Productivity of Neoclassical Compounds

Bulaja, Petra

Sveučilište u Zadru

Odjel za anglistiku

Sveučilišni prijediplomski studij

Anglistika

Petra Bulaja

Productivity of Neoclassical Compounds

Završni rad

Zadar, 2024.

Sveučilište u Zadru
Odjel za anglistiku
Sveučilišni prijediplomski studij
Anglistika

Productivity of neoclassical compounds

Završni rad

Student/ica:
Petra Bulaja

Mentor/ica:
Doc. dr. sc. Frane Malenica

Zadar, 2024.

Izjava o akademskoj čestitosti

Ja, **Petra Bulaja,** ovime izjavljujem da je moj **završni** rad pod naslovom **Productivity of Neoclassical Compounds** rezultat mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na izvore i radove navedene u bilješkama i popisu literature. Ni jedan dio mojega rada nije napisan na nedopušten način, odnosno nije prepisan iz necitiranih radova i ne krši bilo čija autorska prava.

Izjavljujem da ni jedan dio ovoga rada nije iskorišten u kojem drugom radu pri bilo kojoj drugoj visokoškolskoj, znanstvenoj, obrazovnoj ili inoj ustanovi.

Sadržaj mojega rada u potpunosti odgovara sadržaju obranjenoga i nakon obrane uređenoga rada.

Zadar, 6. rujan 2024.

Table of contents

1. Introduction

Compounds are words made out of two or three different bases. They can be spelled in various ways. Neoclassical compounds have an element derived from either Greek or Latin language. That is why they are called *neoclassical*, because within themselves they contain a neoclassical element. In this paper, I will research the productivity of neoclassical compounds, based on their neoclassical element, either on the initial or final position. In the third chapter, I will also mention some previously done research that dealt with similar topics. The conducted research, analyzed samples and methodology will be further explained in chapter four. All of the collected data was sampled using corpus British Web 2007 (ukWaC). I researched neoclassical compounds based on 10 initial combining forms and 10 final combining forms. The list of every word was checked to make sure it fits the description of the sample I needed to conduct my research.

2. Compounds and compounding

Compounds are words made of two or more different bases. They can be written as one word e.g. *bedroom* or with a blank space e.g. *high school* serving as a separating element and they can also be written with a hyphen, for example, *word-formation.* This example can be written with and without a hyphen, we can say that it is an individual preference. Plag (2003, p. 5) states that in some languages, like German, compounds are only treated as single words, but in English, they can be in the form of more than one word. For example, *girlfriend* (girl + friend) is written without a blank space whereas *apartment building* is written as two separate words. However, it is still seen as a compound. The orthography of compounds is not completely flexible. For instance, *rail way* with a blank space would be considered unusual. Bauer (2017, p. 6) states that whether we write compounds with or without a hyphen could be a matter of the length of the word. For example, we write *textbook,* but not *librarybook.*

Moreover, Plag (2003, p. 133) argues that defining compounding as a process of combining two words is open for discussion because there are examples that question this restricted definition, as seen in the example *power source requirement.* English grammar even enables compounds to be stacked one on the other, as in, for example, *university teaching award committee.* This process can go on and other words can be repeatedly added to the already existing compound, which is known as recursion. These kinds of compounds are not very desirable as listeners and speakers find them hard to process. Plag (2003, p. 135) states that compounds can consist of the first element which is a root, a word or a phrase, and the second element which is either a word or a root. Complex words can have either one free root or one bound root (Carstairs-McCarthy, 2002, p. 21), but compounds can have two roots, for example, *bookcase.* Because they can be coined by roots or, to some extent, by phrasal words, they have a structure of phrases but function as words (Carstairs-McCarthy, 2002, p. 59). While stress in noun phrases is on the right side of the phrase, in compounds the left side is always stressed, for example, *greenhouse.* This is how we can differentiate compound nouns from noun phrases. The most common compounds in English are in the form of nouns. This can be explained by the need to form new vocabulary because of cultural or technological change.

Bisetto and Scalise (2005, p. 321) state that we can divide compounds into two groups - endocentric and exocentric. Endocentric compounds are compounds that have a semantic head, whereas exocentric compounds do not. For example, *bluebird* is a type of bird, the head

is a right-hand element and it is an endocentric compound, whereas *scarecrow*, which is an exocentric compound, is not a type of crow, but a mannequin used to scare away birds from crops. A head is a part of a compound that represents the main meaning of the word and it is the same word class as the compound (Fabb, 1998, p. 67). For example, in *sneak-thief,* the head of the compound is a *thief,* while *sneak* only represents the kind of a thief. In this case, both original words are nouns. Endocentric compounds in English tend to have heads on the right side. Exocentric compounds do not have heads, for example, *redhead* describes a type of hair, not a type of head. Fabb (1998) states that regarding the endocentric and exocentric compounds the distinction between them is usually based on the interpretation. Furthermore, he provides the example *greenhouse* which can fall into both categories, endocentric (a type of house that is green) and exocentric (a glass house in which plants that need protection are grown). It is important to mention that determining the head here is of small relevance.

Neoclassical compounds are formed using elements from either Greek or Latin language. In order to understand the meaning of the word we have to be able to identify both the head and the combining form which was used in the coining of a compound. For example, *biology* consists of *bio* which means life and *logy* which means the study of something, therefore the meaning is "the study of life". Also, neoclassical compounds are mostly used in the fields of science, medicine and technology. This enables us to figure out their meaning using the context they belong to.

Compounds can be formed by joining two nouns, a verb and a noun, an adjective and a noun, a preposition and a noun, etc. Nominal compounds are compounds in which the head is a noun, e.g. *bedroom.* In this case, two nouns are combined together and the compound also functions as a noun. Nominal compounds can also be formed with an adjective and a noun, e.g. *goldsmith* or with a verb and a noun, e.g. *swimwear.* There are also adjectival compounds which function as adjectives. They modify nouns and can be formed in multiple ways, e.g. *worldwide* (N + ADJ). Verbal compounds are formed using verbs and they can form compounds that belong to either verbs or nouns, e.g. *babysit* (N + V). Adverbial compounds include adverbs and they usually modify verbs and adjectives, e.g. *overcook* (ADV + V).

Additionally, compounds belong to lexical words. The order of the base words and their word classes will determine the word class of the newly coined compound. For instance, if we want to coin a compound that is a noun, its first root can be of any word class, but its second root must be a noun, for example (1). The same principle applies to verb compounds (2). If

we coin an adjective compound, the first base word can be anything but a verb and the second one must be an adjective (3).

1. adjective + noun- *high school*
2. noun + verb- *babysit*
3. noun + adjective- *ice-cold*

As we can see, there are many ways of forming new compounds with many possible word class combinations. This enables constant coining of new compounds with new meanings and it is the reason why compounding is seen as one of the most prolific word formation processes.

2.1. Neoclassical compounds

Neoclassical compounds are compounds formed by using Classical Greek or Latin elements. They are a relatively new word formation in modern languages and English. Since they are modern formations used in modern languages, they were not produced for usage in classical languages (Bauer, 1998, p. 405). Examples of such compounds would be *geology, photograph, anthropology, psychology* and *neurolinguistics*. Bauer (1998, p. 406) states that there is a problem with neoclassical compounds because of the affix *-o-* that appears between the two elements. In Greek, it was a thematic vowel, which is a vowel that does not belong to the root of a word rather it stands at the ending of a base. Bauer (1998, p. 406) writes that there are four possible solutions to why the affix is so common:

a) it is seen as a linking element, for example between *phot-* and *–graph*
b) it is seen as the part of the first element
c) it is seen as the part of the second element and when the second element is attached to the first one it takes the *-o-* with it, e.g. *phraseograph*
d) it belongs to both the first and the last element and -oo- is shortened to -o-

Therefore, we can conclude that neoclassical compounds differ from native compounds in the usage of classical stems and in using a linking element. However, they are considered as compounds because they are formed in English by two bases.

Sometimes, it is hard to decide whether a neoclassical compound is a compound or a derivative. For example, *sociolinguistics* has the first part *soci(o)-,* which is a neoclassical base. The argument for treating this lexeme as a derivative is that *soci(o)-* is seen as a prefix here. However, the argument for looking at this lexeme as a compound is that the first part of

the word is seen as a base because it has a semantic value similar to lexemes. Because of this kind of formation, neoclassical compounding is sometimes seen as affixation in English.

Another controversy about neoclassical compounds is that they are often confused with clippings or blends, which are also hard to classify because of their irregular morphology. Bauer (1998, p. 407) gives an example of the lexeme *telethon.* It can either be seen as a product of clipping the word *television* fused with the splinter of the word *marathon,* or it can be seen as a blend of those two words. He states that because of this, it is not clear whether the first part *tele-* has enough semantic meaning to be considered as a whole lexical element.

Another example of confusion between neoclassical compounds and other word formations is the word *Eurocrat.* This lexeme might be a combination of the Greek affix *kratos,* which means "power" or it might be a clipping that was added to the splinter of the word *bureaucrat.* It could also be a blend deriving from the words *Europe* and *bureaucrat.* The same case is happening with the lexeme *technophobia,* which could really be a neoclassical compound but could also be a Greek combining form added to an English word, therefore it is hard to conclude whether it is a case of neoclassical compounding or prefixation. There are also words with clear analysis, such as *securocrat,* but they are a combination of two word formation processes, in this case, clipping with a combining form. Bauer also states that another reason why neoclassical compounds are problematic is because classical elements are sometimes paired with English affixes. For example, in the word *hydroponic*, derived from *hydro*, which means "water" and *phone,* which means "sound", with the English suffix "-ic", the meaning may be difficult to decipher. It is a system in which plant growth is stimulated by sound and it has no connection to an instrument called hydrophone. Furthermore, Bauer (1998, p. 408) states that, sometimes, affixes are seen as neoclassical bases. He gives an example of the lexeme *superette* (a small supermarket). *Super-* is not a base anywhere else, so it seems that this word is made up of a prefix and a suffix with no base. It may be that the first part of the word is taken from the lexeme *supermarket,* but it just does not carry enough meaning to be considered a base. Bauer (1998, p. 409) concludes that neoclassical compounds are not well defined and distinct enough from other word formations because we find them being used with other combining forms, affixes and other combining lexical items. However, they do differ because of their foreign parts. Unlike French suffixes, Latin and Greek suffixes are considered foreign. Their elements, such as, for example, *-ography* cannot stand alone. The main question of this paper is whether neoclassical compounding is a productive process and so it will be tried to explain later in this

paper by using corpus. However, Bauer (1998, p. 420) says that if neoclassical compounding is not recognized as its own category it cannot be productive. If it is seen as a whole category, it can be considered productive because there are many words in English today using Latin and Greek bases and many words have been formed by using this pattern.

2.2. Affixation and combining forms

Neoclassical compounds are formed using elements from the Greek or Latin language. According to Plag (2003, p. 74), these elements are considered to be bound roots. They usually come at the beginning or the end of a word and because of that, they have either initial or final position. If they were to be considered as affixes it would mean that each word containing them consists of a prefix and a suffix with no root in the middle, e.g. *biology*. Because of that, it is important to explain the main difference between affixation and compounding.

Affixation is one of the most productive word formation processes in the English language. The most frequent morphemes in English are affixes. Most words that came from Latin and Greek, or that have been coined from Latin or Greek roots, have one or more affixes. Some of the affixes are very frequent, which means that they have high productivity. However, unlike roots, affixes do not always have clear meanings as they often undergo the process of semantic bleaching. Minkova and Stockwell (2001) explain that it is a process in which the primary meaning of the word changes due to generalization. They provide the following example, the word *awful* completely changed its meaning, it used to mean "full of awe" and now it is a synonym for "terrible". Because of this process, during which their original meaning had completely faded, it is almost impossible to define them. For example, this happened with the English suffixes *–dom, -ly, -hood*. While it is usually easy to differentiate suffixes from roots, prefixes tend to be more root-like (Minkova, Stockwell, p. 157). Prefixes could be put into categories according to their meaning, but that does not mean that prefixes will only ever be used in that sense. Concerning suffixes, most of the time it seems that they only convert nouns into other word classes. While usually the last suffix of the word does that, some suffixes have more specific meanings than others. They can be put in groups, but again, they do not fit in only one specific context.

Plag (2003, p. 157) states that the main difference between affixes and combining forms is their ability to form a new word based on another element they are supposed to combine

with. For example, an affix can combine with a bound root (*prob-able*), and a root can take an affix (*baptism*) when creating new words. However, two affixes cannot combine together (*dis-able*), and a root cannot combine with some other bound root to create a new word. On the other hand, combining forms can combine with another combining form (*astrology*), words (*morpho-syntax*) or bound roots (*scientology*) to form a new word.

3. Previous corpus-based research

In the following chapter, I will analyze some of the previously done research on morphological productivity using corpus. The first one, which was conducted by Pustylnikov and Schneider-Wiejowski (2009), deals with the morphological productivity of German noun suffixes. They assumed that productivity is measured by affixes. After they collected the data using newspaper articles from different time periods they checked their hypothesis. Amiot and Dal (2005) researched neoclassical combining forms in French using lexeme-based morphology. They defined the categories describing neoclassical compounds. Although that criterion was highly successful, they came to the conclusion that not all neoclassical compounds can be described using the same method. Mititelu (2018) did her research on the productivity of affixes thinking that the derived word should contain both the meaning of the base word and the affix. She collected the sample and detected the most frequent affixes. Diaz-Negrillo (2014) researched the morphological behavior of neoclassical compounds and morphological development in this compound class. She studied some of the most important features of neoclassical compounds. Using the BNC she collected the data to research the productivity of combining forms in neoclassical compounds, the presence of the linking vowel and the behavior between old and new formations. My research deals with the productivity of neoclassical compounds based on their either initial or final combining form. I used the BNC to collect my sample. According to both the topic and the methods I used, it is similar to all of the pieces of research I mentioned.

In their research Pustylnikov and Schneider-Wiejowski (2009) deal with derivational morphology and morphological productivity of German noun suffixes. They compared three different German corpora- two of these were newspaper corpora from different periods of time and one was a speech corpus. They investigated the differences between written and spoken corpora. The newspaper corpora consisted of 2000 words from five different regions, all of which are in Germany, and within the periods between 1650-1700, 1701-1751 and 1751-1800. They compared these data with the newspaper corpus from the 20th century. This corpus consisted of 975 526 tokens. Afterwards, they measured the degree of productivity within the speech corpus, which consisted of 362 795 tokens. They assumed that productivity is measured by affixes when it comes to coining new words, or in other words, the contingency of the affix usage when it constructs a new word. Some of the suffixes they used were *–heit*, *-keit* and *–ung*. Those suffixes proved to be really productive because they are

used in many German nouns and because it is really common and easy to create new words with them. However, they were not productive in the earliest period they researched. The most productive suffix from that period was *–nis.* Concerning the spoken corpus, *-ung* was the most productive one, followed by *–nis.* They concluded that their results match the results of previous research concerning these suffixes in German.

Amiot and Dal (2005) analyze neoclassical combining forms in French using lexeme-based morphology. Since neoclassical compounds greatly differ from one another they can be described using various categories. They defined these categories as:

a) Lexematicity in the source language (Latin or Greek), which means these lexemes were associated with grammatical words.
b) The absence of syntactic realization in the targeted language (English, German, French), which means that neoclassical compounds will exclusively appear like bound constituents of lexemes and will not receive associated grammatical words in their targeted language.
c) The type of vocabulary they are being used for (field of technology, physics, biology, medicine).
d) The presence of a linking vowel between constituents, as well as its position. For example, in French, the linking vowel *–o-* will appear when there is a constituent of Greek origin, while *–i-* will appear when there is at least one constituent of Latin origin. However, *-o-* is far more common.

Although this criterion is highly successful in describing analogous elements of neoclassical compounds, there are still some examples that would be classified as heterogenous. For example, the position of combining forms can vary from the initial, middle and final position. They came to the realization that not all neoclassical combining forms can be analyzed in the same way. Also, the usage of lexeme-based morphology proved to be a sufficient method of analyzing neoclassical combining forms. On the other hand, they also found that based on the current knowledge, the distinction between allomorphy and suppletion cannot be made, at least in this kind of analysis. Besides that, there is a problem with differing neoclassical compounding and native compounding in French and other languages of the Romance group. They found that not all affixes are problematic, for example, *micro-* and *–logue* are well grammaticalized. The problem occurs with affixes like *–phage, –cide* and *–vore* because some of them have grammatical features while others have lexical features. Those with lexical features usually have a specific meaning and they usually cannot be

merged with many other words, while those with grammatical features, such as *micro-*, can serve as a compounding pattern base for many different compounds because it has gone through a process of grammaticalization. These grammaticalized items become highly productive, which is why they are seen as affixes.

Another research dealing with the productivity of affixes was covered by Mititelu (2018). Using the Princeton WordNet, she took pairs of base-derived words and studied the affixes out of which the derived words were made. She studied the affixes based on different derivation types. Her hypothesis was that the derived word should contain both the meaning of the base word and the affix. If the new word does not have that compositional meaning, it probably does have an idiomatic meaning. She does not consider these idiomatic words as derived pairs but presumes that the idiomatic meaning in the derived word comes from a semantic change. Mititelu's sample consisted of 40,318 pairs (base words and derived words) within all parts of speech. She found that the most frequent affixes are *–ly, -less, -ness, -y, -al, -ic, -ity* and *–able.* These results match former research conducted by Hay and Baayen. They also proved their hypothesis that affixes can form different meanings with no relation to the base word. According to this paper, this means that affixes have the power to make a semantic shift.

In addition, Diaz-Negrillo (2014) researched the morphological behavior of neoclassical compounds and morphological development in this compound class. She tried to determine whether the morphological behavior of neoclassical compounds differs from that of earlier compound classes. She collected data and studied some of the most important features of neoclassical compounds- their internal configuration, productivity and presence or absence of a linking vowel within them. She used both synchronic analysis for their morphological behavior and diachronic for any proof of how these formations have developed or to find some possible hints about their morphological tendencies. She used quantitative analysis in order to understand the behavior of neoclassical compounds and she did that on a sample of 425 neoclassical compounds. She found the compounds in the *British National Corpus* (BNC) and classified them according to 10 combining forms. The results showed that recently coined neoclassical compounds mostly have bound initial bases. Compounds that have free bases date back to the 19th century. She also found that most compounds have a linking vowel and those that do not, usually have bases that end with vowels already. This can be explained by the fact that there is a tendency to avoid two consonants occurring at the borderline. There is one exception when it comes to linking vowels, it does not occur when there is *–mania.*

Concerning the productivity of final combining forms, she proved that not all forms are equally productive and that their productivity changes depending on time. Today, the most productive are *–mania, -phile* and *–phobia,* while the least productive are *–crat, -logy* and *–cide.* The most frequent ones form at least 50% of all neoclassical compounds today. Suffixes *–crat, -morph* and *–lith* have the most regular morphological behavior, although they are the least productive. In addition, the most common ones show very specific behavior. She concluded that older neoclassical compounds are more basic in their behavior, while newer formations are not. Furthermore, neoclassical compounds usually do have a linking vowel. She found that *–mania* and *–phobia* are the most unusual elements.

4. Research on productivity

My research deals with the morphological productivity of neoclassical compounds. In the second chapter, I mentioned the means of their formation as well as the key elements they consist of. Furthermore, I described how to understand their meaning by analyzing the formation of the compound itself. I also explained the common confusion between affixation and compounding and how to distinguish one from the other.

As for the previously done pieces of research I described in the third chapter, they can be used to better understand my own. Pustylnikov and Schneider-Wiejowski (2009) conducted a research that can relate to my own because part of it deals with the productivity of noun suffixes in German. Amiot and Dal (2005) did a research that is a bit more similar regarding the topic because it deals with the productivity of neoclassical combining forms but in French language. Mititelu (2018) can relate to my research by examining the productivity of affixes in pairs of base-derived words. Diaz-Negrillo (2014) conducted a research on the morphological behavior of neoclassical compounds, as well as some of their most important features, including their productivity.

I used the corpus British Web 2007 (ukWaC), which is a large English corpus consisting of more than 2 billion tokens. It is one of the largest resources provided when doing any type of linguistic research. I researched the productivity of neoclassical compounds based on their frequency. I used this corpus to find compounds containing either initial or final combining forms, which is also a neoclassical element. In other words, my research was based on compounds that either start or end with combining forms derived from Greek or Latin. The list of initial combining forms is: *acro-* (beginning, end, tip, top, peak, height), *bio-* (life, living organisms), *crypto-* (covered, hidden), *electro-* (electrical, electricity), *geo-* (earth, ground, soil land, country), *hydro-* (water, liquid), *neuro-* (nerve, nervous system), *poly-* (many, several, much), *syn-* (together with, at the same time as, along with) and *theo-* (god, gods, God). The list of final combining forms is: *-cide* (killer, slayer), *-cyte* (cell), *-ectomy* (surgical removal), *-gram* (writing, record), *-graphy* (writing or representation in a specified manner or by specified means), *-itis* (inflammation, disease), *-logy* (oral or written expression, theory, doctrine, science), *-meter* (instrument, means of measuring something), *-philia* (friendly feeling, liking for something, tendency toward something), *-scope* (means for viewing or observing something).

4.1. Research questions

This paper and research are based on these questions.

1. Research question: What is the most/least productive initial combining form in the sample of this paper?
2. Research question: What is the most/least productive final combining form in the sample of this paper?
3. Research question: Are initial combining forms more productive than final combining forms in the sample of this paper?
4. Research question: Is the measure of productivity in both the initial and final combining forms linear in the sample of this paper?

4.2. Methodology and research sample

When using the previously mentioned corpus in Sketch Engine I first selected the option 'Concordance', which is a feature providing the user with a list of all examples of the searched word found in a corpus. I then selected the feature CQL which enables the user to search for some more complex lexical patterns or to use specific search criteria. When searching for compounds that started with acro-, I used the following query: [lemma="acro.*" & tag="N."]. This query helped me in the search for words that started with *acro-* and came in the form of a noun. When searching for compounds that started with *bio-*, *crypto-* or any other initial combining form I used the same query and instead of *acro-* I wrote any other element I searched for. When looking up final combining forms I used a slightly different query: [lemma=".*ectomy" & tag="N."]. The same analogy is applied when looking up words that contain other final combining forms.

Once I had a list of all the words I needed, I downloaded the given lists to Excel where I checked if all the words fit correctly to the query. The list also contained the frequency of each word, which I would later use. Each word was checked in a dictionary (OED), as well as its etymology. Whilst checking the words I noticed a few false positives, which are some examples I cannot use for research purposes mostly because they are either not compounds or in some other way do not fit the description. In order to calculate the productivity of the previously mentioned combining forms I used 20 words from each list. All in all, I analyzed the data of 10 initial and 10 final combining forms and each of these lists contained 20 words.

Once all the necessary data was collected, I used that information to calculate the productivity. Each list contains 20 words that represent types. In other words, type is a number of different words which appear in a text. Frequencies of each word added together equal the total number of tokens, or in other words, the total number of all words that appear in a text. Therefore, the productivity ratio would equal a total number of types (X) divided by a total number of tokens (Y):

$$P = X{:}Y$$

This was described by Baayen (1992) in his work on the quantitative aspects of morphological productivity. He concluded that the usual equation for measuring productivity which is a number of hapaxes (words appearing only once) divided by number of tokens, can be replaced by another one in the right circumstances. If the research sample is not very big the query used to analyze the measure of productivity is derived from number of types divided by number of tokens. He concluded that this query is closely connected to the frequency levels. The final results may vary from those that refer to the bigger sample, but the ratios of the required values are complementary. He concluded that this equation satisfies conditions for quantitative measure of productivity.

Neoclassical compounds I used came in the form of singular nouns. The sample contains 10 combining forms which take the initial position and 10 which take the final position of a word. When I collected my research sample of all the initial and final combining forms, I noticed a great difference in the length of each list. I decided each list will contain the first 20 words which fit the necessary description. In order to do that I had to remove all of the false positives and filtrate my research sample. I wanted to create a research sample in which each compound came in the form of a singular noun. Some of the words I removed were not compounds e.g. *neuro,* other words were of different word classes e.g. *electromagnetic* or they were in the plural form e.g. *biosciences* instead of *bioscience*. Some of the words on the lists were spelled the wrong way and were clearly typos e.g. *sucide* instead of *suicide*. I also removed all of the words which were abbreviations e.g. *bio-tech* instead of *bio-technology*.

5. Results

5.1. Productivity of initial combining forms

In order to calculate the productivity I chose 10 initial combining forms, meaning that they appear at the beginning of a word. Using corpus I extracted 20 lexemes which are compounds in forms of singular nouns. When making the following table I listed all of the initial combining forms and provided the examples for which one. In the table, I also listed the frequency levels for each compound. It is the number of appearances in the corpus I used. When calculating the productivity ratio for each combining form the list of examples I provided represents the number of types, whereas the frequency levels added together represent the number of tokens or all occurrences.

Table 1. List of initial combining forms

acro-		**bio-**		**crypto-**		**electro-**		**geo-**		**hydro-**		**neuro-**		**poly-**		**syn-**		**theo-**	
Lemma	Freq.	Lemma	Freq.	Lemma	Freq	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.
acronym	2587	biology	21885	cryptography	1906	electron	11453	geography	21614	hydrogen	15122	neuroscience	1916	polymer	5526	syndrome	18654	theory	123393
acrobat	1374	biodiversity	16697	cryptogram	374	electronics	3991	geometry	11366	hydrocarbon	1787	neurology	1684	polyester	2677	synthesis	13885	theology	18540
acropolis	231	biography	16225	cryptosporidiosis	186	electrode	2754	geology	8203	hydroxide	1344	neuron	1456	polystyrene	1883	syntax	12663	theorem	4033
acromegaly	217	biotechnology	10384	cryptosporidium	134	electronica	1392	geologist	1239	hydrology	1158	neuropathy	1249	polythene	1514	synopsis	7919	theologian	2363
acrostic	204	biomass	5171	cryptology	130	electrolyte	1327	geomorphology	838	hydrotherapy	1120	neurologist	1228	polyethylene	1429	synagogue	2687	theorist	1698
acrobatics	112	biopsy	3515	cryptographer	70	electrolysis	832	geographer	656	hydrolysis	812	neurotransmitter	815	polypropylene	1370	synergy	2624	theosophy	387
acromion	68	biochemistry	2910	cryptosystem	60	electrophoresis	717	geochemistry	489	hydrochloride	643	neuroblastoma	678	polymorphism	1367	syndicate	2301	theocracy	378
acrolein	19	biologist	2027	cryptozoology	46	electromagnetism	426	geoscience	461	hydropower	373	neurosis	602	polygon	1366	synchronisation	1938	theodolite	283
acrophobia	19	biographer	1912	cryptologist	31	electrocardiogram	286	geophysics	330	hydrocephalus	346	neurosurgery	481	polyurethane	1349	synchronization	1768	theoretician	266
acrosome	15	biodiesel	1713	cryptorchidism	20	electrocution	281	geomancy	135	hydrocortisone	324	neuropsychology	464	polymerase	1228	synod	1493	theophylline	246
acropora	8	bioscience	1047	cryptosporidia	15	electrophysiology	272	geophysicist	125	hydrofoil	214	neurobiology	396	polynomial	861	synonym	1418	theobromine	230
acrocephalus	6	biosphere	1032	cryptozoologist	11	electrochemistry	228	geopolitics	100	hydrogenation	209	neurosurgeon	384	polypeptide	683	synthesizer	1137	theodicy	203
acroter	5	biosecurity	870	cryptococcosis	10	electronegativity	184	geodesy	90	hydrogel	204	neurophysiology	335	polyphony	539	synchrotron	1089	theorisation	143
acrophony	3	biosynthesis	718	crypto-fascist	8	electrometer	172	geometer	80	hydrogeology	181	neuroimaging	328	polymerisation	519	syndication	835	theorbo	118
acroterion	3	bioavailability	664	cryptorchid	8	electrospray	165	geodiversity	79	hydrophone	172	neuropathology	308	polysaccharide	416	synthesiser	809	theosophist	93
acrodermatitis	3	bioinformatics	623	cryptomnesia	7	electromagnet	150	geochronology	75	hydroponics	150	neurotoxicity	198	polygamy	416	synthase	731	theophany	78
acrocyanosis	3	biofeedback	550	cryptoanalysis	7	electroencephalogram	138	geoarchaeology	63	hydrography	133	neuroscientist	188	polytunnel	408	synapse	530	theorization	69
acromioplasty	3	biofuel	476	cryptograph	7	electroscope	123	geomagnetism	59	hydrophobicity	102	neurodegeneration	171	polycarbonate	390	synchronicity	472	theosis	27
acrolect	2	biosafety	468	cryptomeria	7	electrotherapy	109	geolocation	56	hydrometer	94	neuroanatomy	143	polyhedra	298	syncretism	458	theoria	24
acrophobe	2	bio-diversity	435	cryptococcus	6	electromyography	108	geomorphologist	55	hydroelectricity	93	neurogenesis	137	polytheism	283	synthetic	431	theogony	17

Table 2. Productivity of initial combining forms

COMBINING FORM	NUMBER OF TOKENS	PRODUCTIVITY
crypto-	3043	0.006572
acro-	4884	0.004095
neuro-	13161	0.00152
poly-	24522	0.000816
hydro-	24581	0.000814
electro-	25108	0.000797
geo-	46113	0.000434
syn-	73842	0.000271
bio-	89322	0.000224
theo-	152589	0.000131

Once I analyzed each initial combining form and calculated productivity. I created a table sorted by productivity (from highest to lowest) based on the sample i researched. Since the first table contains 20 words, the final results may vary from those which contain a bigger word list.

When comparing *crypto-* which has the highest productivity ratio, with *theo-*, which has the lowest, I can conclude that combining forms that have highest productivity ratios mostly form words in the science field (medicine, biology, chemistry). Also, initial combining forms with a lower number of tokens have higher productivity ratios than those with a higher number of tokens.

Also, the combining forms *crypto-* and *acro-* which have the highest productivity ratios form compounds which are used in various fields of science, as well as everyday life. They are not restricted in their usage. On the other hand, this analogy can't be applied to those initial combining forms with the lowest productivity ratios, such as *bio-* and *theo-*.

Regarding the initial combining form *acro-* the most frequent compounds are *acronym* and *acrobat*. We can notice a significant fall in frequency levels, after which follow the compounds *acropolis* and *acromegaly*. All of the other compounds listed have very low-frequency levels, the last one being *acrophobe*, with only 2 appearances in the corpus. Also, the initial combining form *acro-* has the second highest productivity ratio 0.004095.

The most frequent compounds starting with the combining form *bio-* are *biology*, *biodiversity* and *biography*. All of the words listed in the table above have a high number of

occurrences. Because of that initial combining form *bio-* has the second biggest number of tokens and very small value of productivity ratio, falling at the penultimate place. Its value of productivity is 0.000224.

Compound *cryptography* has a frequency value of 1906, after which follows *cryptogram* with 374 numbers of occurrences in the corpus. The initial combining form *crypto-* has the smaller number of tokens and the highest value of productivity ratio. It is the most productive initial combining form whose value in productivity is 0.006572.

As for the initial combining form *electro-*, the most frequent word appearing in the corpus is *electron*, followed by *electronics*. Although listed words have a gradual fall in their frequency levels, *electro-* falls somewhere in the middle regarding the total number of tokens and productivity values, which is 0.000797.

The most frequent compound with the initial combining form *geo-* is *geography*, followed by *geometry* and *geology*. The fall in the frequency levels appears to be gradual. Initial combining form *geo-* has a high total number of tokens and low value of productivity. It falls in seventh place and its productivity value is 0.000434.

Compounds that start with the initial combining form *hydro-* are all used in chemistry and their frequency levels have small values in differences, e.g. *hydrocarbon*, *hydroxide* and *hydrology*. All of the compounds listed are used in science. The initial combining form *hydro-* falls in the fifth place with values of productivity of 0.000814.

The initial combining form *neuro-* has the third highest value of productivity, 0.00152. Similar to the initial combining form *hydro-*, *neuro-* forms compounds which are used in the field of science. The compound having the highest frequency level is *neuroscience*, followed by *neurology*.

Compounds starting with the initial combining form *poly-* which have the highest number of occurrences are *polymer*, *polyester* and *polystyrene*. The initial combining form *poly-* falls in fourth place regarding its value of productivity, which is 0.000816.

The initial combining form *syn-* has the third largest number value in a total number of tokens, but it is not as productive because the productivity value it has is 0.000271, situating it in the eighth place. The compound with the most occurrences is *syndrome*, followed by *synthesis*.

The initial combining form *theo-* has the biggest total number of tokens. It is also the least productive initial combining form with productivity value being 0.000131. Compounds such as *theory*, *theology* and *theorem* have the biggest number of occurrences.

5.2. Productivity of final combining forms

In order to calculate the productivity of the final combining forms I used the same methodology. I chose 10 final combining forms, meaning that they appear at the end of a word. When calculating the productivity ratio for each combining form the list of examples I provided represents the number of types, whereas the frequency levels added together represent the number of tokens or all occurrences.

Table 3. List of final combining forms

-cide		-cyte		-ectomy		-gram		-graphy		-itis		-logy		-meter		-philia		-scope	
Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.	Lemma	Freq.
suicide	23125	lymphocyte	878	hysterectomy	1643	program	143047	photography	29347	arthritis	14076	technology	227347	diameter	22462	haemophilia	1153	telescope	9698
genocide	5214	leukocyte	327	prostatectomy	585	diagram	16579	geography	21614	hepatitis	8207	methodology	26413	parameter	19660	paedophilia	404	microscope	9301
pesticide	4961	oocyte	264	vasectomy	455	histogram	2044	biography	16225	meningitis	4009	psychology	22335	perimeter	5279	necrophilia	115	kaleidoscope	1051
herbicide	2449	erythrocyte	262	mastectomy	425	telegram	1925	bibliography	14323	osteoarthritis	2728	biology	21885	thermometer	1950	eosinophilia	107	horoscope	926
homicide	1950	leucocyte	191	cholecystectomy	264	kilogram	944	autobiography	6183	dermatitis	2517	archaeology	19363	spectrometer	1429	hemophilia	82	stethoscope	486
insecticide	1484	granulocyte	115	splenectomy	194	hexagram	899	choreography	3952	bronchitis	1485	sociology	18883	barometer	1403	thrombophilia	65	oscilloscope	417
fungicide	999	chondrocyte	106	tonsillectomy	194	hologram	746	topography	3765	encephalitis	1223	theology	18540	altimeter	657	pedophilia	46	periscope	373
infanticide	460	monocyte	100	cystectomy	143	datagram	653	pornography	2874	colitis	1144	terminology	11920	speedometer	596	arachnophilia	42	endoscope	327
regicide	265	hepatocyte	92	nephrectomy	134	anagram	647	historiography	2389	pancreatitis	899	ideology	11588	micrometer	568	neutrophilia	33	gyroscope	222
fratricide	165	T-lymphocyte	84	gastrectomy	132	monogram	513	ethnography	2352	rhinitis	887	ecology	10530	kilometer	549	biophilia	22	ophthalmoscope	174
biocide	165	oligodendrocyte	74	orchidectomy	116	subprogram	497	cryptography	1906	mastitis	800	biotechnology	10384	interferometer	508	technophilia	22	electroscope	123
microbicide	133	keratinocyte	73	thyroidectomy	112	angiogram	402	iconography	1644	poliomyelitis	750	analogy	8333	magnetometer	503	cinephilia	19	laryngoscope	109
rodenticide	127	melanocyte	65	laryngectomy	99	cryptogram	374	typography	1522	conjunctivitis	672	geology	8203	pedometer	346	europhilia	15	stereoscope	106
deicide	93	phagocyte	46	lymphadenectomy	96	microgram	360	discography	1500	cystitis	615	astrology	7008	calorimeter	313	francophilia	13	spectroscope	88
nematicide	64	myocyte	37	lobectomy	95	Polygram	317	crystallography	1445	sinusitis	605	physiology	7008	millimeter	310	scopophilia	12	colonoscope	68
parricide	63	reticulocyte	35	endarterectomy	92	electrocardiogram	286	cinematography	1416	vasculitis	535	apology	6479	chronometer	308	coprophilia	12	arthroscope	60
biopesticide	60	adipocyte	33	oophorectomy	88	epigram	281	tomography	1389	otitis	493	anthropology	6401	anemometer	271	anglophilia	10	laparoscope	50
matricide	52	thymocyte	31	appendectomy	83	mammogram	257	calligraphy	1363	spondylitis	489	pathology	6199	ergometer	243	xenophilia	9	bronchoscope	47
acaricide	43	astrocyte	29	colectomy	82	pictogram	232	cartography	1256	appendicitis	471	mythology	5890	odometer	241	gypsophilia	8	otoscope	45
bactericide	42	B-lymphocyte	16	appendicectomy	81	pentagram	229	radiography	1189	gastroenteritis	436	anthology	5634	potentiometer	235	zoophilia	6	stereomicroscope	43

Table 4. Productivity of final combining forms

COMBINING FORM	NUMBER OF TOKENS	PRODUCTIVITY
-philia	2195	0.009112
-cyte	2858	0.006998
-ectomy	5113	0.003912
-scope	23714	0.000843
-cide	41914	0.000477
-itis	43041	0.000465
-meter	57834	0.000346
-graphy	117654	0.00017
-gram	171232	0.000117
-logy	460343	0.000043

The productivity ratios of this table show that final combining forms which are mostly used in coining words related to medicine have bigger values. On the other hand, forms which are on the bottom half of the table are not used only in the field of medicine but science in general.

When comparing -*philia* which has the highest productivity ratio, with -*logy*, which has the lowest productivity ratio, I can conclude that the final combining forms with a lower number of tokens have higher productivity ratios than those with a higher number of tokens.

Also, the final combining forms I chose to conduct my research do not have differences regarding their usage. Most of these compounds are used in medicine, math, physics and different fields of science.

Regarding the final combining form -*cide* the most frequent compounds are *suicide* and *genocide*. We can notice a gradual fall in frequency levels. The final combining form -*cide* falls in fifth place regarding the value of productivity, which is 0.000477.

As for the final combining form -*cyte*, the most frequent word appearing in the corpus is *lymphocyte*, followed by *leukocyte*. The listed words have a gradual fall in their frequency levels and are all used in the field of medicine. The final combining form -*cyte* has the second highest value in productivity, which is 0.006998. Because of that, we can conclude that it is highly productive.

Compounds formed by the final combining form -*ectomy* are also used in medicine for procedures that involve the removal of a certain body part. The compound with the highest frequency level is *hysterectomy*. This combining form has the third highest value of productivity, which is 0.003912.

The most frequent compounds ending with the combining form -*gram* are *program* and *diagram*. All of the words listed in the table above have a high number of occurrences. Because of that final combining form –*gram* has the second biggest number of tokens. On the other hand, it falls in the ninth place regarding its value of productivity, which is 0.000117. This means that it's not highly productive.

Compounds ending in the final combining form -*graphy* also have high frequency levels. The most frequent compound is *photography*, followed by *geography*. This final combining form is also not very productive. Its value of productivity is 0.00017, situating it in eighth place.

The most frequent compound formed with the final combining form -*itis* is *arthritis*, followed by *hepatitis* and *meningitis*. The fall in the levels of frequency is gradual. Its value in productivity is 0.000465, situating it in the sixth place.

The final combining form -*logy* has the third highest levels of frequency. The most frequent word is *technology* with a total number of 227347 occurrences. It is followed by *methodology* and *psychology*. The final combining form -*logy* has the highest total number of tokens, but it is the least productive combining form with its value being 0.000043.

Compounds *diameter*, *parameter* and *perimeter* have the highest frequency levels regarding the final combining form -*meter*. This final combining form has the value of productivity 0.000346, situating it on the seventh place.

Compounds formed by the final combining form -*philia* generally have low frequency levels, the highest being *haemophilia*, followed by *paedophilia*. This final combining form has the lowest number of tokens, but it is the most productive. Its productivity value is 0.009112.

Compounds with the highest frequency levels, formed by the final combining form -*scope* are *telescope* and *microscope*, after which we can notice drastic fall in the frequency level, followed by *kaleidoscope*. This combining form has a productivity value of 0.000843 and it falls in fourth place.

5.3. Final results

In this final table where the productivity ratios of initial and final combining forms are compared, we can conclude that the initial combining forms have higher productivity ratios than the final ones. The initial combining form *crypto-* is less productive than the final combining form *-philia*. The first three examples of the productivity of final combining forms have higher values than initial combining forms. However, following the rest of the list it is obvious that the initial combining forms have higher values. While observing this table we can also notice that the fall in the productivity rations of the initial combining forms happens gradually, while the one of the final combining forms is more drastic, with bigger value differences between the first and the last form. Also there is a significant difference in the productivity, when comparing the least productive initial and final combining form. We can conclude that the initial combining form *theo-* is more productive that the final combining form *-logy*.

Table 5. Comparison of productivity values

Initial combining forms	**Productivity ratios**	**Final combining forms**	**Productivity ratios**
crypto-	0.006572	-philia	0.009112
acro-	0.004095	-cyte	0.006998
neuro-	0.00152	-ectomy	0.003912
poly-	0.000816	-scope	0.000843
hydro-	0.000814	-cide	0.000477
electro-	0.000797	-itis	0.000465
geo-	0.000434	-meter	0.000346
syn-	0.000271	-graphy	0.00017
bio-	0.000224	-gram	0.000117
theo-	0.000131	-logy	0.000043

6. Conclusion

To summarize, the first chapter of the paper explains what are compounds and how are they made. It states the differences in orthography and what types of compounds exist in a language. Furthermore, it explains what are neoclassical compounds and how are they formed. Later on, the difference between affixation and combining forms is explained, because those two terms can be easily mixed up together. Because of this, I decided to conduct my research using neoclassical compounds. I split my research into two parts: based on initial or final combining forms which are used in the formation of neoclassical compounds. Those combining forms can sometimes be misinterpreted as suffixes or prefixes. In the following chapter, I mentioned some of the previously conducted corpus-based research which dealt with similar topics. Some of those even researched neoclassical compounds and their elements. After I analyzed pieces of research I conducted my own. I researched the productivity of neoclassical compounds which I put into two categories, based on their combining form and their position. In order to find the words I was looking for, I had to come up with the correct query, which would then list every word from the corpus I used. After I listed the research sample and explained the meaning of each initial and final combining form, I listed all of the collected data to calculate the productivity ratios. I had to filtrate my research sample to make sure every compound on the list fit all of the criteria. The next step was finding the correct query for measuring productivity, using a number of tokens and types. Once I calculated all of the productivity ratios I had to make a table presenting the results. Then I listed all of the forms in a table, starting with those having the highest productivity ratio to the lowest.

When presenting the results I first explained the relation between the total number of tokens and the value of productivity. I concluded that those two are inversely proportional. I also mentioned what are some of the compounds with the highest number of occurrences. Then I analyzed the values of productivity for each combining form. In the end, while comparing the initial and the final combining forms I concluded that the initial combining forms all together have higher productivity ratios than the final combining forms.

7. References

Amiot, D., & Dal, G. (2005). *Integrating neoclassical combining forms into a lexeme-based morphology*. *Morphology*, *15*(1), 3-17.

Barbu Mititelu, V. (2018). *Investigating English affixes and their productivity with Princeton WordNet*. In F. Bond, P. Vossen, C. Fellbaum (Eds.), *Proceedings of the 9th Global Wordnet Conference* (pp. 53-60). Global Wordnet Association.

Bauer, L. (1998). *Is there a class of neoclassical compounds, and if so is it productive? Linguistics, 36*(3), 403-417.

Bauer, L. (2017). Compounds and words. In *Compounds and compounding (pp.3-27)*. Cambridge University Press.

Baayen, R. H. (1992). *Quantitative aspects of morphological productivity*. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Kluwer Academic Publishers.

Bisetto, A., & Scalise, S. (2005). *The classification of compounds. Lingue e Linguaggio, 4*(2), 319–332.

Carstairs-McCarthy, A. (2002). *An introduction to English morphology: Words and their structure*. Edinburgh University Press.

Díaz-Negrillo, A. (2014*). Neoclassical compounds and final combining forms in English.* Linguistik Online, *68*(6).

Fabb, N. (1998). *Compounding*. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology* (pp. 66–84). Blackwell Publishers Ltd.

Stockwell, R., & Minkova, D. (2001). *Semantic change and semantic guesswork*. In *English words: History and structure* (pp. 147-160). Cambridge University Press.

Plag, I. (2003). *Word-formation in English*. Cambridge University Press.

Pustylnikov, O., & Schneider-Wiejowski, K. (2009). *Measuring morphological productivity*. In R. Köhler (Ed.), *Issues in Quantitative Linguistics* (pp. 1-9).

Productivity of Neoclassical Compounds

Abstract

This paper provides a corpus-based research on the productivity of neoclassical compounds. The research was divided into two categories. The paper investigated the productivity of neoclassical compounds with either initial or final combining forms. After getting the results I concluded which initial combining form is the most productive and which is the least productive. The same was done for the final combining forms. This research gave me an insight into the difference in the productivity values between initial and final combining forms. Also, some of the previously done corpus-based research were mentioned because they can help better understand my own.

Key words: neoclassical compounds, compounding, productivity, corpus-based research, initial combining form, final combining form

Produktivnost neoklasičnih složenica

Sažetak

Ovaj rad pruža korpusno istraživanje o produktivnosti neoklasičnih složenica u engleskom jeziku. Istraživanje je podijeljeno u dvije kategorije, pri čemu je istraživana produktivnost neoklasičnih složenica s početnim tvorbenim oblicima i sa završnim tvorbenim oblicima. Dobivenim rezultatima može se zaključiti koji je početni oblik najproduktivniji, a koji najmanje. Isto vrijedi i za završne tvorbene oblike. Ovo istraživanje pruža uvid u razlike unutar vrijednosti produktivnosti između početnih i završnih oblika. Također, spomenuta su i neka prijašnje provedena korpusna istraživanja, kako bi se ovo lakše razumjelo.

Ključne riječi: neoklasične složenice, tvorba složenica, produktivnost, korpusno istraživanje, početni tvorbeni oblik, završni tvorbeni oblik