The morphological productivity of selected combining forms in English

Pro gradu thesis

Department of modern languages
University of Helsinki
26 September 2013
Eeva Rita-Kasari

Contents

1. Introduction

Morphological productivity, the ability of speakers to coin new words using the resources of a given language, has received a great deal of attention in the past few decades, especially in the context of the English language. Bauer (2001) and Plag (1999), for example, offer book-length accounts of this phenomenon. According to Bauer, "productivity remains one of the most contested areas in the study of word-formation" (1983: 62). Even though Bauer's book was published almost thirty years ago and many new theories on morphological productivity have been proposed since then, this statement still seems to be valid. Even though the concept of productivity seems to be easy to grasp intuitively, there are several practical and theoretical issues that have not yet been tackled.

Baayen and his collaborators in particular have developed a number of empirical corpus-based methods to gauge productivity quantitatively (see, e.g., Baayen 1992, 1993, 2009; Baayen and Lieber 1991; Chitashvili and Baayen 1993; Baayen and Renouf 1996). In addition, the psycholinguistic aspect of productivity, i.e., the way speakers store complex words in their mental lexicon, has received considerable attention since the early nineties. Important contributions have come from Frauenfelder and Schreuder (1992), Baayen (1993), and Hay and Baayen (2002).

In addition to being an important theoretical concept, studying morphological productivity also has practical relevance. Measuring degrees of productivity is important from the point of view of writing grammars, for example. While unproductive processes are listed in the lexicon, productive patterns can be described by rules, and thus measuring productivity offers valuable information on what processes should be given prominence in the case of limited resources (Lüdeling et al. 2000: 57). The importance of the concept of productivity can also be seen in the fact that computational tools cannot function properly unless they take productive word-formation into account (Baayen 2008: 900). In addition, knowledge of the productivity of different word-formation processes may also be relevant in language teaching. Since combining forms are particularly frequent in scientific registers, this study might also have some

relevance in the context of English for special and academic purposes (see Fradin 2000: 22–23).

Studies concentrating on the productivity of English affixes abound. Most of them have focused on suffixes (see, e.g., Baayen and Lieber 1991; Baayen and Renouf 1996; Bauer 1992; Plag et al. 1999; Plag 2006, Palmer 2009, Säily and Suomela 2009, Säily 2011). This might be due to several reasons, one of which being their class-changing nature. Prefixation, on the other hand, has not received the same amount of attention. Plag writes that in spite of the fact that English is one of the best-described language in the world, "the exact properties of many affixes are still not sufficiently well determined" (2003: 86), and that there is still a great need for more detailed investigation.

Neoclassical compounding and what the *Oxford English Dictionary* calls combining forms have been studied even less, and the few exceptions (e.g., Warren 1990, Prćić 2005, Prćić 2007, Prćić 2008, McCauley 2006, and Kastovsky 2009) have mainly been rather theoretical, and nothing really explicit has been said about their productivity. The productivity of native N + N compounds has been studied by Férnandez-Domínguez (2009), but neoclassical compounding differs from them in many ways. This gap in research might partly be due to the complex status of combining forms as a part of the English word-formation system (see, e.g., Bauer 1998): they seem to be situated somewhere between affixation and compounding (see section 2.2 for discussion).

To make the issue even more complicated, even the *OED* seems to be rather inconsistent in the way it classifies certain elements as either combining forms or traditional affixes. This inconsistency seems to concern initial combining forms in particular. For example, the *OED* defines *hyper-* and *-ultra* as prefixes and *quasi-* and *pseudo-* as combining forms, even though there seems to be no apparent structural difference between these two groups of elements. It has also been argued that the category of combining forms in general is superfluous and that other, less marginal types of word-formation could well be used to describe this word-formation process (Kastovsky 2009).

In this paper, I will measure the morphological productivity of selected English initial combining forms (ICFs) and final combining forms (FCFs), three of each, and see whether there are any differences between these two rather closely related categories. The focus is on the so-called classical

combining forms, as opposed to modern combining forms (see Prćić 2005, 2007, 2008). The initial combining forms studied are *hyper-*, *quasi-*, and *pseudo-*, and the final combining forms are *-logy*, *-graphy*, and *-nomy*. The study applies corpus methods, and the data comes from the written part of the *British National Corpus* (henceforth the *BNC*). My hypothesis is that the initial combining forms will prove more productive and that they will combine more easily with native bases, one of the reasons being that prefixes in general are often more lexeme-like than suffixes, the latter often being used to convey grammatical information.

The main goal of this study is to clarify the status of combining forms in the English word-formation system and to provide information on their productivity. As for methodology, I will mainly rely on previous approaches that have been widely tested in several studies. I will apply several indicators to assess the morphological productivity of the selected combining forms. The first one is the measure called productivity in the strict sense *P* (also called potential productivity), developed by Baayen and his collaborators (see Baayen and Lieber 1991 in particular). In addition, two measures from lexical statistics, vocabulary growth curves and frequency spectra, are employed. An open source statistics software package, R, is used to obtain information on the selected combining forms.

Section two provides the reader with the theoretical background relevant for this study, defining and discussing the basic concepts related to English word-formation. The category of combining forms is introduced, as well as that of neoclassical compounds. The status of combining forms as a part of the English word-formation system is of great relevance here, and it is therefore thoroughly discussed. A brief history of diachronic approaches to the study of affixation is also introduced. The notion of morphological productivity and the theories attempting to describe it are also discussed, as well as the various constraints that affect the application of various word-formation processes. In addition, the basic theoretical background for lexical statistics and LNRE modelling are introduced. Psycholinguistic approaches to productivity are also discussed, since they have proved to capture an important aspect of the nature of productivity. Furthermore, many quantitative measures of productivity are also psycholinguistically motivated.

Section three discusses the method and the data chosen for the present study, as well as potential problems that might turn up in the research process. Important questions include the reliability of corpus evidence, as well as the principles on which the units of analysis are determined. Section four presents the results, which are then analysed and discussed in the fifth section. Section six provides a brief summary of the most important findings of the thesis and suggests some ideas for possible future studies.

2. Background

2.1. Word-formation

2.1.1. Basic concepts of word-formation

The concept of *word* is, in spite of its apparent simplicity, rather problematic in linguistics. The notions of *lexeme* and *word-form* have proved much more useful in the study of morphology. Haspelmath defines lexemes, or dictionary words, as abstract entities that consist of several word-forms. Word-forms, or text words, on the other hand, are the concrete realizations of a lexeme that can be pronounced and used in texts (2002: 13). Plag uses the terms *inflection* and *derivation* to distinguish between word-forms and lexemes: according to him, word-forms are created with inflectional suffixes, while new lexemes are produced by derivational affixes (2003: 14; see also Kastovsky 2001: 218). Since lexemes consist of several word-forms, it is necessary to have a specific *citation form,* under which the lexical entry is listed in dictionaries (Haspelmath 2002: 14).

From the point of view of morphological productivity, another central distinction in morphology, the one between *possible* (or potential) *words* and *actual words*, is also relevant. Plag defines possible words as words whose structure follows the rules of the language. Actual words, on the other hand, are in use in the language, although it is not clear how "in use" should be defined (whether it be in the vocabulary of an individual speaker, or in a dictionary) (Plag 1999: 7). Plag writes that the usefulness of the possible-actual dichotomy has been

widely criticized, but he considers it a useful notion nevertheless: according to him, studying large numbers of actual words is necessary in order to find out about the mechanisms that are related to the properties of possible words (1999: 8–9).

The notion of *mental lexicon* is also relevant in the possible/actual dichotomy. Anshen and Aronoff define the speaker's mental lexicon as "the list of irregular items that the speaker/hearer carries around in his or her head" (1998: 238). Words that are stored in the speaker's mental lexicon are *existing words*, while *potential words* are words that meet all the criteria for being a genuine word in a language but that are not listed in the mental lexicon of that person (though they may exist for another speaker) (Anshen and Aronoff 1998: 238). According to Haspelmath, most morphologists agree that all simple and at least some complex words are listed in speakers' mental lexicons, but it is difficult to state which complex words are listed and which are not, because most word-formation processes cannot be described as either unequivocally productive or unproductive (2002: 41).

Since this study applies corpus linguistic methods, another distinction concerning the nature of words is worth mentioning, namely that of *types* and *tokens.* In a sense, types represent different word-forms and tokens different orthographic units or strings of characters in a text (see, e.g., Carstairs-McCarthy 2002: 5–6). In other words, types are distinct word-forms, and tokens all the running words in a text (Pápai 2004: 157). Thus, for example *pseudo-science* and *pseudo-sciences* are considered as separate types, even though they are strictly speaking manifestations of a single lexeme. Baayen and Lieber define tokens as all the instances of a certain type (1991: 803). Both type and token frequencies are always determined in relation to a particular corpus (Bauer 2001: 47). A common notation to describe the number of tokens, i.e., *corpus size*, is *N*, while the *vocabulary size*, i.e., the number of types, is often referred to as *V* (Baroni 2008: 805). It must also be noted that the number of tokens in a text or in a corpus is always necessarily higher than the number of types in the same corpus.

2.1.2. The basic notions of the English word-formation system

There are two main processes of word-formation in English, namely affixation by *derivation* and *compounding*. There are other, more marginal types of word-formation in English, such as *blending*, *clipping*, and *conversion*, but this section will mainly concentrate on derivation and compounding, since they bear the strongest resemblance to the process of coining new words with combining forms.

Affixes are bound morphemes that are attached to bases,[1] often modifying the meaning or the part of speech of the base in some way. They are further divided into prefixes and suffixes: prefixes precede the base (*co-* in *co-occur*), while suffixes are attached at the end of the base (*-less* in *childless*). Minkova and Stockwell (2009: 101–102) state that affixes are relatively typical in loanwords, since almost all the Greek and Latin loanwords in English contain one or more affix. The third form of affix is called an *infix*, which is inserted into another morpheme, as in *abso-bloody-lutely* (Plag 2003: 11). Infixes are generally considered as a marginal type of word-formation in English (Plag 2003: 101).

With respect to semantics, the affixal meaning is often not as clear as the meaning of the root (Minkova and Stockwell 2009: 71). Biber et al. state that suffixes in particular often convey meanings that would be difficult to paraphrase, a fact that might also be partially interconnected with their class-changing nature. Prefixes, on the other hand, often express meanings that could also be conveyed by using such things as adjectival premodifiers or other lexical words (Biber et al. 1999: 324). For example, the meaning of such suffixes as *-ist* or *-ian* can be explained by saying that they are used to form agentive nouns (*scientist*, *physician*). On the other hand, the meaning of prefixes (e.g., *neo-*, *micro-*) can often be explained by using adjectives (‘new’, ‘small’).

There are several ways to classify affixes in English, each of which has its own advantages and disadvantages. Various classifications are introduced in Plag (2003: 85–86). In addition to the obvious prefix-suffix dichotomy, he mentions such criteria as the grammatical category of the base word, the semantic properties of the affix, and the ability of the affix to change the grammatical

[1] *Root* and *stem* are concepts that are easy to confuse, and in fact there is some disagreement as to how they interrelate. Minkova and Stockwell define root as the central element of derivational process that can either stand free (*host)* or be bound (*seg* in *segment*) (2009: 69). Base, on the other hand, may either contain a root or a root plus affix, while the term stem is usually used in inflection (2009: 71).

category of the base word. Plag states that one of the most unproblematic classifications, at least with respect to suffixes, is the syntactic category of the derived word. Some suffixes, such as *-age* and *-ess*, always result in nouns, while *-ify* and *-ize* are used to derive verbs from nouns or adjectives. Prefixes, on the other hand, are, because of their "more semantic" nature, often classified according to their meaning (see, e.g., Plag 2003: 98–99, Marchand 1969: 134, Minkova and Stockwell 2009: 102–105). As a final remark, it must be noted that inflection, which allows the expression of various grammatical categories such as number (with nouns) or tense (with verbs) also makes use of affixes, but it is generally considered as part of syntax rather than word-formation (Plag 2003: 14).

Compounding, as opposed to derivation by affixation, involves combining two (or possibly more) free forms into a single unit. In general, compounding is the most important source of new lexemes in the English lexicon (Minkova and Stockwell 2009: 9). As for their internal structure, compound lexemes are usually divided into *endocentric* and *exocentric compounds.*[2] Endocentric compounds involve what is usually called a *modifier-head structure*, where one element, the head, is somehow modified by another element (Plag 2003: 132). In English, compounds tend to be right-headed (Carstairs-McCarthy 2002: 61). For example, in *doll house, house* is the head that is modified by *doll*. The head determines most of the semantic and syntactic properties of the compound as a whole, so if the head is a noun (as in *doll house*), the whole compound must be a noun (Plag 2003: 132).

Exocentric compounds, or *bahuvrihi compounds*, on the other hand, exhibit a structure where the semantic head is outside the compound (Plag 2003: 145). For example, *pickpocket* is not a kind of pocket but a person who picks pockets, and *turncoat* is not a kind of coat but a person betrays their original cause by shifting to the opposing side. However, as Plag points out, these compounds do have a head on the structural level, since the part of speech of the compound is again determined by the right-hand member of the compound as a whole (2003: 145–146). Thus, *pickpocket* is a noun, not a verb. A third type of compounds

[2]Carstairs-McCarthy uses the terms *headed* and *headless* compounds, respectively (2002: 64). Minkova and Stockwell, on the other hand, distinguish between *syntactic compounds*, which are formed by regular rules of grammar and are therefore not listed in dictionaries (e.g., *birthplace*), and *lexical compounds*, the meaning of which cannot be computed from the meanings of its constituent elements (*ice cream*) (2009: 10).

includes words such as *singer-songwriter* and *poet-translator.* They are called *copulative compounds,* or *dvandva compounds,* and can be said to have two semantic heads, neither of them being more prominent than the other (Plag 2003: 146).

2.1.2. The diachronic perspective on the English word-formation

The primary subject of this study is not the diachronic development of the English word-formation system. However, a diachronic approach may explain some of the differences between different word-formation processes in terms of morphological productivity (see Carstairs-McCarthy 2002: 100). As Nevalainen has put it, the rate of language change is most easily seen in the development of lexicon, despite the fact that the core vocabulary of a language is usually relatively stable (1999: 332). None of the combining forms studied here are of native origin, but belong to the foreign strata of the English lexicon.

Affixes have originally been independent words that have had an auxiliary function, often derivational or inflectional (Warren 1990: 123–124). Eventually, they have gone through a grammaticalization process, which has resulted in semantic and phonetic erosion of varying degrees. Minkova and Stockwell call this phenomenon *semantic bleaching* (2009: 102). For example, the original meaning of the suffix *-hood* (‘condition’, ‘state of affairs’) is almost completely bleached out and is almost impossible to specify in contemporary English. Other similar examples in modern English include the native suffixes *-dom* and *-ship,* which are derived from nouns (Minkova and Stockwell 2009: 72). On the other hand, sometimes the meaning of a word develops in the opposite direction and a bound morpheme may acquire a new meaning as a free root, like the noun *ex* (‘a former spouse’) (Minkova and Stockwell 2009: 72).

In general, the assumption is that new non-native affixes are born when a sufficient number of complex lexemes are borrowed into a language and they thus start to be interpreted as morphologically transparent (Nevalainen 1999: 377, Marchand 1969: 129). Transparency, which means the fact that a lexeme is

formally and semantically analysable into its constituent parts, seems to be at least indirectly linked to productivity (Dalton-Puffer 1994: 252–253, Pounder 2000: 134). The relationship between transparency and productivity is discussed in more detail in section 2.3.1.

In the course of history, the English vocabulary has been shaped by several waves of loanwords from various languages, such as French, Latin, Greek, and the so-called Neo-Latin, which was a mixture of Latin and Greek vocabulary and which served as the lingua franca during the Middle Ages and the Renaissance (Kastovsky 2009: 1). Besides the vocabulary, this considerable borrowing has also affected the patterns of derivational morphology available to speakers (Cowie and Dalton-Puffer 2002: 410).

In terms of such morphological strata, the lexicon of Old English was very homogeneous, with only about three percent of the lexical items being of non-native origin (Kastovsky 2006: 167). The Norman Conquest, however, brought about enormous changes in the English word-formation system, and many Germanic word-formation processes became obsolete (Marchand 1969: 130). These non-native derivational patterns became productive gradually (Kastovsky 2006: 167–168). First, individual lexical items were borrowed, and after a certain point "a formal-semantic relationship" could be established so that the pattern could be extended to new formations by analogy, until the word-formation process finally became productive.

The number of native prefixes in contemporary English is very small (Marchand 1969: 129). They include *a-*, *be-*, *fore-*, *mid-*, *mis-*, and *un-*, and they have their origin in independent words. During the Middle English period, a great number of affixes, such as *dis-*, *en-*, *inter-*, *non-*, *-age*, *-ation*, *-ive* and *-ous*, was borrowed from French and Latin, and by the Early Modern English period they had become very frequent (Kastovsky 2006: 168). This abundance of borrowing affected the whole word-formation system by causing competition, not only between native and non-native affixes, but also between the non-native elements themselves (e.g. the various negative prefixes *a-*, *dis-*, *in-* and *non-*) (Kastovsky 2006: 169). As for suffixes, the effect of the invasion of French, Latin and Greek loanwords has not been as radical (Marchand 1969: 227).

During the Modern English period, the extensive borrowing has continued, this time because of the development of scientific terminology,

introducing elements such as *bi-, di-, hyper-, hypo-, meta-, micro-* and *multi-* – the number of prefixes outnumbering that of suffixes (Kastovsky 2006: 170). In addition, any learned affixes, or combining forms,[3] have acquired new meanings (Hughes 2004: 347). For example, *extra-* used to mean 'out of', as in *extraordinary* ('outside the scope of ordinary') but is nowadays used to denote a meaning 'more than the usual', as in *extra strong.* Similarly, *super-* 'above' and *ultra-* 'beyond' are increasingly used to stress
size.

Kastovsky summarizes that the overall structure of English vocabulary has changed and the number of derivational patterns available has increased enormously from Old English to the present day, "partly rivalling each other, partly restricting each other according to etymological domains" (2006: 170). In a similar vein, Hughes writes that different word-formation patterns are used with greater flexibility in contemporary English than they were before: recent coinages that exploit the suffix *-able* include, for example, such words as *doable*, *puttable* and *liveable* (2000: 346).

2.2. Combining forms

2.2.1 Defining combining forms

It was stated above that affixes are always bound morphemes that attach to a root or a base. What should then be thought about such words as *neuro-logy* or *bio-logy?* At first sight it seems that lexemes like these are formed with a prefix and a suffix but no root, which of course runs counter to the basic assumption that words should always contain a root (see Bauer 1983: 213–214).

Carstairs-McCarthy offers a partial solution by stating that affixes are always bound but roots are not always free (2002: 20). Thus the first element in words like *audi-ence, magn-ify,* or *applic-ant* is actually a bound root, since they cannot occur alone in English (Castairs-McCarthy 2002: 19). The common denominator with all these words is that they are all loanwords, borrowed either

[3]The distinction between affixes and combining forms will be discussed in more detail in section 2.2.

from Latin or Latin via French.[4] The same "boundedness" applies to such formations as *neurology* and *biology* as well, but in this case not only the first element but also the final constituent is bound. This is what Castairs-McCarthy means when he points out that words can in fact consist of more than one bound root (2002: 21). In this sense, the constituents in words like *neurology* and *biology* could also be called bound roots, but it has become a common practice by linguists and dictionary-makers to call them *neoclassical elements,* or, more commonly, *combining forms* (Plag 2003: 74; Carstairs-McCarthy 2002: 21).

The term combining form was first used in the *New English Dictionary* (1884–1928), which was the predecessor of the *Oxford English Dictionary* (Kastovsky 2009: 2; see McCauley 2006 for the use of the term combining form in different editions of *NED* and *OED*). The current version of the *OED Online* has a total of 2244 entries that are categorized as combining forms. However, many entries under the label of combining forms have been first published as early as in 1899, but have not yet been fully updated. As for the definition of the term itself, the *OED Online* cites Bloch and Trager (1942: 66):

> In Latin and other languages, many words have a special combining form which appears only in compounds (or only in compounds and derivatives)... The foreign-learned part of the English vocabulary also shows a number of special combining forms; cf. *electro-,* combining form of *electric*, in such compounds as *electromagnet.*

For the third edition of the *OED,* the category of combining forms is defined in the following way:

> [A]n element used, either initially or finally, in combination with another element to form a word. For the purposes of *OED3,* a combining form differs from a prefix or suffix by being generally noun-like or adjective-like

[4]It is not always easy to differentiate between loanwords of Latin and French origin, partly because these two languages are etymologically related and because it is not always clear at which point in history a certain word has been borrowed into English. When borrowing from Latin/French was at its most active phase, England was in fact a trilingual community, which makes it even trickier to trace the etymology of some words (see McConchie 2006). Nevalainen in fact uses the term *latinate* to describe words that are of either Latin or French origin (1999: 371).

> and having a relatively full lexical meaning. (McCauley 2006: 96)

Marchand writes that the practice of coining new lexemes with learned prefixes and combining forms dates back to the 16th century (1969: 131). This was mainly due to the Renaissance and the revival of sciences in Western Europe, which produced an increasing need for scientific vocabulary (see also Carstairs-McCarthy 2002: 66). However, the practice of coining new words with combining forms did not become common until the 18th and 19th centuries, along with the industrial revolution and the development of modern science that it brought about (Marchand 1969: 131; Carstairs-McCarthy 2002: 66–67).

In general, the proportion of specialist terms in the lexical intake grew steadily throughout the eighteenth century, which contributed to the fact that neoclassical formations started to be associated with technical registers (Nevalainen 1999: 365). The development of learned vocabulary usually goes hand in hand with the development of the corresponding domains of science and scholarship (Fradin 2000: 23). Quirk et al. point out that in Present-Day English, in addition to scientific registers, combining forms are also common in commercial brand names aimed at international markets (1985: 1535). Many neoclassical compounds are also "in international currency", which means that they have been adopted in some form in several languages (Quirk et al. 1985: 1575).

It must be noted that there is some variation in terminology. So far we have only talked about what has been usually called *classical combining forms* (see, e.g., Fradin 2000: 11; Prćić 2005: 317). However, the term combining form has also been extended to cover elements such as *-gate* in *Watergate* or *-ware* in *shareware* (Lehrer 1990: 14, Fradin 2000: 11). Lehrer calls them "productive splinters", claiming that they have more in common with combining forms than with affixes or roots (1990: 14). Prćić distinguishes between classical combining forms, which have been adopted from Greek or Latin, and modern combining forms, which are based on modern English words but which have been modified to look like combining forms by, for example, the addition of a thematic vowel or by clipping (2005: 317; 2007: 383). Examples include *jazzo-phile*) (from *jazz*), or *Thatcher-nomics* (from *economics*). Another type of formation that Prćić defines

as combining forms includes various other bound forms that are independent words in their source languages (e.g., *scandal-monger* and *dulls-ville*) (2007: 384). Prćić calls combinations like these, where one element is modern and the other neoclassical (either with modern ICF + neoclassical FCF or neoclassical ICF + modern base) "semi-neoclassical compositions" (2005: 321).

This study will concentrate on what has been called classical combining forms, which are said to represent the "core of established combining forms" and are usually listed under the label of combining forms in dictionaries as well (Fradin 2000: 12). Warren defines them as allomorphic variants of Greek or Latin model words (1990: 115). They can either be derived from lexemes proper (*-logy*) or Greek or Latin prepositions (*hyper-, quasi-*). Due to the limited scope of this study, I have therefore excluded formations that are clippings from pre-existing words (e.g., *euro-*) and elements that are not strictly speaking new morphemes at all (e.g., *-gate*; see Warren 1990: 115). The etymology of the selected combining forms is discussed in more detail in section 3.1.

Bauer distinguishes between initial combining forms (ICF) and final combining forms (FCF), depending on whether they occur in initial or final position (2003: 214). This dichotomy has become widely adopted in studies concentrating on combining forms (see, e.g., Kastovsky 2009). Quirk et al. write that "the first element [of a compound] may be in its special 'combining form' (as in the noun *trouserleg* or the adjective *socioeconomic*)", and thus exclude final combining forms from this definition (1985: 1567). In a similar vein, some editions of the *OED* reserve the term combining form for the first element of a compound and use the term *terminal element* for the final element (Kastovsky 2009: 4). However, this distinction has generally been replaced by the one between ICFs and FCFs, which will be adopted in this study as well.

2.2.2. Neoclassical compounds

Formations that make use of combining forms are called *neoclassical compounds* (Plag 2003: 74). Bauer mentions several characteristics that neoclassical compounds have in common with native compounds in English (1998: 405). For example, the meaning relationship between the constituent parts varies and cannot

always be "deduced in isolation from its context". Secondly, they follow the condition of hyponymy in the sense that they are hyponyms of their head element (which is the rightmost element of the compound). These same word-formation strategies were functional in the donor languages (Greek and occasionally Latin) as well, the only difference thus being the fact that neoclassical compounds are modern formations that are not attested in Greek and Latin as such (Bauer 1998: 405).

What makes neoclassical compounds atypical in English is that the English language seems to prefer compounds that consist of free roots, such as *bookcase* or *motorbike* (Carstairs-McCarthy 2002: 21). Combining forms, on the other hand, have "little or no currency" as separate words, and they are not normally the stressed part of a complex word (Quirk et al. 1985: 1520). Since they consist of bound elements but otherwise behave like compounds, they have sometimes been described as situated between compounding and affixation in the English word-formation system (Quirk et al. 1985: 1520).

Many neoclassical compounds contain a central linking vowel, usually *-o-*,[5] which cannot unequivocally be assigned to either of the constituent combining forms (Carstairs-McCarthy 2002: 66). If the vowel is considered as part of the first constituent, the final combining form in such words as *anthropology* and *cosmology* will be *-logy*. However, in such words as *toyology* (two occurrences in the *BNC*) or *grandadology* (one occurrence), the initial element is of native origin, and thus the final element might be considered to be *-ology*, which could then be considered as a morphologically conditioned allomorph of *-logy*, in which case the central *-o-* will be part of the final constituent (Kastovsky 2009: 6).

The primary stress of neoclassical compounds is usually on this central *-o-*, a feature that distinguishes them from native compounds that usually have the stress on the first element (Carstairs-McCarthy 2002: 66). In the majority of cases, the *-o-* does not show up if the initial combining form ends with a vowel or the final combining form begins with a vowel (e.g., *tele-scope* vs. *laryngo-*scope) (Plag 2003: 158). However, in some cases, such as in *bio-* and *geo-*, the *-o-*

[5]Since the linking vowel *-a-* has been gaining ground especially in modern formations, Prćić calls it a "modern linking vowel" (2007: 384). It must also be noted that the orthographically, other forms are accepted as well, provided the appropriate phonetic criteria are met, as in *bureaucracy*.

seems to be part of the inherent structure of the combining form, hence *bio-energy* and *geoarchaeological* (Plag 2003: 158). According to Kastovsky, the vowel has originally been either a stem-formative or an inflectional ending (and thus associated with the first element), but they have both become opaque in English (2009: 6, 10). He suggests that the most useful interpretation of *-o-* would be as a simple linking vowel.

Quirk et al. note that compounding with initial combining forms may also be recursive, i.e., involve more than two elements, such as in *neurolymphomatosis* or *nephrolithotomy* (1985: 1576). Final combining forms, on the other hand, cannot be used recursively (Prćić 2007: 384). In terms of semantics, the meaning of neoclassical compounds is usually predictable from the meaning of its constituent combining forms (*anthrop(o)-* 'human' and , *-(o)logy* 'science or study' equal 'science or study of human beings'), which is important especially in coining new technical terms (Carstairs-McCarthy 2002: 66).

There are three ways in which a neoclassical compound may enter a language (Cottez 1985: XVII, cited in Fradin 2000: 24). The lexeme may already exist in the donor language, in which case it is borrowed with the same meaning and corresponding form (*phlebotomy* vs. Greek *φλεβοτομία*). The second way is adaptation: the lexeme exists in the donor language, but with a different meaning (*physiology* vs. Greek *φυσιολογία* 'natural sciences'). The third way is creating new lexemes by combining elements that exist in the donor language as separate units, not as a whole (*chronology* from *χρονο* and *λογία*).

2.2.3. The status of combining forms

Bauer argues that there is plenty of evidence for defining combining forms as a category distinct from affixes: combining forms can combine with each other (an ICF + FCF) but not with affxes, and they usually look more like lexical items than affixes (1983: 214–215). The latter feature also applies to their phonological structure, since they carry their own stress, contain full vowels, and are generally disyllabic (Bauer 2003a: 35). In addition, affixes constitute a closed set of morphemes, while combining forms form a relatively open set that grows when

new lexemes are created with the help of resources from Greek and Latin – or English (Prćić 2005: 316).

Yet another feature that has been suggested as distinguishing combining forms and affixes is that the same affix can never function as both prefix and affix, while some combining forms can be seen both as ICF and FCF (e.g., *patho-* in *pathology -pathy* in *cardiopathy,* both from Greek *πάθος* 'emotion') (Harastani et al. 2012: 74). On the other hand, combining forms also have plenty of shared features with traditional affixes: for example, both can be separated from the remaining element of the lexeme without their meaning or form being changed, and both produce "output formations of a binary structure" that are morphologically and semantically analysable into their parts (Prćić 2005: 315–316).

The line between combining forms and other non-native bound roots is not always clear either: some combining forms can be combined with traditional affixes as well (Carstairs-McCarthy 2002: 66). For example, *socio-* and *-logy* can be used with the suffixes *-al* and *-ic(al)* to form such words as *(psycho-)social* and *logic(al).* Prćić calls this phenomenon suffixal expansion, stating that in such cases, the combining form becomes composite in nature. The added suffix becomes the head of the combining form, and the word as a whole can be seen to display a dual headedness where the head at the CF level (e.g., *-ic*) can be termed the *micro-head* and the head at the word level (*logic*) a *macro-head* (2007: 386–387). In this study, such suffixal expansions will not be taken into account in analysing the selected final combining forms, for reasons explained in section 3.3.

To summarize, the status of combining forms in the English word-formation system is not clear. They have features that distinguish them from compounding on the native basis, but they are also distinct from affixes. The problem is that neoclassical combining forms have not been studied very systematically (see Kastovsky 2009: 1). Marchand criticizes the *OED* for being inconsistent with the term and for not properly discussing their exact role in English word-formation (1969: 132, 218). What he claims is certainly true, and the *OED*'s, definitions for the category itself are vague at best.

However, Marchand himself is rather inconsistent in defining combining forms and does not provide any solution as to their status. For example, he writes that elements that are attached to full English words should be

termed prefixes, and others "are of a purely dictionary interest in any case" (1969: 132). This, as Kastovky has noted, is "a very questionable argument", since they are extremely frequent, especially in scientific registers (2009: 4).

Bauer argues that neoclassical compounding is in fact a rather arbitrary category (they can be a part of clippings and blends, like in *telethon*, a blend of *television* and *marathon*) and that it should be defined in terms of more or less prototypical members of the category (1998: 407–409, 419; see also Kastovsky 2009: 8). He suggests that problems with classifying neoclassical compounds and combining forms might be due to the falsely unambiguous label "neoclassical compound" itself (1998: 414). Kastovsky presents a similar view, but goes even further by claiming that the term might as well be abandoned, since "modern word-formation theory can well do without it" (2009: 2). He explains this by the fact that even though modern English word-formation is dominated by word-based morphology, [6] English has recently adopted some stem-based morphology as well in the form of borrowed non-native elements, such as the *soci-al* and *logi-al* mentioned above (2009: 9).

His solution thus involves simply introducing the concept of stem as a possible lexeme representation (2009: 11), thus elaborating on the ideas presented by Carstairs-McCarthy (2002: 21). The stems can either combine with affixes (*electr-ify*) or with another stem (*socio-logy*), where the latter process produces what has been called neoclassical compounds. As for the modern combining forms, Kastovsky writes that such established word-formation processes as blending (*chunnel*, from *channel* + *tunnel*) and clipping (*euro-*, as in *Eurocentric, eurocrat*) can be used to account for them (2009: 11–12).

Kastovsky's proposition is certainly appealing, and would make the make-up of the English word-formation system much more simple and consistent. It is, however, the ability of combining forms to combine with affixes that causes some trouble. Bauer presents such ungrammatical examples as **electroness* and **electronization* to justify the claim that the combinations are in general not possible. It is true that such combinations *electr-ic* or *soci-al* are perfectly grammatical, which would indicate the contrary. It must be also kept in mind that

[6]In word-based morphology, the base form is directly accessed and thus not an abstraction, while in stem-based morphology, the base does not occur without a derivational (or inflectional) morpheme, such as in *scient-ist* (Kastovsky 2006: 157).

not all affixes can attach to all combining forms, just as it is not possible for any affix to attach to any base. However, there are also some (usually initial) combining forms that do not seem to be able to combine with any affixes, such as *pseudo-* and *hyper-*. This might indicate that there is a cline between more affix-like and more lexeme-like combining forms.

In addition, Kastovsky himself seems to be somewhat inconsistent by terming such elements as *auto-*, *hypo-* and *mono-* as prefixes instead of combining forms. But what is the exact quality that would make them prefixes, not combining forms? The *OED* defines *auto-* and *mono-* as combining forms, and *hypo-* as a prefix (the former being developed from lexical words, the latter from a preposition).

If Kastovsky's view is accepted, it will raise a number of questions. What is the benefit of studying the productivity of a category that does not exist in its own right? Still, the category of combining forms (at least in the classical sense) might be useful from the point of view of the study of scientific terminology. Furthermore, the productivity of these selected elements has not been studied yet, and it may still reveal something new about them and help situate them in the big picture of the English word-formation system. In addition, it is the differences between initial and final combining forms that we are interested in, and in this kind of study a term like combining form might not be irrelevant. All in all, it can be said that none of the suggested solutions concerning the status of combining forms presented in literature has proved completely satisfactory so far. Thus, the term combining form will be used in this study, partially for lack of a better term.

2.3. Morphological productivity

2.3.1. Defining morphological productivity

Morphological productivity is an important concept in the study of word-formation and morphological theory in general (Cowie and Dalton-Puffer 2002: 411). Productivity is often defined as the speakers' ability to create new lexemes by exploiting the resources of a language. Speakers are able to create an almost unlimited number of new words, and the lexicon of a language is thus never fixed

(Haspelmath 2002: 39). Therefore, word-formation processes can be called either productive or unproductive, depending on the extent to which they can be applied to form new morphologically complex words (Plag 2006: 537).[7]

To illustrate the difference between productive and unproductive word-formation processes, Plag mentions the verb-forming suffix *-ize* (*grammaticalize*) and the prefix *en-* (*enlist*), which are both used to form verbs in English (2006: 537). The former can be considered fairly productive, since it is frequently used to form new lexemes that have not yet made their way in the established lexicon. The latter, on the other hand, cannot be so easily used in new formations in contemporary English, and it is thus felt to be unproductive.[8]

Haspelmath writes that unproductive rules are a unique property of morphology, since there is no direct equivalent to them in the area of syntax (2002: 40). Carstairs-McCarthy explains this phenomenon by mentioning the nonexistence of clearly ill-formed words (**ion-trans-al-at-form*), but also of some words that are in theory well-formed (**arrivation, *ridiculosity*), as well as the unpredictability and non-compositionality of the meaning of some existing words (*recital* vs. *recitation*) (1992: 32). Bauer states that it is the productivity of word-formation processes that is at least partially responsible for the huge vocabulary of English, a fact that can be verified by consulting various dictionaries of neologisms (1983: 63).

Another way to illustrate the difference between productive and unproductive categories comes from Baayen, who defines productive categories as ones with a growing membership, while unproductive categories have a fixed or declining membership (2009: 900). Whether productivity should be understood as an essentially synchronic or diachronic phenomenon (or both), is controversial, however. Bauer claims that it is possible to talk about productivity in synchronic terms, or about *changes* in productivity in diachronic terms, but not of productivity as such as a diachronic phenomenon (1988: 61). Cowie and Dalton-

[7]Terms such as *word-formation process, word-formation rule* (often abbreviated as WFR), *morphological rule,* and *morphological category* are often used more or less synonymously in the literature. They are usually used to refer to the way morphemes can be combined to form lexemes (Plag 2003: 30, 179). The rule must contain information at least about the phonology and semantics of the morpheme as well as about possible base morphemes (Plag 2003: 31). This is what distinguishes "rules" or "processes" from the rather mechanical concept of *morpheme*. This study will mainly use the term word-formation process.

[8] One of the few exceptions is *encrypt,* which was first attested in the 1950's, according to the *OED.*

Puffer write that if productivity is to be taken as "a design feature of language in general", it should be understood as a synchronic notion (2002: 417).

On the other hand, it has been suggested that using type frequency, for example, as an indicator of productivity (i.e., counting all the types formed between two points in time) is inherently diachronic, even though the aim was to study productivity from a synchronic point of view (see Cowie and Dalton-Puffer 2002: 418 for discussion). Be that as it may, changes in the productivity of certain word-formation patterns have been frequently studied as well (see, e.g., Dalton-Puffer 1994, 1996; Cowie 1998; Cowie and Dalton-Puffer 2002; and Säily and Suomela 2009).

Plag asks whether productivity should be seen as a theoretical primitive (i.e., an inherent property of word-formation rules) or whether it results from other properties of these rules (2006: 537). According to Dressler and Ladányi, productivity is "a primitive and prototypical property" of word-formation rules, comparable to inflectional, syntactic, or phonological rules (2000: 119). On the other hand, Férnandez-Domínguez (2009: 53) claims that since linguists often list a number of features associated with productivity (e.g., the various constraints on it), the phenomenon might be considered as being of composite nature and not an indivisible notion.

The distinction between inflection and derivation is also relevant in defining morphological productivity. Although productivity is often considered more relevant for the study of word-formation rather than that of inflectional processes, Plag states that, at least to some extent, productivity also has to do with inflection (2006: 538). For example, the regular past tense affix *-ed* might not be considered fully productive, since some verbs form past tense forms with ablaut or take no overt suffix, although the formation of the past tense is fully productive as a category. According to Bauer, concatenative processes such as conversion and back-formation should also be understood as morphological processes and thus relevant for productivity as well (2005: 316). The productivity of compounds has also been researched, though to a lesser extent (see Férnandez-Domínguez 2009). Bauer goes as far as to suggest that productivity even concerns syntax: some syntactic structures are used more frequently than others to form new sentences, and thus the notion of syntactic markedness might be understood as some form of productivity or as relating to it (1995: 21).

Morphological productivity and *linguistic creativity* are two concepts that are easy to confuse. There is in fact some confusion and overlapping between the two terms among linguists themselves (see, e.g., Bauer 2001: 1, 62ff). According to Bauer, productivity is an inherent feature of human language, which allows speakers to exploit the rules of word-formation to create new lexemes, while creativity allows for a native speaker "to extend the language system in a motivated, but unpredictable (non-rule-governed) way" (1983: 63; see also Booij 2005: 67 and Dressler and Ladányi 2000: 105).

Renouf writes that creativity is often employed to achieve a stylistic effect, and is typically seen as a departure from the rules of grammar, which can manifest itself in such instances as punning or wilful error (2007: 70). Evert and Lüdeling state that words that result from productive processes are often not recognized as new by speakers, while creative formations are perceived as new (2001: 475). Bauer also uses the term "semi-productivity" to describe linguistic creativity, claiming that it differs from productivity by requiring some sort of encyclopaedic, extra-linguistic knowledge (1995: 26, 28). Plag writes that productivity is often considered unintentional and creativity intentional, but he also claims that the intentional/unintentional dichotomy is not very useful, since speakers have varying levels of awareness of the structure of the lexicon. In addition, even productive rules may be used intentionally (1999: 13–14).

It must be noted, however, that productivity should not be confused with *regularity* either, since all productive patterns are rule-governed and thus regular, but so are many unproductive rules as well (Dressler and Ladányi 2000: 105). For example, the suffix *-ess* is perfectly regular, but cannot be used to coin neologisms (Haspelmath 2002: 111). The same holds true for *transparency*: a lexeme can be perfectly transparent but not necessarily productive (Pounder 2000: 134–135; Bauer 2001: 54).[9] On the other hand, productive processes are usually formally and semantically transparent, unless the process becomes lexicalized (Bauer 2001: 135). Dalton-Puffer points out that the term transparency is often

[9]Transparency and regularity are closely related concepts, and sometimes used synonymously. Pounder (2000: 135) suggests that *transparent* emphasizes the "analysis side of morphology", while *regular* has more to do with the fact that the formation is not isolated in the morphological system of a language from the speaker's point of view. Another related concept is *analysability*, which is used by Dalton-Puffer 1994, who uses it more or less synonymously with *transparency*.

used in a vague sense in the literature, since it is usually not specified whether the focus is on the linguist's or the speaker's point of view (1994: 253).

2.3.2. Qualitative versus quantitative approaches to productivity

The approaches to defining the nature of productivity can roughly be divided into two, qualitative and quantitative. If productivity is considered a qualitative phenomenon, a given word-formation process either has this character or not (Plag 2006: 538–539). The qualitative approach to productivity has mainly concentrated on various kinds of phonological, morphological, syntactic and semantic restrictions of different word-formation processes, the goal having been to provide an intensional description of all the possible bases of a word-formation process (Evert and Lüdeling 2000: 167). The most typical constraints on the productivity of word-formation processes are discussed in section 2.3.3.

It has also been claimed that the productivity of word-formation processes can be described as a continuum rather than as an all-or-nothing property in the sense that these processes are productive to a varying extent (Plag 2006: 539). Baayen calls this kind of productivity *scalar*, as opposed to an *absolute* view of productivity (1992: 185). Most scholars currently seem to assume at least some degree of scalar productivity, whether it means that productivity is infinitely variable on a scale or that there is a certain number of steps on that scale (see, e.g., Bauer 2001: 15–16, 126). Even though the qualitative and quantitative approaches to productivity can be seen as a dichotomy, the suggestion is that they are nevertheless closely related. As Plag has put it, "the idea of potentiality, which is central to qualitative definitions of productivity, can be expressed in the statistical terms of probability" (1999: 22).

The distinction between the *availability* and *profitability* of morphological processes is quite useful in grasping the difference between qualitative and quantitative aspects of productivity. The terms come originally from Corbin, who calls them *disponibilité* and *rentabilité*, respectively (1987: 177).[10] The availability of a word-formation process means whether it can be used

[10] The English translation is by Carstairs-McCarthy (1992: 37).

to form new lexemes, and the profitability of a process shows how many bases the process can affect (Carstairs-McCarthy 1992: 37). Availability is thus a qualitative notion, while profitability can be seen as quantitative (Plag 2006: 122).

Profitability is often used as synonym for *type frequency*, that is, the number of words that have been produced using a certain word-formation process (the number of actual words). In this sense, profitability is a rather straightforward concept, since it is possible to test it empirically (Férnandez-Dominguez 2007: 60). Availability, on the other hand, is a trickier concept. Again, the problem lies in the synchronic versus diachronic dichotomy. Processes that are being used, i.e., available, at present, are rather unproblematic, but problems arise with processes that have ceased to be used productively. Férnandez-Domínguez suggests that availability corresponds to the paradigmatic axis of productivity and profitability to its syntagmatic axis (2007: 63–64). In other words, availability concerns the selection between competing suffixes, such as the noun-forming suffixes *-ance*, *-al*, or *-ion*.

2.3.3. Measuring morphological productivity

As was stated above, quantitative approaches to productivity often assume that there is a way to operationalize the problem and determine the exact degree of productivity of a particular word-formation process. According to this view, the productivity values of different processes can also be compared. Productivity as a quantitative phenomenon is probably mentioned first by Bolinger, who defines it as "the statistically determinable readiness with which an element enters into new combinations" (1948: 18). Since then a number of methods to measure productivity have been proposed. The most important measures will be discussed next.

Type frequency has been suggested by several authors as a rather good measure for gauging productivity. Baayen calls this kind of measure *realized productivity* (2009: 901–902) or extent of use (1992). Haspelmath uses the term *degree of generalization* (2002: 109). The basic assumption with realized productivity is that a productive morphological category has many members. Thus, it can be measured simply by counting the number of different types (that is, different word forms) within a certain morphological category.

The method is not devoid of problems, however. Type frequency does not tell us anything about the availability of a process, i.e., whether the process can be used to form new words in the future or whether it has been "saturated" (Bauer 2001: 144–145). For example, the suffix *-ment* has a high type frequency, but there are only a few neologisms with *-ment* in the *OED* that have been recorded in the twentieth century (Haspelmath 2002: 109). In this sense, type frequency can only be used as an indicator of past productivity (Cowie and Dalton-Puffer 2002: 416). In addition, if type frequency is assumed to equal the degree of productivity of a certain word-formation process, it may lead to counterintuitive results. There are many classes of words that have several members but that are not productive, such as pronouns.

Type frequency has been used as an indicator of productivity, especially in historical studies (as in Dalton-Puffer 1996, Cowie and Dalton-Puffer 2002), since many statistical methods do not function properly if there is scarcely data available (see below for discussion). Another method used in diachronic studies is counting the number of neologisms that are attested during a specified period of time (Plag 2006: 541). The neologisms can be determined with the help of a good historical dictionary, but the reliability of the dictionary may become an issue. For example, lexicographers may overlook new words that are formed with a very productive pattern (Haspelmath 2002: 110). Bauer also mentions the danger of confusing the actual process of word-formation from the results of the word-formation: in English, for example, there are several complex words (such as *payable*) which are not created in English but are loans (2001: 145). The advantages and disadvantages of dictionaries, as opposed to corpora, are discussed in section 3.2.

A more refined view of measuring productivity comes from Aronoff (1976), who has proposed that productivity should be understood as a ratio of actual to possible words: the higher the ratio, the greater the productivity of a given word-formation rule. However, it is not a simple task to count the number of possible words with that word-formation process. It could, in theory, be determined by identifying all the possible restrictions on that pattern, but in practice there are several unproductive rules in English that do not seem to have any identifiable restrictions (Haspelmath 2002: 110). In addition, since the possible bases may themselves be complex words that are formed productively,

the set of possible words becomes practically infinite (Haspelmath 2002: 110). Anshen and Aronoff have also found out that for highly productive word-formation processes, the model gives a rather low productivity index, since the number of possible words is, in theory, infinite (1981: 64). The most fundamental problem with Aronoff's model, though, is how to actually count the number of possible words. Baayen (1992, 1993) has attempted to formalize Aronoff's model by introducing the potentiality measure *I*, which is the ratio of the number of possible words *S* and the number of actual words *V* (the reverse of Aronoff's index) with a given affix.

The methods discussed above can be defined as quantifications, as opposed to *probabilistic measures* (see Plag 1999: 24). Baayen and his collaborators in particular (see, e.g.,Baayen 1992, 1993, 2009; Baayen and Lieber 1991;Chitashvili and Baayen 1993; Baayen and Renouf 1996) have developed a number of statistical methods to measure productivity that make use of large text-corpora.

Most of the statistical measures to gauge productivity rely on the notion of the hapax legomenon. The words *hapax legomenon* constitute a Greek expression that has been adopted from classical studies, and it literally means 'said (only) once' (Carstairs-McCarthy 2002: 96). Hapax legomena (or just 'hapaxes') are words that occur only once in a certain corpus (Bauer 2005: 325).

With productive morphological processes a large number of low frequency words is to be expected, because our mental lexicon "allows the decomposition of the newly encountered word ... and thus the computation of the meaning on the basis of the meaning of the parts" (Plag 2006: 542). Chitashvili and Baayen call this kind of phenomenon a Large Number of Rare Events (LNRE) distribution (1993: 57).[11] Unproductive processes, on the other hand, tend to result in a large number of high-frequency words (Plag 2006: 542). In a similar vein, the fact that a productive word-formation rule is available in the mental lexicon guarantees that all complex words with that affix can be produced or understood, even if they have a very low frequency (Baayen and Renouf 1996: 74). This implies that hapaxes can be said to be a good indicator of productivity.

[11] How this distribution can be further used in the statistical study of productivity will be discussed in section 2.3.3.

It is not important whether the hapax legomena are familiar to the speaker or not, but rather the fact that they "represent the rate at which new words are being coined in the language as a whole" (Bauer 2005: 325).[12]It is also worth noting that not all hapax legomena are necessarily neologisms – while some hapaxes are just old or obsolete words, some neologisms may occur more than once in a corpus or be completely unattested (Szymanek 2005: 430–431). Thus, hapaxes are not a goal in itself: they merely function as a statistical tool for estimating productivity (Baayen 2009: 905–906).

The most famous and widely cited of Baayen's measures also makes use of hapaxes. The measure has several names; most often it is called *potential productivity* (Baayen 2009: 901–902), *productivity in the strict sense* (Baayen and Lieber 1991: 821), or *the category-conditioned degree of productivity* (Baayen 1993: 200). Baayen and Lieber (1991: 809) have formalized the measurement of potential productivity as follows:

$$P = \frac{n_1}{N}$$

where n_1 represents the number of hapax legomena with a given affix and N the number of tokens of all words with the same affix. The resultant figure is a decimal that has a value between 0 and 1. The higher the value, the more productive the process is, and vice versa. For example, if there are 100 tokens with the prefix *re-* in the corpus, and 6 types that occur only once, i.e., are hapaxes, the *P* value of *re-* is 0.06. For the sake of simplicity, this measure will be mostly referred to as *P*.

Baayen and Lieber state that *P* allows measuring "the rate at which new types are to be expected to appear when *N* tokens have been sampled" and estimating the probability of encountering new types (1991: 809). Therefore, if the sample is representative enough, "*P* makes a statement of what potentially

[12]Cowie and Dalton-Puffer state that the"newness" of words is usually defined in terms of dictionaries:"what is not in a given comprehensive dictionary is taken to be new". This must not be confused with the psycholinguistic aspect of newness, since a word that is well established in a dictionary may be new to individual speakers (2002: 419).

could have been in the sample but has not been actualized for some reason or other" (1991: 811). In other words, the measure captures the probability that the next token in the sample is a hapax: according to Baroni, the proportion of hapaxes at the *N*th token should be a good enough estimate of how likely a word $N + 1$ will be a hapax (2008: 818).

Baayen and Renouf state that if the size of the corpus is increased, the number of hapaxes increases as well, but not in a linear way (1996: 75). This is explained by the fact that while the sample is still small, the probability of encountering a so far unattested word is relatively high. On the other hand, the larger the sample is, the more likely it already contains that particular word. Thus, the measure does not allow the comparison of *P* figures between different corpora (Baayen 1993: 191).[13] The growth rate of the vocabulary and hence the slope of the vocabulary growth curve varies at different sample sizes. Vocabulary growth curves will be discussed in section 2.3.3.

Potential productivity is not the only statistical method developed by Baayen and his colleagues. The *hapax-conditioned degree of productivity P** was introduced in Baayen (1993). Also called *expanding productivity*, it captures the rate at which a category is "expanding and attracting new members", and is calculated by dividing the number of hapaxes of a given category by the number of all the hapaxes in a corpus (Baayen 2009: 901–902). In other words, *P** can be used to study the contribution of a certain word-formation process to the expansion of the vocabulary in general. Baayen suggests that *P* and *P** should be seen as complementary processes, the former being used to distinguish between productive (even though this is not simple at all) and unproductive processes and the latter to rank productive processes (1993: 194).

The measures introduced in the work of Baayen and his colleagues are not the only statistical methods that have been developed to gauge productivity. Säily and Suomela (2009) have designed a model that makes use of accumulation curves and statistical significance testing. In their method, the corpus is divided into samples that preserve discourse structure. A growth curve for a morphological category (such as a particular affix) is then plotted on a

[13] This is a specific instance of a more general problem of "variable constants" in lexical statistics (Tweedie and Baayen 1998), cf. type/token ratio.

figure.[14]A sample is picked randomly, the number of types or hapaxes is calculated, and the sample is plotted on a figure (see Figure 1). Another (random) sample is picked and added to the previous one, until the whole corpus has been sampled. The procedure is then repeated, for example, a million times with the help of a suitable computer software, and the permutation, i.e., the order of the samples changes every time. The result looks something like Figure 2: it is possible to indicate visually the area covered by a certain percentage of the curves, which helps us to see the extent to which the productivity rates of a certain feature in a sub-corpus differ from those of the corpus as a whole (Säily 2011: 125).

According to Säily, type accumulation curves are a particularly good measure in comparing productivity between different (e.g., gender-specific) subcorpora (2011: 127, 135). There are two ways to plot a type accumulation curve: either the number of types as a function of corpus size, or the number of hapaxes as a function of the number of types. Hapax accumulation curves, however, require a large corpus in order to produce statistically valuable data in which the results are not skewed (Säily 2011).

[14]The curves can be plotted either for types (as a function of token frequencies) or hapaxes (as a function of running words in the corpus) (see Säily 2011: 126).

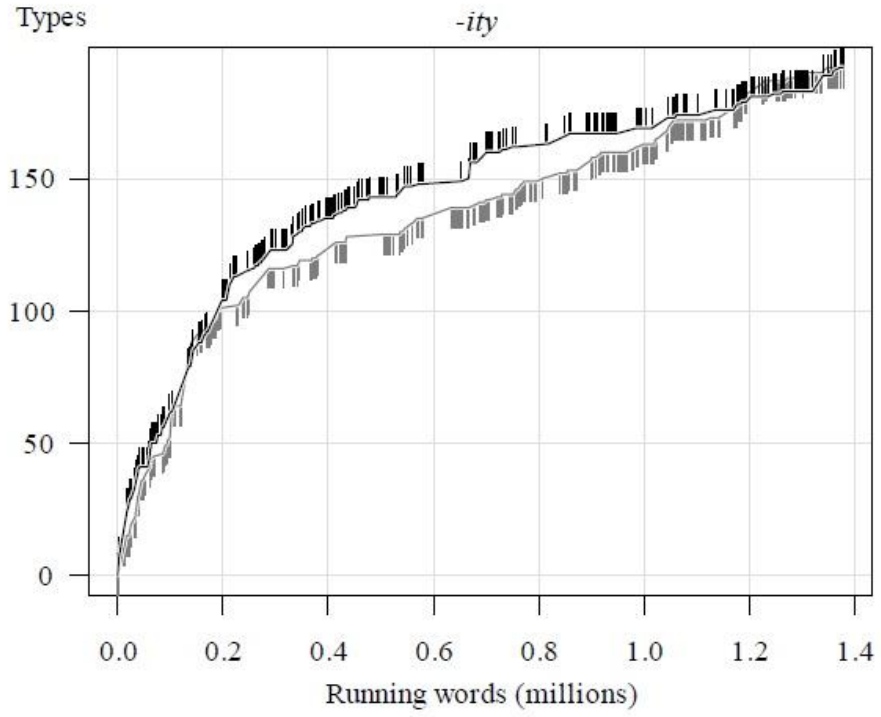

Figure 1. Two type accumulation curves. Each tick mark represents the addition of one sample. From Säily and Suomela (2009: 102).

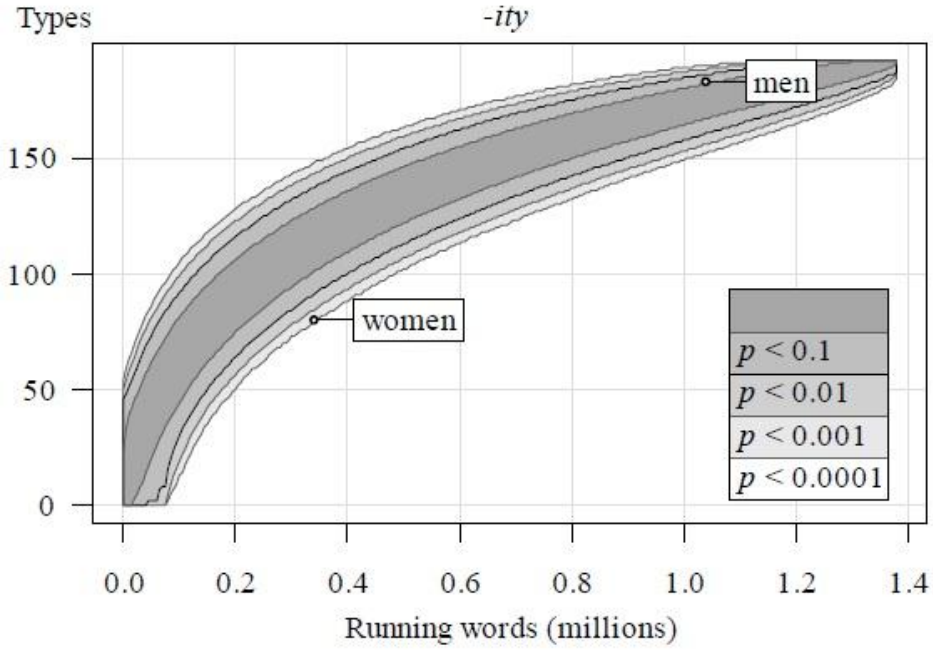

Figure 2. Bounds for 1,000,000 accumulation curves for the suffix *–ity* in the 17th-century part of the *CEEC*.Gender-based subcorpora plotted on the curves, which show that women have a significantly lower type frequency. From Säily and Suomela (2009: 100, as cited in Säily 2011: 126).

2.3.3. Lexical statistics and productivity

As was shown with the example of type accumulation curves, the study of productivity can benefit from what is usually called *lexical statistics*, i.e., the study of word frequency distributions and the probabilities of occurrence of different types across texts and corpora (see Baroni 2008: 803). How different word types are distributed in texts usually follows some pattern. For example, it was pointed out above that productive word-formation processes are characterized by a large number of low-frequency words, such as hapax legomena or *dis legomena* (words that occur twice in a corpus). Interestingly, samples consisting of naturally occurring text, independently of size, language, or textual typology, are characterized by a similar pattern of few very frequent words (typically function words such as *a, the, of,* etc.) and a long tail of low-frequency types (Baroni 2008: 810). A statistical model like this is called a "large number of rare events" (LNRE) distribution (Baayen 2001: 51).

Such a skewed distribution with "few giants ... and an army of dwarves" (Baroni 2008: 814) can be predicted by *Zipf's law* (Zipf 1949, 1965).[15] The model shows how frequency is a non-linearily decreasing function of the rank of a word in the text. For example, the most frequent word in the text has the rank 1, the second most frequent word has 2, and so on. Zipf's law is formalized as follows:

$$f(w) = \frac{C}{r(w)^a}$$

where *f(w)* is the frequency of a word *w*, and *r(w)* the rank of the same word, while *C* and *a* are constants that depend on the corpus (Baroni 2008: 813). If *a* is assumed to be 1, then *C* is the frequency of the most frequent word in a corpus (having thus rank 1). If the most frequent word occurs 3000 times in the corpus, then *C* = 3000. The second most frequent word then has a frequency of *C/2* = 1500, the 100th most frequent word *C*/100 = 3, and so on. Thus the frequencies of words decrease rapidly among the high ranks, but slows down as the rank decreases.

[15] Due to the limited scope of the study, the deep mathematics underlying the Zipf-Mandelbrot model are discussed here only cursorily. A more thorough account of the mathematical motivation of the model is given by Baayen (2001).

Even though Zipf's law is sound in theory, it does not provide a perfect fit for expected values of word frequencies (Baroni 2008: 815). A better fit is obtained when the extra parameter *b* is added to the formula:

$$f(w) = \frac{C}{(r(w) + b)^a}$$

The resulting formula is called *Zipf-Mandelbrot's law,* of which the original Zipf's law is a special case with $b = 0$ (Baroni 2008: 815). The optimal parameters can be determined with the help of a computer. The zipfR package, available for the R software, has been used in this study (see Evert and Baroni 2007).

Table 1. A toy rank/frequency profile (left panel) and frequency spectrum (right panel)

r	f
1	13
2	7
3	5
4	2
5	1

f	V(f)
1	10
2	5
3	3
4	2
5	1

Baroni mentions two useful tools that help capture word frequency distributions in a visual way: *rank/frequency profiles* and *frequency spectra* (2008: 806). A rank/frequency profile resembles a type frequency list on which the types in a corpus and their respective frequencies are placed in a descending order. In a rank/frequency profile, the type labels are replaced by their frequency-based ranks, 1 referring to the most frequent type and so on. A frequency spectrum, on the other hand, reports how many types have a certain frequency. Table 1 offers an example of a rank/frequency profile and a frequency spectrum

which can be seen as reverse sides of the same phenomenon. A more visual way to illustrate those two frequency distributions is to present them as a bar plot graph. An example is offered in figure 3, which draws from data for the present study.

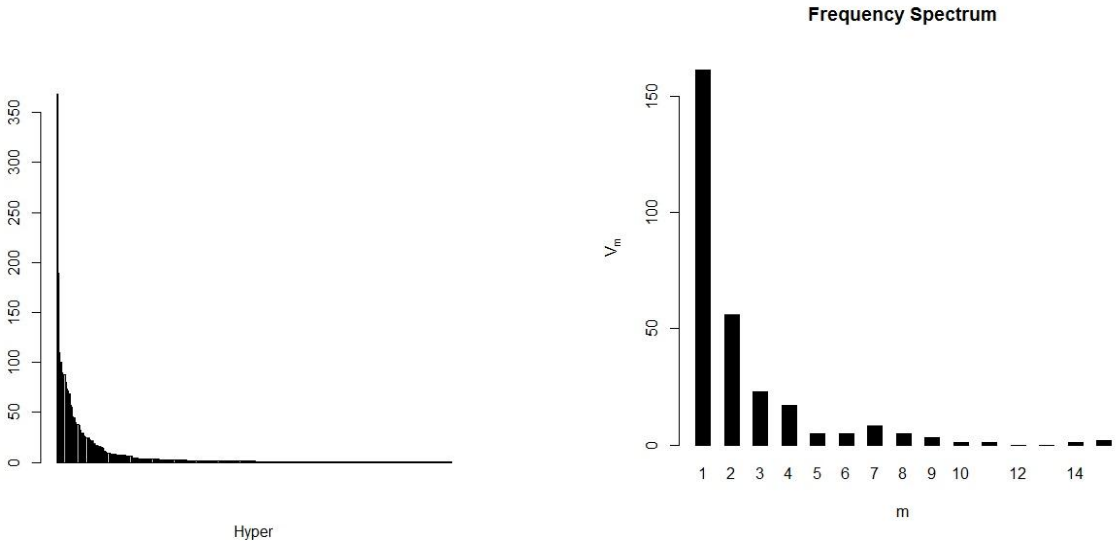

Figure 3. A rank/frequency profile and a frequency spectrum for the initial combining form *hyper-* in the written part of the *BNC.*

What makes rank/frequency profiles and frequency spectra interesting from the point of view of productivity is that if we know the distribution of type probabilities of a certain word-formation process, the expected vocabulary size *V* and the number of hapax legomena V_1 can be calculated for any sample size *N* (Baayen 2001: 41–51). Since potential productivity *P* essentially measures the vocabulary growth rate of different processes and consequently the slope of the growth curves for that process, comparing the slopes will give us information about the differences in productivity between different processes. The vocabulary growth of a process can be visualized by *vocabulary growth curves*, which are essentially the same thing as the type accumulation curves (Säily and Suomela 2009, Säily 2011), displaying the type frequency *V* of a certain word-formation process as the function of its token frequency *N.* An exemplary vocabulary growth curve is presented in Figure 4.

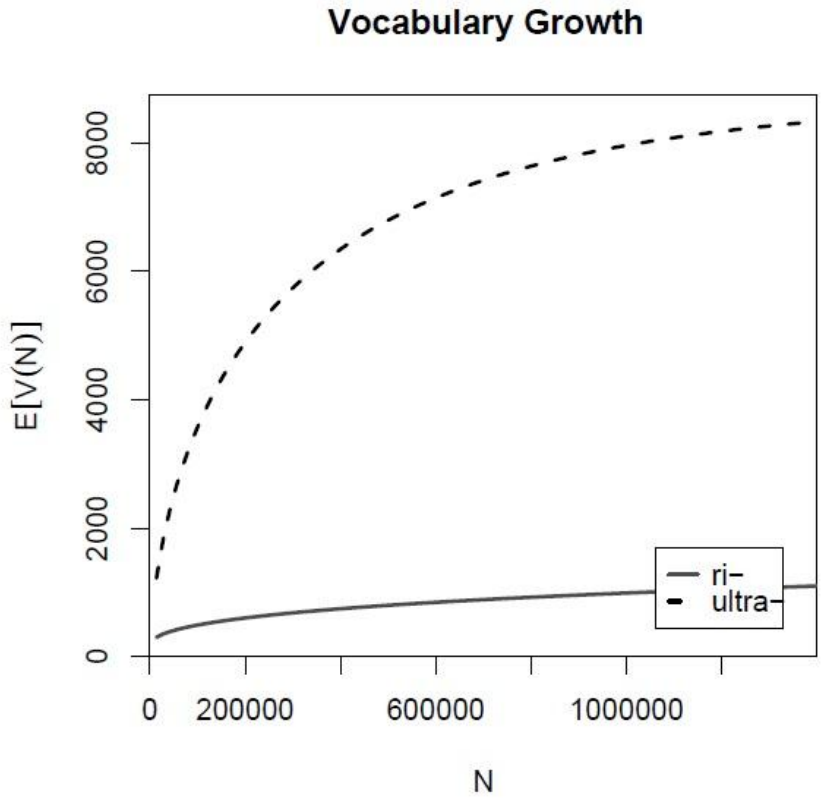

Figure 4. Vocabulary growth curves for the Italian prefixes *ri-* (interpolated) and *ultra-* (extrapolated). The x axis displays the corpus size *N* and the y axis the number of types *V* in relation to *N*. From Evert and Baroni 2007.

However, since the vocabulary does not grow linearily, direct comparison of differently sizes samples, i.e., different word-formation processes, is not reliable. In the case of samples of different sizes, it is therefore necessary to extrapolate the growth curve of the smaller sample and predict the growth rate for an arbitrary sample size. This involves computing the expected sample size for a smaller corpus in order to "stretch" it to the length of a larger corpus (Baroni 2008: 820). This can be done with the zipfR package for the R statistics software. [16]

Figure 4 (from Evert and Baroni 2007) illustrates the difference between the productivity of two Italian prefixes. The prefix *ultra-* has a smaller population that the prefix *ri-*, but when the sample is extrapolated to equal the size

[16] For a more detailed account of the zipfR package and its implementation, see Evert and Baroni 2007.

of *ri-*, and the resulting vocabulary growth curve is compared to that of *ri-*, it can be seen that it in fact has a steeper curve. This tells us that it has a higher growth rate, i.e., a higher *P* value. In general, it is typical for a vocabulary growth curve to grow more steeply at lower sample sizes, because if a corpus is small, it is more likely to encounter new, unattested types. As the corpus grows, the likelihood decreases.

2.3.4. The psycholinguistic aspect of productivity

Attempts to measure degrees of productivity constitute only one aspect in the study of this phenomenon. Another issue is how speakers process, storage, retrieve and produce morphologically complex words in terms of their mental lexicon (cf. Bauer 2001: 101, 112), and what the role of productivity is in this schema. The primary interest of this study is not the psycholinguistic aspect of productivity, but it is worth discussing nevertheless, since many quantitative measures of productivity are also psycholinguistically motivated, as will be shown below.

A crucial question at this point is whether complex words are stored as single units or as distinct morphemes in the mental lexicon, while the latter way involving morphological parsing.[17] The former way is more adequate from the point of view of processing, the latter facilitates the storage of the forms (Frauenfelder and Schreuder 1992: 166–167). In Frauenfelder and Schreuder's model, each complex word is simultaneously processed in both ways, the faster route winning the race (hence the term *morphological race model*) (1992: 175). Factors that influence which route will win include token frequency (high token frequency favours the direct route) and transparency (1992: 176–177). The model is related to Baayen's statistical productivity measures by predicting that the parsing route wins with transparent, infrequent words – the very properties of words that are assumed to be of high productivity in Baayen's model.

Relying on Frauenfelder and Schreuder's work, Baayen (1993) discusses the *activation level* A, a measure that attempts to determine which route

[17]Parsing has several different (although related) senses. In this context, by parsing one means the way speakers analyse complex words into their constituent parts (see, e.g., Frauenfelder and Schreuder 1992: 175).

to choose in the race. The activation level of a word-formation process is formalized as the number of tokens containing that process with a frequency less than a threshold θ (1993: 196). If the frequency is higher than the threshold, the word is processed by the direct route, and while words with a lower frequency are parsed (1993: 196). For example, in the present data, the activation level for *pseudo-* would be 491, if θ = 8 (the same as Baayen uses). Thus, all types with a frequency less than 8 are parsed. Baayen does not tell us how exactly the threshold θ should be determined, but he writes that if the value of θ is high, more opaque forms will be included in the frequency counts (1993: 203).

Baayen's model is further improved by Hay (2001, 2003), who states that instead of the absolute frequency of a word, the degree of decomposability of complex words depends on the relative frequency of the derived word and the base, i.e., the frequency of the derived word divided by the frequency of its base. For example, the derivative *business* is more frequent than its base *busy,* which means that it has a high relative frequency and that it is accessed as whole (Plag 2006: 548). By contrast, words with a low relative frequency are often morphologically productive and semantically transparent. Plag summarizes that the semantical transparency of productive word-formation rules is the consequence of cognitive processing (2006: 548). Hay and Baayen (2002) attempt to further refine the model by defining the specific parsing line above which derived words are most likely to be parsed.

Neither Baayen's activation level nor Hay's relative frequency model will be applied in this study as such. However, psycholinguistic evidence might prove useful in determining the appropriate units of analysis, i.e., how to deal with pseudo-affixed words. Baayen actually offers some guidelines on how to determine appropriate units of analysis when studying productivity (1993: 200–201). Questions like this will be further discussed in section 4.

2.3.4. Constraints on productivity

A word-formation pattern cannot apply to any word in a language, but only to a certain subset of words (Rainer 2005: 335). In other words, there are various factors that restrict the potential combinations of affixes. It has even been proposed that all morphological rules are equally productive, so that the

differences in (apparent) productivity are caused by restrictions on the rules. However, there are many rules that do not seem to be restricted in any general way, which corroborates this idea.[18] The study of the various restrictions on word-formation processes is related to the qualitative aspect of productivity.

The subset of possible bases for a word-formation process is called a *domain* (Rainer 2005: 335). This phenomenon can be called a restriction, but most scholars use the term *constraint* instead. Bauer, for example, justifies the use of *constraint* by the fact that the restrictions are not always absolute (2001: 126). There are several ways to classify different kinds of constraints. Rainer, for example, divides them into *universal* and *language-specific* (2005: 335). The former apply to all languages, while the latter depend on the language in question.

Another division of constraints is often made between *external factors* (factors that are mainly related to fashion, knowledge, and attitudes) and *internal factors* (caused by the internal structure of a language) (Férnandez-Domínguez 2007: 69). I will mainly follow Rainer's model, although my focus will be on the language-specific constraint relevant for the English language and especially on the various structural constraints that comprise what Férnandez-Domínguez calls "all the levels of linguistic description" (2007: 69). All the structural constraints (i.e., phonological, morphological, syntactic, semantic, pragmatic) are thus language-specific, and the only constraints treated here that fall under the category of universal constraints are the various types of blocking.

Phonological constraints on morphological compatibility are a very common type, illustrated by phonologically conditioned allomorphs, such as the complementary set of plural markers /iz/, /s/ and /z/ (Bauer 2001: 128). Bauer mentions three properties of the base that may constrain what affixes may attach to it (2001: 129–130). Sometimes the segmental make-up of the base is important: the verb-forming *-en* only attaches to bases that end in a stop or a fricative). In other cases the suprasegmental make-up, i.e., stress, is what matters: the noun-forming suffix *-al* is only attached to bases that have the stress on the first syllable. Sometimes the number of syllables in the base determines which affixes may attach to it. As regards the third case, Bauer does not mention any examples

[18] For discussion, see Haspelmath (2002: 111).

in English. Phonological restrictions are a common example of language-specific constraints.

With *morphological constraints* on productivity, the morphological structure of the base determines what kinds of affixes may attach to it (Bauer 2001: 130). Again, Bauer lists three options (2001: 130–132). Some affixes apply only to a certain morphologically defined set of bases. The set can be defined in terms of etymology, for example. This is often called *level ordering* (see below for discussion). Sometimes a certain affix may be added only to a derived or to an underived base. The base may also have to end in a particular affix before any further affixes can be added. In general, prefixes have fewer word-class restrictions than suffixes, because they do not usually alter the word-class of the base they attach to (Nevalainen 1999: 355).

The distinction between morphological and *syntactic constraints* is not always clear-cut. Bauer draws the line by defining morphological constraints as ones that concern the internal structure of the word, while syntactic constraints deal with the way a word is used in context (2001: 133). A typical example of a syntactic constraint is the fact that many affixes can only be added to a certain word class as a base (Bauer 2001: 133). Sometimes the transitivity of the verb base is of importance: for example, the suffix *-able* may only be used on transitive verbs.

Level ordering is one of the language-specific constraints on productivity. Level ordering arises because many languages have several morphological strata, a native and a non-native one, which both have some combinatory restrictions (Rainer 2005: 340). For example, in English, many Latinate affixes are restricted to Latinate bases only (Haspelmath 2002: 108). However, since speakers do not necessarily know whether a stem belongs to the native or non-native stratum, the restriction on the basis of morphological strata is sometimes unstable: the suffix *-ous* has been added to some native bases as well (*murderous, thunderous*) (Haspelmath 2002: 108).

Finally, pragmatic and sociolinguistic factors may also restrict the productivity of word-formation processes. While semantic constraints depend on the linguistic nature of the base, *pragmatic constraints* (as opposed to structural constraints) involve the way in which words are actually used (Bauer 2001: 135). Certain patterns are only used in particular situations, and it requires language

competence to know in which situations to use which patterns (Rainer 2005: 349). One of the clearest cases of pragmatic restrictions is fashion and the "extra-linguistic developments in the society that make certain words or morphological elements desirable to use" (Plag 1999: 39).

Another pragmatic constraint is the general need for a certain word (Plag 1999: 39). The need may either relate to the naming or labelling of a new concept or entity, or to the speakers' need to condense information due to stylistic purposes or text cohesion. Yet another, a rather closely related restriction is the nameability requirement: all new lexemes must denote something nameable in the language (Plag 1999: 40).

Semantic constraints are not very common, and are mainly concerned with cases where the referent of the base must follow certain semantic requirements. For example, the referent of the nouns with the suffix *-ee* must be sentient (Barker 1998; cited in Rainer 2005: 349). Semantic restrictions are a common type of constraint, *semantic blocking* being one example. Carstairs-McCarthy defines semantic blocking as a situation in which "the existence of a word (...) with a particular meaning inhibits the morphological derivation, even by formally regular means, of another word with precisely that meaning" (2002: 91). His example makes use of the adjectives *curious* and *glorious.* A formally regular way to form nouns from these adjectives would be to employ the suffix *-ity.* With *curiosity,* this seems to work, the result being *curiosity.* However, with *glorious,* the corresponding noun is *glory,* not **gloriosity*: the existence of *gloriosity* is thus blocked by the word *glory.* Even a formally regular word-formation process can be blocked (ibid.). One reason for the existence of the phenomenon of semantic blocking is the tendency to avoid exact synonymy in most languages (Carstairs-McCarthy 2002: 85).

Rainer calls this kind of blocking *token blocking,* and writes that children and non-native speakers often produce formations like *gloriosity,* which indicates that the "existence" of a synonymous word refers to the mental lexicon of the speaker, not to language as a social institution (2005: 336–337). Besides token blocking, Rainer mentions *type blocking,* in which case the unacceptability of a complex word is due to a synonymous pattern that takes precedence (2005: 337–338). He illustrates type blocking by the set of rival suffixes in German, namely *-heit, -ität,* and *-ie* (2005: 338). The suffix *-heit* is used with adjectives

with final stress, and it is fully productive. However, its domain is diminished by *-ität* (with learned bases) and, more rarely, *-ie* (with adjectives ending in *-phil).* Férnandez-Domínguez defines both type blocking and token blocking as examples of *synonymy blocking*, as opposed to *homonymy blocking*, which prevents the coining of a new word because of structural overlapping with an existing word (2007: 71–73). For example, the word *liver* 'inner organ' prevents the coining of **liver* 'someone who lives' (Plag 1999: 50; cited in Férnandez-Domínguez 2007: 73).

To conclude, it can be said that new words are always formed for a specific need, and thus constraints may not tell us everything about the profitability of word-formation processes (Bauer 2001: 143). Férnandez-Domínguez goes as far as to state that the naming need on the part of the language community is so strong that it may even "knock out" the restrictions caused by various constraints (2007: 84). In any case, the application of different word-formation processes is seldom straightforward, and there are several exceptions to the rules.

3. Material and methods

3.1. A closer look at the selected combining forms

As was stated earlier, combining forms can be divided into initial and final combining forms. In this study, I will compare the two and try to find out whether they differ in productivity. To do so, I will compare the values of Baayen's potential productivity *P* of three initial combining forms (*hyper-*, *ultra-*, and *pseudo*) as well as three final combining forms (*-logy*, *-graphy*, and *-nomy*), in order to determine whether there are any major differences in the productivity rates between these two subtypes. These six combining forms were chosen because they are among the most frequent ones in the *BNC.* Sample size is important in performing statistical analyses on the combining forms, since many of the measures used in this study do not work properly if the sample is too small (see section 2.3.3. for discussion).

It must be noted that some neoclassical elements have two kinds of combining forms (with allomorphic variation) derived from the same lexical word, those that can occur in both initial and final position (cf. Fradin 2000: 13). For example the Greek *λόγο-ς* 'word, speech' occurs in words like *logopedics* and *biology*. A similar example is the Greek form *-γράφος*, which is used in formations like *graphology* and *biography*.

Unless otherwise indicated, the *OED* constitutes the source in the following description of the selected combining forms. The problem with the *OED*, however, is that even though some of the selected combining forms, the final ones in particular, have rather sketchy entries. McCauley writes that the amount of information in each entry is to some extent determined by the frequency, longevity, and complexity of the word or element in the English language (2006: 99). In addition, most entries are rather old and date back to the late 19th century (see section 3.2. for discussion). However, the *OED* is currently undergoing a thorough revision that will also affect entries for combining forms: for example, in the second edition of the *OED*, there is sometimes ambiguity in whether a certain element is an affix or a combining form (McCauley 2006: 97).

hyper-

Hyper- is often classified as a prefix instead of a combining form, probably because it has its origin in the Greek preposition (*ὑπέρ* 'over, beyond, over much, above measure') instead of a lexical word. Nevalainen classifies *hyper-* as an intensifying prefix, and writes that it is the Greek cognate of Latin *super-*, having the meaning of 'over, too much' (1999: 388). The *OED* states that *hyper-* has been extensively used in English since the 17th century in the formation of new compounds (often on Greek analogies), and that it combines relatively freely with nouns and adjectives.

pseudo-

Pseudo- has its origins in the Greek root ψευδο- 'false, falsity, falsehood', and is also quite frequently attached to Latin bases as well. The first attestations of formations with *pseudo-* in English date back to early 15th century in a religious context, but it did not become common until the 17th century. The first native formations (e.g., *pseudo-prophetical*) date back to that period as well. From the 19th century onwards, *pseudo-* combined with nouns and adjectives becomes frequent in scientific terminology in particular.

ultra-

Unlike the majority of the combining forms in this study, *ultra-* has its origins in Latin instead of Greek. It is derived from a Latin preposition that has a meaning 'beyond', and like *hyper-*, it is often classified as a measurement prefix as well. The original meaning in English is 'lying spatially beyond or on the other side of' (as in *ultra-terrestrial, ultra-zodiacal*), or 'going beyond, transcending the limits of' (*ultra-human, ultra-rational*). However, from the early 19th century onwards, a third sense, 'an excessive or extreme quality of an adjective' (*ultra-cautious, ultra-cosmopolitan*) has steadily become more common (Hughes 2000: 346–347).

-logy

The origin of *-logy* is the Greek element *λόγο-ς* 'word, speech' or the corresponding root *λόγ-*.[19] Derivatives of logy have two distinct senses in English. The former group includes words with the sense 'saying or speaking' (*eulogy, necrology*), while the latter sense can be seen in the names of sciences or disciplines (*biology, astrology*). All the modern formations of *-logy* can be said to have a corresponding formations using the combining forms *-logical* and *-logist* (called suffixal expansion by Prćić (2007: 386–387). Kastovsky states that *-logy* is actually much more suffix-like than lexeme-like (2009: 5).

[19] Sometimes the thematic vowel *-o-* is included by dictionary-makers, which results in the form *-ology* in English The status of the linking element *-o-* is discussed in section 2.2.2.

-graphy

The form *-graphy* is closely related to the variant *-graph.* However, formations with *-graph* are not included in this study. The combining form *-graphy* is derived from the Greek –γράφος, which denotes writing. Often the words that have the component *-graphy* refer to sciences (*geography, bibliography*), but they can also denote writing, drawing, or graphic representation (*calligraphy, lithography, photography*). Hybrid formations with *-graphy* are rare, *stratigraphy* being one of the few examples. As with *-logy,* words with *-graphy* can have the corresponding derivatives *-graphic* and *-graphical.*

-nomy

The combining form *-nomy* is derived from the ancient Greek element –νομία, a variant of the base νέμειν 'to deal, distribute, hold, manage'. The earliest attestations of *-nomy* are *astronomy* and *economy* (both from the 16th century). There is also a small number of words (from the 16th century onwards) that denote 'law', 'rule', or 'government', as in *autonomy.*

3.2. Corpora versus dictionaries in the study of productivity

It was stated above that the most widely-used methods to measure productivity are either corpus-based or dictionary-based. Both types of methods have their advantages and disadvantages. Corpora are generally considered to be a more representative sample of language than dictionaries, but the issue is not so straightforward, as will be shown.

Historical dictionaries in particular are often partial or even incomplete, even though they are useful in providing information about the early attestations of lexemes (Palmer 2009: 8). Nevalainen mentions the various problems with the *OED* as regards chronological study of words. For example, issues such as chronological coverage are often ignored (1999: 337). Therefore, the citations in the *OED* do not necessarily reflect the full variety of texts in a

given period. For the earliest stages of language, the sampling is often sparse, and a word may have been in use long before its first written record (Baayen 2008: 909–910). In a similar vein, Schäfer writes that the *OED* reflects the frequency and status of a word at the moment of compilation of the dictionary instead of the period of origin (1989: 69).

According to Baayen and Lieber, some of the benefits of corpora as opposed to dictionaries are that they reflect actual language usage (1991: 803). Indeed, as Baayen and Renouf point out, "it is commercially unattractive to print thousands of words the meaning of which is immediately clear to anyone familiar with the basic meaning of productive affixes" (1996: 69). According to Booij, dictionaries only register which new complex words have been established in the lexicon, while true morphological productivity is reflected by "complex words that never make it into dictionary", i.e., hapaxes (2005: 69). In a similar vein, Baayen and Renouf state that dictionary-based methods may underestimate the productivity rates of affixes (1996: 70). In addition, the inclusion of a word into a dictionary is always arbitrary to some extent, which means that a dictionary cannot be considered as a representative sample of a language (Plag 1999: 27).

The size of the corpus is an important factor in all kinds of linguistic analysis, and in the study of productivity in particular: in a small corpus, most hapax legomena will be well-known words of the language, but if the corpus size increases, the proportion of neologisms among the hapaxes increases as well (Plag 2006: 542–543; see also Baayen and Lieber 1991: 813). For unproductive processes, increasing the corpus size does not add new types after a certain "saturation point" has been reached (Lüdeling et al. 2000: 58). In this vein, the use of the *BNC* is justified, since it is one of the largest English corpora available. Other corpora generally considered large enough include the COBUILD / Bank of English corpus and the COCA corpus of contemporary American English (see Plag 1999: 25–26). Historical corpora, on the other hand, are often too small for studying potential productivity *P* (Palmer 2009: 8–9).

The data for this study was obtained from the written part of the *The British National Corpus* (*BNC*). In total, the *BNC* comprises just under 100 million word tokens of which roughly 89 million words represent written British English that falls into various genres, or registers, such as fiction, academic prose, and personal letters (see Burnard 2007). One of the possible disadvantages with

the *BNC* is the fact that since it does not cover the last twenty years or so, the most recent changes or trends in the English word-formation are not covered. However, in this study, the use of the *BNC* is justified, because it has already been used in many studies on the productivity of various affixes. This enables the comparison of the productivity rates to the ones calculated for the selected combining forms. When necessary, and especially to provide supplementary information on the behaviour of these combining forms, the *Corpus of Contemporary American English* (*COCA*) will also be consulted, especially in the discussion section.

It must be noted, however, that a corpus-based productivity measure is not devoid of problems, either. Lüdeling et al. criticize Baayen for assuming that all electronic corpora are automatically perfectly prepared (2000: 59). There are, however, at least three corpus pre-processing problems that must be taken into consideration: mistagged items, typographical errors, and repetition of the same texts in corpus composition. Similar problems are even more common in historical data, since historical corpora often contain several spelling variants of the same affixes (see Säily and Suomela 2009: 93). Tagging is not a problem in the present study, since we are only interested in the internal make-up of lexemes. Even though the *BNC* has been claimed to have some inherent structural problems, such as wrongly classified texts and overly broad categories (see Lee 2001), the sampling has been done carefully (see Burnard 2007: section 1.4.4 for the selection procedures employed), so that the repetition of texts is highly unlikely.

Typographical errors are, however, a problem that must be taken seriously. Furthermore, errors often occur only once and consequently add to the number of hapaxes, and may therefore considerably affect the results (Lüdeling et al. 2000: 59). Plag expresses the same kind of concern, adding that some qualitative decisions are necessary even in purely quantitative analysis (1999: 29). Thus, in order to properly analyse the data for the present study, a certain amount of manual sorting was necessary.

In general, corpora constitute a better source for the purposes of the present study. However, as has been shown, corpora as a source are far from unproblematic. Even the largest, most balanced and representative corpus will never equal real language use, even though it may provide a useful tool to make

certain estimations or predictions based on statistics. As Baroni and Evert have put it, "corpora are finite samples from the infinite sets that constitutes a language" (2008: 777). These samples let us make generalizations and inferences, but this requires that the problem at hand is accurately operationalized in quantitative terms (Evert and Baroni 2008: 777–778). A good strategy in the study of productivity is to combine corpus and dictionary evidence to provide a comprehensive view of the productivity of different patterns (see Plag 2006).

3.3. How to determine units of analysis and other methodological problems

One has to be careful when making inferences based on corpus evidence. Lüdeling et al. claim that even seemingly error-free corpora may cause trouble in this respect: for example, words may accidentally contain strings of characters that look like affixes but are not (2000: 59). For example, not all words that begin with *sub-* actually contain the prefix *sub-* in contemporary English (e.g., *subtle*), even if they might be derived from a word that was morphologically complex in the donor language (Latin *sub* 'under' + *tēla* 'web') (Plag 2006: 543). In psycholinguistics, this phenomenon is referred to as *pseudo-affixation* (see, e.g., Baayen 1993: 199). Another good example of pseudo-affixation is the status of the constituent-*fer* in such words as *infer*, *confer*, *prefer* and *transfer* – it might be analysed as a bound root, but it does not carry a meaning that would be the same in all these words (see Plag 2003: 24–25). In the context of the present study, examples of words that seem to consist of several elements but that actually are mono-morphemic include *hyperion* and *hyperoid.* Therefore, they are excluded from the type count.

Another similar issue is the fact that many affixes have adopted idiosyncrasies that may blur their morphological make-up: for example, words that end in *-ity* occur in many transparent, complex words, but also in words like *entity*, which might not be considered complex (see also Plag 1999: 28). In a generous count, this word would probably be counted as a unit of analysis and included in the type count, but in that case it would skew the productivity rate of a modern word-formation process.

Fortunately, the combining forms studied in this paper are quite distinctive in form, at least compared to affixes in the more traditional sense. However, several qualitative decisions have to be made concerning which elements to include and which to exclude from the type count. For example, Prćić points out that *-logy* and *-holic* can be analysed as combining forms in such words as *morphology* and *foodaholic*, but not in *apology* and *alcoholic* (2007: 382). In a similar vein, Warren suggests that *hyper-* can be considered a prefix in *hyperactive* but a combining form in *hypertrophy* (1990: 113).[20]

However, since it is not likely that speakers are aware of such a distinction, formations such as *hypertrophy* are not excluded from the data. Baayen points out that if an affix has several different, although related, senses, the goodness-of-fit of the word frequency distribution might be lacking (2001: 145). How the decision not to exclude the "pure neoclassical compound" instances of the selected combining forms affects their productivity values and the word frequency distributions will be assessed in the Discussion section.

One of the problems of Baayen's potential productivity index *P* is that none of the individual studies that apply the measure explains how the units of analysis are determined and how the data is treated in the study in question – this makes it difficult to compare the productivity rates between different authors, even though they used the same data (see Plag 1999: 29). This is even more problematic with combining forms (at least if they are to be considered as a word-formation process distinct from affixation), since methods associated with affixation might not be fully applied to combining forms.

In his study on the productivity of N + N compounds, Fernández-Domínguez deals with problems in applying the affix-centred productivity measures to compounds in English (2009: 142). The problem with compounds is to determine which part corresponds to the base and which to the affix. The problem can be tackled by for example calculating the average frequency of the components (Férnandez-Domínguez 2009: 143). However, in this study I will rely on the affix-based approach, since native compounds differ from neoclassical compounds and since combining forms are used in a more affix-like way (i.e., they can be combined with free roots as well).

[20]In fact,Quirk et al. use the term combining form only with neoclassical compounds, and reserve the term *neoclassical affix* for the affix-like use of the same elements (1985: 1545).

Determining the appropriate units of analysis is an important issue as well. For example, whether one counts *ultra-violet* and *ultraviolet* as a single type or separate types has consequences in terms of the *P* values one obtains. Sometimes an affix can also have several distinct, although to some extent interrelated meanings (see, e.g., McConchie 1998 for the different senses of the prefix *dis-* in English). Sometimes words generated with an apparently single affix are in fact generated by several distinct word-formation processes (Lüdeling and Evert 2003). For example, the German *-lich* may attach to adjectives, verbs and nouns alike. Plag also points out that lexicalization, i.e., a process in which a complex word develops a new, idiosyncratic meaning, may cause problems of classification (2006: 543). However, the combining forms selected for this study turned out to be fairly unproblematic in this respect.

In this study, I decided to count the spelling variants of a single lexeme (*quasi-judicial* vs. *quasijudicial*) as tokens of a single type. The spelling of English compounds is by and large unregulated, so that some compounds can have as many as three different spelling variants that Férnandez-Domínguez calls open (*pseudo scientific*), solid (*pseudoscientific*), and hyphenated (*pseudo-scientific*) (2009: 32–33). In a similar vein, words that have different spellings according to the different national (mainly British vs. American) spelling standards (e.g., *palaeontology* vs. *paleontology*) are counted as a single type.

Of course, the number of types is not as relevant as the number of tokens in this particular measure of productivity. However, the type *hyper-tension*, for example, occurs only once in the *BNC*, while *hypertension* has 369 occurrences. Thus *hyper-tension* increases the productivity of *hyper-*, unless it be counted as a variant of *hypertension.* For a more detailed discussion of counting tokens and types, see Baroni (2008: 804–805).

Multiply affixed words constitute another problem in determining the units of analysis. Should such word as *conventionalizable* be understood as an example of *-ize* or *-able* – or perhaps both (Plag 1999: 28–29)? The decision one makes has consequences for the productivity values and word frequency distributions of the elements selected. In the context of the present study, the problem is more relevant for final than for initial combining forms.

In the present study, I decided to count formations with a prefix or an additional initial combining form attached as separate types. Therefore, *neuro-*

physiology (with its spelling variants) and *physiology* are separate types, as well as *aero-technology* and *technology*. Another way would have been to add up the frequencies of *aero-technology* and *technology*. However, Baayen states that such cumulated token frequencies are more appropriate in estimating the activation level A than in calculating the values for *P* or *P** for at least two reasons. First, it is difficult to estimate the weight of the various word-formation processes in multiply complex words. Second, if a word is assumed to be an independent free form in the lexicon, it will be counterproductive to sum up the token frequencies of these two (1993: 200–201). If a word were assigned to two morphological categories, the observations in those two categories would no longer be independent, and the statistical tests for comparing productivity rates would no longer be valid (Baayen 2008: 903). In their study on the productivity of selected nominal suffixes (such as *-al* and *-ist*), Plag et al. have chosen to exclude prefixed formations (*unavailable)* and compounds (*performance-artist)* completely (1999: 214).

It must be noted again that in some respect the rules concerning affixes cannot be fully applied to the study of combining forms. For example, it is not always easy to determine which component is the base and which one corresponds to the affix. However, it is intuitively clear that adding a prefix (*aero-*) to a neoclassical compound (*technology*) involves lexical creativity, and thus contributes to the productivity of *-logy*. On the other hand, since formations with neoclassical compounds are unclear cases in many ways, the opposite view could be justified as well.

It must also be noted that there are several final combining forms that have related forms, such as *-logical* from *-logy* and *-graphical* or *-graphic* from *-graphy* (called suffixal expansion by Prćić (2007: 386–387)). These forms will not be dealt with in this study, for the reasons mentioned above. *P* is a category-conditioned degree of productivity and is thus calculated on the basis of a single morphological category (see Baayen 1993: 200).

4. Results

4.1. Potential productivity *P*

In order to find all the instances of the initial combining forms, the wildcard * was used to include all the formations that begin with a certain sequence of letters, in this case *hyper*, ultra*, pseudo**, and *quasi**. With respect to final combining forms, the query was formed as *log[y,ies], *graph[y,ies], etc., in order to include the plural form of the selected combining forms as well, which were all noun-forming. The queries were restricted to cover only the written part of the *BNC*. This makes it easier to compare the results to those obtained in previous studies (e.g., Plag 1999, 2002, 2006).

Table 2 summarizes the type frequency *V*, the token frequency *N*, as well as the number of hapax legomena $n_1$and the values of potential productivity *P* of each combining form that was selected for this study.

Table 2. Type frequency *V*, token frequency *N*, the number of hapax legomena n_1, and the productivity rate *P* of selected combining forms in the written part of the *BNC*.

	V	*N*	n_1	*P*
hyper-	323	2,859	163	0.0570
ultra-	317	1,628	188	0.1155
pseudo-	325	818	231	0.2824
-logy	375	36,077	111	0.0031
-graphy	166	7,440	57	0.0077
-nomy	46	14,240	20	0.0014

In terms of type frequency *V*, the initial combining forms generally seem to rank higher than the final combining forms. The only exception is *-logy*, which has the highest type frequency of all the selected combining forms. On the

other hand, the final combining form *-nomy* has a particularly low type frequency, when compared to the others. It must be noted that the difference in type frequencies between initial and final combining forms would be even more drastic if multiply-prefixed words were counted as a single type. In this study, such words as *aero-technology* and *gastro-technology* (both hapaxes) have been counted as separate types, not as occurrences of *technology*, and thus the type frequency (and consequently the number of hapax legomena) is higher than with a less generous count.

Some tentative conclusions can be already made by looking at the type frequencies of the selected combining forms, since they clearly measure some aspect of productivity. As has been stated earlier, this aspect has been called profitability, extent of use, or realized productivity. Thus, at this point, one could suggest that *-logy* has a high extent of use, while with *-nomy* the extent is not so substantial. The three initial combining forms and *-graphy* are situated somewhere in between on the scale.

As regards the token frequency *N*, there is more variation, and again some interesting tendencies can be observed there as well, even though token frequency in itself has not been generally considered as a very good direct measure of productivity (see, e.g., Bauer 2001: 147). The most frequent combining form is *-logy*, with 36,077 occurrences, which is more than twice as frequent as the next frequent combining form, which is *-nomy* with 14,240 occurrences. Both have some extremely frequent types that contribute greatly to their total token frequencies. For example, *technology* has 11,353 occurrences, which equals more than 30 percent of the total number of tokens. If formations with *technology* plus a prefixal element or another initial combining form attached (such as *aero-technology* or *multi-technology*, both hapaxes) are counted as a single type, the total token frequency is even higher. As regards *-nomy*, the relative weight of the most frequent type (*economy* with 9,997 occurrences) is even more substantial (approximately 70 percent of all the tokens with *-nomy*). In general, the initial combining forms have far lower token frequencies than the final combining forms.

In terms of the number of hapax legomena n_1, the initial combining forms display much higher frequencies than the final combining forms. The hapaxes also include many lexemes that are not attested in the *OED*. These

include such items as *hyperalertness, pseudo-Scottish,* and *ultra-Zionist*. As for the final combining forms, such hapaxes include *ghettology* and *dreamography.* There are at least two reasons why they are not listed in the dictionary. Firstly, they are quite rare in the language, and secondly, they are semantically rather transparent. For example, *hyperalertness* need not be listed in a dictionary, because *alertness* already has an entry, and because its meaning can be constituted from the meaning of its parts. This can be seen as another indicator for the productivity of the combining forms examined, since productivity and transparency are at least indirectly linked (see Dalton-Puffer 1994: 252–253, Pounder 2000: 134). As was stated earlier, transparent, infrequent words are parsed, instead of being accessed as a whole (Frauenfelder and Schreuder 1992). The only exception in this group of words is *-nomy,* since all hapaxes with *-nomy* have an entry in the *OED.* Besides, most of them are scientific words, such as *chirognomy* and *aeronomy.*

At this point, it might be justified to examine the ratio of the number of hapaxes to the overall type frequency of each combining form. This ratio has not been used in any of the previous literature as a measure of morphological productivity, but it might offer some tentative insights on the relative importance of hapaxes. In this sense, it certainly describes some aspect of productivity by being indirectly linked to potential productivity *P*.

Table 3 displays the proportion of hapax legomena out of the total type frequency for each combining form. The ratio is higher with initial combining forms, being over 50 per cent in all the cases. The final combining forms do not seem to lag very far behind, however. The table can help make predictions about the *P* values of the combining forms, which will be discussed next.

Table 3. The ratio of hapax legomena n_1 to the type frequency V of selected combining forms in the written part of the *BNC.*

	n_1 / V
hyper-	0.5046
ultra-	0.5931
pseudo-	0.7108
-logy	0.296
-graphy	0.3434
-nomy	0.4347

The last column in Table 2 presents the *P* values for the selected combining forms. Again, the initial combining forms score higher. The lower *P* values of the final combining forms are partly explained by the presence of those few very frequent types mentioned above. With the initial combining forms, such extremely frequent types are nonexistent. For example, with *hyper-*, which is the least productive initial combining form as regards potential productivity *P*, the most frequent type is *hypertension* (368 occurrences) and with *ultra-*, *ultraviolet* (332 occurrences). These are only a fraction of the corresponding numbers for the final combining forms.

What can be said about the overall productivity of the selected combining forms? The problem with interpreting the results is that the *P* values alone cannot give us any definitive answers. Baayen and Lieber emphasize the fact that *P* is only a relative measure and cannot therefore in itself be used to determine whether a particular word-formation rule is productive or not (1991: 816). The method is statistical, but it does not give any threshold values that would indicate the boundaries between productive and unproductive processes. As Plag has noted, *P* only provides "a continuum of more or less productive processes", so that it is difficult or even impossible to pin down where exactly the productive processes end and the unproductive processes begin (1999: 33).

Baayen and Lieber suggest that in order to decide whether a process is productive, its *P* value should be compared to the *P* value of a relevant set of

simplex words (1991: 816). However, calculating the *P* value of, say, an inflectional process would be beyond the scope of this study. Another way to estimate the productivity of the selected combining forms is to compare it to a set of words formed with another word-formation process, such as suffixes as opposed to prefixes, or vice versa. Plag, for example, provides the *P* values for seven different suffixes in English (2006: 545, based on data from Plag et al. 1999 and Plag 2002). The data come from the written part of the *BNC* as well (see Plag et al. 1999: 210), so the figures should be more or less comparable to those of the present study, even though Plag does not offer very rigorous details about the basis on which he has determined the units of analysis (see section 3).

Table 4. Type frequency *V*, token frequency *N*, the number of hapax legomena n_1, and the productivity rate *P* of selected suffixes in the written part of the *BNC*. From Plag (2006: 545).

	V	*N*	n_1	*P*
-ness	2,466	1,369,116	943	0.061
-ion	2,392	371,747	524	0.0338
-ity	1,372	106,957	354	0.0096
-ist	1,207	98,823	341	0.0088
-less	682	28,340	272	0.0036
-ish	491	7,745	262	0.00092
-wise	183	2,091	128	0.00038

Table 4 presents the type and token frequencies as well as the number of hapax legomena and the potential productivity values of selected suffixes in the written part of the *BNC*, calculated by Plag (2006: 545). The suffix *-wise* has the highest *P* value (0.061), and *-ion* the lowest (0.00038). If these values are compared to those of the combining forms that were examined in this study, it can be seen that the suffixes have significantly lower *P* values than the combining forms do. Even the least productive prefix, *hyper-*, (*P* = 0.0570) has a

somewhat higher *P* value than the most productive suffix in Plag's study. As for the final combining forms, they score lower than the most productive suffixes, but higher than the least productive ones. One of the disadvantages with Plag's study is that is since it does not account for prefixes, it is not possible to see whether prefixes in general would rank higher than suffixes.

To summarize, the initial combining forms seem to rank somewhat higher than the final combining forms, both in terms of the number of different types *V* and potential productivity *P*. On the other hand, the final combining forms in general have higher token frequencies *N*, which is often caused by the existence of a few very frequent types. What this section has discussed constitutes the basic productivity measures that have been established and widely adopted in the study of productivity. There is, however, a set of measures more rarely employed in studies on the productivity of different word-formation processes. These measures rely on lexical statistics and word frequency distributions, and thus illustrate a slightly different aspect of productivity.

4.2. LNRE modelling

It was stated above that Zipf(-Mandelbrot)'s law fits the word frequency distribution of naturally occurring texts reasonably well. The same holds true for productive word-formation processes as well, which makes LNRE modelling particularly useful for the study of productivity. There are several methods that can be used to obtain information on the productivity of the combining forms selected.

As was stated above, one of the inherent problems with *P* is that the proportion of hapax legomena does not grow linearily as the corpus size *N* increases. This makes it difficult to compare the *P* values between different word-formation processes, since the sample sizes for each process varies. A solution that was offered in the previous section was to compare the sets of *P* values of different word-formation processes. However, there is a more elegant way to determine the productivity of the combining forms: the zipfR package of the R statistics software (Evert and Baroni 2007) allows the prediction of the shape of

the vocabulary growth curve at arbitrary sample sizes.[21] Another method that is particularly useful in the study of productivity is calculating the expected frequency spectra for the combining forms and comparing them to the observed values. This allows us to assess the goodness-of-fit of the statistical model, i.e., how well it fits the data. Therefore, it can help us detect possible outliers, i.e., data points that deviate significantly from the rest of the sample.

Table 5 presents the observed and expected *P* values for each combining form. For *-logy*, the values are identical, since *-logy* has the highest token frequency and therefore serves as the "base" against which the other combining forms are compared. For the other combining forms, the differences between the observed and expected values vary, depending on their sample size. In addition to their *P* values, the ordering of the combining forms after the extrapolation operation changes as well. However, *pseudo-* has the highest *P* value in both cases, and *-nomy* the lowest.

Table 5. The observed and expected *P* values for the selected combining forms in the written part of the *BNC*.

	P (observed)	*P* (expected)
hyper-	0.0570	0.0052
ultra-	0.1155	0.0047
pseudo-	0.2824	0.0347
-logy	0.0031	0.0031
-graphy	0.0077	0.0028
-nomy	0.0014	0.0007

[21] For a sample session illustrating the necessary operations for obtaining the expected *P* values for different word-formation processes, see Evert and Baroni 2007.

A more visual way to compare the productivity rates is to look at their vocabulary growth curves, which are shown in figure 4. A vocabulary growth curve can be used to show changes in vocabulary size *V* as the sample size *N* increases (Evert and Baroni 2007: 32). The steeper the curve, the more rapidly the vocabulary (i.e., the number of types) is growing, and the more productive a particular word-formation process can be considered. In a sense, a vocabulary growth curve works in a similar way to the potential productivity *P*, since both attempt to capture the rate at which the vocabulary of a given word-formation process grows. The only difference is that a vocabulary growth curve relies on types, while *P* makes use of hapax legomena.

As can be seen in Figure 4, *pseudo-* holds the highest place, which is shown by the steepest curve. On the other hand, *-nomy* has the lowest productivity, which is represented by a moderate, almost flat curve. In general, the initial combining forms retain the highest productivity ranks, while all the final combining forms cluster at the low end of the graph. All the slopes follow the general tendency to grow more rapidly at low *N* and then gradually decrease. At small sample sizes, it is more likely to encounter a yet unattested type, but the likelihood decreases as the sample size increases.

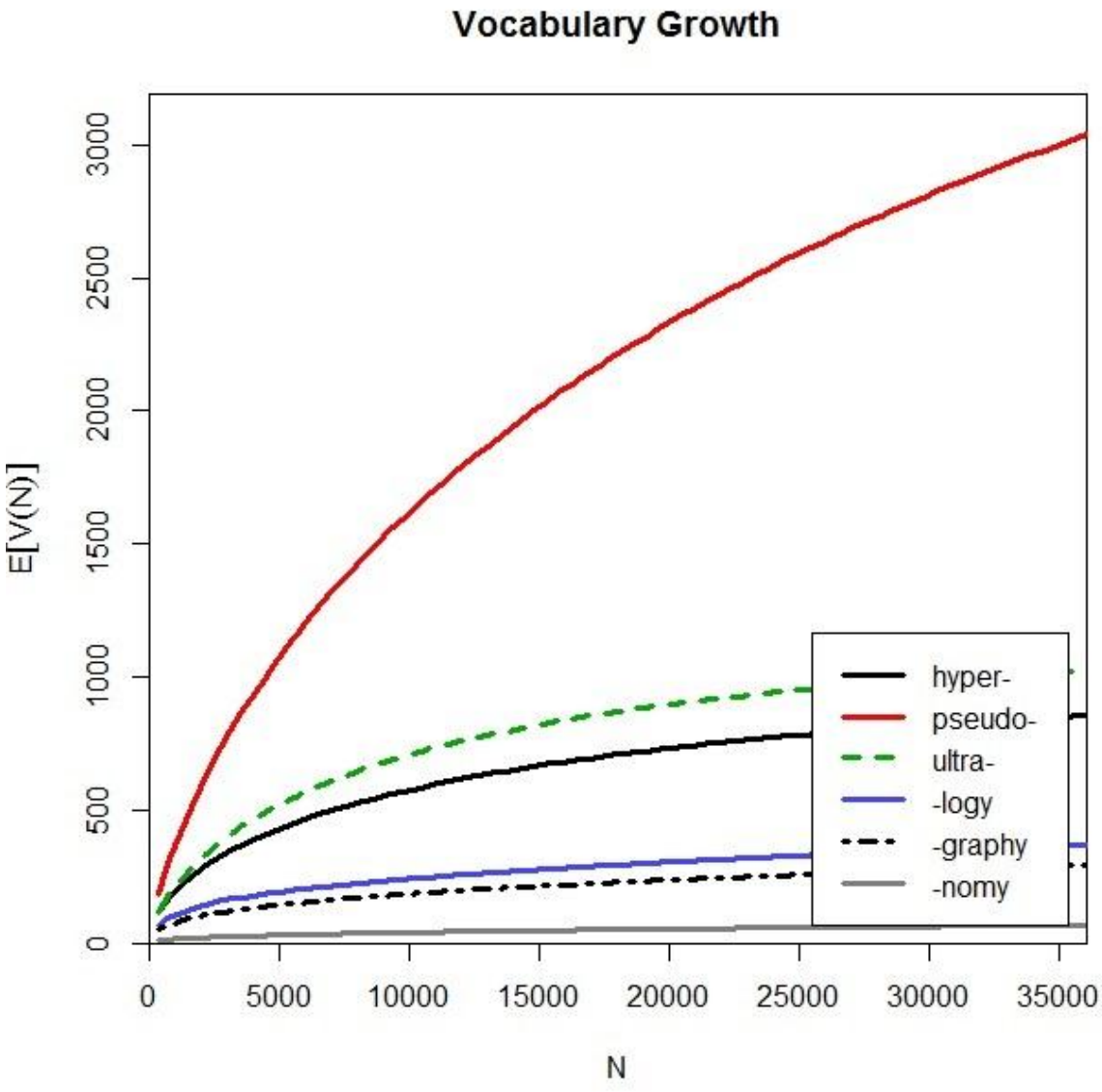

Figure 4. Extrapolated vocabulary growth curves for the selected combining forms. The x axis displays corpus size *N*, and the y axis the expected type frequencies at different sample sizes.

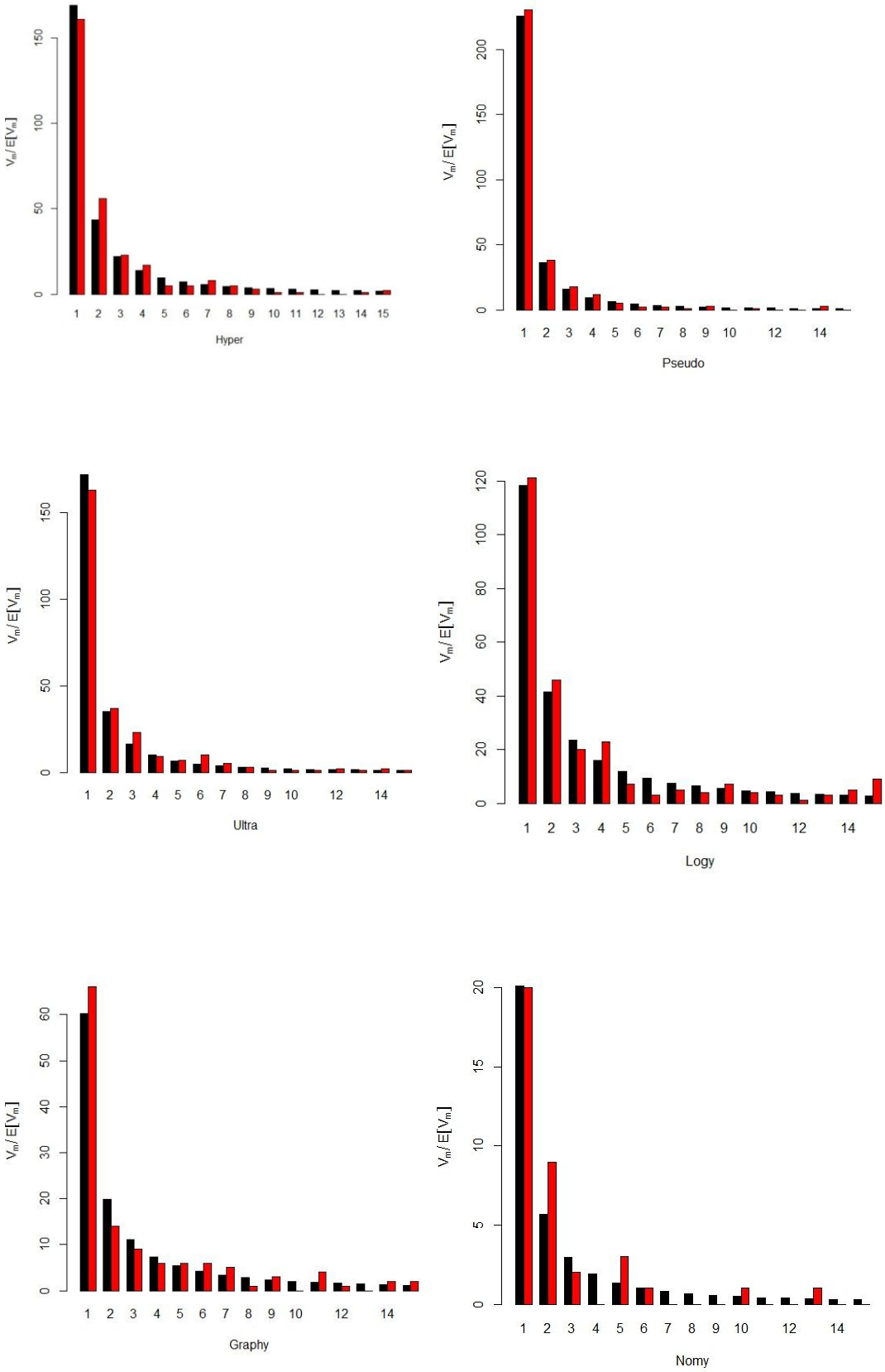

Figure 5. The expected (black) versus observed (red) frequency spectra for the selected combining forms. The x axis displays frequency classes, and the y axis displays the number of types in each class.

Not only vocabulary growth curves, but also frequency spectra provide a useful tool to assess the productivity of different word-formation processes. In particular, a frequency spectrum provides information on the distribution of the type frequencies of a given word-formation process, because productive ones tend to behave in a certain way in this respect, producing a skewed, rapidly decreasing plot. Even looking at the frequency spectra per se can reveal interesting facts, but it is even more useful to compare the observed values to the expected, average values. Again, this can be done with the ZipfR package of the R statistics software. The actual formula to calculate the expected values for a frequency spectrum is explained by Baayen (2001: 92).

Figure 5 shows the expected (black) versus observed (red) frequency spectra for the combining form data. The figure only displays the first fifteen frequency classes, but it can nevertheless give a useful visualization of how the different types are distributed with each combining form. Even a brief glance at the figure shows that all the combining forms display similar, rapidly decreasing plots with a long tail of high frequency classes that are realized by only few types. In addition, as the figure shows, almost all the observed frequency spectra line up relatively beautifully with the expected, average values.

The only exception is *-nomy,* which displays quite drastic differences between the expected and the observed values at times. For example, frequency class 2 (i.e., the number of types that occur twice in the corpus) has many more members than expected. The same holds true for frequency class 5. Scarcity of data cannot be the reason for this inconsistency, since the token frequency of *-nomy* is the second largest of all the selected combining forms. On the other hand, despite its high token frequency, *-nomy* has a particularly low type frequency, only 46 types (the most frequent combining forms having over 300 types). In a situation like this, even a small deviation from the expected values might have an impact. Another striking feature with *-nomy,* which might have an impact on the results is that the most frequent type, *economy,* accounts for over 70 per cent of all the words with *-nomy.* With *-graphy,* the corresponding percentage is roughly 20% (the most frequent type being *geography*), and with *hyper,* no more than 13% (*hypertension*).

One explanation for the lack of goodness-of-fit might also be sampling error, in which case the irregularities would be outliers, i.e., atypical data points with exceptionally high or low values (see Baayen 2008: 91). However, one has to be careful when discarding data as potential outliers. As Baayen has put it, "you should not tweak the data by removing data points so that a non-significant effect becomes significant" (2008: 237). Baayen also points out that the statistical LNRE models rely on the assumption that words occur randomly and independently of each other in texts, which is of course a simplification and might cause the lack of goodness-of-fit (2008: 233). Of course, the randomness factor might not be as pervasive in lexical studies as in studies on syntax.

The slight out-of-datedness of the *BNC* mentioned in section 3.2 is a factor that must not be overlooked either. In fact, there are several cases where a more up-to-date corpus would certainly have yielded different results. The word *euro-cracy* offers a good example, even though it is not among the combining forms selected for this study. It is a hapax in the *BNC*, but in view of the extra-linguistic development of the last two decades, it is certainly not a very rare word in British English. On the other hand, *eurocracy* also turned out to be a hapax in the *COCA* corpus, which is updated every year and should thus cover even the most recent changes in the English language. However, this might partially be explained by the fact that the *COCA* records American English. Therefore the out-of-date nature of the *BNC* shows up and skews the results. It is also possible to find examples of words which have been fashionable at the moment of corpus compilation, but which have since then become rare or even obsolete, as a result of changes in fashion or technological development. *Hypersparc* (100 occurrences), which refers to a 1990's microprocessor, is a good example.

5. Discussion

It is a relatively easy task to extract figures and tables from the data and count the *P* values for a set of word-formation processes – even if the methodological problems discussed in section 3.3 are taken into consideration. In a similar vein,

calculating frequency spectra and vocabulary growth curves with R is a relatively straightforward process. In fact, in many studies that rely on *P* as an indicator of productivity, the discussion of results is rather superficial and the evidence drawn from the *P* values of affixes is taken for granted.

However, careful qualitative analysis of the results is necessary in order to get a realistic view of the productivity of these combining forms. According to Evert and Lüdeling, quantitative and qualitative analyses can complement each other, and while quantitative approaches provide useful information on the nature of different word-formation processes, they cannot really be used without qualitative interpretation (2000: 167–168). There are several linguistic and extra-linguistic factors that must be taken into account. The factors may also shed light on eventual differences in productivity rates obtained with various productivity measures. It is also worthwhile to discuss the reasons that might explain those differences. Issues like these will be the topic of this section.

In general, all the measures employed in this study seem to indicate that the initial combining forms have higher productivity rates than the final combining forms. They have higher *P* values, and they produce steeper vocabulary growth curves. On the other hand, it seems that there are no major differences in the goodness-of-fit of the frequency spectra between the initial and the final combining forms comparing the observed and expected frequency spectra. However, based on the evidence brought about by potential productivity *P* and vocabulary growth curves, we might well claim that that the initial combining forms score higher than the final combining forms.

What might the reasons for the higher productivity rates obtained by the initial combining forms be? The differences between initial and final combining forms from a semantic or structural point of view have not been considered in previous literature, but the prefix-suffix distinction might prove a useful analogy, since it is not certain that speakers make a distinction between affixes and combining forms. Nevalainen writes that prefixes have fewer word-class restrictions on their input range than suffixes do (1999: 355). From the semantic point of view, the main linguistic function of prefixes is often semantic, while suffixes tend to be class-changing, but more abstract in meaning (see section 2.1.).

Studies concentrating on or touching upon the productivity of combining forms have mainly been theoretical. Warren, for example, rather cryptically writes that prefixes have "productive force", while combining forms "need not have productive force" but can be nonce formations (1990: 123).[22] Similarly, Prćić states that the productivity of prefixes is systematic, since they are used regularly in "ready-made morpho-syntactic patterns", and that each prefix can be assigned a value of high, restricted, or low productivity (2005: 325). Combining forms, on the other hand, "are simply there to be used if/when need for them arises", and thus display non-systematic productivity, comparable to that of elements in compounding (Prćić 2005: 325). Fischer states that dictionaries can be used to assess the productivity of combining forms since if a combining form has its own entry in a dictionary, it can be said to have at least some degree of productivity (1998: 63). However, the fact that all the elements studied in this paper have an entry in the *OED* is too weak an indicator in itself to gauge their productivity.

It is also worthwhile to consider the general tendencies or underlying changes that might be going on in terms of the English word-formation system. Bauer argues that since many of the native prefixes in English have lost ground, having been replaced by learned "prefix-like elements", a gradual typological shift from prefixation towards "something more like compounding" might be taking place (2003a: 35, 37). Even though he does not state it explicitly, by these "learned prefix-like elements" Bauer seems to specifically mean neoclassical combining forms. Szymanek seems to hold a similar view as well, claiming that the neoclassical compounds have gained ground in the English word-formation system, especially during the last few decades (2005: 432). However, Bauer and Szymanek do not discuss whether these tendencies apply to suffixes and final combining forms as well.

Szymanek also points out that coining new words, at least nouns, by compounding, is easy for speakers, due to recursion and the absence of any major grammatical restrictions (2005: 432). The same holds true for prefixes, as they,

[22] Bauer (1983: 45–46) defines nonce formations as complex words that are created "on the spur of the moment", usually to cover an immediate need. Thus they are usually unique and rare. Nonce formations must not be confused with hapax legomena: hapaxes are always defined in terms of a certain corpus. When nonce formations start to be used regularly in the speech community and become institutionalized, they become neologisms (Fischer 1998: 15–16).

too, can be used recursively, and usually do not affect the stress pattern of their bases and do not have segmental phonological effects on their bases (Lieber 2005: 389).

In a similar vein, as mentioned earlier, Hughes has noted that many learned prefixes, such as *extra-, super-* and *hyper-* have gained new, creative meanings and are thus used with greater flexibility in contemporary English than they were before (2003: 346–347). Yet another fact suggesting the high productivity of prefixes or initial combining forms comes from Bauer, who uses the term *word families* to illustrate the fact that a single prefixing operation may result in several members of the same family (as in *biodegradability, biodegradable, biodegrade, biodegradation*), which of course increases the *P* value of the element in question. Kastovsky notes that besides scientific terminology, neologisms that exploit combining forms also occur in science fiction novels, which also demonstrates the productivity of this word-formation process (2009: 1). He mentions such examples as *cryosleep, holomovie,* and *exo-biology.* With these observations in mind, combined with evidence from the *BNC*, it is not too bold to state that all the combining forms studied in this paper are fairly productive.

Etymology is a factor that might play an important role in the differences between the productivity rates between initial and final combining forms. The latter all come from lexical words (mostly nouns), while some initial combining forms come from prefixes (hyper) – do they resemble prefixes in other senses as well? Another question is how readily the selected combining forms combine with native bases as opposed to Greek and Latin elements – a possible topic for a future study.

All this raises the question whether speakers of English actually differentiate between affixes and combining forms in general and whether the distinction between these two word-formation processes has any relevance from the point of view of cognitive processing and the mental lexicon. As has already been demonstrated above, it is not always easy to distinguish between affixes and combining forms (or combining forms and bases) even in terms of morphological theory. The distinction between prefixes and initial combining forms in particular does not seem to be so clear-cut. Prćić states that modern morphology in general has been unable to make a consistent distinction between these two types of word-

formation, and they are also often labelled inconsistently in dictionaries as well (2005: 313–314).

Marchand writes that learned affixes and combining forms are "on the same footing" in the English word-formation system, since they are both borrowed elements that do not exist independently as words (1969: 131–132). Similarly, Bauer, when suggesting that English is in the process of going through a gradual typological change, in fact defines prefixes and initial combining forms as a single category of "prefixal elements" and treats elements like *anti-*, *sub-*, *super-*, *mini-* and *micro-* as members of a similar category, although the *OED* defines the first three as prefixes, and the latter two as combining forms (2003a). In a similar manner, Nevalainen seems to treat all the prefixal elements, native and non-native alike, as a single category that she calls prefixes (1999: 383).

Prćić suggests that certain elements that have traditionally been defined as combining forms (including *pseudo-* and *quasi-*) may become more like affixes after repeated use: they acquire a "ready-made morphosemantic pattern" (2005: 329).[23] He goes as far as to suggest that elements like *mega-*, *micro-*, *multi-*, *pseudo-* and *quasi-*should be recategorized as prefixes for various reasons, one of them being the fact that they modify the meaning of the base in a regular way (2005: 329). Warren has come to a similar conclusion, claiming that affixes traditionally result from a grammaticalization process while combining forms were originally bound lexical elements und thus "better equipped to resist grammaticalization" (1990: 123–124). However, she also writes that combining forms are not immune to grammaticalization either, and thus some of them have moved towards an affixal status. Fundamentally, this all comes down to Kastovsky's claim (2009) that that a category like combining forms is altogether unnecessary and that processes like affixation, compounding, blending and clipping could be used to deal with that category.

In any case, it is evident that productivity in the strict sense only measures one aspect of productivity. A more comprehensive study would involve a combination of different methods. Baayen and Lieber themselves acknowledge the fact that the type frequency of a given affix is also an indicator of its extent of

[23] Carstairs-McCarthy writes that the recent proliferation of combining forms in association with free Germanic roots can actually be seen as a rather peculiar trend, since Latin- and Greek-derived words are otherwise quite rare in present-day English (2002: 109).

use and may thus be of interest in such fields as lexicography or applied linguistics (1991: 817).

6. Conclusion

> Sometimes ignorant but pretentious people take to coining words, re-interpreting foreign words in their own way. They vaguely feel that there is some characteristic termination in a Greek or Latin word which they then attach to some English basis to give a combination a 'learned' tinge. As a result, we get barbarisms in *-athon,* coined after *Marathon*, such as *danceathon, swimathon,* etc... (Marchand 1969: 213, emphases orig.)

Whether Marchand's rather extreme claim is justified or not, he may well consider this battle as lost. What was considered "a purely dictionary interest" (see Marchand 1969: 132) a few decades ago, is now an integral part of the English lexicon without doubt. The main results of this study indicate that what has here been called combining forms are indeed frequent in English, and they combine quite freely with both native and non-native elements.

This study has attempted to measure the productivity of six neoclassical combining forms (three initial and three final) in English, using several complementary methods. First, a quantitative, statistical measure *P* developed by Baayen and his co-workers has been applied to these combining forms. Secondly, two measures of lexical statistics, i.e., vocabulary growth curves and frequency spectra, have been used in order to take as many aspects of productivity into account as possible. The motivation for using these methods has been discussed, as well as the various problems that have occurred during the research process and the appropriate solutions.

As regards their *P* values, the initial combining forms seemed to score somewhat higher than the final combining forms. In terms of vocabulary growth curves, the initial combining forms produced steeper curves than the final combining forms, which indicates that their vocabulary grows at a greater pace

and that they are therefore more productive. The third measure, i.e., the comparison of observed and expected frequency spectra, did not yield any major differences between the initial and the final combining forms. The observed and the expected values lined up rather nicely with almost all of the selected elements.

The findings are not surprising, given the latest tendencies in English word-formation that seem to favour prefix-like elements over suffixes. It has even been suggested that a typological shift might be under way in the English word-formation system (Bauer 2003a). In addition, the proliferation of scientific texts seems to contribute to the high productivity of these elements. In order to make broader generalizations, though, a larger set of combining forms should be compared. However, the problem is that in performing an analysis that draws from lexical statistics, sample size matters: many combining forms do not produce many hits even in a large corpus like the *BNC*. This makes their statistical analysis hard. R might fail to produce expected frequency spectra for some combining forms if the sample size is too small. For example, the initial combining form *ultra-* has 2000 occurrences in the *BNC*, which is not enough for R to produce an extrapolated vocabulary growth curve.

It must also be noted that most of the combining forms selected turned out to be rather affix-like by nature, their function being quantifying rather than that of resembling a lexeme. This is probably related to their high frequency in English, as opposed to more lexeme-like combining forms. However, these combining forms with more clearly lexeme-like meanings (such as *bio-* 'life' or *astro-* 'star') would probably have acquired different, probably lower, *P* values.

Another important aspect that must be kept in mind at this point is psycholinguistics. Whether speakers actually differentiate between affixes and combining forms from the point of view of mental lexicon and cognitive processing is a question worth investigating. As Plag has noted, "speakers of English can and do master English morphology without etymological knowledge" (2002: 286). It is the boundary between prefixes and initial combining forms that seems to be rather blurry. Final combining forms, on the other hand, seem to be more distinct from affixes as a category.

During the research process, it has also become evident that Baayen's quantitative method measures only one aspect of productivity, i.e., the growth rate of the vocabulary of a certain morphological category (see Baayen

2008: 902). As Plag has put it, it is "only a more or less accurate statement of the problem and not a solution to it" (1999: 35). Therefore the potential productivity *P* cannot be the only way to assess the productivity of the selected prefixes and combining forms. In this study, methods of lexical statistics were used to offer an account that is as diverse as possible and that takes into account the different aspects of morphological productivity.

Another interesting topic for future research could involve comparing the productivity rates of combining forms across different registers, following the example of Plag et al. (1999). Many Greek prefixes in particular are relatively recent loans in English and are mainly used in word-formation in specialized scientific registers (Biber et al. 1999: 324). The problem with that kind of research is that the samples compared must be of equal size, which, as Säily has noted, may lead to discarding valuable data (2011: 125). On the other hand, studying the correlation of productivity and different sociolinguistic variables, such as gender, age or social rank of the speakers (see Säily and Suomela 2009, Säily 2011), is also worth further study. Another aspect of productivity that has not been thoroughly investigated yet is the competition between different word-formation processes that express the same need to name various concepts and phenomena (see Lüdeling et al. 2006).

In the future, more theoretical accounts of productivity are also needed: lately, studies of productivity have mainly concentrated on the productivity of individual word-formation processes (see Dressler and Ladányi 2000: 104). Besides, even though the development of different empirical and statistical methods to measure productivity has flourished in the past few years, there are still many questions that concern the very nature of productivity that need to be tackled. Plag, for example, mentions such problems as allomorphy and the relationship between rival morphological processes, as well as how morphology interacts with phonology, syntax, and semantics (1999: 1). In addition, the study of productivity might have practical applications beyond morphology, measuring the "generative potential" (Evert/Baroni) of such things as syntactic processes or collocations/idioms. LNRE models in particular are also an area that has not so far been widely employed in the study of productivity.

Kastovsky argues that instead of considering the demarcation between combining forms and affixation, a more relevant question would actually

be the difference between compounding and affixation in general (2009: 10–11). He points out that a number of native affixes has developed from members of compounds (e.g., *-dom, -hood,* and *-wise*), and the same phenomenon can be seen with combining forms (2009: 10). Kastovsky suggests that the theoretical framework of grammaticalization would probably be productive in this kind of analysis (2009: 11). From the semantic point of view, a crucial distinction in this model is also the one between "lexically specific" and "lexically non-specific" word-formation processes, which can also be seen as a cline between words and stems on the one hand and affixes on the other hand (Kastovsky 2009: 11). Another research topic not yet fully covered is the so-called modern combining forms, such as *jazzo-*(*phile*) or *speedo-(meter)* (Prćić 2005, 2007, 2008).

Primary sources and software

BNC = The British National Corpus, version 4.1 (*BNC* XML Edition). 2008. Distributed by Oxford University Computing Services on behalf of the *BNC* Consortium. Online: http://www.natcorp.ox.ac.uk/

COCA = The corpus of contemporary American English. Compiled by M. Davies. 2008– Online: http://corpus.byu.edu/coca

The R project for statistical computing. http://www.r-project.org

References

Amiot, D. and G. Dal. 2007. Integrating neoclassical combining forms into a lexeme-based morphology. In G. Booij.et al. (eds.), *On-line proceedings of the fifth Mediterranean Morphology Meeting (MMM5)* Fréjus 15–18 September 2005, University of Bologna. Online: http://mmm.lingue.unibo.it/mmm-proc/MMM5/232-336-Amiot-Dal.pdf [accessed on October 26, 2012]

Anshen, F. and M. Aronoff. 1981. Morphological productivity and phonological transparency. *The Canadian Journal of Linguistics* 26 (1), 63–72.

Aronoff, M. 1976. *Word formation in generative grammar.* Cambridge, MA: MIT Press.

Aronoff, M. and F. Anshen. 1998. Morphology and the lexicon: Lexicalization and productivity. In A. Spencer and A.M. Zwicky (eds.), *The Handbook of morphology*. (Blackwell Handbooks in Linguistics). Cambridge, MA: Blackwell Publishers, 237–248.

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In G. Booij and J. van Marle (eds.), *Yearbook of morphology 1991*. Dordrecht: Kluwer, 109–149.

Baayen, R. H. 1993. On frequency, transparency and productivity. In G. Booij and J. van Marle (eds.), *Yearbook of morphology 1992.* Dordrecht: Kluwer, 182–208.

Baayen, R.H. 2001. *Word frequency distributions.* Dordrecht: Kluwer.

Baayen, R.H. 2003. Probabilistic approaches to morphology. In R. Bod et al. (eds.), *Probabilistic linguistics.* Cambridge, MA, MIT Press, 229–287.

Baayen, R. H. 2008. *Analysing linguistic data: a practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. H. 2009. Corpus linguistics in morphology: morphological productivity. In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: An international handbook,* ed. by Anke Lüdeling and Merja Kytö. Berlin: Walter de Gruyter, 899–919.

Baayen, R. H. and R. Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29 (5), 801–843.

Baayen, R. H. and A. Renouf. 1996. Chronicling the Times: Lexical innovations in an English newspaper. *Language* 72 (1), 69–96.

Barker, C. 1998. Episodic *-ee* in English: a thematic role constraint on new word formation. *Language* 74: 695–727.

Baroni, M. 2008. Distributions in text. In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: an international handbook.* Berlin: Mouton de Gruyter, 803–822.

Baroni, M. and S. Evert. 2008. Statistical methods for corpus exploitation. In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: an international handbook.* Berlin: Mouton de Gruyter, 777–803.

Bauer, L. 1983. *English word-formation.* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Bauer, L. 1988. *Introducing linguistic morphology.* Edinburgh: Edinburgh University Press.

Bauer, L. 1992. Scalar productivity and *-lily* adverbs. In G. Booij and J. van Marle (eds.), *Yearbook of morphology 1991*. Dordrecht: Kluwer, 185–191.

Bauer. L. 1995. Is morphological productivity non-linguistic? *Acta Linguistica Hungarica* 43 (1–2), 19–31.

Bauer, L. 1998. Is there a class of neoclassical compounds and if so is it productive? *Linguistics* 36 (3), 403–422.

Bauer, L.. 2001. *Morphological productivity.* (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.

Bauer, L. 2003a. English prefixation – a typological shift? *Acta Linguistica Hungarica* 50 (1–2), 33–40.

Bauer, L. 2003b. *Introducing linguistic morphology.* 2nd edition. Edinburgh: Edinburgh University Press.

Bauer, L. 2003c.The productivity of (non-)productive morphology. *Rivista di Linguistica* 15, 7–16.

Bauer, L. 2005 Productivity: theories. In Štekauer and R. Lieber, 315–334.

Biber, D. et al. 1999. *Longman grammar of spoken and written English.* London: Longman.

Bloch, B. and G.L. Trager. 1942. *Outline of linguistic analysis.* Baltimore: Linguistic Society of America.

Booij, G. E. 2005. *The grammar of words: an introduction to linguistic morphology.* Oxford: Oxford University Press.

Burnard, L. (ed.). 2007. *Reference guide for the British National Corpus (XML edition).* Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. Online: http://www.natcorp.ox.ac.uk/docs/URG/ (accessed on 1 October 2012).

Burnley, L. 1992. Lexis and semantics. In Blake, N. (ed.) *The Cambridge history of the English language, Vol. 2: 1066–1476.* 409–499.

Carstairs-McCarthy, A. 1992. *Current morphology.* London/New York: Routledge.

Carstairs-McCarthy, A. 2002. *An introduction to English morphology: words and their structure.* (Edinburgh Textbooks on the English Language). Edinburgh: Edinburgh University Press.

CEEC = Corpus of Early English Correspondence. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of English, University of Helsinki.

Chitashvili, R.J. and H. Baayen 1993. Word frequency distributions. In L. Hrebícek and G. Altmann (eds.) *Quantitative text analysis*. Trier : WVT, 54-135.

Corbin, D. 1987. *Morphologie dérivationnelle et structuration du lexique.* Tübingen: Max Niemeyer.

Cottez, H. 1985. *Dictionnaire des structures du vocabulaire savant: éléments et modèles de formation.* Paris: Le Robert.

Cowie, C. 1998. The discourse motivations for neologising: Action nominatization in the history of English. In J. Coleman and C. J. Kay (eds.), *Lexicology, semantics and lexicography. Selected papers from the fourth G. L. Brook Symposium.*(Amsterdam Studies in the Theory and History of Linguistic Science. Series 4, Current issues in linguistic theory, vol. 194). Amsterdam: John Benjamins.179–207.

Cowie, C. 2003. "Uncommon terminations": Proscription and morphological productivity. *Rivista di Linguistica* 15 (1), 17–30.

Cowie, C. and C. Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In J. E. Díaz Vera (ed.) *A changing world of words: Studies in English historical lexicography, lexicology and semantics.* Amsterdam: Rodopi.410–437.

Dalton-Puffer, C. 1994.Productive or not productive? The Romance element in Middle English derivation. in F. Férnandez et al. (eds.) *English Historical Linguistics 1992: Papers from the 7th International Conference on English Historical Linguistics, Valencia, 22–26 September 1992.* Amsterdam: John Benjamins.247–260.

Dalton-Puffer, C. 1996.*The French influence on Middle English morphology: A corpus-based study of derivation.* (Topics in English Linguistics 20). Berlin: Mouton de Gruyter.

Dressler, W. U. and M. Ladányi. 2000. Productivity in word formation (WF): A morphological approach. *Acta Linguistica Hungarica* 47 (1–4), 103–144.

Evert, S. and A. Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In P. Rayson et al. (eds.) *Proceedings of the Corpus Linguistics 2001 Conference.* Lancaster: University Center for Computer Research on Language. 167–175.Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.9162&rep=rep1&type=pdf [accessed on October 15, 2012]

Evert, S. and A. Lüdeling. 2003. Linguistic experience and productivity: Corpus evidence for fine-grained distinctions. In D. Archer et al. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference.* UCREL technical paper 16. Lancaster: UCREL. 475-483.Online: http://citeseer.ist.psu.edu/viewdoc/similar?doi=10.1.1.13.2813&type=ab [accessed on January 13, 2013]

Evert, S. and M. Baroni. 2007. zipfR: Word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Posters and Demonstrations Session.* Prague, Czech Republic. Online: http://www.stefan-evert.de/PUB/EvertBaroni2007.pdf [retrieved on April 13, 2013]

Férnandez-Domínguez, J. et al. 2007. How is low morphological productivity measured? *Atlantis* 29 (1), 29–54.

Férnandez-Domínguez, J. 2009.*Productivity in English word-formation: An approach to n+n compounding.* (European University Studies, series 21, Linguistics, vol. 341) New York: Peter Lang.

Fischer, R. 1998. *Lexical change in present-day English: A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms.* (Language in Performance). Tübingen: G. Narr.

Fradin, B. 2000. Combining forms, blends and related phenomena. In U. Doleschal and A. M. Thornton (eds.), *Extragrammatical and marginal morphology.* Munich: LINCOM Europa, 11–60.

Frauenfelder, U. H. and R. Schreuder 1992.Constraining psycholinguistic models of morphological processing and representation: the role of productivity. In G. E. Booij and J. van Marle (eds.), *Yearbook of morphology 1991.* Dordrecht: Kluwer Academic Publishers, 165–183.

Harastani, R. et al. 2012. Neoclassical compound alignments from comparable corpora. In A. Gelbukh (ed.), *Computational linguistics and intelligent*

text processing 13th international conference, CICLing 2012, New Delhi, India, March 11–17, 2012, Proceedings, Part II. Heidelberg: Springer, 72–82.

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

Hay, J. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39, 1041–1070.

Hay, J. 2003. *Causes and consequences of word structure.* (Outstanding Dissertations in Linguistics). New York: Routledge.

Hay, J. and H. Baayen. 2002. Parsing and productivity. In G. Booij and J. van Marle (eds.), *Yearbook of morphology 2002,* 203-235. Dordrecht: Kluwer.

Hohenhaus, P. 2005. Where word-formation cannot extend the vocabulary: Creativity, productivity and the lexicon in synchronic and diachronic morphology. In D. Kastovsky and A. Mettinger (eds.), *Lexical change and the genesis of English vocabulary.* Berlin/New York: Mouton de Gruyter,

Huddleston, R. and G. Pullum. 2002. *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Hughes, G. 2003. *A history of English words.* Oxford: Blackwell Publishing.

Kastovsky, D. 2001. English morphology: A typological reappraisal. In C. Schaner-Wolles et al. (eds.). *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*. Turin: Rosenberg and Sellier.215–224.

Kastovsky, D. 2006. Typological changes in derivational morphology. In A. van Kemenade and B. Los (eds.), *The handbook of the history of English.* (Blackwell Handbooks in Linguistics). Malden: Blackwell Publishing. 151–176

Kastovsky, D. 2009. Astronaut, astrology, astrophysics: About combining forms, classical compounds and affixoids. In R. W. McConchie et al. (eds.) *Selected proceedings of the 2008 Symposium on New Approaches in English Historical Lexis (HEL-LEX 2).* Somerville, MA: Cascadilla Proceedings project. 1-13. Online: http://www.lingref.com/cpp/hel-lex/2008/paper2161.pdf [accessed on January 14, 2013].

Krott, A. et al. 1999. Complex words in complex worlds. *Linguistics* 37. 905–926.

Lee, D. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the *BNC* jungle. *Language Learning & Technology* 5 (3), 37–72.

Lehrer, A. 1998.Scapes, holics, and thons: The semantics of English combining forms. *American Speech* 73 (1), 3–28.

Lieber, Rochelle. 2005. English word-formation processes: Observations, issues, and thoughts on future research. In Štekauer and Lieber (eds.), 375–427.

Lipka, L. 1990. *An outline of English lexicology.* (Forschung & Studium Anglistik 3). Tübingen: Niemeyer.

Lüdeling, A. et al. 2000. On measuring morphological productivity. *Proceedings of the KONVENS 2000*, 57–61. Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.4819&rep=rep1&type=pdf [accessed on November 16th, 2012].

Lüdeling, A. et al. 2002. Neoclassical word-formation in German. In Booij.G. and J. van Marle (eds.) *Yearbook of morphology 2001*.Dordrecht: Kluwer, 253–283.

Lüdeling, A. and S. Evert. 2003. Linguistic experience and productivity: corpus evidence for fine-grained distinctions. *Proceedings of the Corpus Linguistics 2003 Conference.* Online: http://www.ims.uni-stuttgart.de/projekte/corplex/paper/evert/LuedelingEvert2003.pdf [accessed on November 26, 2012].

Lüdeling, A. et al. 2006. Need and competition: Deconstructing quantitative productivity. Online: http://cogsci.uni-osnabrueck.de/~qitl/abstracts/luedeling_baroni_evert.pdf [accessed on November 25, 2012].

Marchand, H. 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach.* 2nd edition. Munich: Beck'sche.

Marle, J. van. 1985. *On the paradigmatic dimension of morphological creativity.* (Publications in language sciences 15). Dordrecht: Foris Publications.

McCauley, J. 2006. Technical combining forms in the third edition of the *OED*: Word-formation in a historical dictionary. In R. W. McConchie et al. (eds.) *Selected proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX).*Somerville, MA: Cascadilla Proceedings Project. 95–104. Online: http://www.lingref.com/cpp/hel-lex/2005/paper1350.pdf [accessed on January 14, 2013]

McConchie, R. W. 1998. The vernacularization of the negative prefix *dis-* in Early Modern English. *Lexicology, semantics and lexicography.* In J. Coleman and C. J. Kay (eds.) *Selected papers from the fourth G. L. Brook Symposium.*(Amsterdam Studies in the Theory and History of Linguistic Science. Series 4, Current issues in linguistic theory, vol. 194) Amsterdam: John Benjamins. 209–227.

McConchie, R. W. 2006. *Disseisin:* The lexeme and the legal fact in Early Middle English. In R. Dury el al. (eds.) *English historical linguistics 2006: Lexical and semantic change.* Amsterdam: John Benjamins. 203–216.

Minkova, D. and R. Stockwell. 2009. *English words: Structure and history.* 2nd edition. Cambridge: Cambridge University Press.

Miller, D.G. 2012. *External influences on English: From its beginnings to the renaissance.* Oxford: Oxford University Press.

Nevalainen, T. 1999. Early Modern English lexis and semantics. In Lass, R. (ed.). *The Cambridge history of the English language, Vol. 3: 1476–1776.* Cambridge: Cambridge University Press. 332–458.

*OED = The Oxford English dictionary online.*2007 (Online version 2013).3rd edition. Oxford: Oxford University Press.Online: http://www.oed.com

Palmer, C. C. 2008. Borrowed derivational morphology in Late Middle English: a study of the records of the London Grocers and Goldsmiths. In Fitzmaurice, S. et al. (eds.). *Studies in the history of the English language IV: empirical and analytical advances in the study of English language change.* (Topics in English linguistics 61). Berlin/New York: Mouton de Gruyter, 231-64.

Palmer, C. C. 2009. *Borrowings, derivational morphology, and perceived productivity in English, 1300–1600.*Ph. D. thesis. University of Michigan. Online: http://deepblue.lib.umich.edu/bitstream/2027.42/64624/1/palmercc_1.pdf [retrieved September 18, 2012]

Pápai, V. 2004. Explicitation: A universal of translated text? In A. Mauranen and P. Kujamäki (eds.) *Translation universals: Do they exist?* Amsterdam: John Benjamins Publishing.143-164.

Plag, I. 1999. *Morphological productivity. Structural constraints in English derivation.*(Topics in English Linguistics 28). Berlin: Mouton de Gruyter.

Plag, I. 2002. The role of selectional restrictions, phonotactics, and parsing in constraining suffix ordering in English. In G. Booij and J. van Marle (eds.), *Yearbook of morphology 2001*, 285–314.Dordrecht: Kluwer.

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

Plag, I. 2006. Productivity. In B. Aarts and A. McMahon (eds.), *The handbook of English linguistics*. (Blackwell Handbooks in Linguistics). Malden: Blackwell Publishing.537–556.

Plag, I. et al. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3 (2), 209–228.

Pounder, A. 2000. *Processes and paradigms in word-formation morphology.* Berlin: Mouton de Gruyter.

Prćić, T. 2005. Prefixes vs. initial combining forms in English: A lexicographic perspective. *International Journal of Lexicography* 18 (3), 313–334.

Prćić, T. 2007. Headhood of suffixes and final combining forms in English word-formation. *Acta linguistic hungarica* 54 (4), 381–392.

Prćić, T. 2008. Suffixes vs. final combining forms in English: A lexicographic perspective. *International Journal of Lexicography* 21 (1), 1–22.

Quirk, R. et al. 1985. *A comprehensive grammar of the English language.* London: Longman.

Rainer, F. 2005. Constraints on productivity. In Štekauer and Lieber (eds.), 335-352.

Renouf, Antoinette. 2007. Tracing lexical productivity and creativity in the British Media: 'The Chavs and the Chav-nots'. In J. Munat (ed.) *Lexical creativity. texts and contexts.* (Studies in Functional and Structural Linguistics). Amsterdam: Benjamins. 61–89.

Säily, T. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7 (1), 119–141.

Säily, T. and J. Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In A. Renouf and A. Kehoe (eds.) *Corpus linguistics: Refinements and reassessments*. (Language and Computers: Studies in Practical Linguistics 69.) Amsterdam: Rodopi. 871–909. Authors' version available from http://www.cs.helsinki.fi/u/josuomel/doc/icame-2007-paper.pdf [retrieved on October 27, 2012]

Schäfer, J. 1989. Early Modern English: *OED, New OED, EMED.* In R. W. Bailey (ed.), *Dictionaries of English: prospects for the record of our language.* Cambridge: Cambridge University Press. 66-74.

Štekauer, P. and R. Lieber (eds.) 2005.*Handbook of word-formation.* Dordrecht: Springer.

Štekauer, P. et al. 2005.Word-formation as creativity within productivity constraints: sociolinguistic evidence. *Onomasiology Online* 6, 1–55. Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.9286&rep=rep1&type=pdf [Accessed on November 18, 2012]

Szymanek, B. 2005. The latest trends in English word-formation. In Štekauer and Lieber (eds.), 429–448.

Tweedie, F. J. and R. H. Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32, 323–352.

Warren, B. 1990.The importance of combining forms. In W. U. Dressler et al (eds.) *Contemporary morphology.*Trends in linguistics: Studies and monographs 49. Berlin: Mouton de Gruyter.111–132.

Zipf, G. K. 1949. *Human behaviour and the principle of least effort.* Cambridge, MA: Addison-Wesley.

Zipf, G. K. 1965. *The psycho-biology of language.* Cambridge, MA: MIT Press.