Filozofická fakulta Univerzity Palackého

Katedra anglistiky a amerikanistiky

The Core of the English Vocabulary and Its Origin

(Bakalářská práce)

2017

Natálie Wimmerová

The Core of the English Vocabulary and Its Origin
(Bakalářská práce)

Autor: **Natálie Wimmerová**
Studijní obor: Česká filologie - Anglická filologie
Vedoucí práce: **Mgr. Michaela Martinková, Ph.D.**
Počet stran (podle čísel): 56
Počet znaků: 82 229
Olomouc 2017

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a uvedla úplný seznam citované a použité literatury.

V Olomouci dne 2. 5. 2017 ……………………………….

Děkuji Mgr. Michaele Martinkové, Ph.D. za vstřícný přístup, užitečnou metodickou pomoc a cenné rady při zpracování práce.

Abstract

The present thesis is concerned with the origin and structure of the core of the English vocabulary. It is a common claim that the majority of the English lexicon is of Romance origin, the most frequently used words are, however, supposed to be Germanic. The aim of this paper is to analyse the 500 most frequent items presented in the *New General Service List* by Václav Březina and Dana Gablasová in terms of their origin, the amount of function and content words among them, and the lexical fields the content words belong to.

Key words

centre, center, core, periphery, English vocabulary, origin, Germanic, Romance, OED

Anotace

Tato bakalářská práce se zabývá původem a strukturou jádra anglické slovní zásoby. Běžně se uvádí, že většina anglických slov je románského původu, zároveň se ovšem věří, že ta nejfrekventovanější slova jsou germánská. Tato práce si klade za cíl analyzovat 500 nejfrekventovanějších položek uvedených v seznamu *New General Service List* sestaveném Václavem Březinou a Danou Gablasovou s ohledem na jejich původ, množství autosémantik a synsémantik a zastoupení lexikálních polí.

Klíčová slova

centrum, jádro, periferie, anglická slovní zásoba, původ, germánský, románský, OED

OBSAH

1. Introduction

Though it is often argued that "over 60% of English words are of Romance origin" (Emonds and Faarlund 2014, 29), it is, at the same time, believed that "[Anglo-Saxon lexemes] provide almost all the most frequently used words in the language" Crystal (2003, 124).

The aim of the thesis is to subject to scrutiny what is often called the Centre, or the Core[1] of English vocabulary. In Sections 2 and 3 I will present what linguistic literature has to say on the Centre/Core – Periphery distinction: Section 2 will very briefly outline the linguistic history of the Centre/Core – Periphery distinction, starting from the early German and Prague structuralist theories, and finishing with the Prototype theory within the cognitive linguistics framework (Skrebtsova 2014).

In Section 3, attention will be paid to the lexical plane. I will provide several definitions of the Centre and the Periphery of the lexicon as presented in the linguistic literature and explain the challenge the study of the phenomenon represents. Finally, I will introduce the results of various research projects performed in the same field.

Sections 3.1 to 3.3 will introduce the concept of Basic vocabulary and various approaches to it; attention will be paid to the way basic words are selected and marked in the top five learner dictionaries and basic vocabulary lists. Special attention will be given to Michael West's *General Service List* and the *New General Service List,* created by Václav Březina and Dana Gablasová.

In the research part, I will try to identify the origin of the 500 most frequent words, as presented in the *New General Service List*. The main source of the data will be the etymological information provided for these 500 words in the *Oxford English Dictionary*. Additionally, I will attempt to describe the structure of the Core of the English vocabulary in terms of a) proportion of function and content words, and b) the lexical fields represented there. Finally, I will present and discuss the data.

[1] Čermák points out that apart from the traditional distinction Centre – Periphery, American linguistics, in particular, also works with the dichotomy Core - Periphery.
SOURCE: https://www.czechency.org/slovnik/CENTRUM%20A%20PERIFERIE. Accessed April 24, 2017.

2. Approaches to Centre and Periphery in linguistic literature

According to Skrebtsova (2014, 148), the distinction between the Centre and the Periphery, which recognises certain asymmetries in language, can be traced back to the theories of German and Prague structuralists.

In German linguistics, the terms originated in the field theory, which says that words in language do not exist individually but rather in semantic groups or fields (Skrebtsova 2014, 144). As Skrebtsova (2014, 145) reports, Günther Ipsen, who coined the term *field*, describes the situation rather metaphorically, comparing the semantic field to a mosaic in which the contours of words, similarly to the contours of stones in the mosaic, merge.

In Trier's conception (quoted in Skrebtsova 2014, 145), each word has its specific position within the field. Its position depends on the word's semantic relations with other words in the field; some are closer to others and some are farther from them. Trier consequently claims that fields contain areas such as the Centre (where the words are closely related) and the Periphery (where the words are considerably farther from each other, i.e. share fewer attributes).

Czech linguistics, on the other hand, traditionally credits the phenomenon of the Centre and the Periphery to the Prague Linguistic Circle. According to Šimková (2013, 138), the assumed "father" of the idea is František Daneš, who claims it older, but does not give information about its real author. To be more concrete, Daneš (1966, 9) admits that the discussion of the relation between the Centre and the Periphery can be found in the works of the Prague school as well as elsewhere. Daneš (1966, 10) and Vachek (1966, 27), both members of the Prague school, for example refer to C. F. Hockett's *The problem of universals in language*.

Linguists agree (e.g. Daneš 1966, 9 and Němec 1976, 118) that the Centre and the Periphery can be distinguished at all linguistic levels. However, according to Daneš (1966, 13), the relation between them appears to be especially well explored in the phonic plane. Vachek (1964, 8) sees some crucial ideas on this matter in André Martinet's *Économie des changements phonétiques.*

Martinet discusses the fact that some phonemes are fully integrated in the system while others are not. A non-fully integrated phoneme "is not linked by oppositions of its distinctive features to a larger number of other phonemes co-existing with it in the same system of phonemes" (Vachek 1964, 8).

As Vachek (1964, 8) points out, Martinet's terms "fully" and "non-fully integrated" correspond to what the Prague school calls central and peripheral. In addition to this finding, he highlights another factor, i.e. a low "functional yield"[2], which is also characteristic for peripheral elements. Daneš develops Vachek's conception and suggests considering also "the criterion of the utilization in utterance contexts, i.e. in principle, the frequency of occurrence of the given language unit" (1966, 13).

Last but not least, Skrebtsova (2014, 144) further points out that nowadays the terms centre and periphery are commonly used in cognitive linguistics.

> [A cognitive psychologist Eleonor Rosch] challenged the conventional, classical view which holds that categories have clear and fixed boundaries and are defined by a set of necessary and sufficient attributes, all members having equal status within the category. …. Her experimental data showed that some category members were judged to be more representative ("good examples"), others less representative ("bad examples").
>
> Skrebtsova (2014, 148)

Although Rosch did not use the terms Centre/Core and Periphery herself, Skrebtsova (2014, 148) informs that other cognitive scientists who followed Rosch's theory (e.g. George Lakoff and John Taylor) mention the more central or more peripheral status of the examples within a given category.

[2] In the early Prague school days, the functional yield (also functional load) was used to refer to the importance of phonemic contrasts in making distinctions between words in a language. If a phoneme has a low functional yield, it means that it hardly ever distinguishes words from one another.

3. The Centre and the Periphery of the lexicon

When discussing the necessity of taking the systemic approach to the lexical plane, Jakobson (quoted in Němec 1996, 223) admits the task is significantly more difficult than identifying the phonological and morphological core. Veselovská's lecture notes briefly remark that the core of the lexicon contains "grammatical elements and basic words" (2016, 25), but do not specify what she means by the latter. Carter admits that "[i]t is very much an intuitive notion" (1982, 39).

Čermák (2010, 202) suggests the Centre is relatively small and its lexemes are supposed to be shared by all speakers of the language.[3] The periphery, on the other hand, contains less frequent items, e.g. dialects, terms, and elements which most speakers consider diachronic.

Cvrček suggests two major views on the identification of the basic (or core) vocabulary: First, these might be "elements which are common to all or to the majority of texts" (2011, 2). This "smallest set of elements capable of fulfilling basic communicative needs" (2011, 2) is usually captured in vocabulary lists. However, this conception covers only the core which is useful for pedagogical purposes. Such a core vocabulary, as Cvrček himself points out, would definitely not suffice the second purpose he suggests, which is constructing small or medium-sized dictionaries. This core would consist of "all grammatical words (i.e. tens or hundreds of units) and some basic lexical words (i.e. hundreds or thousands of words)" (Cvrček 2011, 3).[4]

Apart from Sgall (2011, 25), who sees the main distinction between the Centre and the Periphery as the distinction between marked and unmarked (in accordance with Trubetzkoy's approach, as Sgall himself points out), all definitions operate with the frequency of appearance as the indicator of "closeness" to the core. Even though the frequency of items as a possible distinction between the central and the peripheral seems to be merely the consequence of the qualities of the elements, it "plays a crucial role in determining what the core elements are" (Cvrček

[3] He himself admits that this is uneasy to prove.

[4] More information on basic vocabulary and different approaches to it will be provided in Section 3.1.

2011, 1). Čermák (2004, 7) even simplifies the issue and defines basic words as the most frequent ones.

In terms of structure, Sgall (2011, 25) sees the Centre as simply structured whereas the Periphery consists of many layers. It is diverse and rather unstructured. Němec (1976, 120) discusses again the compactness of the Centre and the diffuseness of the Periphery and points out that this is due to the fact that the elements of the Periphery do not have all the features which are common to the elements of the Centre. Thus, the peripheral items are less integrated in the system.[5] As Němec (1976, 120) continues to explain, the Periphery is thus less stable over time and peripheral words are more likely to infiltrate into the core than vice versa.

A great role in the possible diffusion is also played by domestication. According to Němec (1976, 121) domestication is a process during which the borrowed items, initially located in the periphery of the lexicon, are gradually adapting the features typical for the lexemes which are central to the language.

In Peprník's words, the Centre consists of "lexical units with greatest stability and frequency and with greatest independence from the changeable extralinguistic reality" (2006, 22). The Periphery, on the other hand, includes "words limited in frequency, restricted as to territory and period" (Peprník 2006, 22).

Algeo and Pyles claim that "the core vocabulary of English is, and has always been, native English" (2004, 271). However, the authors do not define what they mean by "native English," which makes their statement rather problematic. Based on the history of the English language, it can be assumed, though, that Algeo and Pyle's "native English" refers to Anglo-Saxon vocabulary.[6]

On the other hand, they promptly acknowledge that "an overwhelming majority of the words in any large dictionary and a large number of words we use every day were either borrowed from other languages or made up using the elements of borrowed words" (Algeo and Pyles 2004, 271).

[5] The criterion of integration in the system is typical for the approach of the Prague school and was also discussed by Vachek and Daneš (see Section 2.).

[6] Emonds and Faarlund (2014, 22) argue that Modern English is not "(…) descended from the language of the Anglo-Saxons", however, also admit that it is undeniable Germanic, among other reasons, in as much as its core vocabulary. This view will be discussed in more detail towards the end of section 5.1.2.

More concrete information is provided by Finkenstaedt and Wolff (1973), who estimates that “over 45 per cent of commoner words (25 per cent of the general lexis) in Present-Day English are of Germanic origin …. Latin and French each account for a little more than 28 per cent of the lexis recorded in the *Shorter Oxford English Dictionary*” (paraphrased in Burnley 1992, 415). In contrast, Čermák’s brief analysis of one thousand most frequent Czech words indicated only 11.5 per cent of loanwords among them (Čermák 2010, 203).

The aim of my thesis is to examine the English core vocabulary and state the approximate percentage of Germanic vocabulary compared to Romance words. The results will be compared with Finkenstaedt and Wolff’s statistics. Additionally, the words will be sorted out into function and content ones. The content words will then be evaluated with respect to the lexical fields they belong to.

One of the difficulties I expect to face is the fact that in many cases the word origin can be assigned only approximately, as Čermák (2010, 203) points out with respect to his research. The other problem is the already mentioned fact that there is no commonly accepted definition and delimitation of the core vocabulary. Daneš stresses out the vagueness of the terms Centre and Periphery. “[T]he concepts … are not defined in exact terms but rather in an intuitive and symbolic manner” (Daneš 1966, 14). There is no strict boundary but rather a transitional zone between them (Daneš 1966, 14). In his view, the “compact core” progressively changes into “the diffuse periphery” (Daneš 1966, 11).

Sgall (2011, 25) agrees with Daneš’s rejection of dichotomy and locates the transitional zone mainly on the edge of the Centre. Čermák (2010, 202), however, emphasises the drawbacks of this idea by stating that, the transitional zone does not provide any new information about the concept of the Centre and the Periphery, as nobody can denote its borders.

In order to avoid the difficulty, I will examine the words which have already been selected as basic by other researchers.

3.1. Basic vocabulary

When teaching a language, or writing a dictionary, one must decide which words are the most important for a learner to acquire. Unfortunately, “there is no universally accepted, ready-made list of “the core words of the English language” to be found” (Lee 2001, 250). This, as Lee argues, is at least partly caused by different approaches to the topic.

Major efforts to identify basic vocabulary can be dated back to the period closely preceding WW2. This is also the time when the concept of Basic English was invented (or discovered)[7] by Charles Kay Ogden.

I. A. Richards, Ogden’s associate, defines Basic English as “English made simple by limiting the number of its words to 850, and by cutting down the rules for using them to the smallest number necessary for the clear statement of ideas” (Richards 1943, 23).

The words were chosen to suffice in everyday communication. Particular stress was put on the ease of learning and reduction of meaning to the central one. Other criteria to consider were “simplicity”, “economy”, “regularity”, “scope”, “clarity”, “naturalness”, and “grace”.[8] Basic English was designed as a vocabulary set sufficient to define twenty thousand other English words, and as such it was used in *General Basic English Dictionary* (Richards 1943, 23-27).

To determine the set of vocabulary, Ogden used what they called “the Panoptic Method”. In this process, a word was put in the middle of a circle with radial lines representing different relation of the tested word with other words. All these related words are clearly not necessary in the basic vocabulary as their meanings can be explained using the appropriate word in the middle together with the appropriate relation word.[9]

[7] Richards (1943, 26) justifies his usage of the word “discovery” rather than “invention” by claiming that Basic English was something inherent to the English language due to its development. He argues (45) that the invasion of Danish people allowed a significant degree of analysis in the language, as a result of the similarity of the two languages and the will to understand each other. To advocate his belief, Richards (46) translates Jespersen’s quote from *Growth and Structure of the English Language* into Basic English: “In fact, the most necessary parts of the language are the very ones on which the effect of Scandinavian languages has been greatest.” This also supports Algeo and Pyle’s claim that the core vocabulary of English should be mainly Germanic.

[8] The criteria are very subjective and, in some cases, quite unclear.

[9] SOURCE: http://ogden.basic-english.org/panoptic.html. Accessed February 11, 2017.

This way he established the list of names of “things”, names of “qualities”, and “operations”. In terms of traditional parts of speech, the first two groups correspond to nouns and adjectives respectively; operations are supposed to express relations between them. They include a very limited repertoire of verbs, prepositions, pronouns, adverbs, articles, and others (Richards 1943, 28-37).

Lee (2011, 251) suggests that the readers of his paper might have “differing conceptions” of basic vocabulary. Lee (2011, 252-255) proposes seven working definitions: the most frequent words in the language as a whole; the most frequent words in terms of a particular medium; the most frequent words for a particular demographic grouping; words that are cognitively basic or most salient; words that, in their most general sense, have the most widespread usage across a wide range of genres; words that are most general, or unmarked, or central to the language and words useful for dictionary definitions. In his view, Ogden’s Basic English e.g. fulfils the last two criteria.

Carter (1982) also intends to provide some basic definitions of core vocabulary. Generally, Carter’s view of core vocabulary is similar to Lee’s. In addition, Carter puts great emphasis on the neutrality and unmarkedness as the criterion for identifying the core items. As Carter explains, the contrast between the neutral core items and the marked peripheral ones is crucial for distinguishing expressive meaning.

> Degrees of expressivity would be impossible to perceive unless there were some neutral norm or unmarked set of features against which deviation can be measured by both addresser and addressee.
>
> (Carter 1982, 39)

However, as Carter concludes, it is never a single criterion which would establish an item as central. Additionally, different criteria are preferred for different purposes (1982, 46). One of the purposes is the pedagogical one, which is to be discussed in the following section focusing on basic vocabulary in learner’s dictionaries and service lists. Bogaards’s overview of basic vocabulary lists (2008) in five learner’s dictionaries is taken as a starting point. The dictionaries discussed are: *Oxford Advanced Learner’s Dictionary (OALD7), Macmillan English Dictionary for Advanced Learners (MEDAL2), Collins COBUILD Advanced*

Learner's Dictionary (Cobuild5), Longman Dictionary of Contemporary English (LDOCE4), and Cambridge Advanced Learner's Dictionary (CALD2).

3.2. Basic vocabulary in learner’s dictionaries

Bogaards (2008, 1231) claims that “[t]he vocabulary that is described in these [five learner’s] dictionaries is selected on the basis of frequency of appearance in English.” However, he additionally reports the criterion of frequency was not the only one used (Bogaards 2008, 1232) and points out the inconsistency of the frequency data in the “big five” learner’s dictionaries.

The different strategies and different types of marking the basic vocabulary in these dictionaries are presented in the following sections.

3.2.1. Oxford Advanced Learner’s Dictionary

The frequency information used for *Oxford Advanced Learner’s Dictionary* (*OALD7*)[10] is based on the data from the *British National Corpus* and the *Oxford Corpus Collection*. The most important lexemes are marked with a key and included in the *Oxford 3,000 list of important words* (Bogaards 2008, 1231-2). Further discussion on this vocabulary list is provided below.

3.2.1.1. Oxford 3000

The 3,000 key words included in the *Oxford 3000* were, reportedly, selected by a group of experts on linguistics and language teaching as the most useful. The usefulness of the vocabulary is evaluated according to its frequency in the corpora and text coverage.

On top of that, several words were added as “a panel of over seventy experts in the fields of teaching and language study” claimed them “very familiar to most users of English. These include, for example, words for parts of the body, words used in travel, and words which are useful for explaining [the meaning of other words]”[11].

[10] The eighth edition of *OALD* uses the same system of marking the key words as *OALD7*.

[11] SOURCE: http://www.oxfordlearnersdictionaries.com/about/oxford3000. Accessed February 23, 2017.

As the website states, all word definitions in the *OALD* are written using the repertoire of the *Oxford 3000* together with their list of "language study terms".

Unlike the *New General Service List* (which will be presented in section 3.3.2.) the *Oxford 3000* neither provides the rank information with their headwords, nor separates them into categories with respect to their frequency.

3.2.2. Macmillan English Dictionary for Advanced Learners

The authors of the *Macmillan English Dictionary for Advanced Learners* (*MEDAL2*) characterise the core vocabulary as "common words". The base for their research was the *World English Corpus*, which (with its 200,000 items) is the smallest of all the corpora used for this purpose.

MEDAL2 uses the system of one to three stars for marking the core vocabulary. One star marks the "fairly common" words while three stars indicate the 2,500 "most basic words of English". *MEDAL2*, unlike *OALD7*, distinguishes the frequency of e.g. *bank* as a verb from *bank* as a noun and assigns them a different number of stars. It, however, does not distinguish between the homonymic nouns *bank* as a financial institution and *bank* of a river (Bogaards 2008, 1231-2).

According to Bogaards (2008, 1234-5), it is a well-known fact that the most frequent words are often polysemous and some of the senses are more frequent than others. *MEDAL2* deals with the situation by accompanying certain senses of words with a list of their frequent collocations.

3.2.3. Collin's COBUILD Advanced Learner's Dictionary

Collin's COBUILD Advanced Learner's Dictionary (*Cobuild5*) emphasises the impact of frequency on their choice of basic vocabulary. The information was taken from the *Bank of English*.

Similarly to the star system used by *MEDAL2*, the authors of *Cobuild5* mark the most frequent words with one to three diamonds. There is, however, a very limited set of words marked with three diamonds; only 650 (compared to the 2,500 three-star items of *MEDAL2*).

Reportedly, *Cobuild5* also pays attention to different meanings of a word and evaluates them separately (Bogaards 2008, 1231-2).

3.2.4. Longman Dictionary of Contemporary English

The basic vocabulary selection for the *Longman Dictionary of Contemporary English* (*LDOCE4*) is based on the *Longman Corpus Network*. Similarly to the *OALD* team, the authors of the Longman dictionary created their list of the 3000 most important English words called *Longman Communication 3000*.

Unlike the other four dictionaries, *LDOCE4* differentiates the frequency of use in written language from the frequency of use in spoken language. The most frequent words are marked with the letter W for written and S for spoken and the number 1-3 printed in red letters, which signals that they belong among the one thousand, two thousand or three thousand most frequent words, respectively (Bogaards 2008, 1231-2; *Longman Communication 3000,* 2009, 2044).

Additionally, the *LDOCE4* in some cases includes graphs showing imparity between the frequency of appearance in British and American texts (Bogaards 2008, 1235). *LDOCE*'s latest edition is the sixth one (2014). However, I only managed to examine the 5th edition (2009), which, in addition to the imparity between British and American texts, also includes visuals showing the difference between the frequency in written and spoken language.

3.2.5. Cambridge Advanced Learner's Dictionary

The *Cambridge Advanced Learner's Dictionary* (*CALD2*) marks the "most important words" by the letter E for "essential". Additionally, the dictionary also marks a number of words with I for "improvers", and A for "advanced" (Bogaards 2008, 1232).

According to the authors of *CALD*, the "improver" vocabulary is "also common in native speaker English … [and] include[s] less common words which express useful concepts" (Walter 2008, VIII). "Advanced" vocabulary is supposed to be "still highly significant". It aims at advanced students in order to "make their English more fluent and natural" (Walter 2008, VIII). The marks can be attached to the whole entry or to one of its senses.

As for the numbers, *CALD2*'s common vocabulary is quite extensive; it consists of 4,900 essentials, 3,300 improvers and 3,700 advanced words (Bogaards 2008, 1321-2).

The third edition[12] of *CALD* (published in 2008) uses the same system as Bogaards described for the second edition. Additionally, the authors of *CALD* claim "[t]he frequency information in this dictionary is special because it shows the relative importance not only for words, but also of their meanings, and of individual phrases" (Walter 2008, VIII). The researchers reportedly worked with the data from the *Cambridge International Corpus*. "They extracted all the high-frequency words and then coded examples of them to work out the frequency of their different meanings" (Walter 2008, VIII). Moreover, several words (e.g. some basic grammar words) were included despite their lower frequency due to their high importance to language students (Walter 2008, VIII).

[12] The latest edition of *CALD* is the fourth one, which was published in 2013. I was working with the third edition (2008).

3.3. Service lists

3.3.1. General Service List

The topic of word lists in language pedagogy is discussed in Nation and Waring (1997). According to the authors, the earliest attempt to extract the basic vocabulary for second language learners is *The Teachers Word Book of 30,000 words* developed by Thorndike and Lodge in 1944. From today's perspective, it is mainly remarkable for its wide range of vocabulary and the amount of manual work done on counting the frequency of words.

Bauman (2002) states that the frequency data were then used by Michael West for his *General Service List* (*GSL*) from 1953. Březina (2014), on the other hand, claims that the *GSL* is, in fact, a revision of the *Interim Report on Vocabulary Selections* from 1936. Regardless its origin, Michael West's *GSL* was successfully used for generations. Nation and Waring (1997), for instance, call it the "classic" core vocabulary list.

The *GSL* consists of 2,000 headwords, each followed by a list of its related forms. As Březina (2014) observes, the units are organized (with respect to their morphology) into word families. Nation and Waring (1997) add that the headwords are given in alphabetical order and the frequency information is provided for the headword, usually along with all its meanings and parts of speech.

Despite its popularity, the *GSL* has been repeatedly criticised for several reasons. Firstly, Bauman (2002) notices that "[t]he inclusion of related form under a headword is not consistent." Secondly, it is only based on one corpus from 1936 (Browne 2014, 1 and Březina 2014). It can be argued that even in West's time it was obsolete.

Furthermore, Březina (2014) highlights the insufficient size of the corpus (2.5 – 5 million words) and the subjectivity of some of the criteria West applied. Aside from the necessity and neutrality of the vocabulary, the criterion of ease of learning caused words which were low in frequency, like *timely*, to be included in the list. West's argument in support of the addition was its morphological relation to the much more common word *time* (both are members of the same word family). In his view, it is easy to remember the word *timely* once the student knows the word *time* (Březina 2014).

Finally, in Engels's view (quoted in Nation and Waring 1997) the words outside the top 1000 provide a rather poor text coverage.

Some innovations were made to the list by John Bauman and Brent Culligan in 1995. Their objective was to provide a steadier system of sorting words into word families and update the frequency data. The first problem was dealt with using Bauer and Nation's *Word Families* and the new frequency data were taken from Francis and Kučera's *Brown Corpus* compiled in the 1960s (Bauman 2002).

Březina and Gablasová (2015) inform that in fact "[t]here are different versions of West's *GSL*." They, for instance, mention Nation's version, which "[was extended] for the use in the *RANGE* program[13]."

However, due to the criticism of the West's original version, a completely new version of the list was released in 2013. For the *New General Service List*, Charles Browne and his team (including Brent Culligan) intended to determine the narrowest possible set of the most important words with the highest text coverage. Emphasis was put on working with a more recent and larger corpus (273 million words) and a clearer definition of *word* (Browne 2014, 1-2). The latest update of Browne's list was provided in 2016. According to the authors, they mainly "update[d] … the frequencies for the spoken subsection of the *NGSL*"[14].

As Browne (2014, 2) claims, approximately 6 months after the publication of their first version, Václav Březina and Dana Gablasová released another updated version of the original *GSL*, also called *New General Service List*. In the next section I will focus on their version.

3.3.2. *New General Service List*

While Michael West's *GSL* used word family as the organising principle, the *New General Service List* (*new-GSL*) is based on the lemma principle. Březina and Gablasová (2015) criticise the word-family division for the questionable assumption that "the meaning of a derived word is largely transparent and can be understood on the basis of the knowledge of the individual morphological components." To illustrate the problematic cases, they use the pairs of words such

[13] RANGE is a software application developed for "analysing the vocabulary load of texts". SOURCE: http://www.victoria.ac.nz/lals/resources/range. Accessed May 1, 2017.

[14] SOURCE: http://www.newgeneralservicelist.org/. Accessed February 16, 2017.

as *train* and *trainers* or *part* and *particle* (cf. also *time* and *timely,* mentioned above). They explain that especially beginner learners may not fully understand the word formation processes that take place there. Therefore, the lemma principle was preferred.

Březina and Gablasová worked with four corpora (the *Lancaster-Oslo-Bergen Corpus*, the *British* National Corpus, the *BE06 Corpus of British English*, and *EnTenTen12*); ranging from 1 million-token corpora (the *Lancaster-Oslo-Bergen Corpus* and *BE06*) to *EnTenTen12*'s 12 billion. The smaller ones were included for the fact that they "reflect a variety of written English genres, including newspapers, fiction, essays, and scientific writing" (Březina and Gablasová 2015). Additionally, they reflect the situation from 1960s and late 2000s. The *British National Corpus* provides data from the later part of the 20th century, while *EnTenTen12* can be dated to 2012. Together, the four corpora cover the period from the 1960s until present.

As for text types, Březina and Gablasová used a variety of written texts from different genres (especially in the case of *EnTenTen12*, which consists of web documents) as well as the 10 million words of spoken English in the *British National Corpus*.

The *new-GSL* was created in order to cover the common lexical core but also the "recent development" in the English lexicon. In the first step, the authors created word lists from all four corpora, using the criteria of frequency and dispersion. To establish the recent vocabulary, the two latest corpora's wordlists were compared and the shared items were selected and subsequently also compared with the word lists compiled in the first step. Březina and Gablasová describe the final step as follows.

> The compilation of the *new-GSL* involved combining the wordlists created in the previous two steps: the common lexical core items were put together with current words shared by the two corpora reflecting the present-day use of the English language (*BE06* and *EnTenTen12*). The items in the *new-GSL* were then marked (…) and the list was alphabetized.
>
> (Březina and Gablasová, 2015)

The final list consists of the "base part" (including 2,116 items), and 378 "current words". The base part is divided, according to the frequency, into three types. The first type, which is marked by bold red capital letters, includes the 500 most frequent lemmas, the "bold type" (in bold writing but lacking colour coding) consists of the following 500 lemmas (rank 501-1000). The rest of the base vocabulary is called the "plain type" and is typed in plain writing. The "current words" are written in italics (Březina 2014, Březina and Gablasová 2015).

The *new-GSL* recognises twelve distinct "word classes". These are: noun; verb; modal; adjective; adverb; adverbial particle in phrasal verbs; preposition or conjunction; pronoun; determiner, quantifier or particle; abbreviation; existential there; and to as infinitive marker.

4. Methodology

As it was suggested in Section 3, my research will be compared to the study performed by Finkenstaedt and Wolff (paraphrased in Burnley 1992, 415). As mentioned before, they analysed the vocabulary listed in the *Shorter Oxford English Dictionary*, which is an abridged version of the *Oxford English Dictionary* (*OED*). This fact present a major problem.

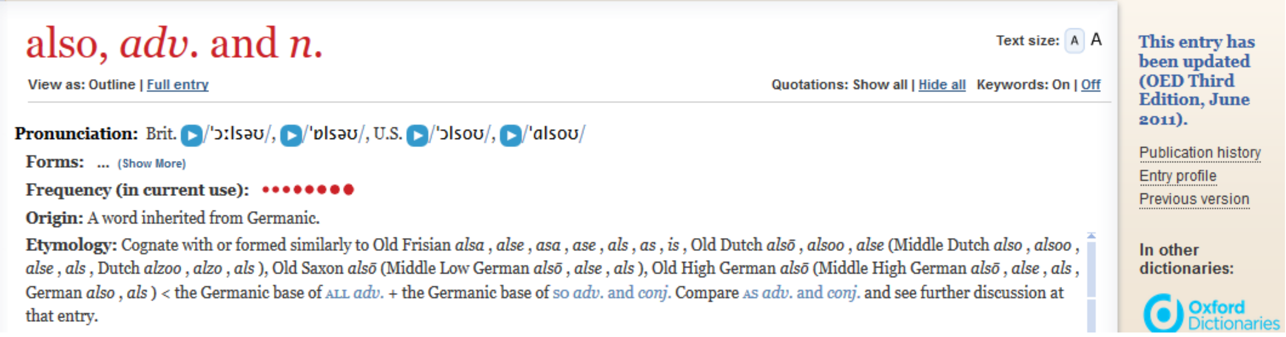
When Finkenstaedt and Wolff discuss the origin of "commoner words", they, in fact, seem to refer to the whole content of the 3rd edition of the *Shorter Oxford English Dictionary*, which contains thousands of words. My research, on the contrary, concentrated on what could be called the "opposite extreme", because it was limited to 500 most frequent words ("the bold red type") listed in the *New General Service List* (*new-GSL*) by Březina and Gablasová.

Referring to the conceptions of the core vocabulary proposed by Cvrček (2011) and quoted at the beginning of Section 3, the objective of my research was to reflect the contemporary core of the language which Cvrček (2011, 2) defines as "the smallest set of elements capable of fulfilling basic communicative needs", which "has been typically delimited from the perspective of a user … for pedagogical reasons" (Cvrček 2011, 2). Cvrček (2001, 3) claims that "[this] core vocabulary … does not exceed hundreds of types". Finkenstaedt and Wolff, on the other hand, worked with the other conception of core vocabulary Cvrček discusses, i.e. the core constructed for lexicographic purposes, which must contain "at least several thousands of elements" (Cvrček 2011, 3).

I created an Excel sheet with the 500 words (lemmas) provided by the *new-GSL,* each followed by the information about their rank and "word class", which, as reported in the previous section, only roughly corresponds to the word's part of speech. Instead of the original alphabetical order, the words in the Excel sheet were ordered according to their rank. Each word was then searched for in the online version of the *Oxford English Dictionary* (*OED*) to obtain the information about its first recorded appearance and origin. Unfortunately, this turned out to be far from a straightforward task.

The first volume of *OED* was published in 1884, i.e., more than 130 years ago. Naturally, there have been some updates since then. The authors of the online dictionary claim it is updated four times a year[15]. However, many entries I worked with have, reportedly, not been fully updated yet, so they might possibly contain the same or similar information they did when published for the first time; i.e. in some cases as late as the 1880s (see Figure 3). The assumption can be supported by the official website's claim that "[t]oday, the dictionary is in the process of its first major revision."[16]

The lack of a complete revision results in the fact that concrete formulations of the same information included in entries updated in different time periods vary. Compare Figure 1, showing the information provided for the entry for *also* (updated in June 2011), with Figure 2, which shows the entry for *some* (not fully updated since its first publication in 1913), and with Figure 3, showing the entry for *by* (not fully updated since its first publication in 1888).



also, *adv.* and *n.*

View as: Outline | Full entry

Text size: A A

Quotations: Show all | Hide all Keywords: On | Off

Pronunciation: Brit. /ˈɔːlsəʊ/, /ˈɒlsəʊ/, U.S. /ˈɔlsoʊ/, /ˈɑlsoʊ/

Forms: ... (Show More)

Frequency (in current use): ••••••••

Origin: A word inherited from Germanic.

Etymology: Cognate with or formed similarly to Old Frisian *alsa* , *alse* , *asa* , *ase* , *als* , *as* , *is* , Old Dutch *alsō* , *alsoo* , *alse* (Middle Dutch *also* , *alsoo* , *alse* , *als* , Dutch *alzoo* , *alzo* , *als*), Old Saxon *alsō* (Middle Low German *alsō* , *alse* , *als*), Old High German *alsō* (Middle High German *alsō* , *alse* , *als* , German *also* , *als*) < the Germanic base of ALL *adv.* + the Germanic base of so *adv.* and *conj.* Compare AS *adv.* and *conj.* and see further discussion at that entry.

This entry has been updated (OED Third Edition, June 2011).

Publication history

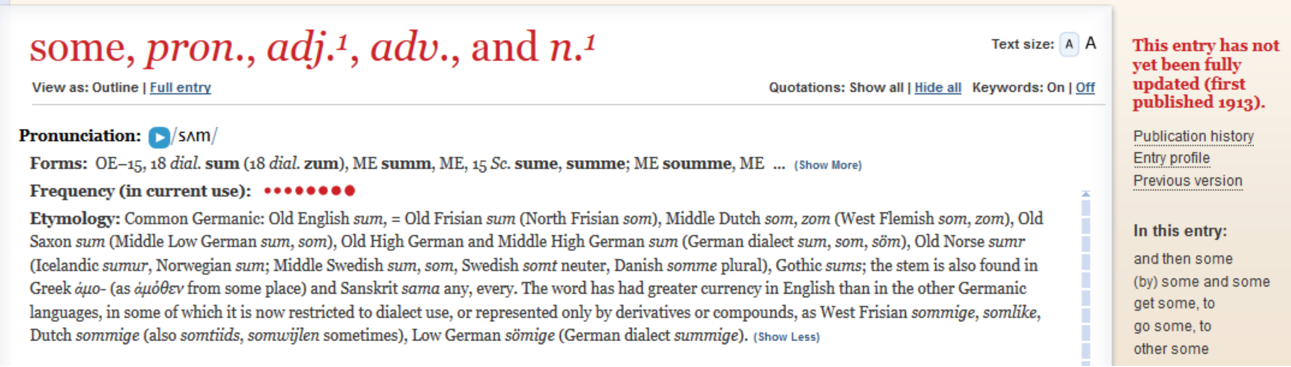
Entry profile

Previous version

In other dictionaries:

Oxford Dictionaries

Figure 1: The *OED Online* entry for *also*.



some, *pron.*, *adj.*¹, *adv.*, and *n.*¹

View as: Outline | Full entry

Text size: A A

Quotations: Show all | Hide all Keywords: On | Off

Pronunciation: /sʌm/

Forms: OE–15, 18 *dial.* **sum** (18 *dial.* **zum**), ME **summ**, ME, 15 *Sc.* **sume, summe**; ME **soumme**, ME ... (Show More)

Frequency (in current use): ••••••••

Etymology: Common Germanic: Old English *sum*, = Old Frisian *sum* (North Frisian *som*), Middle Dutch *som*, *zom* (West Flemish *som*, *zom*), Old Saxon *sum* (Middle Low German *sum*, *som*), Old High German and Middle High German *sum* (German dialect *sum*, *som*, *söm*), Old Norse *sumr* (Icelandic *sumur*, Norwegian *sum*; Middle Swedish *sum*, *som*, Swedish *somt* neuter, Danish *somme* plural), Gothic *sums*; the stem is also found in Greek *ἀμο-* (as *ἀμόθεν* from some place) and Sanskrit *sama* any, every. The word has had greater currency in English than in the other Germanic languages, in some of which it is now restricted to dialect use, or represented only by derivatives or compounds, as West Frisian *sommige*, *somlike*, Dutch *sommige* (also *somtiids*, *somwijlen* sometimes), Low German *sömige* (German dialect *summige*). (Show Less)

This entry has not yet been fully updated (first published 1913).

Publication history

Entry profile

Previous version

In this entry:

and then some

(by) some and some

get some, to

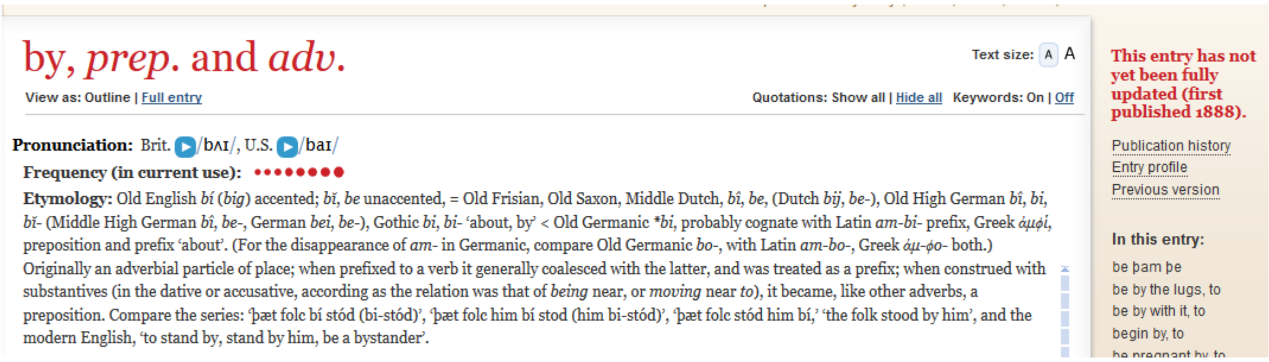
go some, to

other some

Figure 2: The *OED Online* entry for *some*.

[15] SOURCE: http://public.oed.com/about/#ViewingSecondEdition. Accessed April 24, 2017.

[16] *ibid.*



by, *prep*. and *adv*.

Text size: A A

View as: Outline | Full entry

Quotations: Show all | Hide all Keywords: On | Off

Pronunciation: Brit. /bʌɪ/, U.S. /baɪ/

Frequency (in current use): ●●●●●●●●

Etymology: Old English *bí* (*big*) accented; *bĭ*, *be* unaccented, = Old Frisian, Old Saxon, Middle Dutch, *bî*, *be*, (Dutch *bij*, *be*-), Old High German *bî*, *bi*, *bĭ*- (Middle High German *bî*, *be*-, German *bei*, *be*-), Gothic *bi*, *bi*- 'about, by' < Old Germanic **bi*, probably cognate with Latin *am*-*bi*- prefix, Greek *ἀμφί*, preposition and prefix 'about'. (For the disappearance of *am*- in Germanic, compare Old Germanic *bo*-, with Latin *am*-*bo*-, Greek *ἀμ*-*φο*- both.) Originally an adverbial particle of place; when prefixed to a verb it generally coalesced with the latter, and was treated as a prefix; when construed with substantives (in the dative or accusative, according as the relation was that of *being* near, or *moving* near *to*), it became, like other adverbs, a preposition. Compare the series: 'þæt folc bí stód (bi-stód)', 'þæt folc him bí stod (him bi-stód)', 'þæt folc stód him bí,' 'the folk stood by him', and the modern English, 'to stand by, stand by him, be a bystander'.

This entry has not yet been fully updated (first published 1888).

Publication history
Entry profile
Previous version

In this entry:

be þam þe
be by the lugs, to
be by with it, to
begin by, to
be pregnant by, to

Figure 3: The *OED Online* entry for *by*.

Notice that Figure 1, unlike the latter two, provides the label for the origin of the word; namely, it claims the word was "inherited from Germanic". Figure 2, does not contain the label of origin at all, however, the authors provide information about the word's origin in the etymology section. To be concrete, the word is classified as "common Germanic". Finally, Figure 3 neither includes the label of origin, nor does it provide a systematic classification under the etymology label.

I proceeded as follows. If an entry did not include the label of origin, I looked up the information about the origin of the word in the etymology section. All the three examples presented above were classified as Germanic.

As for the classification according to origin itself, I distinguished three main types: Germanic, Romance and unknown or uncertain. Further division was attempted at in the Germanic group as *OED* claims several English words to be Scandinavian borrowings. I additionally compared the information with the list of Scandinavian borrowings provided by Baugh and Cable (2002), Emonds and Faarlund (2014) and Algeo and Pyles (2004). The Romance group was roughly divided according to the direct source language of the words into the Latin subsection, French subsection, and the multiple origins subsections. More information concerning the possible division inside the Germanic and Romance groups is given in the respective chapters of the data analysis.

The next step following the data collection was to divide the analysed words into function and content words. All function words of the language (or at least the majority of them) are expected to be found in the Centre. However, the division is problematic in many cases, especially in the situation when no context is provided. Before giving a deeper analysis, Corver and Riemsdijk (2001) offer a simplified definition of function and content words. "Content words are … those lexical items which have a relatively 'specific or detailed' semantic content and as such carry the principal meaning of the sentence. …. [F]unction words have a more 'non-conceptual' meaning and fulfil an essentially 'grammatical' function" (Corver and Riemsdijk 2001, 1).

As reported in Corver and Riemsdijk (2001, 6), in the study *A Unified Theory of Syntactic Categories* (1985) Emonds points out the existence of "grammatical nouns, verbs, adjectives and prepositions". To name a few examples, this category includes the nouns *thing, place,* and *time*; the verbs *be, have, bring, take, do, make,* and *say*; several adjectives like *such* or *same* and the "adpositions" *out* and *up*.

I decided to apply a rather intuitive division for the disputable cases, so whereas the noun *time* was classified as a content word as well as all verbs, except for those which may serve as auxiliaries and modals, all preposition and conjunctions were classified as function words.

Furthermore, I was also interested in determining the lexical fields the content words belong to. My findings were then compared with the information presented by Algeo and Pyles (2004), Baugh and Cable (2002), and Emonds and Faarlund (2014).

Next, I created pie charts to illustrate the ratio between the words of Germanic origin, Romance origin and unknown or uncertain origin, and the ratio between content and function words. After that, the information about the origin and the ratio between content and function words were combined to compare the amount of content words and the amount of function words in the Germanic group with the Romance group and the "unknown" group.

Finally, I tried to find out if the ratio between the different origins as well as the one between content and function words is changing if the scope of the core vocabulary is being gradually enlarged. For this purpose, I counted the number of words of a given quality (i.e. Germanic origin, Romance origin, uncertain origin; content word, function word) in the first, second, third, fourth and fifth hundred (divided according to the rank) separately and compared the results by creating 100% stacked bar charts.

5. Data analysis

5.1. Germanic vocabulary among the 500 most frequent words

English is a Germanic language. This means that it, along with Danish, Dutch, German, Norwegian and others, go back to the mutual ancestor, i.e. Germanic or Proto-Germanic (PG). As Baugh and Cable report, the Germanic tribes of the Angles, the Saxons and the Jutes began invading Britain around 449. The Angles and the Jutes were originally settled in the Danish peninsula, while the Saxons probably came from the area between the Elbe and the Elms (Baugh and Cable 2002, 41-42). The first and main source of Germanic vocabulary is supposed to be the dialects these tribes brought and which evolved in the language called Old English.

As stated already, the labels *OED* provides for these cases are "inherited from Germanic" or "common Germanic". If the dictionary entry has not been updated recently, the etymology information provided would start with the form which the word had in Old English.

The pie chart in Figure 4 shows that **65 per cent (324 words) of the analysed core vocabulary (the first 500 words in the *new-GSL*) are indisputably or with high probability of Germanic origin**. The "common Germanic", or "Old English" vocabulary forms a clear majority of the words I evaluated as Germanic.

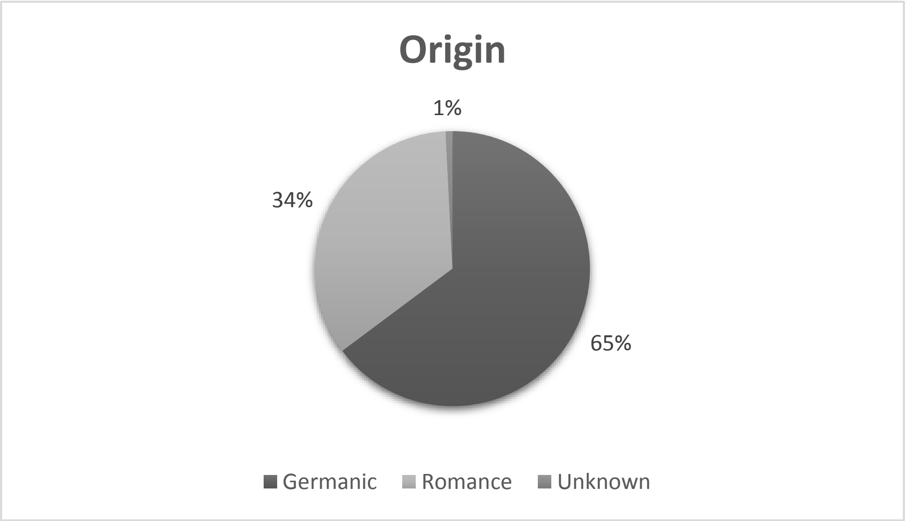


Figure 4: Pie chart illustrating the proportion of the vocabulary of the three main types of origin.

The second source of Germanic vocabulary detected in the sample was Old Norse – the language of the Vikings. The Norwegians started attacking Britain in the middle of the eighth century and partially colonised it (Baugh and Cable 2002, 83). Emonds and Faarlund inform that:

> [t]he Scandinavian-speaking descendants of the Vikings were increasingly predominant until 878, when the successes of the Saxon King Alfred led to a roughly equal division of the country (English control in the south and west and Danish control in the north and east). …. The situation for nearly 200 years was then that England consisted of two countries with a highly unstable border, the Danelaw and Wessex."
>
> Emonds and Faarlund (2014, 35)

Baugh and Cable (2002, 89) note that "[t]he number of Scandinavian words that appear in Old English is … small." Emonds and Faarlund (2014, 38) explain this fact by stating that "a native language borrows terms for novel concepts introduced by the newcomers, but not for those already expressed in its own vocabulary" and add that the 6th edition of Baugh a Cable's *A History of the English Language* states that there are only three words Old English borrowed form Scandinavian which survive in Modern English. Those are *law*, a *hold* of land, and *boatswain*.

As could be expected, the first 500 words in the *new-GSL* contain only the noun *law* with the rank of 461. According to *OED*, the word had been in use before the year 1000 and indeed was adopted from "prehistoric Old Norse" during the late Old English period.

However, *OED* also claims that the verbs *take* (with the rank of 46), *call* (with the rank of 143), and *run* (with the rank of 200), which are marked as possible early Scandinavian borrowings, appeared already in Old English. On top of that, I recorded three more words which *OED* pronounced either as "common Germanic" or as "formed within English" but other sources claimed undoubtfully Scandinavian. Specifically, these were the words *give* (rank 74), *both* (rank 202) and *though* (rank 255).

According to *OED*, the word *give* was first used already before 855. Although the entry does not include the label of origin, the etymology section claims the word is of common Germanic origin and lists all the cognates found in Germanic languages, including Old Norse. Emonds and Faarlund (2014, 51) state *give* is Scandinavian.

The same statement is made about the word *both* (Emonds and Faarlund 2014, 51). The origin and etymology sections of the entry in *OED*, however, claim the word was "formed within English" by combining the adjective *bo* and the pronoun and adjective *tho*. *Bo* appears in Old English as *begen*, *ba*, *bu* and corresponds to the Gothic stem *ba-* and the Germanic stem *bo-*. *Tho* appears in Old English as *þá*, which corresponds (among others) to Old Norse *þeir*.

Finally, the word *though* is considered Scandinavian not only by Emonds and Faarlund (2014, 52) but also, as again reported in Emonds and Faarlund (2014, 51), by Barbara Strang and her *A History of English* (1970). The authors of *OED* found the first record of this word around 888. As in the case of *give*, the entry does not include the label of origin. The etymology section lists the Old English forms, which are *ðéah*, *þéah*, and *þéh*, and the corresponding forms in other Germanic languages. The Old Norse one is *þó*.

If I count the words *give*, *both* and *though* as well as the words *take*, *call*, and *law*, it can be stated that 6 out of 324 words I evaluated as Germanic can be Scandinavian borrowings adopted during the Old English period.

Surprisingly, most Scandinavian borrowings entered English during the Middle English period (ca. 1150-1500) although the Viking invasions were already over and the country was under French control. *OED* proposes that fourteen words from my vocabulary sample might be Scandinavian borrowings adopted between the years 1150 and 1500. These are *they* (rank 28), *their* (rank 31), *get* (rank 55), *want* (rank 106), *same* (rank 120), *seem* (rank 130), *until* (rank 227), *happen* (rank 261), *low* (rank 302), *big* (rank 310), *upon* (rank 420), *raise* (rank 451), *die* (rank 464), and *likely* (rank 470). A few of the entries, however, suggest that considerable doubt exists as to the precise origin.

Firstly, the origin section of the entry *happen* claims it was actually formed within English by derivation. The stem *hap* is probably Scandinavian, while the suffix *-en* is Germanic.

Secondly, the word *big* is claimed to be of unknown origin. However, the author of the entry informs it was only recorded in North Midland and Northern sources (the area where the Scandinavians were settled) and suggests comparing the word with a Norwegian regional expression *bugge* (meaning "a mighty man") and a rare adjective *bugga* (meaning - "rich", "wealthy", or "powerful").

Thirdly, while no clear statement is made about the word *upon*, the entry again suggests an influence of Old Norse: "[t]he compound may have partly arisen from uses of *upp on* or *uppe on* in Old English …, but the date at which it appears, and the locality of the texts in which it is first prominent, suggest that it was mainly due to the influence of Old Norse *upp á*"[17].

Finally, according to *OED*, *likely* might be either an early Scandinavian borrowing, or the combination of Old English *ylike* (which was inherited from Germanic) and the Germanic suffix -*ly*. Nevertheless, Emonds and Faarlund (2014, 51) cast no doubt upon the word's Scandinavian origin. This could be confirmed by the "retention of the hard pronunciation of *k*" (Baugh and Cable 2002, 87).

Additionally, having said that *though* was borrowed from the Scandinavian vocabulary, I suggest putting the word *although* on the list as well since *OED* claims it was formed within English by combining the etymons *all* (reportedly inherited from Germanic) and *though*. According to the dictionary, the word firstly appeared in written records around 1325.

Altogether, the core vocabulary I analysed contains **twenty possible early Scandinavian borrowings**, which means they only account for a little over **six per cent of the Germanic vocabulary.**

Emonds and Faarlund (2014) introduce a different view on the matter. They believe that the "mysteriously late 'borrowings'" (as they call them) recorded in Middle English were not borrowed at all. It is argued that Middle English did not developed from Old English (as it is commonly believed) but instead it descended from the language spoken by the Scandinavian people.

These Scandinavian, as Emonds and Faarlund go on to argue, then had to borrow a great deal of vocabulary from the areas that the English were relatively

[17] Oxford English Dictionary Online, s.v. "upon", accessed March 30, 2017, http://www.oed.com/view/Entry/220029.

familiar with; e.g. Christian practices and beliefs, monastic life, road-building and building construction in general, crop and food production, inheritance, property, schools, metal-working, embroidery, etc. (Emonds and Faarlund 2014, 40-41).

While the traditional view supposes Middle English lexicon was more English than Scandinavian, Emonds and Faarlund highlight that "[t]here is no burden of proof on claiming that Middle English words derive from Norse cognates rather than from Old English" (Emonds and Faarlund 2014, 54). According to their statistics (Emonds and Faarlund 2014, 54), Old English and Norse cognates account for 50 per cent of the Middle English vocabulary (if Romance languages are not included). In fact, my list contains only two words of Germanic origin for which *OED* explicitly states they do not have cognates in Scandinavian languages, namely the verb *speak* and the adverb *soon*.

In conclusion, the share of Scandinavian vocabulary in the English language is still being researched. The aim of my work was not to perform a detailed etymological analysis but rather to describe the situation mainly as presented in *OED*.

5.1.1. Content and function words – Germanic descent

Algeo and Pyles (2004, 271) claim that "everyday things (…), relationships (…), responses and actions (…), basic numbers and directions (…) and grammatical words (…) are all native English." As discussed in Section 3, "native English" is a rather vague identification of the language. The previous section presented two different views on the origin of today's English. Both views, however, claim it descended form a Germanic language; be it a Western one (Old English) or a Northern one (Old Norse).

As for the grammatical words, Algeo and Pyles's examples include personal pronouns (*I*, *you*, *he*), prepositions (*to*, *for*, *from*, *after*), auxiliary verbs (*be*, *have*) and conjunctions (*but*, *and*). My classification, which was described in the Methodology section, agrees with the one by Algeo and Pyles.

It follows from Figure 5 that the core vocabulary I pronounced as Germanic contains **47 per cent of function words**. It can be stated that function words constitute one half of the Germanic core vocabulary found among the 500 most frequent items.

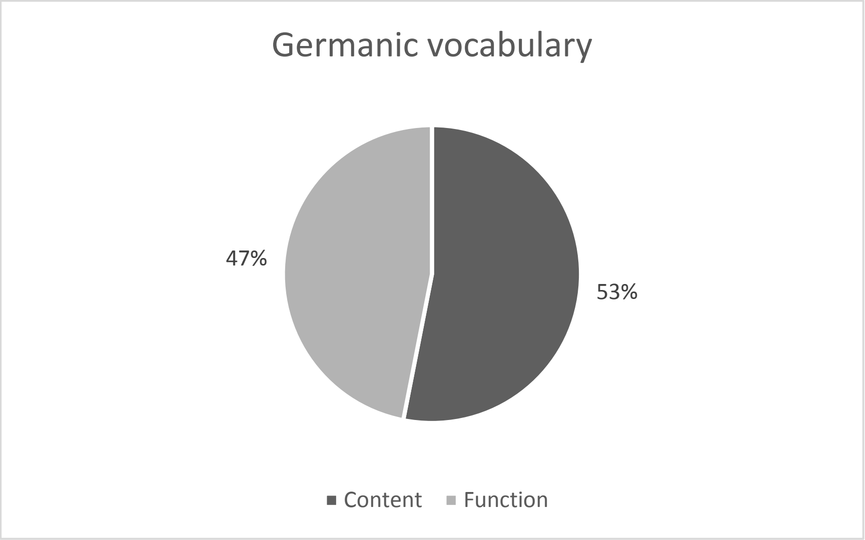


Figure 5: Pie chart illustrating the share of content and function words in the Germanic core vocabulary found in the sample

Moving on now to consider the **content words**, which account for **53 per cent** of the Germanic vocabulary, I would like to compare the lexical fields linguistic literature mentions when speaking of Germanic vocabulary with the lexical fields identified in my sample. Emonds and Faarlund provide similar examples to the ones given by Algeo and Pyles. They claim that language sub-familes usually share "daily life vocabulary". This is supposed to cover "basic counting, kinship terms, familiar body parts, and vocabulary for natural things" (Emonds and Faarlund 2014, 18).

My sample of Germanic core vocabulary contains several **basic numbers** (mentioned in both, Algeo and Pyles, and Emonds and Faarlund), namely *one* (rank 36), *two* (rank 71), *three* (rank 125), *four* (rank 250), *five* (rank 344), and *six* (rank 386). As the *new-GSL* mostly distinguishes individual parts of speech, *one* as a pronoun was counted separately from the numeral and given the rank of 314. Interestingly, the higher the number is, the lower its frequency of appearance is. Additionally, I identified the ordinal number *first* (rank 437) and the adverb *once* (rank 239).

The next category, named again in both sources, is “kinship terms” or, more generally, **“relationships”**. It is represented only by the nouns expressing the closest relations; i.e. *friend* (rank 273), *mother* (rank 375) and *father* (rank 477).

As for the “familiar **body parts”**, mentioned by Emonds and Faarlund, the list includes the word *body* itself (rank 291) and the nouns *hand* (rank 155), *head* (rank 204), *eye* (rank 244), *foot* (rank 439), and *heart* (rank 488). It can be noted that these words are frequently used metaphorically, e. g. in phrases such as *give somebody a hand* or *the head of the company*.

A great deal of the analysed vocabulary is represented by “**responses and actions”** (using Algeo and Pyles’s words). To be concrete, verbs account for 23 per cent of the Germanic vocabulary. They concern e.g. communication (*say*, *ask*, *speak*, *talk*) movement and “every-day activities” (*make*, *work*, *help*, *run*, *live*, *fall*, *sit*), sensing (*see*, *feel*, *hear*), “commercial activities” (*buy*, *spend*) and fighting or competition (*lose*, *win*).

The “vocabulary for **natural things**” (mentioned by Emonds and Faarlund) can be exemplified by nouns related to time, its measuring and phases (*time*, *year*, *today*, *day*, *week*) and life in general (*water*, *life*, *food*, *death*).

Additionally, the sample contains nouns describing **human inventions and “institutions”** (e.g. *house*, *door*, *road*, *war*, and *law*), names for **people** (e.g. *child*, *man*, and *woman*) and nouns related to **communication** (e.g. *word* and *book*).

Turning now to adjectives, it can be summarised that they mostly describe **“basic qualities”**. It is not unusual to come across cases of synonymy (e.g. *small* and *little*) and especially antonymy, which can be illustrated e.g. by the following pairs: *new* (rank 87) or *young* (rank 224) versus *old* (rank 160); *high* (rank 145) versus *low* (rank 302); *small* (rank 166) or *little* (rank 174) versus *big* (rank 310); and *short* (rank 394) versus *long* (rank 395).

As for the adjectives describing colours, the only one present in the sample was *white* with the rank of 463.

5.2.Romance vocabulary among the 500 most frequent words

The Romance language family covers all languages which evolved from Vulgar Latin. As it was illustrated in Figure 4, **34 per cent of the core vocabulary analysed were identified as Romance**. My vocabulary sample contained those words of Romance origin which were either adopted from Latin, adopted from French or it is uneasy to determine which of the two served as the direct source. In addition, the etymology of one word was too problematic to be included in any of the three categories established. More information about this particular case is to be provided later.

Returning to the first type, i.e. the direct influence Latin had on the English language, Algeo and Pyles (2004, 272) claim, it can be "seen in every period of the language's history."[18] As they report, Latin first influenced the language before English was even separated from Germanic, thus various forms of the borrowings can be found in all Germanic languages. "[These early loanwords] are mostly concerned with military affairs, commerce, and agriculture or with refinements of living that the Germanic peoples had acquired through a fairly close contact with the Romans" (Algeo and Pyles 2004, 272). Algeo and Pyles (2004, 273) identify approximately 175 of these early borrowings, many of which have remained in the language until now.

Moving on to the Old English, Baugh and Cable (2002, 74-75) recognise two periods of Latin influence. The first one can be traced back to the arrival of the Germanic tribes following the exodus of the Romans. Baugh and Cable (2002, 74) explain the situation as follows:

> It is probable that the use of Latin as a spoken language did not long survive the end of Roman rule in the island and that such vestiges as remained for a time were lost in the disorders that accompanied the Germanic invasions. There was thus no opportunity for direct contact between Latin and Old English in England, and such Latin

[18] Algeo and Pyles (2004) follow the traditional view on the history of English, which supposes that Middle English evolved from Old English.

> words as could have found their way into English would have had to come in through Celtic transmission.
>
> Baugh and Cable (2002, 74)

The influence of the first period was, however, minimal and, apart from some five words, the borrowings survived almost exclusively as a part of place names (Baugh and Cable 2002, 74). The second period of Latin influence on Old English is connected to the Christianising of Britain, which began in 597. Since then, as Baugh and Cable (2002, 77) claim, Latin words were gradually entering English until the very end of the Old English period.

The first 500 words in the *new-GSL* include altogether five possible Latin borrowings from the Old English period, namely: *place* (rank 139), *turn* (rank 172), *school* (rank 232), *study* as a verb (rank 420) and *history* (rank 495). The noun *case* (rank 156) could be added as its grammatical sense and the sense of "particular circumstance or situation" were also borrowed in Old English times from Latin.

The origin of the noun *place* is slightly complicated, as *OED* claims it is a Latin borrowing, however, "modelled on a French lexical item". Some doubt is also casted upon the origin of the noun *history* since it was firstly borrowed from classical Latin and later reborrowed from Old French or Latin.

All the six possible Latin borrowings were compared with a handful of examples provided by Algeo and Pyles (2004, 272-274) and the more extended list of examples by Baugh and Cable (2002, 78). The first source did not include any of them, however, Baugh and Cable, without any doubt, listed the words *place*, *turn*, and *school* as Latin loanwords (2002, 78). The word s*tudy* was pronounced as a French borrowing from the Middle English period, however, they only referred to it as a noun (Baugh and Cable 2002, 160). The words *case* and *history* were not discussed.

The Middle English period was rich in Latin borrowings. As Algeo and Pyles (2004, 275) report, there were "hundreds of Latin words adopted before 1500". My vocabulary simple contains 17 Latin words which were borrowed in Middle English times and 4 words which were first recorded after 1500, i. e. not in the Middle English period. To be concrete, the four words are *area* (rank 187), which appeared in written records before 1552; *result* as a noun (rank 231), which first appeared in 1626; *expect* (rank 278), which has been known since 1535; and *suggest* (rank 298), which was first recorded in 1526. All of them, with the exception of the noun *result*, were found in written records only shortly after 1500 which means they may have been already used in spoken language in the Middle English period. The noun *result* was, according to *OED*, first recorded in the 17th century, the same source, however, claims it was created by conversion from the verb *result* which first appeared probably before 1425. Therefore, I suggest that all the Latin words appearing among the first 500 words in the *new-GSL* could have been possibly borrowed before the end of the Middle English times.

Altogether, I have **identified 27 Latin borrowings, which account for about 16 per cent of the whole Romance vocabulary analysed**. Apart from the slightly doubtful cases mentioned earlier, the group of Latin borrowings includes two other words with problematic etymology; these are *involve* and *across*. Although the authors of *OED* conclude the verb *involve* is of Latin origin, they also suggest comparing it with Old French *involver*.

As for the word *across*, *OED* informs it was constructed by the combination of the Germanic prefix *a* and the stem *cross*, which was derived by different channels from Latin *cruc-em*. Algeo and Pyles (2004, 271) define borrowing as a process in which "speakers imitate a word from a foreign language and at least partly adapt it in sound or grammar to their native speechways." Similarly, Němec (1976), as mentioned before, explains that borrowed items are gradually adapting the features typical for the lexemes which are central to the language. I believe that taking part in derivational processes with domestic affixes is one of the features confirming the item was fully adopted in the English language. Consequently, I decided to classify the word *across*, along with others which were formed this way, according to the origin of the stem.

In the Middle English period, as Algeo and Pyles (2004, 274) put it, English borrowed many words for which "it is [frequently] impossible to tell whether [they] are from French or from Latin ". Emonds and Faarlund (2014, 18-19) explain there are three criteria linguists consider when studying the relationships between languages. Those are "regular sound changes", shared morphosyntax, and, finally, shared core vocabulary. A great number of cognates appear among the core vocabulary of Romance descent and linguists are not able to state the precise origin of such words as both (or all) potential etymons are very similar.

Most French words English has borrowed originally come from Latin. There was, however, one problematic word as for the ultimate source, namely the verb *wait*. The authors of *OED* inform that the word's ultimate source is common Germanic. English, however, borrowed this word from Old North French, which acquired it by way of Old High German. Although this work intends to classify words with respect to their direct origin, in this case, the imparity between the direct and ultimate source was too large to ignore, hence the word was classified as Germanic.

Altogether, my sample contains **49 words *OED* claims to be of "multiple origins"**; they account for **29 per cent of the Romance vocabulary**. This group also includes two words which were formed by combining a Romance stem with the Germanic suffix -*ly*, i.e. *probably* (rank 319) and *particularly* (rank 475).

As for the somewhat problematic words included in this group, the noun *line,* as the authors of *OED* claim, originated when "two words, ultimately of the same etymology, coalesced"[19]. The first one was an early Germanic adoption of the Latin *līnea* (*líne* in Old English). The second one was adopted into Middle English by the way of French as *ligne* or *line*. The other problematic word is the noun *quality,* which was borrowed from French or Latin, but either way, it was modelled on a Greek item.

To move on now to the French vocabulary, the greatest spur of French borrowings took place in the Middle English period, when Britain was under French control. Baugh and Cable (2002, 156) inform that "[i]n this movement two stages can be observed, an earlier and a later, with the year 1250 as the approximate dividing line." As they later explain, "the years from 1250 to 1400 mark the period when English was everywhere replacing French. During these 150 years 40 percent of all the French words in the English language came in" (Baugh and Cable 2002, 165).

In total, I have identified 94 French items in my sample. All but three of them were firstly recorded before the year 1500 or shortly after. The exceptions are the words *development, develop* and *plan,* which were recorded later. The noun

[19] Oxford English Dictionary Online, s.v. "line", accessed March 30, 2017, SOURCE: http://www.oed.com/view/Entry/108603.

development (rank 328) was firstly recorded in 1724, its immediate etymon, i.e. the verb *develop* (rank 339), first appeared in 1653. The first record of the noun *plan* (rank 433) comes from 1635.

The French group also contains words which were formed by combining a French stem and a Germanic affix. Firstly, they are adverbs derived from adjectives by adding the suffix *-ly*[20]; namely *usually* (rank 413), *simply* (rank 431), *certainly* (rank 444), and *especially* (rank 479). Secondly, it is the word *around*, which was derived by adding the Germanic prefix *a-* to the French stem *round*, which, as it is claimed in *OED*, might have been modelled on an English item.

Finally, the word *per cent* (rank 409), which was first recorded in 1569, has such a complicated etymology it was not included in any of the three types I recognised within Romance vocabulary.[21] *OED* states that *per* is partly a borrowing from French, partly a borrowing from Latin. *Cent* is also of multiple origins. It may from the French *cent* or the Latin *centum*. *OED*, moreover, suggests it could have been formed within English by clipping or shortening. The authors of the dictionary also propose comparing English *per cent* with Middle French and French *pour cent*, Spanish *por ciento*, Dutch *per cent* (historical *per cento* and contemporary *pro cent*), and early modern German *per cento* or *percento*

Additional reference is provided to the adverb *per centum*, which is classified as a borrowing from Latin, modelled on an Italian lexical item.

To summarise the numbers within the vocabulary of **Romance origin**, the whole group accounts for about **34 per cent (approximately one third) of the whole core vocabulary sample**. The Romance group is then formed by **16 per cent by Latin borrowings**, by **55 per cent by words of French origin**, and by a little more than **29 per cent by words of multiple origins**. The unique word *per cent*, which did not quite match any of the types, accounted for less than one per cent. The proportions are illustrated in Figure 6.

[20] This claim refers to the traditional view on *-ly*, which sees it as a derivational suffix

[21] These are Latin borrowings, French borrowings, and words of multiple origin (Latin or French).

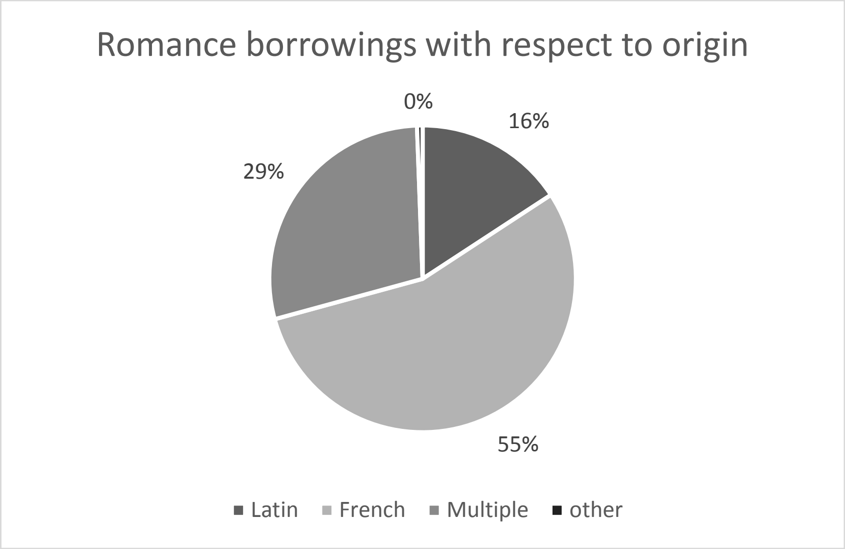


Figure 6: Pie chart illustrating the proportions of vocabulary of various origins within the Romance group.

5.2.1. Content and function words – Romance descent

Although Algeo and Pyles (2004, 271) claim that all grammatical words are native English, it is not entirely true. Unsurprisingly, the core vocabulary of Romance origin is by 95 per cent made up of content words. The remaining 5 per cent, however, stand for nine words I evaluated as function. These are the adverb *just* (rank 75), the adverb *very* (rank 86), the preposition *during* (rank 179), the preposition *around* (rank 262), the adverb *perhaps* (rank 292), the preposition *per* (rank 324), the determiner *several* (rank 357), the preposition *across* (rank 373), and the preposition *according to* (rank 410). The ratio between content and function words is illustrated in Figure 7.

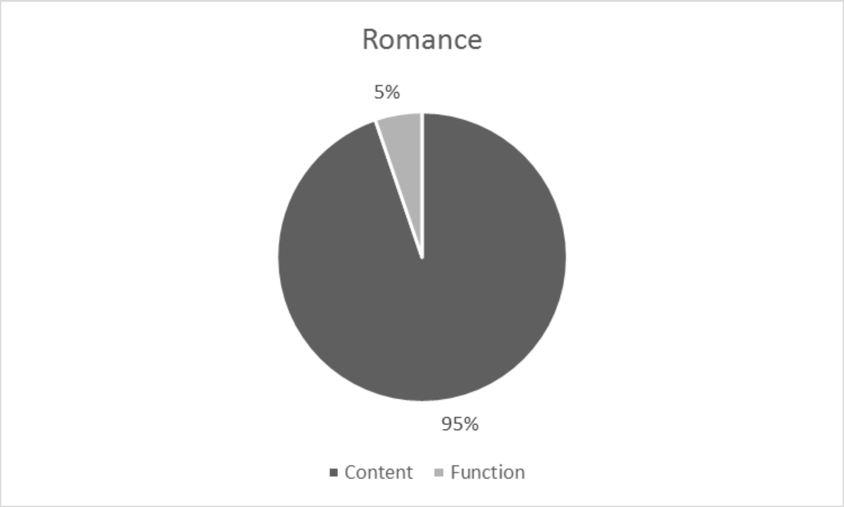


Figure 7: Pie chart illustrating the ratio between content and function words in Romance vocabulary.

As for the content words of Romance origin, Baugh and Cable (2002, 157-160) talk about six major lexical fields influenced by French vocabulary; these are: governmental and administrative words; ecclesiastical words; law; army and navy; fashion, meals; and social life; and finally, art, learning, medicine. These lexical fields include not only French borrowings but also Latin ones. Algeo and Pyles (2004, 275) claim that Latin gave English many “words having to do with religion”, “legal terms”, “words having to do with scholastic activities or science” and a number number of verbs and adjectives. In general, it can be said that Romance vocabulary is concerned with more “cultivated activities”.

Governmental and administrative words are richly represented in the sample. To name a few examples, they include the words *order* (rank 222), *government* (rank 235), *state* (rank 295), *office* (rank 374), *public* (rank 382), *force* (rank 416), *subject* (rank 441), and *authority* (rank 467), which are all French.

The following three lexical fields listed by Baugh and Cable, i.e. ecclesiastical words, law, and army and navy, are probably not represented in the sample at all. This is perhaps due to the fact that many of the words are in fact technical terms and the fields they belong to, especially in the case of church, nowadays play a less important part in people’s lives.

The sample does contain several words related to social life or civilisation in general. To name a few examples, these are *people* (rank 79), *country* (rank 194) and *city* (rank 414). The words related to learning are also quite frequent. They are represented e.g. by the French *course* (rank 189) or *story* (rank 499) and the Latin *fact* (rank 181), *school* (rank 232), and *idea* (rank 233).

As for verbs, several “basic” ones can be identified in the Romance core vocabulary; for instance, *try* (rank 154), *move* (rank 220), *change* (rank 253), and *carry* (rank 272). However, formal verbs prevail; e.g. *consider* (rank 241), *expect* (rank 278), *suggest* (rank 298) and *support* (rank 397). Sometimes, the sample also contains their less formal synonyms of Germanic origin. Compare e.g. *provide* (rank 177) and *give* (rank 74) or *remain* (rank 263) and *stay* (rank 442). The same is true for the Romance adjective *large* (rank 178) and the probably Germanic *big* (rank 310).

5.3.Problematic cases

The present section provides a detailed discussion on the etymology of four words found among the first 500 words of the *new-GSL* whose origin is, according to *OED*, unknown or uncertain. This is to advocate the fact they were not included in either of the types I have recognised. The words discussed are the adjective *bad* (rank 305) and the nouns *job* (rank 308), *girl* (rank 385), and *boy* (rank 435).

Firstly, *OED* suggests that the Modern English *bad*, which was first recorded in 1203, might be related to Old English *bæddel*, meaning “hermaphrodite, effeminate or homosexual man”. *Bæddel* was, however, recorded solely in glossarial sources.

Another proposal says that *bad* may have been derived directly from *gebǣded*, which is the past participle of the verb *bǣdan*, meaning “to force, constrain, impel, to require, demand, exact, to urge, incite”. Nevertheless, the authors of *OED* are rather sceptical and consider this suggestion less likely than the former one. Though both possible etymologies *OED* presents see *bad* as a Germanic word, there is no direct evidence of its origin, hence *bad* was not included in the Germanic group.

Secondly, concerning the noun *job*, several entries are provided by the authors of *OED*. I believe that nowadays the word *job* is most frequently used in the sense of “a piece of work”, “task” or “occupation”, which is presented in the second entry. The third entry represents the noun *job* which has two senses: 1) “a cartload” and 2) “a stamp, a block” or “a tassel”. The authors go on to theorise the following:

> In sense 1 the original meaning was perhaps ‘an amount the carrying of which constitutes a single job (i.e. task)’, in which case the word would apparently show a spec[ific] use of [the noun *job* presented in the second entry]. However, if that were the case, it would not be easy to explain the semantic development apparently represented by sense 2, which suggests an underlying concrete sense, perhaps ‘piece’ or ‘mass’.[22]

Consequently, it is suggested that that the meaning presented in the second entry “originally itself represents a spec[ific] sense development of [the noun job presented in the third entry] in the phrase *job of work*”[23]. However, no further etymology is proposed.

The authors of *OED* finally add that “a connection with *gob* [which is “apparently a borrowing from French”] has also been suggested but poses phonetic and semantic difficulties”.[24]

Thirdly, Diensberg (1984, 473) claims “[i]t is widely known that the words for *boy*/*girl* represent etymological puzzles in English as well as in other European languages”. According to *OED,* many linguists have reconstructed an Old English form of the Modern English *girl,* i.e. **gyrela.* No further etymology of this form was, however, found plausible by recent scholars.

As the authors of *OED* further inform, F. C. Robinson’s article *European Clothing Names and the Etymology of Girl* implies the word *girl* is related to Old English *gyrela,* which means “robe” or “dress”. Robinson’s explanation was, though, criticised by certain scholar. By way of illustration, Diensberg (1984) argues Robinson’s hypothesis is based on wrong connections and lacks evidence.

[22] Oxford English Dictionary Online, s.v. “job”, accessed March 30, 2017, http://www.oed.com/view/Entry/101395.

[23] *ibid*.

[24] *ibid*.

Another theory was proposed suggesting a parallel between the English *girl* and the Middle Low German *Gör, Göre* (meaning "girl" or "small child"). However, the authors of *OED* declare "this explanation encounters chronological difficulties and also fails to account well for the variation in stem vowel shown by the Middle English word."

Last but not least, *OED* informs that the origin of the Modern English *boy* is uncertain "as is the early development of the word"[25]. Two etymologies have been suggested. The first one considers *boy* a French borrowing while the other one claims it is of Germanic origin.

Two French etymons have been suggested. As reported in *OED*, the more widely accepted hypothesis by E. J. Dobson proposes a connection with Anglo-Norman *boie*, meaning "a male servant". The authors of *OED* believe that "[s]uch an origin would account well for the variation shown by the English word, and also gives an entirely plausible explanation of the sense 'male servant'"[26], which was one of the former senses of the English word.

Diensberg (1981), however, criticises Dobson's theory, especially on the phonological level. He argues the word *boy* is related to Old French *boiasse*, an infrequent variant of Old French *baiasse*, meaning "female servant". He assumes that the ending *-asse* was interpreted as the feminine suffix *-esse*. Consequently, a masculine form *boie* was formed by analogy. The authors of *OED*, however, question the evidence Diensberg offers.

Other scholars, who are convinced the word's Germanic origin, suggest comparing Modern English *boy* with West Frisian *boi* and German regional (Low German) *boi, boy* with the meaning of "boy" or "young man". As the authors of *OED* add, the explanation was further developed in *De etymologie van Fries "boai", Engels "boy" en Middelnederlands "boye"* by K. Roelandts, where he argues that all these forms go back to a mutual etymon; i.e. a familiar or nursery form of "brother".

[25] Oxford English Dictionary Online, s.v. "boy", accessed March 30, 2017, http://www.oed.com/view/Entry/22323.

[26] *ibid*.

5.4.Content and function words in the whole sample

Having discussed the share of content and function words in the Germanic and Romance groups separately, I will now briefly compare the ratio between them in the two groups and present the amount of function and content words in the whole sample. The illustration of the ratio in the whole core vocabulary sample is provided in Figure 8. The sample contains **twice as much content words as function words**. Content words account for 68 per cent of the vocabulary while function words account for 32 per cent.

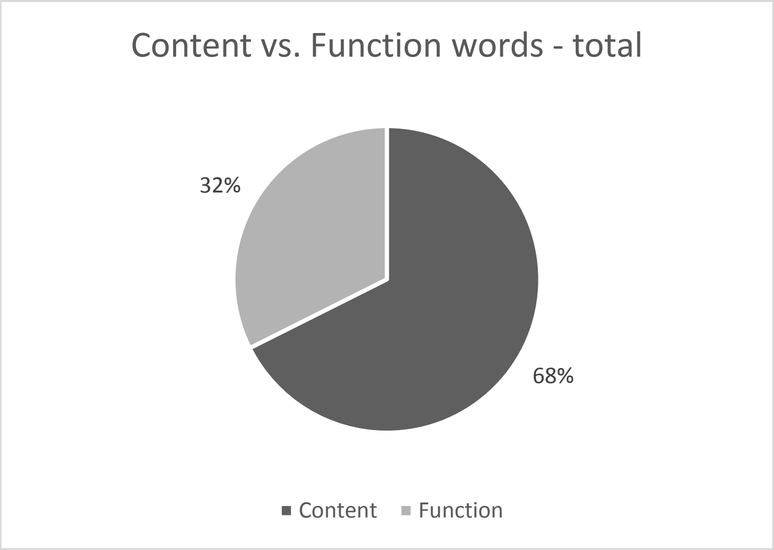


Figure 8: Pie chart illustrating the ratio between content and function words in the whole vocabulary sample.

Figure 9 illustrates the differences in the ratio between content and function vocabulary in the groups of various origins. **The amount of content words found within the Germanic vocabulary is almost equal to their amount found in the Romance group. The function vocabulary, on the other hand, is almost exclusively of Germanic origin.** Note that the graph also shows four words of unknown origin which are all content.

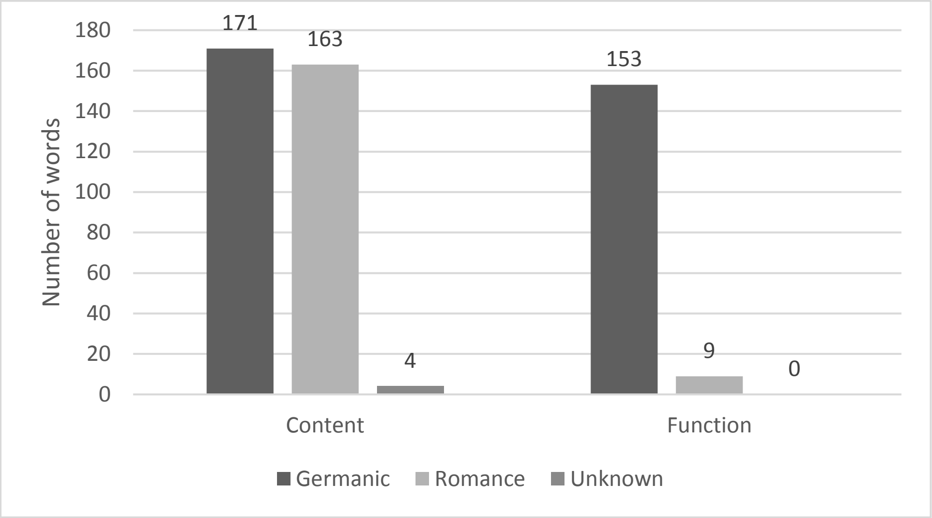


Figure 9: Bar charts comparing the amount of content and function words with respect to their origin.

5.5.Enlarging the scope of the core

Crystal (2003, 124) states that "[i]n the million-word Brown University corpus of written American English …, the 100 most frequently used items are almost all Anglo-Saxon [with the exception of a few Scandinavian loans]; there is nothing from Romance sources until items 105 (*just*) and 107 (*people*)." To illustrate this claim and describe the situation in my sample, I have counted and compared the amount of words of Germanic, Romance, and unknown origin in the first, second, third, fourth and fifth hundred of the first 500 items in the *new-GSL*. The results are illustrated in Figure 10. The **first 100 words are almost exclusively Germanic** until items 75 (*just*) and 76 (*use* as a verb), followed by the noun *people* (rank 79) and the adverb *very* (rank 86).

In summary, there has been a **gradual increase in the number of Romance words in the core vocabulary** until the third hundred where the amounts of Germanic and Romance words are close to equal. The fourth and fifth hundred, on the other hand, show a slightly increasing predominance of Romance sources.

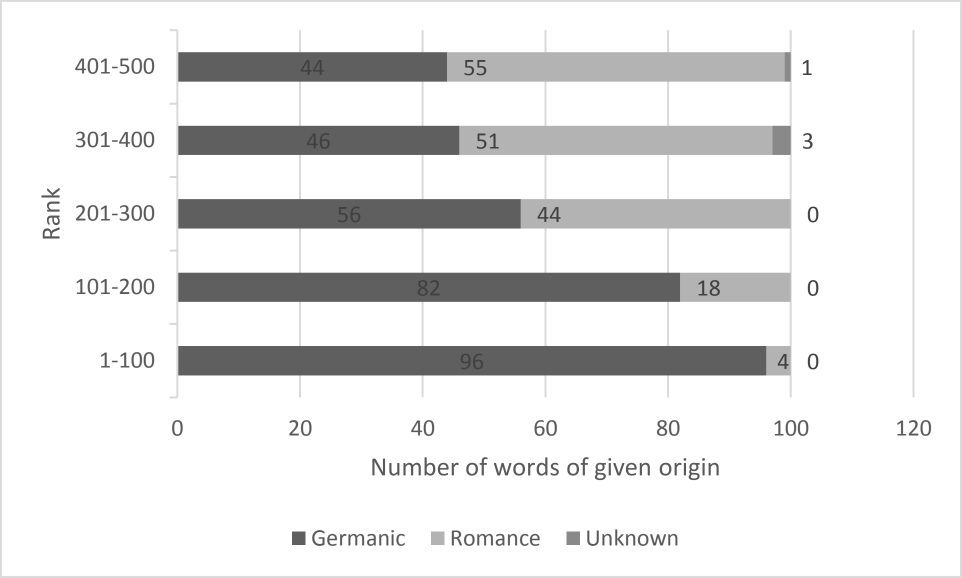


Figure 10: 100% stacked bar charts illustrating the change of the ratio between the words of given origin with increasing frequency.

Having discussed enlarging the scope of the core with respect to origin, let me now consider how the ratio between function and content words changes when the core is being gradually enlarged. As mentioned before, the majority of function words are supposed to be located in the very core of the lexis. The amount of function words is then expected to decline steadily as the frequency of use drops.

As Figure 11 reveals, function vocabulary accounts for more than two thirds of the first hundred words. However, there has been a steep rise of content words in the second hundred; content words account for a little over two thirds of it. A slight rise continues in the third and fourth hundred but stagnates in the fifth hundred.

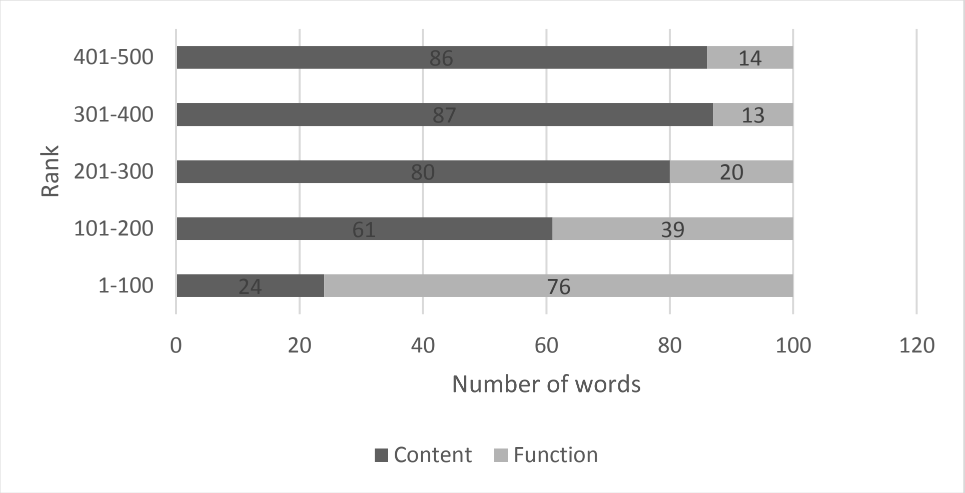


Figure 11: 100% stacked bar charts illustrating the change of the ratio between content and function words with increasing frequency.

6. Conclusion

The aim of this thesis has been to inspect the English core vocabulary with respect to its **origin** and describe the **structure of the lexical core** in terms of a) the amount of content and structure words and b) the lexical fields represented.

With regards to the first research question, I examined the etymology of the first 500 items in the *New General Service List* by Václav Březina and Dana Gablasová. The main source of the etymological information was provided by the online version of the *Oxford English Dictionary*. Additionaly, I consulted Algeo and Pyles's *The Origins and Developments of the English Language*, Baugh and Cable's *A History of the English Language*, and Emonds and Faarlund's *English: The Language of the Vikings*.

I have identified two sources of loanwords, i.e. Germanic and Romance languages. A further discussion was provided as to the possible identification of etymological sub-types within the two main groups and the uncertain etymology of some items.

I further intended to count the words of the two main types found in the sample and compare the results with Finkenstaedt and Wolff' research (paraphrased in Burnley 1992, 415). My study claims that **65 per cent of the 500 most frequent English words listed in the *new-GSL* are of Germanic descent. Romance loans represent 34 per cent of it and words of uncertain origin account for 1 per cent.**

Finkenstaedt and Wolff declared that nearly half of English "commoner words" are Germanic while Latin and French loans each account for a little more than 28 per cent. The differences in the results of the two studies are probably caused by different approaches to the delimitation of core vocabulary. Whereas my sample contained 500 words, Finkenstaedt and Wolff may have possibly dealt with thousands of items. The imparity may, however, be helpful in projecting trends in samples of core vocabulary larger than 500 words. In larger samples, the share of Romance vocabulary is supposed to increase while the share of Germanic vocabulary will decrease.

To summarise the findings regarding the second research question, i.e. the structure of the core, it was observed that **a vast majority of English function words found among the first 500 items of the *new-GSL* are of Germanic origin. Romance words, on the contrary, are almost exclusively content**. Enlarging the scope of core vocabulary implies a rise in the number of content words and words of Romance descent.

The lexical fields identified in the **Germanic content vocabulary** are mainly connected to **every-day life, describing basic actions, basic qualities, names for people and their relations**. The **Romance words**, on the other hand, are generally concerned with more sophisticated occasions.

A remarkably high number of Romance words in the sample were to do with administration. Additionally, several words I worked with had a surprisingly high rank; e.g.: *suggest* (rank 298), *per cent* (rank 409), *authority* (rank 467), *quality* (rank 478), compared to e.g. *mother* (rank 375), *morning* (rank 465) *food* (rank 468), *father* (rank 477). That raises the question if the *new-GSL* really captures the very core concepts of the language.

To answer this question, I consulted the basic vocabulary lists presented in the top five learner's dictionaries discussed in Section 3.2. All these dictionaries mark the words *mother*, *morning*, *food*, and *father* as the most frequent. Regarding the words with a surprisingly high rank, the information slightly differs. *LDOCE5* marks the word *per cent* with the abbreviations S3 and W2, which means it belongs to the third thousand of the most frequent words in spoken language and the second thousand in written language. The word authority is one of the one thousand most frequent words in written language, however, the entry lacks information about the spoken language, which implies the word is not among the 3000 most frequent items used in spoken language. The other words discussed were all found among the most used items in both written and spoken language.

All the other learner's dictionaries agreed the words *suggest*, *per cent*, *authority* and *quality* are among the most basic words of the English language. The most likely causes of this fact are the large amount of words the dictionaries pronounce as the most basic (in the case of *CALD2*, it is as many as 4,900 items) and the importance attached to frequency as a criterion for the delimitation of basic vocabulary.

Future studies on the current topic are recommended in order to obtain more relevant information about the origin and structure of the lexical core. I suggest expanding the research to the whole range of the *new-GSL* (a little less than 3000 items) and comparing the data with the content of the *Oxford 3000*, which does not only include words high in frequency but also a number of words which are not as frequent but, according to language experts, are still very important for language users.

7. České resumé

Cílem této bakalářské práce bylo prozkoumat centrální slovní zásobu anglického jazyka z hlediska původu a popsat strukturu anglického lexikálního jádra se zřetelem k množství autosémantik a synsémantik a k zastoupeným lexikálním polím.

Za účelem zodpovězení první výzkumné otázky jsem podrobila etymologické analýze prvních 500 položek uvedených v *New General Service List* Václava Březiny a Dany Gablasové. Hlavním zdrojem etymologických informací byla online verze slovníku *Oxford English Dictionary*. Dále jsem využila publikaci *The Origins and Developments of the English Language* autorů Algea and Pylese, *A History of the English Language* autorů Baugha and Cablea a knihu *English: The Language of the Vikings* od autorské dvojice Emonds a Faarlund.

Byly identifikovány dva zdroje výpůjček, a to germánské a románské jazyky. V rámci těchto dvou skupin bylo navrženo dělení do několika podtypů. Několik slov ze vzorku nemohlo být zařazeno k žádnému z typů, protože jejich původ je nejasný.

Mým záměrem bylo spočítat slova germánského a románského původu a porovnat tyto výsledky s výsledky studie provedené Finkenstaedtem a Wolffem (parafrázovanými Burnleym 1992, s. 415). Dle mé studie je vzorek 500 nejfrekventovanějších anglických slov podle *New General Service List* z 65 procent tvořen slovy germánského původu. Románské výpůjčky jsou zastoupeny 34 procenty a slova neznámého původu tvoří 1 procento.

Naproti tomu Finkenstaedt a Wolff tvrdí, že téměř polovina „běžnějších slov" je germánského původu, zatímco latinské a francouzské výpůjčky jsou shodně zastoupeny o něco málo vice než 28 procenty. Rozdíly mezi výsledky obou studií jsou patrně způsobeny rozdílným přístupem k vymezení centrální slovní zásoby. Můj vzorek obsahoval 500 slov, kdežto Finkentsaedt a Wolff pravděpodobně pracovali s tisíci položek. Rozdíl mezi těmito výsledky může být využit k popsání tendencí ve vzorcích lexika větších než 500 slov. Ve větších vzorcích bude pravděpodobně stoupat zastoupení románských slov a klesat zastoupení germánských slov.

Poznatky týkající se druhé výzkumné otázky, tj. struktury jádra, lze shrnout následovně: Bylo vypozorováno, že valná většina anglických synsémantik figurujících mezi 500 nejfrekventovanějšími slovy je germánského původu. Na druhou stranu románská slova jsou téměř výlučně autosémantická. Pokud se zvětšuje rozsah jádra, narůstá ve vzorku počet autosémantik a slov románského původu.

Lexikální pole zastoupená mezi slovy germánského původu se povětšinou vážou ke každodennímu životu, základním dějům, základním vlastnostem, názvům osob a pojmenování vztahů mezi nimi. Naopak románská slova se obecně vážou k „sofistikovanějším" činnostem.

Pozoruhodně vysoký počet slov románského původu byl spojen s administrativou, navíc měla některá slova ze vzorku překvapivě vysoký rank. Srovnej např: *suggest* (rank 298), *per cent* (rank 409), *authority* (rank 467), *quality* (rank 478) a *mother* (rank 375), *morning* (rank 465) *food* (rank 468), *father* (rank 477). Nabízí se tedy otázka, jestli *New General Service List* skutečně zachycuje základní koncepty jazyka.

K ověření mohou být použity seznamy základních slov podle výkladových slovníků, které byly prezentovány v části 3.2. Ve všech těchto slovnících jsou slova *mother, morning, food* a *father* označena jako nejfrekventovanější. U skupiny slov s překvapivě vysokým rankem se údaje mírně liší. *LDOCE5* uvádí u slova *per cent* údaje S3 a W2, což znamená že bylo zařazeno až do třetí tisícovky nejužívanějších slov v mluveném jazyce a druhé tisícovky v jazyce psaném. Slovo *authority* zde sice patří mezi tisíc nejfrekventovanějších slov v psaném jazyce, ale údaj o mluveném jazyce není uveden, tudíž lze předpokládat, že v něm slovo tak frekventované není. Ostatní uvedená slova patří mezi nejfrekventovanější jak v mluveném, tak v psaném jazyce.

Ostatní prezentované slovníky uvádí všechna slova s překvapivě vysokým rankem mezi nejfrekventovanějšími, což může být způsobeno jejich velkým rozsahem základních slov (v případě *CALD2* až 4900 položek) a nadřazením frekvence výskytu nad jiná kritéria používaná k vymezení základní slovní zásoby.

Za účelem získání relevantnějších informací o původu a struktuře lexikálního jádra doporučuji budoucí výzkum rozšířit na celý rozsah *New General Service List* (téměř 3000 položek) a porovnat získaná data s obsahem seznamu *Oxford 3000*, který zahrnuje nejen slova s vysokou frekvencí výskytu, ale i slova, která nejsou tak frekventovaná, zato jsou odborníky považována za důležitá pro uživatele jazyka.

8. References

Algeo, John, and Thomas Pyles. 2004. *The origins and development of the English language*. 5th ed. Boston, Massachusetts: Thomson Wadsworth.

Baugh, Albert C., and Thomas Cable. 2002. A History of the English Language. 5th ed. London: Routledge.

Bauman, John. 2002. “About the General Service List”. Accessed February 23, 2017. http://jbauman.com/aboutgsl.html

Bogaards, Paul. 2008. "Frequency in learners' dictionaries." Proceedings of the XIII EURALEX International Congress, Barcelona, July 15-19.
Accessed February 23.
http://www.euralex.org/elx_proceedings/Euralex2008/015_Euralex_2008_Paul%20Bogaards_Frequency%20in%20Learnes%20Dictionaries.pdf

Browne, Charles. 2014. “A new general service list: The better mousetrap we’ve been looking for?” *Vocabulary Learning and Instruction*. 3 (1): 1-10. doi: 10.7820/vli.v03.1.browne

Březina, Václav. 2014. “The New General Service List”. Filmed [October 2014]. YouTube video, 18:22. Posted [October 2014]. https://www.youtube.com/watch?v=UDSqTsEPziU

Březina, Václav, and Dana Gablasová. 2015. "Is there a core general vocabulary? Introducing the New General Service List." *Applied Linguistics*. 36 (1). Accessed December 4, 2016. doi: 10.1093/applin/amt018

Burnley, David. 1992. “Lexis and Semantics.” In *Middle English*, vol. 2 of *The Cambridge History of the English Language*, edited by Norman F. Blake. Cambridge: Cambridge University Press.

Carter, Ronald. 1982. “A note on core vocabulary.” *Nottingham Linguistic Circular* 11(2): 39-50.

Corver, Norbert, and Henk van Riemsdijk. 2001 “Semi-lexical categories.” In *Semi-lexical Categories: The Function of Content Words and the Content of Function Words* (pp.1-19), edited by Corver, N. and H. van Riemsdijk. Berlin: Mouton de Gruyter.

Crystal, David. 2003. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press

Cvrček, Václav. 2011. "How large is the core of language." *Corpus Linguistics* 2011.

Čermák František, 2010. *Lexikon a sémantika*. Praha: Nakladatelství Lidové noviny

Čermák, František. 2010. Introduction to *Frekvenční slovník češtiny* by František Čermák et al.. Praha: Nakladatelství Lidové noviny

Daneš, František. 1966. "The relation of centre and periphery as a language universal." *Travaux linguistiques de Prague* 2:9-21.

Diensberg Bernhard. 1981. “The etymology of modern English “boy”: A new hypothesis.” *Medium Ævum* 50(1): 79-87.

Diensberg Bernhard. 1984. The etymology of modern English “girl”: An old problem reconsidered.” *Neuphilologische Mitteilungen* 85(4): 473-475.

Emonds, Joseph Embley, and Jan Terje Faarlund. 2014. *English: The Language of the Vikings*. Olomouc: Palacký University. http://anglistika.upol.cz/vikings2014/

Lee, David Y. W. 2001. “Defining Core Vocabulary and Tracking Its Distribution across Spoken and Written Genres.” *Journal of English Linguistics* 29(3): 250-278.

Longman Communication 3000 In *Longman dictionary of contemporary English*. 2009. Harlow: Pearson Longman.

Nation, Paul and Robert Waring. 1997. “Vocabulary size, text coverage and word lists.” In *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19), edited by Schmitt, N. and M. McCarthy. Cambridge: Cambridge University Press. http://www.lextutor.ca/research/nation_waring_97.html

Němec, Igor. 1976. "Vztah centrum-periférie v lexikálním vývoji." *Naše řeč* 59:118-124.

Němec, Igor. 1996. “Lexikální význam ve světle teorie Pražské školy.” *Slovo a slovesnost* 57(3):218-225.

Oxford English Dictionary Online. http://www.oed.com/

Peprník, Jaroslav. 2006. *English lexicology*. Olomouc: Univerzita Palackého

Richards, I. A. 1943. *Basic English and its uses*. New York: W. W. Norton and Company. Inc.

Sgall, Petr. 2011. *Jazyk, mluvení, psaní*. Praha: Karolinum Press.

Skrebtsova, Tatiana. 2014. “The concepts *centre* and *periphery* in the history of linguistics: from field theory to modern cognitivism” *Respectus Philologicus* 26 (31):144-151.

Šimková, Mária. 2014. “Centrum a periférie ještě jednou” In *Jazyk a slovník. Vybrané lingvistické studie*, edited by František Čermák. Originally published in *Člověk a jeho jazyk 3. Inšpirácie profesora Jána Horeckého*, edited by Mária Šimková. Bratislava: Veda, 2013.

Vachek, Josef. 1964. "Peripheral elements in the structure of language" In *On peripheral phonemes of Modern English,* 7-9. Brno Studies in English 4. Praha.

Vachek, Josef. 1966. The Linguistic School of Prague: an introduction to its theory and practice. London: Indiana University Press.

Veselovská, Ludmila. 2016. *English Morpohology*. Olomouc: Univerzita Palackého.

Walter, Elizabeth. 2008. Introduction to *Cambridge advanced learner's dictionary*. 3rd edition. Cambridge: Cambridge University Press.