SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA ANGLISTIKU

Ak. god. 2015./2016.

Nikola Dubović

English Neologisms Since 2001

diplomski rad

Mentor: dr. sc. Mateusz-Milan Stanojević, doc.

Zagreb, 2016.

Abstract

The purpose of this paper is to investigate English word formation since the beginning of the 21st century and draw a comparison with earlier periods to determine how, if at all, word formation patterns have evolved. A sample of neologisms created in the period 2001-2013 was gathered, with etymological analyses, and frequencies of the various word formation processes in the sample were calculated. The results showed a significant reduction in shortenings, relative to the results of the research of previous periods, and an even more significant increase in blending. There are some questions, however, about the validity of the results due to possible bias in the sample that was analyzed.

Key words: neology, neologisms, English neology, English neologisms, word formation, English word formation, word formation processes, derivation

Sažetak

Rad istražuje tvorbu riječi u engleskom od početka 21. stoljeća i uspoređuje ju sa tvorbom riječi u engleskom u prethodnim periodima. Istraživanje je provedeno na uzorku riječi nastalih u periodu 2001.-2013. Rezultati pokazuju znatno umanjenu ulogu kraćenja u odnosu na rezultate istraživanja prethodnih perioda, a još veća promjena je puno češća uporaba srastanja u tvorbi novih riječi. Postoje razlozi, međutim, za sumnju o vjerodostojnosti rezultata zbog nesigurnosti o reprezentativnosti uzorka.

Ključne riječi: tvorba riječi, tvorba riječi u engleskom, neologija, novotvorenice, novotvorenice u engleskom, neologizmi, neologizmi u engleskom

Contents

1. Introduction

The purpose of this paper is to investigate English word formation since the beginning of the 21st century. A sample of neologisms that have been created since 2001 will be gathered and etymologically analyzed in order to determine the productivity of each word formation process in the period 2001-2013. A comparison will be made with previous research investigating productivity of various word formation processes in different periods of the 20th century to determine what, if any, changes have happened since.

There are a number of difficulties that are encountered when trying to conduct research into the productivity of the different word formation processes of a language during a given period of time.

The first challenge is in gathering a sample of words representative of English neology for the period that the researcher wants to examine. Past research has mostly used dictionaries of new words or general dictionaries. (Algeo 1998, 84) Any sample derived from such sources is going to have a bias that will depend on the editorial policy of the dictionary.

Dictionaries differ in the kind of sources they monitor when searching for candidate words for inclusion. *The Longman Register of New Words*, for example, takes words almost exclusively from newspapers and other periodical publications (Ayto 2003, 186), while *The Oxford English Dictionary* is biased towards literature, particularly that of canonically enshrined authors, and gives less regard to folk language. (Algeo 1998, 63) The language used in these different sources is likely to vary in degrees of formality, presence of subculture-specific argot and other ways. There are also varying policies when it comes to the inclusion of nonce words. Some dictionary makers will discriminate a great deal on the basis of whether they believe a word has staying power in the language or is likely to be a short lived fad, others will have a more inclusive policy. *The Chambers Dictionary* will consider for inclusion words with as few as 10 corpus citations, (O'Donovan and O'Neill 2008, 576) while the dictionaries of Oxford University Press require a word to be found in a variety

of different sources by different writers, not be limited to only one group of users, to have a long history of use and be likely to be used in the future. (Taylor 2015, 42)

The second challenge involved in doing this sort of research is establishing the correct etymology for a neologism. The following quote from Algeo illustrates the difficulties involved:

> For example, unconscious 'that part of the mind not available to introspection, which nevertheless affects behavior' might reasonably be thought to be either a shift of use from the adjective or a clipping of the collocation unconscious mind, or even a reformation with the prefix *un-*. The OEDs first citation, dated 1884, is from Mark Pattison's *Memoirs*: 'T cannot help observing the remarkable force with which the Unconscious — *das Unbewusste* — vindicated its power.' That citation suggests that the English word is a calque on German and therefore a borrowing. Such uncertainty is far from unusual. (Algeo 1998, 83-84)

Obtaining accurate results requires that both of these challenges be met in a satisfactory manner.

The structure of this paper is as follows: The second section deals with the theoretical framework, explaining the taxonomy of word formation processes used in the research and discussing the ways in which research of neology may be conducted in the future by making use of computer technologies. The third section describes the methodology used in conducting the research reported on in this paper and the fourth section gives the results as well as the results of past research of English neology so that a comparison can be made. In the fifth section the results are briefly discussed and the sixth section is the conclusion.

2. Theoretical background

New words are generated in languages constantly and represent the most readily observable form of language change. According to one estimate based on the texts collected by Google Books the rate at which English is producing new words is increasing quite dramatically. In the last 50 years the word stock has increased by over 70%, growing at a clip of approximately 8,500 words annually. The estimate puts the size of the English vocabulary in 2000 at 1,022,000. (Michel 2011)

New words are created when they are needed for a new product, technology, cultural phenomenon, event, or sometimes, simply on a lark. The motivation for the creation of some words, like for example “radar,” is shortening a phrase that is too much of a mouthful.

Words are created through a number of different processes. Over the many centuries of the study of new words a traditional taxonomy of word making processes has emerged, however the taxonomy has not been defined clearly or consistently. As a result some words will represent very clear and typical examples of a category, but others will seemingly fall in an ambiguous space between two categories. The following quote from Algeo gives examples that illustrates the problem well:

> Thus, everyone would agree that scuba is an acronym for self contained underwater breathing apparatus; and probably there would be agreement on radar for radio detecting and ranging; but what about Nabisco for National Biscuit Company; or sit com for situation comedy; or prof for professor? Are all of those acronyms, or at some point did we leave the class of acronyms and enter some other class of words? There is no sure answer to that question because there is no consensus on what an acronym is. Some linguists would consider all of those items acronyms; some, not. Thus we find ourselves in the odd position of being sure that some things are acronyms, but not being sure what sort of thing an acronym is. (Algeo 1978, 123)

The word categories I will discuss here will be the ones defined by Algeo, as they were the ones used for the purpose of the research this paper reports on. (1998, 59-61)

Algeo's taxonomy of word forming processes pays particular attention to the relationship between a word and the sources from which it is constructed, its etyma. The categories are defined by asking four questions:

1. Does the word have an etymon, which is to say, is it based on any preexisting words?

2. Does the word omit any part of an etymon?

3. Does the word combine two or more etyma?

4. Do any of the word's etyma come from a language other than English?

These questions define six major classes of words:

1. Creations

Words that are created through onomatopoeia or made up, rather than created through modifying a preexisting word or morpheme. "Bang" would be an example of a word created through onomatopoeia. Examples of words that are entirely made up are always uncertain, as it is possible that the the word does have some connection to a preexisting word or morpheme that we fail to see. An often cited example of a made up word is "googol."

Creation happens very rarely, most words have some relation to an already existing etymon.

2. Shifts

Words that neither combine nor shorten etyma. These are words that have been transferred from one grammatical category to another. For example the word "Google" was initially used as a noun, but later began to be used as a verb.

3. Shortenings

Words that omit part of their etyma. In many taxonomies this category is further segmented into abbreviations, clippings and backformations.

Backformation is a process which operates by analogy. An oft-cited example of back formation is the creation of “edit” from “editor.” This word was created by analogy with the many words containing the suffix *–or* which denote the performer of an action, such as “protector,” “administrator,” “legislator,” etc. By analogy with these forms, the ending *–or* was removed from “editor,” yielding “edit.”

Clipping, unlike backformation, does not change the meaning of the word, but merely shortens an existing form. Clipping deletes a part of a word, usually retaining the first part of the base word (“photo” from “photography,” or “demo” from “demonstration”), or, much less frequently, part of a stressed syllable (“phone” from “telephone”).

Abbreviations are words formed by taking the initial letters (and sometimes non initial letters) of a multi word sequence to form a new word. Some abbreviations may come to resemble blends by combining larger sets of initial and non-initial letters. (Plag 2003, 126)

Among abbreviations a distinction is further typically made between words which are pronounced as normal words and those in which each individual letter is pronounced as if in isolation. The former are called acronyms and the latter initialisms.

4. Composites

Words that combine two or more etyma. Algeo further distinguishes between compounding and affixation.

Compounds are words consisting of multiple elements (roots or words) joined together to create a new word. Ortographically a space can remain between the elements (“White House”), they can be separated by a dash (“blue-eyed”) or they can be joined together (“greenhouse”).

Affixation involves attaching bound morphemes to bases. Not all affixation creates new words, some affixes merely change the grammatical properties of words. These are called inflectional affixes. The affixes which produce a new lexical item are called derivational affixes. Derivational affixes can be prefixes, attaching to the beginning of a base, such as *re-*,

used in “replay,” “rebuild,” “reuse,” or suffixes, attaching to the end of a base, such as *–ish*, used in “longish,” “tallish,” “slowish.”

5. Blends

Words that combine at least two etyma and omit part of at least one. The distinction between blending and compounding being that in compounding all the etyma involved remain intact.

Examples of blends are “brunch,” from “breakfast” and “lunch,” or “motel,” from “motor” and “hotel.”

6. Loanwords

Words with at least one non-English etymon, excluding etyma which are non-English in origin but have been in the language for a long time, such as the etyma used in the creation of neoclassical formations like “biochemistry,” or “geology.”

Algeo further distinguishes between adoption and adaptation. Adoption is a popular process in which words are borrowed with minimal change such as the French “baguette.” Adaptation, on the other hand, involves modifying the word to better fit the patterns of English. “Snorkel” from German “Schnorchel” is an example of adaptation.

In addition to these six major classes, Algeo mentions “native developments,” which are words that are phonological and semantic developments of earlier words from English, like “town” from Old English “tun” meaning “an enclosed space.”

There has not been much research into the productivity of each of these processes at various times in the history of English. The research that has been done invariably faces the problem that the accuracy of the results is uncertain because of questions about the representativeness of the word stock analyzed. If a dictionary is used to form a sample the researcher has to wonder about whether the dictionary makers monitored equally the different possible sources, such as newspapers, magazines, fiction, non-fiction etc. or whether an excessive focus was given to one particular type of source, thus biasing the sample.

Furthermore, a group of people working on a dictionary can only monitor so many sources and any dictionary must omit certain words in the interest of concision. The previously referenced study performed on Google Books concluded that 52% of the English lexicon is not documented in standard reference books. (Michel 2011)

These are difficulties that linguists in the past were simply stuck with, but in recent years computers have been making it possible to make very substantial methodological improvements when conducting research into new words and productiveness of various word formation processes.

Already in 1990 the University of Birmingham began working on what they referred to as the Analysis of Verbal Interaction and Automatic Text Retrieval (AVIATOR) project, the goal of which was text searching software that would be capable of automatically monitoring the changes in the word stock of English. (Renouf 1993)

AVIATOR maintains a "master word list" which represents all the words already in the language. The list is compiled automatically by the software from corpus data that is fed into it. Every time corpus data is fed into it all of the words in the data are checked against the master list in order to find words that are not in it and in that way detect new words. AVIATOR also automatically subdivides new items into proper nouns, abbreviations and acronyms, numerals and "ordinary words," and notes first and last date of appearance.

After AVIATOR, from 1997 to 2000, the University of Birmingham developed APRIL (Analysis and Prediction of Innovation in the Lexicon), whose purpose is, among other things, to automatically determine the word formation process that produced a new word and classify the word grammatically.

A similar kind of automated system for detection of neologisms is used by the makers of the *Chambers Dictionary*. Their process involves compiling a corpus on a monthly basis, always from the same set of sources, which include newspapers, magazines and websites, in which words that are not already in their dictionary are automatically detected. A list of new words that is automatically generated by this process is then made available to lexicographers who then make judgments about which words should be included in the dictionary, which ones should be discarded and which ones are to be monitored further. The list includes a KWIC

type citation with further options to link through to concordances in the Chambers Harrap International Corpus and ukWaC and hits in Google and Wikipedia to make work easier for the lexicographers.

The automated part of the process, before the intervention of the lexicographers, generates about 600 words a month. The *Chambers Dictionary* also maintains a directed reading program in which commissioned readers scan publications that have been identified as being potentially productive sources of new words. This produces around 200 new words a month. This is one indication of the improvement that automation brings about over old methods of studying neology, although it should be said that the output of the automated process does contain some noise as well, such as identifying misspellings, common nouns or leftover HTML elements as new words. (O'Donovan and O'Neill 2008)

It has also been shown that it is possible to automatically detect not only new forms, but new senses, as well. One method of doing this is by monitoring changes in collocational patterns of a word, but other methods have been proposed, as well. (Renouf 1993b) (Cook 2013)

The possibility of building very large corpora of diverse texts through automated web crawling combined with these kinds of programs of automated detection of neologisms could greatly improve the quality of research into new words. O'Donovan and O'Neill's paper which describes how *Chambers'* automated system works states that it compiles the monthly corpora in which it checks for neologisms from magazines, newspapers and websites, but it doesn't specify which kind of websites. The *Chambers Harrap International Corpus*, for example, is compiled from daily web crawling as well as PDFs of books provided by cooperating publishers. Newspapers and magazines make up a little less than 50% of it. The rest is books (fiction and non-fiction), blogs, websites and even spoken language.

That is quite a diverse mix of sources and the *Chambers Dictionary's* neologism detection process would benefit from the inclusion of all of them, but ideally a corpus would include informal internet communication as well, such as that conducted over forums or social networks. Language of a different register is likely to be found there that would not be found on websites of newspapers or magazines, or even blogs. In fact, ideally, all types of websites

would be crawled daily, developing a vast monitor corpus. This would be a big step towards developing a truly representative sample in which all types of sources are given an equal examination. The combination of automated detection tools with human oversight applied to such a sample would yield a far more accurate and comprehensive image of neology in a language at a given time than had previously been possible, with much less bias in the sample.

3. Methodology

A differentiation was made in the sample of words gathered for the research between words which achieved a degree of longevity in the language and those that did not. This yielded two separate data sets; one which held all the neologisms, regardless of whether they stayed in the language or quickly disappeared, and one which only held words which had longevity. Separate analyses were done for each dataset so that a comparison could be made showing how taking into account longevity affected the productivity of each word formation process.

"Longevity" is an inexact term and could be defined many ways, but for the purpose of this research it was used to exclude nonce formations or words with very short life spans due to being very odd or atypical or sometimes obvious examples of "stunt coining." Therefore, a word was taken to have achieved longevity if it could be attested at least 10 times in the second year since its entry into the language or later. Since the research involved a comparison between the two datasets, all of the neologisms since 2001 and the neologisms since 2001 which achieved longevity, this definition of longevity limited the pool of neologisms that could be used in the research to those that had been coined up to 2013, because the number of times a neologism from 2014 was attested in the second year since its coining could not be determined, as this research was done in 2016.

10 instances of use may seem like a small number to use as a criterion for determining whether a word was still used in a language, but any corpus, no matter how large, represents a tiny fraction of actual language use in a given year, so if 10 instances can be found in corpora there will surely have been many more times that the word was actually used by the speakers of the language. There is no way to determine some correct number of minimum citations and the matter can be debated.

For example, the lexicographers who compile the *Chambers Dictionary* have developed a set of guidelines in their work in which words with at least 25 corpus citations are considered strong candidates for inclusion in the dictionary, while words with between 25 and 10

citations are considered weaker candidates, but are not automatically dismissed. (O'Donovan and O'Neill 2008, 576)

Attention was paid to the sources of the citations to ensure that they came from multiple sources. A word could potentially have a high frequency in a corpus, but all the citations could be from a single book in which the word was coined. It was necessary to recognize such cases and count them as only one instance of use.

Three corpora and Google were used to check whether a word was still being used a year after entering the language. The corpora used were the Corpus of Contemporary American English (COCA), EnTenTen and the News of the Web (NOW) corpus.

NOW and EnTenTen were chosen because of their size, 3 billion and almost 23 billion words, respectively, and COCA was chosen because it covered the years the other two did not (EnTenTen starts from 2008 and NOW from 2010) and because EnTenTen and NOW draw texts exclusively from the Web, while COCA draws texts from a variety of sources including academic sources, fiction and spoken language.

Scientific terms were not included in either data set, unless they clearly entered the general lexicon. The burden of proof here was greater than 10, but was not strictly defined, because, as O'Donovan and O'Neill point out, scientific terms can show a high frequency in a corpus when they are used in newspapers following a major scientific discovery, but their use in non scientific publications ends with the news cycle. (2008, 577) Because of this dynamic it is not possible to define a frequency threshold which a scientific term would need to break in order to be considered to have entered general use and subjective judgment has to be introduced. Fortunately, there were only two scientific words in the data set so any subjectivity could not significantly skew the results.

Among the New Words, a regular column in the journal *American Speech,* was used as a source of neologisms. An advantage *Among the New Words* has over some dictionaries of new words is that it records semantic change, not only new forms. Most of the words in *Among the New Words* come from texts, but words are gathered from spoken language as well.

All of the editions of *Among the New Words*, beginning with the year 2001 and ending with 2013, were reviewed and all neologisms listed in them were included in the data set provided they were first attested in 2001 or later. Exceptions were the annual *Word of the Year* editions and the occasional thematic editions in which the editors would provide neologisms relating to a certain theme (often politics, as there would frequently be an influx of politics-related neologisms during election season). These were excluded because they were not an unbiased sample of neologisms. This process yielded 408 words.

Among the New Words provides an etymology for each entry and while in a small number of instances I found the etymology dubious I chose to defer to *Among*'s editors as I could not find information to determine with certainty the correct etymology between two possibilities.

The taxonomy of derivational categories used was the one defined by Algeo. (1980) It was selected because of its similarity to other research conducted by Cannon (1987) and Barnhart (1987), so that the results of this research were as compatible as possible with theirs as well as Algeo's and a comparison could be made.

4. Results

When the criterion of longevity was applied to all of the words collected from *Among the New Words* 34.56% of words failed to meet the standard, bringing the size of sample B down to 267 from 408. The results of the analysis of the samples are given in Table 1, represented as "2001-2013 Sample A (all words)" and "2001-2013 Sample B." Additional information on the *Among* sample is given in Tables 3 and 4.

Table 1 also gives the results of previous research on the productiveness of various word formation processes of English for various periods. The years in brackets under each study represent what period that study covered, that is to say during what years the words analyzed as part of that study first appeared. The studies are organized chronologically based on the initial year of the period with the earliest on the left and latest on the right, but the periods don't fit neatly and there is overlap.

Table 2 gives the results of Laurie Bauer's research on a sample of words first attested during the period 1880-1982 taken from the first edition of the *Oxford English Dictionary*.

Additional information on the research in Tables 1 and 2 is given after the tables.

Table 1. Word formation processes in various studies

	OED2 (1776-1989)	NEWS (1900-1988)	6,000 words (1961-1976)	Cannon (1961-1981)	Barnhart (1963-1973)	Simonini (published 1966)	BDC (1982-1985)	Longman (1989-1990)	2001-2013 Sample A (all words)	2001-2013 Sample B
CREATING	0.3	0	0	0.6	0	3	0	0	0	0
SHIFTING	23.4	30.8	2.52	19.7	14.2	14	9.6	19.4	20.59	20.22
SHORTENING	1.8	17.5	9.91	17.1	9.7	8	9.7	10	1.47	2.25
COMPOUNDS	19.8	12	35.51	29.6	29.8	50	57.6	36.3	33.82	33.71
AFFIXATION	32.3	25.6	43	24.2	34.1	14	15.9	18	11.03	15.73
BLENDS	3.3	1.1	0.91	1	4.8	3	0.5	9.8	30.15	23.6
LOANWORDS	18.8	6.9	7.41	7.5	6.9	8	6.2	4.3	2.94	4.49
UNKNOWN	0.3	0	0	0.4	0.5	0	0.5	2.2	0	0
OTHERS	-	6.1	0.73	-	-	-	-	-	-	-
SAMPLE SIZE	393	500	3,985	16,570	1,000	Not given	2,688	1,220	408	267

Table 2. Word formation processes from Bauer (1983)

	1880-1913	1914-38	1939-82	Total
NAMES	2.06%	3.59%	3.59%	2.98%
SHORTENING	1.82%	2.61%	4.68%	2.94%
COMPOUNDS	16.02%	15.82%	18.4%	16.7%
AFFIXATION	43.08%	47.63%	45.4%	45.14%
BLENDS	0.49%	2.28%	2.5%	1.78%
LOANWORDS	31.3%	22.3%	19.2%	25%
OTHERS (corruptions, word-manufacture, reduplication, onomatopoeic words, phrases)	4.37%	5.38%	5.93%	5.49%
SHIFTS	Not included in the sample	-	-	-
SAMPLE SIZE	824	613	641	2,078

Table 1

The studies given in Table 1 are the following:

OED2 (1776-1989) - 393 words from *The Oxford English Dictionary* (OED), 2nd edn., a sample consisting of the first form or sense on each of the 1,019 pages of volume 1 (A—Bazouki), provided that form or sense had an earliest citation date of 1776 or later, analyzed by Algeo (1998)

NEWS (1900-1988) - about 500 words beginning with the letter A and first attested after 1900, taken from NEWS (*New English Words Series*), a collection of some 5,000 words not in the OED or *A Supplement to the Oxford English Dictionary* (OEDS), as analyzed by John Simpson (1988)

6,000 Words (1961-1976) – 3,985 words that entered the language 1961-1976, taken from *6,000 Words: A Supplement to Webster's Third International Dictionary,* analyzed by Cannon and Mendez Engle (1979)

Cannon (1961-1981) – 16,570 words analyzed by Cannon (1987), consisting of 4,927 words in *The Barnhart Dictionary of New English Since 1963* (1973), 4,536 words in *Second Barnhart Dictionary of New English* (1980), and 7,107 words in the addenda of the 1981 printing of *Webster's Third International Dictionary* (1961)

Barnhart (1963-1973) – 1,000 words from *The Barnhart Dictionary of New English since 1963*, a sample of about one fifth of the words in that dictionary, analyzed by Algeo (1980)

Simonini (published 1966) – An article on English word formation which gives percentages for each word formation process, but does not give the source or size of its sample, nor what time span it coversBDC (1982-1985) – 2,688 words from volumes 1—4 of *The Barnhart Dictionary Companion,* analyzed by David K. Barnhart (1987)

Longman (1989-1990) – 1,220 words in *The Longman Register of New Words*, all the words in that dictionary, analyzed by Ayto (1989)

2001-2013 Sample A (all words) – The results of this research. All of the words from the issues of *Among the New Words* published 2001-2013, provided that the word was first attested in 2001 or later

2001-2013 Sample B – Sample B is formed by excluding from sample A the words which fail to meet the longevity criterion, that is to say it was not possible to find 10 instances of their use in the second year after they entered the language or later

Table 2

Laurie Bauer analyzed a sample taken from OEDS (1972-86) using the following method: Every fifth word was taken from each double page of the OEDS, providing that the word was not an addition to an entry in the first edition of the *Oxford English Dictionary* (OED1) and the word was not spelled in precisely the same way as a word already listed in OED1. Words with first citations before 1880 were discarded. This left a sample of 2,078 words. These were divided into three groups, according to the date of first appearance: 1880-1913, 1914-1938, 1939-1982. The dates for the divisions were chosen on political, not linguistic grounds.

Note that Bauer's sample does not include semantic shifts and the category "other" is made up of a large group of formation processes including corruption, onomatopoeia, reduplication, creation and others. Phrases were also placed in the "other" category, which would potentially include some formations that the research in Table 1 would categorize as compounds, but Bauer does not explain his taxonomy in much detail. Presumably all items made up of multiple unbound words were classified as phrases. Obviously he used a taxonomy that is quite different than the one used in this research, which is why it was given in a separate table. In the representation of Bauer's results in Table 2, whenever possible, categories were joined together to make the taxonomy closer to Algeo's, the one used in this research, to make for an easier comparison. For example. Bauer gave the numbers for abbreviations and other kinds of shortenings as separate categories, Algeo counted both under shortenings, so the numbers were added up and the sum was given in the shortenings category in Table 2.

The "others" category which is featured in the analyses of the NEWS research and the analysis of Webster's *6,000 Words* is comprised of various smaller categories which would in fact fit into one of the major categories in the taxonomy used in this research, but because the researchers only reported them aggregated as "others" it was not possible for the purpose of this research to disaggregate the different categories and distribute them appropriately.

Table 3. Field or topic the neologism is related to (Sample A)

CURRENT EVENTS	14.63
POLITICS	8.13
TECHNOLOGY	7.32
INTERNET	11.38
HUMOR	1.62
ADVERTISMENT	1.7
PRODUCT/BUSINESS NAME	4.07
BUSINESS	7.32
CULTURE (lifestyle, customs, tradition – not art, literature, music)	17.07
GAMES (card games, chess etc.)	0.81
SCIENCE	1.62
MEDICINE	4.07
VIDEO GAMES	2.44

MILITARY	0.84
FILM	9.76
MUSIC	1.75
NATURAL PHENOMENA	0.77
PERSONAL RELATIONSHIPS	0.86
FASHION	1.2
OTHER	3.25

Table 4. Field or topic the neologism is related to (Sample B)

CURRENT EVENTS	8.86
POLITICS	1.19
TECHNOLOGY	7.59
INTERNET	15.19
HUMOR	1.23
ADVERTISMENT	1.32
PRODUCT/BUSINESS NAME	5.06
BUSINESS	8.86
CULTURE (lifestyle, customs, tradition – not art, literature, music)	20.25
GAMES (card games, chess etc.)	1.27
SCIENCE	0
MEDICINE	6.33
VIDEO GAMES	3.8
MILITARY	1.31
FILM	8.86
MUSIC	2.53
NATURAL PHENOMENA	0
PERSONAL RELATIONSHIPS	1.27
FASHION	1.34
OTHER	3.8

The most noticeable trend in the 21st century relative to previous periods which the *Among* sample seems to suggest is the dramatic increase in the number of blends. The magnitude of the disparity raises suspicions that it might, in part at least, be due to a sampling bias. Algeo, who compared the results of OED2, NEWS, Cannon, Barnhart, BDC and Longman, suggested that blends might be overrepresented in Longman, relative to the other samples, because the *Longman Register of New Words*, from which the sample for that particular research was taken, includes a great number of voguish and nonce words, which tend to be created through blending. (1998, 86)

The assertion that blending tends to create a lot of voguish and nonce words seems to be supported by the fact that in the *Among* sample the number of blends drops off significantly once the criterion of longevity is applied. Looked at in terms of percentages, blends fall from 30.15% to 23.6%, while none of the other categories drop, or not beyond a single percentage point. However, even the 23.6% represents a substantial increase relative to previous periods, but one might wonder how much of a further decrease would be brought about by a stricter definition of longevity. It seems clear, though, that Algeo's assertion is correct - blending creates a disproportionate number of faddish words which do not have staying power.

Furthermore, Ayto, who is the editor of Longman, points out that Longman gathered its words almost exclusively from newspapers. 99% of its word stock is derived from newspapers, unlike the other samples which Algeo analyzed, which had a greater variety of sources. Ayto goes on to discuss the evidence, granted, some of which is anecdotal, which suggests that a large number of blends originate in newspapers. (2003) This is an assertion other linguists, including Algeo, have made. (1980, 271) Among the evidence Ayto mentions is a study he performed in which he examined a hundred blends originating after 1900 taken from the pages of OED in which he found that 54 of them had an earliest citation from a newspaper or some other kind of periodical publication. By contrast a sample of a hundred non-blend neologisms from the OED originating after 1900 had only 45 items with earliest citations in periodicals. (2003, 185)

When a word is accepted for publication in *Among the New Words* the editors try to establish the first instance of its use and that information is given with each entry. That information was gathered as part of this research with the interest of potentially giving an overview of what medium words are most frequently coined in. The aim was specifically not just to find earliest record of the word's use, but the moment of coining. Of course such claims are always very unreliable and involve a great deal of assumption and this was not one of the primary aims of this research, but the information was given with each entry so the data was gathered. Of course, as was expected, the editors were not able to determine the instance of use which could be assumed to have been the moment of coining for such a large portion of the sample that it makes the data gathered on this of limited use. Very often

the earliest record of a word would be found in a text that mentioned an appearance of a new word, but did not originate it. Again, this was to be expected, as most words are probably coined in everyday speech. Nevertheless, the data is given in Tables 5 and 6 as it does show something of relevance to the point about the prevalence of blends in newspapers.

Table 5. The medium the neologism was coined in (Sample A)

SPEECH (recorded in journalistic reportage)	4.41
NEWSPAPER OR MAGAZINE (including online editions)	48.53
BLOG	0.74
BOOK	0.74
TELEVISION	1.47
SONG	0.74
INTERNET FORUM	2.94
COMIC BOOK	0.74
UNKNOWN	39.71

Table 6. The medium the neologism was coined in (Sample B)

SPEECH (recorded in journalistic reportage)	3
NEWSPAPER OR MAGAZINE (including online editions)	43.68
BLOG	0
BOOK	1.15
TELEVISION	0
SONG	1.15
INTERNET FORUM	3
COMIC BOOK	1.15
UNKNOWN	48.28

The data shows that the *Among* samples are nearly in line with Ayto's blends-free sample in terms of the representation of words first attested in newspapers. An overrepresentation of words from newspapers could possibly account for the large percentage of blends in the *Among* sample, but such overrepresentation does not appear to be present in the sample. Newspaper words account for 48.53% of sample A and 43.68% of sample B. Again, Ayto's results were 45% for a sample that did not include blends and 54% for a sample made exclusively of blends, so an unbiased sample would be expected to be somewhat above 45%. There is a distinction, however, between Ayto's methodology and mine. If a text mentioned

a word, but it was clear that it did not originate it, I filed the word in the "UNKNOWN" category. The OED, which Ayto used, would cite it in the word's entry and, presumably, Ayto would categorize it according to the type of text it was.

It is very difficult to scrutinize *Among the New Words* for bias because its word stock is not gathered through a directed reading program, but through submissions from readers. Without knowing what kind of sources are being monitored, and how much of any particular type of source, it is difficult to analyze potential biases.

Ayto also conducted a study of frequency of blends in the lexicon of each decade of the twentieth century. Each decade was represented by a stock of 500 words that were first recorded in that decade. He does not give the source of the words. Interestingly, his study revealed a significantly greater ratio of blends then in any of the samples in Table 1 other than the *Among* sample and significantly greater than in any of Bauer's samples. Perhaps like Longman he gathered the words for his sample in large part from newspapers.

Oddly, Ayto does not give a numerical representation of the results, only a graphical one, but after the initial two decades which hover in the 5%-10% range, the 20s reach up to the 10%-15% range, and then, after the 30s, which Ayto refers to as "the decade of the blend," because they reach into the 20%-25% range, higher than any other decade, the ratio of blends does not fall below 10%, with the 60s nearly reaching 20%. The 90s end the century somewhere in the 15%-20% range. (2003, 185)

There is a significant number of what Bauer called "analogical formations" among blends. (1983, 96) An analogical formation, according to Bauer, is "a new formation clearly modeled on one already existing lexeme, and not giving rise to a productive series." A non-blend example of this would be *whitelisting* from *blacklisting* or *earwitness* from *eyewitness*. However, the blends found in the *Among* sample clearly show that analogy DOES give rise to productive series. The most productive example would be the many blends modeled after *blaxploitation.* There is a large number of genre names that are produced by blending a word with the word *exploitation*. The first of these was *blaxploitation* after which, by analogy, many others followed, such as *mexploitation, aussiesploitation* (or *ozsploitation*), *jewsploitation*, *pulpsploitation*. The analogy even moved out of film genres and into music genres and beyond with *popsploitation*, *geeksploitation*, *fansploitation* etc.

Another example found in the *Among* sample which speaks to the productivity of analogical formations is *cowboynomics*, a blend of *cowboy* and *economics* referring to the economic policies of George W Bush and made by analogy with *Reganomics*. Other words produced by the same pattern (but not in the *Among* sample) are *Abenomics* (economic policies of Shinzo Abe, the Prime Minister of Japan) and *Trumponomics*.

Plag makes the assertion that analogical formations should be distinguished from words formed through standard word formation processes. (1999, 20) An argument for this could be made on the basis of, for example, the phrase *homicide bomber,* which appears in the compounds category of the *Among* sample. Obviously, the phrase is made by analogy with *suicide bomber* and it is highly unlikely that *homicide bomber* would have ever existed if not for *suicide bomber*. *Suicide bomber* was made by joining *suicide* and *bomber*, but in *homicide bomber* the phrase was made with reference to *suicide bomber* by replacing *suicide* with a related, but contrasting concept. The formation process was different. It was not another instance of compounding but a modification of an existing compound. If we were to accept this argument and such a distinction were to be made it would somewhat affect the number of blends, but analogical formations are found in other categories, as well, such as the previously mentioned *homicide bomber*, which is a compound, so it is unclear whether it would bring the numbers of blends down relative to other categories, even if we were to accept Plag's view. A further problem would be encountered in that Algeo's taxonomy does not recognize the category of analogical formations and neither does any of the past research, so to use such a category would make a comparison with past research difficult.

The pre-21st century research also reveals a decrease in affixation in the 80s relative to the period of the 60s-70s. When compared to the BDC and Longman samples (which roughly cover the 80s) the *Among* sample shows either a continuation of that decrease, in the case of sample A, or a stagnation in the case of sample B. It should be said that most dictionaries, including dictionaries of new words, use some criterion of longevity when determining whether a neologism should be included in the dictionary, whether that be the lexicographer's subjective prediction on whether a word is likely to become a part of the lexicon or drop out, or, with more recent dictionaries, often corpus frequencies. Because of this it is perhaps more legitimate to draw a comparison between sample B and other research.

Almost all of the affixes in the 408 word *Among* sample are quite unproductive, yielding only one or two words. Among the suffixes the exceptions are *–er* and *–ing*, owning to the fact that certain neologisms produce further neologisms once they enter the language through suffixation with *–er* or *–ing*. *–ing* is used in creating gerunds like *Skyping* and *vlogging* and *–er* is used when a neologism that denotes an activity enters the language to produce a word for a person who performs that activity such as *vlogger* or *teak surfer*.

By contrast, Cannon, in his analysis of *The Barnhart Dictionary of New English Since 1963* (published 1973*)*, *The Second Barnhart Dictionary of New English* (published 1980), and the 7,107 words in the addenda of the 1981 printing of *Webster's Third New International Dictionary of the English Language*, found the most common suffixes in the production of English neologisms to be *-er*, *-ist*, *-ism*, *-ize*, and *-ic*, in descending order.

In his analysis of 1,000 neologisms from *The Barnhart Dictionary of New English since 1963*, Algeo found that, while prefixes are more numerous, they produce 15.6% of the new words while suffixes account for 18.5%. (1980, 274) This close to equal ratio of prefixes and suffixes is quite different than the numbers Bauer finds, although none of the periods Bauer investigates overlap neatly with Algeo's. Bauer's research does not include shifts so a comparison between his and Algeo's research in terms of how much of the new vocabulary is due to prefixation or suffixation in terms of percentage of all neologisms cannot be made, but the relationship which Bauer finds between prefixation and suffixation is roughly a 1 to 5 ratio in 1880-1913, and 1 to 4 in 1914-1938 and 1939-1982. Obviously, very different than Algeo's findings. In the *Among* samples, the ratios are almost exactly 1 to 3 for the A sample and exactly 1 to 2 for the B sample.

Shifts appear to be on a continuous rise throughout the 20th century and the trend seems to be continuing at the beginning of the 21st century. Functional shift or conversion is a very productive process, although only two types of conversion were found in the *Among* sample; noun to verb and verb to noun. Noun to verb was the most productive type, which is the same state of affairs that Simonini (1966) and Algeo (1980) found. Almost every instance of noun to verb conversion in *Among* involves proper nouns, the great majority of those proper nouns being product or company names (Enron, Google, Skype).

English used to maintain a reputation as a borrower language, but that seems to have changed by the 20th century and it seems like there might be a slight ongoing decline in the 20th and the 21st century, although differences of 2-3% should probably be considered within the margin of error given the small sizes of the samples investigated.

The findings on shortenings in the *Among* sample also represent a dramatic change with respect not only to the two chronologically closest samples, BDC and Longman, but all the other samples in Table 1, with the exception of OED2. However, Algeo makes some remarks that qualify some of the numbers in the Table. The dramatic deviation in the OED2, he says, is to be expected given that the OED lists abbreviations under the initial letter of the alphabet and the OED2 sample was derived by taking the first entry from each page of the first volume of the dictionary, provided that form or sense had an earliest citation date of 1776 or later. This would lead to undersampling of abbreviations. Furthermore, he points out that almost 7% of Cannon's shortenings are words that could be analyzed differently. (1998)

However, Bauer's research also differs drastically from all of the results in Table 1 and is much closer to the *Among* results, and Bauer's research doesn't even include shifts, which account for up to 30% of neologisms in the other research. Had shifts been included in Bauer's research the percentage of shortenings would be lower yet. Algeo's comment about the OED listing abbreviations under the initial letter of the alphabet applies to Bauer as well, since he derived his sample by taking every fifth word from every double page of the OEDS, but determining whether that leads to a bias towards or against their inclusion in the sample, or no bias at all is an overly complicated proposition that would not clarify things sufficiently to justify working out the complications involved in the work that would need to be done. What is telling, however, is that Bauer finds very few abbreviations in all three of the periods he examines, 0.4% in 1880-1913, 1.1% in 1914-1938 and 2.5% in 1939-1982. Algeo's analysis of Barnhart, which covers the period 1963-1973, finds 8.3%. This does suggest the possibility that something is amiss, although there are 25 years between Algeo's sample and Bauer's chronologically closest sample.

Furthermore, the lion's share of shortenings in Algeo's sample is abbreviations. Shortenings are 9.7% of Algeo's sample and 8.3% of the sample is abbreviations, while only 1.4% are

backformations. Bauer's findings with regards to non-abbreviation shortenings are similar; 2.3% in 1880-1913, 2.3% in 1914-1938 and 3.3% in 1939-1982.

It does appear that the reason for the great disparity in shortenings between Algeo's and Bauer's sample, and, presumably, between Bauer's sample and the other samples in Table 1, is the same as the reason for the disparity between OED2 and the other samples; undersampling of abbreviations due to the method of sampling. In the *Among* samples, abbreviations account for 0.74% of sample A and 1.15% of sample B. Again, this would suggest some sort of bias against the inclusion of abbreviations in the *Among* sample.

Algeo has comments with regards to compounds, as well. He points out that the very high number for compounds in the BDC is partly due to its practice of listing forms like *telework, teleworking, teleworker* as independent compounds, whereas the other sources in Table 1 would treat them as related to each other through affixation or backformation, which is how I analyzed the *Among* sample, as well. (1998) Following the latter practice however leads to the problem of establishing which form was first and originated the others, because this affects the analysis and ultimately the results the analysis produces. *Among* lists dates for when each form in any such cluster of related words was first attested and I based my analysis on that, but such related words would have originated very quickly one after the other in a short span of time so one can easily imagine that the order in which they were recorded for the first time in texts where lexicographers can find them and construct a chronology need not necessarily reflect the order in which they were actually coined. Not to mention the fact that the coining of all three of the words may have happened in spoken language.

In the *tele-* trio that Algeo cites as an example, if we take it that *telework* is the originator that would mean we would count two instances of suffixation in *teleworking* and *teleworker*. If we take *teleworking* as the originator we would count an instance of backformation in *telework* and an instance of suffixation in *teleworker*, that's presuming *teleworker* was derived from *telework*, otherwise further complications arise.

Algeo further points out, with regards to the high number of compounds in the BDC, that idioms like *keep one's feet to the fire,* were counted as compounds. The OED's number, on the other hand, is affected by the fact that it lists compounds as run-in rather than main

entries, which, again, would lead to undersampling. The low number of compounds in NEWS he finds puzzling, but doesn't have a definitive explanation for. He suggest that perhaps part of the explanation is that Simpson counted only noun compounds of three patterns (noun + noun, adjective + noun, verb + noun) and adjectives of two patterns (noun + adjective, adjective + noun) and that the "other" category, which accounts for 6.1% of the sample, may hold a substantial number of compounds of other kinds. (1998) However, even if the "other" category was made up entirely of compounds of other kinds that would still get NEWS up to only 18.1%; much less than any of the others.

Simonini (1966) is another drastic deviation from the norm in terms of compounds, but, again, he doesn't give the source of his sample so it is impossible to scrutinize it.

Among is in line with Longman, the chronologically closest sample, showing no significant change in rate of compounding. Bauer's research shows a steady decrease in neo-classical compounds, albeit the effect is rather small starting from the 1914-1938 period when neo-classical compounds accounted for 5.1% of the whole sample (not of compounds), to 3.6% in 1914-1938 and 2.3% in 1939-1982. The *Among* sample shows a continuation of this trend with neo-classical compounds making up 0.75% of the A sample and 1.15% of the B sample.

5. Discussion

Algeo concludes his survey of neologisms in 1980 with the following:

> Nearly two-thirds of the new words (to be precise, 63.9 percent) were composites - that is, they were compounds or forms derived by affixation - new words constructed by combining morphemes already present in the language. Compounding and affixation were doubtless also the chief sources of new words in the Old English period, more than ten centuries ago; they remain the favored processes for adding to the lexicon. Perhaps they are the dominant processes in all languages, but without studies of neology in Chinese, Swahili, Hebrew, and a great many other languages, we cannot be sure of that. What we can be fairly sure of, is that modern English preserves an historical continuity with its Germanic ancestry in a strong preference for compounding and affixation as a source of new words. (Algeo 1980)

Cannon (1987, 265) seems to imply that this view that English vocabulary grows mostly through compounding and affixation was commonly held. Indeed, the two samples included in Table 1 which cover roughly a similar period as the one Algeo analyzed, 6,000 Words (1961-1976) and Cannon (1961-1981), find that compounding and affixation are the two dominant processes of wordmaking which together account for a very large share of newly produced words, although exactly how large varies.

This, however, appears to begin changing towards the end of the 20th century. By 1989-1990, the Longman sample, the two most prolific wordmakers are no longer compounding and affixation, but compounding and shifting, albeit by a very small margin. While compounding is more or less maintaining its place and significance, affixation appears to be losing ground in favor of blending.

This trend seems to be continuing with the turn of the millennium, and blending in particular appears to have become much more frequent. The upturn in the frequency of blending is so dramatic, however, that it does raise suspicion about possible bias in the sample. It should be said, however, that, according to Ayto's research, there had been a previous "decade of the blend," the 1930s, as mentioned in the preceding chapter, which lands in about the same area as where sample B of Among puts the 2001-2013 period. Again, Ayto, doesn't give

numerical results, only graphical ones, but he puts the 1930s in the 20-25% range; sample B of *Among* is 23.6% blends.

Furthermore, two decades after the "decade of the blend," Ayto has the 1950s at 10-15%. That is about the same difference in the same time frame that exists between Among sample B and the Longman sample, which covers 1989-1990. (2003)

This makes the *Among* result seem plausible. It needs to be noted, however, that Ayto's research shows a significantly higher percentage of blends than any of the other research except for Longman and *Among* itself. It is quite possible that Algeo's explanation of the high percentage of blends in Longman applies to *Among* as well. The high percentage of blends, according to Algeo, is due to the inclusion of a great number of nonce and faddish words. This implies that the other dictionaries do not include as large a number of such words. It is quite possible that *Among* has a substantially more liberal policy with regards to the inclusion of words that very much seem either of the moment or faddish than an average dictionary of new words does. As was mentioned in the previous chapter, once the criterion of longevity is applied the percentage of blends goes down from 30.15% to 23.6%. Which suggests that, indeed, there is a significant number of faddish words among the blends. One can wonder whether the percentage would undergo a further significant decrease if one were to discount all words that do not show a substantial frequency in corpora five years after they are first attested. Unfortunately the fact that the period covered in the *Among* sample is so recent did not allow for putting that question to a test.

Most editors of dictionaries of new words do exercise judgment about a word's potential longevity and are not inclined to include in their dictionary words that they expect will quickly disappear from the language. It is understandable that *Among the New Words,* being a regular feature in a journal intended for linguists, would have a greater tolerance for peculiar and unusual formations and short lived novelties than a dictionary.

If there is indeed a difference in the readiness to include nonce or faddish words between *Among the New Words* and the average dictionary of new words, that could account for the dramatic difference in the percentage of blends found in the *Among* sample and the other research since all of the other research used dictionaries or addenda to dictionaries.

The other category in which *Among* diverges significantly from most of the other research is shortening, with 1.47% in sample A, and 2.25% for sample B. The only two samples which agree with those numbers are Bauer's and OED2. With the exception of those two, the numbers range from 8% to 17.5% in the other samples. However, as Algeo pointed out and as has been discussed in the previous chapter, the numbers for OED2 and Bauer are certainly highly biased due to the way in which the sample was derived. (1998, 85) The magnitude of the disparity between the number in *Among* and the numbers in other research does suggest the possibility that it, too, is a reflection of bias rather than a genuine change in the derivational patterns of English. It is difficult, however, to imagine what in the way that *Among* goes about their work would produce a bias against the inclusion of abbreviations.

These kinds of uncertainties about the sample are one of the difficulties in doing this sort of research. The researcher is always forced to analyze and guess at the possible biases that exist in his sample due to the editorial policy of the dictionary he drew from (or whatever happens to be his source), what the sources were that the dictionary used, whether there was a good enough ratio of different sources representing different kinds of language use and, as it happens, how strict is the editorial policy on the inclusion of nonce or faddish words. All of this ends up raising, often unanswerable, questions about the accuracy of his findings. The progress of computers, however, promises to improve upon these weaknesses of past research.

6. Conclusion

The analysis of neologisms from the period 2001-2013 shows, most notably, a decrease in the productivity of shortening and an even greater increase in the frequency of blending. The increase in blends appears to possibly be a continuation of a trend starting in the late 20th century in which affixation began to lose ground in favor of blending. This is quite a notable finding as it represents a break from a pattern of derivation which English appears to have held to since its inception in which affixation has continuously, along with compounding, been the most frequent method of deriving new words. Questions, however, do remain about the accuracy of the findings because of how drastically they diverge from the findings of all of the other research.

7. References

Algeo, John. "Vocabulary." *The Cambridge history of the English language Volume IV 1776-1997*, edited by Suzanne Romaine, Cambridge University Press, 1998, 57-91.

Algeo, John. "Where do all the new words come from?." *American Speech* 55.4 (1980): 264-277.

Ayto, John. "Newspapers and Neologisms." *New media language*, edited by Aitchison, Jean, and Diana M. Lewis, Psychology Press, 2003, 182-187.

Bauer, Laurie. *English word-formation*. Cambridge University Press, 1983.

Cannon, Garland, and Beatrice Mendez Egle. "New borrowings in English." *American Speech* 54.1 (1979): 23-37.

Cannon, Garland. *Historical change and English word-formation*. New York: Lang, 1987.

Cook, Paul, et al. "A lexicographic appraisal of an automatic approach for detecting new word senses." *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. 2013.

Michel, Jean-Baptiste et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science (New York, N.y.)* 331.6014 (2011): 176–182. *PMC*. Web. 26 Sept. 2016. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742

O'Donovan, Ruth, and Mary O'Neill. "A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary."*Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*. 2008.

Plag, Ingo. *Morphological productivity: Structural constraints in English derivation*. Vol. 28. Walter de Gruyter, 1999.

Plag, Ingo. *Word-formation in English*. Cambridge University Press, 2003.

Renouf, Antoinette. "A Word in Time: first findings from the investigation of dynamic text'." *English Language Corpora: Design, Analysis and Exploitation, Rodopi, Amsterdam* (1993): 279-288.

Renouf, Antoinette. "Making sense of text: automated approaches to meaning extraction." *International online information meeting*. 1993b.

Simonini, R. C. "Word-making in present-day English." *The English Journal* 55.6 (1966): 752-757.

Taylor, John R., ed. *The Oxford handbook of the word*. OUP Oxford, 2015.