finland. Retrieved on 12.10.2020 from https://www.kielipankki.fi/development/korp/corpus-input-format/
- *Korpin korpusannotaatio: TDT.* (n.d.) The Language Bank of Finland. Retrieved on 12.10.2020 from https://www.kielipankki.fi/tuki/korp-tdt/
- Kotimaisten kielten keskus (n.d.). *Heittomerkki*. Retrieved from http://www.kielitoimistonohjepankki.fi/ohje/10
- Kotimaisten kielten keskus 2006. Kaksoispiste :. *Kielikello 2/2006.* Retrieved from https://www.kielikello.fi/-/kaksoispiste-
- Kotimaisten kielten tutkimuskeskus (Kielitoimisto) 1979. *Uudissanasto 1980*. Söderström.
- Käenmäki, Katriina 2019. *"Bombertakki fotoprintillä" - The Translation of Anglicisms Back into English in Finnish Online Stores for Children's Clothing.* Master's thesis. University of Vaasa.
- Leech, Geoffrey, Paul Rayson & Andrew Wilson 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus.* Longman, London.
- Lison, Pierre & Jörg Tiedemann 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 923-929.
- Losnegaard, Gyri Smørdal and Gunn Inger Lyse 2012. A data-driven approach to anglicism identification in Norwegian. In Andersen, Gisle (ed.): *Exploring Newspaper Language. Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*, 131–154. Amsterdam: John Benjamins.
- Lyse, Gunn Inger & Gisle Andersen 2012. Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text. In Andersen, Gisle (ed.): *Exploring Newspaper Language. Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*, 79–110. Amsterdam: John Benjamins.
- McEnery, Tony, Richard Xiao & Yukio Tono 2006. *Corpus-based Language Studies: An Advanced Resource Book.* Routledge.
- McEnery, Tony, and Andrew Hardie 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Mäkelä, Eetu, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mikko Hämäläinen, Samuli Kaislaniemi & Terttu Nevalainen 2020. Wrangling with non-standard data. In Reinsone, Sanita, Inguna Skadiņa, Anda Baklāne & Jänis Daugavietis (eds): *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference: Riga, Latvia, October 21-23, 2020* (CEUR Workshop Proceedings, vol. 2612). Available at http://ceur-ws.org/Vol-2612/paper6.pdf

- *Neologism*. The Oxford Companion to the English Language. Eds. McArthur, Tom, Jacqueline Lam-McArthur, and Lise Fontaine. 2018 Oxford University Press, Oxford Reference. Retrieved on 24 Sep 2020 from https://www-oxfordreference-com.libproxy.helsinki.fi/view/10.1093/acref/9780199661282.001.0001/acref-9780199661282-e-831
- Onysko, Alexander 2007. *Anglicisms in German: Borrowing, lexical productivity, and written codeswitching*. Vol. 23. Walter de Gruyter.
- Porter, Martin F. 1980. An algorithm for suffix stripping. *Program 14.3*: 130-137.
- Pulcini, Virginia, Cristiano Furiassi and Félix Rodríguez González 2012: The lexical influence of English on European languages. In Furiassi, Cristiano, Virginia Pulcini, and Félix Rodríguez González (eds.) 2012. *The anglicization of European lexis.* John Benjamins Publishing.
- Räikkälä, Anneli 1995. Menneiltä vuosilta. *Kielikello* 1.
- Sadeniemi, Matti (ed.) 1951-1961. *Nykysuomen sanakirja*. WSOY.
- Sajavaara, Paula 1989. Vierassanat. In Vesikansa, Jouko (ed.) 1989. *Nykysuomen sanavarat*. WSOY. 64-109.
- Sijens, Hindrik, and Hans Van de Velde 2020. The Formation of Neologisms in a Lesser-used Language: The Case of Frisian. *Dictionaries: Journal of the Dictionary Society of North America 41*, no. 1: 45-67.
- Suomi, Kari, Juhani Toivanen, and Riikka Ylitalo 2008. Finnish sound structure. *Studia humaniora ouluensia* 9.
- Säily, Tanja, Eetu Mäkelä & Mika Hämäläinen 2018. Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition 25(1).* 30–49. Retrieved from https://www.semanticscholar.org/paper/20255167be74b4e8f6a86e08927e38cedc48b856
- Vatanen, Tommi, Jaakko J. Väyrynen & Sami Virpioja 2010. *Language Identification of Short Text Segments with N-gram Models.* LREC.
- VISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho 2004. *Iso suomen kielioppi.* Helsinki: Suomalaisen Kirjallisuuden Seura. Retrieved from: http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7

7.1. CORPORA

- *Finnish N-gram Corpus, version 1* (FNC1). 2014. [text corpus] Distributed by the University of Helsinki on behalf of the FIN-CLARIN Consortium. URL: http://www.helsinki.fi/finclarin/fnc1
- Huovilainen, T. (2019a). *Finnish OpenSubtitles 2017, source* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2019110801
- Huovilainen, T. (2019b). *Finnish Wikipedia 2017, source* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2019110803
- National Library of Finland (2014). *The Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2014073038
- University of Helsinki (2019). *Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s (VRT), Version 2* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-201908191

- Yleisradio. *Yle Finnish News Archive 2011-2018, source* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017070501
- Ylilauta (2011). *The Downloadable Version of the Ylilauta Corpus* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016101210