Morphological regularities and patterns in English word formation

Kateryna Krykoniuk

Thesis submitted in partial fulfilment of the degree of Doctor of Philosophy

Centre for Language and Communication Research

School of English, Communication and Philosophy

Cardiff University

October 2021

Abstract

The aim of this study is to identify the main morphological constructions, patterns, regularities and paradigms involved in English word formation. The study is based on a sample of 32,000 words, which are analyzed morphologically and etymologically by means of formal morphological analysis and which, together with their corresponding metalinguistic morphological patterns, constitute a morphological metacorpus—the key practical undertaking of this thesis. With this metacorpus, different quantitative and qualitative characteristics of the English lexicon have been obtained.

The main methodology of this thesis is formal morphological analysis (Bratchikov 1958; Tyshchenko 1969, 2003). This method includes distinguishing the morphological elements of a lexeme and encoding them with a metalanguage, specifically designed for this purpose. The inner morphological structure of each word is verified with the help of the Oxford Etymological Dictionary. The metacorpus is classified and analyzed with the help of Visual Basic macros and the method of matrix optimization, which reveal precise morphological constructions and regularities of English word formation. Their different quantitative and qualitative aspects are further explored with different statistical techniques (e.g. graph analysis, regression, relative entropy and clustering). Thus, generalizations about English word formation are made on the basis of the analyzed data and with the metalinguistic terminology framework that has emerged in the course of the analysis, which means that some aspects of the description are new to the field and not previously discussed in the literature. Nevertheless, the thesis findings also verify some assumptions about English word formation postulated in other word-formation theories. The major problems substantiated in this thesis concern (a) constraints of suffix ordering (Plag 1996), (b) Unitary-Base Hypothesis (Aronoff 1976), (c) the impact of type frequency of morphemes and of morphological patterns on English word-formation grammar, and (c) the degree of the expression of such typological morphological characteristics as agglutination, isolation and fusion (Greenberg 1960; Sapir 1921) in English word formation. In addition to their theoretical value, the findings of this thesis have potential to be used for designing learning vocabulary at different stages of second language acquisition, for teaching English derivational morphology and for developing a morphological parser.

Acknowledgements

A journey—whether it is physical, spiritual or scientific—depends not only on our own resources and will, but also on the external circumstances and people who accompany us and come our way. Although, in any long journey, there are both moments of doubts, disorientation and failure and moments of confidence, clarity and success, it is mostly the people who shape the memory of that journey, allowing for the enjoyable reminiscences to last and the unease to fade away.

My scientific journey which has led to this doctoral thesis has been an extraordinary one, for the most part because of the people who supported me during this time. Hence, I would like to devote the first page of my thesis to acknowledging everyone without whom this undertaking would not have been possible. First and foremost, I want to thank Cardiff University for providing a welcoming and rich (both in terms of content and material resources) environment for academic growth. During my interaction with its different departments, I was touched by the professionalism, attentiveness and compassion of all the university's staff. My special gratitude goes to my supervisors Dr. Lise Fontaine and Dr. Michelle Aldridge-Waddon for their constant scientific and moral support, and their wise, motherly guidance. In particular, I would like to show appreciation to my first supervisor, Dr. Lise Fontaine, who has vigorously encouraged me not to be afraid to explore new horizons. Further, I am indebted to Prof. Nikolai Leonenko for consulting me on mathematical and statistical problems that have arisen in the course of this research, as well as to Prof. Geeraerts, whose company I had the chance to enjoy during my research stay at KU Leuven, for offering a heuristic insight on how to overcome the excessive formalism of the thesis methodology. In addition, I would like to express my gratitude to Dr. David Schönthal for his inspirational introduction to the world of English academic writing and for his meticulous comments on the earlier drafts of this thesis. On the financial side, I want to acknowledge the generous contribution of the Houtan Foundation led by Dr. Mina Houtan and the Professionals Aid Guild (PAG) and would like to thank them for the partial sponsorship of this research.

Finally, I wish to pay tribute to my family for their enormous support over the course of my doctoral journey. Specifically, I am tremendously grateful to my husband Amir Rostambeihi who has never lost his faith in me, even in the moments when I seemed to lose mine. I hope that this piece of research will substantiate kindness, trust and assistance of all those I have mentioned.

Table of Contents

List of Tables

Table of Figures

Abbreviations and annotations

N	noun
Verb	verb
Verb*2	past tense verb
Verb*3	past participle
Aj	adjective
Num	numeral
Ad	adverb
Prep	preposition
Conj	conjunction
Intj	interjection
Pron	pronoun
Abbr	abbreviation
BM	bound morpheme
CC	conversive class
a	affix
C	word base
{}	boundaries of a morphological construction
{{}}	boundaries of a morphological meta-construction
R	radical/word
R_A	formed by analogy
R_ACR	acronym
R_AL	altered lexical item
R_APH	aphetic
R_BF	back-formation
R_BL	blending
R_C	corruption
R_CN	conversion
R_D	dialectic form
R_E	echoic/expressive
R_EPH	euphemism
R_F	frequentative form
R_I	imitative
R_IN	inversion
R_O	onomatopoeic word
R_RD	reduplication
R_S	semantic split
R_SRT	shortened

R_SX	syntax formation
R_V	phonetic variant of a word
R-ind	individual coining
R-pn	formed from a proper name
RQ	research question
L	Level
(')	2 & 3 levels of word formation
|*a*|	allomorph of an affix
TF	type frequency
TokF	token frequency
P	potential productivity
P.	expanding productivity
TTR	type-token ratio
TV	type valency
MQ	MorphoQuantics
ML	MorphoLex
C	CELEX
OED.s	the sample of words used in this research
OE	Old English
OEG	Old English from Germanic
OF	Old French
OFL	Old French from Latin
GT	Gothic
FG	French from Greek
FL	French from Latin
L	Latin
AN	Anglo-Norman
O	Origin

1 Introduction

First, this chapter introduces the overall linguistic and philosophical context of the current thesis. Second, it discusses the goals of the study and its research questions, and gives an overview of each chapter.

1.1 The general linguistic and philosophical context of the current thesis

Language is still a phenomenon we do not yet fully understand, although in recent decades there has been incredible progress in the fields of linguistics and language processing. We know that language is governed by some internal rules, which we call 'grammar' and which allow us to build meaning from the smallest linguistic units. A rule-based approach,[1] initiated by Chomsky, states that the language production is organized linearly or hierarchically: phonemes combine into words and words into sentences following some deep underlying rules. They operate sequentially or parallel in time: some of them through language production, and others through social or cognitive constraints. In accordance with this approach, the aim of linguistics is to discover these hidden cogs, wheels and barrels of language, which are as well-tuned as an intricately designed clock mechanism.

However, how good is the concept of language as a rule-based system? Language is strongly influenced by our tastes, trends and social environments and displays a high degree of divergence from rules, and Universal Grammar does not seem to capture these effects. Because of this external influence, there are innumerable exceptions in language that do not lend themselves to the mechanistic rule-based explanations. Within the strictly rule-based approach, it is difficult to account for metaphors, metonymies, idiomatic expressions and widely used deviant collocations (e.g. *Him a lawyer?*, *It's amazing the lies he told her* cited in Hilpert 2014) that comprise a significant portion of language. The Achilles heel of this approach is that it mainly focuses on language's form whose rules allow for the production of grammatical but meaningless sentences (such as Bertrand Russell's 1940 *Quadraplicity drinks procrastination*). Thus, when the rule-based approach attempts to integrate meaning into its grammatical description, it becomes as cumbersome and complex as language itself, which contradicts an initial Chomskyan objective set

[1] The rule-based approach is challenged by cognitive and construction-grammar theories.

for linguistics: to pinpoint an optimal universal system able to explain a phenomenon in its entirety. This is not to say that linguists should avoid studying language's form and using different formalization techniques to capture regularities in language: formalization is a powerful tool to inform researchers about language's structure, which are inevitable in the performance of different statistical analyses. However, what is important here is a shift of the focus from an abstract rule-based structure of language to an empirically-grounded model that attempts to account for a multifaceted nature of language.

Thus, linguists realized that the methodological tools of generative grammar had reached their limit—yielding, nevertheless, remarkable findings for historical records. The end of the hegemony of generative grammar is metaphorically expressed in the following rhetorical question by Dąbrowska (2015: 12): "Is it [generative grammar] a fruitful approach? (Or perhaps a better question might be: Was it a fruitful approach?)". Nowadays, research focus shifted to the problems that for long years remained unexplored. There is no language without humans. Hence, it is not feasible to understand language without considering human biology, psychology and society for whose needs language is produced. Indeed, language resembles a living being: it is constantly changing and displays many unpredictable patterns. What if these patterns can be captured by probability and statistical approaches which have proven themselves exceedingly helpful in other scientific fields, such as biology and physics, and which might point to some new associations and relations between different aspects of language? To explore this question, new subfields of linguistics have emerged sharing its borders with biology, statistics, as well as social, neuro- and computational sciences. Linguistics has become more diverse and heterogeneous.

For the most part, this new type of linguistics views language not as an abstract entity, but rather how language is realized in its innumerous instances of various modalities. This collection of language's instances has been termed a 'corpus'. It provides rich evidence for various linguistic phenomena and allows linguists to take a more objective stance, avoiding conclusions based only on their own intuition and insight. Corpora have become central to much linguistic research. Corpora have embodied the revolutionary and controversial idea that language is not a system located somewhere in Plato's idealistic realm. It is a summation of its individual occurrences.

To compile and explore corpora, linguists have equipped themselves with new methodologies and tools developed in cooperation with specialists from other fields. Linguistics—a once highly isolated and narrowly specialized area—has widely opened its doors to

mathematicians, biologists, statisticians, clinicians, programmers and software developers. The problem of language has been recognized as part of a more general and mystical problem of human cognition. In order to understand language, scientists need to first understand human consciousness. Hence, recording different activities of the brain to different linguistic stimuli, collecting frequencies of various linguistic events and units from corpora, as well as performing different psychological experiments and statistical analyses of data have all become a key strategy for explaining different linguistic behaviors and changes. The current research also follows this multidisciplinary trend by incorporating new entropy methods to the study of word formation.

Yet, it is still a difficult task to capture the phenomenon of language in all its entirety, even when approached from different angles: rule-, cognitive-, corpus-based or all of them combined. If we were to ask linguists why this is the case, the most frequent answer we would probably hear is that it is because of the complexity of language. However, this is not the whole story. Language is complex indeed, but it is the *non-linearity* of its structure that makes language extremely hard to study. As Gleick (1987: 24) puts it: "Non-linearity means that the act of playing the game has a way of changing the rule [… it] is like walking through a maze whose walls rearrange themselves with each step you take". Thus, similar to chaotic systems, in the study of language, it is a 'demon of non-linearity' (ibid.) that countermands the unified description.

In the linguistic realm, this kind of non-linearity is most vividly manifested in the relation between the content and expression planes of language (Hjelmslev 1961). The ideal linguistic theory (if such a thing will be developed in the future!), thus, congruently and optimally unifies these planes on the syntagmatic and paradigmatic axes of language. It establishes various sorts of connections between units of paradigms and explains their syntagmatic realization. It seeks to answer the question of what is formally distinguishable and what is formally identical in paradigms and syntagmas.

Hence, paradigms—in the sense of a set of related linguistic units that allow for their mutual substitution which eventually leads to "a contrast on the relevant linguistic level" (Bauer 2019: 153)—have become central to modern linguistics. In particular, this trend is obvious in the study of morphology and word formation which is the main research theme of this study. Although, according to Bauer (2019: 173), there are only two types of paradigms in morphology—inflectional and derivational—there is also a large number of sub-paradigms, because "there are many types of relationships which can count as paradigmatic".

How, then, are paradigms helpful in dealing with non-linearity of language and in uniting the content and the expression planes of language? Through *isomorphism*. This term, initially coined in mathematics, was transposed to linguistics by Kurilovich (1949: 79) to denote a deep parallelism or similarity between various linguistic structures and units of the content and expression planes (e.g. parallelism between syllables and sentences). Isomorphism implies that some paradigms repeat themselves on different linguistic levels, from micro- to macro-levels—starting with the phonological level up to the syntactical or even pragmatic. Therefore, it may be that universality of grammar lies not in a set of some basic rules which transform phonological material into meaningful utterances, but rather in a panoply of paradigms—paradigms that are condensed abstractions and skeleton of grammar, paradigms that disclose regularities in the arrangement of linguistic units, paradigms that account for all the existing typological diversity of languages, paradigms that are isomorphic, and, despite the seeming change in linguistic forms, their underlying construct and inner relations remain the same. Speaking in the language of geometry, such paradigms have a fractal structure with a self-referential and self-replicative quality.

1.2 The goals and research questions of the current thesis

Thus, in light of the above discussion, the main objective of this thesis is to identify the fundamental formal paradigms, patterns, relations and regularities that determine how morphemes are organized in English word formation. This objective determines two idiosyncratic features of this study: it has a broad scope and it is empirically-driven. The bigger picture of English word formation has become possible with a large sample of words (32,000) and by adopting a wide theoretical perspective. The empirically-oriented nature of this study implies that many of its concepts and discussions have arisen from empirical observations and are new to the literature of word formation (e.g. the concept of ‘type valency’). With such a frame of reference, this thesis tries to partially address the linguistic problem outlined by Bybee (2007a: 68) that “[m]orphological systems have not in general been subject to explanation, nor have they been treated as natural objects whose properties follow from their functions”.

In order to reach the main objective of the thesis, the following more specific research questions (RQ) will be addressed:

1) What is the general picture of word formation the English lexicon?
2) What formal morphological regularities, patterns, constructions and paradigms are found in English word formation?
3) What are the effects of type frequency[2] in English word formation?
4) How English word-formation processes evolved over a period of time?
5) What typological and clustering characteristics do we find in English word-formation processes?

The empirical nature of this study defines the choice of its research methodology which is mainly *inductive* (i.e. a generalization about English word formation is made on the basis of the analyzed data). However, some empirical observations of this study have also substantiated theoretical problems of English word formation with regards to (a) constraints of suffix ordering and the interaction between affixes and word bases (e.g. base-driven (Plag 1996) and Unitary-Base (Aronoff 1976) hypotheses), (b) the impact of type and token frequency of morphemes and of morphological patterns on English word-formation grammar (usage-based theories), and (c) such typological characteristics of English as isolation, agglutination and fusion (Sapir 1921; Greenberg 1960). These are *deductive* methodological threads of this thesis.

Heretofore, the discussion has been focused on the general methodological ground of this research. Narrowing the focus to practical techniques for reaching the set objectives, in this research I deploy different formalization and statistical tools. In particular, a sample of lexemes is formalized with the help of formal morphological analysis, which converts lexemes into morphological patterns. This procedure is done manually while consulting the OED for the etymology and the morphological structure of each lexeme. Further, morphological patterns are complemented with the year of a lexeme's first appearance in a text. Then, the morphological patterns are classified and organized using macros written in the Visual Basic (VB) in the form of a morphological metacorpus—a set of formally represented morphological patterns of English words. Lastly, different statistical analyses are used to explore various aspects of this metacorpus (e.g. graph networks, correlation, cluster analyses and Poisson regression).

[2] The effect of frequency is a concept introduced by Bybee (2007a), which is understood as an impact of the high/low type or token frequency of a particular unit or pattern on a grammatical/cognitive representation of language.

1.3 The overview of the chapters

The order of the above-mentioned methodological procedures determines the organization of this thesis. In particular, Chapter 2 examines the history of the development of morphology and word formation from the time of coinage of the term 'morpheme' by de Courtenay up to the present day. Focus is placed on such theoretical frameworks as structuralism, generative and construction grammars, Marchand's synchronic-diachronic theory, Dokulil's onomasiological approach, usage-based theories and, as I term it, a cognitive stat-rule approach, because these theories have made a significant contribution to our understanding of morphology and word formation, in particular. As indicated earlier, this broad scope of the literature review has been chosen deliberately. The first reason is that this study is empirically-driven, and an extensive awareness of various theoretical stances allows for a better and deeper explanation of the research's quantitative and qualitative findings. Secondly, with such a comprehensive perspective, it is much easier to avoid reinventing the wheel, namely, coining terms and notions which have already been introduced to linguistic theories—the risk inherent to a narrow-scope stance.

After exploring how a linguistic understanding of morphology and word formation has changed over the course of time, this thesis, then, looks at the methodological tool used to formalize the data under study in this thesis. Hence, Chapter 3 is devoted to formal morphological analysis—the formalization technique that isolates morphological information about lexemes. Its metalinguistic symbols are fully explained and justified. Because this method implies a strict categorization of each morpheme and morphological process, its biggest challenge is how to qualify, for example, unproductive obsolete morphemes, unconventional word-formation processes or lexemes with an uncertain morphological structure. Further, another challenge of assigning morphemes and morphological processes to a particular category—the challenge intrinsic to the English language—is how to draw a clear-cut distinction between morphemes which are part of foreign word-formation processes and the same morphemes which have become part of English native morphology. For instance, consider the words *collection* and *authentication*. Their morphological structure looks identical, comprising the verb and the suffix *-ion* (*collect* + *-ion*, *authenticate* + *-ion*). However, these words have different etymological histories: the former has been borrowed from French as a full word with the first record in 1387, whereas the latter has been formed within English in around 1612. Hence, one of the ways to overcome this problem, adopted in this research, is to complement the formal morphological analysis with a

diachronic perspective—as put forward by usage-based theories. In addition to precluding the artificiality in morphological parsing, these tactics also make it possible to track the historical process of morpheme development and to answer the questions of when originally Latin and French morphemes came to be perceived as native morphological building blocks, and which dynamics can be identified in these processes. Lastly, Chapter 3 also introduces the terminology used in this research to describe patterns and regularities of English word formation.

Chapter 4, subsequently, is concerned with the description of the statistical methods used to quantitatively explore data reformatted with the formal morphological analysis. They include Poisson regression modelling, cluster analyses and hypothesis testing with three Kullback-Leibler Divergence estimators. In addition, the principles of graph theory and their application to morphological formulas are discussed. The goal of this chapter is to acquaint the reader with the statistical tools deployed in this research, and to show how the chosen methods suit the study's dataset. For this reason, the description is kept more general and is aimed at readers who have some background in statistical and graph analyses. Nevertheless, although this review of the statistical methods is concise, the interested reader is referred to further literature where they can find more information on the topic.

Chapter 5, then, presents an overall structural analysis of the morphological metacorpus. It gives various quantitative and qualitative characteristics of different word-formation aspects (e.g. the proportion of different word classes in the morphological metacorpus and their origins). The major focus is placed on simplexes, which are taken to form a zero morphological level of English word formation. The chapter also identifies the type-frequency effect of suffixes on the orthographical/morphonological properties of word bases.

Subsequently, Chapter 6 introduces morphological regularities and paradigms of English word formation by word classes. It profiles morphological constructions in the form of optimized matrices which have led to the identification of different morphological regularities and in the form of network graphs which have visually captured the formal structure of paradigms. These network graphs also embody the main typological features of English word formation. Further, the level structures of the paradigms are presented with the help of graphs, and their qualitative and quantitative features are considered.

This discussion is followed by Chapter 7, which explores different aspects of the morphological metacorpus with statistical methods. It establishes two effects of type frequency in

English word formation: first, the type-frequency effect of suffixes on their type valency, and, second, the type-frequency effect of word bases on suffix combinations. Further, this chapter investigates the development of the type-frequent word-formation processes in a timescale and identifies the clusters of morphological constructions showing similar characteristics.

Lastly, Chapter 8 summarizes the findings. It lists the most important findings of this research and draws theoretical conclusions as to the nature of English word formation. The chapter also outlines the limitations of this study and its potential for further development.

To summarize, in the current chapter, I have briefly introduced the thesis. First, the present-day linguistic trends have been outlined, and it has been shown how the thesis fits into this context. After presenting the thesis's main objectives, the methods to reach them have been generally sketched. Then, the structure of the thesis has been put into focus, and every chapter has been briefly introduced. The thesis provides innovative solutions to several linguistic problems of word formation, and its morphological metacorpus can serve as a basis for some sort of morphological parser in English, if the study is taken further into the domain of natural language processing.

2 Literature review

One of the most generally accepted definition of morphology has been formulated by Melčuk (1997: 30): "Morphology is a part of linguistics that studies word in all its relevant aspects".[3]

The object of morphology concerns the minimal bilateral units of language (having both meaning and form), i.e. morphemes and sets of these units that have a number of specific features. Thus, it would be right to say that some linguistic problems of morphology stand in the place between a morpheme and a word-form. However, the latest discoveries in neuroscience and artificial intelligence have added a new range of problems to the study of morphology, namely how the morphological plane of language is organized and represented in the human brain. What are the mental and biological mechanisms which control the organization of morphemes in words and the morphologic representation of grammar in syntax? Seen from this aspect, morphology plays a crucial role in bridging the gap between phonology and syntax. Shedding light on these problems may help to understand the phenomenon of language as a whole.

In this chapter, I will take a look at the history of morpheme and morphology and what place these notions occupy in different linguistic theories that have emerged with the development of linguistics. To present the complete picture of the field, I will describe how the understanding of morphology (which, in this thesis, is understood as a general term for word formation and inflection) has evolved in linguistic studies over the years. In fact, every linguistic framework has brought a new look on morphology and, consequently, on word formation. Hence, a narrow focus on only the development of word formation would overlook the wider connections of word formation to the grammatical mechanisms of language.

Taking such a broad perspective is important for various reasons. First, in my view, modern linguistics has accumulated a vast amount of knowledge. This is why it would benefit more from unifying the existing concepts and valuable insights towards creating an optimal unified *metatheory of language* (e.g. Tyshchenko 2000), rather than introducing another incomplete theory to the field with potential for future development. I view my research as the first steps towards this goal. Hence, it is crucial to be familiar with the history of morphology. Secondly, some concepts,

[3] The quote is my translation from Russian: "Морфология есть часть лингвистики, занимающаяся словом во всех его релевантных аспектах".

terms and arguments used in this study are better understood if readers have some background knowledge of how the field of morphology has evolved.

To sum up, in Section 2.1, I will shortly describe the emergence of the term 'morpheme'; Section 2.2 is dedicated to morphology in early Western European tradition; Section 2.3 discusses morphology and word formation in Distributional Linguistics. In Section 2.4, morphological views in Copenhagen School of Structural Linguistics are considered, and Section 2.5 gives an account of morphological assumptions in Prague Functional Linguistics. Section 2.6 deals with the word formation approach of Hans Marchand, and Section 2.7 with morphological and word-formation theories in generative grammar. Finally, Section 2.8 briefly introduces Construction Morphology, and Section 2.9 reviews the basic principles of an onomasiological approach to word formation. Section 2.10 looks at Usage-based Theory, and Section 2.11 at a cognitive stat-rule approach. Finally, Section 2.12 summarizes the history of the word-formation studies.

2.1 The birth of the term 'morpheme'

The term 'morpheme' was coined by Baudouin de Courtenay. He defined it as "a part of the word, which is endowed with psychological autonomy and is for the very same reason not further divisible. It consequently subsumes such concepts as root […], all possible affixes […], endings which are exponents of syntactic relationships, and the like" (de Courtenay 1895[1972]: 153).

Noticeably, 'the father of modern linguistics', Ferdinand de Saussure, was acquainted with the term 'morpheme' introduced by de Courtenay, but avoided it, likely because of its unclear and ambiguous definition, linking a sign only to a word. This fact led Anderson (2015: 3) to suggest that two different attitudes toward morphological structure in modern linguistics—that is, 'morpheme-based' analysis and 'rule-based' conception, which define the field today—have derived from these distinctive linguistic paradigms of de Courtenay and de Saussure.

Another prominent linguist, Lev Scherba was the first to raise the question of how the distinction between lexicon and grammar was visible on the morphological level of grammatical analysis. He invented a Russian sentence in around 1928 (Uspenskyi 1962), which is akin to Lewis Carroll's Jabberwocky: *Гло́кая ку́здра ште́ко будану́ла бо́кра и курдя́чит бокрёнка* (*Glokaya kuzdra shteko budlanula bokra i kurdyachit bokryonka*). In this sentence, all lexical morphemes, i.e. word stems, are meaningless combinations of phonemes, whereas the lexical-grammatical morphemes, i.e. suffixes, are used correctly. Although a native Russian speaker may not

understand the complete meaning of this sentence, as there are no such word stems as *glok-*, *kuzdr-*, *shtek-*, *budl-*, *bokr-*, *kurd-*, the overall meaning of the sentence is grasped due to the correct suffixes. We can deduce the following meaning from this sentence: a kind of female animal with a particular feature did something probably unpleasant, by applying force in some particular way to, most probably, a male animal and its baby (Apresyan 1966: 147). According to Scherba (in Uspenskyi 1962: 202), there are only several possible interpretations of the sentence which serves as a vivid example of the fact that language consists of formulas (schemas) that must satisfy certain conditions.[4] Because Russian is an inflective language, these certain conditions are explicitly realized by morphology. Admittedly, Sherba's ideas stood at the roots of the theory of grammar, pointing to the importance of morphology in meaning-making.

2.2 Morphology in early Western European grammatical traditions

Linguists of early Western European grammatical traditions (most notably, Otto Jespersen, Hermann Paul, Henry Sweet and Adolf Noreen) embraced the notion of morpheme. They were convinced that morphology as a study of the linguistic forms is opposed to a study of the language functions (which in different works was called 'syntax', 'semantics' and 'semiology') (see Bulygina 1977: 20, who makes reference to Sergievskyi 1940, and Gukhman 1968: 21). These scholars argued with slight insignificant variations in views that the object of the study of morphology should be a formal aspect of grammatical structure, whereas the second part of grammar, i.e. the study of meaning, should aim at the exploration of its content plane, namely at grammatical meaning (Bulygina 1977: 20-21).

According to Jespersen (1992[1924]: 31), a differentiation between morphology and syntax lies not in the distinction between the objects of study, but in whether a researcher chooses to study form or meaning. This point of view is justified by the fact that in natural languages there is no perfect harmony between the morphological and the syntactic ways of expressing the same fact (Jespersen 1992[1924]: 41). It is because in a language, there are "innumarable overlappings as if one district belonged at the same time to two or three different states" (Jespersen 1992[1924]: 41).

[4] Interestingly, because of this linguistic view, in the beginning of 20th century Scherba was accused of formalism (Uspenskyi 1962: 204).

In the first part of grammatical structure, i.e. morphology, a linguist should proceed from form to meaning. At this stage, all the linguistic units with similar formal expressions are examined. However, as Jespersen (1992[1924]: 40) states, it does not imply that their content should be ignored: "It should be a grammarian's task always to keep the two things in [their] mind, for signification, form and function, are inseparable in the life of language, and it has been the detriment of linguistic science that it has ignored one side while speaking of the other".

Furthermore, Jespersen does not make a distinction between derivation and inflection. He considers it reasonable to treat these two fields together,[5] because "on closer inspection, it will be seen that it is hard, not to say impossible, to tell exactly where the boundary has to be drawn beween flextion and word formation", as follows from the below examples in French—*paysan*/*paysanne*, *bon*/*bonne* (Jespersen 1992[1924]: 42)—where the morpheme *-e* that denotes feminine gender can be logically assigned to both an inflective and derivational category. Most interestingly, he considers the grammatical order of words in a sentence—what nowadays is peceived as a domain of syntax—as a morphological phenomenon.

Nowadays, some of Jespersen's and his contemporaries' beliefs on the grammar of language might seem naïve from the standpoint of modern linguistics, as, for instance, the described above division of grammar into two parts—morphology and syntax. However, they were the first who understood the importance of treating the language system as a whole and offered interesting solutions to a range of lingustic problems.

2.3 Morphology in Distributional Linguistics

A significant contribution to the development of morphology was made by American linguists in Distributional Linguistics. In fact, in the heights of structuralism in the USA (1930s to 1950s) morphology and phonology were the main objects of linguistic study. The American structuralists (e.g. Bloomfield, Trager, Harris, Hockett) worked out in detail a method of distributional analysis, the essential point of which is to obtain a compact description of the structural properties of language (with regard to its form). Distributional Linguistics is not a theory in the orthodox sense of the word. It is better understood as a scheme of procedures that allow the disclosure of grammar, or as experimental techniques for collecting and initial processing of the raw linguistic material (Apresyan 1966: 44).

[5] This idea echoes with the perspective on inflection and derivation in usage-based theories.

Although the morphological views of the American structuralists underwent changes during the developments of Distributional Linguistics, the main general principles on morphology may be concisely described as follows:

1) Morphemes are seen as a unit of the expression-plane of language "which bears no partial phonetic-semantic resemblence to any other form" (Bloomfield 1935: 161). Morphemes are "the smallest individually meaningful elements" (Hockett 1958: 123).
2) Each morpheme in a word is represented by one and only one allomorph; and each allomorph represents one and only one morpheme (Anderson 2015: 4). In other words, morphemes are viewed as distributional classes of its invariants (morphs and allomorphs) (Apresyan 1966: 46).
3) Morphemes are built from the units of the lower rank, e.g. phonemes (the underlying principle is that the units of the higher level are made of the units of the lower level). Thus, the morphological level is an intermediate level between the phonological and syntactic levels. The syntactic level of language (phrases and sentences) can be analyzed through a combinatorial morphological analysis (Apresyan 1966: 45). To put it differently, syntax was generally thought of as the concatenation of morphemes (Spencer 2005: 74).

From the perspective of the modern state of the art, it is possible to pinpoint the apparent shortcomings of Distributional Linguistics. Overall, the American structuralists overestimated the potential of distributional analysis. In other words, in this approach, language was perceived mainly through the prism of how linguistic units are distributed in a text. At times, the main units of the linguistic description were overfomalized (consider an example of words such as *conceive, deceive, receive* described as bimorphemic by Bloomfield, Harris and Nida (see Marchand 1960: 6). Moreover, the American structuralists did not sufficiently take into account the paradigmatic relations in the morphological level of the language system and did not discern the functional categories of morphemes. More importantly, the grammarians of Distributional Linguistics undervalued the significance of word-formational morphology in grammar (Kubryakova 1974: 9).

Nevertheless, despite some methodological limitations, Distributional Linguistics raised the critical questions about the notion of morpheme, thus, deepening a general understanding of morphology.

2.4 Morphological views in glossematics

Another famous structural approach is known as glossematics. Also called Copenhagen School of Structural Linguistics, it was founded by the Danish linguist Louis Hjelmslev in 1931. Hans Uldall, Viggo Brødal and Eli Fisher-Jørgensen made a further valuable contribution to the developing of this linguistic framework.

Although not acknowledged widely, the glossematic ideas had an immense impact on the general development of modern linguistics. The most influential thought was introduced by Hjelmslev in 'La Structure Morphologique' (1939): he proclaimed the inductive method futile and incomplete, saying "that a semantic value is never to be found by inductively collecting all particular significations" (Siertsema 1955: 46) and then concluding that "car l'induction reste par définition incomplète" (Hjelmslev 1939: 69, cited in Siertsema 1955: 46). Instead, methodological importance was given to the deductive empirical method, which allows "one to start from the general terms possible" − "la méthode déductive exige qu'on part des termes les plus généraux possibles" (Hjelmslev 1939: 87, cited in Siertsema 1955: 47). This methodological trend defined the field of linguistics until the end of the last century.

Given this methodological background, glossematics, unlike Distributional Linguistics, was created as a universal linguistic theory, not as a network of experimental techniques (Apresyan 1966: 54). The linguists of the Copenhagen School argued that the structural deductive approach to language rescinds the distinctions between the linguistic units—sentences, words, morphemes, syllables, and phonemes. In their view, all linguistic units share the same set of features that can be analyzed by only one procedure—therefore, when looked from this perspective, the need to discern lexicology from grammar, grammar from phonology, etc., simply disappears. Hjelmslev (1969[1961]: 59) explains the benefits from deploying this frame of reference: "Through the whole analysis, this method of procedure proves to result in great clarity and simplification […] From this point of view it will be easy to organize the subsidiary disciplines of linguistics according to a well-founded plan and to escape at last from the old, halting division of linguistics into phonetics, morphology, syntax, lexicography, and semantics—a division that is unsatisfactory and involves some overlapping". Alternatively, he proposes to obtain the sets of features of all linguistic units through the division of the language system into four dichotomous strata or layers: content ÷ expression, substance ÷ form (Murat 2006: 202).

Because of the universalistic linguistic nature of glossematics and the disregard for the traditional division of linguistics, little attention was paid to morpheme and morphology, if at all. In 'Prolegomena to a theory of language', Hjelmslev (1969[1961]: 26) understands the term 'morpheme' mainly as inflectional element. In another paper,[6] he defines the notion of 'morpheme' even more vaguely: "We use the notion of 'morpheme' in the sense which it has acquired in the European science" (Hjelmslev 2006[1961]: 177, my translation from Russian).[7]

To understand how Copenhagen structuralists treated the morpheme, we have also to keep in mind the fact that the philosophical basis of glossematics was logical positivism which denied the real existence of the objects in the material world believing them to be a bundle of the intersection of their relationships (Ivanova 2006: 141). Accordingly, the morpheme was defined through its syntagmatic and pragmatic relations and dependencies, like for example, the interconnection between inflectional morphemes and prepositions: "there is very often solidarity between morphemes of different categories within a "grammatical form", such that a morpheme of one category within such a grammatical form is necessarily accompanied by a morpheme of the other category and vice versa" (Hjelmslev 1969[1961]: 26) (e.g. as seen in Latin and Russian: the interdependency between *sine* and the ablative case and *на* (/nə/) and the prepositional case respectively). The various linguistic dependencies were interpreted in a mathematical sense as functions. In fact, a focus on the recognition of the linguistic dependencies meant a shift from the description of the linguistic units (pursued in the previous structural schools) to the revealing of their function and role in the language system.

In light of semiotics, Hjelmslev (1969[1961]: 44) views both morpheme and word as a sign: "Words can be analyzed into parts which, like words, are themselves bearers of meaning as words: roots, derivational elements, inflectional elements […] When, for example, the analysis of the English word like *in-act-iv-ate-s* is carried through this way, it can be shown to contain five distinguishable entities which each bear meaning and which are consequentely five signs". In this context, Hjelmslev acknowledges that there are different types of meanings. In other words, although morphemes and words are meaningful elements of language, their meanings are not the same.

[6] 'Dans quelle mesure les significations des mots peuvent-elles être considérées comme formant une structure?'
[7] "Мы используем слово «морфема» в том смысле, какой этот термин получил в европейской науке".

As mentioned before, Hjelmslev (2006[1961]: 176) believes that it is necessary to abandon the traditional opposition of syntax and morphology since the correlations (morphological relations) and dependencies (syntagmatic relations) determine each other. Hence, a core grammatical mechanism of language lies in a complex interaction between the morphological categories and the syntactic units. This idea leads Hjelmslev to the conclusion stated for the first time by Sapir that "morphemes are the main elements of a sentence because of the interdependencies between them" (Hjelmslev 2006[1961]: 176, my translation from Russian).[8] Since the dependencies within words are analogous to that within sentences, they can be analyzed in a similar way.

Noticeably, because the traditional division between the levels of grammar was dismissed, word fomation was seen through the prism of content vs. expression and substance vs. form. Consequently, Hjelmslev does not make a clear distinction between inflection and word formation. As in the case with inflectional morphemes, he interprets the derivational elements through the nature of the relationship and dependencies between them introducing in this regard the universal 'selection' principle. In general, it reflects the degree of dependency between linguistic units and as concerns word formation – between morphemes:

> "The structure of a language may be such that a word stem can appear both with and without derivational elements. Under this condition, there is then selection between the derivational element and the stem […] a derivational element necessarily presupposes a stem but not vice versa. The terms of conventional linguistics (morphology) are thus, in the last resort, inevitably based on selection, just like, for example, the term 'primary clause' and 'secondary clause'" (Hjelmslev (1969[1961]: 27).

Accordingly, the same selection principle applies to the syntactic relations between words, as, for instance, in any attributive phrase where existence of an adjective presupposes a noun and not the other way round.

All in all, the Copenhagen structuralist theory brought many new heuristic ideas to linguistic studies, the most important of which, in my view, is an understanding of language as a network of relations. However, despite its considerable theoretical contributions (largely forgotten

[8] "Следовательно, морфемы оказываются основными элементами, образующими предложение в силу реляций между ними".

nowadays), glossematics was hardly applicable for the practical description of the natural languages. For this reason, some of its extreme doctrinarian views were metaphorically called by Andre Martine ‘an ivory tower’ (in Apresyan 1966: 64).

2.5 Morphology in Prague Functional Linguistics

At the end of the 19th – the beginning of the 20th century, at the dawn of modernism and scientific revolution, a new philosophical concept called relativism emerged. Over the course of time, it evolved into a number of different schools, however, all of them shared a basic belief that a certain phenomenon (e.g. epistemic, aesthetic and ethical norms, experiences, judgments, and even the world) is somehow dependent on and co-varies with some underlying, independent variable (e.g. paradigms, cultures, conceptual schemes, belief systems, language) (Baghramian & Carter 2017). Noticeably, the prominent figures of the Linguistic Circle of Prague—V. Mathesius, B. Trnka, B. Havranek, V. Skalička, R. Jakobson, N. Trubetzkoy and S. Karcevsky—tried to apply the ideology of relativism to the study of grammar (Plungyan 2003: 230). In the context of language study, it meant the assessing of linguistic units not on the basis of how they correspond to entities in the real world, but on account of how they are correlated to other elements within the language system (Plungyan 2003: 229).

Another revolutionary thought of the Prague Linguistic School was that language must be investigated in its nature of ‘functional system’, whereby ‘functional’ refers to the communicating function of language (Luraghi 2005: 471). Roman Jacobson was the first in the history of linguistics who gave a voice to the idea that language serves for communication (in Waugh 2005: 549) and thus its internal structure has to be studied from the standpoint of the tasks it performs. In times of an ultimate domination of orthodox structuralist approach that viewed language as a closed system of pure realtions, he proclaimed a speaker and a hearer to be a central objects of grammar. Consequently, this meant that without the external (in relation to language) parameters, such as a situation of utterence of a specific text, it is impossible to determine even such ‘purely linguistic’ phenomenon as grammatical categories (Plungyan 2003: 237).

The Prague structuralists addressed a wide range of linguistic problems. However, the main focus of their works was phonology (Malmkjær 2010: xxviii). The interest in this topic was initiated by Trubetzkoy’s (2010[1939]) work ‘Grundzüge der Phonologie’ where, for the first time, the difference between phonetics (parole) and phonology (langue) was recognized. The

characterization of the phoneme itself as a 'bundle of distinctive features' also derived from Prague and was taken to America by Jacobson in 1942 and incorporated in publications with Morris Halle and others (Malmkjær 2010: xxxv).

As far as morphology is concerned, the Prague linguists believed that the basic unit of the morphological system is the morpheme; it is possible to represent every morpheme as a chain of the elementary morphological oppositions which were regarded as binary, such as, for instance, oppositions of the cases or oppositions of verb tense forms, etc. For example, Jacobson describes the system of cases in Russian not on the basis of a set of six features (or eight features) as it is accepted in traditional grammar, but by splitting it into three sets with two opposite features in each. Every set was named in accordance with the characteristics it bears: peripherality/non-peripherality (dative, instrumental, local vs. nominative, genitive, accusative), orientation/non-orientation (dative and accusative vs. nominative and instrumental) and dimensionality/non-dimensionality (genitive and local vs. nominative, dative, accusative and instrumental) (Apresyan 1966: 72). A 'repertoire' of the semantic features that serves as a basis for these oppositions was defined in a way quite similar to that in phonology (Apresyan 1966: 72).

In the pursuit of functional description of grammar, the Prague structuralists focused mainly on the description of inflectional morphology, whereas lexicology and word formation stood in the margin of the research activities of the generation of the founders of Prague Linguistic School (Dokulil 1994: 123). Only after the disintegration of the Prague Circle, Mathesius (1942, 1947) introduced a few ideas on the description of the lexicon concentrating his attention on the fundamental difference between 'descriptive' (i.e. derivationally motivated) and 'simple' (unmotivated, isolated) words (Dokulil 1994: 124) and naming units. These views laid the foundation of the onomasiological theory of word formation later developed by Dokulil (1962), Horecký (1983; 1989) and Štekauer (1998).

2.6 The 'father' of modern word formation theories

Proponents of different word-formation theories would probably agree on one fact: that Hans Marchand is one of greatest figures in the history of word formation. His contribution to the development of modern word-formation theories can hardly be overestimated. Marchand (1960) created one of the fullest descriptions of English word formation summarized in 'The Categories and Types of Present-Day English Word Formation' that has not lost its topicality nowadays.

Marchand published his studies in times when the subject of word formation was on the margins of linguistic interests. This fact is reflected in the preface of the first edition of his book: "[I]t somewhat surprising to see how very few there are that deal with word formation. This subject has been greatly neglected in grammatical works while the parts of phonetics, accidence and syntax have always received full attention" (Marchand 1960: i). Thus, Marchand's research has started a new chapter of word-formation studies in modern linguistics.

Marchand's word-formation theory is based on the Saussurean tradition. That is, he views both word and morpheme as a sign: "A word and, for that matter, any morpheme is a two-facet sign, which means that it must be based on the significate/significant (*signifié*/*signifiant*) relationship posited by Saussure" (Marchand 1960: 1). However, contrary to Saussure who considers only grammatical syntagmas to be motivated signs, Marchand also treats complex words as motivated signs defining them as "intellectually motivated by the significates" whose "certain form goes with a certain underlying concept" (Marchand 1960: 2). The object of a word-formation study, thus, should be morphological complex words formed by the combination of full signs by means of compounding, affixation, back-derivation and conversion, as well as complex words coined through the combination of incomplete signs by the way of blending, clipping, rime, etc. Hence, simplexes are excluded from the word-formation domain as unmotivated and unanalyzable signs. In addition, Marchand believes that word formation should only deal with synchronically productive patterns.

Despite the prevailing synchronic focus, Marchand takes into account historical aspects of word formation. However, this is done not with the purpose of describing the evolution of morphemes in the course of the history of the English language, but, rather, with an intention to justify a certain theoretical concept or classification. This is the main reason why Marchand (1960: 8) names his approach 'synchronic-diachronic'. In modern usage-based linguistics, such a position is highly justified and even desirable, because it allows us to capture the transitional nature of language, thus, avoiding artificiality in a linguistic description. This characteristic distinguishes Marchand's word-formation theory from structuralist approaches of the day.

Another cognitive feature of Marchand's framework, although not consistent, is that, at times, he acknowledges the importance of human perception in the process of coining new words, as can be conceived, for example, from the following quote: "The principle of combining two words arises from the natural human tendency to see a thing identical with another one already

existing and at the same time different from it" (Marchand 1960: 11). The idea of explaining a linguistic phenomenon with human cognitive abilities was brand new in the literature of that time where the systematic view on language was prevailing.

Furthermore, Marchand makes a distinction between a native vs borrowed basis of coining. In his view, borrowed lexemes may have a place in a study of word formation if they become analyzable, as in the case with such English words as *dis-agreeable* and *trans-alpine* (Marchand 1960: 4). More importantly, to deal with all of the complexity involved with borrowed morphological bases in English, Marchand (1960: 7) introduces different degrees of foreignness for morphemes. This also resonates with the concept of a graduation continuum of morphological patterns sustainable in modern usage-based approaches.

The synchronic-diachronic orientation of Marchand's approach also led to an interesting solution to the problem of how to treat foreign words with identical roots that entered English independently (usually pairs of words from French and Latin, e.g. *deceive: deception, resume: resumption*). Marchand (1951) proposes to differentiate two types of phonological alternations within morphemes: pure lexical (allomorphic) and morpho-phonological. In his view, the criterion for the distinction between these two types should not be the etymology of words, but whether the word is formed in accordance with productive native patterns or with the patterns of the source language. By way of illustration, the words *cultivatable* and *educatable* are formed within English by derivation in the nineteenth century. Therefore, the alternation in the pairs of words *cultivate: cultivatable* and *educate: educatable* (which is /eit/~/əbl/) should be considered as morpho-phonological. On the other hand, such words as *navigable* and *communicable* are non-native (Kastovsky 2005: 116). Hence, the phonological alternation in the pairs of words *navigate*: *navigable* and *communicate*: *communicable* is allomorphic. Although, this interpretation might appear somewhat artificial as, from the synchronic perspective, these words look morphologically identical, nevertheless, this is still a valid way to avoid morphological incongruities in the study of English word formation so that to make a linguistic description consistent.

In addition to the above-mentioned artificial categorization, at times, Marchand's account of the English word-formation system looks a bit descriptive from the perspective of modern linguistics. On the other hand, it does not provide the generalizations necessary for establishing a full theory. Thus, although Marchand himself does not evaluate his study as structuralist, its

objectives still remain congruent with other structuralist works which emphasize the importance of descriptive techniques and neglect a universal perspective on language.

In view of these considerations, 'The Categories and Types of Present-Day English Word Formation' by Marchand is still one of the most profound works on English word formation. Nowadays, the book has risen to the status of a classic that everyone who is interested in word formation should read.

2.7 Morphology in generative grammar

The emergence of dynamic models in linguistics is closely related to the beginning of the computer era. A problem of creating a machine that simulates human behavior has required a transition from the classification of linguistic observable units to the modelling of synthesis of sentences' sound forms in accordance with certain rules and from 'deep units', stored in the memory of a speaker. Analysis of the historical development of linguistics ideas gives grounds to assert that the study of language after Chomsky can be metaphorically called the era of modelling—he was among the first who shifted the focus of attention from describing separate linguistic phenomena to the modelling of language.

The core methodological framework of generative grammar was developed in 1960s and had an immense influence on linguistic studies in Europe and the USA. Chomsky and his followers ardently argued against the taxonomical classificatory approach in structural linguistics (Revzin 1977: 16) using *structuralism* as a swear word to denote everything they disliked about the earlier linguistic works. However, it is obvious that in the original sense of the term, Chomsky's theory is no less structuralist than other approaches (Trask & Stockwell 2007: 277). Indeed, virtually all serious work in linguistics in the twentieth century has been structural in outlook, though many contemporary linguists continue to regard structuralism as a term of abuse and would not apply the term to their own work (Trask & Stockwell 2007: 278). Nowadays, it is an accepted fact among many linguists—however, stated more in private conversations than in open articles—that in the second half of the 20th century the theoretical effectiveness of a method or a linguistic belief was much evaluated on the basis of *ad hominem* argument and not *ad rem*, as it fairly should be.

Generative grammar was inspired by the mathematical automata theory, particularly a Turing machine model. In his early work, Chomsky showed that 'generative grammars' (or

some of their types) can be regarded as a finite-state machine (Apresyan 1966: 81). It is a device that contains symbols[9] and a finite number of transformational grammar rules according to which phrases from the elements of that set of symbols are built. This grammar is able to construct an infinite set of right sentences of a certain language and ascribe some structural characteristics to them (Chomsky 1961: 8).

So far, it has not been established where a morphological component is allocated in the theory of 'transformational grammar'. According to Spencer (2005: 73), there are many different (sometimes even irreconcilable) morphological theories within the generative grammar tradition that range from the assumptions that only inflection should be regarded as a part of syntax (e.g. Kayne 1994) to the belief that the syntactic rules should be applied to both word-formation and inflection components of grammar (e.g. Halle & Marantz 1993). Further, considerable disagreement exists over the problem of constituent parts of derivation, namely, which elements should be ascribed to the lexicon and which to word-formation rules. Curious about which morphological generative theory Chomsky favours more, I sent him an email asking this question. The answer was short but informative: "Seems to me an open question. I don't have a strong opinion" (email received on 22/12/2017).

Although there are many theories of generative morphology and word formation nowadays, the situation was not that way when the generative grammar approach was born. At the beginning, it did not offer much explanation as to how morphology and word formation should be treated and the theoretical gap was evident. The first reason behind it was the primal universalistic nature of the theory—it sought to explain the phenomenon of language in terms of abstract overall transformational automata. A focus on syntax might be another reason for the interim avoidance of morphology after the emergence of the theory. Generativists' major interest was in units larger than the word: the structure of phrases and sentences (Bauer 1983: 3). Lastly, in the primary version of generative grammar, all morphological phenomena were regarded as 'surface' layers of language that do not provide much information about 'the deep structures'[10] of language. This is the main reason why many early works on word formation, including Panini's (who is an ancient

[9] As Chomsky (1961: 8) puts it: "A grammar is based on a certain vocabulary of symbols used for the representation of utterances and their parts, and including, in particular, the 'a priory' phonetic alphabet".
[10] The deep structures are the abstract representations of a linguistic unit, whereas the surface structures are the morphological/syntactic realizations of the linguistic unit as a string of phonemes/morphemes/words (Desagulier 2017: 1).

Sanskrit grammarian), were dismissed as 'taxonomic' (Lees 1960: xix, cited in Bauer 1983: 3). As a result, all morphological and word-formation phenomena were labeled as uninteresting—noticeably, this way of thinking is more typical for logicians than linguists (Bulygina 1977: 7).

The need to consider morphology and word formation was for the first time vividly expressed in a 1970 Chomsky paper 'Remarks on Nominalization' which, as many linguists believe, had an immense impact on generative word formation (Štekauer 2000: 96). In this paper, Chomsky discusses his views on the notion of word and introduces derivative regularities for nominalization called 'lexical redundancy rules' which differ in their nature from the transformational rules of sentence generation (Štekauer 2000: 95). To put it differently, he argues that because of the irregularity of derivative processes, they must be treated elsewhere in grammar—that is in the lexicon (Štekauer 2000: 96).

When analyzing the lexical redundancy rule, Chomsky proposes two feasible ways of considering derived nominals: either by extending base rules to allocate nominals directly within the lexicon or, alternatively, by simplifying the base structures of words—in this case, words are not listed in the lexicon and nominal derivational processes are the matter of transformational rules. Chomsky inclines towards the first possibility as he thinks that in terms of productivity, word-formation processes are limited, and in terms of semantics, they are irregular (Štekauer 2000: 96). By this token, Chomsky recognizes derivational morphology as an autonomous subsystem of language independent from syntax (Carrier 1979: 415). This idea gave birth to the 'lexicalist' approach where the rules regulating the structure of words and describing the nature of relations between morphemes were considered a constituent part of the lexical component of a language model. The second alternative option offered by Chomsky for considering derivational rules within grammar developed into the 'transformationalist' approach.

Another big step forward in the development of generative morphology was Halle's (1973) famous paper 'Prolegomena to a theory of word formation'. In this short programmatic statement, he points to a vacuum in generative linguistic theory where morphology should be (Spenser & Zwicky 1998: 1). Since then, generativists have come up with a number of morphological and word-formation theories the common features of which are sketched in the following paragraph.

In most generative grammar approaches, morphology is not considered an autonomous component of the grammar; it is split between *morphophonology*, as part of phonology, and *morphosyntax*, as part of syntax (Frawley 2003: para.1). It is conceived as a subpart of the bigger

lexical component of grammar. Consequently, a word-formation rule is understood as a chain of transitions that underlies the transformation of the initial syntax construction into a corresponding derivate. To put it simply, words are treated as a species of phrase or clause structure. Basically, this means that a word formed by either inflection or derivation can be represented as a syntactic node and hence subtend syntactic relations with structures which surface as words or parts of words (Spencer 2005: 73).

Having these methodological guidelines as such that roughly define the terrain of morphology, generativists aim to solve three major problems in word formation. First, they try to find out the extent to which syntactic rules can have access to the internal structure of words. The second problem is related to the scale of the incorporation of syntactic constructions into words. Lastly, the syntactic behaviour of the newly formed words, particularly those realized in argument structures, defines the domain of the third word-formation problem in generative grammar (Spencer 2005: 73). To trace and categorize the derivation of words with the involvement of syntax, many different techniques have been developed by generativists (e.g. Borer 1998, 2003; Lieber 1992, 2004; Selkirk 1982; Marantz 2013; and Harley 2005) that enhance our understanding of relations between grammar and lexicon. These theories argue that there are no principled differences between words and phrases (Don 2014: 101).

In what follows, I will briefly discuss the ideas of the main word-formation theories within generative grammar that had an immense impact not only on the development of this linguistic paradigm, but on the whole linguistic field.

2.7.1 Level-Ordered Morphology and Lexical Phonology

Probably, the first systematic attempt to marry phonology and morphology was made by Siegel (1974) within the framework of Level-Ordered Hypothesis. Later, it was further developed in works of Selkirk (1982) and Kiparsky (1982, 1985) under the title of Lexical Phonology. In general, these approaches explore different phonological mechanisms involved in linear ordering of affixes in English.

Siegel (1974: 12) views morphology as the study of word-formation processes that take place in two distinct components of language: inflectional and derivational. Consequently, inflectional morphology treats the generation of words by the syntactic component of the grammar, whereas derivational morphology is regarded as the study of word-formation processes which occur in the lexicon. In her opinion, each of these morphological processes is governed by

constraints which are characteristic of the components in which they arise. Based on this division, words are either syntactically or lexically derived.

In the models of Level-Ordered Morphology and Lexical Phonology, word-formation processes are ordered in two or more levels which determine the domain of application of the cyclic phonological rules of that level (Don 2014: 43). Different innate properties of affixes define a level to which they belong. For example, such suffixes as *-al*, *-ous*, *-ity* (also termed 'Latinate') attach to both words and stems (Siegel 1974: 103) allowing prosodic change, e.g. a stress change in a word (Siegel 1974: 112). Therefore, they are allocated at level 1 (or 'class I' in the terminology of Siegel). In contrast, such suffixes as *-hood*, *-ness*, *-er*, *-ism*, *-is* are 'stress-neutral' (Siegel 197: 112) and bind only to words. This feature determines their position at level 2 (or 'class II').

Paradoxically, the ordered arrangement of affixes is both a strong and weak point of these two theories. On the one hand, the robustness of these models lies in the fact that they allow us to make trustworthy predictions of the familiar ordering of affixes (Don 2014: 44). In addition, the theories partially describe productivity: irregular and unproductive rules are allocated at level 1, whilst more regular and transparent processes are ordered at higher levels, explaining, thus, why irregular unproductive process come prior to the general productive ones (Don 2014: 45). On the other hand, although Level-Ordered Morphology and Lexical Phonology give a detailed and multifaceted account of ordered interrelations between morphology and phonology, they do not explain the reason why affixes are organized in a particular way. What are the driving forces behind the specific ordering of affixes? One of the plausible answers to this question comes from outside the generative grammar tradition. A recently developed theory by Hay (2000, 2001 and 2002) interprets morphological complexity and ordering as a psycholinguistic phenomenon which largely relies on the parsability of affixes (Plag 2003: 175).

Another arrow of criticism is aimed at the theories' basic tenet. Because language is an immensely sophisticated system, its morphological heterogeneity cannot be described by affix restrictions alone. As shown by Fabb (1988) in his critical discussions of Level-Ordered Morphology and Lexical Phonology, these theories are not sufficient to explain affix ordering in its entirety (Spencer & Zwicky 1998: 2). However, the alternative model of selectional restrictions proposed by Fabb (1988) has also been shown to have weaknesses for the reason of assigning "all kinds of restrictions to stipulated idiosyncrasies of the suffixes" (Plag 1996: 770).

Despite the discussed limitations of these approaches, they provide valuable insights as to how derivational and inflectional processes are organized in English. Level-Ordered Morphology and Lexical Phonology are also supported by considerable empirical evidence that makes them worthy of attention nowadays.

2.7.2 Aronoff's Word-Formation Theory

One of the most recognized theories of word formation in the generativist tradition has been developed by Aronoff (1976). The biggest contribution of his theory is the establishing of morphology as an independent level of the linguistic system that has led to its recognition as a fully-fledged field of study within the generativist framework. In addition, Aronoff's monograph is also prominent for outlining morphological problems that should be addressed in a theory of word formation.

Aronoff fully accepts the traditional division of morphological phenomena into inflection and derivation, the latter being restricted "to the domain of lexical category" (Aronoff 1976: 2). Derivational morphology is seen as organized syntagmatically.[11]

To adjust his theory with the fundamental principles of generative grammar, Aronoff adopts a word-based morphology: he believes that "all regular word-formation processes are word-based" and that a "new word is formed by applying a regular rule to a single existing word" (Aronoff 1976: 21). This is because morphemes cannot always be considered a Saussurean sign, i.e. a union of form and meaning (Aronoff 1976: 16). Instead, within Aronoff's Word-Formation Theory, a morpheme is defined as a phonetic string which can be connected to a linguistic entity outside that string and which does not always have meaning (Aronoff 1976: 15), that is, in terms of its formalistic and syntactic properties. This definition is supported by the morphological components *-fer* and *-mit* in the Latinate verbs *refer*, *defer*, *prefer*, *infer*, *confer*, *remit* and *commit*, *transmit*, *submit*, *admit*. In Aronoff's view, these components are not just mere accidental phonological sequences. Rather, they "must be stated on another linguistic level, the level of the stem or morpheme" (Aronoff 1976: 13). In other words, Aronoff (1976) explains the incongruity and ambiguity that morphemes sometimes display by modifying the traditional definition of 'morpheme'.

[11] Aronoff does not use the term 'syntagmatic' himself, but refers to it strictly as 'non-paradigmatic' throughout.

A morpheme's divorce from meaning, then, has an impact on the whole theory. In particular, word-formation rules are seen as not operating on morphemes, but only on a set of words. When a rule is applied, which is believed to be 'offline' and 'once-only', it designates a phonological operation that is performed on the word, as well as a syntactic label, a subcategorization and a semantic reading of the resulting word. The rules of word formation are perceived as mechanisms of a language's dictionary and are assumed to be separate from the syntactic and phonological rules of the grammar (Aronoff 1976: 22).

Furthermore, to account for the exceptions and irregularities observed in a real language, Aronoff introduces different types of restrictions for word-formation rules: syntactic, semantic, phonological, and morphological. One of these restrictions, for example, is 'morphological blocking' which refers to the fact that the output of a more idiosyncratic, less productive word-formation rule prevents application of a more general and productive rule (Kenstowicz 1994: 210). This concept is illustrated by the word *glory* in English which blocks the derivation of **gloriousity* from *glorious* (Manova 2015: 960).

In addition, in the Unitary-Base Hypothesis, Aronoff discusses the limitation of a morphological rule to 'the syntacticosemantic specification of the base' (Aronoff 1976: 48). This hypothesis assumes that an affix cannot be attached to any lexical category. Instead, it selects words of exclusively one category (Scalise & Guevara 2005: 162). However, many affixes can attach to more than one base, such as the English suffix *-ize* which attaches to both adjectives (*legalize*) and nouns (*unionize*) to form verbs (Lieber 2009: 179), or the high-valent Persian suffix *-i* which combines with nine lexical bases to form nouns: nouns (*بارانی*), adjectives (*بلندی*), adverbs (*شبانه روزی*), present tense verb bases (*هستی*), past tense verb bases (*شگفتی*), the infinitive form of verbs (*نوشیدنی*), pronouns (*منی*), numerals (*بیستی*), and question words (*چندی*) (Krykoniuk 2014: 26).

Another type of restriction concerns affix ordering. In Aronoff and Manova (2010), various aspects of affix order are considered. They distinguish eight types of approaches to deal with affix constraints, the basis of which is specific information deployed by a word-formation rule (phonological, statistical, psycholinguistic, templatic, etc.). For example, in English, the Germanic derivational morphemes (e.g. *-ness*) are considered to be closing suffixes (Manova 2015: 958).

Moreover, in his theory, Aronoff superficially touches on productivity. Notably, the derivational phenomenon of productivity is not seen as an isolated attribute of a word-formation

rule, but as tightly connected to another property: 'semantic coherence'. Semantic coherence concerns the predictability of any complex word from the interaction of the lexicon and the set of word-formation rules, the outcome of which is that the predictable word does not need to be listed in the lexicon (Scalise & Guevara 2005: 159). To calculate an index of productivity, Aronoff (1976: 45, emphasis in original) proposes "to count the number of words which we feel *could* occur as the output of a given WFR [word-formation rule] […], count up the number of actually occurring words formed by that rule, take a ratio of the two, and compare the same ratio for another WFR." However, the problem with this proposed method to estimate derivational productivity is that we can hardly know for sure how many potential bases there are for a given lexeme formation process (Lieber 2009: 67). Hence, this method of calculating productivity is impractical.

Despite the undeniable contribution of Aronoff's theory to linguistics, it has weaknesses that can be extended to the whole methodology of generative grammar. First, the word-formation system is analyzed purely in the synchronic dimension of the language. Failing to account for diachronic word-formation processes, as well as for the probabilistic and transitional nature of the language, makes Aronoff's description artificial at times. Secondly, the theory treats language as some abstract entity (as if it existed independently in a separate dimension), ignoring the fact that there cannot be a natural language outside the human mind.

With all things considered, Aronoff's theory was a big step forward in the development of generative grammar and in the understanding of English word-formation processes as a whole unified system.

2.7.3 The concept of 'argument structure' in generative grammar

In some theories within generative grammar, 'argument structure' constitutes the basic concept. 'Argument' denotes a syntactic element required by a specific verb, and 'argument structure' is a hierarchically represented range of arguments licensed by a specific lexical unit (Matthews 2014). For example, in *I give you my word*, the verb *give* takes three arguments: *I*, *you* and *my word* (i.e. a subject, and indirect and direct objects respectively). Therefore, the key idea behind argument structure is that verbs assign a thematic role to their arguments. The list of thematic roles often includes agent, causer, patient, theme, experiencer, source, goal, location, beneficiary, instrument, and comitative, among others (Mateu 2014: 24).

For the most part, theories based on argument structure consider how some word-formation processes manipulate different argument structures (Don 2014: 100). Although there is a

considerable body of literature on this topic, and the methods are developed in detail, I will outline only the main ideas of these approaches due to the limited scope of this chapter.

Simplifying significantly the state-of-the-art, two main approaches to the study of argument structure can be distinguished: endo-skeletal ('projectionist') and exo-skeletal ('constructionist'). In the first approach propagated by Williams (1981), di Sciullo and Williams (1987), Levin and Rappaport-Hovav (1995), Grimshaw (1990), Reinhart (1996) and others, the syntactic domain of a lexeme is defined by the semantic features of that lexeme (Don 2014: 101). To put it differently, it focuses on the lexeme's properties "as the skeleton around which the structure is built" (Borer 2003: 33). When a particular verb comes with an object, it is because its lexically specified argument structure says so (Don 2014: 102). According to Borer (2003: 32), such approaches are less restricted and more redundant with the empirical advantages to account for unpredictable syntactic properties and the wealth of other linguistic phenomenon.

On the other hand, the second approach—associated with works of Borer (1994, 2003), Harley (1995), Kratzer (1994) and Marantz (1996, 1997)—is based on the opposite assumption, namely, that it is a syntactic structure that determines both grammatical properties of lexemes and their ultimate fine-grained meaning (Borer 2003: 34). In this view, a verb comes with an object because the syntactic structure of the particular sentence is such that the object forms a constituent with the verb at some point (Don 2014: 102). As a result, exo-lexical approaches consider the lexicon to be highly impoverished, containing "little beyond the sound-meaning pair" (Borer 2003: 34). The biggest advantage of this method is that, from the computational perspective, it is less costly.

To conclude, in addition to considerable theoretical importance, generative grammar theories of argument structure have a number of practical implications. Findings from these theories help us better understand the nature of the lexicon and the relations between sentences and derived words.

2.7.4 Lieber's Lexical Semantics

The account of the development of word-formation studies in generative grammar would be incomplete, if we did not mention the work of Rochelle Lieber. In my view, it is one of the most complete theories in linguistics that treats morphology from the semantic perspective. Generally, she explores the meaning of morphemes and how they combine to form the meaning of complex words.

Morphological description of Lieber's theory is rooted in the works of Jackendoff (1983; 1996), and Levin and Rappaport-Hovav (1988; 1995). However, the approach diverges in that it emphasizes the importance of cross-categorial description. Moreover, in contrast to Jackendoff whose interest is mainly focused around verbs, the inventory of Lexical Semantics is adjusted for the meaning exploration of various parts of speech formed by such word-formation processes as affixation, compounding and conversion (Lieber 2004: 6). On the other hand, the approaches of Wierzbicka (1985), Pesteyovsky (1995) and Szymanek (1988) provide the semantic foundations of Lieber's theory.

The central idea to the theory of Lexical Semantics is that, in word formation, the meaning of a complex lexeme is composed of two 'semantic skeletons' whose relationship to each other is either of the juxtaposition type or subordination (Lieber 2004: 10). The primary mechanism for the formation of a complex lexeme is called 'co-indexation' that allows for the integration of the referential properties of morphemes. Co-indexation is seen as a device "to tie together the arguments that come with different parts of a complex word to yield only those arguments that are syntactically alive" (Lieber 2004: 45). Each word-formation process has a specific principle of co-indexation, according to which this mechanism applies. More importantly, co-indexation provides a good explanation for the polysemy of an affix (e.g. *-er, -ee*).

Another important notion in Lexical Semantics is 'paradigmatic extension' (Lieber 2004: 72). The theory suggests that a semantic space of the simplex lexicon, including affixes, is organized paradigmatically. When, in a given language, there is a need for a new coinage and none of the existing suffixes meets this need, 'paradigmatic pressure' on the simplex lexicon occurs (ibid.). As a result, the semantically closest productive affix is put to use, even if it requires a violation of the principles of co-indexation (Lieber 2004: 74). The application of a new meaning to an affix, therefore, is termed 'paradigmatic extension'.

The theoretical apparatus of Lexical Semantics is extremely compact and consists of six lexical semantic features—[material], [dynamic], [IEPS] (i.e. 'Inferable Eventual Position or State'), [Loc], [B] (which stands for 'Bounded') and [CI] (which is an abbreviation for 'Composed of Individuals'). Nevertheless, with these primary features, it is possible to distinguish broad semantic classes of different parts of speech formed by derivation, compounding and conversion (Lieber 2004: 12). Furthermore, Lieber (2004) argues that this apparatus allows us to predict semantic content of some derivational affixes that ought to exist in English.

The first two lexical features represent major ontological classes: [material] stands for the conceptual category of substances/things/essences, whereas [dynamic] for the category of situations. The third feature originally developed in Lieber and Baayen (1999)—[IEPS]—captures the major aspectual classes of verbs. The addition of this feature to a skeleton signals the inclusion of the whole semantic component, which is the sequence of Places or States (Lieber 2004: 29). The feature of [Loc] asserts the relevance of place or position in the semantic content of a complex lexeme (Lieber 2004: 99). It mainly explains subclasses of simplex (static) verbs and of adpositions. Finally, quantity—by which 'duration, internal individuation and boundaries' (Lieber 2004: 134) are understood—finds a realization in the features of [B] and [CI]. The former expresses intrinsic spatial or temporal boundaries in a situation or substance, whilst [CI] embodies the relevance of spatial and temporal units implied in the context of a lexeme (Lieber 2004: 136). Additionally, in Lieber's model, the presence or absence of a feature in the body of a semantic representation is marked by the '+' and '−' symbols respectively.

This framework of semantic representation is based on a one-to-one correspondence with morphemes. This is the reason why it can deal with all sorts of exceptions and semantic-morphological incongruities efficiently.

In summary, Lieber has developed an inclusive and consistent theory that accounts for a wide range of word-formation processes and that unifies the expression and content planes of language. However, as Park (2017: 805) points out, the theory would have benefited greatly from incorporating to the theory a cognitive aspect of word formation.

2.8 Construction Morphology

If a linguist were to summarize Construction Grammar in one sentence, they would probably say that it is a study of speakers' knowledge of language, consisting of different types of 'constructions', i.e. form-function pairings (Goldberg & Suttle 2010: 468). Construction Grammar began as an attempt to account for such constructional idioms as *the X-er the Y-er* (e.g. *the more the merrier*) which have idiosyncratic properties not predictable from general rules or principles of language (Dąbrowska 2015). Nowadays, however, the concept of constructions is extended to the whole domain of language. Thus, grammar and lexicon are seen as an inseparable continuum of constructions (Bybee 2007a: 280) which properties—phonological, syntactic, semantic, etc.—can be uniformly represented as features with values (Croft 2007: 479).

With the rise of Construction Grammar, Booij (2010), whose early works can be classified as such that have emerged from the generativist tradition, has developed Construction Morphology. The idea of constructions in this approach looks like a compromise between the generative grammar's notion of rules and the concept of 'constructional schemas' (Langacker 2007: 441) established in cognitive linguistics. One of the direct indications for this statement comes from the basic principle of Construction Morphology—that constructions are generalizations over a set of words (Booij 2010: 2). It vividly echoes with Jackendoff's (1975) view that word-formation rules are redundancy rules[12] over a list of words (Don 2014: 61). In addition, formal representations of constructions—called constructional schemas—are borrowed from the Parallel Architecture framework, developed by Lerdahl and Jackendoff (1983). Constructional schemas are assigned to a word on the basis of its systemic paradigmatic relations with other words (Booij 2015: 2).

Construction grammarians (Hilpert 2014: 75) bring three main arguments in favour of derivational constructions (which behave similarly to syntactic constructions): first, they are selective to their input elements; secondly, constructions exhibit 'coercion effects', i.e. when a context influences elements of constructions and changes their meaning (Hilpert 2014: 17). In Hilpert's opinion, these two features are difficult to explain by the machinery of rules. Thirdly, as compared to rules, constructions are neutral to production or perception (Booij 2015: 3).

Therefore, derivational constructions as meaning-form parings are seen as the main instrument for word formation (Hilpert 2014: 75). Together with complex words, they constitute a hierarchical lexicon (Booij 2015: 1), in which abstract schemas dominate their instantiations (Booij 2015: 3). According to Booij's model, the lexicon bestows base words (roots) with meaning, whilst the meaning of affixes is specified by constructional schemas (Booij 2015: 2).

The constructional schemas have internal morphological and external—'holistic'—properties. Internal properties of the constructions are defined by their morphological elements, as can be seen in affixation. On the other hand, holistic properties are determined by whole constructions, as evidenced by reduplication (Booij 2015: 7). In this context, Booij (2015) suggests

[12] In generative grammar, redundancy rules are believed to have a function in a grammar that stores whole forms. As has been shown by Bochner (1993), the informational load of a word in the lexicon which is formed fully in accordance with a redundancy rule is close to zero (Don 2014: 10).

a remarkable interpretation of VN exocentric compounds[13]: their idiosyncratic semantic features are explained by the morphological configuration of a construction as a whole.

Moreover, Construction Morphology construes a paradigmatic relation between syntax and morphology by finding similarities between derivational and syntactic constructions. It means that a noun phrase (with no morphological expression) also become an object of word-formation morphology. This principle might work for English which has relatively poor inflectional morphology. However, it is not clear how this principle would be implemented for the study of highly inflective languages where syntax is radically different from word formation. My intuition[14] as a native speaker of Ukrainian and Russian tells me that, because of the free order of words in these languages (compensated by a high number of inflectional forms), there would be considerably fewer conventionalized noun phrases that fall into the category of a word-formation construction in inflective languages than in analytic languages (e.g. English). This is not to say, however, that there are no such phrases in inflective languages, but the question is whether these instances are sufficient to make a decisive theoretical generalization. In my view, further empirical research on a greater number of languages should be done to substantiate/explain this assumption.

Construction Morphology also attempts to address the problem of morphological productivity—the notion which, in modern linguistics, has perhaps the highest number of interpretations (as shown in Bauer 2001). Some productive constructions are explained with a new type of productivity—'embedded'—which occurs when a productive word-formation process is boosted by a syntactic construction (Booij & Audring 2017: 293). Embedded productivity supports the idea that all constructions constitute a continuum, which makes it impossible to draw a sharp line between lexicon and grammar.

To conclude, as evidenced above, Construction Morphology integrates three perspectives on morphology— semantic, morphological and syntactic—offering well-grounded explanations to many word-formation phenomena and establishing relations between morphology, semantics and syntax.

[13] Exocentric compounds are compounds with an absent semantic head.
[14] To the best of my knowledge, there are no linguistic studies on this subject.

2.9 The onomasiological theory of word formation

The definition of morphemes *per se* envisages the fusion of meaning and form. Although most morphological approaches discussed so far regard this statement as fundamental, not all of them, in my view, treat semantics as thoroughly as the onomasiological approach. This is because it studies language from meaning to form.

In modern cognitive lexicalist approaches, the study of the content plane of language evolves in three key directions: analyzing the semantic properties of words with regards to their morphological structure, as well as the class to which they belong; investigating the semantic relations between morphological constituents of words; and elucidating the problem of a choice, namely why a particular morpheme is selected for the formation of a new word. A heuristic undertaking to unify these three lines of research was made by the founders of the onomasiological theory, Dokulil (1962), Horecký (1983), Štekauer (1998). The theory revolves around the idea that word formation starts on the semantic level with the combination of 'semes'—the instantiations of the respective logical predicates of the pre-linguistic level. Then, the Morpheme-to-Seme-Assignment Principle matches semes to the potential morphemes in the lexicon and checks for the semantic and formal compatibility and restrictions (Štekauer 2005: 216).

According to Štekauer (2005), there are five onomasiological patterns in word formation, which are termed 'Onomasiological Types'. The basic assumption behind this distinction is that every word potentially consists of two components: an onomasiological 'base' and 'mark'. The base specifies a class to which a word belongs (usually, the base is realized by an affix), whereas the mark contains specific semantic information about a named object/action/phenomenon. Moreover, the mark may be further divided into a 'determining' and 'determined' constituent. Štekauer (2005: 215-16) believes that the latter is always represented by the category of Action in one of its three modifications: Action, Process or State. The onomasiological types have the corresponding semantic patterns that reflect the semantic categories of the morphemes. Semantic categories are understood as conceptual categories that constitute a more general cognitive schema of a particular situation (Haselow 2011: 56).

On a larger scale, the major undertaking of the onomasiological theory is to marry the structuralist tradition of the Prague Linguistic School, generative grammar and modern cognitive linguistics, which makes it a leading candidate among modern approaches to word formation.

2.10 Usage-based approach to morphology

One of the modern approaches to language within the framework of cognitive linguistic is Usage-based Theory (Bybee 2006), which acknowledges the impact of experience with language on its cognitive representation. Therefore, this theory is centered around the study of frequency effects of various sorts, since these frequencies are believed to determine the nature of grammatical organization of language. Within this approach, linguistic structure is perceived as "emergent—governed by certain regular processes, but always changing as it is re-created in the individual and in specific usage situations" (Bybee 2013: 50). For this reason, the definitions of linguistic units (morphemes among them) have no strict and rigid definition in this theory: their manifestation is gradational. Further, instead of bringing into light smaller linguistic units, the theory places a major focus on constructions, which are understood as "processing units or chunks—sequences of words (or morphemes) that have been used often enough to be accessed together" (Bybee 2013: 51).

Different types of frequencies have different effects on language. Hence, the distinction between 'type' and 'token' frequencies has an important place in Usage-based Theory. Token frequency "counts the number of times a unit appears in a running text" (Bybee 2007a: 9), whereas type frequency "refers to the number of distinct items that can occur in the open slot of a construction or the number of items that exemplify a pattern (Bybee 2007a: 14).

Consequences of high type and token frequencies of different units within constructions occur due three major effects of frequencies: the Conserving and Reducing Effects (associated with token frequency) and Autonomy (associated with type frequency). The Conserving Effect stems from the established cognitive fact that repetition strengthens memory representation, suggesting that, within a certain paradigm, high frequency tokens resist changes and serve as a base upon which new forms are created (Bybee 2007a: 10). The Reducing Effect, then, explains sound change in language: high-frequency combinations of token words are prone to phonetic reductions (Bybee 2007a: 13). Lastly, Autonomy is defined as "the extent to which a word is likely to be represented in the speaker's lexicon as a whole and separate unit (Bybee 2007a: 50). The effect of Autonomy has been observed in derivational morphology (Hay 2001; Hay & Baayen 2002) in the phenomenon of the morphological parsability of words: it has been shown that the meaning of the derived words distances from that of their bases if the derived words are more frequent than their bases.

Another important idea offered by Bybee (1985) concerns the problem of distinction between inflectional and derivational morphology, which has implications for a more general discussion of the nature of the connection between grammar and lexicon. Unlike many preceding linguistic theories, Usage-based Theory does not contrast inflection and word formation, considering them different points on the gradational continuum of the expressions of morphological features of language. Although this approach is very appealing in detaching from a rigid and mechanistic view of linguistic categories, it might face some problems in the typological analysis of languages, where it is not always justified to view grammar entirely from a morphological perspective. Nevertheless, with its strong focus on psychological and cognitive phenomena, the Usage-based Theory provides 'powerful explanatory possibilities' (Bybee 2013: 69) to many linguistic facts and is, thus, made a central explanatory framework to this thesis.

2.11 The cognitive stat-rule approach

There are influential studies in morphology and word formation in modern linguistics, which are not distinctly unified under the umbrella of a particular theory. However, they share a lot of similarities: they engage with the literature, morphological problems and terminology related to the rule-based ground of generative grammar; they attempt to provide a cognitive explanation to the observed phenomena; and they draw on the data from corpora and psycholinguistic experiments and use statistical methods to analyze it and to support their claims. For this reason, I have termed this trend in linguistics 'the cognitive stat-rule approach' to word formation. The most prominent figures of this approach are Ingo Plag, Harold Baayen and Jennifer Hay. Their new insights into many morphological problems have revolutionized the field.

Ingo Plag has made two significant contributions to unlocking the mystery of suffix ordering and morphological constraints. Firstly, in contrast to Fabb's (1988) model, Plag (1996) proposes a comprehensive and empirically-driven account of the combinatorial properties of derivational suffixes, which arise as a result of base-driven selectional restrictions, paradigmatic morphological processes, and independent principles and constraints of English derivational morphology. Further, through an investigation of 15 English suffixes and their potential 210 two-suffix combinations, Hay and Plag (2004) show that the selectional restrictions of suffixes and their parsing restrictions coincide most of the time, and they propose a model of suffix ordering along a hierarchy of processing complexity. In another pioneering study, Plag and Baayen (2009)

provide evidence for a correlation between a higher rank of suffixes and their increased productivity in complexity-based ranking. Secondly, Plag (1996) has deepened our understanding of the nature of constraints on morphological processes. With a focus on verb derivation and employing both a dictionary-based and a text-corpus-based account, he depicts a complex, multifaceted and interactive picture of English morphology (Cetnarowska 2001).

Harold Baayen has contributed to many linguistic fields, in particular by introducing various statistical models that provide an account of different aspects of language. In the field of morphology, his greatest impact perhaps lies in introducing different measures of productivity, which allow for the quantification of this morphological process (Baayen & Lieber 1991; Baayen 1993). These findings have been an immense step forward, as they have revealed new surprising aspects of the behaviour of affixes (for example, the fact that affixes with a relatively high type and token frequency have a low potential productivity). Further, Hay and Baayen (2002) have demonstrated that there is a positive correlation between the productivity of an affix and its parsability in perception. Moreover, in the area of morphological processing of words, Moscoco del Prado Martin, Kosctić and Baayen (2003) have brought to light an impact of the information residual of a word on the processing of inflectional and derivational paradigms. Lastly—narrowing down an impressive list of Baayen's contributions to morphology due to the limited scope of this study—he has also summarized unusual statistical properties of word frequencies and their distributions in large corpora (Baayen 2001).

Hay's (2000) PhD thesis, with a major focus on the effects of speech perception strategies upon morphological structure of words, has also influenced the field of morphology. She has established that the likelihood of a word's parsability, when accessed in memory, determines many aspects of its long-term representation, such as, for example, degree of semantic transparency, polysemy, phonetic detail and suffix ordering (Hay n.d.). Further, Hay (2001) has presented evidence for the frequency effect of words that derived words semantically split from their bases, if their relative frequency is higher than that of bases. She has also proposed the 'segmentability hypothesis' (Hay 2003) that predicts less phonetic integration and, consequently, phonetic reduction, in newly derived words, as evidenced by a negative correlation between morphological segmentability and phonological integration.

In sum, the works of these three researchers have transformed the modern field of morphology and have defined new horizons for its development.

2.12 Conclusion

This chapter discusses the development of different morphological and, in particular, word-formation theories in the history of linguistics—from the coinage of the term ‘morpheme’ in the nineteenth century to the present time. In the beginning, the main morphological concepts from the early Western European grammatical traditions are explored; then, I consider morphology in Distributional Linguistics of American structuralists, in the theory of glossematics developed by Danish linguists and in Prague Functional Linguistics founded by Russian and Czech scholars. In a separate section, Marchand’s synchronic-diachronic approach to word formation is analyzed that stimulated the interest of other linguists in word formation. Then, the distinct theories of generative grammar—Level-Ordered Morphology, Lexical Phonology, Aronoff’s word-formation theory, Lexical Semantics and theories based on argument structure—are considered in some detail. The next two sections briefly describe the tenets of theories construed on the concept of Construction Morphology and of the onomasiological approach. This literature review is finished by the discussion of the main contributions of Usage-based Theory and the studies of the cognitive stat-rule approach.

3 The procedure of the formal morphological analysis

The central method of the current study is a formal morphological analysis. It has been chosen as the core methodology, because it provides tools for the precise categorization of morphemes, as well as for their short and informative annotation, which is particularly useful in statistical analyses and in the assessment of the English word-formation system as a whole—thus, assisting me in addressing the main goals of this thesis. This chapter discusses theoretical premises of the formal morphological analysis. In what follows, Section 3.1 introduces the method of formal morphological analysis, and Section 3.2 outlines the scope and the sample of this research. Section 3.3, then, elucidates the meta-apparatus and the basic methodological tenets of the formal morphological analysis. Section 3.4 focuses on the problem of morphological parsing of words and presents the adopted criteria for resolving some morphological ambiguities that have arisen during the parsing of words in the sample. In Section 3.5, the importance of etymology for morphological parsing is discussed. Section 3.6 looks at the parameters, which are analyzed in greater detail in Chapter 7 of this study. The chapter ends with Section 3.7 which summarizes the discussion.

3.1 The formal morphological analysis: the prolegomenon

The remarkable profile of the formal morphological analysis in offering accurate, concise and consistent solutions to a morphological description of languages—although not known in Western linguistics—has motivated my choice of this method. The formal morphological analysis was developed by Russian and Ukrainian linguists (e.g. Bratchikov et al. 1958; Tyshchenko 1969, 2003) for the needs of machine translation. It implies distinguishing the elements of a lexeme and assigning them to a particular morphological category (Tyshchenko 1969: 25). This definition is encapsulated in the following formula:

R” = mR’m1mR’m2...

Here, **R** stands for a radical or a morphologically indivisible word annotated by its word class (N, Ad, Aj, etc.); and **m** stands for grammatical or lexical-grammatical morphemes (affixes) denoted unchanged by means of the Latin alphabet. The prime (’), double prime (”), triple prime (”’) symbols indicate the number of morphemes in a formula: two morphemes are encoded by one prime symbol, three morphemes by two, etc. For example, *brighten* is parsed as V’=Aj+en,

breadwinner as N"=N+N', and *affectionately* as Ad"'=Aj"+ly. These individual realizations of the above-given formula are termed 'morphological patterns'. The list of the metalinguistic annotation used in this formula is given below:

N	noun	Intj	interjection
Verb	verb	Part	particle
Aj	adjective	Da	definite article
Num	numeral	Abbr	abbreviation
Ad	adverb	BM	bound morpheme
Prep	preposition	Verb*2	past form of verb
Conj	conjunction	Verb*3	past participle form of verb

Therefore, according to this method, all words and morphemes are assigned to a particular morphological category, i.e. a word class. Three types of word classes are distinguished in this study: morphological (N, Aj, Verb, Ad, Num and Intj), conversive (which include words that belong to two or more morphological classes at the same time, e.g. N/Aj, Aj/Ad, N/Aj/Ad and N/Intj) and grammatical (e.g. Conj, Part, Intj, Da and Prep). Bound morphemes (BM) are also treated as a morphological class, but they do not show syntactic involvement. Further, the categories of all morphemes and words have been determined by the Oxford Etymological Dictionary, which eliminates any subjectivity in category assignment.

Further, from the computational perspective, the formal morphological analysis is a procedure restricted in time and space. This is because, first, it is performed on a dictionary[15] which has a limited number of entries, and, second, there are phonological, morphological and semantic constraints that limit the possibilities of morphemes' combinations.[16] Therefore, the procedure of the morphological analysis consists of a finite number of steps.

One step up in the generalization, we reach the level of constructions. Constructions, as defined by Croft (2007: 472), 'are fundamentally symbolic units', and they 'can be thought of as a linguistic pattern' (McArthur et al. 2018). On the morphological level, it can be generalized that a polymorphemic word with a clear etymology may consist of different combinations of two types of morphemes, namely a root and an affix. If the root is encoded with 'C' and an affix with 'a', then the following distinct combinations are possible in concatenative languages: {C-a}, {a-C},

[15] A list of words for this study has been taken from a dictionary, because, I believe, any trustworthy dictionary provides a credible sample of words in a language. Further, as shown in Krykoniuk (2020), the distribution of type frequency of suffixes in a dictionary is similar to that in a corpora.
[16] The limitation of morphemes' combinations also justifies the use of a dictionary, which provides informationally rich data.

{a-C-a}, {C-C}, {C-C-a}, {C-a-C}, etc. We can consider these combinations as a formal representation of general morphological constructions, where the schematic slot determines one of the two types of morphemes in a morphologically complex word. These morphological constructions represent a higher level of abstraction in the sequencing of morphemes of the lexicon (Krykoniuk 2020: 4). Furthermore, for the convenience of the analysis, the variation of this annotation has also been used in some tasks, where the slot for affix 'a' has been replaced with the actual affix (e.g. {C+ness}).

The highest level in this hierarchy of generalization is formed by meta-constructions which represent three major morphological word-formation processes: prefixation, suffixation and compounding. They are formalized as follows: {{a-C}}, {{C-a}} and {{C-C}}. These meta-constructions are discussed in Chapter 6.

3.2 The research area of the study

The previous section has established that the formal morphological analysis is a tool for exploring sequences of morphemes. In the current section, this area of the application of the formal morphological analysis is outlined.

3.2.1 The sample

Two Oxford dictionaries have been chosen for compiling a sample of this study: the Pocket Oxford English−German dictionary (4th ed.; online version) and the Oxford Etymological Dictionary (OED). The reason behind selecting Oxford dictionaries is that Oxford University Press is a widely-accepted authority in the field of lexicography. Further, the choice of the Pocket Oxford English−German dictionary is motivated by its comparatively smaller size. However, this dictionary largely contains relatively frequent present-day words (i.e. in an eight-band logarithmic scale of frequencies given in the OED, the band frequency of most words in this dictionary, as has been realized during the analysis, ranges from 4 to 8), and it omits words which are less frequent. For this reason, in order for my sample to be fully representative, I have randomly added words from the OED with a band frequency of 1 to 3.[17] By this token, a sample of 32,000 words has been created from the word list of the Pocket Oxford English−German dictionary and randomly chosen

[17] These adjustments have been made, because I did not consider a dictionary to be a random sample. I believe that the question of randomness of a linguistic sample deserves more attention from linguists, as there are some inconsistencies with the application of some statistical methods to language that still remain unresolved (e.g. Kilgarriff 2007).

entries of the Oxford Etymological Dictionary. In this study, a word is understood as a unit between two blank spaces.

In the course of the analysis, phrases[18] have been excluded from the list, except for hyphenated words. Despite the popular view in modern linguistics (e.g. adopted by Construction Grammar) that considers phrases with no grammatical expression to be part of derivational constructions, this study, in the spirit of a formalistic approach, draws a clear distinction between words and phrases. The latter are seen as units of the syntactic component of language. Therefore, the sample of 32,000 words contains only mono- and multimorphemic words.

3.2.2 Simplexes as morphological building blocks

Some approaches (e.g. Marchand 1960) exclude simplexes from the study of word formation, considering them unmotivated signs with no morphological structure. However, in view of morphology as the study of a lexicon, the merit of exploring them becomes obvious: simplexes are building blocks of a lexicon that contain useful information about syntactic, phonological and etymological processes in a language. One of the challenges facing word-formation theories is to establish the factors that determine what concepts in a language are denoted by simplexes and what concepts by derived words (Plank 2018). With this view, the study of simplexes seems unavoidable. Chapter 5 is devoted to the analysis of simplexes.

3.2.3 The level structure of word formation

The number of morphemes in a morphological pattern defines its level of word formation: simplexes constitute a zero-level of word formation, words with two morphemes the first level, words with three morphemes the second level, etc. By this token, a multilevel word-formation structure emerges that provides insight into word-formation processes with different degrees of complexity.

3.3 The meta-apparatus of the formal morphological analysis

I have termed the system of notions relevant to different aspects of the formal morphological analysis the meta-apparatus. It encompasses a number of terms that concern the main methodological tenets of the formal morphological analysis, procedures of the morphological description and the structure of word formation. They are as follows: metalanguage; morphological

[18] In this study, phrases are treated formally as units consisting of two or more words, separated by a blank space.

metacorpus; initiale, mediale and finale; hapax; unique/recurrent segments of the morphological metacorpus; and formal morphological paradigms. These are the tools for the description of morphological properties of English words.

3.3.1 The morphological metacorpus

With the help of the formal morphological analysis, all words in the sample have been manually converted into morphological patterns—formal units containing morphological, morphonological and etymological information about words. The totality of the morphological patterns is dubbed as the morphological metacorpus which, similar to a POS tagging system in corpus linguistics, records morphological patterns in a language. One of the indispensable functions of any text corpus is its POS-tagging system that allows for extracting grammatical categories of words. Similarly, the name of the 'morphological metacorpus' has been motivated by the fact that it permits the extraction of morphological information about words. The morpheme 'meta-' indicates that this corpus concerns a metalinguistic representation of linguistic units. Hence, the morphological metacorpus is useful in computing different quantitative characteristics of morphemes, morphological patterns and constructions.

A unique morphological pattern (i.e. produces only one word) is referred to as a morphological hapax. Due to its unproductivity, it can be viewed as a lexical unit. The share of all hapaxes in the metacorpus constitutes a unique segment of the morphological metacorpus. On the other hand, morphological patterns that spawn two or more words are defined as recurrent. They form a recurrent segment of the morphological metacorpus.

The entirety of morphological patterns within a morphological meta-construction is termed a formal morphological paradigm. These paradigms are represented in the form of graph networks in Chapter 6.

3.3.2 Initiale, mediale, finale

The notions of 'initiale', 'mediale' and 'finale' have been introduced in the analysis of morphological patterns to indicate the location of morphemes. I have borrowed these notions from a description of Chinese grammar (Švarný 1997: 172, 216, 483), where they specify the position of the initial, middle and final phonemes in a syllable. Accordingly, in the context of this study, the first slot in a morphological pattern is termed as the initiale, the middle slot as the mediale, and the last slot as the finale. This distinction of slots is useful in an optimal matrix analysis because it

helps establish combinatory properties of morphemes. Since an optimal matrix is three-dimensional (for the explanation of how to read the matrix, see Section 6.1.1), it is assumed that the initiale and finale slots allow only for one morpheme, whereas the mediale slot can take on a range of values—from a zero morpheme to three (or more) morphemes.

It is also important to mention that this formal distinction of slots in a morphological construction does not abolish the traditional division of morphemes into prefixes, roots and suffixes. The initiale, mediale and finale are merely the metalinguistic terms that are used for the description of the morphological metacorpus.

3.3.3 The methodological tenets of the formal morphological analysis

Four basic tenets provide the methodological foundation for the formal morphological analysis: completeness, simplicity, uniformity, and formalization. The completeness of the morphological description is achieved by considering all morphological and non-morphological phenomena and processes observed in the sample in the initial stage of the research. This means that non-morphological formations such as blending, corruption, split of senses and formations by analogy have been included into the description. In what follows, the annotations for the non-morphological word-formation processes are given.

R_A	formed by analogy	R_F	frequentative form
R_ACR	acronym	R_I	imitative
R_AL	altered lexical item	R_IN	inversion
R_APH	aphetic	R_O	onomatopoeic word
R_BF	formed by back-formation	R_RD	reduplication
R_BL	blending	R_S	semantic split
R_C	formed by corruption	R_SRT	shortened
R_CN	formed by conversion	R_SX	syntax formation
R_D	dialectic form	R_V	phonetic variant of a word
R_E	echoic/expressive	R-ind	individual coining
R_EPH	euphemism	R-pn	formed from a proper name

Some of these processes fall into broader word-formation categories of simplexes explored in Chapter 5: conversions (R_CN), phonological formations (R_A, R_AL, R_APH, R_C, R_D, R_F and R_V), contractions (R_BF, R_SRT), semantic formations (R_S, R_EPH, R-ind and R-pn) and onomatopoeic words (R_E, R_I and R_O). Others are considered non-morphological instances of compounding (R_ACR, R_BL, R_RD and R_SX).

The simplicity of the morphological description is attained by using a concise and transparent metalanguage, which is seen as a formalized semiotic code. According to the principle of simplicity, the metalanguage must not be more complex than the observed phenomena.

Further, the formalization principle is implemented through adopting a concise and laconic system of symbols. In addition to the metalinguistic annotation discussed above, the colon ':' is used to denote morphonological or orthographic changes if followed by brackets (e.g. *clumsily*: Ad'=Aj:(y→i)+ly), with the arrow '→' showing the direction of change. When the colon is succeeded by a letter or letters, it indicates the repetition of the last letter in a stem/word (e.g. *digger*: N'=Verb:g+er). The cent sign '¢' followed by a letter/letters encodes their omission (e.g. *actress*: N'=N(¢o)r+ess; *activation*: N'=Verb¢e+ion). Further, morphemes given in brackets before a suffix indicate 'virtual' morphemes, meaning that they have been added to a suffix for the reason of their tight, restricting connection. For example, the noun *conurbation* was formed in 1915 as *con-* + *-urb-* + *-(at)ion* and corresponds to the following morphological pattern: N"=con+BM+(at)ion. There is no such verb as **conurbate*, but the word-formation constraint of *-ate* + *-ion* has led to the emergence of the virtual morpheme (*-at-*) before *-ion*. Finally, the principle of uniformity implies a consistent application of the formal morphological analysis to the whole sample.

3.4 The importance of a diachronic perspective to morphological parsing

The main focus of this study is synchronic, which is defined by its goal: to provide a general and comprehensive description of word formation in present-day English. Therefore, only those morphemes that are meaningful to present-day native speakers have been assigned to grammatical categories.

Throughout the history of linguistic studies, a considerable bulk of French, Greek and Latin borrowings has posed challenges to a uniform description of the English lexicon and has led to multiple interpretations of the lexicon structure. Simplifying the state-of-the-art, morphological frameworks are classified in two camps: morpheme-based and word-based. Morpheme-based approaches (e.g. Distributional Linguistics) view morphemes as fundamental units of the lexicon and reduce linguistic analysis down to the study of morphemic sequences. In contrast, word-based approaches (e.g. Aronoff 1976) consider words and their phonological covariations to be the basic units of morphology (Haspelmath & Smith 2010: 41). The widely used argument against

morpheme-based approaches is that there are morphemes which do not fall into the definition of morphemes as smallest meaningful units of language. For example, in such words as *cranberry*, *boysenberry* and *huckleberry*, the morphological components *cran-*, *boysen-*, *huckle-* are morphologically distinguishable, but have no meaning, and, thus, cannot be considered morphemes. On the other hand, in the words *refer*, *defer*, *prefer*, *infer*, *confer* and *transfer*, less morphologically parsable elements can be identified: the shared component /fər/ and the prefixes *re-*, *de-*, *pre-*, *in-*, *con-* and *trans-*. Hence, if the morpheme is a unit that occasionally does not have a meaning (as shown in the first example), it is feasible to classify /fər/, given in the second example, as a morpheme. From the purely synchronic perspective, this chain of reasoning is logical.

However, language is not Kant's thing-in-itself—an isolated system independent of observation and experience. Rather, it is a dynamic entity whose "synchronic states are the result of a long chain of diachronic developments" (Bybee 2007b: 945). This is also true for morphemes which are at "a crossroad between diachronic and synchronic morphology" (Bolinger 1948: 18). The diachronic aspect of word-formation morphology in a synchronic dimension is reflected in, for example, different degrees of productivity of morphemes: some of them are highly productive, and some others are unproductive to the point that their meaning is lost. Thus, it is the ability of a morpheme to "enter into new combinations" (Bolinger 1948: 21) that bestows it with a synchronic legacy. With this view, /fər/ cannot be regarded as a morpheme, because it has never gained productivity in English.

The complex and dynamic nature of language is best explained with the metaphor of 'an airport terminal' used to describe a living organism in the famous book 'Descartes's Error: Emotion, Reason and the Human Brain' by Damasio (2000: 87, emphasis in original):

> "Imagine yourself in a large airport terminal, looking around, inside and outside. You see and hear the constant bustle from many different systems: people boarding or leaving aircraft, or just sitting or standing; people strolling or walking by with seeming purpose; planes taxing, taking off, landing; mechanics and baggage handlers going about their business. Now imagine that you freeze the frame of this ongoing video or that you take a wide-angle snapshot of the entire scene. What you get in the frozen frame or in the still snapshot is the image of a *state*, an artificial, momentary slice of life, indicating what was going on in the various organs of a vast organism during the time window defined by the

> camera's shutter speed. (In reality, things are a bit more complicated than this. Depending on the scale of analysis, the state of organisms may be discrete units or merge continuously)."

'An artificial, momentary slice' is an eloquent description of the synchronic dimension of a language. Although, by looking at this slice, it is still possible to distinguish some patterns and regularities, it would say nothing about the causes that have led to the emergence of certain grammatical and lexical features or a certain ordering of linguistic elements. In Roman Jacobson's (1951, cited in Apresyan 1966: 104) humoristic words elucidating the problem of a linguistic description, we can certainly chop off a hen's head and make valuable observations of its behavior in this state. However, it would be imprudent to assert that this state is natural for the hen and that by observing and studying it, we will learn all the essentials. These metaphoric accounts pinpoint the biggest problem with an exclusively synchronic approach—the artificiality.

Therefore, in order to comply with the methodological tenets of the formal morphological analysis on the one hand and to overcome the challenges of morphological and etymological heterogeneity of the English lexicon on the other, each word in the sample has been supplemented with a year of its first record, and the origin and the morphological structure of each word in the sample has been verified in the Oxford Etymological Dictionary. As a result, some formally similar words have been parsed differently, for the reason of their different history: e.g. *acceptance* (N-fr=BM+ance) and *utterance* (N'=Verb+ance). Moreover, words which have been formed in Old or Middle English and whose present-day phonological form have diverged from the original one have been assigned to a zero morphological level (e.g. the adverb *ghostly* formed as *gástlíce* in Old English: gást *ghost* n. + *-líce*, *-ly*).

As evident from the given example, the morphological parsing bias has been resolved by expanding the functional domain of bound morphemes and by assigning formally similar words to different levels of word formation, if their history differs (e.g. *acceptance* is assigned to the zero level of word formation and *utterance* to the first level). Hence, the category of bound morphemes is viewed as a broad class of morphemes, including (i) morphemes whose meaning has been forgotten by present-day speakers (e.g. *mutiny* formed from the obsolete noun or verb *mutine* and suffix *-y*; (ii) relatively modern morphemes of the Latin and French origin whose role is that between a suffix and a root (e.g. *aero-* in *aerosphere*; *auto-* in *autocross*); (iii) words which are formed from bound morphemes by analogy to Latin or French words or containing Latin/French

roots (*hypnot-* in *hypnotism*); and (iv) other kinds of creative formations (e.g. *shipping* coined from *-ship* in *relationship* and the suffix *-ing*; *Xerox* formed from *xero-* in *xerography* and *-x*). In short, a bound morpheme is a morpheme which is not used as an independent lexical unit.

3.5 Etymology as an important factor of morphological parsing

In the example above, the etymological information about words helps resolve a morphological bias of parsing. Thus, all simplexes in the sample have been encoded for their etymology. The following system of annotations has been adopted for this purpose. This list represents languages that have contributed to the diversity of the English lexicon.

R-abr borrowing from Australian aborigine
R-ad Adnyamathanha
R-afr Africaan
R-afr African
R-alg Algonquian
R-am American
R-anr Anglo-Romani
R-ar Aryan
R-arb Arabic
R-arm Aramaic
R-ass Assyrian
R-blg Bulgarian
R-bn Bangali
R-br British
R-brm Burmese
R-ch Chinese
R-chp Chinese Pidgin
R-clt Celtic
R-cnt Cantonese
R-cr Cree
R-crn Cornish
R-crt Croatian
R-ct Catalan
R-cz Czech
R-dh Dharuk
R-dn Danish
R-dt Dutch
R-fl Flemish (Belgian Dutch)
R-fn Finnish
R-fr French
R-fz Frisian
R-gl Gaelic
R-glb Galibi Carib
R-gr German
R-gr.pn Pennsylvania German
R-grh Middle High German
R-grk Greek
R-grl Middle Low German
R-grm Germanic
R-hb Hebrew
R-hi Hindi
R-hn Hungarian
R-hw Hawaiian
R-ic Icelandic
R-in Inuit
R-ind Indo-European
R-ir Irish
R-it Italian
R-jp Japanese
R-jv Javanese
R-lat Latin
R-ltz Latinized
R-mic Micmac
R-ml Malay
R-mnd Mandingo
R-mo Maori
R-mon Mon
R-mp Maliseet-Passamaquoddy
R-mrt Marathi
R-no Old Norse
R-np Nepali
R-nrn Norn
R-nrs Old Norse
R-nrth Northumbrian dialect
R-nrw Norwegian
R-ny Nyungar
R-oj Ojibwa
R-pal Pali
R-per Persian

R-pln Polynesian
R-pnj Panjabi
R-pr Peru
R-pt Portuguese
R-rm Romanic
R-rmz Romanized
R-rus Russian
R-sa South African (Dutch)
R-sc Scadinavian
R-sct Scottish
R-shn Shona
R-shr Sherpa
R-slv Slovak
R-smw Sierra Miwok
R-som Somali
R-sp Spanish
R-mx Mexican
R-srn Sranan
R-st Sanskrit
R-sw Swedish
R-swh Swahili
R-sx Saxon
R-tag Tagalog
R-tai Tahitian
R-tb Tibetan
R-tel Telugu
R-tg Tupi-Guarani
R-tha Thai
R-tm Tamil
R-tng Tongan
R-ts Tswana
R-tur Turkish
R-ur Urdu
R-va Virginia Algonquian
R-wl Welsh
R-yd Yiddish
R-yug Yugoslavian
R-yup Yupic
R-? unknown origin

3.6 Morphological parameters of the metacorpus

In the previous sections, the procedures, tenets and notions of the formal morphological analysis have been brought to light, which have allowed for the formation of the morphological metacorpus. This section looks at the parameters, with which the metacorpus has been assessed: type frequency, token frequency, type valency, realized productivity, diachronic productivity, potential productivity and type-token ratio. These morphological parameters have been used variously in regression models, cluster and graph network analyses, and they are the main focus of the study in Chapter 7.

3.6.1 Type frequency

Type frequency, which refers to the number of distinct items that can occur in the open slot of a construction (Bybee 2007a: 14), offers a helpful insight into the leading grammatical forces that drive language change. Type frequency is a relatively well-studied phenomenon in different linguistic frameworks (e.g. Baayen & Lieber 1997; Berg 2014; Bybee 1985, 2007a; Hay 2001; Hooper 1987 and Krug 1998).

The following three key accounts of type and token frequency can be summarized. In the synchronic dimension, the major effects of high type frequency are, on the one hand, an increase in productivity of a linguistic process (Bybee 1985), and the formation of schemas (Taylor 2002)

on the other. From the diachronic perspective, high type frequency may result in lexicon enrichment (Bybee 2001). Finally, Berg (2014) argues that type frequency, as compared to token frequency, displays more extreme values.

In this research, the values of type frequency from four sources are considered. First, the type frequencies of morphological patterns have been calculated from the list of words in the sample. The database (Laws & Ryder 2014) constitutes the second source of frequencies which are integrated into research in cluster analysis of Chapter 7. MorphoQuantics was compiled on the basis of Spoken BNC2014 with 10,000,000 tokens (Love et al. 2017). Above four million of its tokens comprises spontaneous conversational English. According to the BNC's compilers (BNC 2015, para. 3), representativeness in the spoken component of the BNC "was achieved by sampling a spread of language producers in terms of age, gender, social group, and region, and recording their language output over a set period of time". Third, the MorphoLex database (Sánchez-Gutiérrez et al. 2018) served as a source for the third set of type frequency in this study. MorphoLex is based on the English Lexical Project (ELP; Balota et al. 2007), which provides various counts for psycholinguistic studies. Specifically, MorphoLex's variables of the type and token frequencies were calculated from the HAL (Hyperspace Analogue to Language) corpus with 130 million tokens, gathered across 3,000 Usenet newsgroups during February 1995 (Balota et al. 2007: 450). The HAL corpus exemplifies conversational English on the Internet. Finally, the fourth set of type frequency values has been taken from the CELEX corpus, as reported in Hay and Baayen (2001). The corpus is based on the Cobuild corpus and consists of 75% written and 25% spoken language and encompasses broadly general adult language (from 1961 to 1990) without instances of poetry, drama and regional dialects (Baayen & Lieber 1991: 803).

On a more abstract level, type frequency is a major factor determining the degree of productivity of a construction (Bybee 2007a: 14). Therefore, the concept of the 'formal productivity' of a construction has been introduced in this study whereby productivity is defined in terms of the type frequency of slots of a formal construction: the higher the number of formal elements that can be placed in the open slots of the construction, the more formally productive a construction is. In fact, the formal productivity of a morphological construction defined in terms of the type frequency is a 'realized productivity' proposed by Baayen (2009: 902). The only difference is that it applies to a formal morphological construction, as understood in the framework of the formal morphological analysis.

3.6.2 Token frequency

Token frequency considers the number of times a unit appears in running text (Bybee 2007: 9). Similar to type frequency, the effects of token frequency have been widely studied (e.g. Bybee 1985, 2007; Tottie 1991). The two most known token-frequency effects are the Conserving and Reducing Effects. The former is observed in the fact that tokens with high frequency resist reformation on the basis of analogy with other forms (Bybee 2007: 10), whereas the latter leads to the reduction of the phonological form of a high-frequency token. This study uses two sets of token frequency. The first set has been taken from MorphoQuantics, and the second set from MorphoLex (both sources are described in the previous section). These measures are used in cluster analyses discussed in Chapter 7.

3.6.3 Type valency

Type valency is the measure of a morpheme's connectedness to different types of morphemes (Krykoniuk 2020). For example, the English suffix *-er* attaches to verbs (e.g. *rubber*), nouns (e.g. *cricketer*), adjectives (e.g. *deader*) and bound morphemes (e.g. *soccer*). Therefore, its type valency is 5. In this research, the type valency of word-formation processes has been calculated from the studied sample.

Valency as a combinative power of linguistic units has received less attention in linguistic theories. The notion of 'valency' was first introduced in a linguistics context by the Soviet linguist Solomon Katsnelson (1948) to denote the ability of words to combine with other words. In the same year, the British linguist Dwight Bolinger also touched upon the phenomenon of valency (although without using this term). Bolinger (1948: 18) suggested that the valency, or, in his words, 'the statistically determinable readiness with which an element enters into new combinations, is the only sure linguistic evidence that the element has a meaning of its own'. Thus, Bolinger (1948: 21) defines the morpheme in terms of its potential to combine with other morphemes, which laid the foundation for new statistical measures of productivity in later stages of the development of linguistics (see Plag 2003: 51). Thereafter, the notion of 'valency' was incorporated in the study of syntax in a number of theories. Tesnière (1959) first used this concept to describe the idea of how verbs attract 'actants' (i.e. subjects and objects accompanying a verb) in forming clauses. Later, the term 'valency' found its way into a few syntactic theories in generative grammar becoming central to the notion of 'argument structure', e.g. Government and Binding Hypothesis (Chomsky 1981, 1982), and Argument Realization (Levin & Rappaport-Hovav 2005), as well as

into theories which are more oriented towards natural language processing, such as Generalized Phrase Structure Grammar (Gazdar et al. 1985), and Head-Driven Phrase Structure Grammar (Pollard & Sag 1994). It was then borrowed into the theory of Construction Grammar. Used as a synonym for 'argument structure', valency in Construction Grammar is also mainly seen as a property of verbs (Hilpert 2014: 41).

3.6.4 Productivity

The quantifiable measure of morphological productivity (largely known in its two versions—'potential productivity' and 'global productivity') has probably received less attention, as compared to type and token frequency. The discussion of productivity of suffixes was initiated by Aronoff (1976) within the context of generative theory of word formation, and was then taken further by Baayen & Lieber (1991) and Baayen (1991) who developed formulae for capturing the degree of productivity. Despite the valid criticism of these measures (e.g. van Marle 1992; Bauer 2001) that point to the possible counter-intuitive results coming from their application, they are the most plausible quantitative evaluations of productivity of morphemes introduced so far that create an approximate picture of derivational processes in a corpus.

Four measures of productivity are considered in this study: realized productivity (or formal morphological productivity), diachronic productivity, expanding productivity and potential productivity. Realized productivity (Baayen 2009: 901) is estimated by the number of types in a morphological category. The development of the realized productivity over time has been termed 'diachronic productivity'. Further, expanding productivity is a ratio of the number of types in a morphological category and the total number of hapax legomena in a corpus (Baayen 2009: 902), whereas potential productivity is estimated as the hapax legomena in a morphological category divided by the total number of its tokens (ibid.). Specifically, realized productivity is used in this study to describe morphological constructions (Chapter 6). Diachronic productivity is considered in entropy estimation, and expanding and potential productivity in cluster analyses (Chapter 7). The values of potential productivity have been taken from CELEX (as reported in Hay & Baayen 2002) and from MorphoLex (Sánchez-Gutiérrez et al. 2018), and the values of expanding productivity from MorphoLex (ibid.).

3.6.5 Type-token ratio

The measure of type-token ratio has received a large interest among linguists, specifically, in the field of psycholinguistics (e.g. for the study of child language acquisition, see Silverman & Ratner 2002; Demir-Lira et al. 2019). In its broadest definition, it is taken as a measure of how lexically complex/rich/varied the vocabulary of a text (written or spoken) is. Some other scientists (e.g. Popescu 2009) take TTR as a measure of 'information flow' and 'topic deployment'. Various statistical models were developed to measure the index of TTR. Some of them are refined with a parameter (e.g. Yule 1944), others with the help of the cumulative function (Yomans 1991). However, as the weaknesses of these models became apparent, new measures of TTR were introduced (e.g. the moving-average type-token ratio). In the current study, the measures of raw TTR have been used, which is a ratio of type and token frequency. The TTR values, used in cluster analyses, have been calculated on the basis of MorphoQuantics and MorphoLex.

3.7 Conclusions

In this chapter, the basic tenets, terminology and procedures of the formal morphological analysis have been described: e.g. the morphological metacorpus; initiale, mediale and finale; morphological hapax and recurrent pattern. Furthermore, I have discussed the criteria for the morphological parsing of words in the sample. Finally, an overview of the parameters has been given with which the metacorpus has been studied in statistical analyses. The next chapter focuses on statistical methods used in this research.

4 Statistical tools

Perhaps one of the first attempts to use statistical tools in analyzing linguistic data was made by Zipf (1935). Since then, statistical tools have been widely used in linguistics to describe the distribution of various linguistic units, phenomenon and languages. Thanks to these methods we came to know, for example, that the most frequent words are shorter (this regularity is known as Kaeding–Zipf–Flesh's regularity), have higher number of meanings (Zip–Yule's regularity), tend to combine with a larger number of words (Zipf–Herdan's regularity) and are the oldest in a lexicon of languages (fourth Zipf's law) (see Tyshchenko 2007: 147). Further, the number of speakers in each of 18 language families decreases in accordance with Zipf's law, and the steepness of slope in the Zipfean distribution of words within texts depends on their genres and functional styles (Tyshchenko 2007: 60).

The emergence of statistical software has fostered the application of statistical tools to linguistic data. One of the leading roles among these software systems belong to *R* and *RStudio* (R Development Core Team). Therefore, I have chosen *RStudio* as a means to analyze the data of this study.

Thus, the main thrust of this chapter is to introduce the statistical methods used in this research, as well as their realization in *RStudio*, with the purpose that the reader does not have to consult textbooks on statistics to understand the flow of this research and its outcomes. If the reader is familiar with the discussed statistical procedures, they may skip this part. In Section 4.1, the concept of correlations is discussed. Section 4.2 presents Poisson regression fitted to the data of this research for the purpose of studying relationship between the type frequency of suffixes and their type valency. The focus of Section 4.3 is three estimators of Kullback-Leibler Divergence and the *R* package 'kldtools', which have been applied for the comparison of band frequencies between most type-frequent morphological patterns. Section 4.4 looks at three clustering techniques: agglomerative clustering, k-medoids and Principal Component Analysis (PCA). These methods complement each other, allowing for the construal of a general picture from the data. Section 4.5 introduces a method borrowed from a graph theory that has been used to visualize relationships between various morphological categories in formal morphological paradigms. Finally, Section 4.6 concludes the chapter.

4.1 Correlations

We live in a world of causality: a cause produces an effect which itself, over the course of times, becomes a cause. Objects and phenomena of the outer reality are constantly interacting and influencing each other—this is how our world evolves. In science, thus, most research is concerned with various kinds of relationship between the elements of objective reality. By knowing how objects and phenomena are related, we become powerful, because we can affect and change them, if necessary, and get the desirable outcome.

Statisticians have developed different methods of studying the strength of quantitative relations. In this section, I will look at one of the measures expressing relations between variables: the *correlation coefficient*. Generally speaking, it tells us to what extent the change in one variable leads to a change in another. However, it reveals nothing about the causality of this relation, that is, whether one phenomenon is caused by the other. Also, the correlation coefficient does not capture directionality: whether the variable *x* influences the variable *y* or vice versa. Thus, while reporting the results of correlations, researchers should be cautious not to interpret them in terms of causality and directionality.

There are three popular correlation coefficients: the Pearson product-moment correlation coefficient, Spearman's correlation coefficient and Kendall's tau. They have different assumptions and are, therefore, used in different contexts. Nevertheless, all of them are measured within a scale between –1 (a negative relationship: an increase in one variable leads to an increase in the other) and +1 (a positive relationship: an increase in one variable leads to a decrease in the other). Further, the relationship with the coefficient ± .1 is considered weak, ± .3 medium, and ± .5 strong (Field et al. 2012: 209).

Since, in this research, the studied variables are frequencies which are not normally distributed, Spearman's non-parametric correlation coefficient has been applied. This test uses Pearson's formula of the standardization of covariance, but before that it ranks the data and, then, calculates the correlation coefficient.

However, knowing the correlation coefficients for the data is not sufficient to draw conclusions about correlations. We also need to know the statistical importance of an effect in these correlations. For this purpose, statisticians have designed a particular statistical procedure which is widely used for variety of statistical models and which allows us to decide on criteria for whether the effect is significant or not. It is known as null hypothesis significance testing.

Although, recently, this procedure has been highly criticized by some statisticians (e.g. Levine et al. 2008), it is the best that we have for the moment. Hence, while interpreting the results, it is important to remember the conventionality of this procedure and to avoid categorical language.

Before calculating the correlation coefficient, we have to decide on the threshold of probability (e.g. .05) below which we can assume with the confidence of 95% or more (depending on the set threshold) that the effect is statistically significant. The probability below .05 has become known as Fisher's criterion or α-level. The opposite way to interpret this criterion is to say that there is only a 5% chance that the obtained results might be due to chance. In fact, the significance level can be set at any point, but the most widely used conventional significance levels are .01, .025, and .05. In social sciences, the significance level of .05 is considered to be the most appropriate (Baayen 2008: 69) as social phenomena are fuzzy, multidimensional and hard to control. Linguistics follows the same pattern.

Therefore, in order to understand whether the effect in a correlation is significant or not, we design two hypotheses. The null hypothesis states that there is no effect between variables in a sample, while the alternative hypothesis asserts the opposite: that the effect between variables exists. Then, the significance level is set—it is the point where we decide (after the calculations being made) whether to accept or to reject the null hypothesis. If the *p*-value for the correlation turns out to be above the set significance level, there is no sufficient ground to reject the null hypothesis of no correlation, i.e. we must conclude that our variables have no sensible relation. Conversely, the *p*-value below the α-level indicates that the variables are related and the effect of this correlation is statistically significant.

Further, the alternative hypothesis can be directional or non-directional. As follows from its name, the directional hypothesis states that we are interested in the direction of the relationship between variables: whether *x* is greater or less than *y*. Conversely, the non-directional hypothesis only identifies the difference between variables. This distinction is important when deciding on the appropriate statistical test. Normally, for the directional hypothesis, we must choose a one-tailed test, whereas, for non-directional hypotheses, two-tailed tests are usually used. One-tailed and two-tailed tests are different in how they calculate statistics, and if we select an unsuitable test, our decision about the statistical significance of a relation might be wrong.

In addition, when deciding on the statistical significance of a correlation, it is important to be aware of two types of errors that might emerge from an inappropriately chosen significance

level. Type I error arises with Fisher's criterion (the probability of .05) when an effect is identified as statistically significant while in fact it is not. On the other hand, Type II error may creep into one's research when the significance level is set at the so-called *β*-level (the probability of .02) and when we reject the correlation effect which is real. Type I and Type II errors are interconnected in the same way as the two seemingly contrary forces of yin and yang are related to each other: the attempt to decrease one type of error leads to an increase in the probability of the other and vice versa. There are various suggestions as to how to avoid them (e.g. Howell 2007). However, the most common advice is to make an educated guess.

In recent decades, the performance of the correlation analysis has been made easy thanks to various statistical programs. One of these programs is *R*, a free software environment for statistical computing and graphics, which is used in this research. In *R*, the functions of the correlation analysis are available from such packages as 'Hmisc', 'polycor', 'boot', 'ggplot2' and 'ggm'. Basic correlation coefficients are computed with the functions cor(), cor.test() and rcorr(), and the method of the correlation is specified as an argument in brackets. In the current study, the Spearman correlation analysis has been conducted to, first, establish the relationship between the type frequency of a suffix and its type valency and, second, to identify a degree of the association between word bases on the first and higher levels of derivation.

4.2 Poisson regression analysis

One step further in studying relations between variables is regression analysis. Using various regression models, it is possible, first, to explore the dynamics and trends in the studied data, and, secondly and more importantly, to make predictions about phenomena under investigation. As with correlation analysis, regression models do not imply causation, although for the purpose of regression analysis, we label the variable which is influenced as dependent (usually located on the *y*-axis) and the variable which exerts an effect as independent (located on the *x*-axis). Therefore, it is possible to hypothesize that changes in the dependent variable *x* are caused by or are associated with the changes in the independent variable *y* (Sokal & Rohlf 1969: 496).

There are various regression models, and which one to choose depends on the distribution of variables, their central/dispersion tendencies, and their nature (whether they are nominal, ordinal, interval or continuous). Hence, it is said that a regression model has assumptions which need to be verified before regressing variables. Further, after running a regression analysis, the

goodness of a model's fit has to be checked: i.e. how well the fitted model explains analyzed data. Different statistical coefficients from a regressed model, as well as various statistical tests, help us to answer this question.

Most variables of this research are frequencies which are regarded as discrete, i.e. they can take on only certain values (Field et al. 2012: 917) reflecting the number of occurrences of an event in a fixed period of time (Coxe et al. 2009: 121). For this reason, the Ordinary Least Square regression, for example, may potentially pose a problem as it can lead to larger standard errors (thus, increasing chances for Type I error) and to biased significance tests (Coxe et al. 2009: 121). Poisson regression, which is a member of a family of analyses known as generalized linear models (GLM), seems to be a better tool for most of this study's data for two reasons: first, the observed scores of Poisson regression are counts, and, secondly, it is flexible in error structure (Coxe et al. 2009: 122).

Poisson regression does not require that the relation between the dependent and independent variables follow a straight line. In order to transform the non-linear relation into the linear, Poisson regression uses the 'log' link function, as shown in (1) below:

(1) $$\log(Y_i) = (\beta_0 + \beta_1 \mathrm{X}_i) + \varepsilon$$

Poisson regression assumes that observations are randomly sampled and that the occurrence of one event does not influence the occurrence of another, i.e. events are independent (Taeger & Kuhnt 2014: 72). Further, a unique property of Poisson regression is the equality of mean and variance (Salkind & Neil 2006: 772). Hence, before deploying this regression analysis, it is important to verify that the distribution of the dependent variable follows this requirement. When the variance is less than the mean, we observe a phenomenon called underdispersion, and when the variance is larger than the mean, data is considered overdispersed. If the assumption of the equality of mean and variance is violated, quasi-Poisson regression should be used instead.

Moreover, the properties of Poisson distribution are similar to the binomial distribution (Baayen 2008: 54). However, as shown in Baayen (2001: 45), in the study of word frequency distribution, Poisson distribution has an advantage over the binomial for two reasons: usually, there is a large discrepancy between the size of a corpus and the frequency of words, and its mathematical properties seem to be more suitable.

As mentioned, Poisson regression belongs to the category of the generalized linear models (GLM). Hence, the glm() function with the 'Poisson' family and the 'log' link specified as arguments allows us to compute Poisson regression in *R*. The summary() function produces the output of a statistical computation for a model. These functions are available in the *R* 'Stats' package. Assumptions about the residuals of Poisson regression are checked with residuals vs fitted values and Q-Q plots with the help of the plot() function, although the distribution of residuals in Poisson regression of not required to be normal (Faraway 2016: 127). The validation of Poisson GLMs can be performed with bootstrap simulation (the *R* packages 'ciTools', 'trending', 'patchwork' and 'MASS').

Specifically, Poisson regression modelling has been used in the current study to validate the impact of the type frequency of suffixes on their type valency. The fitted models have been checked with bootstrapping.

4.3 KLD non-parametric estimators

Kullback-Leibler Divergence, one of the most important measures in information theory, is a measure of a difference between two probability distributions. It was first introduced by Kullback and Leibler (1951) for use in cryptography, and has since become known under a variety of names: the Kullback-Leibler distance, cross-entropy, information divergence and information for discrimination (Cover & Thomas 2006: 54).

The KLD formula is based on the following assumptions. Let $\mathbf{p} = \{p_k,\ k = 1, \cdots,\ K\}$ and $\mathbf{q} = \{q_k,\ \ k = 1, \cdots, K\}$ be two discrete probability distributions on the same alphabet, $\mathcal{X} = \{l_k, k = 1, \ldots, K\}$, where $K \geq 2$ is a finite integer. Kullback-Leibler Divergence is defined by the following formula:[19]

[19] In the definition of the KLD, any base (*b*) of logarithm can be applied. The commonly used values of *b* are 2, Euler's number ($e \approx 2.7183$), and 10, and the corresponding units of entropy are bits for $b = 2$, nats for $b = e$, and bans for $b = 10$ (Schneider 2013). In this paper, the base $b = e$ is used, which means that the obtained values are nats for units of entropy. The difference between the natural and binary logarithm is a constant, as shown in the following:

$$\log_b x = \frac{\log_k x}{\log_k b}\ , k \neq 1, b \neq 1.$$

$$\log_2 x = \frac{\log x}{\log 2} = \frac{1}{\log 2}\log x\ , \frac{1}{\ln 2} \approx 1.4427 \ldots\ , \log 2 = \ln 2 = 0.69315.$$

$$D = D(\mathbf{p}||\mathbf{q}) = \sum_{k=1}^{K} p_k \ln(p_k/q_k) = \sum_{k=1}^{K} p_k \ln(p_k) - \sum_{k=1}^{K} p_k \ln(q_k). \quad (2)$$

In this formula, the first sum is related to Shannon information or Shannon entropy, one of the most commonly used measures of randomness of a distribution **p**:

$$H = -\sum_{k=1}^{K} p_k \ln(p_k), \quad (3)$$

Here and throughout, the following standard conventions are adopted: $0 \ln\left({}^{0}/_{q}\right) = 0$, if $q \geq 0$ and $p \ln\left({}^{p}/_{0}\right) = \infty$, if $p > 0$.

KLD has two major features: first, it is not metric, since it does not satisfy the triangle inequality, and it is not symmetric (Zhang 2017: 183); secondly, it is always non-negative and is zero if and only if $\mathbf{p} = \mathbf{q}$ (Cover & Thomas 2006: 19). Therefore, KLD is a measure of information (but not a distance), revealing how much information is lost when we try to approximate the **q** distribution to the **p** distribution.

In the following subsections, three statistical estimators are introduced that allow for hypothesis testing with the Kullback-Leibler Divergence. Since this theory is new to linguistics, it is introduced in greater mathematical detail. This method has been applied for comparing the diachronic productivity of the most type-frequent constructions of the morphological metacorpus. This comparison has been performed with the help of the package 'kldtools' (Krykoniuk & Shipunov 2021) which calculates these estimators, as well as carries out bootstrapping.

It is important to mention that, although the purpose of these three estimators is to capture the dependence between the pairs of distributions, they are not equivalent—i.e. they might produce different results. However, as shown in Zhang (2017), the Turing's perspective estimator has a

smaller bias for point estimation, which, empirically speaking, means that this entropy estimator is expected to have a larger power.[20]

4.3.1 Hypothesis testing with KLD

Zhang (2017) has proven theorems that enable hypothesis testing with KLD. In this section, the procedure of hypothesis testing is described, as applied to this study. The method is multifaceted and has various features. Thus, for a more detailed mathematical account of the procedure, I refer the interested reader to Zhang (2017).

The central idea of this procedure is a non-parametric estimation of Kullback-Leibler Divergence. Zhang and Grabchak (2014: 2570) showed that, on any finite alphabet, this estimator is consistent (i.e. its precision increases with the increase of a sample) and is asymptotically normal (i.e. the estimator approximates normal distribution as a sample size approaches infinity).

Assume that there are two independent identically distributed (*iid*) samples of sizes *n* and *m* with the unknown distributions **p** and **q**, respectively. Further, let $\{X_1, \cdots, X_K\}$ and $\{Y_1, \cdots, Y_K\}$ be the sequence of the observed frequencies of letters $\{l_1, \ \cdots, \ l_K\}$ in two samples, respectively, and let

$$\widehat{\mathbf{p}} = \{\hat{p}_k, \ k = 1, \cdots, K\} = \left\{\frac{X_1}{n}, \frac{X_2}{n}, \cdots, \frac{X_K}{n}\right\},$$
$$\widehat{\mathbf{q}} = \{\hat{q}_k, \ k = 1, \cdots, K\} = \left\{\frac{Y_1}{m}, \frac{Y_2}{m}, \cdots, \frac{Y_K}{m}\right\},$$

be sequences of the corresponding relative frequencies (Zhang 2017: 183), where *n* and *m* are the sums of frequencies in the distribution $\widehat{\mathbf{p}}$ and $\widehat{\mathbf{q}}$, respectively.

The following assumptions are further imposed: $p_k > 0, \ k = 1, \cdots, K$, and $q_k > 0, k = 1, \cdots, K$, and there exists a $\lambda \in (0, \infty)$ such that $n/m \to \lambda$, as $n \to \infty$.

The 'plug-in' estimator of KLD (2) is given in formula (4) below (see Zhang 2017: 183; formula 5.88).

(4) $$\widehat{D} = \widehat{D}_n(\widehat{\mathbf{p}}||\widehat{\mathbf{q}}) = \sum_{k=1}^{K} \hat{p}_k \ln(\hat{p}_k) - \sum_{k=1}^{K} \hat{p}_k \ln(\hat{q}_k),$$

[20] I would like to thank Dr. Jialin Zhang for clarifying this point.

The basic statistical properties of formula (4) are described below, following Zhang (2017, in Section 5.3.1).

Define the $(2K-2)$-dimentional vector-columns $\mathbf{v} = (p_1, \cdots, p_{K-1}, q_1, \cdots, q_{K-1})^{\boldsymbol{\tau}}$, $\hat{\mathbf{v}} = (\hat{p}_1, \cdots, \hat{p}_{K-1}, \hat{q}_1, \cdots, \hat{q}_{K-1})^{\boldsymbol{\tau}}$, where the upper symbol τ indicates the transpose vector. Then, the estimate $\widehat{D}$ is a consistent and asymptotically normal estimator of D as $n \to \infty$. This means that, in probability $\left(\overset{p}{\to}\right)$, $\widehat{D} \overset{p}{\to} D$, and, in distributions $(\overset{D}{\to})$,

$$\frac{\sqrt{n}(\widehat{D} - D)}{\sigma} \overset{D}{\to} N(0,1). \tag{5}$$

In the formula above, $N = N(0,1)$ is a random variable with the standard normal distribution, i.e.:

$$P\{N \leq x\} = \int_{-\infty}^{x} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\, dz = \Phi(x), \tag{6}$$

which is tabulated—i.e. using the table of normal distribution, we can always find the quantile $\Phi_{1-\frac{\varepsilon}{2}}$ such that for a given $\varepsilon > 0$,

$$P\left\{|N| \leq \Phi_{1-\frac{\varepsilon}{2}}\right\} = 1 - \varepsilon. \tag{7}$$

For example, if $\varepsilon = 0.10, \Phi_{1-\frac{\varepsilon}{2}} = 1.64485363$; if $\varepsilon = 0.05, \Phi_{1-\frac{\varepsilon}{2}} = 1.95996340$; if $\varepsilon = 0.01, \Phi_{1-\frac{\varepsilon}{2}} = 2.57582930$.

The unknown variance σ^2 in (5) is given by

$$\sigma^2 = \mathrm{g}^{\tau}(\mathbf{v}) \sum(\mathbf{v})\, \mathrm{g}(\mathbf{v}), \tag{8}$$

where $\mathrm{g}(\mathbf{v})$ is the vector

$$\mathrm{g}(\mathbf{v}) = \left(\ln\frac{p_1\, q_K}{q_1\, p_K} \cdots, \ln\frac{p_{K-1}\, q_K}{q_{K-1}\, p_K}\ ;\ -\frac{p_1}{q_1} + \frac{p_K}{q_K}, \cdots, -\frac{p_{K-1}}{q_{K-1}} + \frac{p_K}{q_K}\ \right)^{\tau}. \tag{9}$$

This is a corrected version of the formula (5.94) of the book (Zhang 2017: 185), the inconsistency in which has been identified in the process of the creation of the package ‘kldtools’ and by comparing the results obtained from the application of the formula (5.94) and its simplified version

for k = 2 (Zhang 2017: 187; see formula (11) below). A deeper enquiry into this problem has led to detecting the formatting error in the formula given in the book: it is missing the elements for the $\hat{\mathbf{p}}$ distribution.[21]

In the formula (8), the quasi-diagonal $(2K-2) \times (2K-2)$ matrix $\sum(\mathbf{v})$ is obtained by

(10) $$\sum(\mathbf{v}) = \begin{pmatrix} \sum_1(\mathbf{v}) & 0 \\ 0 & \sum_2(\mathbf{v}) \end{pmatrix}.$$

This matrix consists of two $(1K-1) \times (1K-1)$ matrices given by:

$$\sum_1(\mathbf{v}) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{K-1} \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ -p_{K-1}p_1 & -p_{K-1}p_2 & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix},$$

and

$$\sum_2(\mathbf{v}) = \lambda \begin{pmatrix} q_1(1-q_1) & -q_1q_2 & \cdots & -q_1q_{K-1} \\ -q_2q_1 & q_2(1-q_2) & \cdots & -q_2q_{K-1} \\ \cdots & \cdots & \cdots & \cdots \\ -q_{K-1}q_1 & -q_{K-1}q_2 & \cdots & q_{K-1}(1-q_{K-1}) \end{pmatrix}.$$

Note that, for $K = 2$, the formula (8) can be simplified as follows:

(11) $$\sigma^2 = p(1-p)\left[\ln\frac{p(1-q)}{q(1-p)}\right]^2 + \frac{\lambda(p-q)^2}{q(1-q)}.$$

Using Slustky's theorem, it is possible to show that formulae (5) and (8) imply the following:

$$\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\widehat{D}_n - D\right) \xrightarrow{D} (N(0,1), n \to \infty,$$

where

$$\hat{\sigma}_n = \left[\mathrm{g}^{\tau}(\hat{\mathbf{v}})\sum(\hat{\mathbf{v}})\,\mathrm{g}(\hat{\mathbf{v}})\right]^{1/2}.$$

[21] I would like to thank Dr. Jialin Zhang for verifying the correct formula of the vector in the calculation of the variance of the KLD plug-in estimator.

Based on this, we can construct the asymptotic confidence interval based on the following procedure. Since for a given ε and for large *n*, we can use the following approximation:

$$\mathrm{P}\left\{\left|\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\hat{D}_n - D\right)\right| \leq \Phi_{1-\frac{\varepsilon}{2}}\right\} \approx \mathrm{P}\left\{|N(0,1)| \leq \Phi_{1-\frac{\varepsilon}{2}}\right\} \approx 1 - \varepsilon,$$

then

$$\mathrm{P}\left\{\hat{D}_n - \Phi_{1-\frac{\varepsilon}{2}}\frac{\hat{\sigma}_n}{\sqrt{n}} \leq D \leq \hat{D}_n + \Phi_{1-\frac{\varepsilon}{2}}\frac{\hat{\sigma}_n}{\sqrt{n}}\right\} \approx 1 - \varepsilon.$$

This means that, for a given level of confidence $1 - \varepsilon$, the $100(1 - \varepsilon)\%$ confidence interval is of the form:

(12) $$CI_n = \left(\hat{D}_n - \Phi_{1-\frac{\varepsilon}{2}}\frac{\hat{\sigma}_n}{\sqrt{n}};\ \hat{D}_n + \Phi_{1-\frac{\varepsilon}{2}}\frac{\hat{\sigma}_n}{\sqrt{n}}\right).$$

Suppose that, for a given D_0 (the measure of information between two systems of probabilities on the same alphabet), we would like to test the null hypothesis (H_0: $D = D_0$) against the alternative (H_1: $D \neq D_0$). Then, if $D_0 \in CI_n$, we accept the null hypothesis, but if $D_0 \notin CI_n$, we reject H_0.

Since $D(\mathbf{p}, \mathbf{q}) \geq 0$, and $D(\mathbf{p}, \mathbf{q}) = 0$, if and only if $\mathbf{p} = \mathbf{q}$, it is inferred that $D_0 = 0$ if and only if the two systems of distributions $\mathbf{p}$ and $\mathbf{q}$ on the same alphabet are similar. However, if $D_0 \neq 0$, this means that these distributions are different. In other words, if, for example, $\varepsilon = 0.05$,

$$0 \in \left(\hat{D}_n - 1.95996340\ \frac{\hat{\sigma}_n}{\sqrt{n}};\ \hat{D}_n + 1.95996340\ \frac{\hat{\sigma}_n}{\sqrt{n}}\right) = CI_n,$$

then two systems are similar. On the other hand, if $0 \notin CI_n$, two systems are different. This rule can be represented as follows:

H_0: $\hat{\mathbf{p}} = \hat{\mathbf{q}}$, if $\frac{\sqrt{n}(\hat{D}_n - D)}{\hat{\sigma}_n} \leq \Phi_{1-\frac{\varepsilon}{2}}$

H_1: $\hat{\mathbf{p}} \neq \hat{\mathbf{q}}$, if $\frac{\sqrt{n}(\hat{D}_n - D)}{\hat{\sigma}_n} \geq \Phi_{1-\frac{\varepsilon}{2}}$.

4.3.2 Hypothesis testing with the symmetrized KLD

KLD is not symmetric, which means that $D(\mathbf{p}||\mathbf{q}) \neq D(\mathbf{q}||\mathbf{p})$, if $\mathbf{q} \neq \mathbf{p}$. On this account, the symmetrized measure of KLD was also developed:

(13) $$S = S(\mathbf{p}, \mathbf{q}) = \frac{1}{2}(D(\mathbf{p}||\mathbf{q}) + D(\mathbf{q}||\mathbf{p}))$$

It is clear that $S(\mathbf{p}, \mathbf{q}) = S(\mathbf{q}, \mathbf{p})$, even if $\mathbf{q} \neq \mathbf{p}$.

The procedure of the hypothesis testing is similar to that described in Section 4.4, but with two minor differences. First, the plug-in estimator for symmetrized KLD is defined as follows (Zhang 2017: 193):

(14) $$\hat{S} = \hat{S}(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \frac{1}{2}(\hat{D}_n(\hat{\mathbf{p}}||\hat{\mathbf{q}}) + \hat{D}_{\boldsymbol{m}}(\hat{\mathbf{q}}||\hat{\mathbf{p}}))$$
$$= \frac{1}{2}\left(\sum_{k=1}^{K} \hat{p}_k \ln \hat{p}_k - \sum_{k=1}^{K} \hat{p}_k \ln \hat{q}_k\right) + \left(\sum_{k=1}^{K} \hat{q}_k \ln \hat{q}_k - \sum_{k=1}^{K} \hat{q}_k \ln \hat{p}_k\right).$$

Secondly, the asymptotic variance must be calculated differently for symmetrized KLD. We introduce the $(2K-2)$ vector-column:

(15) $$\mathrm{g}_s(\mathbf{v}) = \left(\frac{1}{2}\left(\ln\frac{p_1}{q_1} - \ln\frac{p_K}{q_K}\right) - \frac{1}{2}\left(\frac{q_1}{p_1} - \frac{q_K}{p_K}\right), \frac{1}{2}\left(\ln\frac{p_2}{q_2} - \ln\frac{p_K}{q_K}\right) - \frac{1}{2}\left(\frac{q_2}{p_2} - \frac{q_K}{p_K}\right), \dots,\right.$$
$$\frac{1}{2}\left(\ln\frac{p_{K-1}}{q_{K-1}} - \ln\frac{p_K}{q_K}\right) - \frac{1}{2}\left(\frac{q_{K-1}}{p_{K-1}} - \frac{q_K}{p_K}\right), \frac{1}{2}\left(\ln\frac{q_1}{p_1} - \ln\frac{q_K}{p_K}\right) - \frac{1}{2}\left(\frac{p_1}{q_1} - \frac{p_K}{q_K}\right),$$
$$\left.\frac{1}{2}\left(\ln\frac{q_2}{p_2} - \ln\frac{q_K}{p_K}\right) - \frac{1}{2}\left(\frac{p_2}{q_2} - \frac{p_K}{q_K}\right), \frac{1}{2}\left(\ln\frac{q_{K-1}}{p_{K-1}} - \ln\frac{q_K}{p_K}\right) - \frac{1}{2}\left(\frac{p_{K-1}}{q_{K-1}} - \frac{p_K}{q_K}\right)\right)^{\tau}.$$

Then, based on Theorem 5.13 in Zhang (2017: 194), we can derive that

$$\frac{\sqrt{n}}{\hat{\sigma}_{s,n}}\left(\hat{S}_n - S\right) \xrightarrow{D} N(0,1), n \to \infty,$$

where

(16) $$\hat{\sigma}_{s,n} = [\mathrm{g}_s{}^{\tau}(\hat{\mathbf{v}}) \sum(\hat{\mathbf{v}}) \mathrm{g}_s(\hat{\mathbf{v}})]^{1/2}.$$

It is worth emphasizing that the standard deviation for symmetrized KLD $\hat{\sigma}_{s,n}$ is different from that of asymmetric KLD $\hat{\sigma}_n$.

Given that $\varepsilon > 0$, the asymptotic $(1 - \varepsilon)100\%$ confidence interval or the confidence interval with the $(1 - \varepsilon)$ confidence level for symmetrized KLD is as follows:

$$CI_{s,n} = \left(\hat{S}_n - \Phi_{1-\frac{\varepsilon}{2}} \frac{\hat{\sigma}_{s,n}}{\sqrt{n}};\ \hat{S}_n + \Phi_{1-\frac{\varepsilon}{2}} \frac{\hat{\sigma}_{s,n}}{\sqrt{n}}\right). \quad (17)$$

Hence, for the hypothesis test, we formulate the following statements. If $S_0 \in CI_{s,n}$, we accept the null hypothesis of no difference (H_0: $S = S_0$). Otherwise, if $S_0 \notin CI_{s,n}$, the alternative hypothesis should be chosen instead (H_1: $S \neq S_0$).

Further, when testing with the confidence probability of $1 - 0.05 = 0.95$, and, with $S_0 = 0$, we accept the null hypothesis of no difference between the studied distributions (H_0: $\mathbf{p} = \mathbf{q}$), if

$$0 \in \left(\hat{S}_n - 1.95996340\ \frac{\hat{\sigma}_{s,n}}{\sqrt{n}};\ \hat{S}_n + 1.95996340\ \frac{\hat{\sigma}_{s,n}}{\sqrt{n}}\right) = CI_{s,n}.$$

Otherwise, the alternative hypothesis which states the difference between the distributions is selected (H_1: $\mathbf{p} \neq \mathbf{q}$).

4.3.3 The Turing's perspective estimator

In the sections above, the mathematical details of hypothesis testing by means of the 'plug-in' estimator have been discussed. Yet, there is another statistical tool called the Turing's perspective estimator, proposed by Zhang and Grabchak (2014), which is believed to yield more precise results. Its procedure of hypothesis testing is similar to that of the 'plug-in' estimator. The major difference is that the coefficient of KLD is calculated with the following formula:

$$\begin{aligned} \widehat{D}^{\#} &= \widehat{D}^{\#}(\hat{\mathbf{p}}||\hat{\mathbf{q}}) \\ &= \sum_{k=1}^{K} \hat{p}_k \left[\sum_{v=1}^{m-Y_k} \frac{1}{v} \prod_{j=1}^{v}\left(1 - \frac{Y_k}{m-j+1}\right) - \sum_{v=1}^{n-X_k} \frac{1}{v} \prod_{j=1}^{v}\left(1 - \frac{X_k - 1}{n-j}\right)\right], \end{aligned} \quad (18)$$

where for each fixed $v = 1, \dots, m - Y_k$,

$$\frac{1}{v} \prod_{j=1}^{v}\left(1 - \frac{Y_k}{m-j+1}\right) = \frac{1}{v}\left(1 - \frac{Y_k}{m}\right)\left(1 - \frac{Y_k}{m-1}\right)\left(1 - \frac{Y_k}{m-2}\right) \dots \left(1 - \frac{Y_k}{m-v+1}\right)$$

and

$$\frac{1}{v}\prod_{j=1}^{v}\left(1-\frac{X_k-1}{n-j}\right)=\frac{1}{v}\left(1-\frac{X_k-1}{n-1}\right)\left(1-\frac{X_k-1}{n-2}\right)\ldots\left(1-\frac{X_k-1}{n-j}\right).$$

The Turing's perspective estimator is consistent and asymptotically normal (Zhang 2017: 183). The main advantage of the Turing's perspective estimator is that it has an exponentially fast decaying bias in the sample of sizes n and m, as compared to the 'plug-in' estimator whose bias tends to zero as a power function:

(19) $$\frac{\sqrt{n}}{\hat{\sigma}_{\widehat{D}^{\#}}}\left(\widehat{D}^{\#}-D\right)\xrightarrow{D}(N(0,1).$$

Further, when data contains zeros, an augmentation should be added to the estimator (19), as shown below.

In formula (19), the standard deviation is calculated as follows:

$$\sigma^2{}_{\widehat{D}^{\#}}=\mathrm{g}^{\tau}(\hat{\mathbf{v}}^{*})\sum(\hat{\mathbf{v}}^{*})\,\mathrm{g}(\hat{\mathbf{v}}^{*}),$$

where

$$\hat{\mathbf{v}}^{*}=(\hat{p}_1,\ldots,\hat{p}_{K-1},\hat{q}_1^{*},\ldots,\hat{q}_{K-1}^{*})^{\tau},$$

(20) $$\hat{q}_K^{*}=\hat{q}_K+\frac{1[Y_K=0]}{m},K=1,\cdots,K$$
$$=\begin{cases}\hat{q}_K\ \text{if},Y_K\neq 0\\ \dfrac{1}{m}\ \text{if}\ Y_K=0\end{cases}$$

and g^{τ} and $\sum(\hat{\mathbf{v}})$ are given in formula (9) and (10).

Then, similarly to the 'plug-in' estimator, the confidence intervals based on the Turing's perspective is as follows:

$$CI_{\widehat{D}^{\#}}=\left(\widehat{D}^{\#}-\Phi_{1-\frac{\varepsilon}{2}}\frac{\hat{\sigma}_{\widehat{D}^{\#}}}{\sqrt{n}};\ \widehat{D}^{\#}+\Phi_{1-\frac{\varepsilon}{2}}\frac{\hat{\sigma}_{\widehat{D}^{\#}}}{\sqrt{n}}\right).$$

The procedure of hypotheses testing with the confidence intervals is the same as described in the previous two sections. For the confidence probability of $1 - 0.05 = 0.95$, and, with $\widehat{D}^{\#}{}_0 = 0$, the null hypothesis of no difference is true (H_0: $\mathbf{p} = \mathbf{q}$), if

$$0 \in \left(\widehat{D}^{\#} - 1.95996340\ \frac{\hat{\sigma}_{\widehat{D}^{\#}}}{\sqrt{n}};\ \widehat{D}^{\#} + 1.95996340\ \frac{\hat{\sigma}_{\widehat{D}^{\#}}}{\sqrt{n}}\right) = CI_{\widehat{D}^{\#}}. \quad (21)$$

If zero is not included in the confidence interval, the alternative hypothesis is selected (H_1: $\mathbf{p} \neq \mathbf{q}$).

4.4 Cluster techniques

To identify the clusters of suffixes with similar properties, the following cluster techniques have been used in this study: agglomerative clustering, k-medoids and Principal Component Analysis (PCA). These methods complement each other, allowing for the construal of a general picture from the data. The methods of agglomerative clustering and k-medoids have revealed the homogeneous clusters of the suffixes. On the other hand, PCA has unveiled the interactions between the studied parameters.

In general, all cluster techniques are designed such that they reduce the number of dimensions in data. They are based on arithmetic, geometric, graph-theoretic and statistical (minimizing within-group variance) models which detect similarity and dissimilarity in the studied data set (Legendre & Legendre 2012: 341). Most of them (e.g. PCA, t-distributed Stochastic Neighbor Embedding, correspondence analysis and hierarchical cluster analysis) are regarded as exploratory, as opposed to predictive and explanatory (Desagulier 2017: 239). This is because they do not require assumptions as to the possible underlying groupings, and are therefore also termed 'unsupervised' (Baayen 2008: 118). They are used for the generation of hypotheses (Desagulier 2017: 239), as is demonstrated in this research.

More specifically, agglomerative clustering is a bottom-up hierarchical cluster analysis: its algorithm first defines separate clusters, which then are merged into a compound cluster. By this token, the grouping of clusters continues until one giant cluster emerges. This clustering tree is known as a dendogram, and its height signifies the order in which clusters have been arranged. Hence, it is possible to choose different cutting points for a cluster solution (Salkind & Neil 2006: 235).

In contrast, the method of k-medoids (Hastie et al. 2001), also termed as Partitioning Around Medoids (PAM), belongs to the group of non-hierarchical cluster analyses. Its optimization technique is based on minimizing a sum of pairwise dissimilarity values of data points (Lai & Fu 2011: 765). K-medoids seems to be more robust to noise and outliers, for example to the similar clustering method of k-means (Alok & Bradford 2019, Ch. 1). It is also believed to form more homogeneous clusters (Oliveira et al. 2020: 9).

Finally, PCA finds important tendencies in the data and expresses them "in the form of a handful of new orthogonal variables called principal components" (Desagulier 2017: 243). Within the context of the current research, PCA detects, firstly, the Euclidean distance between the studied derivational processes and, secondly, correlations between the variables discussed in Section 3.6. The former is displayed via a distance biplot of data points scattered across the first two principal components, and the latter via a biplot of a correlation circle with the arrows of vectors, each of which represents a variable. The length of a vector indicates the significance of a variable in explaining the variance observed in principal components (Ter Braak 1994: 135): the longer the vector of a variable, the more important it is for the PCA model and the more influence it exercises on the data points. Finally, the angles between vectors point to the correlation between variables: an acute angle describes a positive correlation between variables, a right angle no correlation, and an angle close to 180 a perfect negative correlation (Legendre & Legendre 2012: 441).

There are several requirements for the application of PCA. First, the variables should have reasonable symmetrical distribution (Baayen 2008: 125) and should be numerical, because the algorithm of PCA is based on Pearson's correlation coefficient. Secondly, the variables need to be standardized, i.e. centered and scaled (Desagulier 2017: 245). Thirdly, the number of observation n in the PCA analysis should not be smaller than or equal to the number of variables p, because, in this case, the eigen-decomposition of a full-rank dispersion matrix $\boldsymbol{S}$ produces $(n - 1)$ real and $[p - (n - 1)]$ null eigenvalues (Legendre & Legendre 2012: 450). Therefore, in order to make the data of this study appropriate for the PCA analysis, as well as to avoid zeroes and negative values, it was log-transformed with the following formula: $log(x - (min(x) - 1))$.

The above-described cluster methods were applied in *R* (R Core Team 2021). The hierarchical cluster analysis was performed with the package 'shipunov' (Shipunov et al. 2020), which allows for the selection of an appropriate method for the plotting of a dendogram, as well

as for the bootstrap of the identified clusters. For the PAM method, the package 'cluster' (Maechler et al. 2019) was used. Finally, PCA is available in the package 'FactoMineR' (Husson et al. 2008).

4.5 Graph theory

Another useful framework for the exploration of the data is the Graph Theory. Many real world situations can be described by means of a diagram consisting of "a set of points together with lines joining certain pairs of these points" (Bondy & Murty 1976: 1). Within a linguistic context, the frequency of a particular unit or type can be associated with the importance of the node, and the co-occurrence frequency of nodes corresponds to the number of edges between nodes (Desagulier 2017: 275). By this token, the studied phenomenon emerges as a network with nodes and edges of different sizes.

Two packages in *R* allow for creating various directed and undirected graphs: 'network' (Butts 2015), 'igraph' (Csárdi & Nepusz 2006). In this thesis, the latter package was used to visualize the strength and the distance between the studied suffixes and different morphological classes in formal paradigms. Specifically, with the codes written by Desagulier (2018: 285-287), different graphs were built where all morphemes of the English word-formation system are perceived as a morphological network with nodes ('vertices'), representing morphemes, and with lines ('edges'), conveying the distance and connection between them (Figure 4.1). These graphs are based on the measurement of eigenvector centrality which assigns a higher weight to nodes connected to important nodes. The colour map of the graphs illustrates the score of centrality which "is a measure of how important the node is in the context of the entire graph" (Desagulier 2018: 286). Consequently, big red circles at the center of the graphs, symbolizing high scores, can be thought of as morphological hubs, whereas yellow and white circles represent nodes with lower scores of centrality. Further, the size of the circle stands for the type frequency of a morpheme.

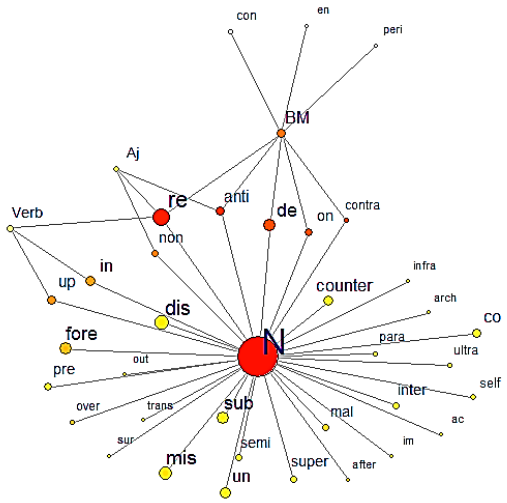


Figure 4.1. A preview of Figure 6.54 as an example of a graph of a formal meta-construction {{a-C}}

The morphological network visualized in this graph is an optimized exemplification of English word-formation processes. In addition to their theoretical significance, they have practical applications in English language teaching and in developing software for morphological parsing.

4.6 Conclusions

This chapter has explained various statistical methods used in this study and the motivation for their usage. It has first looked at the method of identifying relations between variables: i.e. correlation. Then, Poisson regression has been introduced, followed by the description of three estimators of relative entropy. Finally, the chapter has engaged in the discussion of cluster and graph network techniques. Different angles of the morphological metacorpus of this study have been explored with these methods in Chapter 7.

5 The structural analysis of the morphological metacorpus

In this and the following chapters, the morphological metacorpus obtained with the help of the formal morphological analysis is presented. More specifically, this chapter introduces a general picture of English word formation with different quantitative and qualitative characteristics (RQ1). It also identifies morphological patterns and constructions in the studied sample (RQ2) and addresses the problem of level structure of word formation across different word classes. Some effects of high type frequency are established (RQ3).

Similar to any other formal model, the model of word formation presented in this study is an approximation of the reality for a number of reasons. First, many word-formation processes do not have clear boundaries which makes categorization challenging and indirect, whereas formalization implies a strict and succinct definition of categories. Further, the model considers the first written record of words, which gives approximate information about their emergence: only a limited number of words have been 'born' and introduced at once, while most words have complicated and ambiguous origin. Another difficult conundrum has been the fluid nature of the word classes in English. Lastly, some words have two or more routes of origination that have ultimately converged into their present form.

To overcome these formalistic challenges, first, fine-grained categories have been introduced to the discussed word-formation model that captured nuances of a word's origin and its conversion properties, as described in the OED. Specifically, for this purpose, the tool of prime symbols for distinguishing native from non-native word formation has proved useful: it has allowed us to trace the origin of morphemes without compromising their early non-native stages of development. Secondly, the introduction of such parameters as the year of the first record, the target language(s) and the band frequency of words have refined the rigid formal picture that emerged from the application of the formal morphological analysis. Altogether, these formal techniques and adjustments have revealed interesting quantitative and qualitative characteristics of the modern English word-formation system, as well as relations between its constituting elements, which are introduced in detail and visualized in this chapter. This model, in combination with other sources, can then be used to suggest generalizations for present-day English.

Hence, in what follows, Section 5.1 introduces the overall composition of the morphological metacorpus from two perspectives: (i) the makeup of the metacorpus's word classes

and (ii) its etymological organization. Section 5.2, then, zooms into simplexes of the metacorpus and devotes a subsection to each distinguished word class, including conversive and grammatical classes. Section 5.3, also organized by word classes, is dedicated to multimorphemic words and their morphological and structural properties. Finally, in Section 5.4, the main findings of this chapter are summarized.

5.1 The overall composition of the morphological metacorpus

This section offers an overall picture of the organization of the morphological metacorpus. In subsection 5.1.1, the general morphological structure of the sample is discussed, whereas subsection 5.1.2 presents the general etymological information of the metacorpus.

5.1.1 The overall morphological structure of the metacorpus

Simple morphological classes, which are formed without the involvement of word-formation processes, make up the largest portion of the metacorpus (Figure 5.1). Multimorphemic words constitute slightly more than one third of all words in the sample.

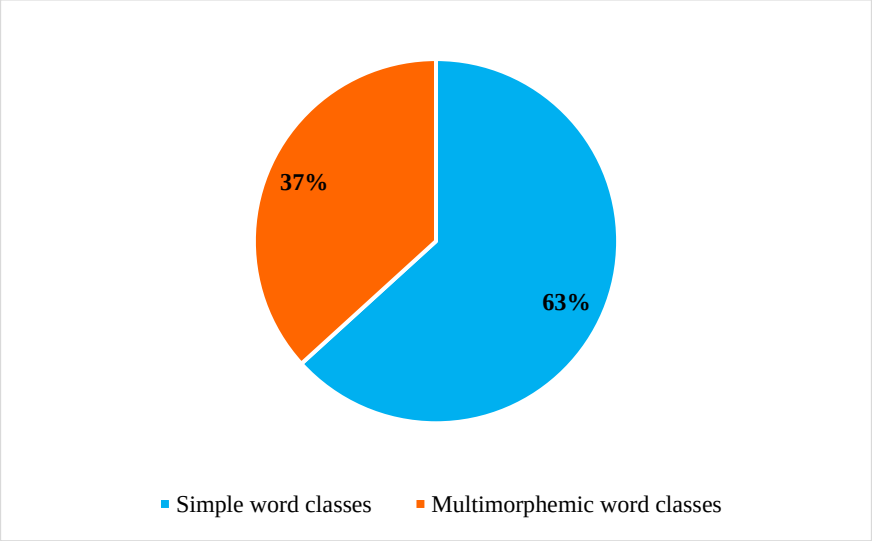
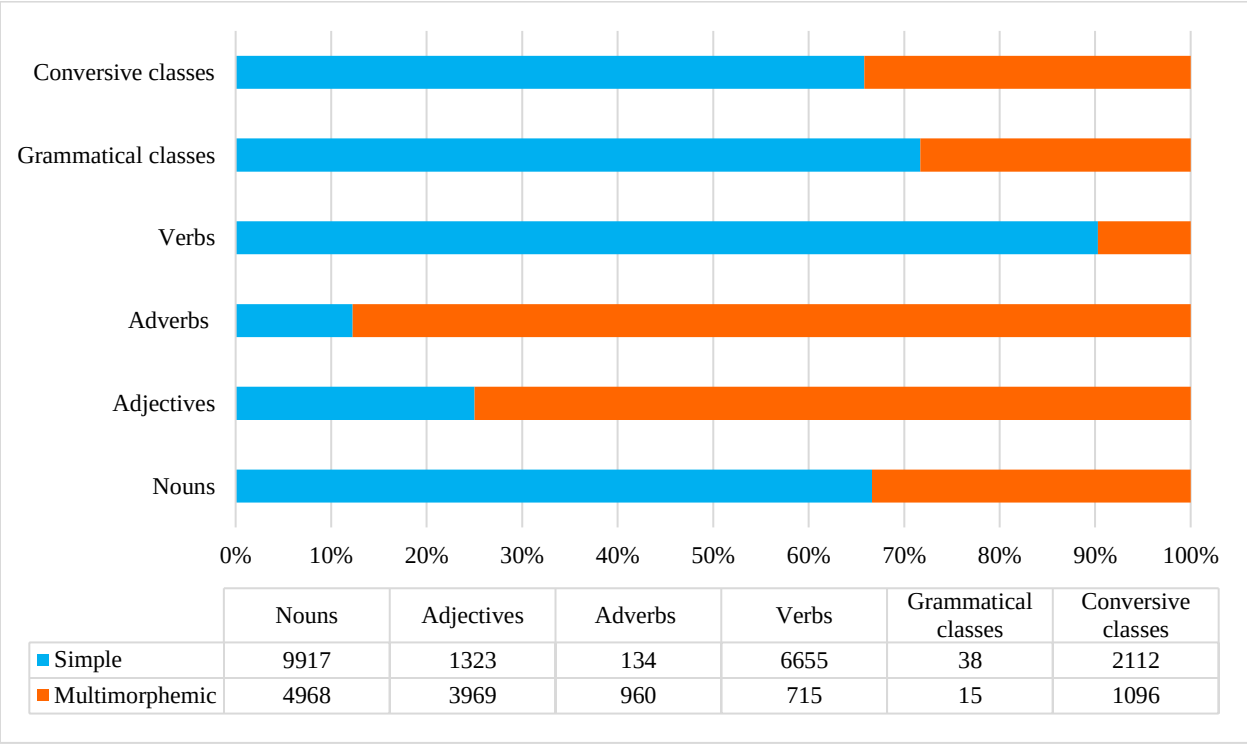


Figure 5.1. Multimorphemic vs simple morphological classes

Although, in general, morphological simplexes prevail in the morphological metacorpus, the proportion of simplex and multimorphemic words is different for each class. Figure 5.2 illustrates trends in each category of words. Simplexes are dominant for verbs and for nouns with only 10% and 33% accounting for multimorphemic words respectively. Simplexes also prevail in grammatical and conversive classes. By contrast, multimorphemic words are more common for

adjectives and adverbs, with the proportion of simplexes constituting 25% and 12% respectively. From these correlations, it can generally be inferred that adjectives and adverbs draw on derivation more heavily than other morphological classes.



	Nouns	Adjectives	Adverbs	Verbs	Grammatical classes	Conversive classes
Simple	9917	1323	134	6655	38	2112
Multimorphemic	4968	3969	960	715	15	1096

Figure 5.2. The proportions of multimorphemic vs simple words in word classes

Another important aspect of the metacorpus is the ratio of word classes within the categories of simple and multimorphemic words (Figure 5.3 and Figure 5.4). Nouns are the dominant word class for both simplex and multimorphemic words (49% and 42% respectively), whereas verbs have the second largest share only for simplexes. Morphologically complex verbs constitute nearly 7% of all multimorphemic words. Further, adjectives form the second largest portion for multimorphemic words (34%), which is significantly smaller for simplexes (7%). Similarly, adverbs are more frequent in the category of multimorphemic words (8%) and less frequent in the category of simplexes (1%). Finally, the share of conversive and grammatical classes is almost the same for simplexes and multimorphemic words (10% and 9% for conversive classes respectively, and under 1% for grammatical classes in both categories). The overall composition of word classes in the morphological metacorpus is presented in Figure 5.5.

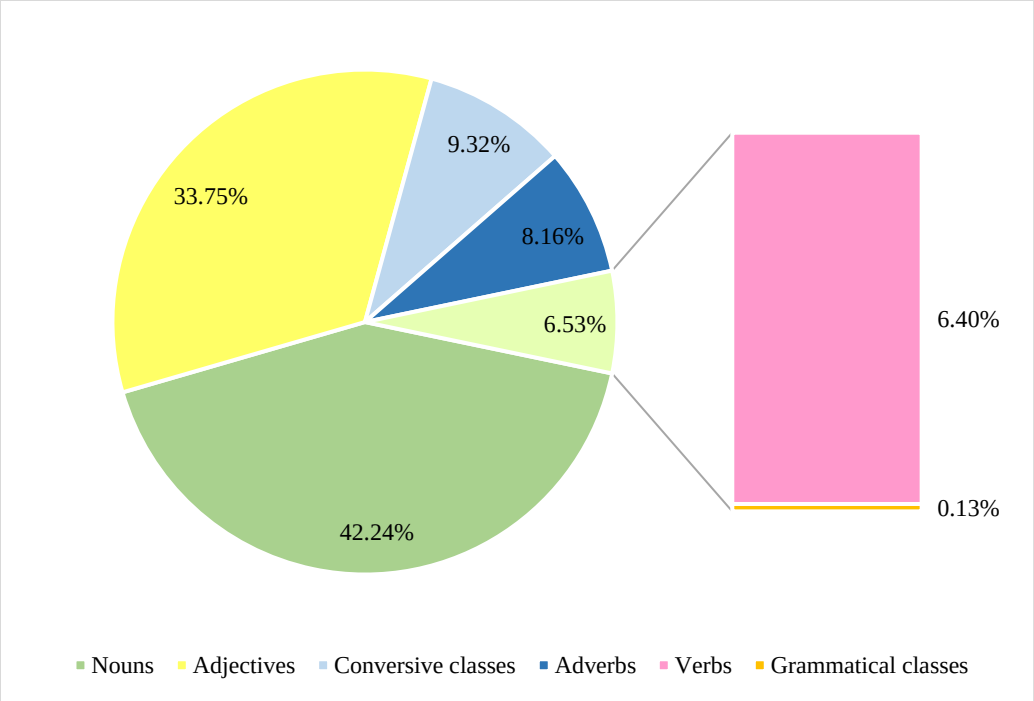


Figure 5.3.The proportions of word classes in the category of multimorphemic words

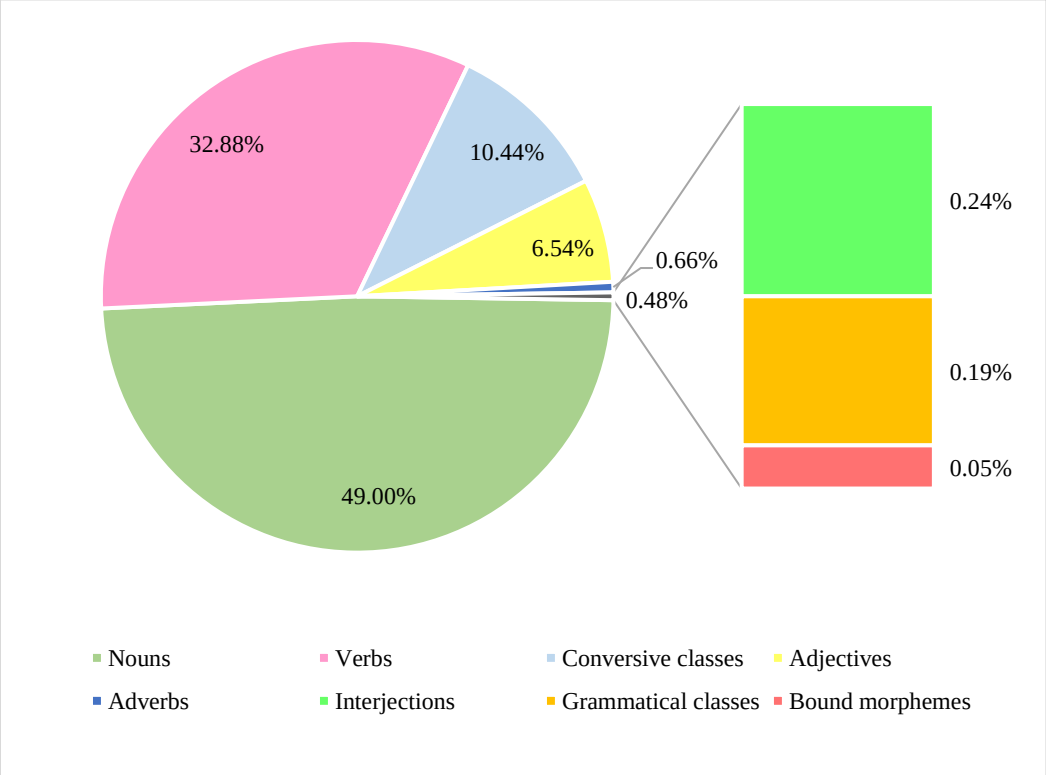


Figure 5.4. The proportions of word classes in the category of simplexes

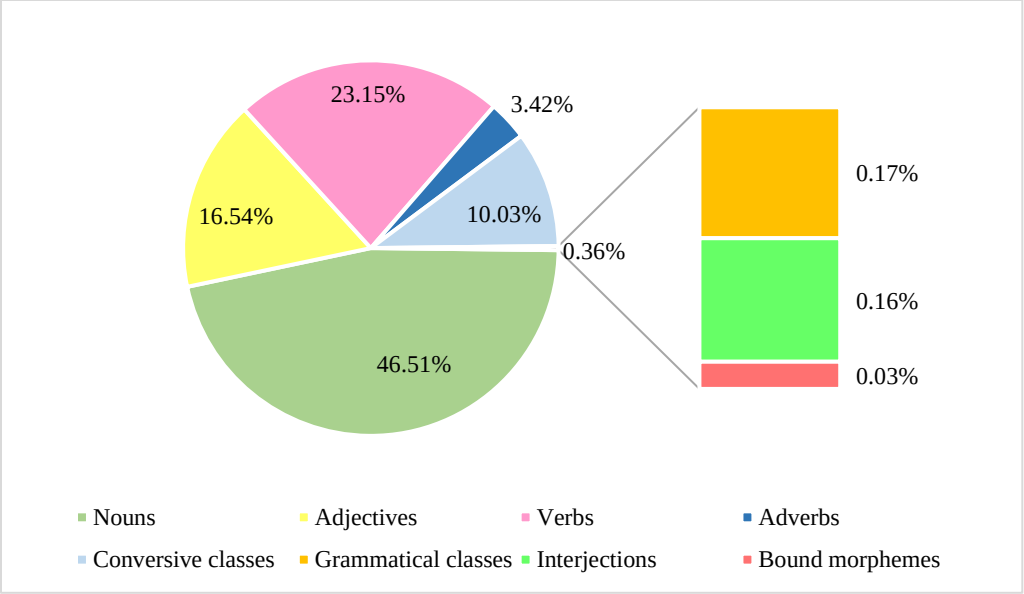


Figure 5.5. The overall proportions of word classes in the sample

5.1.2 The overall etymological structure of the metacorpus

Figure 5.6 sheds light on the etymology of all loan words in the sample. French borrowings make up the largest share of simplexes in the metacorpus, followed by Latin borrowings, by words inherited from ancient Germanic, by parallel borrowings from Latin and French and by borrowings from other languages. Words borrowed from Scandinavian form 2% of all loan words. The detailed account of loan words in English is provided in the corresponding subsections of Section 5.2, as well as in Appendices A, B and C.

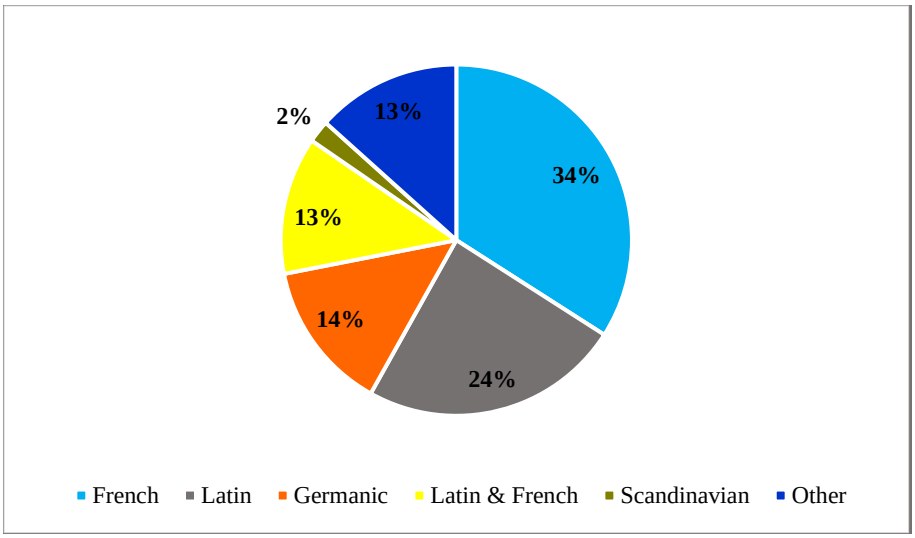


Figure 5.6. The overall picture of the origins of simplexes in the metacorpus

5.2 Simplexes

This section presents the structural analysis of simplexes which make up a zero-level of the discussed word-formation model and which constitute 63% of the sample. Sections 5.2.1 to 5.2.5 look at simple nouns, verbs, adjectives, adverbs, and simple interjections, respectively. Section 5.2.6 focuses on simple grammatical word classes, i.e. prepositions, conjunctions and pronouns. Further, Section 5.2.7 analyses conversive morphological classes such as N/Aj, N/Ad, N/Verb, N/Aj/Ad, and Num/Aj/Ad. Lastly, Section 5.2.8 offers a general discussion of the identified quantitative and qualitative morphological trends.

Since the structural analysis given below involves categories on different levels, different types of pie chart have been used to facilitate their reading. Specifically, to represent the shares of word-formation processes in a word class, a large 3D exploded pie chart has been chosen. Further, a 2D colourful pie chart and a pie-of-pie chart illustrate proportions of borrowings from different languages, whereas a wide donut chart with shades of one colour shows a further fine-grained distinction within the categories of conversion, phonological formation, contraction, semantic formation and onomatopoeia for nouns (in green), verbs (in blue), adjectives (in orange) and adverbs (yellow). Finally, a narrow donut chart depicts categories on the lowest level of the analysis.

5.2.1 Simple nouns

Figure 5.7 presents an overall picture of the origin of simple nouns. They comprise nearly 31% of all words in the sample. Half of the simple nouns are loan words, which reflects the rich history of interaction of English with other languages and which has had ‘far-reaching repercussions’ for English phonology and morphology (Kastovsky 2006: 202)—for example, the noun-verb stress alternation (Sherman 1975) or the fact that many of English present-day productive derivational morphemes are of foreign origin. Then, the second largest share of simple nouns are formed by conversion. Germanic component (words inherited from ancient Germanic and words borrowed from Germanic languages) constitute the third largest share. A small share is made of words formed by phonological alternations of other lexical units and contractions of original forms. Words coined on the native ground constitute 4% of all simple nouns, and onomatopoeic and semantic formations account for a minute portion of all simple nouns. Lastly, 5% are of an uncertain origin. The following subsections describe these categories in more detail.

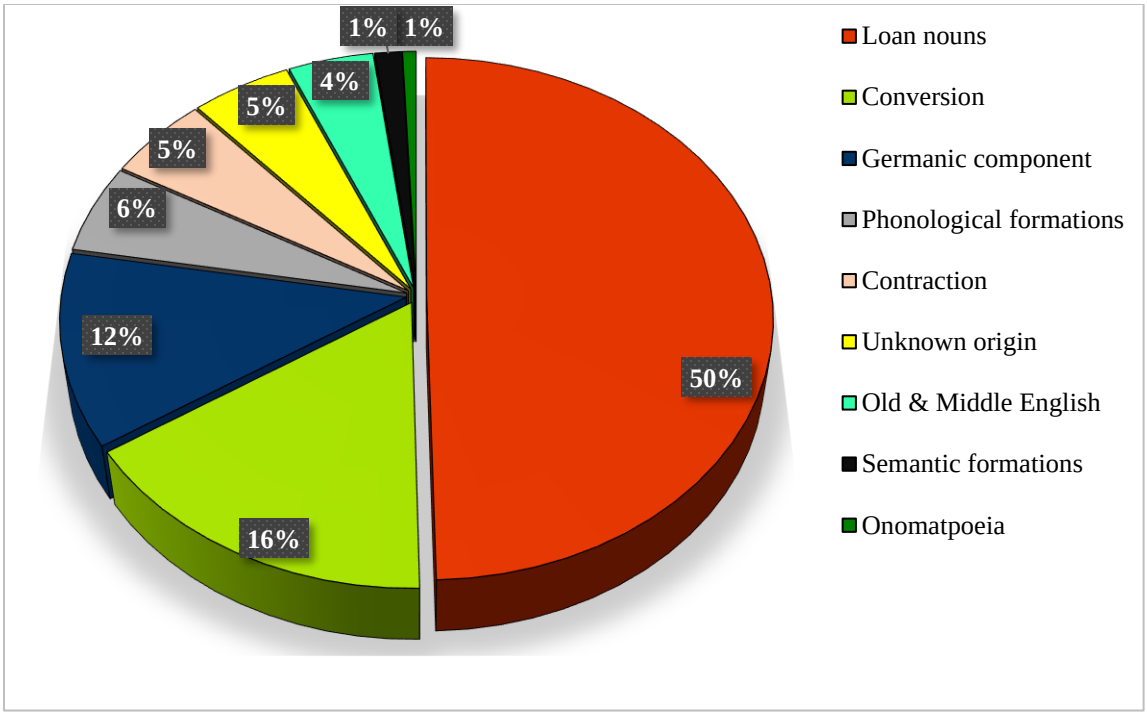


Figure 5.7. The origin of English simple nouns

5.2.1.1 Loan nouns

The English lexicon is impressively diverse and contains nouns borrowed or inherited from different language families. However, as shown in Figure 5.8, it is mainly made up by borrowings from Romance, Latin and Germanic languages. A small share of English simple nouns are of Greek origin. Only 5% were borrowed from other languages: this portion accounts for etymological diversity the most.

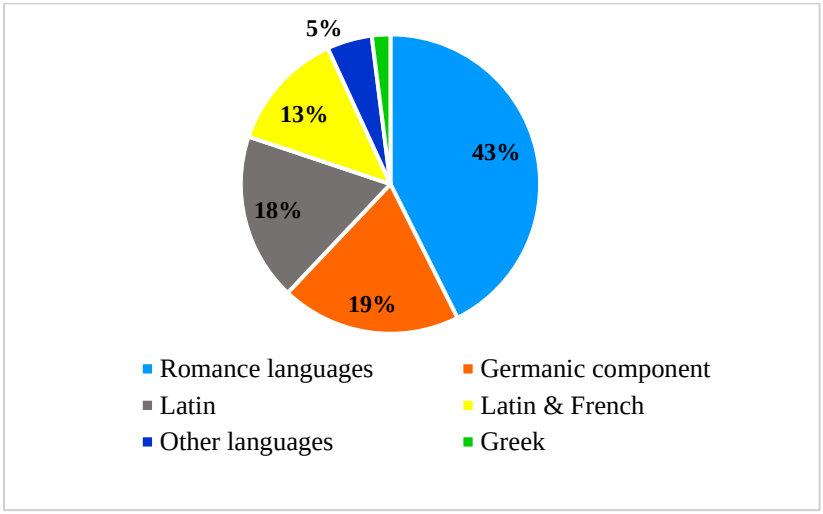


Figure 5.8. The general picture of loan nouns

This etymological diversity is illustrated in Figure 5.9. The highest number of nouns in the category of "Other languages" came to English from Indo-Aryan, Celtic, Semitic and Japonic languages. A small number of nouns were borrowed from other language families. A detailed account of frequencies listed by languages is given in Table 1 (Appendix A).

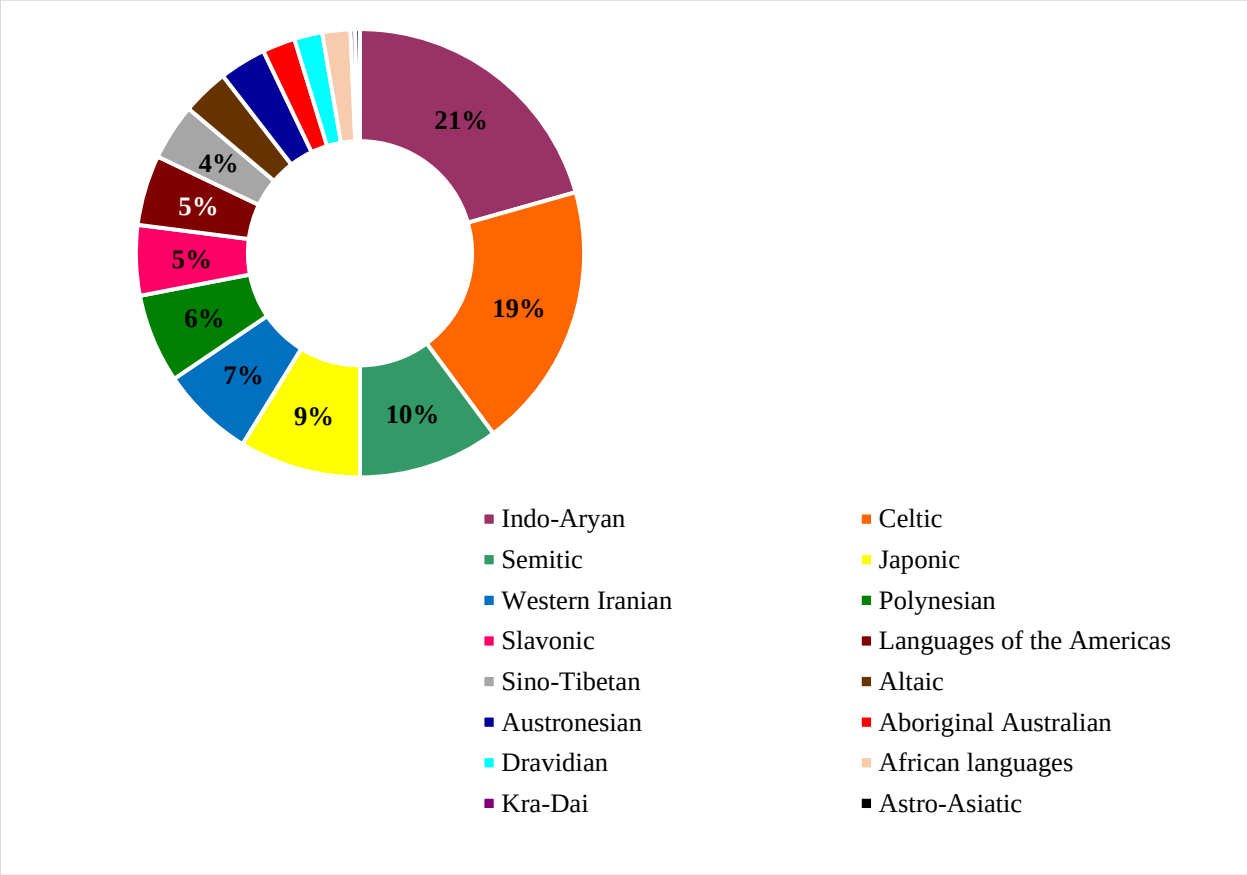


Figure 5.9. Borrowings from other languages
grouped by language families and geographic locations

Romance languages

Figure 5.10 illustrates the distribution of lexic shares among Romance languages. The largest portion in this category constitutes borrowings from French. Further, small, almost equal shares belong to Anglo-Norman, Italian and Spanish. A tiny portion is made up of loan nouns from Portuguese. In the subcategory of dual route, illustrated in the pie-of-pie[22] in Figure 5.10, borrowings from Italian and other routes have the highest portion, followed by the category of

[22] A pie-of-pie is a type of chart which shows proportions of a studied phenomenon in two circular graphs. The advantage of this graph is that it gives a separate pie chart for proportions with small percentages, which facilitates the reading of the small shares of a pie chart.

Spanish and other parallel routes, Anglo-Norman and other routes, and Portuguese and other routes. For a chart that describes dual routes with the involvement of Romance languages in detail, see Appendix A (Figure 1).

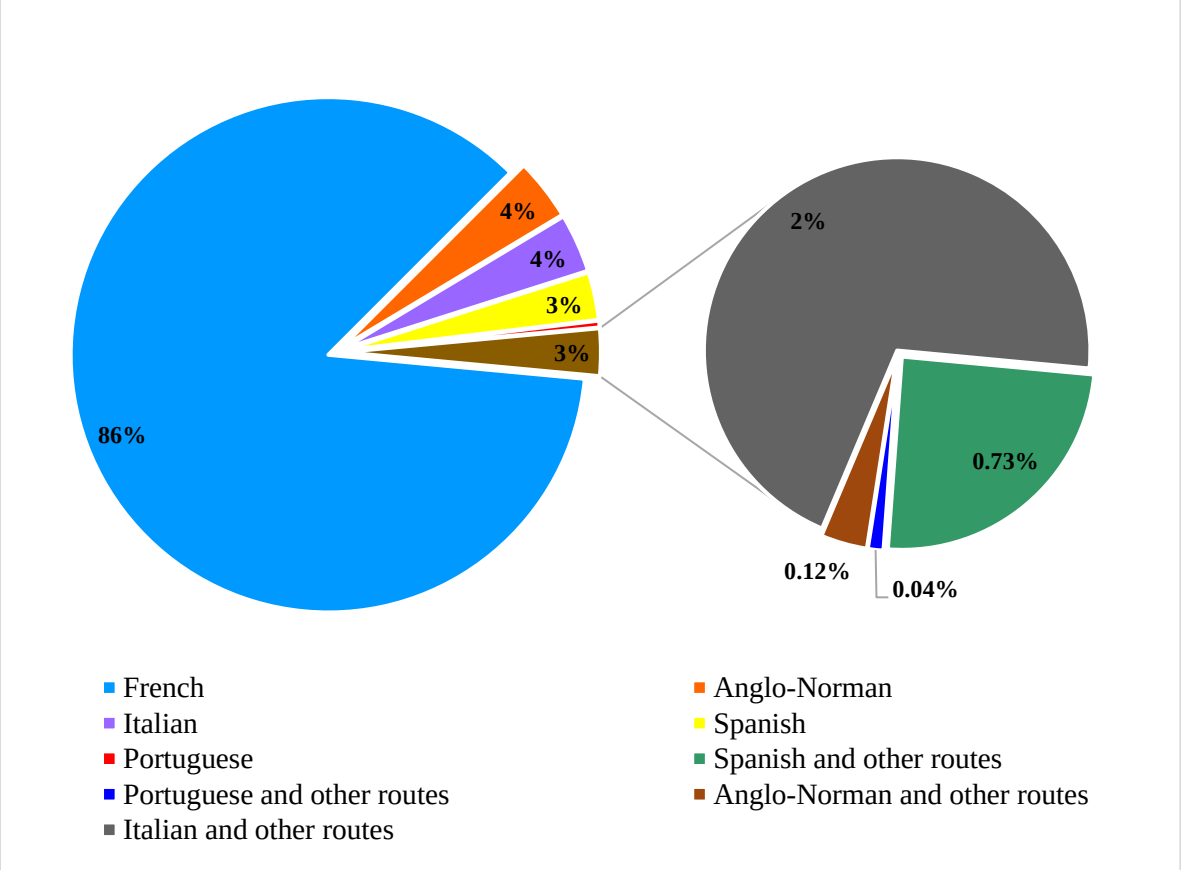


Figure 5.10. Borrowings from Romance languages

French borrowings: zero-level morphological analysis

As mentioned above, French borrowings constitute the largest portion of simple nouns. Although within the framework of this thesis, these words are considered simplexes, they also contain important information about the basic morphological material that in the later stages of the development of English became native. Table 5.1 shows zero-level morphological patterns, their type frequency and percentage shares. The most frequent morphemes among French borrowings in the sample that have been reanalyzed as native are the following suffixes (in descending order of frequency): *-ion*, *-ity*, *-y*, *-ment*, *-er*, *-ance*, *-age*, *-ence*, *-or*, *-ine*, and *-ure*. Morphological hapaxes (patterns that produce only one word) are not given in the table.[23]

[23] These patterns represent French derivation.

Table 5.1. Morphemes represented in French borrowings: a zero morphological level
(Key: Morphological patterns are given with one lexical example and in the order of descending type frequency)

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	N-fr *mullusc*	1377	61.67	18	N-fr=BM+ty *deputy*	14	0.63
2	N-fr=BM+ion *compassion*	171	7.66	19	N-fr=BM+ism *optimism*	14	0.63
3	N-fr=BM+ity *gratuity*	81	3.63	20	N-fr=BM+al *official*	13	0.58
4	N-fr=BM+y *harmony*	81	3.63	21	N-fr=in+BM *influx*	11	0.5
5	N-fr=BM+ment *agreement*	65	2.91	22	N-fr=BM+ry *sanctuary*	9	0.4
6	N-fr=BM+er *jeweller*	54	2.42	23	N-fr=BM+ate *frigate*	9	0.4
7	N-fr=BM+ance *maintenance*	45	2.02	24	N-fr=BM+ic *fabric*	8	0.36
8	N-fr=BM+age *mirage*	42	1.88	25	N-fr=BM+ess *duchess*	9	0.4
9	N-fr=BM+ence *influence*	34	1.52	26	N-fr=dis+BM *disloyalty*	7	0.31
10	N-fr=BM+or *creator*	26	1.16	28	N-fr=BM+ist *integrist*	5	0.22
11	N-fr=BM+ine *medicine*	26	1.16	29	N-fr=BM+ble *constable*	5	0.22
12	N-fr=BM+ure *enclosure*	26	1.16	30	N-fr=BM+BM *typography*	5	0.22
13	N-fr=BM+ant *contestant*	19	0.85	31	N-fr=BM+ian *musician*	4	0.18
14	N-fr=BM+ade *brigade*	17	0.76	32	N-fr=mis+BM *misnomer*	3	0.13
15	N-fr=BM+ery *embroidery*	17	0.76	33	N-fr=BM+et *pamphlet*	3	0.13
16	N-fr=BM+ette *etiquette*	15	0.67	34	N-fr=BM+ee *debauchee*	3	0.13
17	N-fr=N+ice *hospice*	15	0.67				

Latin borrowings: zero-level morphological analysis

In Table 5.2, Latin zero-level morphemes, their frequency and percentage shares are listed. The most frequent of them include the suffixes *-ion*, *-or*. *-ity*, *-y*, and *-um*.

Table 5.2. Morphological patterns represented in Latin borrowings

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	N-lat *abyss*	464	42.11	16	N-lat=BM+sis *analysis*	9	0.82
2	N-lat=BM+ion *magnification*	231	20.96	17	N-lat=BM+on *lexicon*	8	0.73
3	N-lat=BM+or *dedicator*	80	7.26	18	N-lat=BM+al *herbal*	8	0.73
4	N-lat=BM+ity *credibility*	46	4.17	19	N-lat=BM+osis *neurosis*	7	0.64
5	N-lat=BM+y *galaxy*	36	3.27	20	N-lat=in+BM *injury*	5	0.45
6	N-lat=BM+um *gymnasium*	36	3.27	21	N-lat=BM+acy *conspiracy*	5	0.45
7	N-lat=BM+ia *regalia*	24	2.18	22	N-lat=BM+ent *continent*	5	0.45
8	N-lat=BM+ary *summary*	20	1.81	23	N-lat=BM+ism *stoicism*	4	0.36
9	N-lat=BM+ence *prominence*	16	1.45	24	N-lat=BM+ic *panic*	4	0.36
10	N-lat=BM+ency *frequency*	16	1.45	25	N-lat=BM+ance *significance*	3	0.27
11	N-lat=BM+ment *augment*	14	1.27	26	N-lat=BM+ant *elephant*	3	0.27
12	N-lat=BM+ate *estimate*	14	1.27	27	N-lat=BM+ancy *infancy*	3	0.27
13	N-lat=BM+ure *gesture*	14	1.27	28	N-lat=BM+an *librarian*	2	0.18
14	N-lat=BM+er *meander*	13	1.18	29	N-lat=BM+ist *jubilist*	2	0.18
15	N-lat=BM+ory *inventory*	10	0.91				

Latin-French parallel borrowings: zero-level morphological analysis

In the zero morphological level of Latin-French borrowings, the morphological patterns presented in Table 5.3 can be distinguished. The most frequent of them include suffixes *-ion*, *-ity*, *-y*, *-or*, and *-ence*. The table also gives one lexical example for each pattern, as well as its type frequency and percentage share in the category.

Table 5.3. Morphemes represented in Latin-French borrowings

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	N-lat/fr *anthem*	307	38.81	11	N-lat/fr=BM+ance *fragrance*	7	0.88
2	N-lat/fr=BM+ion *administration*	261	33	12	N-lat/fr=BM+ory *oratory*	7	0.88
3	N-lat/fr=BM+ity *nobility*	47	5.94	13	N-lat/fr=BM+er *minister*	5	0.63
4	N-lat/fr=BM+y *irony*	48	6.07	14	N-lat/fr=BM+ive *motive*	4	0.51
5	N-lat/fr=BM+or *orator*	27	3.41	15	N-lat/fr=BM+ist *artist*	3	0.38
6	N-lat/fr=BM+ence *opulence*	26	3.29	16	N-lat/fr=BM+age *foliage*	3	0.38
7	N-lat/fr=BM+ure *pressure*	12	1.52	17	N-lat/fr=BM+ar *scholar*	3	0.38
8	N-lat/fr=BM+ty *safety*	10	1.26	18	N-lat/fr=BM+ent *president*	2	0.25
9	N-lat/fr=BM+ment *ornament*	8	1.01	19	N-lat/fr=BM+ism *baptism*	2	0.25
10	N-lat/fr=BM+ate *certificate*	8	1.01				

Anglo-Norman borrowings: zero-level morphological analysis

Among Anglo-Norman borrowings, as shown in Table 5.4, the most frequent suffixes on the zero-level morphological analysis are *-er* and *-or*.

Table 5.4. Morphemes represented in Anglo-Norman borrowings

No	Morphological pattern	Example	Type frequency	%
1	N-fr/en	*apostle*	62	61.39
2	N-fr/en=BM+er	*commissioner*	12	11.88
3	N-en/fr=BM+or	*director*	11	10.89
4	N-fr/en=BM+ty	*treaty*	5	4.95
5	N-fr/en=BM+ion	*exception*	5	4.95
6	N-fr/en=BM+age	*voyage*	2	1.98
7	N-fr/en=BM+ment	*battlement*	2	1.98
8	N-fr/en=BM+ity	*brevity*	2	1.98

The Germanic component of English nouns

The Germanic component constitutes 19% of all noun borrowings in the metacorpus. In this share, 61% of nouns were borrowed into English from ancient Germanic languages (Figure 5.11), and 12% and 8% were borrowed from Scandinavian and Dutch, respectively. Borrowings from other Germanic languages constitute the remaining part of this share. A detailed account of the dual routes with the involvement of Germanic languages is given in Table 2 of Appendix A.

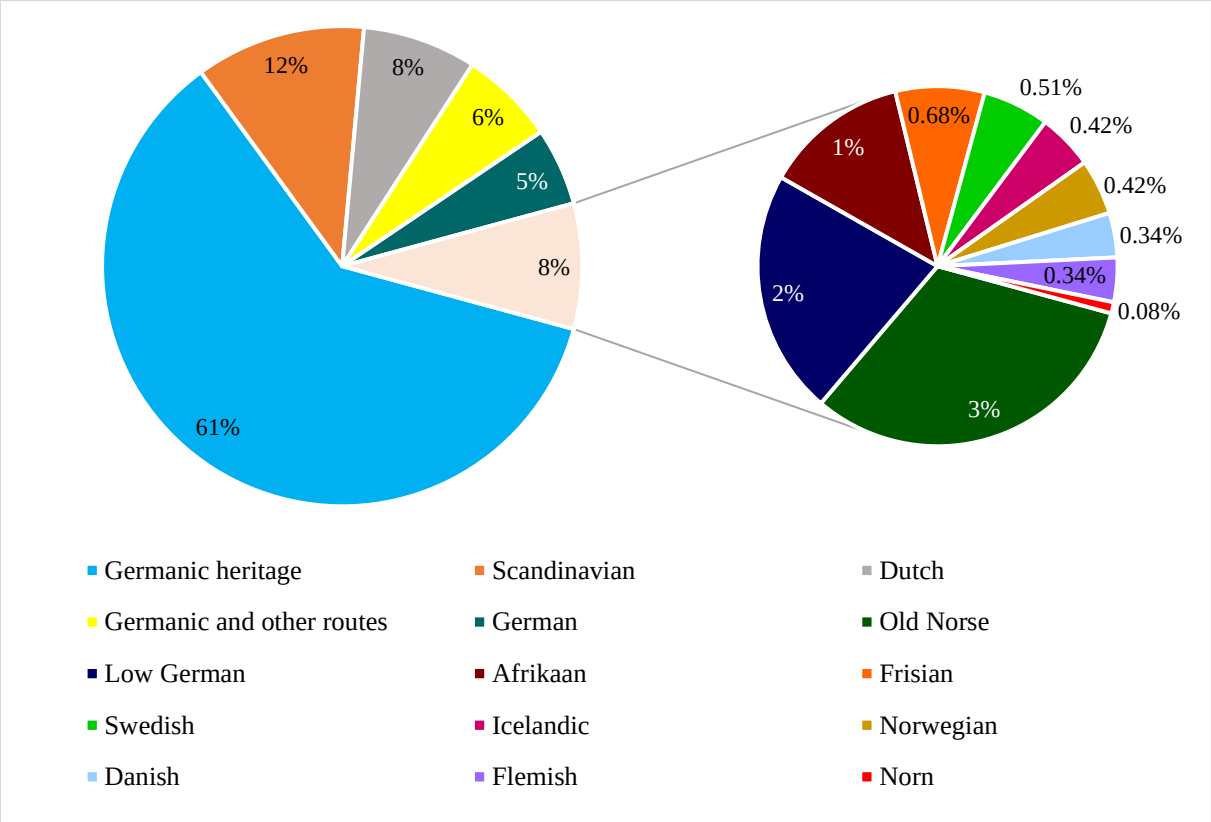


Figure 5.11. The Germanic component in nouns of the sample

Germanic words: zero-level morphological analysis

Table 5.5 offers a morphological account of Germanic morphology on the zero-level of word formation, captured by the metacorpus. As can be seen from the table, most of the Germanic nouns are monomorphemic. The suffixes *-er*, *-ness*, *-ing*, *-dom*, *-ship*, *-th*, the prefix *mis-*, as well plural forms of nouns and compounding, represent Germanic morphology in present-day English.

Table 5.5. Morphemes inherited from Germanic, represented in the sample

No	Morphological pattern	Example	Type frequency	%
1	N-grm	*beard*	705	97.65
2	N-grm=BM+er	*miller*	4	0.55
3	N-grm=BM+ness	*highness*	3	0.42
4	N-pl/grm	*lice*	3	0.42
5	N-grm=BM+BM	*nostril*	2	0.28
6	N-grm=mis+N	*misdeed*	1	0.14
7	N-grm=BM+ing	*opening*	1	0.14
8	N-grm=BM+dom	*freedom*	1	0.14
9	N-grm=BM+ship	*friendship*	1	0.14
10	N-grm=BM+th	*warmth*	1	0.14

5.2.1.2 Conversion

Conversion is the second most productive word-formation process for simple nouns (15%). As illustrated in Figure 5.12, 95% of all nouns in this category are formed by pure conversion and the rest of them by dual routes with the involvement of conversion (see Table 3, Appendix A). Dual routes mean that a word does not have a single origin but has been formed by the convergence of two or more processes. For example, the word *defame* has two origin routes: it was borrowed from French as a noun and it was parallelly formed by conversion from the earlier verb form *defame*, (also borrowed from French).

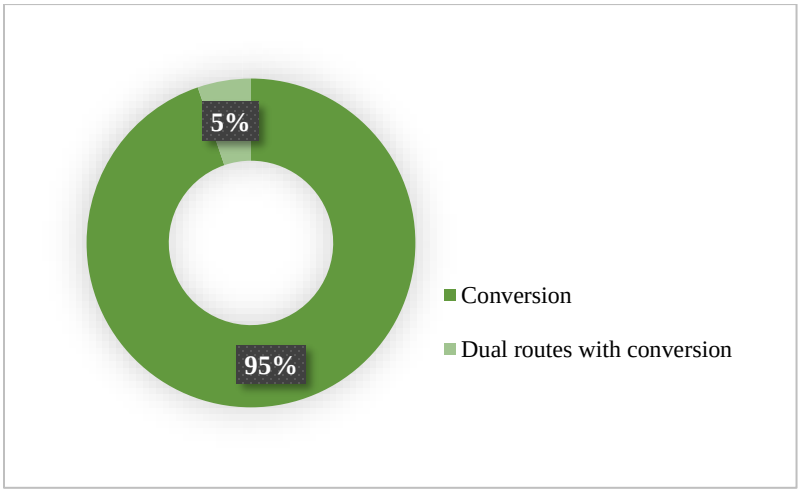


Figure 5.12. Shares of words formed by conversion

5.2.1.3 Phonological formations

There are around 7% of simple nouns which are formed by phonological changes of original forms. Figure 5.13 shows that most nouns in this category are formed by various kinds of phonological alternations of original forms of words. A small share is accounted for by aphetic forms, dialect variants and dual routes with the involvement of phonological changes. The dual routes are presented in Figure 2 of Appendix A.

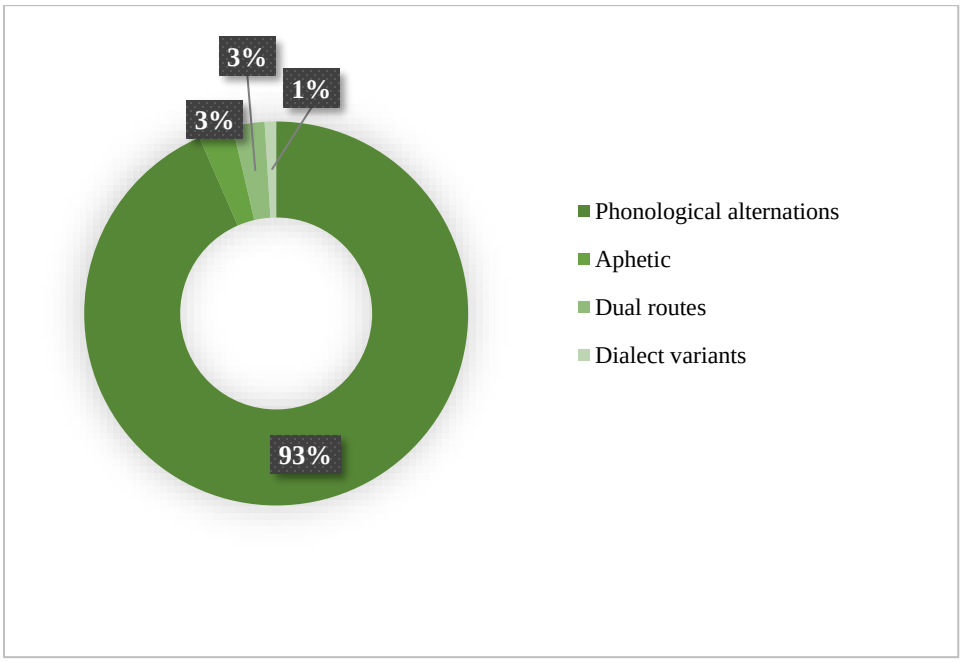


Figure 5.13. Simple nouns formed by phonological changes in original forms

5.2.1.4 Old and Middle English formations: zero-level morphological analysis

This 5% share of the sample represents the 'common core of the language' (Quirk et al. 1985: 16). A zero-level Old and Middle English morphology, which is still present in contemporary English, is shown in Table 5.6. The most frequent complex morphological pattern in this category is compounding. Other morphological processes involve the suffixes *-le*, *-ing*, *-s*, *-er*, *-ness*, *-ock*, *-th*, *-ship*, as well as plural forms of nouns and conversion from past participle.

Table 5.6. Morphemes represented in Old and Middle English

No.	Morphological pattern	Example	Type frequency	%
1	N	*bee*	380	88.17
2	N=BM+BM	*cranberry*	21	4.87
3	N=BM+le	*whistle*	6	1.39
4	N=BM+ing	*willing*	5	1.16
5	N-pl=N:(f→ve)+s	*elves*	4	0.93
6	N=BM+er	*shipper*	4	0.93
7	N-pl	*dice*	3	0.7
8	N=BM+ness	*witness*	3	0.7
9	N=BM+ock	*yolk*	2	0.46
10	N=BM+th	*wealth*	1	0.23
11	N=BM+ship	*worship*	1	0.23
12	N=Verb*3	*lent*	1	0.23

5.2.1.5 Contractions

Another 5% of simple nouns in the sample are formed by contraction of original forms. Figure 5.14 illustrates the trends in this category. The largest share of contractions is formed by shortening. A small portion belongs to back-formations and dual routes. The dual routes with the involvement of conversion are analyzed in Figure 3 in Appendix A.

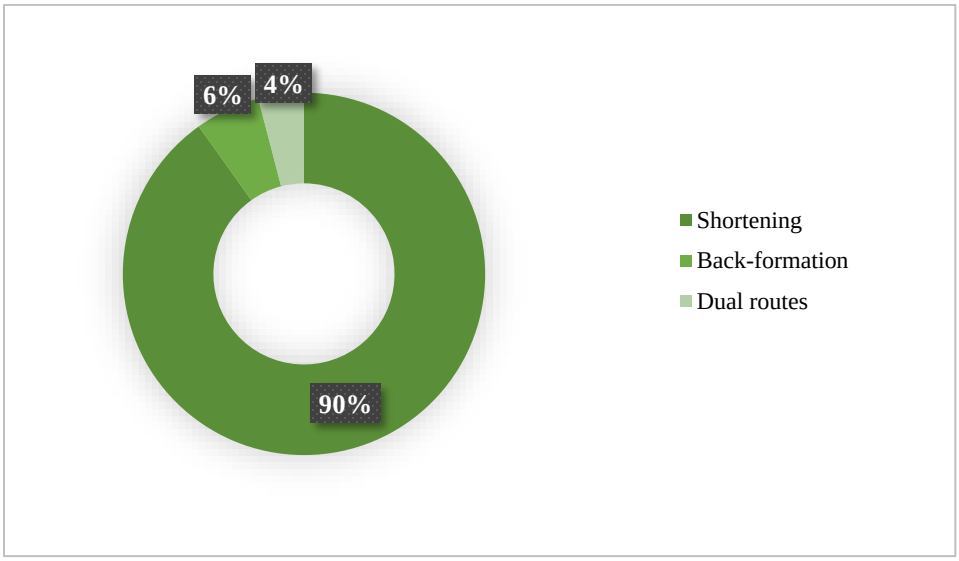


Figure 5.14. The proportion of nouns formed by contraction

5.2.1.6 Semantic formations

Above 1% of simple nouns are formed by the split of meaning[24] in original forms. Figure 5.15 illustrates that the largest formation processes in this category are semantic splits from nouns and from proper names. Only a minute number of nouns involve dual routes.

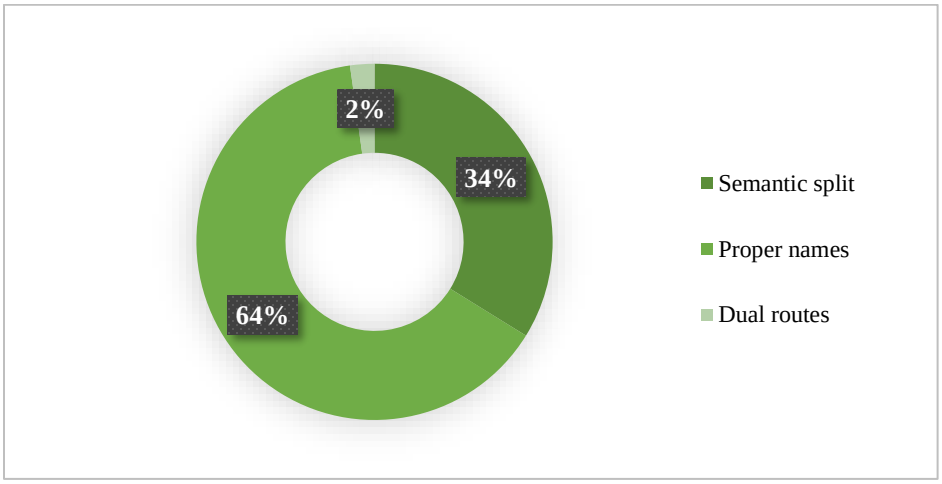


Figure 5.15. Semantic noun formations

5.2.1.7 Onomatopoeia

Onomatopoeic formations constitute below 1% of all simple nouns (Figure 5.16). Expressive and imitative formations make up the largest share in this category. A small portion is formed by dual routes with the involvement of onomatopoeia (Figure 4, Appendix A).

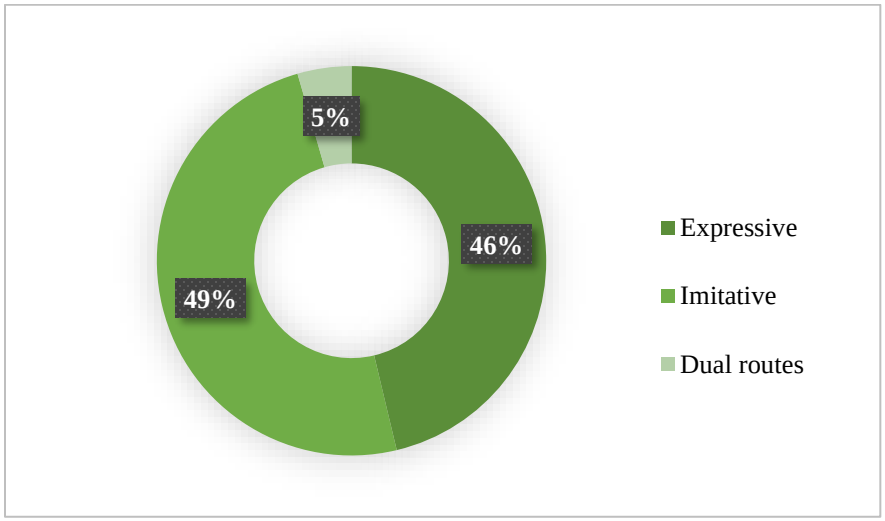


Figure 5.16. Onomatopoeic noun formations

[24] The split of meaning means that a word is formed semantically by diverging from the original meaning.

5.2.2 Simple verbs

Simple verbs constitute 20.66% of the sample. Figure 5.17 depicts a general picture of simple verbs' formations. The largest part of simple verbs has been formed by conversion and by processes of borrowing from other languages. The remaining quarter share belongs to simple verbs originated in Old and Middle English, as well as words formed by phonological alternations, by back-formations/contractions of original forms, by onomatopoeia and by semantic processes. The origin of 4% of simple verbs is unknown. In the following sections, I will look at each of these categories one by one.

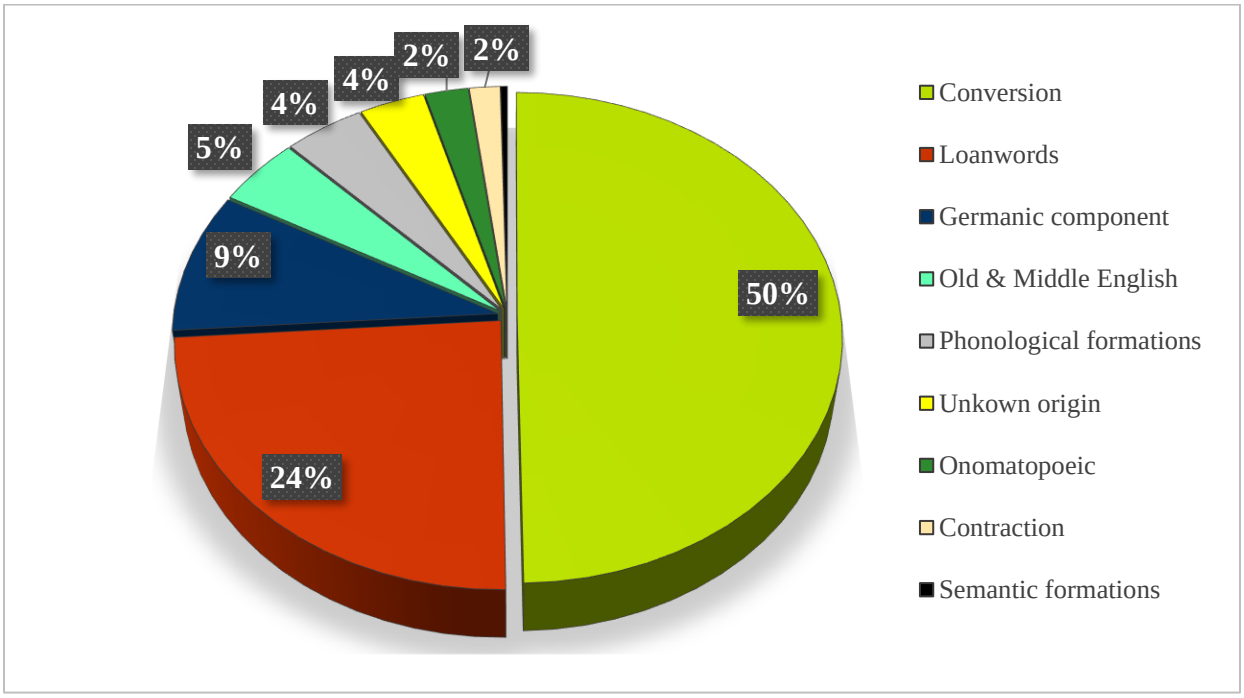


Figure 5.17. The origins of English simple verbs

5.2.2.1 Conversion

As mentioned above, the most frequent simple verb-formation process is conversion (50% of all simple verbs). Figure 5.18 demonstrates that only 4% of all simple verbs formed by conversion have originated from the convergence of several parallel forms of words, whereas 96% are pure conversions. The dual routes are given in Figure 5 of Appendix B.

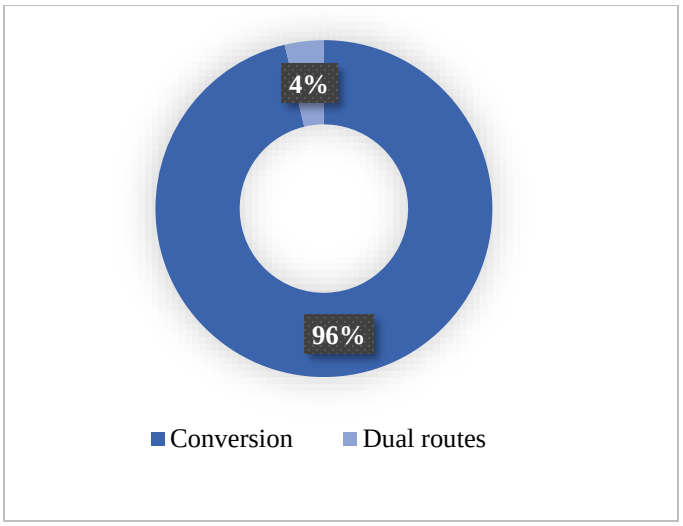


Figure 5.18. Simple verbs formed by conversion

5.2.2.2 Loan verbs

Loan verbs have the second largest share in the category of simple verbs (Figure 5.19). The highest number of verbs have been borrowed from French and/or Latin. The next largest share of simple verbs have been inherited from Germanic. Finally, a small proportion of verbs have been adopted from such languages as Scandinavian, Dutch, Anglo-Norman, Low German, Scots, Old Norse, German, Italian, Spanish, etc. (see Figure 5.20) and by the convergence of parallel borrowings (see Figure 6, Appendix B).

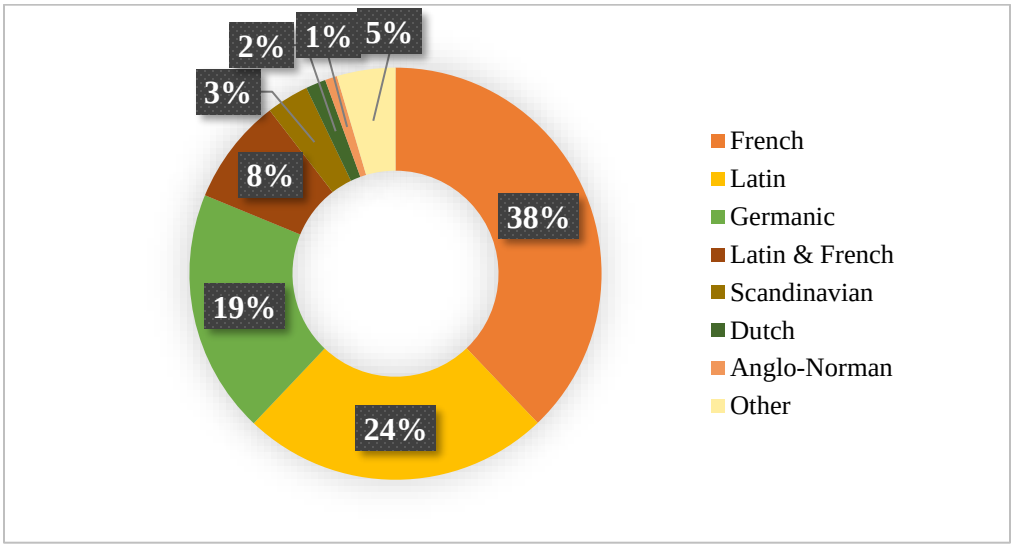


Figure 5.19. Loan verbs

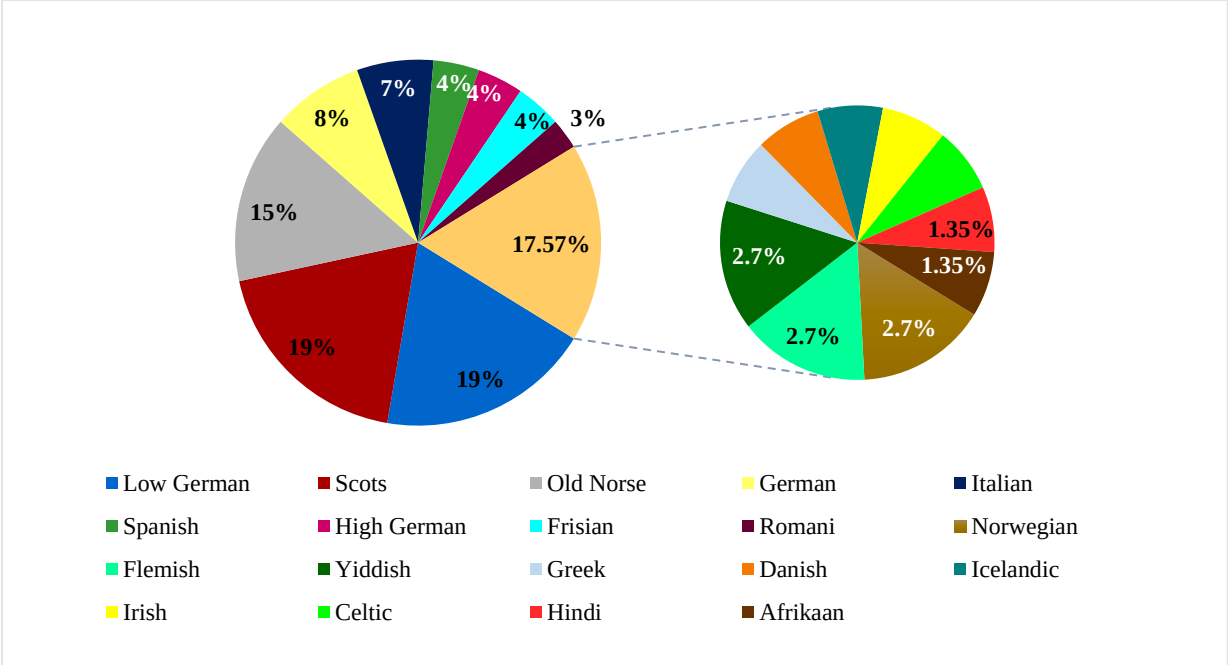


Figure 5.20. Loan verbs: other languages

French borrowings: zero-level morphological analysis

As informed by Table 5.7, the most frequent verb-forming morphemes borrowed from French are the prefixes *re-*, *de-*, *dis-*, *en-* and the suffixes *-ish*, *-ify* and *-ize*.

Table 5.7. Morphemes represented in French borrowings

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	Verb-fr *esteem*	618	73.48	7	Verb-fr=BM+ize *recognize*	15	1.78
2	Verb-fr=re+BM *recover*	51	6.06	8	Verb-fr=en+BM *enrich*	14	1.66
3	Verb-fr=de+BM *degrade*	41	4.88	9	Verb-fr=BM+ise *disguise*	12	1.43
4	Verb-fr=dis+BM *discover*	32	3.8	10	Verb-fr=in+BM *inherit*	7	0.83
5	Verb-fr=BM+ish *establish*	25	2.97	11	Verb-fr=BM+ate *evaluate*	5	0.59
6	Verb-fr=BM+ify *purify*	18	2.14	12	Verb-fr=BM+fy *defy*	3	0.36

Latin borrowings: zero-level morphological analysis

The most frequent morphemes in verbs borrowed from Latin include the suffix *-ate* and the prefixes *in-* and *de-* (Table 5.8).

Table 5.8. Morphemes represented in Latin verb borrowings

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	Verb-lat=BM+ate *donate*	232	43.2	7	Verb-lat=sub+BM *subside*	5	0.93
2	Verb-lat *ascend*	221	41.15	8	Verb-lat=BM+ify *testify*	4	0.74
3	Verb-lat=in+BM *inhibit*	27	5.03	9	Verb-lat=pre+BM *prescribe*	4	0.74
4	Verb-lat=de+BM *deduce*	24	4.47	10	Verb-lat=trans+BM *transcend*	3	0.56
5	Verb-lat=dis+BM *dissect*	7	1.3	11	Verb-lat=BM+ize *canonize*	2	0.37
6	Verb-lat=re+BM *repress*	6	1.12	12	Verb-lat=ir+BM *irrigate*	2	0.37

Latin & French borrowings: zero-level morphological analysis

Among Latin and French verb borrowings, the most frequent morpheme is the prefix *re-* (Table 5.9).

Table 5.9. Morphemes represented in Latin and French parallel borrowings

No.	Morphological pattern	Type frequency	%
1	Verb-lat/fr *adhere*	124	67.03
2	Verb-lat/fr=re+BM *reform*	42	22.7
3	Verb-lat/fr=BM+ify *magnify*	11	5.95
4	Verb-lat/fr=BM+ate *collaborate*	4	2.16
5	Verb-lat/fr=BM+ize *baptize*	4	2.16

5.2.2.3 Onomatopoeic verbs

Onomatopoeic verbs form 2.4% of all simple verbs in the sample. As shown in Figure 5.21, they comprise mostly imitative and expressive formations. The dual routes are presented in Figure 7 in Appendix B.

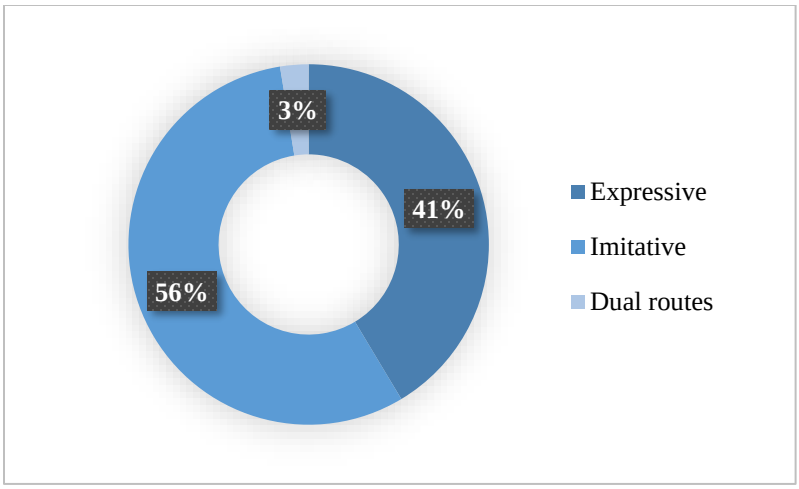


Figure 5.21. Onomatopoeic verb formations

5.2.2.4 Phonological formations

Phonological formations make up 0.44% of all simple verbs in the sample. The largest portion of verbs in this category are formed by phonological alternations—usually, by the alternation of one or two phonemes (Figure 5.22).

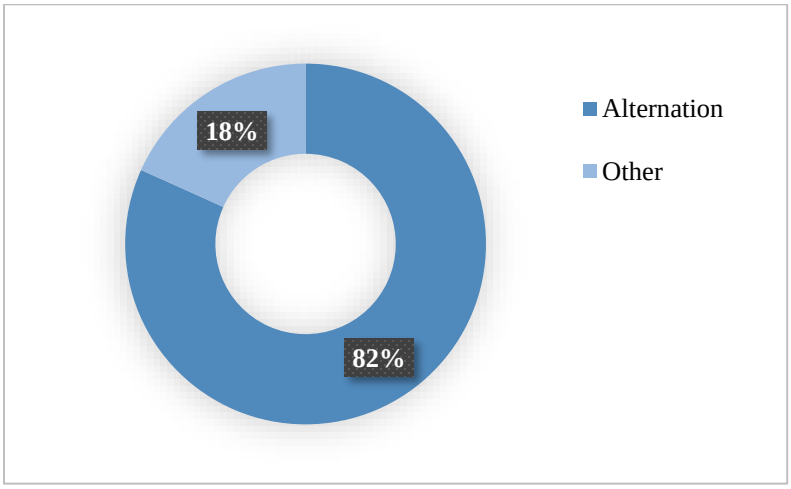


Figure 5.22. Phonological verb formations: the overall picture

Furthermore, as informed by Figure 5.23, verbs formed by the loss of an initial vowel have the largest share in the category of 'Other phonological formations'. Corrupted forms of verbs and their frequentative and dialect variants constitute a smaller portion. The percentages of dual routes are given in Figure 8 of Appendix B.

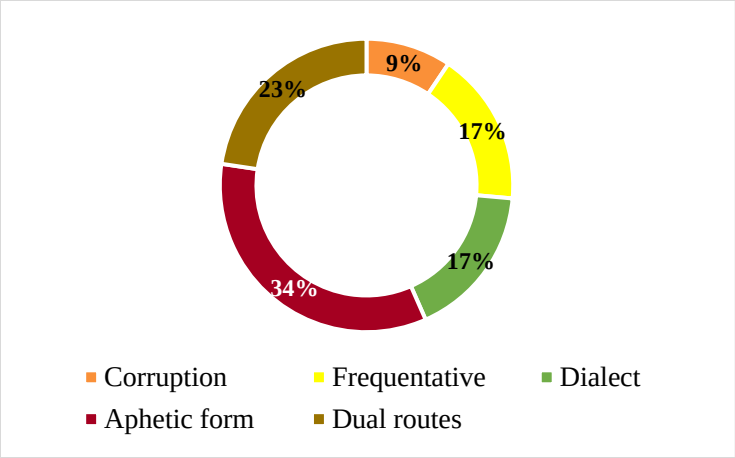


Figure 5.23. Other phonological formations
(the order of annotation in the charts is sequential from left to right)

5.2.2.5 Semantic formations

The smallest proportion of simple verbs have been formed by semantic processes—0.4% of all simple verbs in the sample. The largest share is made up by the split of a verb's sense. Simple verbs in the remaining share have developed their sense from proper names (Figure 5.24).

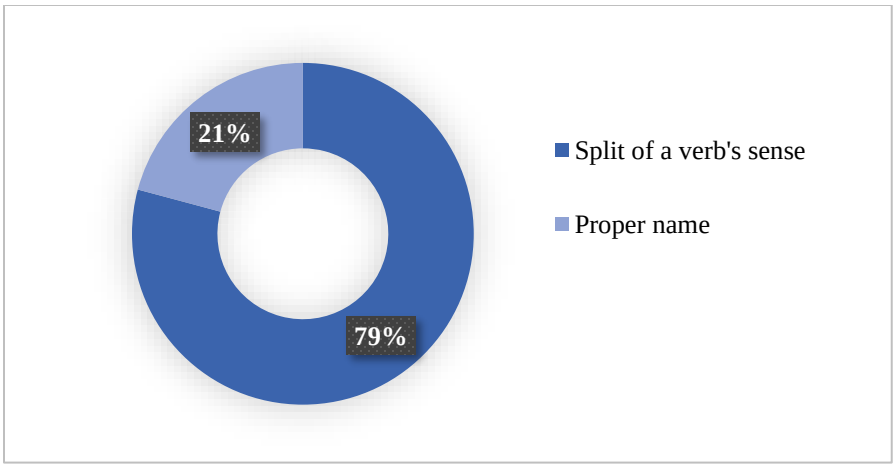


Figure 5.24. Semantic verb formations

5.2.2.6 Back-formations/Contractions

Back-formations, contractions, or dual routes with the involvement of these processes constitute 0.17% of all simple verbs (Figure 5.25). The dual routes are illustrated in Figure 9 of Appendix B.

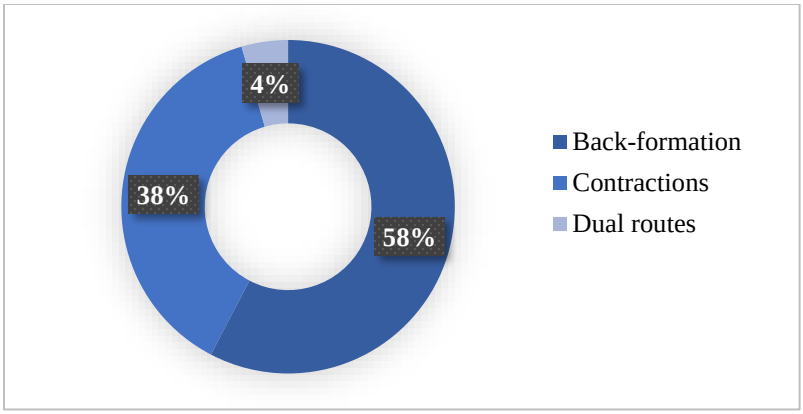


Figure 5.25. Contractions

5.2.3 Simple adjectives

Simple adjectives make up 4.1% of the sample. As informed by Figure 5.26, the largest share in this category is comprised of loan adjectives. Adjectives formed during the period of Old and Middle English constitute almost one seventh of all adjectives in the sample. The third largest share belongs to adjectives formed by conversion. Further, a small portion of adjectives are accounted for by phonological alternations and contraction. The origin of 3% of adjectives is unknown. Finally, just a few simple adjectives are formed by onomatopoeia.

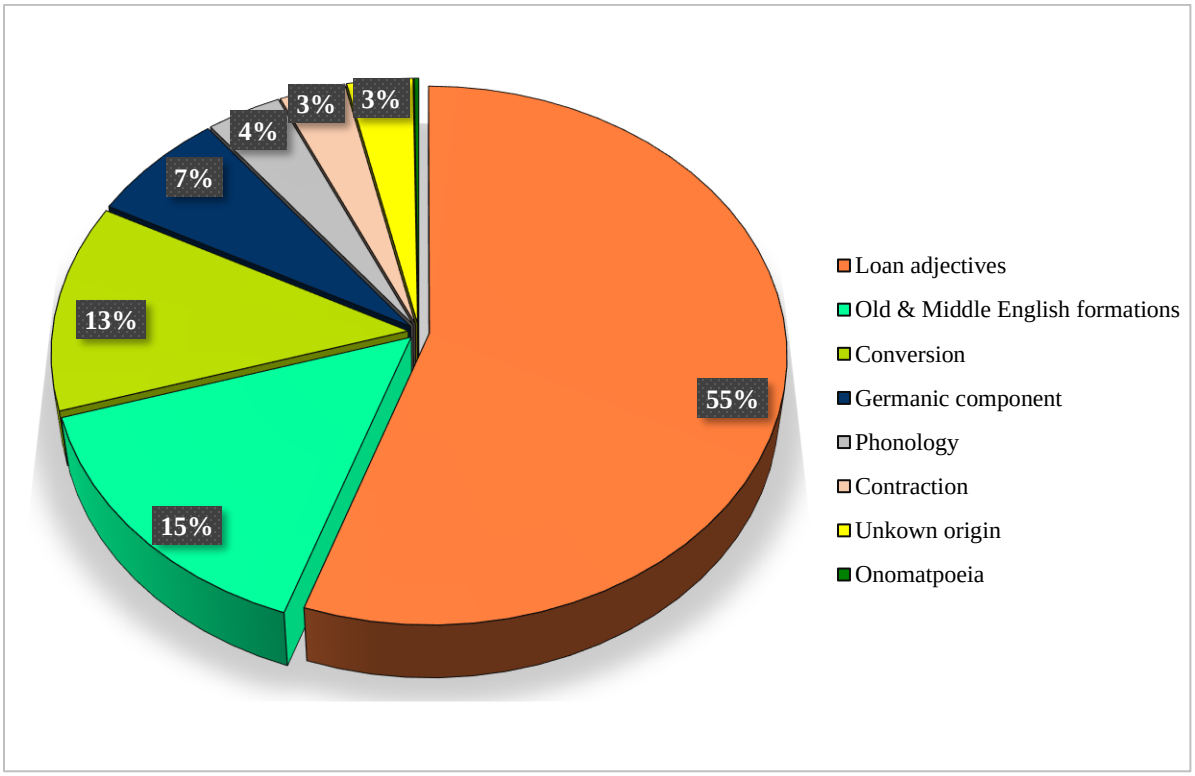


Figure 5.26. The origin of English simple adjectives

5.2.3.1 Loan adjectives

The general picture of loan adjectives is more homogeneous, as compared to that of loan nouns. Near 87% of all English loan adjectives are borrowings from Latin and/or French. Above one-tenth of all loan adjectives are formed by the Germanic component. A small share is composed of borrowings from Greek. Last of all, less than 1% of adjectives are borrowings from other languages (Figure 5.27).

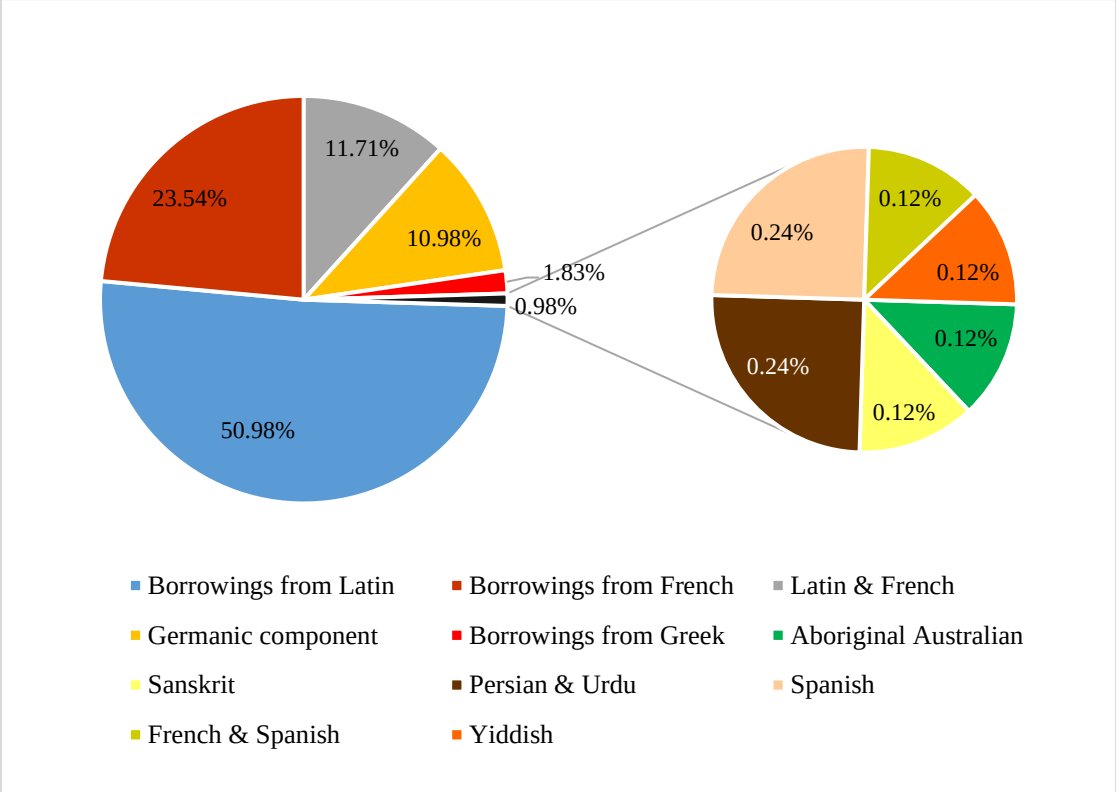


Figure 5.27. The proportions of simple borrowed adjectives

Borrowings from Latin: zero-level morphological analysis

The largest number of English adjectives were borrowed from Latin. Together with this portion of lexicon, the morphemes listed in Table 5.10 have found their way into English. The most frequent of them are the suffixes *-ate*, *-ous*, *-ive*, *-al* and the prefix *in-*.

Table 5.10. Morphemes represented in Latin borrowings

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	Aj-lat *crisp*	75	17.9	10	Aj-lat=BM+ant *resultant*	13	3.1
2	Aj-lat=BM+ate *deliberate*	59	14.08	11	Aj-lat=BM+ic *genetic*	12	2.86
3	Aj-lat=BM+ous *fabulous*	58	13.84	12	Aj-lat=BM+ary *solitary*	11	2.63
4	Aj-lat=BM+ive *adorative*	54	12.89	13	Aj-lat=BM+ible *permissible*	8	1.91
5	Aj-lat=BM+al *herbal*	37	8.83	14	Aj-lat=BM+ar *muscular*	7	1.67
6	Aj-lat=in+BM *innate*	23	5.49	15	Aj-lat=im+BM *impolite*	5	1.19
7	Aj-lat=BM+ory *oratory*	18	4.3	16	Aj-lat=de+BM *deject*	4	0.95
8	A-lat=BM+ent *eminent*	17	4.06	17	Aj-lat=ir+BM *irresolute*	3	0.72
9	Aj-lat=BM+able *viable*	14	3.34				

Borrowings from French: zero-level morphological analysis

French loans form the second largest portion in the category of loan adjectives. The most frequent suffixes borrowed together with this part of the lexicon are the suffixes *-ous* and *-ble* (see Table 5.11).

Table 5.11. Morphemes represented in French adjectival borrowings

No	Morphological pattern	Type frequency	%	No	Morphological pattern	Type frequency	%
1	Aj-fr *cruel*	89	49.44	7	Aj-fr=BM+ible *admissible*	4	2.22
2	Aj-fr=BM+ous *gorgeous*	28	15.56	8	Aj-fr=BM+ant *distant*	3	1.67
3	Aj-fr=BM+ble *voluble*	26	14.44	9	Aj-fr=dis+BM *disjoint*	2	1.11
4	Aj-fr=BM+ive *figurative*	13	7.22	10	Aj-fr=ir+BM *irresistible*	2	1.11
5	Aj-fr=BM+ent *consequent*	6	3.33	11	Aj-fr=in+BM *insupportable*	2	1.11
6	Aj-fr=BM+al *jovial*	5	2.78				

Parallel borrowings from French and Latin: zero-level morphological analysis

Dual-route adjectives involving French and Latin constitute the third largest portion of loan adjectives. As shown in Table 5.12, the suffixes *-ous*, *-ive*, *-ent*, *-able* and *-ic* are the most frequent morphemes in this share of the lexicon.

Table 5.12. Morphemes represented in Latin and French borrowing

No.	Morphological pattern	Example	Type frequency	%
1	Aj-lat/fr	*avid*	27	28.13
2	Aj-lat/fr=BM+ous	*meticulous*	13	13.54
3	*Aj-lat/fr=BM+ive*	*adoptive*	12	12.5
4	Aj-lat/fr=BM+ent	*benevolent*	11	11.46
5	Aj-lat/fr=BM+able	*curable*	10	10.42
6	Aj-lat/fr=BM+ic	*prophetic*	10	10.42
7	Aj-lat/fr=BM+al	*paternal*	9	9.38
8	Aj-lat/fr=BM+ible	*accessible*	2	2.08
9	Aj-lat/fr=BM+ory	*obligatory*	2	2.08

The Germanic component of English adjectives

Figure 5.28 illustrates the distributions of shares between Germanic languages. The largest share of adjectives have been inherited from ancient Germanic languages. Borrowings from Scandinavian, German Dutch and Danish (in descending order of frequency) form the second substantial share in this category. The minute portion is accounted for by other languages.

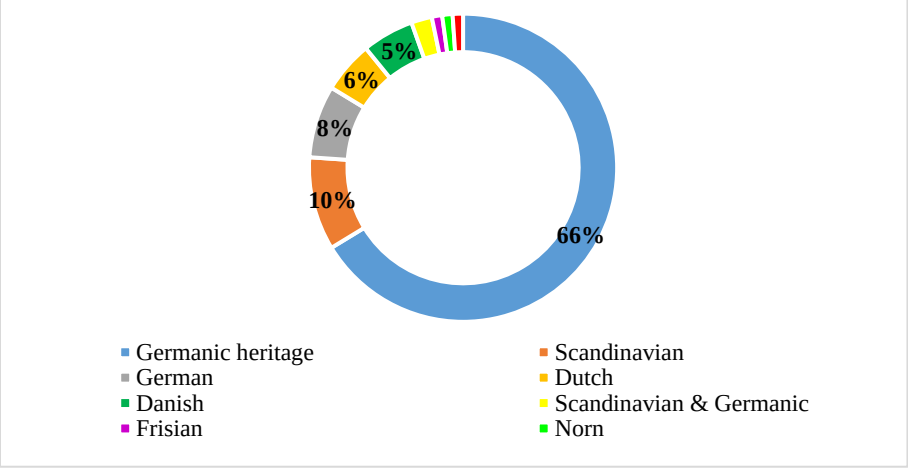


Figure 5.28. The Germanic component for simple adjectives

5.2.3.2 Old and Middle English adjectives

A high number of simple adjectives, as compared to nouns and verbs, have originated in Old and Middle English. Table 5.13 gives an account of morphological processes involved in forming adjectives. The most frequent word-formation process is conversion of past participle verbs to adjectives. Other processes involve the suffixes *-y*, *-ing*, *-ed*, *-ly*, *-ful* and *-ous*.

Table 5.13. Morphemes of Old and Middle English represented in the metacorpus

No.	Morphological pattern	Example	Type frequency	%
1	Aj=Verb*3	*beaten*	131	67.53
2	Aj	*dark*	37	19.07
3	Aj=BM+y	*drowsy*	11	5.67
4	Aj=BM+ing	*gambling*	4	2.06
5	Aj=BM+ed	*shrewd*	3	1.55
6	Aj=BM+ly	*grisly*	3	1.55
7	Aj=BM+ful	*lustful*	2	1.03
8	Aj=BM+ous	*boisterous*	2	1.03
9	Aj=Verb*2	*forsake*	1	0.52

5.2.3.3 Conversion

Conversion is the third most productive word-formation process for simple adjectives (13%). Most adjectives are formed by pure conversion (Figure 5.29), and a small share is accounted for by dual routes (see Table 9, Appendix C).

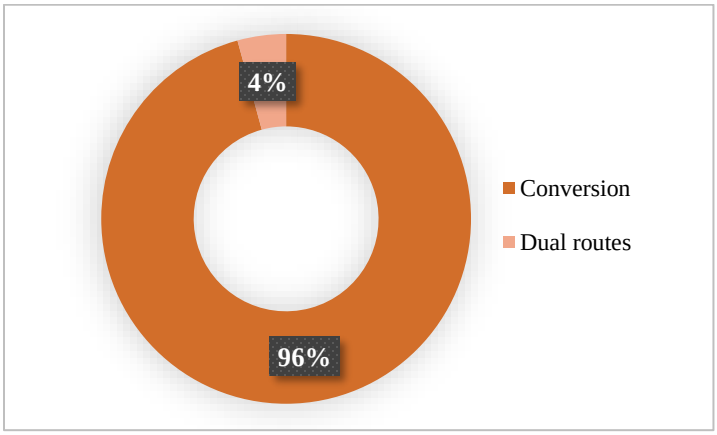


Figure 5.29. Conversion in adjectives

5.2.3.4 Phonological formations

Adjectives formed with the involvement of phonology make up 4% of all simple adjectives. The largest share are adjectives formed by phonological changes of some sort (phonological alternations or variants of original forms), and a small part is comprised of aphetic forms (Figure 5.30).

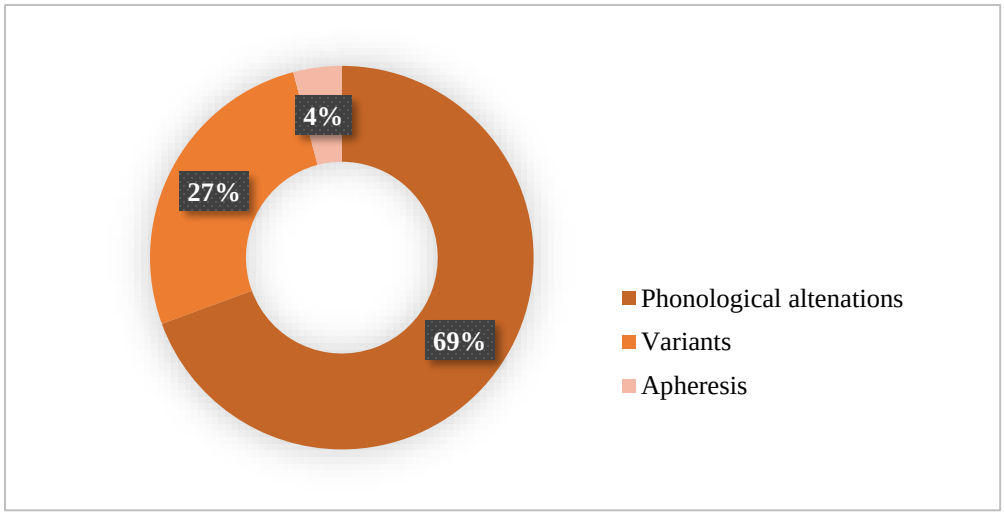


Figure 5.30. Phonological adjectival formations

5.2.3.5 Adjectives formed by contractions

Figure 5.31 illustrates that the highest number of adjectives in this category are formed by shortening of original forms. Back-formations are 14% of adjectives in this category.

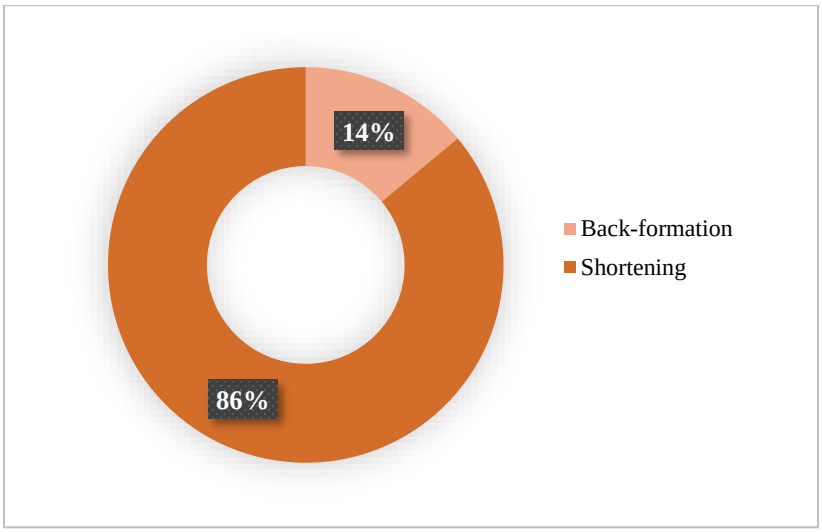


Figure 5.31. Adjectives formed by contractions

5.2.4 Simple adverbs

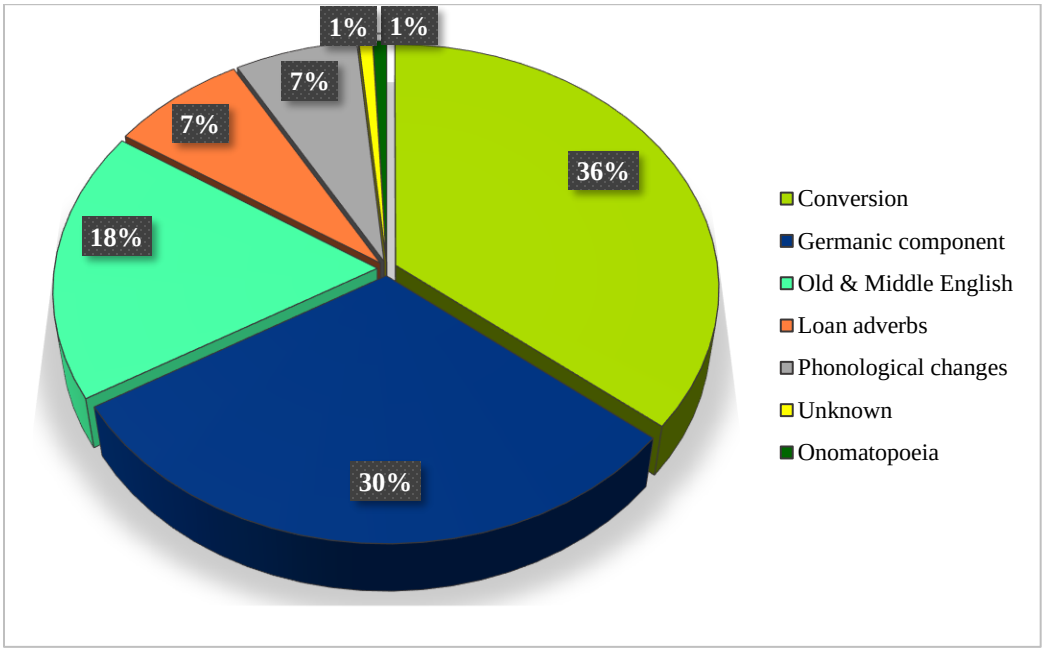


Figure 5.32. The origins of simple adverbs

Simple adverbs constitute 0.42% of all words in the sample. As shown in Figure 5.32, the most common adverb-formation process is conversion. Slightly smaller shares are accounted for by adverbs inherited from ancient Germanic, and by Old and Middle English adverb formations. Loan adverbs constitute one tenth of the simple adverbs in the sample. The smallest portion of adverbs are formed by phonological alternations and onomatopoeia. The origin of only 1% of simple adverbs is unknown. The following subsections give the detailed account of these processes.

5.2.4.1 Germanic heritage

Adverbs inherited from ancient Germanic form the second largest portion of simple adverbs. In Table 5.14, the morphological patterns of adverbs are shown: the first pattern is monomorphemic, and the second pattern contains the suffix *-ly*, which is derived from the Germanic base *-lîko* with an adverb-forming suffix *-ô*. According to the OED (2021), this suffix represents either the ending of the ablative feminine (pre-Germanic *-ād*), the ablative neuter (pre-Germanic *-ōd*), or the instrumental neuter (pre-Germanic *-ōm*).[25]

[25] "-ly, suffix 2." OED Online. Oxford University Press, June 2021.

Table 5.14. Germanic morphemes represented in the metacorpus

No.	Morphological pattern	Type frequency
1	Ad-grm *right*	24
2	Ad-grm=BM+ly *hardly*	14

5.2.4.2 Old English: zero-level morphological analysis

Table 5.15 demonstrates zero-level morphological patterns inherited from Old English. The most frequent patterns involve monomorphemic adverbs and adverbs with the suffix *-ly*, which has developed from the Old English suffix *-líce* (OED 2021). Other morphemes include the preposition *in-* in the third pattern and the suffix *-s*, which was originally *-es*—an Old English adverb-forming suffix, identical with the suffix of the genitive singular of many neuter and masculine nouns and adjectives (OED 2021).[26] The morphological patterns mentioned above have been included in monomorphemic analysis, because they are very early formations in English, and, according to the principle of morphological parsing, the words whose phonological form has diverged from the original, have been considered as such that belong to a zero-level of word formation.

Table 5.15. Morphemes in Old English represented in the metacorpus

No.	Morphological pattern	Type frequency
1	Ad *sharp*	13
2	Ad=BM+ly *sorely*	7
3	Ad=BM+s *hence*	3
4	Ad=in+BM *instead*	1

5.2.4.3 Loan adverbs

The highest number of adverbs were borrowed from Latin and French (Figure 5.33). Other borrowings involve such languages as Scandinavian, Italian, as well as Dutch and Old Saxon.

[26] "-s, suffix 1." OED Online. Oxford University Press, June 2021.

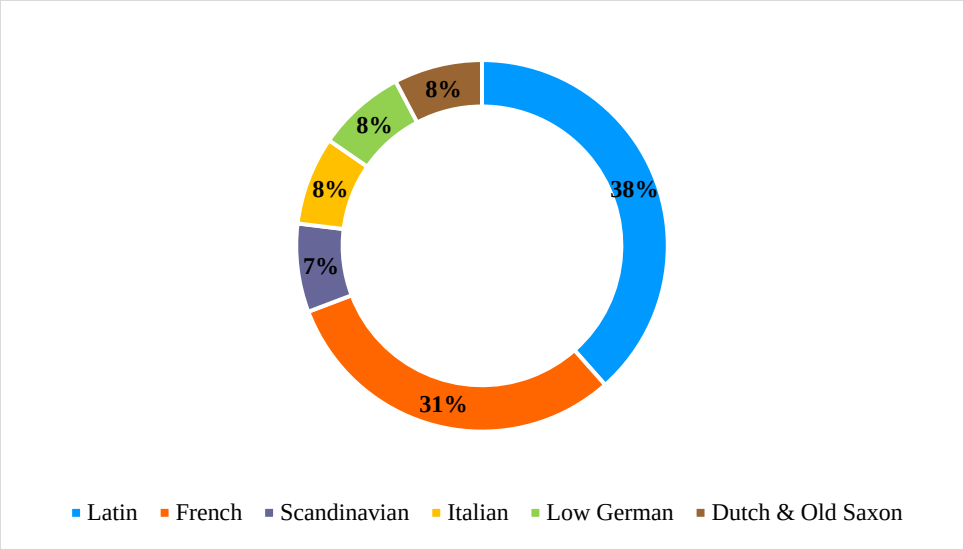


Figure 5.33. The proportions of loan adverbs

5.2.4.4 Adverbs formed by phonological changes

As illustrated in Figure 5.34, the largest portion of simple adverbs are formed by phonological alternations. The second share is accounted for by adverbs formed by the omission of the initial letter, and the third by phonological variants of original forms.

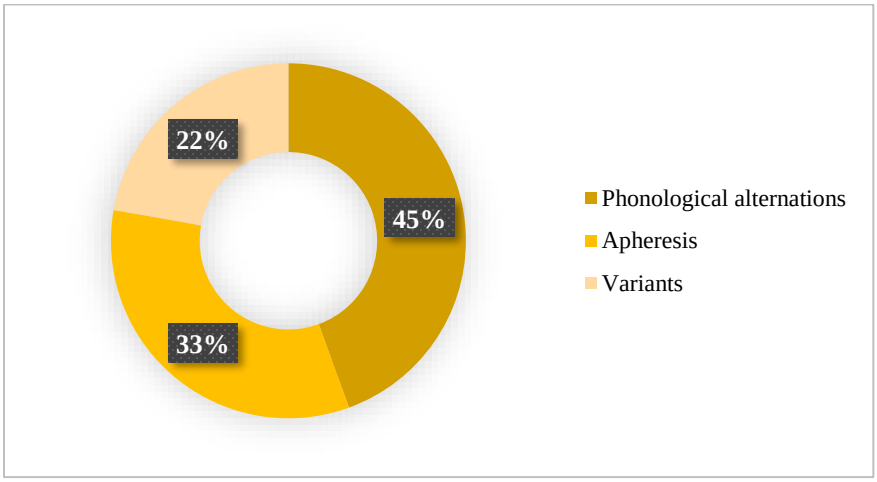


Figure 5.34. Adverbs formed by phonological changes

5.2.5 Simple interjections

Interjections make up 0.15% of the whole sample. The largest share of interjections is formed by onomatopoeia (Figure 5.35). A quarter of all simple interjections are English formations. Conversions constitute the third largest share of interjections. A small portion of interjections are borrowings from Italian, Dutch, French and Irish, and interjections formed by phonological

changes. A minute portion of simple interjections are formed by contraction, and almost 2% of interjections are of unknown origin.

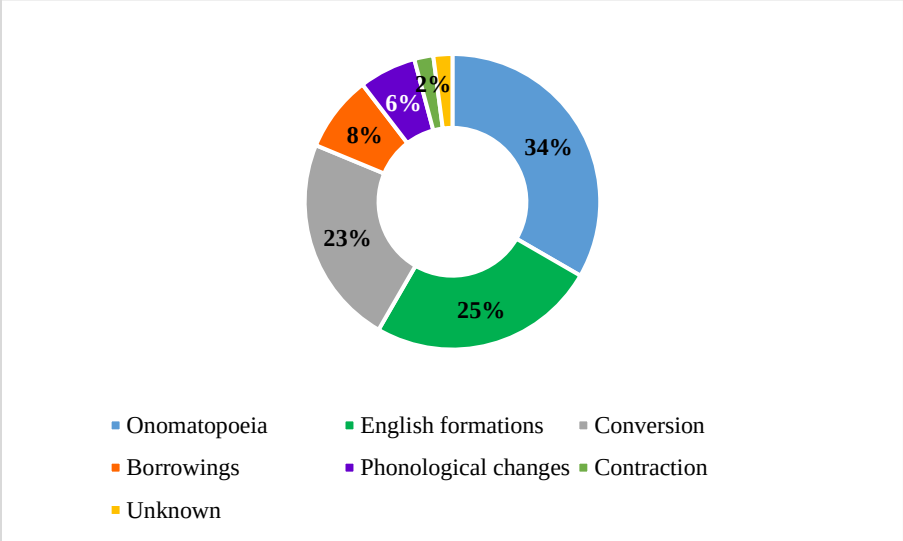


Figure 5.35. Simple interjections

5.2.6 Grammatical word classes: pronouns, conjunctions and prepositions

Simple grammatical word classes constitute 0.12% of all words in the sample. Figure 5.36 illustrates the shares for each simple grammatical class. In the following subsections, a brief overview of these classes is given.

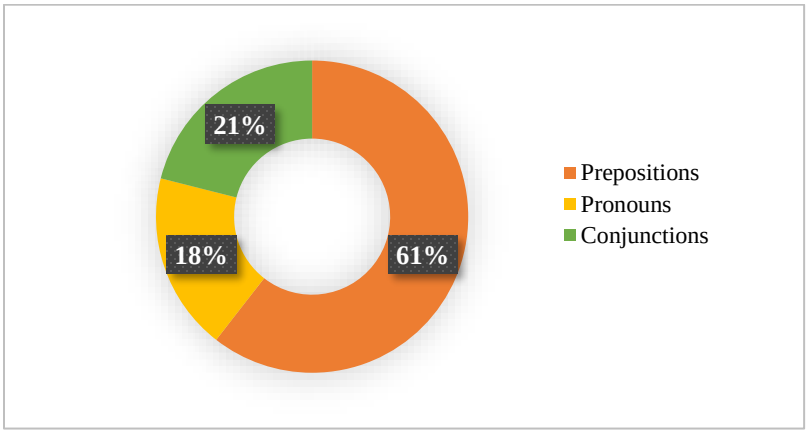


Figure 5.36. The shares of simple grammatical classes

5.2.6.1 Simple prepositions

As shown in Figure 5.37, the largest portion of simple prepositions are borrowings from Latin. Prepositions formed by phonological changes have the second largest share, followed by

prepositions inherited from Germanic. There is an equal portion of prepositions formed in Old English and prepositions formed by conversion. The smallest share is made up of parallel borrowings from Latin and French.

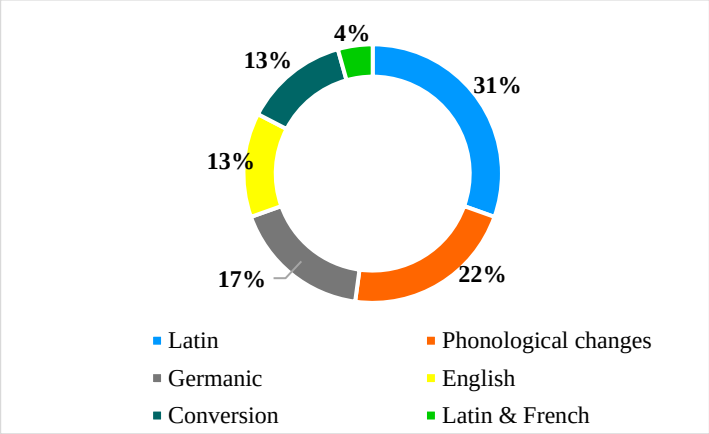


Figure 5.37. The origins of simple prepositions

5.2.6.2 Simple conjunctions

Figure 5.38 displays the shares of origin for simple conjunctions. Half of them are formed by conversion, 38% by phonological changes and 12% are of Old English origin.

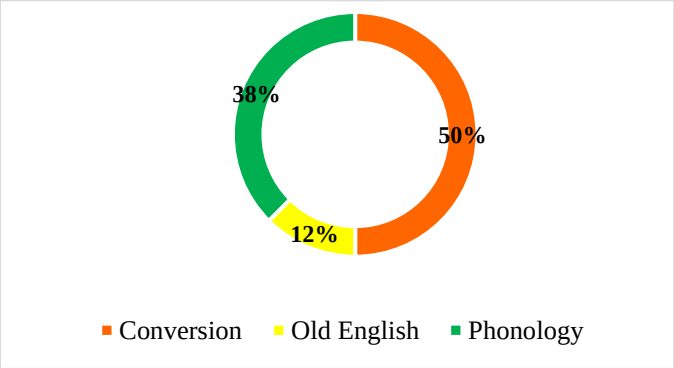


Figure 5.38. The origins of simple conjunctions

5.2.6.3 Simple pronouns

The origin of the major share of simple pronouns in the sample is accounted for by phonological changes (Figure 5.39). The second largest portion is made up of pronouns formed within English. One simple pronoun is formed semantically, and one is inherited from Germanic.

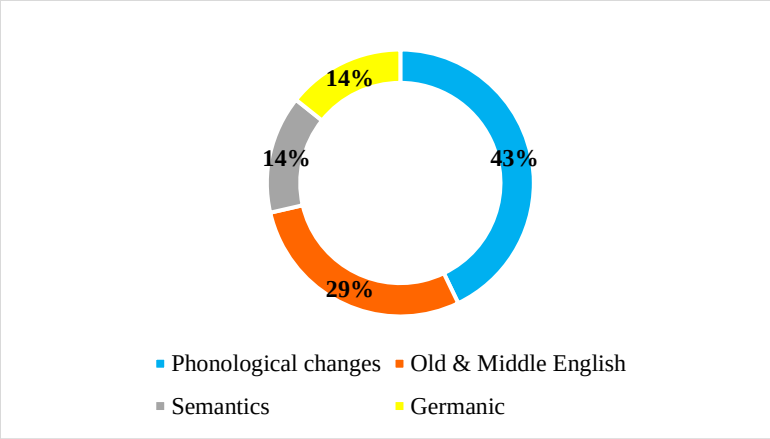


Figure 5.39. The origins of simple pronouns

5.2.7 Conversive classes

Around 6.5% of all words in the sample belong to the category of simple conversive classes. There are 49 conversive classes in total. Their names and type frequency are listed in Table 5.16.

Table 5.16. Morphological patterns for simple conversive classes

No	Morphological pattern	Type frequency	No	Morphological pattern	Type frequency	No	Morphological pattern	Type frequency
1	**N/Aj**	1533	18	**Aj/Ad/Intj**	6	35	**N/Aj/Ad/Pron/Prep**	2
2	**Aj/Ad**	147	19	**N/Verb**	6	36	**N/Aj/Ad/Verb**	2
3	**N/Intj**	84	20	**Aj/Ad/Prep**	5	37	**N/Conj**	2
4	**N/Aj/Ad**	57	21	**N/Ad/Prep**	5	38	**Ad/Part**	1
5	**N/Aj/Num**	51	22	**Ad/Conj**	4	39	**Ad/Prep/Intj**	1
6	**N/Ad**	27	23	**N/Ad/Conj**	4	40	**Ad/Pron**	1
7	**Aj/Pron**	16	24	**N/Aj/Ad/Conj**	4	41	**Ad/Verb/Intj**	1
8	**N/Aj/Ad/Intj**	14	25	**Prep/Conj**	4	42	**Aj/Ad/Prep/Intj**	1
9	**N/Aj/Intj**	14	26	**Aj/Ad/Prep/Conj**	3	43	**Aj/Prep/Conj**	1
10	**Ad/Conj/Prep**	12	27	**Aj/Ad/Pron**	3	44	**Aj/Pron/Intj**	1
11	**Ad/Prep**	11	28	**Intj/Verb**	3	45	**N/Ad/Intj/Part**	1
12	**N/Aj/Pron**	11	29	**N/Pron**	3	46	**N/Ad/Pron**	1
13	**Aj/Intj**	9	30	**Verb/Intj**	3	47	**N/Aj/Prep**	1
14	**N/Aj/Ad/Prep**	9	31	**Aj/Ad/Conj/Pron**	2	48	**N/Aj/Pron/Conj**	1
15	**N/Aj/Ad/Pron**	9	32	**N/Ad/Prep/Conj**	2	49	**N/Intj/Conj**	1
16	**N/Ad/Intj**	7	33	**N/Aj/Ad/Conj/Pron**	2			
17	**Ad/Intj**	6	34	**N/Aj/Ad/Num**	2			

The Venn diagram in Figure 5.40 illustrates the overlap areas of four major word classes: nouns, adverbs, interjections and adjectives. The overlap areas represent conversive classes, while the numbers in the ovals stand for their type frequency.

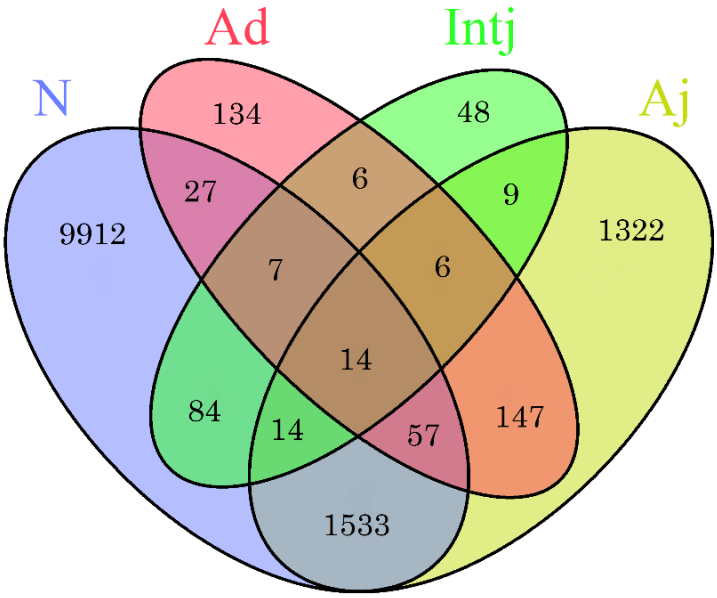


Figure 5.40. The Venn diagram of the overlap area for the four major classes:
nouns, adverbs, interjections and adjectives

The most frequent patterns include N/Aj, Aj/Ad, N/Intj, N/Aj/Ad, N/Aj/Num, and N/Ad. Less than one third of these patterns are morphological hapaxes. In what follows, the origin of words in these classes is analyzed.

5.2.7.1 The conversive class N/Aj

This class is the most frequent among conversive classes. As shown in Figure 5.41, it is largely formed by borrowings from Latin and/or French and by words inherited from ancient Germanic or borrowed from Germanic languages. The detailed account of other origins of words in this conversive class is given in Table 4 in Appendix D.

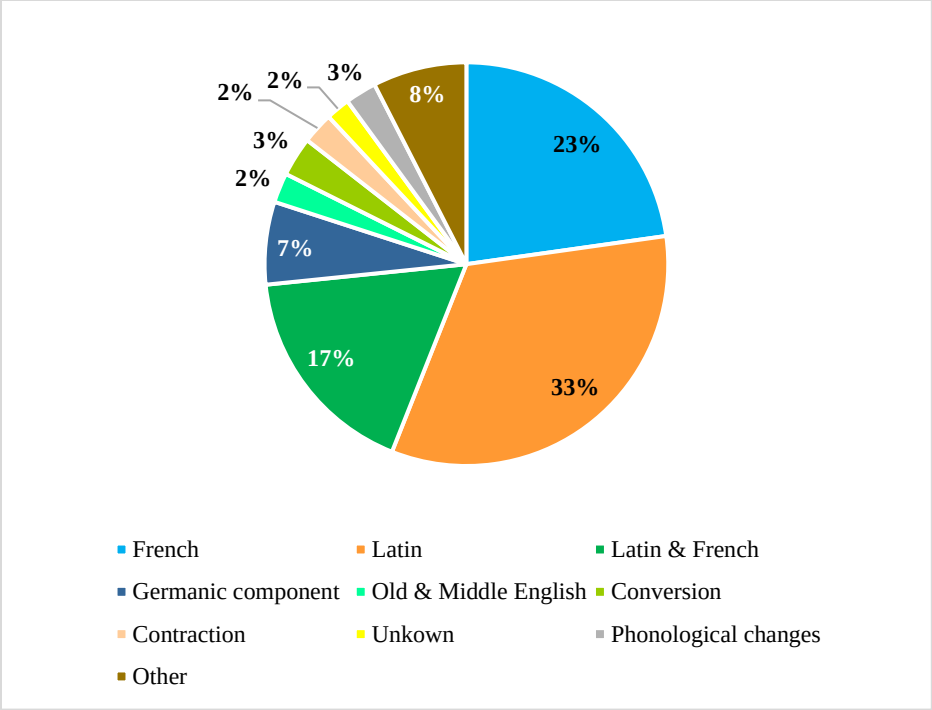


Figure 5.41. The origins of words in the conversive class N/Aj

In the following tables, zero-level word-formation morphology for the conversive class N/Aj is analyzed. For French borrowings, the most frequent morphemes are *-ant*, *-al*, *-ive*, *-ic* and *-y* (Table 5.17), and for Latin borrowings, *-al*, *-ate*, *-ic*, *-ent*, *-ive*, *in-*, *-ary* and *-ant* (Table 5.18). The most frequent zero-level Latin and French borrowings include *-al*, *-ive*, *-ic*, *-ent* and *-ant* (Table 5.19). Lastly, Table 5.20 reveals a zero-level morphological picture for words formed in Old and Middle English.

Table 5.17. Morphological patterns for French borrowings in N/Aj

No	Morphological pattern	Example	Type fr	No	Morphological pattern	Example	Type fr
1	N/Aj-fr	*alert*	154	11	N/Aj -fr=BM+ous	*superstitious*	7
2	N/Aj -fr=BM+ant	*confidant*	33	12	N/Aj-fr=BM+an	*historian*	5
3	N/Aj -fr=BM+al	*metal*	23	13	N/Aj-fr=BM+ist	*purist*	5
4	N/Aj -fr=BM+ive	*sedative*	22	14	N/Aj-fr=BM+ble	*noble*	4
5	N/Aj -fr=BM+ic	*dynamic*	19	15	N/Aj -fr=BM+age	*baggage*	4
6	N/Aj -fr=BM+y	*novelty*	19	16	N/Aj-fr=BM+BM	*stereotype*	3
7	N/Aj -fr=BM+ent	*magnificent*	14	17	N/Aj-fr=BM+ion	*minion*	2
8	N/Aj -fr=BM+able	*debatable*	13	18	N/Aj-fr=BM+ance	*romance*	2
9	N/Aj -lat/fr=BM+ible	*invisible*	14	19	N/Aj-fr=BM+eer	*volunteer*	2
10	N/Aj -fr=in+BM	*inexpert*	9				

Table 5.18. Morphological patterns for Latin borrowings in N/Aj

No	Morphological pattern	Type fr	Example	No	Morphological pattern	Example	Type fr
1	N/Aj-lat	89	*acute*	11	N/Aj-lat=BM+ible	*audible*	12
2	N/Aj-lat=BM+al	72	*diagonal*	12	N/Aj-lat=BM+ar	*solar*	12
3	N/Aj-lat=BM+ate	53	*associate*	13	N/Aj-lat=BM+an	*urban*	11
4	N/Aj-lat=BM+ic	53	*comic*	14	N/Aj-lat=BM+able	*memorable*	8
5	N/Aj-lat=BM+ent	52	*deterrent*	15	N/Aj-lat=BM+or	*exterior*	8
6	N/Aj-lat=BM+ive	37	*locomotive*	16	N/Aj-lat=de+BM	*defunct*	6
7	N/Aj-lat=in+BM	31	*intricate*	17	N/Aj-lat=il+BM	*illegitimate*	3
8	N/Aj-lat=BM+ary	24	*stationary*	18	N/Aj-lat=BM+ous	*studious*	3
9	N/Aj-lat=BM+ant	23	*radiant*	19	N/Aj-lat/fr=BM+ure	*signature*	3
10	N/Aj-lat=BM+ory	17	*trajectory*				

Table 5.19. Morphological patterns for Latin and French parallel borrowings in N/Aj

No	Morphological pattern	Type fr	Example	No	Morphological pattern	Example	Type fr
1	N/Aj-lat/fr	60	*alien*	10	N/Aj-lat/fr=BM+able	*portable*	5
2	N/Aj-lat/fr=BM+al	49	*crystal*	11	N/Aj-lat/fr=BM+ian	*barbarian*	3
3	N/Aj-lat/fr=BM+ive	47	*active*	12	N/Aj-lat/fr=BM+ity	*quality*	3
4	N/Aj-lat/fr=BM+ic	39	*ethic*	13	N/Aj-lat/fr=BM+ure	*mixture*	3
5	N/Aj-lat/fr=BM+ent	18	*orient*	14	N/Aj-lat/fr=BM+ion	*precision*	2
6	N/Aj-lat/fr=BM+ant	13	*assistant*	15	N/Aj-lat/fr=BM+ory	*accessory*	2
7	N/Aj-lat/fr=BM+ary	9	*anniversary*	16	N/Aj-lat/fr=BM+ble	*flexible*	2
8	N/Aj-lat/fr=BM+er	6	*ginger*	17	N/Aj-lat/fr=BM+ate	*primate*	2
9	N/Aj-lat/fr=BM+ous	5	*religious*	18	N/Aj-lat/fr=pro+BM	*profane*	2

Table 5.20. Morphological patterns formed in Old English

No	Morphological pattern	Example	Type frequency
1	N/Aj	*twin*	25
2	N/Aj=Verb*3	*gone*	10
3	N/Aj=BM+ling	*darling*	2

5.2.7.2 The conversive class Aj/Ad

This simple conversive class has the second largest type frequency in the sample. Figure 5.42 illustrates that its major part is made of borrowings from Latin and/or French, words conversed from nouns, bound morphemes and verbs, words formed on native grounds, and words inherited

from Germanic. A small share of this class is accounted for by phonological changes and contraction. More than 8% of words are of unknown origin.

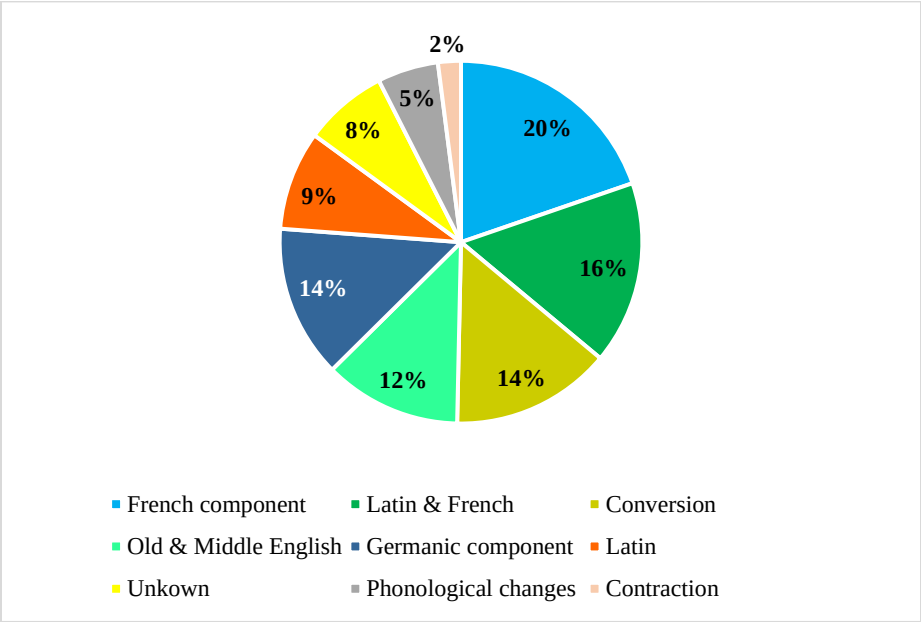


Figure 5.42. The origins of words in the conversive class Aj/Ad

Table 5.21–5.22 show zero-level morphology for French and for Latin and French borrowings in the class Aj/Ad.

Table 5.21. Zero-level morphological patterns for French borrowings in Aj/Ad

No	Morphological pattern	Example	Type frequency
1	Aj/Ad-fr	*honest*	17
2	Aj/Ad-fr=BM+ous	*perilous*	4
3	Aj/Ad-fr=BM+able	*inexplicable*	4
4	Aj/Ad-fr=BM+ant	*instant*	3

Table 5.22. Zero-level morphological patterns for Latin and French borrowings in Aj/Ad

No	Morphological pattern	Example	Type frequency
1	Aj/Ad-lat/fr	*abundant*	18
2	Aj/Ad-lat/fr=BM+ous	*gracious*	3
3	Aj/Ad-lat/fr=BM+al	*natural*	3

5.2.7.3 The conversive class N/Intj

As follows from the name of this class, its most productive word-formation process is onomatopoeia, followed by heritage from ancient Germanic and by conversion. Figure 5.43 illustrates the shares of words' origin in this category. The origin of 4% of these words is unknown.

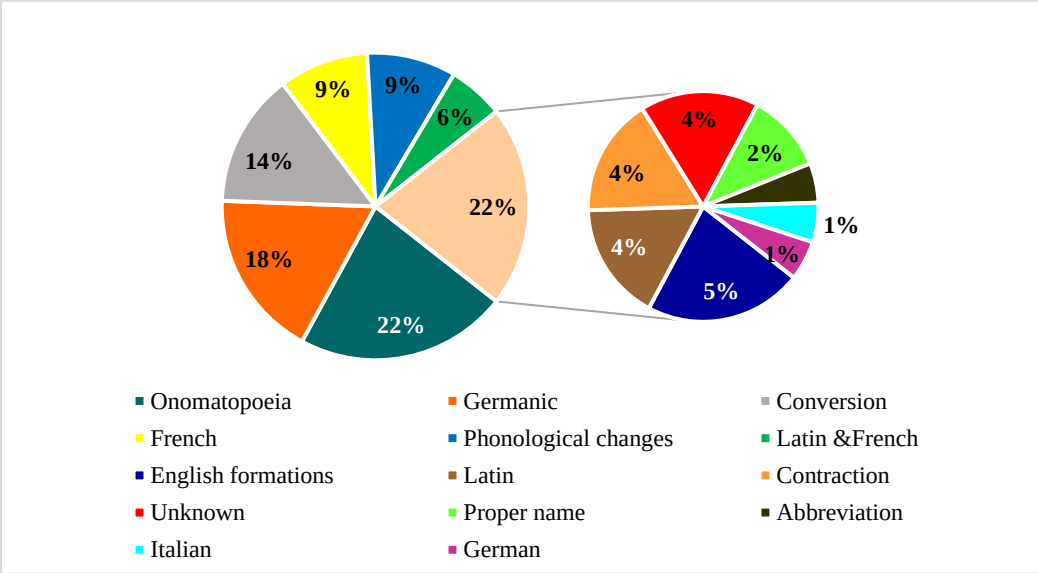


Figure 5.43. The origin of simple conversive class N/Intj

5.2.7.4 The conversive class N/Aj/Ad

As illustrated in Figure 5.44, this class is largely formed by words borrowed from French and by words inherited from ancient Germanic. Around 7% are words formed on native ground. Others include borrowings from Latin and French, Anglo-Norman, and Latin. A small share is accounted for by conversion from a verb.

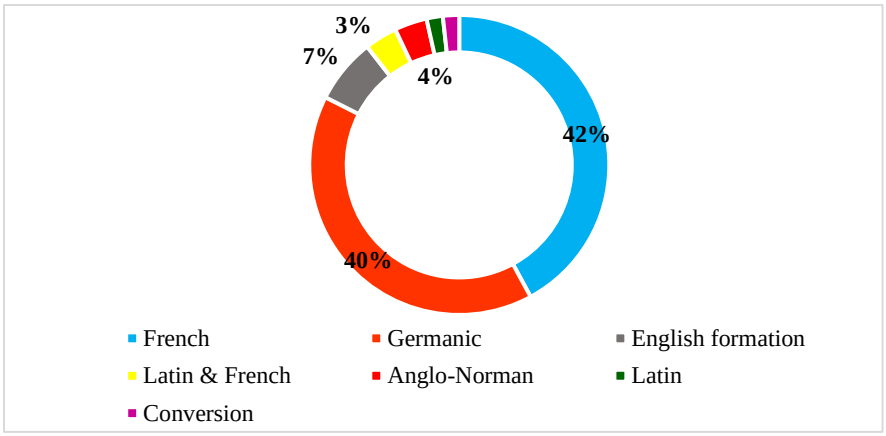


Figure 5.44. The origin of simple conversive class N/Aj/Ad

5.2.7.5 The conversive class N/Aj/Num

This class is mainly formed by numerals. Figure 5.45 shows that the largest share in this class is owned by words inherited from Germanic, whereas a smaller share are formed by words that originated in Old English.

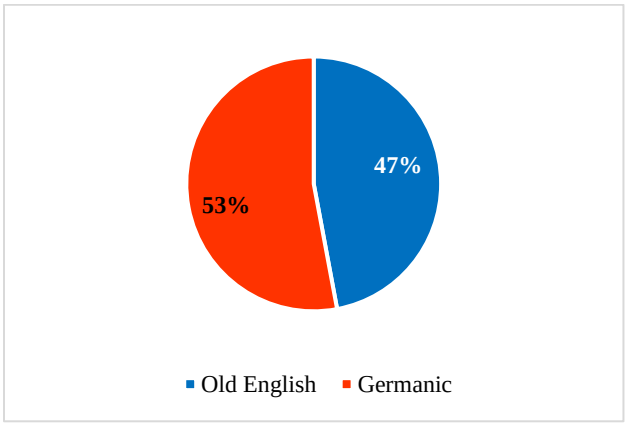


Figure 5.45. The origin of simple conversive class N/Aj/Num

5.2.7.6 The conversive class N/Ad

The origin of words in this conversive class is largely heterogeneous. Figure 5.46 demonstrates the major trends in the formation of this class. The major portion of its words are inherited from Germanic. Around 37% of words are borrowing from Latin, French and other languages. A small share of words are formed by phonological alternations and on native grounds, and a minute portion is accounted for by onomatopoeia and contraction.

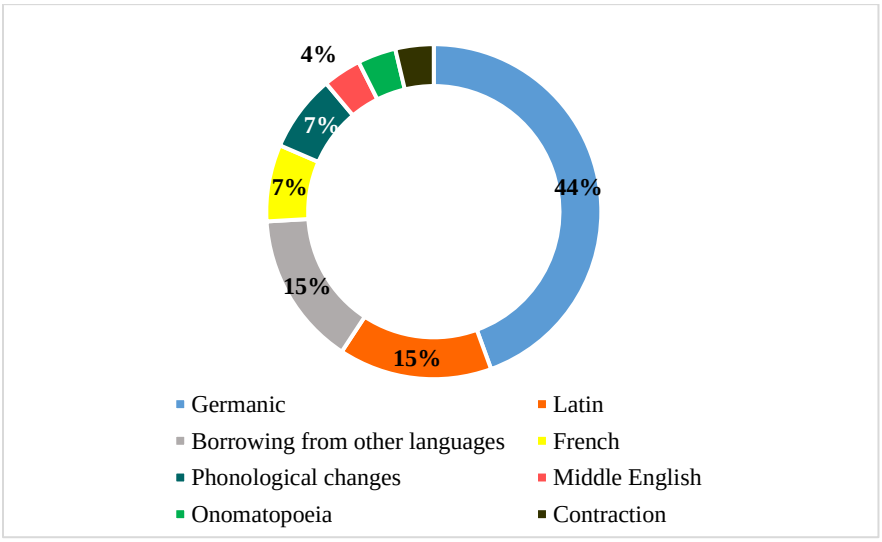


Figure 5.46. The origin of simple conversive class N/Aj/Num

5.2.8 The general trends in the formation of English simplexes

The previous sections have shed light on the origins of English simplexes in the identified word classes. Although the overall tendencies are similar for all classes, they differ in proportion of the identified word-formation processes. This section summarizes these distinctions across the different word classes. All tables with the frequencies of constructions can be found in Appendix E.

All word classes, apart from conjunctions, pronouns, adverbs and low-frequency conversive classes, have a large share of loan words. The highest number of borrowings has been identified for nouns/adjectives (73%) and adjectives (55%), followed by nouns (50%) and verbs (24%). Only 7% of adverbs are of foreign origin, which suggests that this word class tends to rely on the internal resources of language, is less susceptible to changes, and is more grammar-driven than other major word classes. This is also evidenced by the fact that adverbs show the highest number of items in the Germanic component (30%) and in the Old and Middle English formations (18%), which is smaller for adjectives/adverbs (14% and 12% respectively), nouns (12% and 4%), verbs (9% and 5%), adjectives (7% and 15%) and nouns/adjectives (7% and 2%).

Further, conversion is the most productive word-formation process for verbs (50%) and adverbs (36%), whereas for nouns and adjectives it is less productive (16% and 13% respectively). These correlations mean that nouns and adjectives are the main sources of borrowed words in language, which are then passed onto other word classes through conversion. The conversion of nouns and adjectives to verbs is the most productive formation path.

The contraction of original forms of words (both via clipping of syllabic parts or morphemes) is more productive for nouns (5%), less productive for adjectives (3%), verbs (2%), nouns/adjectives (2%) and adjectives/adverbs (2%), and fully unproductive for adverbs. Moreover, in view of types of contraction, back-formation which involves deducing a morpheme from a borrowed word is more common in verbs (58% of all instances of contraction), whereas for nouns syllabic clipping is more frequent.

A small portion of simplexes are formed by phonological alternations of original words. This word-formation process is significant for adverbs (7%), nouns (6%), and adjectives/adverbs (5%). It is less productive for verbs (4%), adjectives (4%) and nouns/adjectives (5%). Specifically, the dominant phonological process in adverbs is apheresis, which is the omission of an initial phoneme.

Phonology is also an important process for grammatical classes: it is the first most productive process for pronouns (43%) and the second for prepositions (22%) and conjunctions (38%). Since grammatical words tend to be the most token-frequent in language, this observation echoes with the literature on the relationship between word frequency and morphonological change (e.g. Schuchardt 1885[1972], Bybee 2007), supporting the view that "sound change initiates in language use" (Bybee 2007: 31).

Furthermore, onomatopoeia is prominent in nouns/interjections (22%). In other classes, it varies from less than 1% (nouns, adjective and adverbs) to 2% (verbs). Finally, semantic formations constitute 1% of words in nouns and verbs. No semantic formations have been identified for adjectives and adverbs, which suggests that these classes are less prone to semantic shift and tend to maintain their original meaning. However, it should be mentioned that some semantic word formations may involve subtle changes in meaning, also known as broadening or narrowing. For example, in Old English the word *brid* or *bird* had a narrow semantic meaning of 'young bird, chicken', which in Modern English has developed a general sense of a feathered animal, able to fly (Kastovsky 2006: 216). These semantic processes require deeper investigation and might not be captured in the broad categories identified in this study due to its different primary research goals.

In light of Latin and French borrowings, French has the upper hand for nouns, verbs and adjectives/adverbs, whereas Latin for adjectives, adverbs and nouns/adjectives. Latin is also a major source of English prepositions (31%). Among the Germanic languages, the highest number of words in all classes was borrowed from Scandinavian and Dutch, mainly during the period of Middle English (Kastovsky 2006: 249).

Lastly, a zero-level morphological analysis of the sample—the purpose of which has been to identify morphemes borrowed/inherited from source languages—has revealed that the most frequent morphemes borrowed together with French words include *-ment*, *-er*, *-ance*, *-age*, *-ine*, and *-ure* for nouns, and *re-*, *dis-*, *-ish*, and *-ify* for verbs. Further, as informed by the sample, Latin borrowings strengthened the representation of nominal suffixes *-um*, *-ia* and *-ary*, the verb-forming morphemes *-ate*, *-in*, and *de-*, and the adjectival suffixes and prefixes *-ate*, *-ive*, *-al* and *in-*. The affixes *-ion*, *-ity*, *-y*, *-ence*, *-or* for nouns, *de-* for verbs, and *-ous* and *-ble/able* for adjectives are strongly present in both Latin and French borrowings. Finally, native zero-level morphology

involves the nominal suffixes *-le*, *-ing*, *-s (pl)*, *-er*, *-ness*, *-ock*, *-th* and *-ship*, and adjectival suffixes *-y*, *-ing*, *-ed*, *-ly* and *-ful*.

5.3 Multimorphemic words of the sample: overall structural analysis

This section deals with multimorphemic words of the sample and presents their structural analysis. Sections 5.3.1–5.3.4 give an account of morphological constructions and corresponding patterns for nouns, verbs, adjectives and adverbs, as well as their quantitative characteristics in the sample. Section 5.3.5 looks at the morphological characteristics of grammatical classes, and Section 5.3.6 at those of conversive classes. Finally, in Section 5.3.7 the main trends in the formation of multimorphemic words are summarized.

5.3.1 Multimorphemic nouns

Multimorphemic nouns constitute over 42% of all multimorphemic words and about 15.5% of the whole sample. English noun formation occurs on four levels (Figure 5.47). In this study, the structural level of word formation indicates how many morphemes are involved in the formation of a morphological pattern (for more detail, see Section 3.2.3).

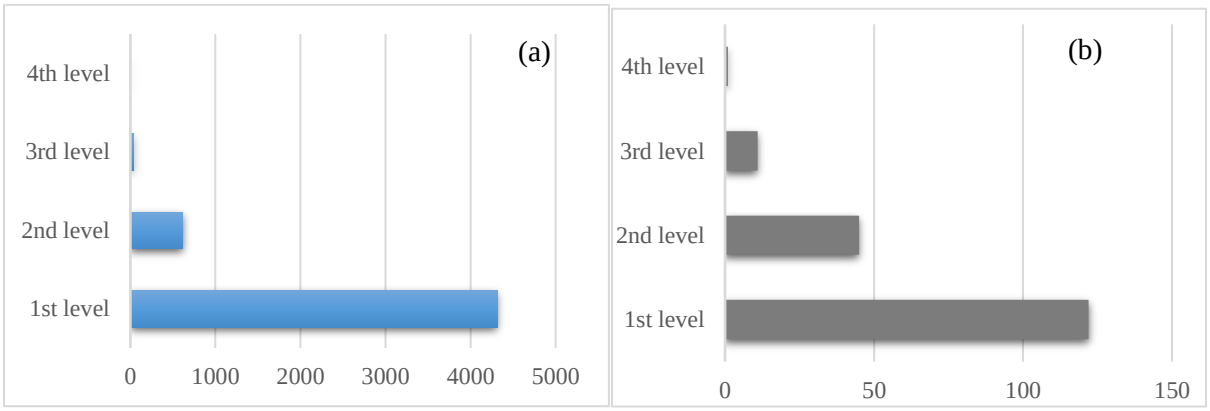


Figure 5.47. The proportion of items in the structural levels of noun formation:
(a) words (b) morphological constructions

The first morphological level of noun formation, which encompasses duomorphemic nouns, is diverse and consists of 122 morphological constructions and 4,314 words (Figure 5.47). The type frequency of noun morphological constructions is presented in Tables 5-7 of Appendix E). These tables also provide the orthographical/morphonological changes observed within each construction, as well as the values of the type valency for each construction with and without the consideration of conversive property of some roots (marked in the tables as ‘with CC’ and ‘without

CC'). For example, the noun *easiness*, which belongs to the construction {C+ness}, has been formed from the conversive class N/Aj/Ad *easy*[27] and the suffix *-ness*. If we calculate the type valency of the suffix *-ness* in this construction taking into account all conversive classes that can be placed in the slot of the root 'C' of this construction, we get a value of 8 (N, Aj, N/Aj, Aj/Ad, Aj/Ad/N, Verb, Ad and BM). However, if we decide to avoid conversive classes in the description, we may choose to assign a conversive root to a single word class, the most type-frequent in a category: e.g. we may assume that the noun *easy* consists of the adjective *easy* and the suffix *-ness*, because the morphological pattern N'=Aj+ness is the most frequent within the {C+ness} construction. In this view, the type valency of the suffix *-ness* is 5 (N, Aj, Verb, Ad and BM).

As inferred from Table 5 (Appendix E), three constructions have the largest share at the first level of noun formation: {C+er} (21%), {C+ing} (21%) and {C+C} (16%). The morphological constructions {C+ness}, {CC} and {C+ism} produce 5%, 4% and 3% of the multimorphemic nouns of the first level respectively, whereas {C+ment}, {C+ist}, {C+age}, (C+ity} and {C+ion} give rise to 2% of bimorphemic nouns. The rest of the noun morphological constructions (excluding hapaxes) seize the share of 1% or less and involve the affixes *-y, -or, -ship, -ee, -ery, re-, -ance, -al, -dis, -ess, mis-, fore-, -ence, sub-, -dom, -et, -ency, -s, -le, -hood, -ie, -ry, -cy/acy, -ful, -let, counter-, -ancy, -ian, -eer, in-, un-, -ling, -ure, -ate, -s-pl, -ant, up-, -in, -th, -ine, -ster, semi-, super-, -ette, inter-, -o, anti-, -ium, co-, de-, mal-, self-, -ide* and *non-*. Finally, almost 32% of all morphological constructions of the first level are hapaxes.

The morphological constructions that display the highest number of orthographical/morphonological changes include {C+er}, {C+ion}, {C+ness}, {C+ing}, {C+y} and {C+ity}. These changes encompass the repetition of consonants (annotated with ':', e.g. ':p'), the omission of letters, sounds and morphemes ('¢'; e.g. '¢e'), the insertion of virtual sounds/morphemes (annotated in brackets; e.g. '(t)') and the transition of sounds/morphemes (shown with the colon and arrow; e.g. ':(b→m)'). Moreover, the highest orthographical/morphonological changes have been observed for the morphological constructions with a high value of type frequency and type valency, which can be viewed as a frequency effect.

[27] In accordance with the OED, the word base *easy* can be assigned to the categories of nouns, adjectives and adverbs. For this reason, it is marked as a conversive class of N/Aj/Ad.

At the second level of noun formation, there are 45 morphological constructions and 616 words. The most frequent constructions are listed in Table 8 (Appendix E) together with their type frequency, type valency and orthographical/morphonological changes. Almost one fifth of this level is made of the constructions {C+C}, {C+ness}, {C+er}, {C+ing} and {C+ity}, owning the share of 29%, 20%, 15%, 12% and 10%, respectively. The morphological constructions with the type frequency of 2-20 include the affixes *de-*, *-ist*, *-ship*, *re-*, *-y*, *un-*, *-s*, *-ess*, *co-*, *on-*, *-hood*, *up-*, *-ery/ry*, *-ance*, *-al*, *-dom*, *-acy*, *-or*, *mal-*, *in-*, *mis-*, *-ian*, *anti-* and *-s-pl* (Table 9, Appendix E). Morphological hapaxes constitute less than 2% of the second-level morphological constructions (Table 10, Appendix E). Lastly, all constructions at this level show a fewer orthographical/morphonological changes.

As for the third level, it contains 11 morphological constructions and 36 words (Table 11, Appendix E). The most type-frequent morphological constructions at this level are {C+ness}, {C+ion} and {C+C}. The fourth level consists of one construction and is formed by compounding (morphological pattern: N””=N’+Aj+N’-pl, which produces the word *daddy-long-legs*).

The morphological constructions that are present in all three levels of noun formation include {C+C}, {C+er}, {C+ing}, {C+ion}, {C+ism}, {C+ity}, {C+ment}, {C+ness}, {C+s}, {C+ship} and {de+C}. Further, most morphological constructions are more frequent at the first level of noun formation, excluding {de+C} and {on+C}, which are more type-frequent at the second level. Finally, the maximum observed value for the type valency in nouns is 7 (i.e. in the morphological construction {C+ness}), but the median for the type valency distribution in nouns is 1.

5.3.2 Multimorphemic verbs

Verb formation involves only two structural levels (Figure 5.48). At the first level, the most frequent constructions are {C+ize}, {re+C) and {C+en} which respectively have shares of 19%, 18% and 9%,. Hapaxes constitute 0.72% of all morphological constructions of the first level. Table 12–13 (Appendix E) summarize the morphological information about the first level of verb formation. As can be seen from these tables, fewer orthographical/morphonological changes are observed for verbs. These changes are the omission of letters/morphemes (e.g. ‘¢e’ and ‘¢ism’), and the reduplication and the transition of sounds (e.g. ‘:t’, ‘:d’; ‘:(se→zz)’), as well as the insertions of virtual sounds (e.g. (t) and (i)).

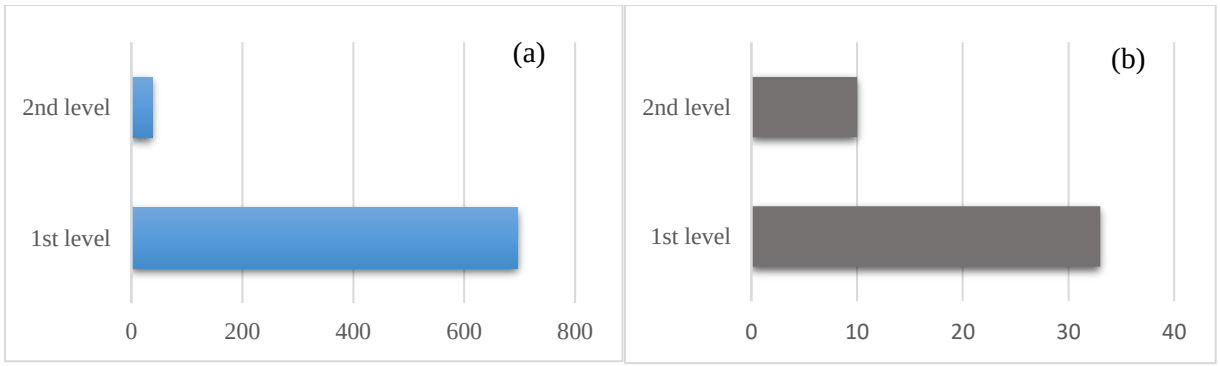


Figure 5.48. The proportions of items in the structural levels of verb formation:
(a) words; (b) morphological constructions

The second level of verb formation contains 10 morphological constructions and 38 words (Table 14 in Appendix E). Its most type-frequent constructions are {C+ize}" and {de+C}". Nearly 16% of the second-level constructions are hapaxes. Furthermore, all constructions producing verbs are more frequent at the first structural level. Finally, their maximum type valency is 5, and the median for the type valency distribution of verbs is 2.

5.3.3 Multimorphemic adjectives

Adjectival formation in the metacorpus consists of three levels, with the major derivation occurring on the first level (Figure 5.49). However, as compared to other word classes, there are a considerable number of morphological constructions on its second structural level.

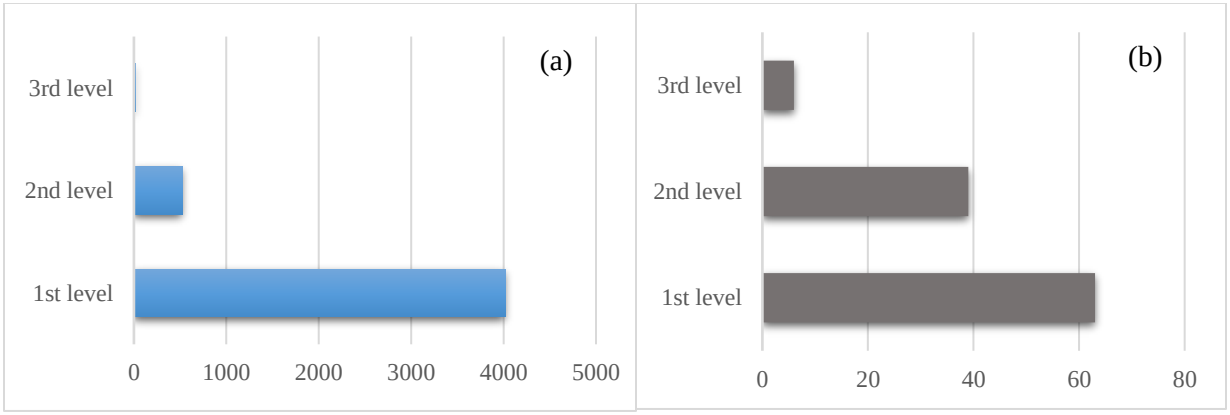


Figure 5.49. The proportion of items in the structural levels of adjectival formation:
(a) words (b) morphological constructions

As illustrated in Table 15 (Appendix E), the most type-frequent pattern in adjectival formation, as well as in the whole metacorpus, is {C+ed} which accounts for 41% of adjectives on the first structural level. It is followed by the constructions {C+ing} (16%), {C+able} (8%), {C+y} (7%),

{C+al} (6%), {C+less} (4%) and {C+ful} (3%). Moreover, most adjectival constructions with a type frequency of above 20 involve a high degree of orthographical/morphonological changes. The lists of adjectival morphological constructions with a type frequency of 2-20 and with morphological hapaxes are given in Tables 16 and 17 in Appendix E.

The second level of adjectival formation contains 39 constructions and 524 words. The most type-frequent constructions at this level are {un+C}, {C+ed} and {C+C} (Table 18). Another interesting property of the second-level adjectival formation is that there are many constructions, whose type frequency is higher on this level (as compared to the first level): {de+C}, {extra+C}, {il+C}, {ir+C}, {mis+C}, {over+C}, {pre+C}, {re+C}, {self+C} and {un+C}. Specifically, the type frequency of {un+C} is almost thrice its first-level type frequency (162 vs 60), which suggests that some adjectival prefixation processes are more dominant on the second level of adjectival formation. Tables 19 and 20 (Appendix E) offer a detailed account of adjectival constructions with the type frequency 2-5 together with morphological hapaxes.

At the third level of adjectival formation (Table 21, Appendix E), there are 6 morphological constructions and 17 words. Most of the third-level constructions (e.g. {C+C}, {C+ed}, {C+ing}, {non+C} and {un+C}) are present in all three levels of adjectival formation. Finally, although the highest observed type valency for adjectives is 6, the median for the type valency distribution in adjectives is 1.

5.3.4 Multimorphemic adverbs

Adverbial formation involves three levels (Figure 5.50). The second level of adverbial formation is prominent in that it spawns a high number of words, as compared to other word classes and as mapped against its first level.

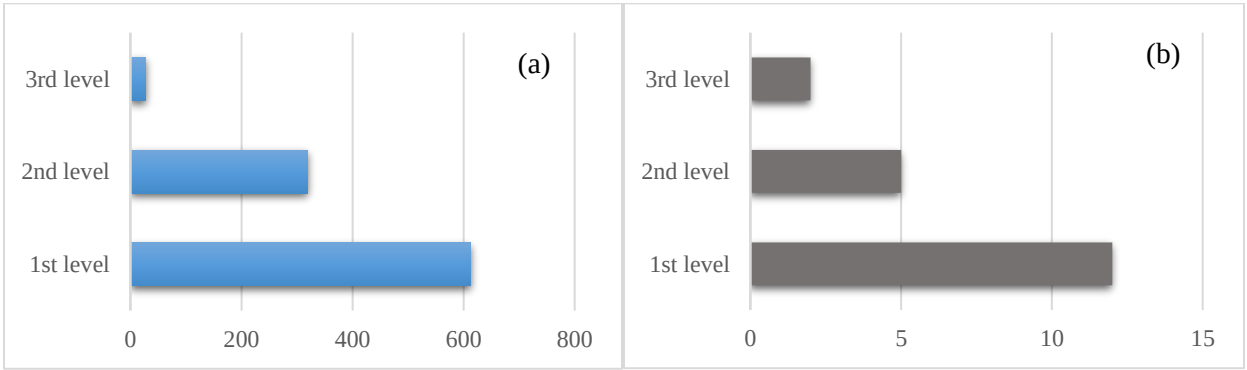


Figure 5.50. The shares of items in the structural levels of adverbial formation:
(a) words; (b) morphological constructions

Tables 22–24 (Appendix E) summarize the morphological information of adverbial formation at three levels. The most type-frequent construction of the first level is {C+ly}, which produces 94% of adverbs. The morphological constructions {C+C} and {a+C} account for 2% and 1.5% of all adverbs at the first level. Prominently, orthographical/morphonological changes on the boundaries of morphemes in adverbial formation are insignificant, which mainly involve the omission of the letter(s)/phoneme(s) *le* and *l* and the diachronic sound change of the prepositions *on* and *of* to the prefix *a-*.

A similar picture is observed at the second level of adverbial formation (Table 23, Appendix E). Almost 95% of all three-morphemic adverbs are formed by the morphological construction {C+ly}". The second type-frequent morphological construction of this level is {un+C}", which forms 3% of adverbs. At the third level of adverbial formation, {C+ly}"' is also the most frequent. A maximum value of 6 has been observed for this most type-frequent adverbial construction. The median for the adverbial type valency distribution is 1.

5.3.5 Multimorphemic grammatical classes

Multimorphemic single grammatical classes constitute a minute portion of the sample. Their constructions together with their type frequency are given in Table 5.23. The dominant morphological construction in the formation of grammatical classes is {C+C}.

Table 5.23. Morphological constructions for grammatical classes

Morphological class	Morphological construction	Type frequency	Examples
Conjunctions	**{C+C}**	1	although
Prepositions	**{C+ing}**	6	concerning
	{C+C}	1	upon
Pronouns	**{C+C}**	5	anyone
	{C+s}	2	ourselves

5.3.6 Conversive classes

This section focuses on the morphological structural properties of conversive classes. Subsections 5.3.6.1–5.3.6.3 look at the conversive classes of nouns/adjectives, adjectives/adverbs and nouns/adjectives/adverbs, respectively, whereas subsection 5.3.6.4 depicts a general morphological picture of the remaining smaller conversive classes.

5.3.6.1 Multimorphemic nouns/adjectives

The structure of nominal/adjectival formation is illustrated in Figure 5.51. It involves three levels, with most word formation occurring on the first level.

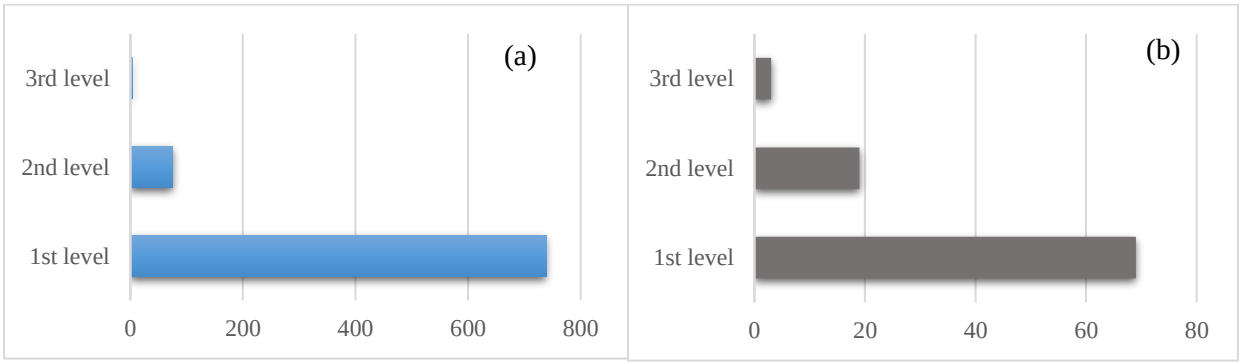


Figure 5.51. The proportion of items in the structural levels of nominal/adjectival formation: (a) words (b) morphological constructions

As shown in Tables 25 and 26 in Appendix E, a prominent feature of nominal/adjectival formation of the first level is that the type frequency of its morphological constructions decreases more consistently and with smaller intervals, as compared to other word classes. The most frequent morphological nominal/adjectival constructions include {C+C}, {C+al}, {C+ed}, {C+an}, {C+y} and {C+ist}, which produce from 5% to 15% of nouns/adjectives, respectively. Hapaxes comprise 3% of the morphological constructions at the first level (Table 27 in Appendix E).

At the second level, the constructions {un+C}[28] and {C+C} have the highest type frequency (Table 28 in Appendix E). The morphological hapaxes of the second level are listed in Table 16 in Appendix E. The third level of nominal/adjectival formation includes three hapaxes: {C+C}”’, {non+C}”’ and {un+C}”’. These are also the constructions that are present in all three levels of nominal/adjectival formation. Finally, the highest value of the type valency observed for this class is 4 with a median of 1 for its distribution.

5.3.6.2 Multimorphemic adjectives/adverbs

Adjectival/adverbial formation encompasses three levels. Table 30 (Appendix E) summarizes the morphological constructions of the first level with a type frequency of above 2 and their orthographical/morphonological changes, which are few (see Table 31 in Appendix E for morphological hapaxes). The second level contains the morphological constructions {C+C}”,

[28] For example, the word *unthinkable* is produced by this construction. The OED qualifies this word as an adjective and a noun. Frederic H. Balfour, a British essayist and sinologist, wrote the essay entitled ‘Unthinkables’.

{anti+C}", {un+C}" and {C+s}", which in total produce 12 words (see Table 32 in Appendix E for morphological hapaxes). The third level is formed by three types produced with the morphological construction {C+C}"'. Lastly, the maximum type valency observed in this class is 3 with a median of 1 for its distribution.

5.3.6.3 Multimorphemic nouns/adjectives/adverbs

The formation of words in this conversive class involves two levels (Tables 33–35 in Appendix E). The high-frequency constructions are {C+C}, {C+ward} and {C+ful} at the first level, and {un+C}" at the second level.

5.3.6.4 Other conversive classes

A morphological picture of the other 29 conversive classes is highly heterogeneous (all conversive classes are listed in Table 5.24). The most frequent conversive classes are represented in the Venn diagram (Figure 5.52). Further, Table 36 in Appendix E provides a detailed morphological account of these classes. Among them, the most productive constructions are {C+C}, {C+s}, {C+ing}, {a+C} and {C+ly}, due to their involvement in the formation of words across many of these classes.

Table 5.24. The list of all multimorphemic conversive classes

No	Conversive class	Type fr	No	Conversive class	Type fr
1	**N/Aj**	817	17	**Pron/N/Ad/Intj**	2
2	**Aj/Ad**	105	18	**Aj/Ad/Pron**	2
3	**N/Aj/Ad**	78	19	**N/Ad/Conj/Prep**	2
4	**Ad/N**	16	20	**N/Aj/Ad/Pron**	2
5	**Ad/Prep**	11	21	**Ad/Conj**	2
6	**Pron/Aj**	7	22	**Ad/Conj/Prep**	1
7	**Aj/Ad/Prep**	7	23	**Ad/Intj**	1
8	**N/Aj/Ad/Prep**	5	24	**Ad/Pron**	1
9	**Pron/N**	4	25	**Verb/Intj/Abbr**	1
10	**N/Intj**	4	26	**Conj/Aj**	1
11	**Prep/Aj**	3	27	**N/Ad/Conj**	1
12	**Ad/Prep/N**	3	28	**Aj/Ad/Verb**	1
13	**N/Ad/Intj**	3	29	**N/Aj/Ad/Prep/Conj**	1
14	**Prep/Conj**	3	30	**Aj/Prep/Conj**	1
15	**N/Aj/Intj**	2	31	**N/Aj/Ad/Intj**	1
16	**Aj/Ad/Intj**	2	32	**Aj/Ad/Prep/Intj**	1

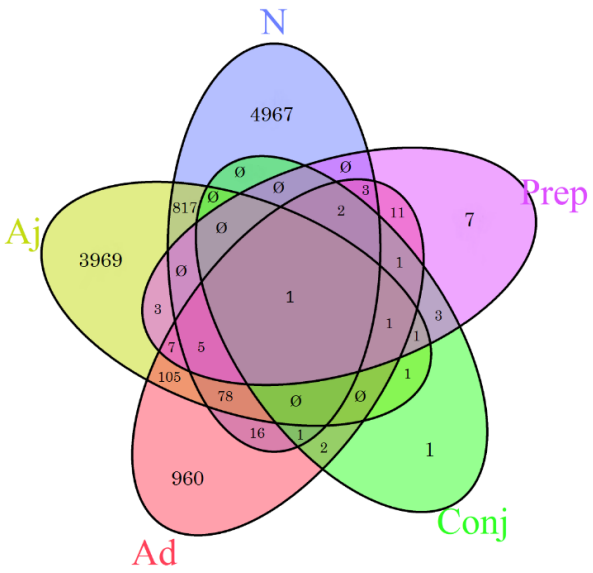


Figure 5.52. The Venn diagram of multimorphemic conversive classes
(Ø represents an empty set)

5.3.7 The general trends in the formation of the multimorphemic words

In the sections above, the structure of the morphological metacorpus has been analyzed in detail. This section seeks to summarize the discussed findings and to identify the general trends in the organization of the metacorpus.

Figure 5.53 provides helpful insight into a major difference between word classes which is rooted in their morphological diversity. In this figure, the numbers of words (e.g. *uproariness*), morphological patterns (e.g. N'=BM+ness) and constructions (e.g. {C+ness}) in different classes are given. Adverbs have the lowest number of morphological constructions and patterns. However, they produce the third highest number of types in the metacorpus (after nouns and adjectives). This large discrepancy between the number of morphological constructions and that of words produced by these constructions suggests that, in the context of word formation, adverbs display a lower degree of morphological diversity. By contrast, conversive classes (i.e. N/Aj and Aj/Ad) show the highest morphological diversity, evidenced by a higher number of morphological constructions and a lower number of word types. Furthermore, adverbs are not only less morphologically diverse, but they also have a larger number of word types at their second structural level, as mapped against the first level (Figure 5.50).

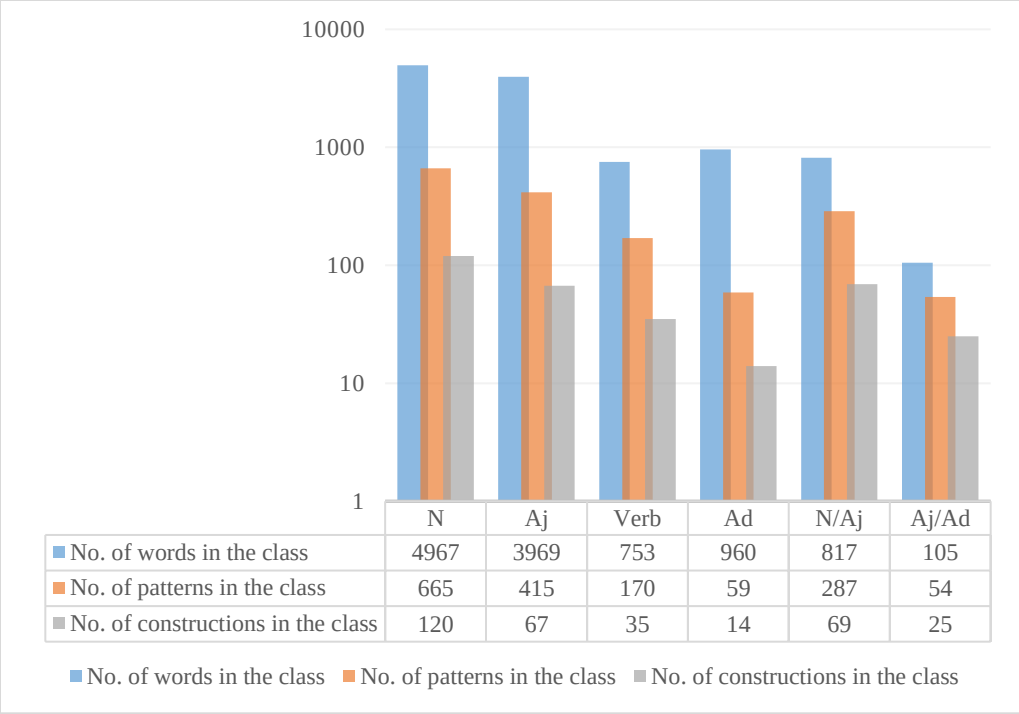


Figure 5.53. The proportions of words, morphological patterns and constructions across major word classes (for multimorphemic words; a logarithmic scale has been applied)

The second structural level has also shown a different property for adjectives. Similar to the overall structure of conversive classes, it has a significantly higher degree of morphological diversity (see Figure 5.49), which is largely accounted for by prefixation constructions. Further, some of them even tend to have a higher type frequency on the second level of formation (with {un+C}" being the most pronounced example). This property is less obvious in nouns and is absent in other word classes, where the type frequency of the first-level morphological constructions is consistently higher than that of the second-level constructions.

On the other hand, with the only two levels of formation, verbs demonstrate less structural complexity than other word classes. This feature may be linked to the fact that half of the simple verbs (on the zero-level of derivation) are formed by conversion, which withers the derivational function of this class. Another distinguished feature of verbs is that their median type valency is 2, whereas that of all other classes is 1. The higher median type valency of verbs indicates that a larger number of verbal affixes attach to two or more types of bases and that there are a smaller number of monovalent verbal affixes.

Table 5.25 presents the morphological constructions of the six major word classes that occupy the first fourteen ranks in the type-frequency list of the metacorpus. It can be generalized

that compounding is a universal and the most formally productive word-formation process in English, because it is highly represented in all word classes. Other important word-formation processes involve the affixes *-er, -ed, -ize/ise, -ly, -ness, -able, -y, -al, -less, re-* and *de-*.

Table 5.25. The first fourteen ranks in the type-frequency list of constructions across six major word classes

r	N	Type fr	Aj	Type fr	Verb	Type fr	Ad	Type fr	N/Aj	Type fr	Aj/Ad	Type fr
1	**{C+er}**	977	**{C+ed}**	1743	**{C+ize}**	148	**{C+ly}**	909	**{C+C}**	126	**{C+less}**	12
2	**{C+ing}**	948	**{C+ing}**	701	**{re+C}**	125	**{C+C}**	16	**{C+al}**	106	**{C+C}**	14
3	**{C+C}**	837	**{C+able}**	336	**{C+en}**	66	**{un+C}**	12	**{C+ed)**	65	**{C+ing}**	10
4	**{C+ness}**	327	**{C+y}**	293	**{un+C}**	40	**{a+C}**	9	**{C+ist}**	62	**{C+ly}**	9
5	**{CC}**	174	**{C+al}**	249	**{C+le}**	39	**{C+s}**	3	**{C+an}**	54	**{a+C}**	8
6	**{C+ism}**	143	**{un+C}**	225	**{C+C}**	38	**{C+ous}**	2	**{C+y}**	48	**{C+ful}**	8
7	**{C+ment}**	137	**{C+C}**	184	**{de+C}**	38	**{in+C}**	2	**{un+C}**	35	**{un+C}**	8
8	**{C+itis}**	135	**{C+less}**	178	**{mis+C}**	33	**{C+al}**	1	**{C+able}**	34	**{C+y}**	7
9	**{C+ion}**	116	**{C+ful}**	115	**{dis+C}**	28	**{C+er}**	1	**{C+ive}**	27	**{C+ish}**	4
10	**{C+ist}**	100	**{C+ic}**	80	**{C+ate}**	27	**{C+ish}**	1	**{C+ic}**	25	**{C+s}”**	4
11	**{C+age}**	87	**{C+ive}**	73	**{en+C}**	24	**{C+ward}**	1	**{im/in+C}**	22	**{C+ous}**	3
12	**{C+y}**	65	**{C+ous}**	68	**{C+er}**	19	**{C+wise}**	1	**{C+ly}**	17	**{in+C}**	2
13	**{C+or}**	53	**{C+ish}**	60	**{out+C}**	15	**{CC}**	1	**{C+ery}**	15	**{anti+C}**	1
14	**{C+ship}**	53	**{C+ly}**	46	**{C+ify}**	11	**{up+C}**	1	**{C+ish}**	13	**{C+able}**	1

Finally, another interesting observation concerns conversive classes. As substantiated by the Venn diagram in Figure 5.52, many multimorphemic word classes overlap. However, there are also empty sets in some areas, largely in the intersection of nouns, prepositions and conjunctions. Although it is difficult to make any definitive conclusion as to the meaning of this pattern of the empty sets, one possible explanation can be that there is a more pronounced distinction between nouns, as a ‘universal and fundamental’ (Langacker 2007: 96) category designating ‘a type of thing’ (Langacker 2007: 265), and prepositions and conjunctions, as expressing ‘nonprocessual relationships’ (Langacker 2007: 100). This distinction is more blurred for adjectives and adverbs.

5.4 The structural description of English word-formation

In summary, this chapter has presented a fine-grained structural analysis of the morphological metacorpus. Here, the main findings of this analysis are summarized and some characteristics of English word formation are contrasted with those of Persian word formation, the data for which are reported in Krykoniuk (2014, 2020). Contrasting these languages is particularly interesting,

since on a more general perspective, both languages can be characterized as relatively isolating (or, in Sapir (1921: 135) terms, analytic, i.e., implying that they combine concepts into single words economically) with some inflective and agglutinative morphological features. Their isolating features include a strict order of words in a sentence (although, in spoken variants of Persian, the word order is more flexible), the absence of the grammatical categories of case and gender, as well as the absence of noun and adjective inflection. Furthermore, inflection in these languages is realized mainly in the grammatical system of verbs, and agglutinative features are, on the whole, specific to plural suffixes. Thus, while the grammatical structures of Persian and English is known to have a lot in common, their derivational morphology is less studied.

Simplexes, which constitute a major part of the metacorpus (63%), are more frequent in nouns, verbs, as well as grammatical and conversive classes, and less frequent in adverbs and adjectives. The major word class for simplexes is noun (47%), followed by verb (23%), adjective (17%) and conversive classes (10%). Simple adverbs and grammatical classes produce 3% and 0.4% of simplexes, respectively. As reported in Krykoniuk (2014: 65), for a Persian metacorpus, the simplexes have a larger portion (70%) than English, with 83% of nouns and 64% of adjectives being simplexes. There are no simple verbs in Persian. Similar to English, simplexes form a small share in adverbs (0.46% of all simple words and 30% of all adverbs). Hence, adverbs in Persian and English are predominantly multimorphemic, which suggests that in adverbial derivation the derivativeness plays a greater role, as compared to other word classes.

The major part of simplexes in word classes are borrowings from other languages (largely from French, Latin and Scandinavian), except for adverbs (7%) and verbs (24%), which prioritize conversion (36% and 50%, respectively) as a major source of word formation. A similar picture is observed in Persian word formation, where there is a large amount of borrowed nouns and adjectives from Arabic (68% in nouns and nearly 80% in adjectives). Contraction is most frequent for nouns and is fully unproductive for adverbs which, in turn, own the largest share of words from the Germanic component. In view of different types of contraction, back-formation is more common in verbs. Further, phonological alternation is a more pronounced formation process in adverbs and grammatical classes, and is less pronounced in verbs and adjectives, whereas onomatopoeia is common in nouns/interjections (22%) and verbs (2%). Lastly, semantic formations make up the smallest portion of simplexes.

A zero-level morphological analysis of simplexes has revealed that French is the major source of such affixes as *-ment*, *-er*, *-ance*, *-age*, *re-*, and *dis-*, and Latin of such affixes as *-um, -ia -ate, -in, de-, ate, -ive* and *-al*. The affixes *-ion*, *-ity*, *-y*, *-ence*, *-or, de-*, *-ous* and *-ble/able* are equally represented in French and Latin borrowings.

Less than 40% of words in the metacorpus are multimorphemic. Multimorphemic nouns involve the highest number of structural levels (four), whereas verb formation occurs only in two levels, with prefixation being the most productive derivation process on the second level. Another distinguished feature of verb formation is that its median type valency in suffixation is 2, which is either evidence for a greater combining power of verb suffixes or a more distinct role of the word bases of nouns and adjectives in a root slot of the suffixation construction within this word class. Conversive classes show the highest degree of morphological diversity, whereas adverbial formation is the least morphologically diverse. The second structural level has a prominent role in adverbs (i.e. by showing a high number of types) and in adjectives formed by prefixation (i.e. a high number of constructions). Furthermore, over 30% of all morphological constructions are hapaxes, and among recurrent constructions, the most universal is {C+C} which represents compounding. Lastly, the structural analysis has brought to light two major frequency effects. The first concerns the orthographical/morphonological changes and the specificities of word formation, which are more distinct in morphological constructions with a higher type frequency, and the second is the impact of type frequency of a morphological construction on the type valency of its elements. The next chapter looks at morphological regularities, in accordance with which elements in the presented constructions are organized, as well as at the formal paradigms defined by these regularities.

6 Formal morphological regularities and paradigms

The previous chapters, describing the metacorpus and revealing its different quantitative characteristics, have built a solid foundation for the identification of the different morphological regularities, patterns and paradigms in English word formation. This chapter and the next are the pinnacle of this study. Whereas the next chapter looks at the statistical trends in English word formation, the current chapter answers the questions related to English formal morphological regularities, patterns and paradigms (RQ2), and what they reveal about the English language and its typological features (RQ5). It also looks at the word-formation level structure of word classes in greater detail and identifies which word bases play the most important role in which word classes.

The formalism of my approach defines the nature of the discussed regularities and paradigms: they are defined by their forms, excluding the semantic relationship between their constituents. The future perspective of this research, hence, is the study of how meaning is mapped against these formally identified structures and how semantics and morphology correlate. In order to arrive at the 'condensed' metalinguistic abstractions, presented in this chapter (see also Section 3.1 for the discussion of different levels of generalizations in the formal morphological analysis: i.e. morphological patterns, morphological constructions and meta-constructions), the method of matrix optimization has been deployed, which is new to the study of morphology. It involves shuffling columns and rows of a matrix to reach the most optimal state, where its elements are structured as close to each other as possible. Further, for the description of the paradigms, graph theory networks are used (see Section 4.5). In the pages that follow, Sections 6.1 introduces formal morphological regularities in multimorphemic nouns, verbs, adjectives, as well as grammatical and conversive classes. Section 6.2, then, presents the major formal morphological paradigms (involving the highest number of items in each construction) across different word classes, and Section 6.3 engages with a detailed analysis of levels of word formation. Lastly, Section 6.4 highlights the main findings of this chapter.

6.1 Formal morphological regularities for multimorphemic nouns

In this section, the formal morphological regularities of multimorphemic nouns are presented according to their word-formation levels. Section 6.1.1 summarizes noun formation at the first

level, and Sections 6.1.2 and 6.1.3 at the second, third and fourth levels, and Section 6.1.4 highlights the main trends in noun formation. Then, Sections 6.1.5–6.1.7 look at verb formation on two levels, Sections 6.1.8–6.1.11 at adjective formation and Sections 6.1.12–6.1.15 at adverbial formation on three levels. Finally, Sections 6.1.16–6.1.20 introduce morphological patterns and regularities in the major conversive classes N/Aj and Aj/Ad. The following routine has been adopted throughout these sections: first, the optimized matrices are introduced which visualize the combinatorial properties of morphemes within constructions of different word classes. The middle cells of matrices (representing the mediale) are coloured such that their visual processing is easier. Then, these properties are described as morphological regularities in the form of tables, supplied with examples for each regularity.

6.1.1 Multimorphemic nouns: the first level

For the identification of the morphological regularities, matrix optimization has been applied. As described in Section 3.3.2 on methodology (Chapter 3), this method presupposes three formal slots in a construction, termed 'initiale', 'mediale' and 'finale'. The first element in the construction (the initiale) is represented in the first column of a matrix, the last element (finale) in the upper row and the middle element (the mediale) in the central rows. For example, in Figure 6.7, the first element of the first column is the prefix *de-*, which combines with the mediales N, BM, Verb (presented in the second column and the second, third and fourth rows of the matrix, respectively) and with the finale *-ion* (presented in the first row of the matrix) to produce the morphological patterns N"=de+N+ion, N"=de+BM+ion and N"=de+Verb+ion, which form such words as *deforestation* and *decipheration* (with the N root), *dehydration* (with the BM root) and *de-escalation* and *demobilization* (with the Verb root). Thus, there is a three-slot limitation in a matrix, which is overcome with the adjustment of the number of items in the slot of the mediale: for the two-slot constructions, the mediale is considered to be an empty slot (formalized as Ø), whereas for the four-slot or higher-number constructions, the mediale slot is assumed to contain two or more items—that is, all morphemes between the initiale and finale. It is also worth mentioning that the identified regularities are not absolute but reflect the word-formation morphology in this study's sample of 32,000 words.

Lastly, there is a difference between a zero morpheme (Ø) and an empty cell of the matrix. The former indicates an empty slot of a construction, showing that the combination of elements has been observed in the sample, where the latter may be considered as a structural zero, suggesting

that the combinations of morphemes are absent from the sample (and, maybe, from the language). For example, in the matrix of Figure 6.2 (Part 1), the suffix *-hood* (represented in the list of finales in the first row of the matrix) combines with nouns and adjectives (represented in the initiale slot in the first column of the matrix). These combinations are marked with a zero morpheme in a corresponding mediale slots. However, the remaining empty slots in the mediale column for *-hood* (i.e. Verb, BM, N/Verb, Ad, N/Aj/Ad, Aj/N, Aj/Ad and Verb/Aj) indicates that the combining power of this suffix is limited to these two word classes (N, Aj) and that it does not occur with other types of bases.

6.1.1.1 The first-level morphological construction {C-Ø-a} or {C-a}

The first-level construction is one of the most formally productive constructions in English in terms of the number of items which are used in its slots (for the definition of formal productivity, see the last paragraph in Section 3.6.1). The inner composition of this construction is visualized in 6.1–6.4.

The initiale of this construction is filled with 10 word classes (with CC) or with 5 classes (without CC)[29], and the finale with 83 suffixes (or 79 if *-ry*, *-ary* and *-ery*, as well as *-ence*, *-ance*, *-ency*, *-ancy*, *-cy* and *-acy*, are considered allomorphs). From the matrices in Figures 56–59, it can be inferred that the largest type valency for initiale is observed for nouns (56), bound morphemes (47), verbs (33) and adjectives (33). In contrast, adverbs have the lowest type valency. Further, the finale suffixes *-ee*, *-ess*, *-let*, *-et*, *-ry*, *-dom* and *-ary* do not combine with adverbs, nouns/adjectives/adverbs, nouns/adjectives, adjectives/adverbs and verbs/adjectives. The monovalent finales in this construction include the following suffixes: *-dom*, *-ary*, *-ty*, *-cy*, *-er*, *-ide*, *-ite*, *-end*, *-ar*, *-ade*, *-cade*, *-ina*, *-ock*, *-t*, *-t*$_2$, *-one*, *-itis*, *-ah o*, *-lic*, *-red*, *-on*, *-ac*, *-el*, *-eme*, *-i*, *-eroo*, *-wards*, *-osis*, *-sy*, *-il*, *-yl*, *-lock*, *-ol*, *-one*, *-oid*, *-kin* and *-ory*. The highest type valency has been observed for the suffix finales *-ness* (8), *-ing* (6), *-er* (6), *-ism* (6), *-age* (6), *-ment* (5), *-ery* (5) and *-y*. Finally, there are no monovalent suffixes that attach to adjectives. All suffixes that attach to adjectives are either duo- or polyvalent.

[29] CC stands for a 'conversive class'. For the explanation of the measure of the type valency with or without conversive classes, see p.115–116. This distinction is interesting for identifying the 'conversiveness' of the type valency in constructions.

C/a	ness	ing	er	ism	age	ity	ist	ment	ery	y	ee	ion	ess	et	let	hood	ship	ry	dom	ful	ary	or	al	ty
N	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø			
Verb	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø								Ø	Ø	
Aj	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø			Ø		Ø	Ø	Ø						
BM	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø		Ø								Ø	Ø	
N/Verb		Ø	Ø		Ø			Ø	Ø	Ø														
Ad	Ø	Ø			Ø																			
N/Aj/Ad	Ø		Ø	Ø																				
Aj/N	Ø			Ø		Ø	Ø																	Ø
Aj/Ad	Ø					Ø																		
Verb/Aj												Ø												

Figure 6.1. The matrix for the morphological construction {C-Ø-a} (Part 1)

C/a	ence	ance	ency	ancy	in	acy	ine	s	ian	o	ate	s-pl	le	ling	ie	ster	ette	cy	eer	th	ant	ure	ide	ite
N	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø					
BM	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø							Ø	Ø	Ø	Ø	Ø
Aj	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø			Ø	Ø	Ø				Ø				
Verb	Ø	Ø	Ø	Ø									Ø		Ø	Ø	Ø			Ø	Ø	Ø		
N/Verb													Ø											
N/Aj														Ø	Ø									

Figure 6.2. The matrix for the morphological construction {C-Ø-a} (Part 2)

C/a	end	ar	ade	t2	our	cade	ium	ard	ina	ock	t	ome	itis	aholic	red	on	ac	ane
Verb	Ø	Ø	Ø	Ø	Ø													
N						Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø					
N/Verb														Ø	Ø			
BM						Ø	Ø									Ø	Ø	Ø
Aj								Ø										

Figure 6.3. The matrix for the morphological construction {C-Ø-a} (Part 3)
(Key: The suffix *-t* is a Germanic suffix (e.g. *thrift*), whereas the suffix *-t2* is an unproductive morpheme which has been formed by analogy as in, for example, *catalyst* created from *catalysis* by analogy to *analyst*)

C/a	el	eme	i	eroo	wards	osis	ia	ese	sy	il	yl	lock	ol	one	oid	kin	ory
N	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø									
BM									Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
Aj							Ø	Ø									

Figure 6.4. The matrix for the morphological construction {C-Ø-a} (Part 4)

In Table 6.1 and 6.2 the morphological regularities of the noun construction {C-Ø-a} identified on the basis of the matrices above are presented. It shows combinatory properties of morphemes within this construction and provides an example for each combination in a sequence.

Table 6.1. Morphological regularities for the noun construction {C-Ø-a} on the first level: NC1_{C-Ø-a} (Part 1)

No	Type in the slot	Attach(es) to	Examples in the data
1	Noun initiale	*-ness, -ing, -er, -ism, -age, -ity, -ist, -ment, -ery, -y, -ee, -ion, -ess, -let, -et, -hood, -ship, -ry, -dom, -ful, -ary, -ence, -ance, -ency, -ancy, -in, -acy, -ine, -s, -ian, -o, -ate, -s-pl, -le, -ling, -ie, -ster, -ette, -cy, -eer, -ium, -ard, -ina, -ock, -t, -ome, -itis, -el, -eme, -i, -eroo, -wards, -osis, -ia, -cade* and *-ese*	*womanness, airing, banker, agism, peerage, mobility, fetishist, atonement, drinkery, bushy, cookee, cricketer, pigmentation, shepherdess, sparklet, hornet, manhood, clientship, heraldry, countdom, officeful, sugary, brilliance, naugatine, politics, electrician, news, dolphinarium, bollard, concertina, paddock, electioneer, yobbo, professorate, starling, catalyst, psychosis, motifeme, smesheroo, journalese, motorcade*
2	Verb initiale	*-ness, -ing, -er, -ism, -age, -ity, -ist, -ment, -ery, -y, -ee, -ion, -ess, -et, -ence, -ance, -ency, -ancy, -le, -ie, -ster, -ette, -th, -ant, -ure, -end,-ar, -ade ,-our*	*adaptness, deserving, screamer, zanyism, leakage, femininity, recordist, employment, smashery, entreaty, payee, insulation, murderess, snippet, convergence, shuttle, movie, lobster, launderette, growth, claimant, seizure, adherend, registrar, blockade, demeanour*
3	Adjective initiale	*-ness, -ing, -er, -ism, -age, -ity, -ist, -ment,-ery, -y, -ee, -ence, -ance, -ency, -ancy, -in, -acy, -ine, -s, -ian, -o, -ate, -ard, -ia and -ese*	*emptiness, rowing, deader, absurdism, adultage, originality, visualist, merriment, finery, goody, redundantee, occurrence, activing, adrenalin, acoustics, magician, wrongo, priorate, drunkard, septicaemia, legalese*
4	Bound morpheme initiale	*-ness, -ing, -er, -ism, -age, -ity, -ist, -ment, -ery, -y, -ee, -ion, -et, -ence, -ance, -ency, -ancy, -in, -acy, -ine, -s, -ian, -o, -ate, -or, -ium, -sy, -il, -yl, -lock, -ol, -one, -oid, -kin* and *-ory*	*uproariness, morphing, soccer, hedonism, petrolage, laxity, florist, attachment, haberdashery, eulogy, nominee, sanitation, punnet, irreverence, penicillin, glycerine, logistics, Paralympian, lingo, incubate, delegator, sodium, pixie, quantile, vinyl, wedlock, xylol, silicone, steroid, napkin, observatory*
5	Ad initiale	*-ness, -ing* and *-age*	*soonness, offing, outage*
6	N/Verb initiale	*-ing, -er, -age, -ment, -ery, -y, -le, -aholic* and *-red*	*lettering, molder, taskage, basement, cookery, slushy, snarl, shopaholic, hatred*
7	N/Aj initiale	*-ness, -ism, -ity, -ist, -ty, -ling* and *-ie*	*idleness, idealism, toxicity, finalist*
8	Aj/Ad initiale	*-ness* and *-ity*	*rashness, spirality*
9	N/Aj/Ad initiale	*-ness, -er* and *-ism*	*easiness, tenner, immediatism*

Table 6.2. Morphological regularities for the noun construction {C-Ø-a} on the first level: NC1_{C-Ø-a} (Part 2)

No	Monovalent finales	Attach to
10	*-end, -ar, -ade* and *-cade*	verbs
11	*-dom, -ery, -cy, -eer, -ina, -ock, -t, -t₂, -one, -itis, -el, -eme, -i, -eroo, -wards* and *-osis*	nouns
12	*-ide, -ite, on, -ac, -one, -sy, -il, -yl, -lock, -ol, -one, -oid, -kin* and *-ory*	bound morphemes

6.1.1.2 The first-level morphological construction {a-Ø-C} or {a-C}

The second productive morphological construction on the second level (see Section 3.2.3 for the definition of the level structure of the metacorpus) is {a-Ø-C} or {a-C} which is introduced in Figure 6.5. In this construction, the initiale is filled with 36 prefixes, and the finale with 7 word classes (with CC) or 4 classes (without CC). The noun finale has the highest type valency (34), and the adjective and conversive classes the lowest (1). It can be generalized that, in noun formation, prefixes tend to attach to nouns. The morphological regularities for this construction are presented in Table 6.3.

a/C	N	BM	N/Verb	Verb	Aj	N/Aj	N/Intj
re	Ø	Ø	Ø	Ø			
con		Ø					
peri		Ø					
on	Ø	Ø					
anti	Ø	Ø					
in	Ø			Ø			
non	Ø				Ø		
fore	Ø					Ø	Ø
dis	Ø						
mis	Ø						
sub	Ø						
counter	Ø						
un	Ø						
pre	Ø						
semi	Ø						
super	Ø						
co	Ø						
inter	Ø						
up	Ø						
de	Ø						
mal	Ø						
self	Ø						
over	Ø						
para	Ø						
ultra	Ø						
ac	Ø						
after	Ø						
arch	Ø						
contra	Ø						
em	Ø						
im	Ø						
infra	Ø						
out	Ø						
sur	Ø						
trans	Ø						

Figure 6.5. The matrix for the morphological construction {a-Ø-C}

Table 6.3. Morphological regularities for the noun construction {a-Ø-C} on the first level: NC1_{a-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Monovalent prefix initiales *dis-*, *mis-*, *sub-*, *counter-*, *un-*, *pre-*, *semi-*, *super-*, *co-*, *inter-*, *up-*, *de-*, *mal-*, *self-*, *over-*, *para-*, *ultra-*, *ac-*, *after-*, *arch-*, *contra-*, *em-*, *im-*, *infra-*, *out-*, *sur-* and *trans-*	N	*disability, miscalculation, subculture, countercharge, unrest, precondition, semingod, superbug, co-pilot, interplay, upset, defusion, malpractice, self-concept, overtone, paratroop, ultrasound, accomplice, aftermath, arch-enemy, contraflow, empathy, imbalance, infrastructure, outward, surname, transbus*
2	Duovalent prefix initiale *on-* and *anti-*	N and BM	*onslaught, onset, anti-hero, antiperspirant*
3	Duovalent prefix initiale *in-*	N and Verb	*instep, inlet*
4	Duovalent prefix initiale *non-*	N and Aj	*nonentity, non-resident*
5	Polyvalent prefix initiale *re-*	N, BM, Verb, N/Verb	*rebirth, reflation, relay, reassurance*
6	Polyvalent prefix initiale *fore-*	N, N/Intj, N/Aj	*forefather, forename, foreword*
7	Noun finale	*re-*, *on-*, *anti-*, *in-*, *non-*, *fore-*, *dis-*, *mis-*, *sub-*, *counter-*, *un-*, *pre-*, *semi-*, *super-*, *co-*, *inter-*, *up-*, *de-*, *mal-*, *self-*, *over-*, *para-*, *ultra-*, *ac-*, *after-*, *arch-*, *contra-*, *em-*, *im-*, *infra-*, *out-*, *sur-* and *trans-*	
8	Bound morpheme finale	*re-*, *con-*, *peri-*, *on-* and *anti-*	
9	Verb finale	*re-* and *in-*	
10	Adjective finale	*non-*	
11	N/Verb finale	*re-*	
12	N/Aj finale	*fore-*	
13	N/Intj finale	*fore-*	

6.1.1.3 The first-level morphological construction {C-Ø-C} or {C-C}

The last morphological construction on the first level is introduced in Figure 6.6. Since its elements are word classes, the construction is limited in its type valency. The initiale slot in this construction is filled with 11 (with CC) or 9 (without CC) word classes, and the finale slot with 9 (with CC) and 7 (without CC) word classes. The most type-valent initiale and finale is a noun (with a type valency of 6 and 7, respectively). The least type-valent initiale is a preposition (1) and the least type-valent finale is a noun/interjection/adverb (1), and the least type-valent finales are pronouns, nouns/interjections/adverbs and conjunctions. Finally, there is no adjective in the position of the finale. Table 6.4 summarizes morphological regularities for this construction.

C/a	N	BM	Ad	Pron	N/Intj/Ad	N/Aj	Verb	Conj
Verb	Ø	Ø	Ø	Ø				
Aj	Ø	Ø	Ø		Ø			
N	Ø	Ø	Ø			Ø	Ø	
BM	Ø	Ø						
Ad	Ø						Ø	Ø
N/Verb	Ø					Ø		
N/Aj/Num	Ø							
Aj/N	Ø							
Pron	Ø							
Pron/Aj			Ø					
Prep							Ø	

Figure 6.6. The matrix for the morphological construction {C-Ø-a}[30]

Table 6.4. Morphological regularities for the noun construction {C-Ø-C} on the first level: NC1_{C-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Verb initiale	N, BM, Ad, Pron	*hovercraft, Dictaphone, breakdown, holdall*
2	Aj initiale	N, BM, Ad, N/Intj/Ad	*busybody, simpleton, close-up, sweetheart*
3	N initiale	N, BM, Ad, N/Aj, Verb	*ashtray, radarscope, cast-off, oatcake, stonehatch*
4	BM initiale	N, BM	*autopilot, astronaut*
5	Ad initiale	N, Verb, Conj	*inroad, intake, nor*
6	N/Verb initiale	N, N/Aj	*creep-hole, scapegoat*
7	N/Aj/Num initiale	N	*hundredweight*
8	N/Aj initiale	N	*smartweed*
9	Pron initiale	N	*somebody*
10	Pron/Aj initiale	Ad	*whatsit*
11	Prep initiale	Verb	*to-do*

6.1.2 Multimorphemic nouns: the second level

There are five morphological constructions on the second level of noun derivation: {a-C-a}, {C-a-a}, {C-C-a}, {C-a-C} and {C-C-C}. The following subsections provide a detailed morphological description for each of these constructions.

6.1.2.1 The construction {a-C-a}

The largest morphological construction (in terms of the number of formal items that can occur in its slots) on the second level is the morphological construction {a-C-a}, which is visualized in

[30] The instructions on how to read matrices are given on p.129.

Figure 6.7. In what follows, the description of this construction and its morphological regularities (Table 6.5) are given.

The initiale of this construction is occupied by the following 26 prefixes (25 if *en-* and *em-* are considered as allomorphs): *de-*, *re-*, *dis-*, *counter-*, *en-*, *fore-*, *per-*, *out-*, *pro-*, *be-*, *em-*, *mis-*, *over-*, *a-*, *un-*, *inter-*, *co-*, *on-*, *in-*, *mal-*, *anti-*, *up-*, *contra-*, *con-*, *sub-*, and *non-*. The finale slot of this construction is open to the following 18 suffixes (16 if *-ance*, *-ency* and *-ence* are considered as allomorphs): *-ion*, *-er*, *-ing*, *-ment*, *-ist*, *-al*, *-y*, *-able*, *-ery*, *-or*, *-ance*, *-ant*, *-ess*, *-acy*, *-ency*, *-ship*, *-ness* and *-ence*. The most polyvalent mediale slot (meaning its potential to combine with finales) occurs in the combination with the prefixes *re-*, *dis-*, *de-* and *en-* in the slot of the initiale. The most frequent type for the mediale is a verb, which occurs in 43 combinations. The initiales *couter-*, *per-*, *out-*, *pro-*, *be-*, *over-*, *inter-*, *mal-*, *contra-*, *con-*, *sub-* and *non-* are monovalent: they attach to one mediale and one finale. In other words, each suffix has only one morphological pattern. Finally, a mediale adjective occurs only for the prefixes which are duo- or polyvalent. It involves the following patterns: N”=re+Aj+al, N”=en+Aj+in, N”=un+Aj+ness, N”=un+Aj+ing, N”=in+Aj+acy and N”=in+Aj+ancy.

a/a	ion	er	ing	ment	ist	al	y	able	ery	or	ance	ant	ess	acy	ency	ship	ness	ence
de	N(at)	N	N	N								BM	N					
	BM		Verb															
	Verb																	
re	Verb(at)	Verb	Verb	Verb	Verb	Aj	Verb	Verb	Verb									
		N																
dis	N	Verb	Verb	Verb		Verb				Verb	Verb							
	Verb																	
counter	Verb																	
en		N	N	N							Verb							
			Aj	N/Verb														
fore		N	Verb															
		Verb																
per		Verb																
out		N																
pro		N																
be			Verb															
em			N	N														
mis			Verb	Verb														
over			Verb															
a			Verb	N		Verb												
un			Verb	Verb												Verb	Aj	
			N															
			Aj															
inter	Verb																	
co		Verb																Verb
on		Verb	Verb															
in			Verb											Aj	Aj			
mal				Verb														
anti					N							Verb						
up		N	Verb			Verb												
contra	BM																	
con	BM(at)																	
sub			N/Verb															
non		Verb																

Figure 6.7. The matrix for the morphological construction {a-C-a}

Table 6.5. Morphological regularities for the noun construction {a-C-a} on the second level: NC2_{a-C-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	*de-*	*-ion*, *-er*, *-ing*, *-ment*, *-ant* or *-ess*	is monovalent for the finales *-er*, *-ment*, *-ant* and *-ess*, is duovalent for the finale *-ing* and is polyvalent for the finale *-ion*	*de-icer, decipheress, deodorant, decipherment, defatting, debunching, de-escalation, demotion, demotivation*
2	*re-*	*-ion*, *-er*, *-ing*, *-ment*, *-ist*, *-al*, *-y*, *-able* and *-ery*	is largely monovalent and is occupied with a verb, except for the finale *-er*, which is duovalent and is occupied with a mediale of a noun or verb. The mediale for the suffix finale *-al* is adjective	*re-former, resounding, redeployable, rediscovery, refashionment, recyclist, renewal, refinery*
3	*dis-*	*-ion*, *-er*, *-ing*, *-ment*, *-al*, *-or*, and *-ance*	is a verb, except for the finale *-ion*, for which it is occupied by a verb and a noun	*disapproval, disintegration, disengager, disliking, disentanglement, disintegrator, disappearance*
4	*out-* and *pro-*	*-er*	is a noun	*outsider, pro-lifer*
5	*contra-* and *con-*	*-ion*	is a bound morpheme	*contraception, conurbation*
6	*counter-*, *per-*, *be-*, *mis-*, *over-*, *inter-*, *co-*, *on-*, *mal-* and *non-*		is a verb	*counteraction, peruser, bewildering, mistreatment, overwhelming, interaction, co-owner, onlooker, maltreatment, non-starter*
7	*en-*, *fore-*, *per-*, *out-*, *pro-*, *be-*, *mis-*, *over-*, *a-*, *un-*, *inter-*, *co-*, *on-*, *in-*, *mal-*, *anti-*, *up-*, *contra-*, *con-*, *sub-*, *non-*, *conter-* and *de-*	*-y*, *-able*, *-ery* and *-or*	does not occur together in the sample	
8	*de-*, *re-*, *dis-*, *counter-*, *en-*, *fore-*, *per-*, *out-*, *pro-*, *be-*, *em-*, *mis-*, *over-*, *a-*, *inter-*, *on-*, *up-*, *contra-*, *con-*, *sub-* and *non-*	*-acy*, *-ency*, *-ship*, *-ness* and *-ence*	does not occur together in the sample	
9	*de-*, *counter-*, *en-*, *fore-*, *per-*, *out-*, *pro-*, *be-*, *em-*, *mis-*, *over-*, *a-*, *inter-*, *co-*, *on-*, *in-*, *mal-*, *contra-*, *con-*, *sub-* and *non-*	*-ist*, *-al*, *-y*, *-able*, *-ery* and *-or*	does not occur together in the sample	

6.1.2.2 The construction {C-a-a}

The second largest morphological construction on the second level as regards the number of types that can be placed in its slots is {C-a-a}. It is visualized in Figure 6.8, and its morphological regularities are listed in Table 6.6.

The initiale of this noun construction is filled with five word classes: nouns, verbs, adjectives, nouns/adjectives and bound morphemes. In other words, the type frequency of its initiale is 4 (without CC) and 5 (with CC). The construction's finale allows for the attachment of the suffixes *-s*, *-ism*, *-er*, *-ing*, *-ity*, *-ion*, *-ist*, *-ship*, *-hood*, *-ry*, *-a*, *-ine*, *-or*, *-y*, *-s(pl)*, *-ness*, *-dom* and *-ess*. Further, the highest type valency for the initiale of this construction to attach to the types

of suffixes has been observed in nouns (17 types), followed by verbs (14 types). The highest type attachability of the finales to the types of suffixes is observed for the suffixes *-ism* and *-er* in combination with the noun finale (the type valency of 6 and 5, respectively), *-ing* and *-er* in combination of the verb initiale (4 and 2, respectively) and *-ness* in combination with adjective finale (9). The bound-morpheme initiale and the finale suffixes *-s*, *-ism* and *-ity* are mediated with the help of the mediale suffixes *-ic*, *-al*, and *-(ic)al*, respectively. Further, the suffix *-er* is the most frequent mediale, if the initiale is a verb. Lastly, the suffix *-ness* shows the highest type valency for the finale (13).

C/a	s	ism	er	ing	ity	ion	ist	ship	hood	ry	a	ine	or	y	s-pl	ness	dom	ess
	ic	an	en	eer	al	ize(at)	ic	er	y	ist	(i)an	ol	ate	er	er	less		
		al	ize	en	(u)al	ate										y		
		ar	ock	ize	ar											ly		
		ent	le	er														
		ee	ing															
N		er																
			er	ize	able	ize(at)	ion	er	er					er	er	ed	er	er
			ize	eer	ive		al		ly							ive		or
				er												y		
				age												ing		
Verb																able		
				en		ize(at)										able		
																less		
																ish		
																ly		
																al		
																ing		
																y		
																ful		
Aj																some		
	ic	al	ish		al											ous		
BM					(ic)al													
N/Aj		al			al													

Figure 6.8. The matrix for the morphological construction {C-a-a}

Table 6.6. Morphological regularities for the noun construction {C-a-a} on the second level: NC2_{C-a-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	N	*-s*	*-ic-*	*aerobics*
2	N	*-ism*	*-an-, -al-, -ar-, -ent-, -ee-, -er-*	*Americanism, commercialism, nuclearism, decadentism, absenteeism, consumerism*
3	N	*-er*	*-en-, -ize-, -ock-, -le-, -ing-*	*hastener, victimizer, buttocker, sparkler, stockinger*
4	N	*-ing*	*-eer-, -en-, -ize-, -er-*	*orienteering, heartening, customizing, dustering*
5	N	*-ity*	*-al-, -(u)al-, -ar-*	*sentimentality, eventuality, molarity*
6	N	*-ion*	*-ize(at)-, -ate-*	*dramatization, formulation*
7	N	*-ist*	*-ic-*	*electronicist*
8	N	*-ship*	*-er-*	*dealership*
9	N	*-hood*	*-y-*	*worthihood*
10	N	*-ry*	*-ist-*	*dentistry*
11	N	*-a*	*-ian-*	*Canadiana*
12	N	*-ine*	*-ol-*	*gasoline*
13	N	*-or*	*-ate-*	*pollinator*
14	N	*-y*	*-er-*	*crockery*
15	N	*-s-pl*	*-er-*	*trousers*
16	N	*-ness*	*-less-*	*homelessness*
17	N	*-ness*	*-y-*	*snippetiness*
18	N	*-ness*	*-ly-*	*weatherliness*
19	Verb	*-er*	*-er-, -ize-*	*potterer, acclimatizer*
20	Verb	*-ing*	*-ize-, -eer-, -er-, -age-*	*acclimatizing, orienteering, rompering, packaging*
21	Verb	*-ity*	*-able-, -ive-*	*performability, adaptivity*
22	Verb	*-ion*	*-ize(at)-*	*acclimatization*
23	Verb	*-ist*	*-ion-, -al-*	*deflationist, removalist*
24	Verb	*-ship*	*-er-*	*leadership*
25	Verb	*-y*	*-er-*	*bakery*
26	Verb	*-s*	*-er-*	*pliers*
27	Verb	*-ness*	*-ed-, -ive-*	*accustomedness, adaptiveness*
28	Verb	*-ness*	*-y-*	*shimmeriness*
29	Verb	*-ness*	*-able-*	*advisableness*
	Verb	*-ness*	*-ing-*	*daringness*
30	Verb	*-dom* and *-ess*	*-er-* or *-or-*	*dealerdom, sailoress*
31	Aj	*-ing*	*-en-*	*fattening*
32	Aj	*-ness*	*-able-, -less-, -ish-, -ly-, -al-, -ing-, -y-, -ful-* or *-some-*	*accountableness, agelessness, childishness, loneliness, criticalness, daringness, hastiness, delightfulness, troublesomeness*
33	Aj	*-ion*	*-ize(at)-*	*femininization*
34	BM	*-er*	is *-ish-*	*Irisher*
35	BM	*-s, -ism or -ity*	*-ic-, -al-* or *-(ic)al-*	*aerobics, serialism, whimsicality*
36	BM	*-ness*	*ous*	*abstemiousness*
37	N/Aj	*-ity or -ism*	*-al-*	*tribalism, brutality*
38	Verb, Aj or BM	*-ry, -a, -ine* and *-or*	does not occur	
39	Verb	*-s, -ism, -ry, -or, -a* or *-ine*	does not occur	
40	N	*-ness, -dom* or *-ess*	does not occur	

6.1.2.3 The construction {C-C-a}

The next formally productive construction on the second level of noun formation is {C-C-a}. Its inner composition is presented in Figure 6.9, and its morphological regularities are given in Table 6.7.

The initiale of the morphological construction {C-C-a} is occupied by a noun, adjective, verb, bound morpheme, preposition, noun/adjective/numeral and adverb, and its finale by the following the following 19 suffixes (18 if *-ance* and *-ency* are considered allomorphs): *-ness*, *-s-pl*, *-er*, *-ing*, *-ery*, *-ship*, *-ment*, *-ance*, *-s*, *-ency*, *-or*, *-ist*, *-y*, *-ity*, *-an*, *-al*, *-ite*, *-ia* and *-on*. The highest attachability for the initiales in this construction is observed for nouns, adjectives and bound morphemes. All suffix finales in this construction allow only for a monovalent mediale, except for the suffixes *-er*, *-ing* and *-s*, which are duovalent.

Table 6.7. Morphological regularities for the noun construction {C-a-a} on the second level: NC2_{C-C-a}

No	**If**		**Then**	**Examples in the data**
	Initiale	***Finale***	***Mediale***	
1	N	*-s-pl*, *-ery*, or *-ship*	N	*gasworks, tomfoolery, kinsmanship*
2	N	*-ness*	Aj	*carsickness*
3	N	*-er* or *-ing*	N and Verb	*breadwinner, quarter-pounder, caretaker, heart-aching*
4	Aj	*-er*, *-ing* or *-ment*	Verb	*ill-treatment, latecomer, broadcasting*
5	Aj	*-s-pl*	N	*lazybones*
6	BM	*-ing*, *-or*, *-al*	Verb	*paragliding, microprocessor, withdrawal*
7	BM	*-ist*, *-y*, *-ite*, *-ia*, *-on*	BM	*biologist, photography, gelignite, hypothermia, Teflon*
8	BM	*-er* or *-s*	Verb or N/Aj	*teleprinter, teetotaler*
9	BM	*-s*	Aj	*bioethics*
10	BM	*-s-pl*, *-ency*	N	*telesales, immunodeficiency*
11	Prep	*-er*	N	*no-hoper*
12	N/Aj/Num	*-ing*	Pron	*thirty-something*
13	Ad	*-ing*	Verb	*forthcoming*
14	Ad	*-ery*	N	*midwifery*
15	N, Prep, N/Aj/Num or Ad	*-ment*, *-ance*, *-s*, *-ency*, *-or*, *-ist*, *-y*, *-ity*, *-an*, *-al*, *-ite*, *-ia* and *-on*	does not occur	
16	Prep, N/Aj/Num, Ad	*-ness* or *-s-pl*	does not occur	

C/a	ness	s-pl	er	ing	ery	ship	ment	ance	s	ency	or	ist	y	ity	an	al	ite	ia	on
N	Aj	N	N	N	N	N													
			Verb	Verb															
Aj		N	Verb	Verb			Verb												
Verb			Verb	Verb				Ad											
BM		N	Verb	Verb					Aj	N	N/Verb	BM	BM	Aj	Aj(i)	Verb	BM	BM	BM
			N/Aj								Verb								
Prep			N																
N/Aj/Num				Pron															
Ad-				Verb	N														

Figure 6.9. The matrix for the morphological construction {C-C-a}

6.1.2.4 The noun construction {C-a-C}

The construction {C-a-C} is less formally productive, as can be seen from Figure 6.10 and Table 6.8. The initiale for the construction {C-a-C} is occupied by 5 morphological classes (a bound morpheme, noun, adjective, verb and adverb), and its finale by 4 morphological classes (a bound morpheme, noun, verb and adverb). The highest type valency for the mediale in this construction is observed for nouns in the slot of the initiale and finale. The mediale infix -o- is the most valent across morphological patterns.

C/C	BM	N	Verb	Ad
BM	o			
N	o	o		
		s		
		i		
		and		
		in		
		a		
Aj	o			
Verb	o	s	and	
		a		
Ad				and

Figure 6.10. The matrix for the morphological construction {C-a-C}

Table 6.8. Morphological regularities for the noun construction {C-a-C} on the second level: NC2_{C-a-C}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	BM or Aj	BM	*-o-*	*hallucinogen, roughometer*
2	N	N	*-o-*, *-s-*, *-i-*, *-and-*, *-in-* and *-a-*	*sellotape, kinswoman, handicap, R&B, son-in-law, tick-a-tick*
3	Verb	BM	*-o-*	*deflectometer*
4	Verb	N	*-a-*, *-s-*	*spokesman, rackarock*
5	Verb	Verb	*-and-*	*hide-and-seek*
6	Ad	Ad	*-and-*	*up-and-up*
7	BM or N	Verb or Ad	does not occur	

6.1.2.5 The noun construction {C-C-C}

This construction is the least formally productive (Figure 6.11). Its initiale slot is occupied by a bound morpheme, and the finale slot by a bound morpheme and a noun. The mediale in this construction is monovalent and is filled with a bound morpheme. The examples of words for this construction are *electrocardiogram*, *povidone* and *chlorofluorocarbon*.

C/C	BM	N
BM	BM	BM

Figure 6.11. The matrix for the morphological construction {C-C-C}

6.1.3 The third and fourth levels

The third level contains five constructions: {C-C-a-C}, {C-a-C-a}, {C-a-a-a}, {C-C-C-a} and {a-C-a-C}. The latter is the most formally productive. Their inner composition of these constructions is given in Figure 6.12 and 6.13, and their regularities in Table 6.9 and 6.10.

The construction {C-C-a-a} produces one morphological pattern N"'=Aj+N+ed+ness and one word in the sample (i.e. *able-bodiedness*). The construction {C-a-C-a} is slightly more formally productive and involves a noun in the slot of the initiale, and 3 suffixes in the slot of the finale (*-er*, *-ing* and *-ship*). In contrast, the construction {C-a-a-a} has a higher type frequency for the initiale (N, BM and Verb), and a lower type frequency for the finale (i.e. *-ion* and *-s*). Finally, the construction {C-C-C-a} has the type frequency of 1 for the finale and the type frequency of 2 for the finale (*-ment* and *-ing*). Finally, the suffix finales *-er*, *-ment* and *-ing* have the common mediale Aj+en, and the suffix finales *-ion* and *-or* have the common mediale N+ate.

C/a	ness	er	ing	ship	ion	ie	s	ment	ing
Aj	N+ed								
N		er+Verb	s+Verb	s+N	an+ize(at)				
					al+ize(at)				
BM					al+ize(at)		ist+ic	BM+BM	BM+BM
Verb					er+ize(at)	ie+Verb			

Figure 6.12. The matrix for the morphological construction {C-C-a-a}, {C-a-C-a}, {C-a-a-a} and {C-C-C-a}

Table 6.9. Morphological regularities for the noun constructions {C-C-a-C}, {C-a-a-a}, {C-C-C-a} and {C-a-C-a} on the third level: NC3_{C-C-a-C}/{C-a-a-a}/{C-C-C-a}/{C-a-C-a}

Construction	**No**	**If**		**Then**	**Examples in the data**
		Initiale	***Finale***	***Mediale***	
{C-C-a-a}	1	Aj	*-ness*	N+ed	*able-bodiedness*
{C-a-a-a}	1	N	*-ion*	an+ize(at) or al+ize(at)	*Americanization, globalization*
	2	BM	*-ion*	al+ize(at)	*decimalization*
	3	Verb	*-ion*	er+ize(at)	*computerization*
	4	BM	*-s*	ist+ic	*linguistics*
{C-a-C-a}	1	N	*-er*	er+Verb	*wheeler-dealer*
	2	N	*-ing*	s+Verb	*painstaking*
	3	N	*-ship*	s+N	*craftsmanship*
	4	Verb	*-ie*	ie+Verb	*walkie-talkie*
{C-C-C-a}	1	BM	*-ment*	BM+BM	*acknowledgement*
	2	BM	*-ing*	BM+BM	*acknowledging*

a/a	**ness**	**er**	**ment**	**ing**	**ity**	**ion**	**or**	**ism**
dis	Verb+able							
	Verb+ed							
	Pron+ish							
un	Aj+ly							
	Aj+ing	Aj+en	Aj+en	Aj+en				
en			Aj+ing					
		Aj+ize						
de		N+ize				Aj+ize(at)		
il					N+al			
in					Verb+able(il)	N+ate		
im							N+ate	
anti								BM+ite

Figure 6.13. The matrix for the morphological construction {a-C-a-C}

The initiale slot of this construction is occupied by 8 prefixes (*dis-*, *un-*, *en-*, *de-*, *il-*, *in-*, *im-* and *anti-*), and its finale by 8 suffixes (*-ness*, *-er*, *-ment*, *-ing*, *-ity*, *-ion*, *-or* and *-ism*). The most type-frequent mediale is for the prefix initiale *en-*, which occurs in combination with four suffixes (*-ness*, *-er*, *-ment* and *-ing*). The most type-valent initiale is the prefix *un-* (which attaches to 3 different types of mediale), and the most type-valent finale are suffixes *-ness* (which attaches to 5 different types of mediale) and *-ment* (3 different types of mediale).

Lastly, the fourth level of noun formation involves the constructions {anti+N}”” and {C+C}””, which produce the words *antiglobalization* and *daddy-long-legs*.

Table 6.10. Morphological regularities for the noun construction {a-C-a-C}

No	If		Then	Examples
	Initiale	***Finale***	***Mediale***	
1	*dis-*	*-ness*	Verb+able	*disreputableness*
2	*un-*	*-ness*	Verb+ed, Pron+ish, Aj+ly	*unlimitedness, unselfishness, unkindliness*
3	*en-*	*-ness*	Aj+ing	*endearingness*
4	*en-*	*-er*	Aj+en	*enlightener*
5	*en-*	*-ment*	Aj+en, Aj+ing	*enlivenment, endearingment*
6	*en-*	*-ing*	Aj+en	*enlivening*
7	*de-*	*-er*	Aj+ize, N+ize	*decentralizer, deodorizer*
8	*de-*	*-ion*	Aj+ize(at), N+ate	*defamiliarization, devaluation*
9	*il-*	*-ity*	N+al	*illogicality*
10	*in-*	*-ity*	N+able(il)	*incognizability*
11	*im-*	*-or*	N+ate	*impersonator*
12	*anti-*	*-ism*	N+ite	*anti-Semitism*

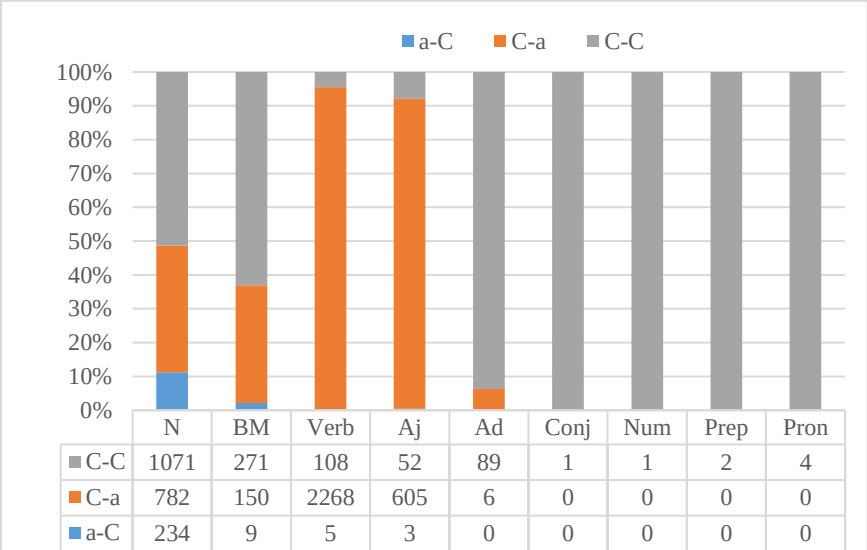
6.1.4 The main morphological trends in noun formation

The previous chapters have summarized the observed combinations of morphemes at four levels of noun formation in the form of morphological regularities. The picture that has emerged from the formal morphological analysis is diverse and has shown that different types of bases are involved in noun formation, leading to polyvalency of some affixes, the most type-valent of which are *-ness*, *-ing*, *-er*, *-ism*, *-ity*, *-ment*, *-ery*, *-age* and *-y*. Hay and Baayen (2002: 8) have established that the high relative frequency of a base word (as compared to the frequency of its derived forms) contribute to the parsability of the derived word. It may be that the type valency of the affix is another factor that influences the parsability of words: *the higher the type valency of an affix, the more parsable the word seems*.

Further, such suffixes *-s*, *-ism*, *-ity*, *-ship*, *-hood*, *-ry*, *-a*, *-ine* and *-ness* have appeared as closing suffixes, whereas the suffixes *-er*, *-ist*, *-ion*, *-ing*, *-or* and *-y* have a prominent role as middle suffixes (featuring in the mediale slot). Other suffixes show a greater involvement on the first level of word formation.

Finally, from the 'virtual' morphemes (see Section 3.3.3 for more detail) that have been recorded during the analysis and from the frequent repetitions of some combinations of morphemes, it can be concluded that some suffixes have closer connections than others. They include such combinations as *-ate* + *-ion*, *-ic* + *-ate* + *-ion*, *-ize* + *-ate* + *-ion*, *-ic* + *-al*, *-al* + *-ism*, *-able* + *-ity*. A considerable body of literature explains these and other combinations of suffixes with different types of restrictions: for example, with etymological (e.g. Marchand 1969; Plank 1981), affix-driven (Fabb 1988) or base-driven (Plag 1996) constraints. The matrix analyses of noun formation performed above, as well as the shares of word bases in three meta-constructions

featured in Figure 6.14, provide evidence for the base-driven explanation of suffix combinations. This observation is studied in greater detail in Chapter 7 (p. 225). Specifically, verb and adjective bases have the greatest contribution in the noun meta-construction {{C-a}} (as evident from Figure 6.14), which justifies the above-mentioned combinations of affixes.



	N	BM	Verb	Aj	Ad	Conj	Num	Prep	Pron
C-C	1071	271	108	52	89	1	1	2	4
C-a	782	150	2268	605	6	0	0	0	0
a-C	234	9	5	3	0	0	0	0	0

Figure 6.14. The shares of word bases in three noun meta-constructions: {{C-C}}, {{C-a}} and {{a-C}}

6.1.5 Formal morphological regularities for multimorphemic verbs

This section describes verb formation at two levels. Similar to noun formation, the first-level formation involves morphological constructions {a-C}, {C-a} and {C-C}. These constructions and their morphological regularities are analyzed in the following subsections.

6.1.5.1 The verb construction {a-C}: the first level

In verb formation, the morphological construction {a-C} is the most formally productive from the perspective of the number of items that fit into its slots. The inner composition of this construction is introduced in Figure 6.15, and its morphological regularities are listed in Table 6.11.

The initiale slot of this construction is occupied with 25 prefixes: *re-*, *de-*, *im-*, *dis-*, *in-*, *per-*, *a2-*, *be-*, *out-*, *en-*, *un-*, *co-*, *extro-*, *mis-*, *pre-*, *up-*, *inter-*, *over-*, *counter-*, *a-*, *mal-*, *sub-*, *under-*, *with-* and *ac-*. In the slot of the finale of this construction, there are 6 (with CC) or 4 (without CC) word classes. The most type-valent initiale are the prefixes *re-* with a type valency of 5 (with CC) or 4 (without CC), *de-* and *en-* with a type valency of 4 (with CC) or 3 (without CC). The most type-valent finales are verbs (23), nouns (8) and bound morphemes (6). On the other hand, the monovalent initiale involves the following prefixes: *extro-*, *mis-*, *pre-*, *up-*, *inter-*,

over-, *counter-*, *a-*, *mal-*, *sub-*, *under-*, *with-* and *ac-*. The least type-valent finales are the conversive classes N/Verb and Verb/Aj (with a type valency of 3 and 1, respectively).

a/C	Verb	N	BM	Aj	Verb/Aj	N/Verb
re	Ø	Ø	Ø	Ø	Ø	
de	Ø	Ø	Ø			Ø
im	Ø	Ø	Ø			
dis	Ø	Ø				
in	Ø	Ø				
per	Ø	Ø				
a2	Ø	Ø				
be	Ø	Ø		Ø		
out	Ø			Ø		
en	Ø		Ø	Ø		Ø
un	Ø			Ø		Ø
co	Ø		Ø			
extro			Ø			
mis	Ø					
pre	Ø					
up	Ø					
inter	Ø					
over	Ø					
counter	Ø					
a	Ø					
mal	Ø					
sub	Ø					
under	Ø					
with	Ø					

Figure 6.15. The matrix for the morphological construction {a-C} or {a-Ø-C}
(Key: the prefix *a2-* is of Old French origin as in the word *appraise*; whereas the prefix *a-* is of Old English origin and the variant of *or-* as in the word *amaze*)

Table 6.11. Morphological regularities for the verb construction {a-Ø-C} on the first level: VC1_{a-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	The prefix initiale *re-*	Verb, N, BM, Aj, Verb/Aj	*reassure, reboot, reflate, renew, relive*
2	The prefix initiale *de-*	Verb, N, BM, N/Verb	*de-escalate, defeature, desecrate, decenter*
3	The prefix initiale *im-*	Verb, N, BM	*impress, imperil, implode*
4	The prefix initiales *dis-*, *in-*, *per-* and *a2-*	Verb and N	*disinfect, discard, input, injelly, peruse, pretension, abound, avail*
5	The prefix initiale *be-*	Verb, N and Aj	*bemuse, behead, belittle*
6	The prefix initiale *out-*	Verb and Aj	*outcry, outsmart*
7	The prefix initiale *en-*	Verb, N, BM, Aj and N/Verb	*enclose, entrust, encrypt, embitter, entangle*
8	The prefix initiale *un-*	Verb, Aj, N/Verb	*unfasten, unstable, unmask*
9	The prefix initiale *co-*	Verb and BM	*co-organize, coordinate*
10	The prefix initiale *extro-*	BM	*extrovert*
11	The prefix initiales *mis-*, *pre-*, *up-*, *inter-*, *over-*, *counter-*, *a-*, *mal-*, *sub-*, *under-*, *with-*	Verb	*misbehave, pre-arrange, update, interconnect, overjoy, counterbalance, appraise, maltreat, subdelegate, underbuild, withhold*

6.1.5.2 The verb construction {C-a}

The second formally productive construction is {C-a}, presented in Figure 6.16, together with its morphological regularities in Table 6.12. The initiale slot of this construction is occupied by 8 (with CC) or 5 (without CC) word classes, and its finale slot by 12 suffixes (if *-ize*, *-ise* and *-yse* are counted as different morphemes) or 10 suffixes (if these suffixes are considered allomorphs). The most type-valent initiales in this construction include a bound morpheme and a noun with a type valency of 8, a verb (6) and an adjective (4). The rest of the initiales are duovalent (N/Verb) or monovalent (Ad, Aj/N and Aj/Ad/Verb). The most type-valent finales are the suffixes *-en* (6), *-ize* (5), *-ate* (3), *-ify* (3), *-le* (4) and *-er* (3). The suffix *-eer* is duovalent, and the suffixes *-ise*, *-ic*, *-yse*, *-ish* and *-age* are monovalent.

C/a	en	ize	ate	ify	le	er	ise	ic	eer	yse	ish	age
BM	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø				
N	Ø	Ø	Ø	Ø	Ø	Ø			Ø	Ø		
Verb	Ø	Ø			Ø	Ø			Ø		Ø	
Aj	Ø	Ø	Ø	Ø								
Ad	Ø											
Aj/N		Ø										
Aj/Ad/Verb	Ø											
N/Verb					Ø							Ø

Figure 6.16. The matrix for the morphological construction {C-a}

Table 6.12. Morphological regularities for the verb construction {C-Ø-a} on the first level: VC1_{C-Ø-a}

No	Type in the slot	Attach(es) to	Examples in the data
1	BM initiale	*-en, -ize, -ate, -ify, -le, -er, -ise, -ic*	*quicken,[31] optimize, nitrate, liquefy, babble,[32] scatter,[33] optimize, paralyse, authentic*
2	N initiale	*-en, -ize, -ate, -ify, -le, -er*	*frighten, idolize, formulate, beautify, fizzle, molder*
3	Verb initiale	*-en, -ize, -le, -er*	*chasten, acclimatize, muzzle, glower*
4	Aj initiale	*-en, -ize, -ate, -ify*	*blacken, equalize, activate, acidify*
5	Ad initiale	*-en*	*uppen*
6	Aj/N initiale	*-ize*	*dentalize*
7	Aj/Ad/Verb initiale	*-en*	*slighten*
8	N/Verb initiale	*-le, -age*	*gamble, rampage*

[31] There are two homophonic verbs that have the form *quicken* (given as two separate entries in the OED). One is formed as Aj+en, and another as BM+en. According to the OED, in the second form, which is the example in the table, the bound morpheme is *quick-* (taken from *quicksilver*).
[32] As informed by the OED, the word *bubble* is apparently formed as the syllable /bʌ/ (which is a characteristic of the early infantile vocalization) and the suffix *-le*. Since, in the current study, the domain of a bound morpheme is expanded and since the diachronic perspective is integrated in the current morphological description as an important criterion of the morphological parsing, this word is considered divisible.
[33] The verb *scatter* is identified by the OED as the word of obscure origin, formed with the iterative suffix *-er*. Because the origin of *scat-* is unknown and because the etymology of the noun *scat* does not seem to be related to the root *scat-*, it is marked as a bound morpheme.

6.1.5.3 The verb construction {C-C}

The least formally productive construction is {C-C} or {C-Ø-C}. Its inner composition is visualized in Figure 6.17 and is explained in Table 6.13. The initiale slot of this construction is occupied by 5 word classes (BM, Aj, N, Verb and Prep) and its finale slot by 4 word classes (Verb, BM, N, Ad and Pron). The most type-valent initiales are nouns (4) and bound morphemes, verbs and prepositions (with a type valency of 2), whereas the most type-valent finales are verbs (4) and nouns (2). The monovalent initiale is an adjective, and the monovalent finales are adverbs and pronouns.

C/C	Verb	BM	N	Ad	Pron
BM	Ø	Ø			
Aj	Ø				
N	Ø	Ø	Ø		
Verb	Ø			Ø	
Prep			Ø		Ø

Figure 6.17. The matrix for the verb morphological construction {C-C}

Table 6.13. Morphological regularities for the verb construction {C-Ø-C} on the first level: VC1_{C-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	BM initiale	Verb and BM	*telecommute, cybercast*
2	Aj initiale	Verb	*broadcast*
3	N initiale	Verb, BM and N	*kidnap, girlcott, toenail*
4	Verb initiale	Verb and Ad	*write-protect, don*
5	Prep initiale	N and Pron	*bay, atone*

6.1.6 The second level of verb formation: {a-C-a}, {C-a-a} and {C-C-a}

There are three constructions on the second level of verb formation: {a-C-a}, {C-a-a} and {C-C-a}. The internal structures of the first two constructions, as well as their morphological regularities, are presented in Figure 6.18–6.19, and Table 6.14–Table 6.15, respectively. The only instance of the construction {C-C-a} is the morphological pattern Verb"=BM+N+ize (the word *vitaminize*[34]).

6.1.6.1 The verb construction {a-C-a}

The initiale slot of this construction is occupied by the prefixes *de-*, *self-*, *re-*, *im-*, *in-*, *dis-* and *en-*, and the finale slot by the suffixes *-ize*, *-ate*, *-ify* and *-en*. The most type-valent initiales are *de-* (3) and *re-* (3), and the most type-valent finales are *-ize* (3), *-ate* (2) and *-en* (2). The monovalent suffix

[34] The word *vitamin*, which is a noun base in *vitaminize*, consists of two morphological components, as identified by the OED: the Latin word *vita-* (meaning 'life') and the noun *amine* (from a mistaken belief about the chemical nature of the compounds). For this reason, the word base *vitamin* has been parsed as BM+N.

is *-ify* (the morphological pattern N"=de+BM+ify). The most frequent mediales in this construction are adjectives and nouns.

a/a	ize	ate	ify	en
de	Aj	N	BM	
	N	BM		
	BM			
self	Aj			
re	Aj	BM		Aj
im		N		
in		N		
dis				N
en				Aj

Figure 6.18. The matrix for the morphological construction {a-C-a}

Table 6.14. Morphological regularities for the verb construction {a-C-a} on the second level: VC2_{a-C-a}

No	If		Then	Examples
	Initiale	***Finale***	***Mediale***	
1	*de-*	*-ize*	Aj, N and BM	*decentralize, de-emphasize, deodorize*
2	*de-*	*-ate*	N and BM	*degranulate, dehydrate*
3	*de-*	*-ify*	BM	*detoxify*
4	*self-*	*-ize*	Aj	*self-actualize*
5	*re-*	*-ate*	BM	*rejuvenate*
6	*im-*	*-ate*	N	*impersonate*
7	*dis-*	*-en*	N	*dishearten*
8	*en-*	*-en*	Aj	*enlighten*
9	*in-*	*-ate*	N	*incapacitate*

6.1.6.2 The verb construction {C-a-a}

The initiale slot of this construction is filled with 3 word classes (BM, N and Verb), and its finale slot with one suffix (*-ize*). The most type-valent initiale is a noun (3). The verb initiale is monovalent in this construction.

C/a	ize
BM	al
	an
N	an
	al
	er
Verb	er

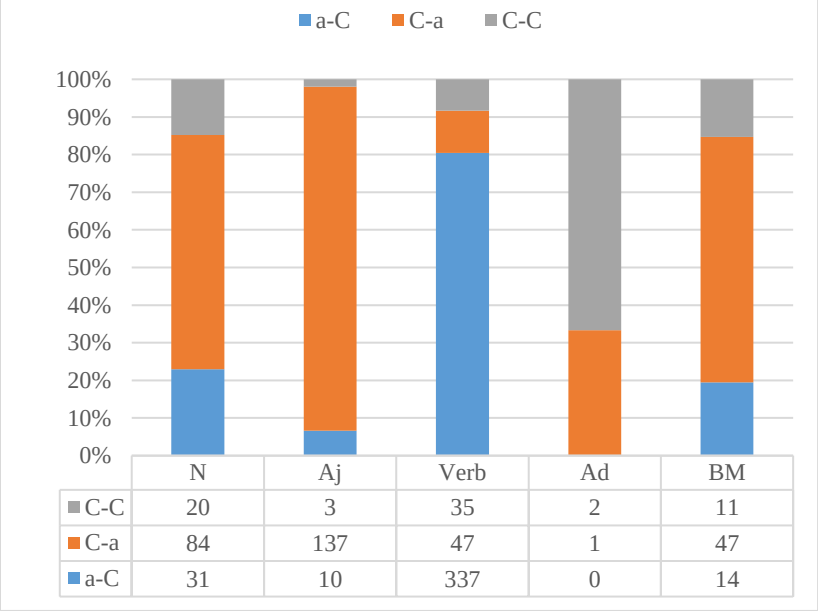
Figure 6.19. The matrix for the morphological construction {C-a-a}

Table 6.15. Morphological regularities for the verb construction {C-a-a} on the second level: VC2_{C-a-a}

No	If		Then	Examples
	Initiale	***Finale***	***Mediale***	
1	BM	*-ize*	*-al-* or *-an-*	*internalize, pedestrianize*
2	N	*-ize*	*-an-*, *-al-* or *-er-*	*globalize, Americanize, computerize*
3	Verb	*-ize*	*-er-*	*crofterize*

6.1.7 The main trends in the formation of verbs

With a lower number of constructions, English verb formation shows a lower degree of word-formation complexity.[35] A small portion of verbs are formed by compounding, with no adjectives observed in the finale slot. In contrast to nouns, the construction {a-C} is more formally productive in verbs, which is evidence for a more pronounced role of prefixation in the formation of verbs. Further, the prefix *de-* is the most type-valent,[36] and is the only prefix observed with verbs ending in *-ize*, *-ate* and *-ify*. Finally, as shown in Figure 6.20, adjective, noun and bound morpheme bases have the greatest contribution in the formation of multimorphemic verbs with the involvement of suffixes, i.e. in the meta-construction {{C-a}}, which justifies the observed frequent combinations of suffixes (e.g. *-al* + *-ize*, *-an* + *-ize*, *-er* + *-ize*) and provides evidence for base-driven selections of suffixes.



	N	Aj	Verb	Ad	BM
C-C	20	3	35	2	11
C-a	84	137	47	1	47
a-C	31	10	337	0	14

Figure 6.20. The share of word bases in verb formation across three meta-constructions

6.1.8 Formal morphological regularities for multimorphemic adjectives

The following subsections are devoted to the analysis of the adjectival constructions on the first level. As with multimorphemic nouns and verbs, the first-level adjectival formation involves three

[35] In the current study, phrasal verbs were excluded from the data. This is because the formal morphological analysis concerns only multimorphemic words, which are defined in the spirit of formal approaches—as an entity between two blank spaces (for more explanation, see p.33 and p.42). Moreover, in English phrasal verbs, the particle is separable from the verb base and can occur after a direct object, which adds another reason to exclude them from the data of this study.

[36] 'The most type-valent' means that an affix has the largest value of the type valency, as compared to other affixes.

constructions: {a-C}, {C-a} and {C-C}. Figures 6.21–6.23 suggest that the formal productivity of the constructions {a-C} and {C-a} is relatively the same, with the initiale slot being more formally productive for the former and the finale slot for the latter.

6.1.8.1 The adjective morphological construction {a-C} or {a-Ø-C}

The initiale slot of this construction is occupied by 31 prefixes and the finale slot by 7 (with CC) or 5 (without CC) word classes (Figure 6.21). The prefixes are predominantly monovalent and duovalent, with an adjectival base displaying the highest frequency. The matrix is described in Table 6.16. Bound morphemes attach only to the prefix *intra-*, adverbs only to the prefix *a-* and verbs to the prefix *non-*.

a/C	Aj	N/Aj	N	N/Aj/Ad	Ad	Verb	BM
im	Ø	Ø		Ø			
un	Ø	Ø					
in	Ø	Ø	Ø				
ante	Ø		Ø				
anti	Ø		Ø				
a2	Ø		Ø				
pre	Ø		Ø				
ex	Ø		Ø				
pro	Ø		Ø				
sub	Ø		Ø				
dis	Ø						
up	Ø						
re	Ø						
ir	Ø						
mis	Ø						
semi	Ø						
ab	Ø						
ad	Ø						
fore	Ø						
il	Ø						
per	Ø						
self	Ø						
ultra	Ø						
non	Ø					Ø	
extra		Ø					
a			Ø		Ø		
over			Ø				
counter			Ø				
de			Ø				
semi			Ø				
intra	Ø						Ø

Figure 6.21. The matrix for the adjective construction {a-C}
(Key: *a2-* refers to the prefix of Greek origin as in *atypical*, whereas *a-* to a non-productive native prefix *on-* as in *alike*)

Table 6.16. Morphological regularities for the adjective construction {a-C}: AC1_{a-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale *im-*	Aj, N/Aj, N/Aj/Ad	*imprecise, impartial, improper*
2	Initiale *un-*	Aj, N/Aj	*unabrupt, uneven*
3	Initiales *in-*,[37] *ante-, anti-, a2-, pre-, ex-, pro-, sub-*	Aj, N	*inedible, in-form, antenatal, ante-post, antisocial, anti-theft, atypical, agender, agender, preconception, ex-focal, ex-directory, ex-focal, proactive, pro-life, subtotal, sub-zero*
4	Initiales *dis-, up-, re-, ir-, mis-, semi-, ab-, ad-, fore-, il-, per-, self-, ultra-*	Aj	*disuniform, uptight, reproductive, irrelevant, misshapen, semi-arid, abapical, adoral, foregone, illegible, pernitric, self-active, ultrathin*
5	Initiale *non-*	Aj and Verb	*non-resident, non-iron*
6	Initiale *extra-*	N/Aj	*extramarital*
7	Initiale *a-*	N and Ad	*alive, alike*
8	Initiales *over-, counter-, de-, semi-*	N	*overweight, counter-camp, decomplex, semi-log*
9	Initiale *intra-*	Aj and BM	*intravital, intra-uterine*

6.1.8.2 The adjective morphological construction {C-a} or {C-Ø-a}

This construction is analyzed in the matrices of Figure 6.22–6.23. Eight (with CC) and four (without CC) word classes are involved in the initiale slot of this construction and 29 suffixes in its finale slot. The most type-valent initiales include verbs, nouns, adjectives and bound morphemes, whereas the most type-valent finales are the suffixes *-ed*, *-ish*, *-less*, *-some*, *-ing*, *-ive*, *-ic* and *-ful*. Table 6.17 looks at the combinatorial properties of morphemes in greater detail.

C/a	ed	ish	less	ing	some	sy	ly	ive	ic	ate	ful	able	ory	ent	en	ible	ant	le
Verb	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
N	Ø	Ø	Ø	Ø	Ø		Ø	Ø	Ø	Ø								
Aj	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø		Ø							
N/Verb	Ø	Ø	Ø	Ø	Ø						Ø	Ø						
BM	Ø	Ø	Ø					Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø			
Aj/Verb	Ø	Ø			Ø													
N/Aj	Ø	Ø									Ø							
N/Aj/Verb	Ø																	

Figure 6.22. The matrix for the construction {C-a} (Part 1)

C/a	y	al	ous	ar	ary	most	en	esque	an	like	id
N	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	
BM	Ø	Ø	Ø	Ø	Ø						Ø
Aj	Ø	Ø	Ø			Ø					
N/Verb	Ø										
Pron	Ø										
Verb	Ø										
N/Aj		Ø									

Figure 6.23. The matrix for the adjective construction {C-a} (Part 2)

[37] The distinction between the allomorphs *im-* and *in-* is formal, and would have not been made if they had had the same pattern of type valency. However, as shown in Figure 6.21, the allomorph *im-* attaches to nouns and nouns/adjectives, and the allomorph *in-* to nouns, nouns/adjectives and adjectives. For this reason, these allomorphs have two separate entries in the table.

Table 6.17. Morphological regularities for the adjective construction {C-a}: AC1_{C-Ø-a}

No	Type in the slot	Attach(es) to	Examples in the data
1	Verb initiale	*-ed, -ish, -less, -ing, -some, -sy, -ly, -ive, -ic, -ate, -ful, -able, -ory, -ent, -en, -ible, -ant, -le, -y*	*faded, diggish, defendless, pleading, crawlsome, tipsy, ghastly, modulative, chameleonic, extortionate, escapeful, scrollable, acceleratory deferent, crouchant, bidden, discussible, brittle, perky*
2	N initiale	*-ed, -ish, -less, -ing, -some, -ly, -ive, -ic, -ate, -y, -al, -ous, -ar, -ary, -most, -en, -esque, -an, -like*	*corridored, clownish, ageless, nursing, quarrelsome, daughterly, documentative, desertic, ovulate, clueful, serviceable, ashen, balmy, lagoonal, hazardous, molar, complimentary, topmost, statuesque, regalian, village-like*
3	Aj initiale	*-ed, -ish, -less, -ing, -some, -sy, -ly, -ive, -ic, -ful, -y, -al, -ous, -most*	*sored, greyish, fledgeless, balding, lonesome, deadly, secretive, serratic, crispy, conical, horrendous, deepmost*
4	N/Verb initiale	*-ed, -ish, -less, -ing, -some, -ful, -able, -y*	*crusted, sluggish, thriveless, quibbling, rattlesome. delightable, flashy*
5	BM initiale	*-ed, -ish, -less, -ive, -ic, -ate, -ful, -able, -ory, -ent, -en, -y, -al, -ous, -ar, -ary, -id*	*convexed, garish, careless, delusive, pelvic, numerate, wistful, decorable, jubilatory, fluorescent, brazen, clumsy, cryptical, anxious, obstacular, monetary, acrid*
6	Aj/Verb initiale	*-ed, -ish, -some*	*diffused, ticklish, wearisome*
7	N/Aj initiale	*-ed, -ish, -ful, -al*	*rectangled, dankish, fanciful, sceptical*
8	N/Aj/Verb	*-ed*	*muted*
9	Pron	*-y*	*naughty*

6.1.8.3 The adjective morphological construction {C-C} or {C-Ø-C}

The compounding adjective construction is the least productive. Its initiale slot allows for 7 word classes, and its finale slot for 5 (with CC) or 4 (without CC) word classes. The most type-valent initiale is a bound morpheme, and the most type-frequent finale is an adjective and a noun. Figure 6.24 illustrates the arrangement of morphemes in this construction, and Table 6.18 provides examples for the observed combinations.

C/C	Aj	N	BM	N/Aj	Ad
BM	Ø	Ø	Ø	Ø	
N	Ø	Ø			
Aj	Ø	Ø			
Verb	Ø				Ø
Ad	Ø				
Prep		Ø			
Part		Ø			

Figure 6.24. The matrix for the adjective construction {C-C}

Table 6.18. Morphological regularities for the adjective construction {C-C}: AC1_{C-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	BM initiale	Aj, N, BM and N/Aj	*hyperactive, cross-party, cucumiform, neoclassical*
2	N initiale	Aj, N	*alcohol-free, bite-size*
3	Aj initiale	Aj, N	*dear-bought, close-range*
4	Verb initiale	Aj, Ad	*rip-off, cock-eyed*
5	Ad initiale	Aj	*roughshod*
6	Prep initiale	N	*in-flight*
7	Part initiale	N	*no-win*

6.1.9 The second-level adjectival constructions

The adjectival formation on the second level is diverse and is comprised the constructions {a-C-a}, {C-C-a}, {C-a-a} and {C-C-C}. They are explored in the following subsections.

6.1.9.1 The adjective construction {a-C-a}

This construction is the most formally-productive on the second level (Figure 6.25).

a/a	ed	ing	able	al	ful	y	ish	ic	ary	ly	ous	ible	ive	ory	ar
	Verb	N	Verb	BM	BM	Verb	Pron	N	N	N					
	N/Verb	Verb	N/Verb	N	N	N				Aj					
un				Aj											
	N	Aj													
	Verb	N													
	Aj														
en	N/Verb														
em	N	N													
over	Verb	Verb													
	Aj	Verb													
	Verb														
be	N														
	Verb	Verb	Verb		N										
mis	N/Verb														
	Verb	N			N								Verb		
dis		Verb													
re	Verb	Verb	Verb		N										
de	N			Aj									Verb		
pre	Verb			Aj											
counter		Verb													
fore		Verb	Verb												
up	N	Verb													
a		N													
	Verb	Verb	BM	N							BM	Verb	BM		
in			Verb												
non				N									BM	BM	
a2				N/Aj				N							
afore	Verb														
im	Verb		Verb	BM											
mal	Verb														
on		Verb													
self	Verb														
ir			Verb		N							Verb			
co		Verb		N											
				N											
sub				Aj											
extra	Verb														BM
il				N											
inter			Verb	N											
								N/Aj							BM
anti								BM							

Figure 6.25. The matrix for the adjective construction {a-C-a}

Table 6.19. Morphological regularities for the adjective construction {a-C-a}: AC2_{a-C-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	*un-*	*-ed*	Verb, N/Verb	*unloaded, unmasked*
2	*un-*	*-ing*	Verb, N	*unappetizing, unnerving*
3	*un-*	*-able*	Verb, N/Verb	*unbelievable, unquestionable*
4	*un-*	*-al*	BM, N, Aj	*unequivocal, unintentional, grammatical*
5	*un-*	*-y*	Verb, N	*unwieldy, unlucky*
6	*un-*	*-ful*	BM, N	*ungrateful, untruthful*
7	*un-*	*-ic, -ary,*	N	*unromantic, uncomplimentary*
8	*un-*	*-ly*	N, Aj	*unearthly, ungainly*
9	*en-*	*-ed*	N, Verb, Aj, N/Verb	*embanked, embroidered, endeared, entangled*
10	*en-*	*-ing*	Aj, N	*endearing, entrancing*
11	*em-*	*-ed, -ing*	N	*embastioned, empowering*
12	*over-*	*-ed, -ing*	Verb	*overwatched, overbounding*
13	*be-*	*-ed*	Aj, Verb, N	*belated, bemused, bespectacled*
14	*mis-*	*-ing, -able*	Verb	*misleading, misleadable*
15	*mis-*	*-ed*	Verb, N/Verb	*miseducated, misfortuned*
16	*mis-*	*-ful*	N	*mistrustful*
17	*dis-*	*-ed, -ive*	Verb	*discontented, disintegrative*
18	*dis-*	*-ing*	N, Verb	*dispiriting, disqualifying*
19	*dis-*	*-ful*	N	*disrespectful*
20	*re-*	*-ed, -ing, -able*	Verb	*reheated, relocating, reusable*
21	*de-*	*-ed*	N	*defatted*
22	*de-*	*-al*	Aj	*delexical*
23	*de-*	*-ive*	Verb	*decorrugative*
24	*pre-*	*-ed*	Verb	*preoccupied*
25	*pre-*	*-al*	Aj	*prehistorical*
26	*counter-*	*-ing*	Verb	*counteracting*
27	*fore-*	*-ing, -able*	Verb	*forewarning, foreseeable*
28	*up-*	*-ed*	N	*uprooted*
29	*up-*	*-ing*	Verb	*uprising*
30	*a-*	*-ing*	N	*amazing*
31	*in-*	*-ed, -ing*	Verb	*incoming, indisposed*
32	*in-*	*-able*	BM, Verb	*incalculable, incognizable*
33	*in-*	*-al*	N	*inconsequential*
34	*in-*	*-ous, -ive*	BM	*inauspicious, inoperative*
35	*in-*	*-ible*	Verb	*incontrovertible*
36	*non-*	*-ive, -ory*	BM	*nondestructive, non-contributory*
37	*non-*	*-al*	N	*non-fictional*
38	*a2-*	*-al*	N/Aj	*apolitical*
39	*a2-*	*-ic*	N	*asymmetric*
40	*afore-, im-, mal-, self-*	*-ed*	Verb	*aforementioned, implumed, maladjusted, self-abandoned*
41	*im-*	*-able*	Verb	*impassable*
42	*im-*	*-al*	BM	*impractical*
43	*on-, co-*	*-ing*	Verb	*oncoming, co-existing*
44	*co-*	*-al*	N	*coeducational*
45	*ir-*	*-ful*	N	*irrespectful*
46	*ir-*	*-able, -ible*	Verb	*irretrievable, irreversible*
47	*sub-*	*-al*	N, Aj	*subcontinental, subtropical*
48	*il-, inter-*	*-al*	N	*illogical, intercontinental*
49	*extra-, anti-*	*-ar*	BM	*extra-curricular*
50	*anti-*	*-ic*	N, BM	*anti-Semitic, antinuclear*
51	*extra-*	*-ed*	Verb	*extra-illustrated*
52	*inter-*	*-able*	Verb	*interdefinable*

The initiale slot of the construction {a-C-a} is occupied by 29 prefixes and the finale slot by 15 suffixes. As evident from the matrix, verbs and nouns have the greatest contribution in the position

of the mediale. The most type-valent initiale is the prefix *un-*, which occurs in the largest number of combinations (with 10 different finales).

6.1.9.2 The adjective construction {C-a-a}

This is the second formally productive construction on the second level (Figure 6.26), with the initiale slot occupied by 4 word classes and the finale slot by 13 suffixes.

C/a	ed	ing	able	al	less	ly	ish	ous	ic	ful	y	ary	ive
	ize	en	ize	ic	er	ward	y	(at)ion	ist		ist		
	ish	ize	ify										
	en	eer											
	ate												
	ify												
	ock												
	let												
N	ist												
	en	en							ist				
Aj	ize	ize											
	er	ize							ist				ate
BM		le											
	ize	le	er	ion	age					ing	th	ion	
	er			ment	ing						er		
Verb				er	er						ed		

Figure 6.26. The matrix for the adjective construction {C-a-a}

Table 6.20. Morphological regularities for the adjective construction {C-a-a}: AC2_{C-a-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	N	*-ed*	*-ize-, -ish-, -en-, -ate-, -ify-, -ock-, -let-, -ist-*	*crystallized, famished, heartened, pollinated, countrified, buttocked, ringleted, touristed*
2	N	*-ing*	*-en-, -ize-, -eer-*	*heartening, patronizing, profiteering*
3	N	*-able*	*-ize-, -ify-*	*journalizable, classifiable*
4	N	*-al*	*-ic-*	*catastrophical*
5	N	*-less*	*-er-*	*customerless*
6	N	*-ly*	*-ward-*	*northwardly*
7	N	*-ish*	*-y-*	*babyish*
8	N	*-ous*	*-(at)ion-*	*flirtatious*
9	N, BM, Aj	*-ic*	*-ist-*	*egotistic, euphemistic, realistic*
10	Aj	*-ed, -ing*	*-en-, -ize-*	*hardened, publicized, darkening, actualizing*
11	BM	*-ed*	*-er-*	*scattered*
12	BM	*-ing*	*-ize-, -le-*	*tantalizing, sprinkling*
13	BM	*-ive*	*-ate-*	*decorative*[38]
14	Verb	*-ed*	*-ize-, -er-*	*acclimatized, whiskered*
15	Verb	*-ing*	*-le-*	*gangling*
16	Verb	*-able*	*-er-*	*trailerable*
17	Verb	*-al*	*-ion-, -ment-, -er-*	*accommodational, developmental, managerial*
18	Verb	*-less*	*-age-, -ing-, -er-*	*luggageless, meaningless, transformerless*
19	Verb	*-y*	*-th-, -er-, -ed-*	*growthy, rubbery, crookedy*
20	Verb	*-ary*	*-ion-*	*deflationary*
21	Verb	*-ful*	*-ing-*	*meaningful*

[38] The verb *decorate* is treated as a duomorphemic word, formed by the Latin participial stem *decorāt-* and the suffix *-ate*. For this reason, the word *decorative* has been parsed as N"=BM+ate+ive.

In the adjective construction {C-a-a}, the noun initiale and the suffix finales *-ed*, *-ing*, *-able*, *-al*, *-less* and *-y* are the most type-valent. The most frequent mediales are the suffixes *-ize-*, *-er-*, *-en-* and *-ist-*. The morphological regularities of this construction are described in Table 6.20.

6.1.9.3 The adjective constructions {C-C-a} and {C-C-C}

This initiale of this construction is filled with 4 word classes and 9 suffixes (Figure 6.27). Bound morphemes and verbs are the most frequent initiales, and the suffixes *-ed* and *-ing* the most frequent finales. Verbs, nouns and bound morphemes have a pronounced role in the position of the mediale. The morphological regularities for this construction are presented in Table 6.21.

C/a	ed	ing	ous	able	ly	al	ic	less	ish
BM	BM	Verb		Verb	N	BM	BM		
						N			
Verb								Ad	Ad
N	Verb	N	BM						
	N	Verb							
Aj	Verb	Verb							
	N								

Figure 6.27. The matrix for the adjective construction {C-C-a}

Table 6.21. Morphological regularities for the adjective construction {C-C-a}: AC2_{C-C-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	BM	*-ed*	BM	*microphoned*
2	BM	*-ing*	Verb	*telecommuting*
3	BM	*-able*	Verb	*biodegradable*
4	BM	*-al*	BM or N	*neurological, permacultural*
5	BM	*-ic*	BM	*bureaucratic*
6	Verb	*-less, -ish*	Ad	*stirrupless, stand-offish*
7	N	*-ed*	Verb or N	*coin-operated, datelined*
8	N	*-ing*	Verb or N	*award-winning, footballing*
9	N	*-ous*	BM	*image-conscious*
10	Aj	*-ed*	Verb or N	*ill-advised, fair-minded*
11	Aj	*-ing*	Verb	*merrymaking*

In the sample of the study, the adjective construction {C-C-C} is realized only by one word: i.e. *made-to-measure* (Aj"=Verb*3+Part+N).

6.1.10 The third-level adjectival constructions

The highest level of adjectival formation is morphologically diverse and encompasses the following 7 constructions: {a-C-a-a}, {a-C-C-a}, {a-a-C-a}, {C-a-a-C}, {C-a-C-a}, {C-C-a-C}

and {a-C-a-C}. They are illustrated in Figure 6.28 and explained in Table 6.22. The most formally productive constructions are {C-a-C-a} and {a-C-C-a}.

a/a	ed	ing	al	y	ic	able	Aj	N	Ad
de	N+ate								
en		Aj+en							
ac		BM+BM							
un	Ad+Verb	pre+Verb		N+N	Aj+ist				Verb+ed
non		N+Verb							
ir						re+Verb			
Aj								er+Conj	
BM	un+Verb								
	al+ize		an+ic				un+Verb		
Verb	ly+N								
N	y+N	s+Verb							
	ium+N/Verb								
Ad		and+Verb							

Figure 6.28. The matrix for the adjective constructions {a-C-a-a}, {a-C-C-a}, {a-a-C-a}, {C-a-a-C}, {C-a-C-a}, {C-C-a-C} and {a-C-a-C}.

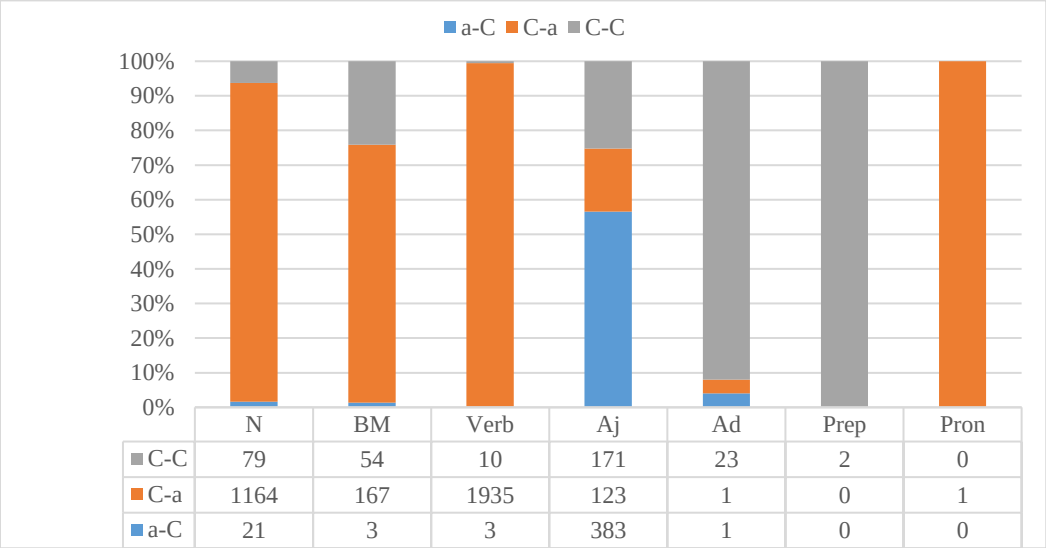
Table 6.22. Morphological regularities for the adjective constructions: AC_3: {a-C-a-a}, {a-C-C-a}, {a-a-C-a}, {C-a-a-C}, {C-a-C-a}, {C-C-a-C} and {a-C-a-C}

Construction	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
{C-a-C-C}	Aj	N	er+Conj	*larger-than-life*
{C-a-a-C}	BM	*-ed*	al+ize	*internalized*
	BM	*-al*	an+ic	*puritanical*
{C-a-C-a}	Verb	*-ed*	ly+N	*curly-haired*
	N	*-ed*	y+N or ium+N/Verb	*starry-eyed, chromium-plated*
	N	*-ing*	s+Verb	*painstaking*
	Ad	*-ing*	and+Verb	*up-and-coming*
	BM	*-ed*	un+Verb	*polyunsaturated*
{a-C-a-a}	*de-*	*-ed*	N+ate	*decaffeinated*
	en-	*-ing*	Aj+en	*enlightening*
	un-	*-ic*	Aj+ist	*unrealistic*
{a-C-C-a}	*un-*	*-ed*	Ad+Verb	*unfulfilled*
	ac-	*-ing*	BM+BM	*acknowledging*
	non-	*-ing*	N+Verb	*non-profit-making*
	un-	*-y*	N+N	*untrustworthy*
{a-a-C-a}	*un-*	*-ing*	pre+Verb	*unprepossessing*
	ir-	*-able*	re+Verb	*irreplaceable*
{a-C-a-C}	*un-*	Ad	Verb+ed	*uncared-for*

6.1.11 The main trends in adjectival formation

A diverse morphological picture has emerged from the matrix analysis presented above, in particular on the second and third levels of adjectival formation. As evident from Figure 6.25, prefixation has a more pronounced role on the second level and suffixation on the first level. Further, the most type-valent adjectival suffixes include *-ed*, *-ish*, *-less*, *-ing*, *-some*, *-ful* and *-y*. Lastly, as also observed for nouns and verbs, the most frequent, established combinations of

suffixes (e.g. *-ize* + *-ed*, *-en* + *-ing*, *-en* + *-ed*, *-ize* + *-ing*, *-ist* + *-ic*, *-er* + *-able*) are motivated by a significant contribution of verbs and nouns to the formation of adjectives.



	N	BM	Verb	Aj	Ad	Prep	Pron
C-C	79	54	10	171	23	2	0
C-a	1164	167	1935	123	1	0	1
a-C	21	3	3	383	1	0	0

Figure 6.29. The shares of word bases in adjectival formation across three meta-constructions

6.1.12 The first-level adverb constructions

Three constructions contribute to the formation of adverbs on the first level: {C-a}, {a-C} and {C-C}. The following subsections provide their detailed analysis. As illustrated in Figures 6.30–6.32, the first-level adverb formation is sparse and is heavily dominated by the suffix *-ly*.

6.1.12.1 The adverb construction {C-a}

This construction is morphologically monotonous (Figure 6.31). Nevertheless, it vividly illustrates the idea that suffixes with a high type frequency tend to attach to a greater number of word bases: on the first level, the suffix *-ly* has a type frequency of 580 (out of 611) and it attaches to 8 (with CC) or 6 (without CC) word classes. Other monovalent suffixes which have a minute share in the formation of two-morpheme adverbs include *-wise*, *-ish* and *-er*. Table 6.23 provides examples for the observed combinations of morphemes in this construction.

C/a	ly	wise	ish	er
Aj	Ø	Ø		
N	Ø			
Ad	Ø		Ø	Ø
N/Aj	Ø			
Aj/Ad	Ø			
BM	Ø			
Verb	Ø			
Num	Ø			

Figure 6.30. The matrix for the adverb construction {C-a} or {C-Ø-a}

Table 6.23. Morphological regularities for the adverb construction {C-Ø-a}

No	Type in the slot	Attach(es) to	Examples in the data
1	Aj initiale	*-ly, -wise*	*abundantly, likewise*
2	N initiale	*-ly*	*beastly*
3	Ad initiale	*-ly, -ish, -er*	*soonly, soonish, upper*
4	N/Aj initiale	*-ly*	*secretly*
5	Aj/Ad initiale	*-ly*	*excellently*
6	BM initiale	*-ly*	*early*
7	Verb initiale	*-ly*	*adaptly*
8	Num initiale	*-ly*	*fourthly*

6.1.12.2 The adverb construction {a-C}

A small portion of adverbs on the first level is formed by prefixation. The initiale slot in this construction is occupied by the prefixes *a-*, *up-* and *un-*, and the finale slot by nouns, adverbs and adjectives (Figure 6.31). The examples for the established combinations of morphemes are given in Table 6.24.

a/C	N	Ad	Aj
a	Ø	Ø	Ø
up	Ø		
un		Ø	Ø

Figure 6.31. The matrix for the adverb construction {a-C} or {a-Ø-C}

Table 6.24. Morphological regularities for the adverb construction {a-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale *a-*	N, Ad, Aj	*aground, afar, afresh*
2	Initiale *up-*	N	*uphill*
3	Initiale *un-*	Ad, Aj	*unalike, unsound*

6.1.12.3 The adverb construction {C-C}

As shown in Figure 6.32, the initiale slot of this construction is less frequent than that of the finale slot: it is filled only with three word classes (adverbs, pronouns and prepositions), whereas five word classes contribute to the finale slot (adverbs, adjectives, prepositions, bound morphemes and nouns). Table 6.25 describes this construction.

C/C	Ad	Aj	Prep	BM	N
Ad	Ø	Ø	Ø	Ø	
Pron	Ø				Ø
Prep					Ø

Figure 6.32. The matrix for the adverb construction {C-C} or {C-Ø-C}

Table 6.25. Morphological regularities for the adverb construction {C-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale Ad	Ad, Aj, Prep, BM	*henceforth, already, thereabout*
2	Initiale Pron	Ad, N	*somehow, somepart*
3	Initiale Prep	N	*indeed*

6.1.13 The second-level adverbial formation

The adverbial formation on the second level involves four morphological constructions: {C-a-a}, {a-C-a}, {C-C-C} and {C-C-a}. Similar to the constructions of the first level, all of them are dominated by the suffix *-ly*.

6.1.13.1 The adverbial construction {C-a-a}

The initiale of this construction is occupied by 7 (with CC) or 5 (without CC) word classes, whereas its finale by only one suffix *-ly* (Figure 6.33). The most type-valent initiale in {C-a-a} is a noun. Most of its mediales include adjective-forming suffixes. The combinatorial properties of morphemes in this construction are analyzed in Table 6.26.

C/a	ly	C/a	ly
	ive		al
	ing		ous
	ed		less
	able		ish
Verb	ant		ful
	ed		ed
	ous		y
	al	**N**	some
	ive		ed
BM	y		y
Pron	some		ish
Verb/Aj		**N/Verb**	able

Figure 6.33. The matrix for the adverb construction {C-a-a}
(Key: Due to the limitation of space, the initiale column has been split into two)

Table 6.26. Morphological regularities for the adverb construction {C-a-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	Verb	*-ly*	*-ive-, -ing-, -ed-, -able-, ant-*	*addictively, cryingly, reportedly, distinguishably, rampantly*
2	BM	*-ly*	*-ed-, -ous-, -al, -ive-*	*debauchedly, acrimoniously, allegorically, abrasively*
3	N	*-ly*	*-al-, -ous-, -less-, -ish-, -ful-, -ed-, -y-, -some-*	*abysmally, hazardously, needlessly, slavishly, manfully, deucedly, angrily, troublesomely*
4	Aj	*-ly*	*-al-, -some-, -ful-*	*clinically, fulsomely, gratefully*
5	N/Verb	*-ly*	*-ed-, -ish-, -able-, -y-*	*rootedly, sluggishly, creditably, creakily*
6	Pron	*-ly*	*-y-*	*naughtily*
7	Verb/Aj	*-ly*	*-some-*	*wearisome*

6.1.13.2 The adverbial construction {a-C-a}

This construction involves 5 prefixes in its initiale slot, and 3 suffixes in its finale slot (Figure 6.34). The most type-valent initiales are the prefixes *in-* and *un-*, and the most-type-valent finale

is the suffix *-ly*. The most frequent mediale is a noun. Table 6.27 looks at the combination of morphemes in greater detail.

a/a	ly	s	ward
im	N/Aj/Ad		
in	Aj	N	
	Aj	Aj	
	N		
un	BM		
dis	N		
a			N

Figure 6.34. The matrix for the adverb construction {a-C-a}

Table 6.27. Morphological regularities for the adverb construction {a-C-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	Initiale *im-*	*-ly*	N/Ad	*improperly*
2	Initiale *in-*	*-ly, -s*	Aj, N	*inadequately, indoors*
3	Initiale *un-*	*-ly*	Aj, N, BM	*unhappily, untimely, unseemly*
4	Initiale *un-*	*-s*	Aj	*unawares*
5	Initiale *dis-*	*-ly*	N	*disorderly*
6	Initiale *a-*	*-ward*	N	*abackward*

6.1.13.3 The adverbial constructions {C-C-a} and {C-C-C}

Both constructions have a low formal productivity. The construction {C-C-a} is visualized in Figure 6.35. Nouns and pronouns fill its initiale slot, and the suffixes *-ly* and *-s* its finale slot. The mediale slot of this construction is occupied by nouns and adjectives. Table 6.28 gives examples for the established combinations of morphemes.

C/a	ly	s
N	Aj	N
Pron		N

Figure 6.35. The matrix for the adverb construction {C-C-a}

Table 6.28. Morphological regularities for the adverb construction {C-C-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	Initiale N	*-ly*	Aj	*headlongly*
2	Initiale N	*-s*	N	*lengthways*
3	Initiale Pron	*-s*	N	*sometimes*

The adverbial construction{C-C-C} is realized in the morphological patterns Ad”=Ad+Da+Ad, producing the word *nevertheless*.

6.1.14 The third- and fourth-level adverbial constructions

The third-level involves the following constructions: {C-C-a-a}, {a-C-a-a} and {C-a-a-a}. They are described in Figure 6.36 and Table 6.29. The combinations of morphemes in these constructions are similar to those found on the first level.

C/a	ly
dis	N+ful
re	Verb+ing
	Verb+ing
	Verb+ed
	N+ed
	N+ful
	BM+ous
un	N+able
sub	BM+ous
in	Verb+ed
	ic+al
BM	BM+al
Aj	en+ing
N	ic+al

Figure 6.36. The matrix for the adverb constructions {C-C-a-a}, {a-C-a-a} and {C-a-a-a}

Table 6.29. Morphological regularities for the adverb constructions {C-C-a-a}, {a-C-a-a} and {C-a-a-a}

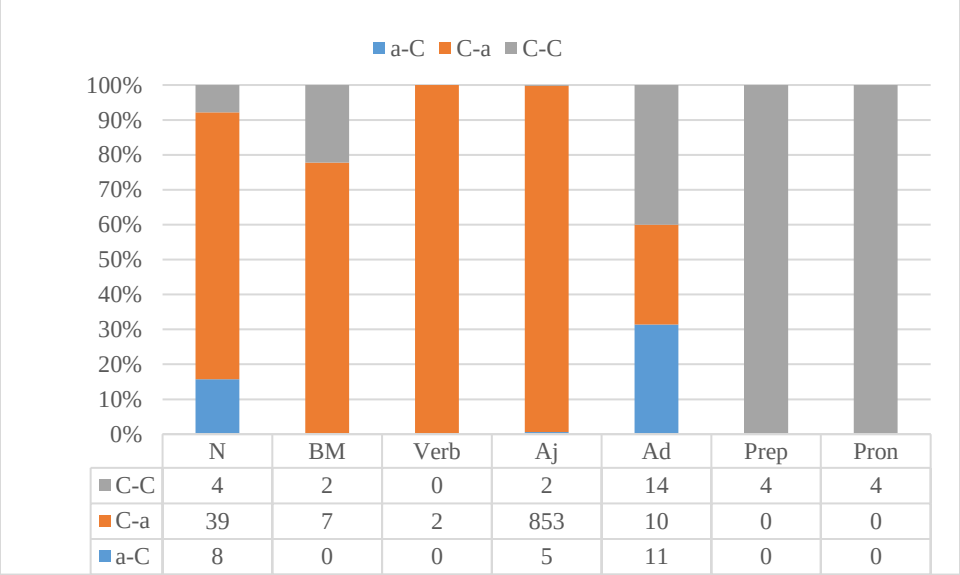
Construction	No	If		Then	Examples in the data
		Initiale	***Finale***	***Mediale***	
{a-C-a-a}	1	*dis-*	*-ly*	N+ful	*distrustfully*
	2	*re-*	*-ly*	Verb+ing	*reassuringly*
	3	*un-*	*-ly*	Verb+ing, Verb+ed, N+ed, N+ful, BM+ous, N+able	*unceasingly, undeservedly, unprecedentedly, unsuccessfully, unconsciously, unmistakably*
	4	*sub-*	*-ly*	BM+ous	*subconsciously*
	5	*in-*	*-ly*	Verb+ed	*inadvisedly*
{C-a-a-a}	1	BM	*-ly*	ic+al	*automatically*
	2	Aj	*-ly*	en+ing	*deadeningly*
	3	N	*-ly*	ic+al	*diametrically*
{C-C-a-a}	1	BM	*-ly*	BM+al	*photogenically*

The fourth level of adverbial formation is represented by a construction {C-a-a-a-a}, realized in two morphological patterns: Ad””=Verb+ist+ic+al+ly (producing the word *deterministically*) and Ad””=N+ist+ic+al+ly (producing the words *journalistically* and *statistically*).

6.1.15 The main trends in the formation of adverbs

The adverbial matrix analyses presented above have shed light on the formation of multimorphemic adverbs. It involves four levels. The first and second levels are the most formally productive. The adverbial-forming suffixes include *-ly*, *-wise*, *-ish*, *-er*, *-s* and *-ward*, with the suffix *-ly* contributing the most. Further, *un-* has the highest type valency among the identified

prefixes. Lastly, the most frequent combinations of suffixes include *-less + -ly*, *-ed + -ly*, *-al + -ly*, *-able + -ly* and *-ing + -ly*. As shown in Figure 6.37, these combinations are driven by a high type frequency of adjective bases in the formation of adverbs.



	N	BM	Verb	Aj	Ad	Prep	Pron
C-C	4	2	0	2	14	4	4
C-a	39	7	2	853	10	0	0
a-C	8	0	0	5	11	0	0

Figure 6.37. The shares of word bases in adverbial formation

6.1.16 The first level noun/adjective formation

Noun/adjective formation on the first level involves the morphological constructions {C-a}, {a-C} and {C-C}. Their morphological account is given in the following subsections.

6.1.16.1 The noun/adjective construction {C-a}

This construction displays a high morphological diversity. Seven (with CC) or 5 (without CC) word classes occupy its initiale slot, and 46 suffixes its finale slot. The most type-valent initiales include a bound morpheme noun and adjective. The suffixes *-al*, *-ist*, *-an*, *-ish*, *-y*, *-able*, *-ed*, *-ly* and *-o* are polyvalent. Almost half of the suffixes are monovalent. This construction is analyzed in Figures 6.38–6.40 and Table 6.30.

C/a	al	ist	an	ish	ly	o	ie	ed	ary	ic	ar	en	ery	ble	ive	ant	ous	ent	oid	ate
BM	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
N	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø								
Aj	Ø	Ø	Ø	Ø	Ø	Ø	Ø													
N/Aj	Ø	Ø	Ø										Ø							
Verb	Ø	Ø		Ø				Ø						Ø	Ø	Ø				
N/Verb								Ø												

Figure 6.38. The matrix for the noun/adjective construction {C-Ø-a} (Part1)

C/a	y	able	ing	ee	ling	less	ful	eer	ry	ese	ine	ock	red	et	th2	hood	i	esque
N	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
Aj	Ø																	
N/Aj	Ø																	
Verb	Ø	Ø	Ø	Ø														
N/Verb	Ø	Ø																
Ad					Ø													

Figure 6.39. The matrix for the noun/adjective construction {C-Ø-a} (Part2)
(Key: *-th2* represents a suffix forming ordinal numbers as in *millionth*)

C/a	er	s	some	most	ory	rel	ern
Aj	Ø	Ø	Ø	Ø			
Verb	Ø				Ø		
N/Verb						Ø	
Ad							Ø

Figure 6.40. The matrix for the noun/adjective construction {C-Ø-a} (Part3)

Table 6.30. Morphological regularities for the noun/adjective construction {C-Ø-a}

No	Type in the slot	Attach(es) to	Examples in the data
1	BM initiale	*-al, -ist, -an, -ish, -ly, -ist, -an, -ish, -ly, -o, -ie, -ed, -ary, -ic, -ar, -en, -ery, -ble, -ive, -ant, -ous, -ent, -oid, -ate*	*decimal, misogynist, pedestrian, Irish, early, video, indie, fond, tributary, psychotic, funicular, harden, nursery, deductible, effusive, rampant, intravenous, malevolent, tabloid, affectionate*
2	N initiale	*-al, -ist, -an, -ish, -ly, -ist, -an, -ish, -ly, -o, -ie, -ed, -ary, -ic, -ar, -en, -y, -able, -ing, -ee, -ling, -less, -ful, -eer, -ry, -ese, -ine, -ock, -red, -et, -th2, -hood, -i, -esque*	*adverbal, creationist, republican, boorish, stately, tango, veggie, valved, revolutionary, robotic, polar, golden, hairy, objectionable, derricking, jobless, hopeful, mountaineer, masonry, Japanese, Plasticine, bollock, kindred, packet, millionth, motherhood, Pakistani, picturesque*
3	Aj initiale	*-al, -ist, -an, -ish, -ly, -ist, -an, -ish, -ly, -o, -ie, -y, -er, -s, -some, -most*	*cubical, actualist, civilian, reddish, lowly, weirdo, smoothie, brawny, hinder, graphics, fulsome, innermost*
4	N/Aj initiale	*-al, -ist, -an, -ery, -y*	*domestical, idealist, cosmopolitan, greenery, scanty*
5	Verb initiale	*-al, -ist, -ish, -ed, -ble, -ive, -ant, -y, -able, -ing, -ee, -er, -ory*	*elemental, determinist, gibberish, accused, reversible, reflective, performant, employable, discerning, absentee, bumper, migratory*
6	N/Verb initiale	*-ed, -y, -able, -rel*	*spotted, smarmy, fashionable, mongrel*
7	Ad initiale	*-ling, -ern*	*underling, eastern*

6.1.16.2 The noun/adjective construction {a-C}

The second formally productive construction on the first level is that of prefixation. Its inner composition is introduced in Figure 6.41. The prominent feature of this construction is that adjectives and nouns have almost an equal share in the formation of this class, which points to its 'conversive' nature. Its initiale slot allows for 21 prefixes, the most type-valent of which are *anti-*, *sub-*, *non-*, *para-*, *a-*, *pre-* and *intra-*. Table 6.31 summarizes morphological regularities in this construction and illustrates the established combinations of morphemes with examples from the data.

sub	Ø	Ø			
anti	Ø	Ø	Ø		Ø
non	Ø	Ø			
para	Ø	Ø			
a	Ø	Ø			
pre	Ø	Ø			
intra	Ø		Ø		
dis	Ø			Ø	
in	Ø			Ø	
trans	Ø			Ø	
infra	Ø				
inter	Ø				
un	Ø				
ir	Ø				
pene	Ø				
up		Ø			
mis		Ø			
per		Ø			
post		Ø			
ultra		Ø			
de			Ø		

Figure 6.41. The matrix for the noun/adjective construction {a-Ø-C}

Table 6.31. Morphological regularities for the noun/adjective construction {a-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale *sub-*	Aj, N	*subconscious, subsoil*
2	*anti-*	Aj, N, BM, Verb	*antiseptic, antihistamine, antibiotic, anti-lock*
3	*non-*	Aj, N	*non-toxic, non-member*
4	*para-*	Aj, N	*paranormal, paramedic*
5	*a-*	Aj, N	*asexual, agnostic*
6	*pre-*	Aj, N	*prehistoric, prerequisite*
7	*intra-*	Aj, BM	*intrapluvial, intravenous*
8	*dis-*	Aj, N/Aj	*disagreeable, dissimilar*
9	*in-*	Aj, N/Aj	*incoherent, insolvent*
10	*trans-*	Aj, N/Aj	*transsexual, transatlantic*
11	*infra-*	Aj	*infra-red*
12	*inter-*	Aj	*international*
13	*un-*	Aj	*unintelligible*
14	*ir-*	Aj	*irresponsible*
15	*pene-*	Aj	*penultimate*
16	*up-*	N	*upstart*
17	*mis-*	N	*misfit*
18	*per-*	N	*peroxide*
19	*post-*	N	*post-war*
20	*ultra-*	N	*ultraviolet*
21	*de-*	BM	*demure*

6.1.16.3 The noun/adjective construction {C-C}

Nine (with CC) or eight (without CC) word classes have been observed in the initiale slot of this construction, and seven (with CC) or six (without CC) word classes in its finale slot. The most type-valent initiales are bound morphemes, verbs and nouns. Nouns and adjectives have the

highest type valency in the finale slot. Figure 6.42 demonstrates the inner morphological composition of this construction, and Table 6.32 provides examples for each combination.

C/C	N	Aj	BM	N/Aj	Ad	Pron	Verb
BM	Ø	Ø	Ø	Ø			
N	Ø	Ø	Ø				
Aj	Ø	Ø			Ø		
Verb	Ø				Ø	Ø	Ø
N/Verb	Ø						
Prep	Ø	Ø					
Part	Ø	Ø					
Num	Ø						
Ad					Ø		

Figure 6.42. The matrix for the noun/adjective construction {C-Ø-C}

Table 6.32. Morphological regularities for the noun/adjective construction {C-Ø-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale BM	N, Aj, BM, N/Aj	*autofocus, polytechnic, binocular, teetotal*
2	Initiale N	N, Aj, BM	*milestone, pea-green, nutmeg*
3	Initiale Aj	N, Aj, Ad	*rapid-fire, lukewarm, grown-up*
4	Initiale Verb	N, Ad, Pron, Verb	*killjoy, pullover, know-all, wannabe*
5	Initiale N/Verb	N	*cocktail*
6	Initiale Prep	N, Aj	*in-box, abovesaid*
7	Initiale Part	N, Aj	*no-go, no-good*
8	Initiale Num	N	*three-colour*
9	Initiale Ad	Ad	*roundabout*[39]

6.1.17 The second- and third level noun/adjective constructions

The second level of the noun/adjective formation encompasses the following constructions: {a-C-a}, {C-C-a}, {C-a-C}, {C-a-a} and {C-C-C}. The most formally productive construction is {a-C-a}, whereas the least productive is {C-C-C}, represented by the morphological pattern N/Aj"=Aj+Verb+Ad (producing the word *merry-go-round*). These constructions are introduced in Figure 6.43–6.46 and in Table 6.33–6.36.

a/a	able	ed	ive	some	ing	ist	ate	ic	ary
un	Verb	Verb	Verb	Aj	Verb				
dis	Verb								
ir	Verb								
in	Verb						BM		
re	Verb	Verb							
	Aj								
de		Verb							
contra			BM						
counter			Verb						
trans						Aj			
non						Verb		N	
con									BM

Figure 6.43. The matrix for the noun/adjective construction {a-C-a}

[39] The etymon of the noun/adjective *roundabout* is the adverb *round about*, which consists of two adverbs. It is also possible to consider that this word was formed by conversion.

Table 6.33. Morphological regularities for the noun/adjective construction {a-C-a}[40]

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	*un-*	*-able*	Verb	*unaccountable*
2	*un-*	*-ed*	Verb	*unmarried*
3	*un-*	*-ive*	Verb	*uncooperative*
4	*un-*	*-some*	Aj	*unwholesome*
5	*un-*	*-ing*	Verb	*unsettling*
6	*dis-*	*-able*	Verb	*disreputable*
7	*ir-*	*-able*	Verb	*irreconcilable*
8	*in-*	*-able*	Verb	*invaluable*
9	*in-*	*-ate*	BM	*innumerate*
10	*re-*	*-able*	Verb, Aj	*rechargeable, renewable*
11	*re-*	*-ed*	Verb	*refined*
12	*de-*	*-ed*	Verb	*demobilized*
13	*contra-*	*-ive*	BM	*contraceptive*
14	*counter-*	*-ive*	Verb	*counteractive*
15	*trans-*	*-ist*	Aj	*transsexualist*
16	*non-*	*-ist*	Verb	*nonconformist*
17	*non-*	*-ic*	N	*non-alcoholic*
18	*con-*	*-ary*	BM	*contemporary*

C/a	ic	ing	er
BM	BM		
N		Verb	
Part			N
Verb			N

Figure 6.44. The matrix for the noun/adjective construction {C-C-a}

Table 6.34. Morphological regularities for the noun/adjective construction {C-C-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	BM	*-ic*	BM	*homophobic*
2	N	*-ing*	Verb	*kidnapping*
3	Part or Verb	*-er*	N	*no-brainer, do-gooder*

C/C	N	Verb	Aj
N	o		
Verb		and	
Aj			ly

Figure 6.45. The matrix for the noun/adjective construction {C-a-C}

Table 6.35. Morphological regularities for the noun/adjective construction {C-a-C}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	N	N	*-of-*	*matter-of-fact*
2	Verb	Verb	*-and-*	*park-and-ride*
3	Aj	Aj	*-ly-*	*newly-wed*

[40]All words in this category have been identified as adjectives and nouns by the OED. This classification means that a primary syntactic function of these words is that of adjectives but occasionally they also act as nouns. The same explanation applies to Table 6.36.

C/a	ic	al	ed	ist	y
BM	ist			al	
N	ist	ic	ify	al	
Verb					er

Figure 6.46. The matrix for the noun/adjective construction {C-a-a}

Table 6.36. Morphological regularities for the noun/adjective construction {C-a-a}

No	If		Then	Examples in the data
	Initiale	***Finale***	***Mediale***	
1	BM	*-ic*	*-ist-*	*linguistic*
2	BM	*-ist*	*-al-*	*serialist*
3	N	*-ic*	*-ist-*	*journalistic*
4	N	*-al*	*-ic-*	*pathological*
5	N	*-ed*	*-ify-*	*classified*
6	N	*-ist*	*-al-*	*factionalist*
7	Verb	*-y*	*-er-*	*confectionery*

The third level of noun/adjective formation involves three constructions {a-a-C-a}, {C-C-C-C} and {a-C-a-a}, producing the following words: *do-it-yourself*, *non-recyclable* and *unsweetened*.

6.1.18 The main trends in noun/adjective formation

The matrix analysis of nouns/adjectives has established that this conversive class combines the properties of nouns and adjectives. First, adjective and noun bases have an almost equal share in the meta-construction {{a-C}} (Figure 6.47), whereas noun and adjective bases are more important in noun and adjective formation, respectively. Moreover, as compared to nouns and adjectives, this conversive class shows less involvement of verb bases and, as compared to adjectives, greater involvement of adjective bases. Finally, the most frequent combinations of suffixes include *-ist* + *-ic*, *-al* + *-ist*, *-ic* + *-al* and *-ify* + *-ed*.

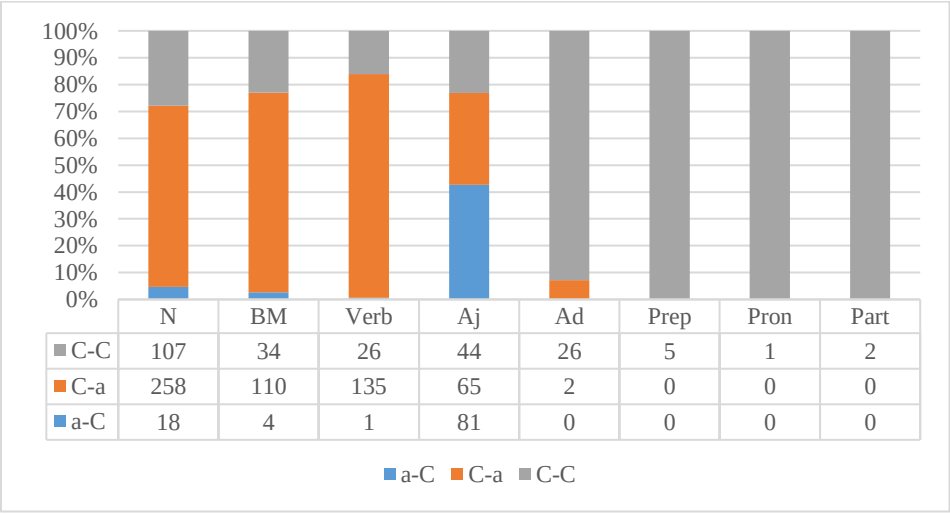


Figure 6.47. The shares of word bases in the formation of nouns/adjectives across three meta-constructions

6.1.19 The adjective/adverb formation

Adjective/adverb formation mainly evolves on two levels. The first level includes the constructions {C-a}, {a-C} and {C-C}. Their morphological composition is given in Figure 6.48–6.51 together with examples from the data in Table 6.37–6.40.

C/a	ish	less	ing	ful	like	y	ous	ed	some	able	ly	ling	fold
N	Ø	Ø		Ø	Ø	Ø	Ø	Ø	Ø				
Verb	Ø	Ø	Ø			Ø				Ø			
Aj	Ø										Ø	Ø	
BM							Ø	Ø			Ø		
N/Aj											Ø		Ø
N/Aj/Ad											Ø		
N/Verb						Ø							

Figure 6.48. The matrix for the adjective/adverb construction {C-a}

Table 6.37. Morphological regularities for the adjective/adverb construction {C-a}[41]

No	Type in the slot	Attach(es) to	Examples in the data
1	N	*-ish, -less, -ful, -like, -y, -ous, -ed, -some*	*sheepish, endless, soulful, warlike, roomy, ravenous, dogged, awesome*
2	Verb	*-ish, -ing, -y, -able*	*fiendish, exceeding, faulty, answerable*
3	Aj	*-ish, -ly, -ling*	*sharpish, entirely, darkling*
4	BM	*-ous, -ed, -ly*	*monotonous, gingerly*
5	N/Aj	*-ly, -fold*	*haggardly, twelvefold*
6	N/Aj/Ad	*-ly*	*suddenly*
7	N/Verb	*-y*	*canny*

a/C	Aj	N
un	Ø	
ir	Ø	
a	Ø	Ø
in		Ø
pre		Ø
up		Ø

Figure 6.49. The matrix for the adjective/adverb construction {a-C}

Table 6.38. Morphological regularities for the adjective/adverb construction {a-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale *un-*	Aj	*unequal*
2	*ir-*	Aj	*irrespective*
3	*a-*	Aj, N	*afoot, aloud*
4	*in-*	N	*indoor*
5	*pre-*	N	*pre-war*
6	*up-*	N	*upmarket*

[41] Similar to the conversive class of nouns/adjectives, words in this category are identified as adjectives and adverbs by the OED.

C/C	Prep	BM	Ad	N
N	Ø	Ø		
BM		Ø		
Part			Ø	
Aj				Ø
Prep				Ø

Figure 6.50. The matrix for the adjective/adverb construction {C-C}

Table 6.39. Morphological regularities for the adjective/adverb construction {C-C}

No	Type in the slot	Attach(es) to	Examples in the data
1	Initiale N	Prep, BM	*herewith, clockwise*
2	BM	BM	*often*
3	Part	Ad	*nohow*
4	Aj	N	*sometime*
5	Prep	N	*online*

The second level of adjective/adverb derivation includes the constructions {C-a-C}, {C-C-C}, {C-C-a}, {a-C-C} and {C-a-a}, which are analyzed in Figure 6.51 and Table 6.40.

C/C	N	BM	ed	ful	able	s
N	a		Verb			
	Prep					
Ad	Prep					
BM						ward
anti		N				
un			Verb	N	Verb	

Figure 6.51. The matrix for the adjective/adverb constructions {C-a-C}, {C-C-C}, {C-C-a}, {a-C-C} and {C-a-a}

Table 6.40. Morphological regularities for the adjective/adverb constructions {C-a-C}, {C-C-C}, {C-C-a}, {a-C-C} and {C-a-a}

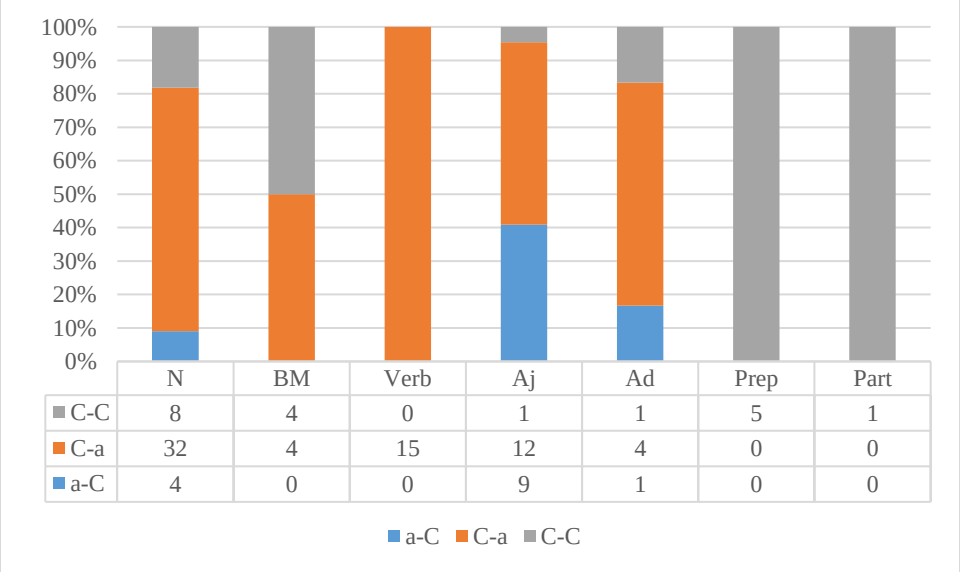
Construction	No	If		Then	Examples in the data
		Initiale	***Finale***	***Mediale***	
{C-a-C}	1	N	N	-a-	*chock-a-block*[42]
{C-C-C}	1	N	N	Prep	*person-to-person*
	2	Ad	N	Prep	*up-to-date*
{C-a-a}	1	BM	*-s*	*-ward*	*outwards*
{a-C-C}	1	*anti-*	BM	N	*anticlockwise*
{C-C-a}	1	N	*-ed*	Verb	*jam-packed*
{a-C-a}	1	*un-*	*-ed*	Verb	*undoubted*
	2	*un-*	*-ful*	N	*unlawful*
	3	*un-*	*-able*	Verb	*unpardonable*

[42] The etymology of the morpheme *-a-* in *chock-a-block* is not known with certainty. It is identified as a connective morpheme by the OED, which has probably evolved from the conjunction *and*. For this reason, it is also possible to assign this word to the construction {C-C-C}.

The third level of adjective/adverb formation is represented by compounding, involving the following morphological patterns: Aj/Ad””=Prep’+Da+N (*across-the-board*), Ad/Aj”’=Ad+Prep+Da+N (*up-to-the-minute*) and Ad/Aj”=Verb+y+Verb+y (*willy-nilly*).

6.1.20 The main trends in the formation of adjectives/adverbs

Adjective/adverb derivation is diverse, especially when considering a small number of words in this class. The first level produces the largest number of words, with the construction {C-a} being the most formally productive. The most type-valent word bases in the construction are nouns (with a type valency of 8) and verbs (with a type valency of 5). The suffixes *-ly*, *-ish*, *-less*, *-y*, *-ous* and *-ed* show polyvalency. In the construction {a-C}, on the other hand, adjectives and nouns have the greatest contribution as word bases (Figure 6.52). The prefix *a-* is the only duovalent morpheme within this construction. Further, nouns and bound morphemes have a more significant role in the construction {C-C}. Finally, the only observed combination of suffixes within this class is *-ward* + *-s*.



	N	BM	Verb	Aj	Ad	Prep	Part
C-C	8	4	0	1	1	5	1
C-a	32	4	15	12	4	0	0
a-C	4	0	0	9	1	0	0

Figure 6.52. The shares of word classes in the formation of adjectives/adverbs

6.2 Morphological paradigms of English word formation as networks

In the previous sections, I have analyzed formal constructions across six major word classes. The optimal matrix analysis has proven useful in establishing the formal productivity of constructions, as well as the type valency of their slots. Further, the matrices have revealed various combinations of suffixes and have brought to light morphemes which have a significant contribution within

constructions. With the information acquired through the optimal matrix analysis, the graph network analysis of morphological constructions has become possible and will be discussed in the following subsections.

Within cognitive and usage-based approaches, language is often viewed as a network with different nodes and types of connections, in which activation of one node spills over to the neighboring nodes (Hudson 2007: 512). If we view morphemes as the special case of a well-entrenched and recombinable connections of a sound string with meaning (Kemmer 2003: 25), then a construction appears as a limiting construct that determines the boundaries of these connections. Hence, the variation of types in the slot of a construction can be perceived as a morphological network, which I have termed 'formal morphological paradigm'. With the view of a morphological construction as a network of morphemes, the tools developed on the basis of graph theory are especially useful (see Section 4.5).

Therefore, the strength and the distance between morphemes within each studied construction are visualized in Figure 6.53–6.70 with the help of a network algorithm based on eigenvector centrality, which 'is a measure of how important the node is in the context of the entire graph' (Desagulier 2017: 286–287). It assigns a higher weight to nodes connected to important nodes. In these graphs, morphemes are represented as nodes ('vertices') and connections between them by lines ('edges'). The type valency of each morpheme is reflected in a number of connections it holds to other nodes, and its type frequency in a size of a node's diameter. The importance of a morphemes in the context of a morphological construction is specified by its location—more significant nodes are located closer to the center of a graph—and by darker colours of a heat map. Consequently, red bigger circles at the center of graphs symbolize high scores of centrality: they can be thought of as morphological hubs. On the other hand, yellow circles represent nodes with lower scores of centrality. All graphs have been created in *R* (R Core Team 2021) with the 'igraph' package and with a piece of code written by Desagulier (2017: 286–287).

The graphs have been created for three meta-constructions—{{C-a}}, {{a-C}} and {{C-C}}—across six largest word classes in the metacorpus (N, Aj, Verb, Ad, N/Aj and Aj/Ad). These meta-constructions capture word formation on all structural levels and with the consideration of the last and most recent process in the formation of a word. For example, the homonyms *reformer* and *re-former* belong to the morphological construction {a-C-a}. However, the former has been

derived as *reform* + *-er*, and the latter as *re-* + *former*.[43] Therefore, these nouns have been assigned to different meta-constructions: *reformer* to {{C-a}} and *re-former* to {{a-C}}. Table 6.41 presents the classification of morphological constructions based on the observed data and considering individual differences of words. As illustrated in the table, the noun construction {a-C-a} belongs to both meta-constructions {{a-C}} and {{C-a}}.

Table 6.41. The distribution of different morphological constructions across three major meta-construction

	{{a-C}}	**{{C-a}}**	**{{C-C}}**
N	{a-Ø-C}, {a-C-a-C}, {a-C-a-a-a}	{C-Ø-a}, {C-a-a}, {C-a-a-a}, {C-C-C-a}, {C-C-a-a}	{C-Ø-C}, {C-a-C}, {C-C-C}, {C-a-C-C-a}
		{C-C-a}, {C-a-C-a}	
	{a-C-a}		
Aj	a-Ø-C}, {a-a-C-a}, {a-C-a-C}	{C-Ø-a}, {C-a-a}, {C-a-a-C}	{C-Ø-C}, {C-a-C-C}
		{C-C-a}, {C-a-C-a}	
	{a-C-a}, {a-C-C-a}		
Verb	{a- Ø-C}	{C-Ø-a}, {C-a-a}	{C-Ø-C}, {C-C-a}
	{a-C-a},{C-C-a}, {a-C-a-a}		
Ad	{a-Ø-C}	{C-Ø-a}, {C-a-a}{a-C-a}, {C-C-a}, {C-C-a-a}, {a-C-a-a}, {C-a-a-a}	{C-Ø-C}
N/Aj	{a-Ø-C}	{C-Ø-a}, {C-C-a}, {C-a-a}	{C-Ø-C}, {C-a-C}
	{a-C-a}		
Aj/Ad	{a-Ø-C}, {a-C-C}, {a-C-a}	C-Ø-a}, {C-C-a}, {C-a-a}	{C-Ø-C}, {C-a-C}, {C-C-C}, {C-a-C-a}

6.2.1 The formal noun formation paradigm {{C-a}}

Figure 6.53 represents the noun formation paradigm {{C-a}}. The morphological network for this paradigm has the most complex structure in English word formation that evolves around five word-class nodes: Verb, N, Aj, Ad and BM. Except for the adjective and adverb nodes, all nodes allow for the attachment of monovalent suffixes. Suffixes that attach only to nouns are represented at the top of the graph (e.g. *-let*, *-dom*), whereas suffixes that attach only to verbs on the left bottom corner (e.g. *-ar*, *-our*). Suffixes that bind only with bound morphemes are shown in the right bottom corner (e.g. *-on*, *-yl*). Among all the word bases, the most frequent is a verb base, as evident from the largest size of the verb node in the graph, although it attaches to a smaller number of monovalent suffixes.

There are fewer duovalent suffixes, which are located closer to the nodes in the center. Suffixes attaching to nouns and adjectives are *-ia*, *-ship*, *-ling*, *-hood*, *-ese* and *-ard*, and the suffixes attaching to nouns and bound morphemes *-s-pl* and *-ium*. The duovalent suffixes *-or*, *-al*, *-ant* and *-ure* bind with verbs and bound morphemes, and the suffixes *-ess* and *-ette* with verbs and

[43] "reformer, n.2." OED Online. Oxford University Press, June 2021.

nouns. The trivalent suffixes are located around the center. They encompass the suffixes *-ate*, *-s*, *-ine*, *-in* and *-ian* that attach to nouns, adjectives and bound morphemes, the suffix *-th* that attaches to verbs, adjectives and bound morphemes, as well as the suffixes *-ster* and *-ie* that attach to nouns, adjectives and verbs. The polyvalent suffixes occupy the very heart of the graph. They have a higher type frequency, as evident from the larger size of their nodes. These suffixes include *-ery*, *-ance*, *-er*, *-ment*, *-ism*, *-y*, *-ee*, *-ity*, *-ion* and *-ist* with a type valency of 4 (N, Aj, Verb and BM) and *-ness*, *-ing* and *-age* with a type valency of 5 (N, Aj, Verb, Ad and BM).

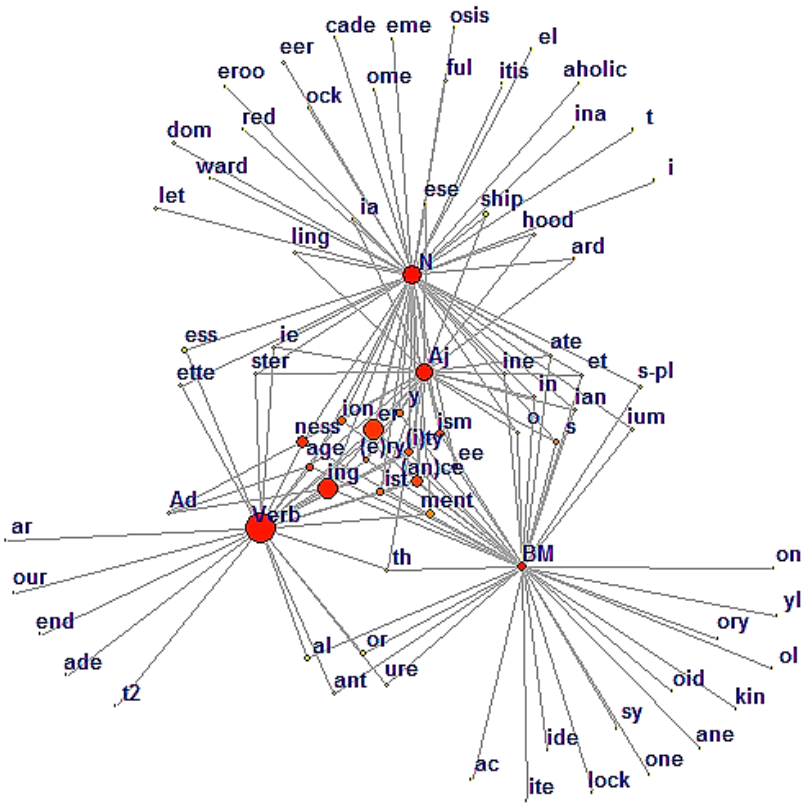


Figure 6.53. The formal morphological paradigm for the meta-construction {{C-a}}
(the suffixes |*-ery*|, |*-ance*| and |*-ity*| are represented as allomorphs)

6.2.2 The formal noun formation paradigm {{a-C}}

A different picture arises for the prefixation in nouns (Figure 6.54). It has one major noun node, surrounded by prefixes. Among the monovalent prefixes, *dis-*, *fore-*, *sub-*, *mis-*, *un-*, *co-* and *counter-* have a more prominent role, as is obvious from the size and colour of their nodes. On the other hand, the contribution of the prefixes *out-*, *trans-*, *sur-*, *after-*, *im-*, *ac-*, *arch-* and *infra-* is insignificant. After the noun node, the second most frequent word base in prefixation is a bound

morpheme that attaches to 8 prefixes, three of which are monovalent (i.e. *con-*, *en-* and *peri-*). Verbs and adjectives do not bind with monovalent suffixes. The most prominent prefix in this paradigm is *re-*, which has the highest type valency (N, Aj, Verb and BM) and the highest type frequency. The prefix *anti-* is trivalent, binding with nouns, adjectives and bound morphemes. Other salient prefixes, which are duovalent, include *up-* and *in-* (attaching to verbs and nouns), *non-* (attaching to nouns and adjectives), *de-*, *on-* and *contra-* (attaching to nouns and bound morphemes).

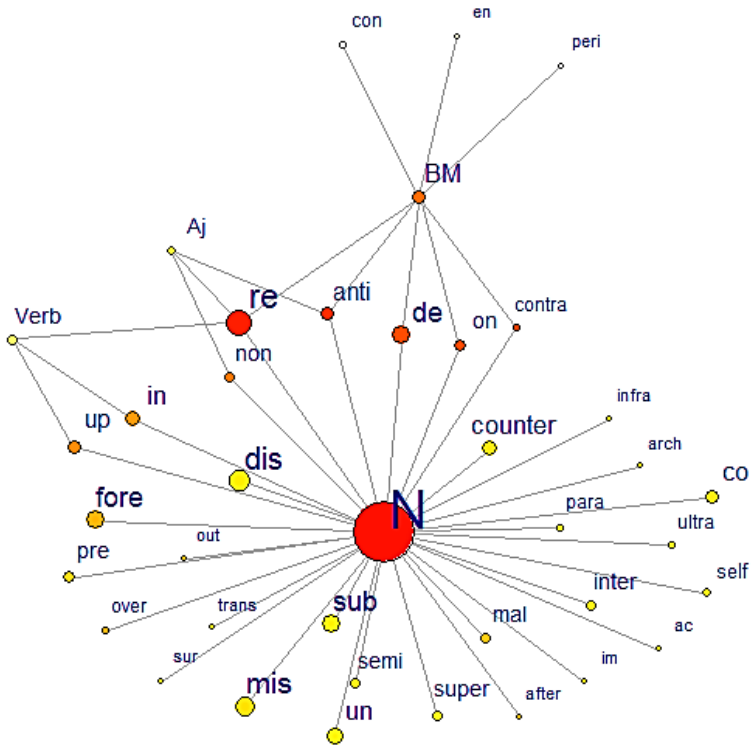


Figure 6.54. The formal morphological paradigm for the meta-construction {{a-C}}

6.2.3 The formal noun formation paradigm {{C-C}}

Although different, the architecture of the two previous graphs for noun suffixation and prefixation are similar in that they have dominant nodes (word classes) surrounded by smaller nodes (suffixes)—bearing a resemblance to the structure of a flower. The nature of compounding is different: because the number of word classes is limited in a language, the combination of morphological elements in compounding is also limited. Figure 6.55 shows the architecture of compounding in English nouns. First and foremost, nouns in the position of the finale (N2) have the biggest size node as well as the highest score of centrality as represented by the darker red

colour, which is evidence for the right-headedness of English noun compounds. In Marchand's (1969) terminology, the role of the *determinantum* (i.e. the finale slot of the construction) can be taken up by nouns, conjunctions, verbs, adverbs, bound morphemes and pronouns. The *determinant* (i.e. the initiale slot) of compounds can be numerals, adverbs, prepositions, nouns, verbs, pronouns, adjectives and bound morphemes. In the sample, no adjectives have been observed in the finale slot of the construction or in the slot of the *determinantum*. Lastly, the most frequent combinations in noun compounding are N+N and Verb+N.

This graph represents only morphological compounding. However, other types of compounding have been observed in the sample, which can be characterized as non-morphological: blending (e.g. *spam*; 22 words), reduplication (e.g. *chit-chat*; 3 words) and initialism (*IQ*; 174 words). Since other processes are involved in their formation (e.g. phonological), they have not been considered in the current analysis.

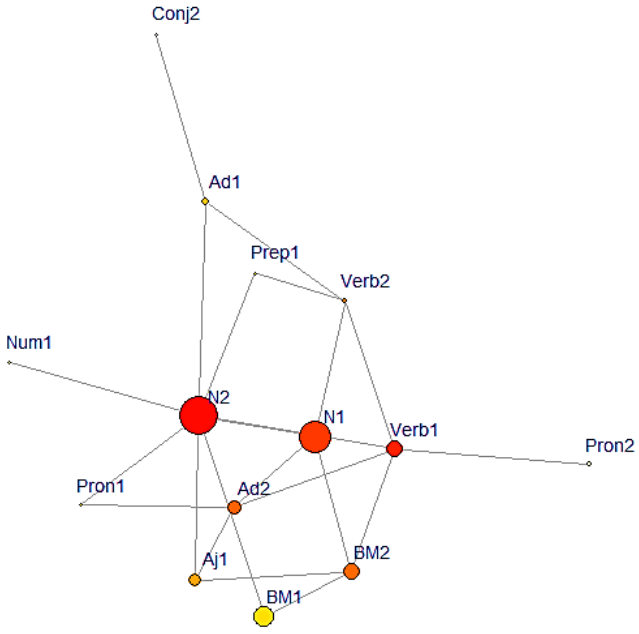


Figure 6.55. The formal morphological paradigm for the meta-construction {{C-C}}(the number next to a word class represents the place in the meta-construction:1 stands for the initiale and 2 for the finale)

6.2.4 The formal verb formation paradigm {{C-a}}

The architecture of suffixation in verbs differs from that of nouns. As illustrated in Figure 6.56, the number of morphemes in this meta-construction is limited—hence, its structure resembles that of noun compounding, discussed in the previous section. The center of the graph is occupied by a noun node: although it has a lower type frequency than that of an adjective, its importance is higher in this graph, because it connects to a greater number of suffixes (i.e. *-eer*, *-en*, *-ize*, *-ate*, *-ify*, *-le*

and *-er*). None of the suffixes which attach to noun bases are monovalent. The verb and bound morpheme nodes are also important, as evident from their colour. In verb formation, verb bases attach to the monovalent suffixes *-ish* and *-age*, to the duovalent suffix *-eer* and the polyvalent suffixes *-en*, *-er* and *-le*, whereas bound morpheme bases attach to the monovalent suffixes *-ise*, *-yse* and *-ic*, and the polyvalent suffixes *-le*, *-er*, *-en*, *-ate*, *-ify* and *-ize*. Similar to nouns, adjective bases attach to polyvalent suffixes. The most polyvalent suffix is *-en* which binds with five word classes (adverbs, verbs, nouns, adjectives and bound morphemes). The suffixes *-ize* and *-ify* are similar in that they attach to nouns, adjectives and bound morphemes. Further, the suffix *-ize* has the highest type frequency.

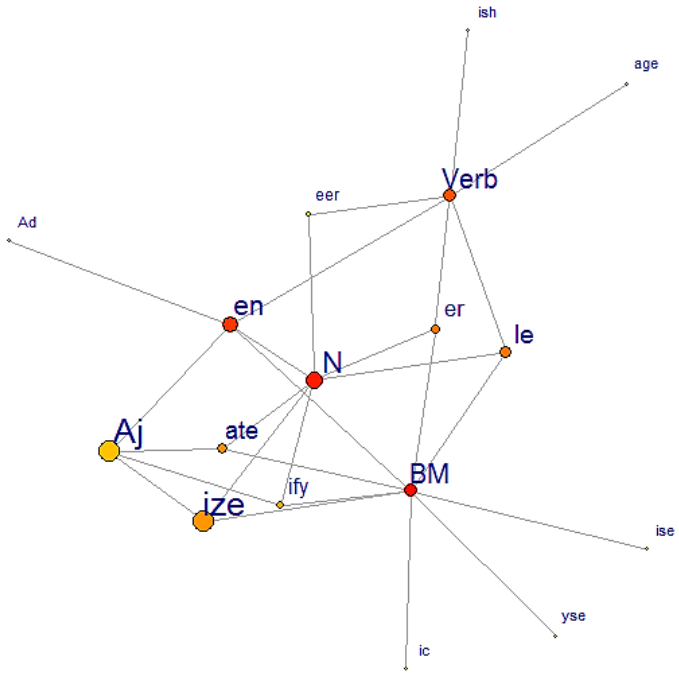


Figure 6.56. The formal morphological paradigm for the meta-construction {{C-a}}

6.2.5 The formal verb formation paradigm {{a-C}}

As evident from Figure 6.57, the architecture of the network in verb prefixation is similar to that of nouns. The major node in the construction is the same word class as that of the construction, i.e. a verb. It attaches to 12 monovalent (the top of the graph) and 12 duo- and polyvalent prefixes (the bottom of the graph). The second important node is that of a noun, which attaches to 9 prefixes, followed by adjective and bound morpheme nodes. Further, noun and adjective bases do not attach to monovalent prefixes. The prefixes *pre-*, *in-*, *dis-* attach to verbs and nouns, the prefixes *un-* and

out- to verbs and adjectives, the prefix *co-* to verbs and bound morphemes, and the prefix *a2-* (of Old English origin) to verbs and nouns. The prefixes *be-*, *re-*, *en-*, *de-* and *im-* are polyvalent. Similar to noun prefixation, the prefixes *re-*, *de-* and *en-* have the most prominent role in this verb construction, both in terms of the type frequency and the type valency.

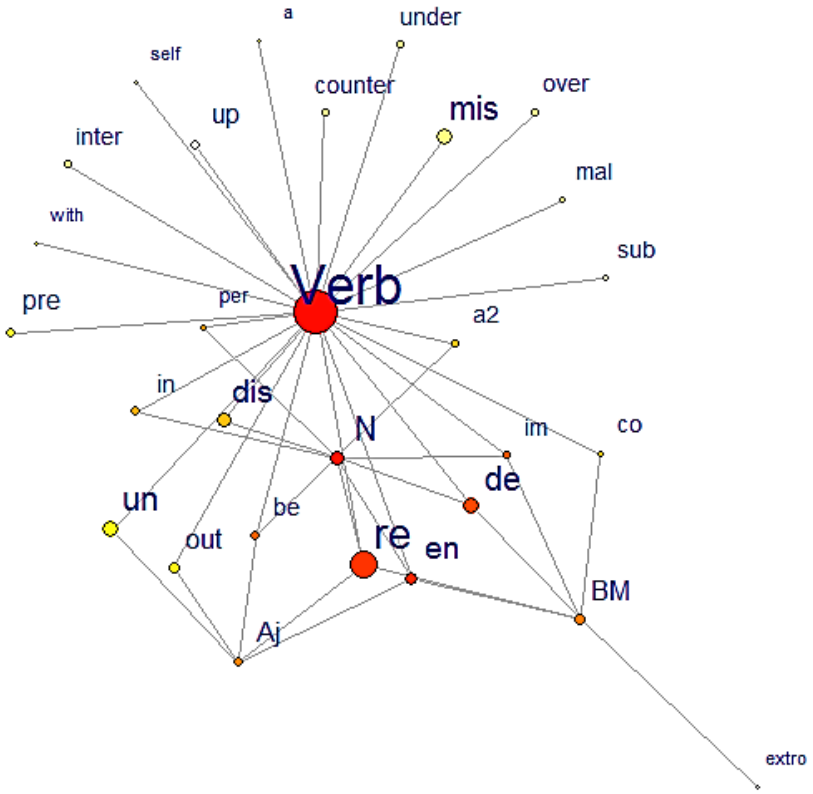


Figure 6.57.The formal morphological paradigm for the meta-construction {{a-C}}

6.2.6 The formal verb formation paradigm {{C-C}}

The architecture of compounding in verbs is simple (Figure 6.58) and involves the word bases of the following morphological classes: nouns, verbs, bound morphemes, adjectives, adverbs, prepositions and pronouns. The simple structure of this paradigm suggests a lower significance of compounding for verb formation. Similar to noun compounding, the major node belongs to that of a finale and is a verb (the same word class as the construction). This node also has the highest type valency (4) by attaching to verbs, nouns, adjectives and bound morphemes. The bound morpheme *determinantum* binds with a bound morpheme and a noun, and the noun *determinantum* with a noun and a preposition. The adverb and the pronoun *determinantums* are monovalent and attach to a verb and a preposition, respectively. As in noun compounding, an adjective is not observed in the slot of the finale.

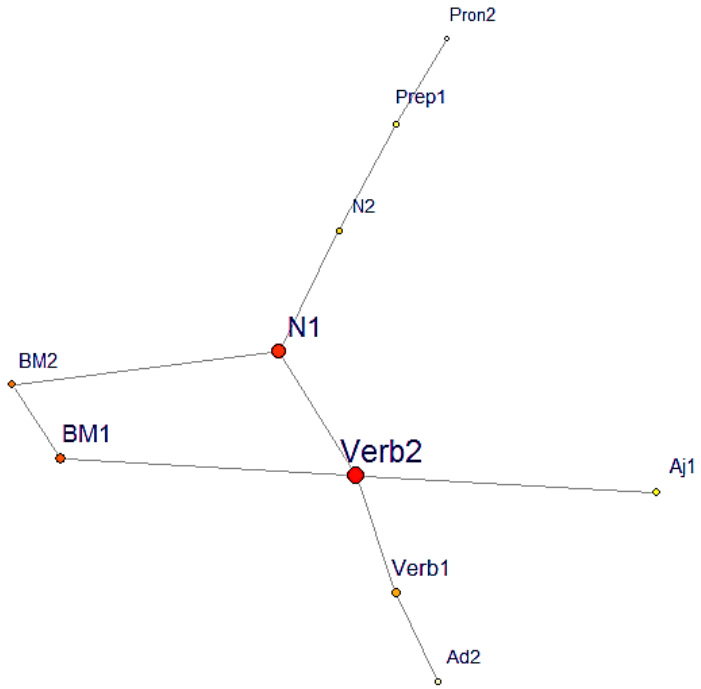


Figure 6.58. The formal morphological paradigm for the construction {C-C} for verb formation

6.2.7 The formal adjective formation paradigm {{C-a}}

The word-formation paradigm of adjective suffixation, visualized in Figure 6.59, shows slightly different properties than that of noun suffixation. Firstly, there is a relatively smaller number of monovalent suffixes in this paradigm. Secondly, the 'division of labour' is more evenly distributed across the paradigm, which is evident from a higher number of the relatively large circles at the heart of the graph. The monovalent suffixes include *-an*, *-esque* and *-like* (attaching to nouns), *-ant*, *-sy*, *-ible* and *-le* (attaching to verbs), and *-id* (attaching to bound morphemes). Other suffixes are either duovalent (located closer to the center) or polyvalent (located at the center). The most type-polyvalent suffixes are *-full*, *-ic*, *-ive*, *-less*, *-ish* and *-y*, and the most type-frequent suffixes are *-ed*, *-ing*, *-al* and *-able*. Among the word class nodes, verbs and nouns have the highest importance, followed by bound morpheme and adjective nodes. The adverb and pronoun nodes have little significance and bind with the suffixes *-ly* and *-y*, respectively.

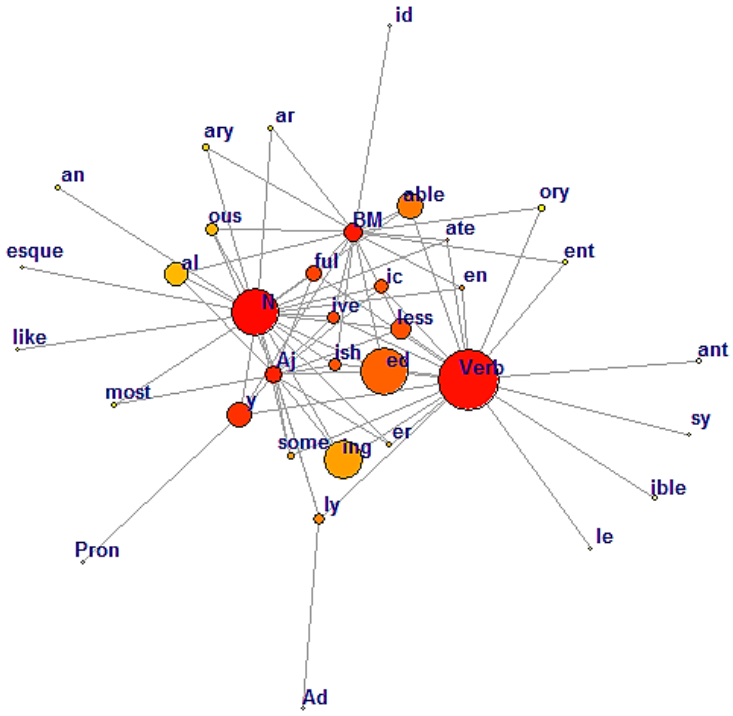


Figure 6.59. The formal morphological paradigm for the meta-construction {{C-a}} for adjective formation (as mentioned earlier, the importance of morphemes in the graph is specified by darker colours in the heat map)

6.2.8 The formal adjective formation paradigm {{a-C}}

The flower-like architecture of the prefixation adjective paradigm looks similar to that of nouns and verbs (Figure 6.60). The dominant node is that of an adjective—the same word class as the paradigm. The monovalent prefixes are mainly represented on the left side and at the top of the graph (e.g. *mis-*, *self-*, *ultra-*), whereas duovalent prefixes are positioned between adjective and noun nodes (e.g. *pre-*, *ex-*, *anti-*). The prefix with the highest type valency is *in-*, binding with adjectives, nouns and bound morphemes, and the prefix with the highest type frequency is *un-* binding with adjectives. Adverb, verb and bound morpheme nodes have little contribution to the whole paradigm, shown as extra-extensions from its structure.

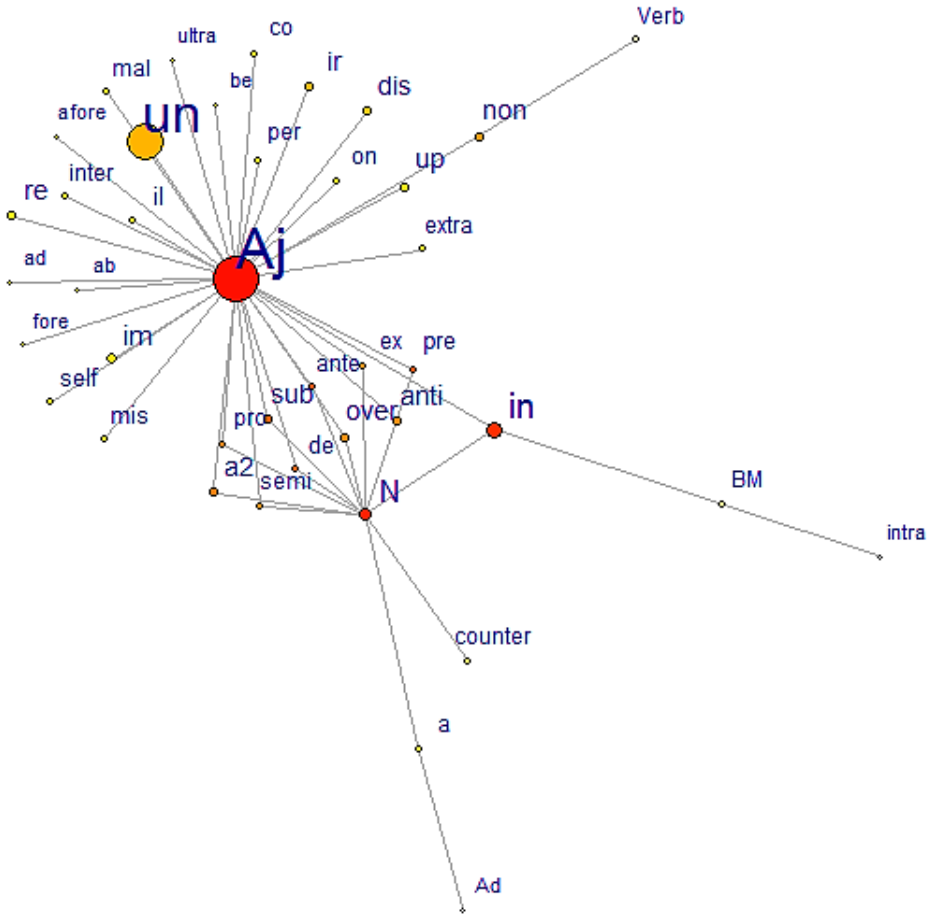


Figure 6.60. The formal morphological paradigm for the meta-construction {{a-C}} for adjective formation

6.2.9 The formal adjective formation paradigm {{C-C}}

The network structure of adjective compounding in the sample has a hexagon structure (Figure 6.61). Its most type-frequent node is that of an adjective in the position of the finale with the type valency of three. However, unlike noun and verb compounding, this *determinantum* slot is assigned a lower centrality score in the graph which is obvious from its orange colour.[44] Bound morphemes and nouns in the slot of the initiale, as well as nouns in the slot of the finale, have higher centrality scores, which highlights their importance for this paradigm. The most type-valent node is that of a noun in the position of the finale, attaching to adjectives, nouns, bound morphemes, prepositions and particles. Bound morphemes in the position of the finale attach to bound morphemes and nouns, and adverbs to adjectives and verbs. No *determinantum* is monovalent in the paradigm.

[44] The meaning of colours in the context of network graphs is explained on page 202.

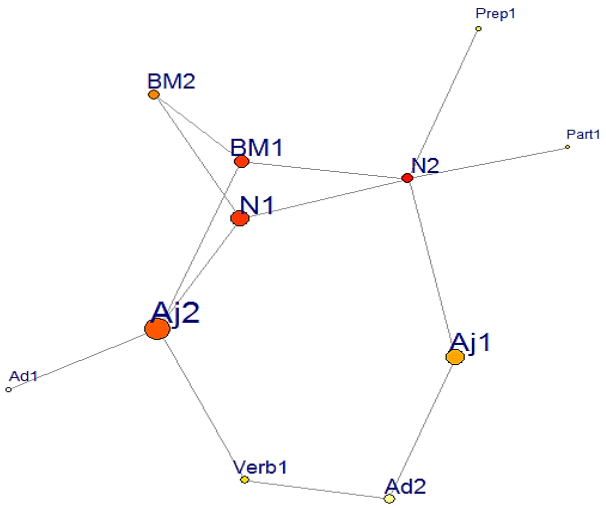


Figure 6.61. The formal morphological paradigm for the meta-construction {{C-C}} for adjective formation

6.2.10 The formal adverb formation paradigm {{C-a}}

The suffixation paradigm in adverbs is interesting in that its dominant node is that of a suffix and not a word class, as observed in the previous graphs (Figure 6.62). The suffix *-ly*, attaching to six word classes, has the highest type valency and is central to the paradigm. The most type-frequent node is that of an adjective, which, in addition to polyvalent *-ly*, attaches to a monovalent suffix *-wise*. Although less frequent, noun and adverb nodes are also important for the paradigm. With its joint link to the suffix *-ly*, the adverb node binds with the monovalent suffixes *-s*, *-ish*, *-er* and *-ward*.

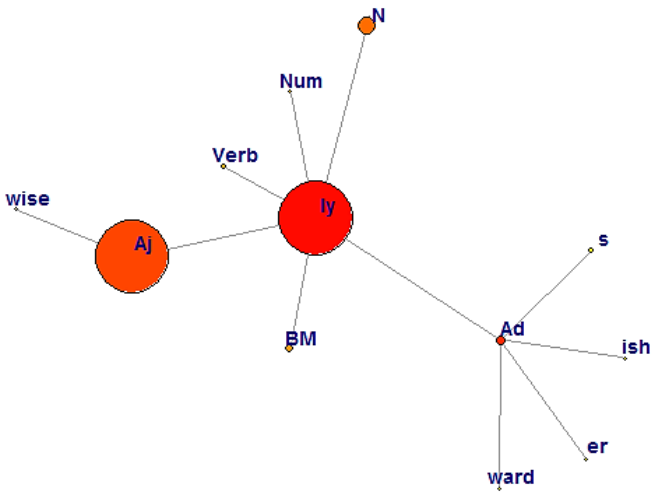


Figure 6.62. The formal morphological paradigm for the meta-construction {{C-a}} for adverb formation

6.2.11 The formal adverb formation paradigm {{a-C}}

The prefixation paradigm in adverbs is simple (Figure 6.63) and is one of the less formally productive in English word formation. Its most significant nodes are that of adverbs and the prefix *a-*. They both have a type valency of 3: adverbs bind with the prefixes *a-*, *un-* and *in-*, and the prefix *a-* with adjectives, nouns and adverbs. The only monovalent prefix is *up-* (connecting to nouns), and the prefixes *in-* and *un-* are duovalent (connecting to nouns and adverbs, and adverbs and adjectives, respectively).

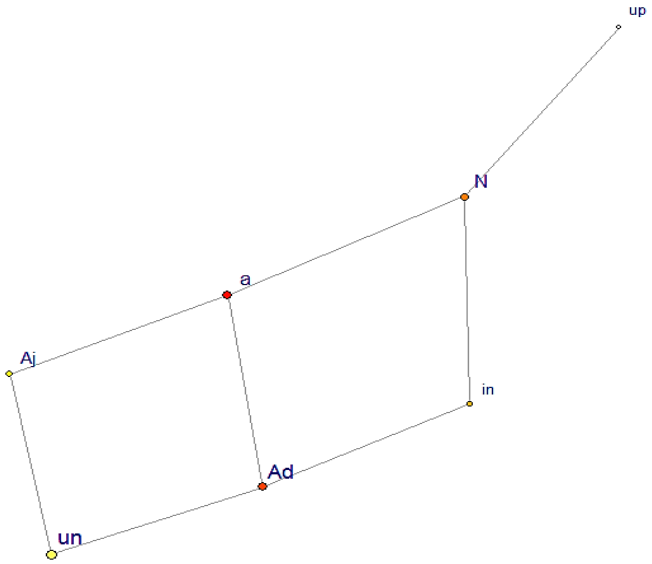


Figure 6.63. The formal morphological paradigm for the construction {{a-C}} for adverb formation

6.2.12 The formal adverb formation paradigm {{C-C}}

As illustrated in Figure 6.64, the network of adverb compounding revolves around an adverb in the position of the initiale and finale, showing the highest scores of centrality. Five word classes occupy the *deteminantum* slot for this construction: adjectives, adverbs, nouns, prepositions and bound morphemes. Noun finale bases in adverb compounding bind with prepositions and pronouns, adverb bases with adverbs and pronouns, adjective bases with bound morphemes and adverbs, and preposition and bound morpheme bases with adverbs.

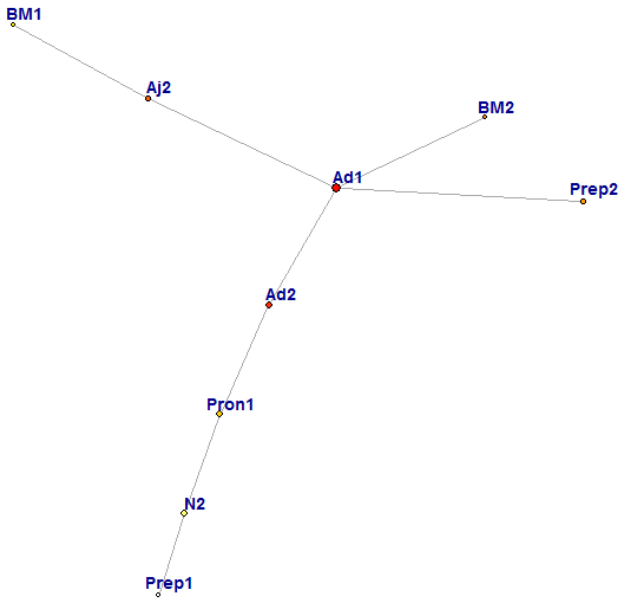


Figure 6.64. The formal morphological paradigm for the construction {{C-C}} for adverb formation

6.2.13 The formal noun/adjective formation paradigm {{C-a}}

The architecture of the noun/adjective suffixation paradigm, presented in Figure 6.65, is similar to that of the noun formation—it has a flower-shape structure with four major nodes—which suggests that the class is more noun-like. The most important node is that of a noun, with monovalent suffixes arranged at the top of the node (e.g. *-hood*, *-ful*). The only exception among these suffixes is *-ling* that also attaches to the adverb node and forms the first ray around it. The second ray for the adverb node belongs to the suffix *-ern*. The bound morpheme node has three monovalent suffixes, located on the left side of the graph (e.g. *-ent*), and the verb node two monovalent suffixes, positioned at the bottom of the graph (e.g. *-ory*). Three monovalent suffixes on the left bottom corner (e.g. *-most*) bind with the adjective node. The duovalent suffixes are positioned on the vertical axis of the bound morphemes node (e.g. *-ary*, *-ar*), as well as on the right side of the graph close to the center (e.g. *-ing*, *-ee*). The polyvalent suffixes form the core of the graph: *-ed, -ie, -ly, -en, -o, -ist, -al* and *-ish*. Among the suffixes, *-al* is the most significant, both in light of type frequency and type valency.

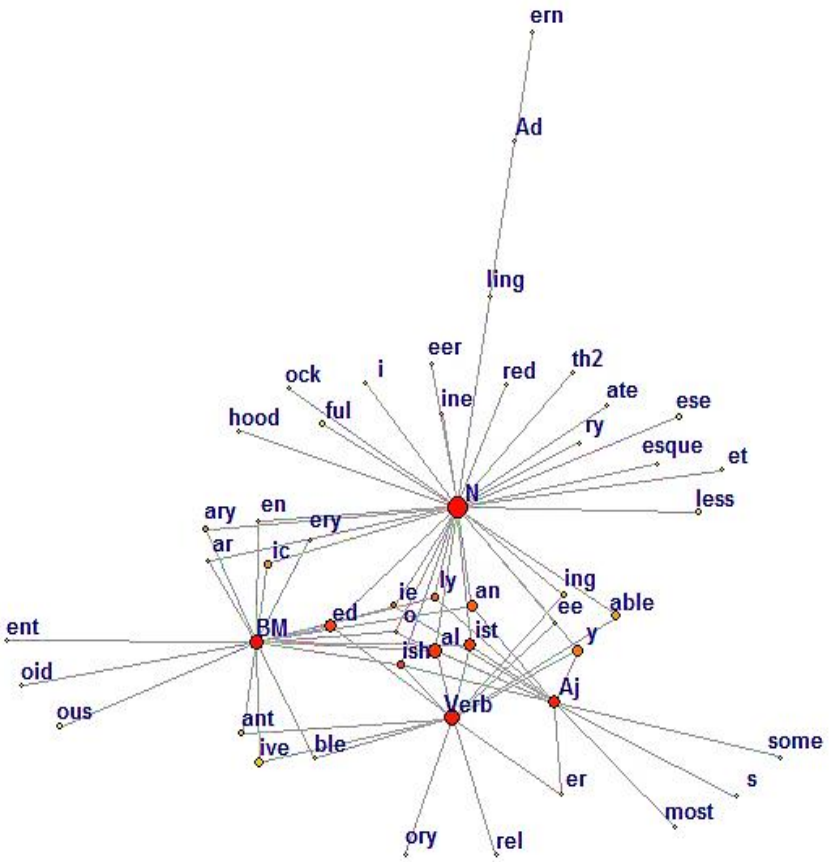


Figure 6.65. The formal morphological paradigm for the construction {C-a} for noun/adjective formation

6.2.14 The formal noun/adjective formation paradigm {{a-C}}

The common feature of the prefixation paradigms for nouns, verbs and adjectives is that their dominant node is the same word class as the paradigm: i.e. the dominant word class for noun prefixation is a noun, for verb prefixation a verb and for adjective prefixation an adjective. It is interesting that in noun/adjective prefixation, visualized in Figure 6.66, there are two large nodes belonging to nouns and adjectives. This observation substantiates the fact that the conversive class noun/adjective combines the features of nouns and adjectives. The adjective node is the most significant in this paradigm binding with the largest number of monovalent prefixes, displayed on the right upper corner of the graph (e.g. *infra-*, *ir-*). The noun node, which attaches to the monovalent suffixes at the bottom of the graph (e.g. *mis-*, *up-*) is of secondary importance. The duovalent prefixes are presented mainly at the center of the graph between the adjective and noun nodes (e.g. *sub-*, *im-*). The prefix *anti-* has the highest type valency in this paradigm, binding with nouns, bound morphemes, adjectives and nouns. The bound morpheme node is shown on the left

upper corner with the connections to the monovalent prefixes *de-* and *contra-*, to the duovalent prefix *intra-* and the polyvalent prefix *anti-*.

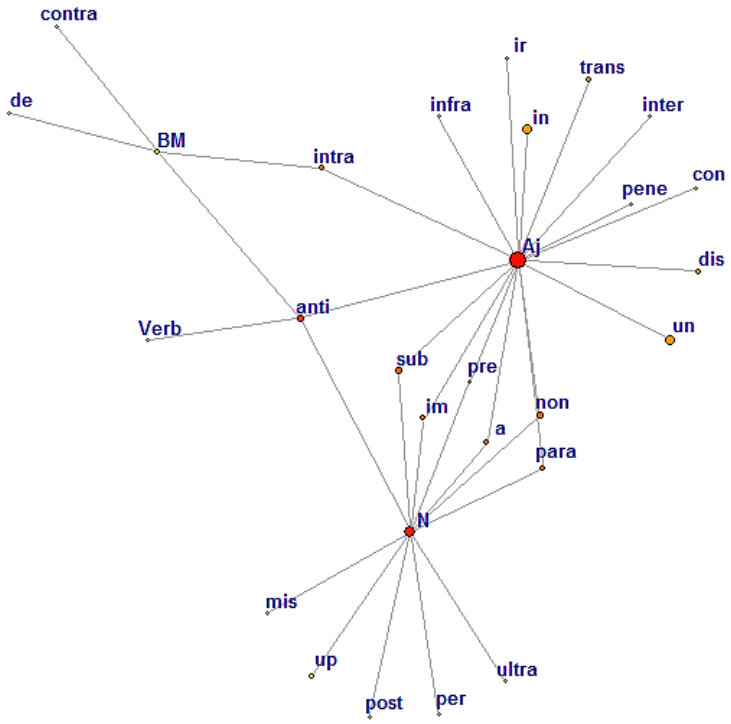


Figure 6.66. The formal morphological paradigm for the meta-construction {{a-C}} for noun/adjective formation

6.2.15 The formal noun/adjective formation paradigm {{C-C}}

Figure 6.67 visualizes the noun/adjective compounding paradigm. The architecture of the paradigm is similar to that of noun compounding. The core node in the graph is that of a noun in the position of the finale. It is located at the center of the graph and connects to the initiale nodes of numerals, prepositions, particles, bound morphemes, adjectives, nouns and verbs. The second significant node is that of an adjective attaching to adverbs, adjectives, prepositions, particles, bound morpheme and nouns. Other classes that occupy the slot for the finale include adverb bases (binding with verbs, adjective and adverbs), verb and pronoun bases (binding with verbs) and bound morpheme bases (binding with nouns and bound morphemes).

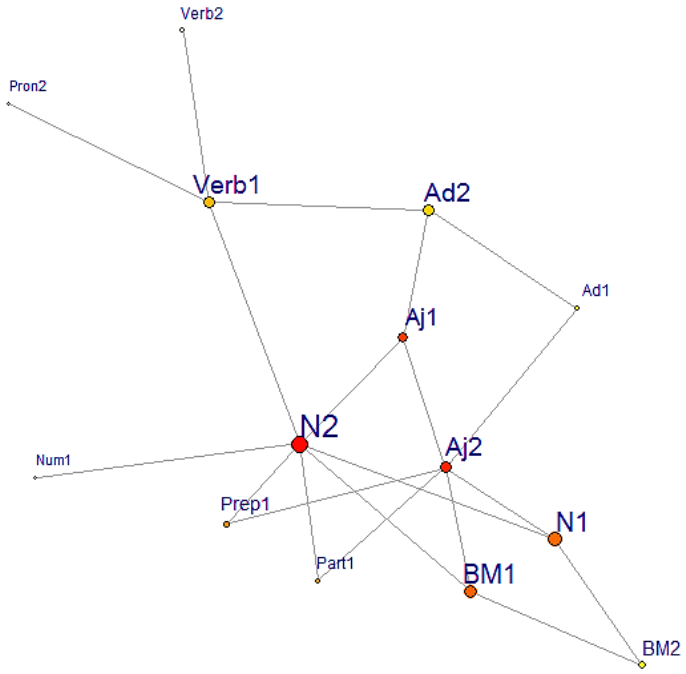


Figure 6.67. The formal morphological paradigm for the meta-construction {{C-C}} for noun/adjective formation

6.2.16 The formal adjective/adverb formation paradigm {{C-a}}

The architecture of the suffixation paradigm in this conversive word class is similar to that of nouns, adjectives and adverbs. It has four major nodes which form a network (Figure 6.68). However, this paradigm displays a feature which is different from all the paradigms described above: its adverb node is detached from the rest of the network.[45] This property of the adjective/adverb paradigm may be the evidence for its less morphological integrity.

The dominant node of this paradigm is formed by nouns which attach to the largest number of monovalent suffixes positioned at the right top of the graph (e.g. *-fold*, *-ed*). The second significant nodes constitute verbs and adjectives, represented on the left upper and bottom corner respectively, together with the rays of their monovalent suffixes. The bound morpheme node is located on the right bottom corner. The highest type valency in this paradigm is 2 which is observed for the following suffixes: *-less*, *-y*, *-ish*, *-ful*, *-ous* and *-ly*.

[45] This piece of the network represents the words *forwards*, *backwards*, *downwards* and *outwards*.

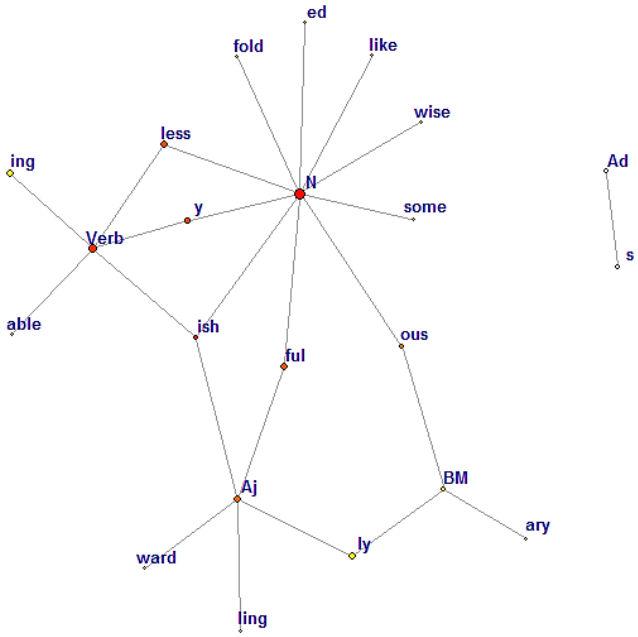


Figure 6.68. The formal morphological paradigm of the meta-construction {{C-a}} for adjective/adverb formation

6.2.17 The formal adjective/adverb formation paradigm {{a-C}}

With three disconnected nodes, the prefixation paradigm in adjective/adverb formation displays even greater morphological disintegrity than the {{C-a}} paradigm (Figure 6.69). Every node in this paradigm attaches to a monovalent prefix. The most significant node belongs to nouns connecting to the prefixes *in-*, *pre-* and *up-*. The second node, with the highest type frequency in the paradigm, binds with the prefixes *un-* and *ir-*. The third node formed by an adverb is monovalent and attaches to the prefix *anti-*.

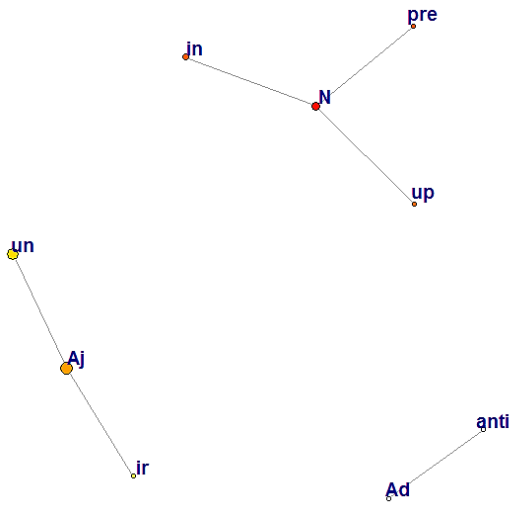


Figure 6.69. The formal morphological paradigm of the meta-construction {{a-C}} for adjective/adverb formation

6.2.18 The formal adjective/adverb formation paradigm {{C-C}}

The last graph in this section describes the adjective/adverb formal paradigm for compounding. It consists of three different pieces (Figure 6.70). The greatest significance in the paradigm is assigned to the bound morpheme node in the position of the finale and the noun morpheme node in the position of the initiale. The former is combined with the *determinant* which is the bound morpheme, and the latter with the *determinantum* which is a preposition. The second important node is formed by a noun in the position of the finale and a preposition and an adjective in the slot of the initiale. Lastly, the third node is formed by an adverb in the position of the finale and a particle in the position of the initiale.

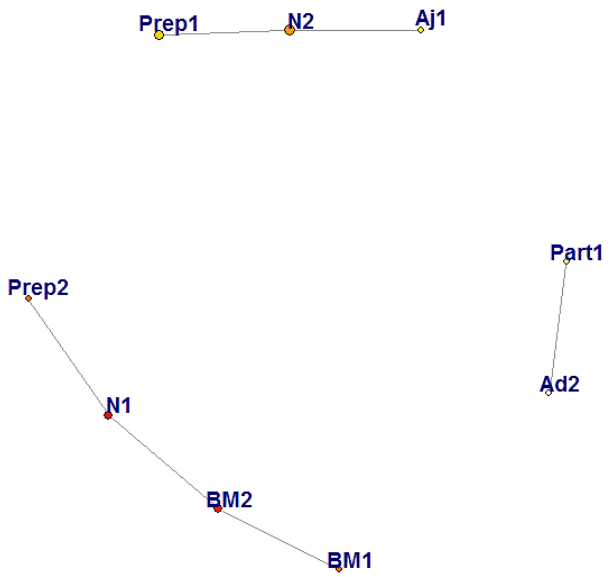


Figure 6.70. The formal morphological paradigm of the meta-construction {{C-C}} for adjective/adverb formation

6.3 Conclusions: the overall features of English word formation

This chapter has presented a detailed description of morphological constructions of English word formation and the regularities of their inner composition. In today's literature on morphology, the term 'construction' has been widely used, but in a very abstract sense, and its description has never been embedded in real language as a whole. The actualization of constructions is a challenging task as it involves dealing with different aspects of a studied phenomenon on a consistent basis.

The attempt made in this thesis to bestow constructions with a real linguistic form has led to establishing their different properties, one of which is formal productivity. The notion 'productivity' is another fuzzy concept in the literature with a plethora of different interpretations.

Thanks to the works of Baayen (1993, 2009), several measures of productivity have been introduced to the field (e.g. potential productivity, global productivity, expanding productivity) which have enabled linguists to quantify this multifaceted phenomenon. Nevertheless, the use of the term productivity in usage-based theories is general and requires further clarification.

The performed analysis of morphological constructions has allowed me to establish a difference between the measures of potential and global productivity (Baayen 1993) and productivity as implied in usage-based theories. As mentioned earlier, one of the hypotheses of the usage-based model of language is that the productivity of a construction is a function of the type frequency of the instances of the construction (Croft 2007: 409): the higher/lower the type frequency of the slots in the construction, the more/less productive it is. Baayen (2009: 901) labels this type of productivity as 'realized'. The difference between realized productivity and other productivity measures is that it concerns the paradigmatic dimension of language, and, as such, can also be termed paradigmatic or formal productivity. On the other hand, the measures of potential, global and expanding productivity are syntagmatic and are realized in real texts/utterances. This distinction is important as it allows linguists to avoid a confusion of terms, and it is one step further in the development of a theory of productivity.

As shown above, different constructions have different degrees of productivity. However, the general trend across almost all word classes is that the construction {C-a} is the most formally productive. The second rank in formal productivity belongs to {a-C-a}, which has the greatest contribution for adjectives.

Further, the analysis performed in this chapter has revealed a high proportion of verb bases in noun formation, which supports argument structure theories (e.g. Grimshaw 1990, Levin & Rappaport-Hovav 1995). These theories acknowledge a central role of verbs and argument structures in the formation of nouns. On the other hand, nouns form the most central and important node in verb suffixations—however, more close attention is required to explain these facts. Another observation is that verb formation largely occurs on the zero and first structural levels. These two pieces of evidence—the dominance of verb bases in nouns and an overall simpler morphological composition of verbs—point to the importance of syntax in word formation as a result of deep intermixing of grammar and lexicon (Bybee 2007b: 980), as well as to the fact that verbs are located on the lower levels of the lexicon's hierarchy.

Furthermore, the significance of certain word bases in the formation of word classes provides evidence for the base-driven hypothesis of derivation (Plag 1996). As has been shown in the preceding sections, a high type frequency of certain word bases explains the most frequent combinations of suffixes (the statistical significance of this correlation is explored in the next chapter). In addition, it is possible to answer the question of whether frequency is a cause or an effect for this particular phenomenon (Bybee 2007: 17-18), if we establish a link between the premises of argument structure theories and those of the base-driven hypothesis (Plag 1996): the cause of the high involvement of certain bases in the formation of words lies in grammar (or our cognition) that projects syntactic connections into word formation, whereas the high type frequency of these bases emerges as an effect of this cause. Then, the high type frequency of certain bases is reverberated in specific combinations of suffixes.

This chapter has also established formal morphological paradigms of English word formation. The variation of the architecture of network graphs for three meta-constructions have demonstrated that some word-formation processes are similar across word classes (e.g. prefixation whose graph, with one dominant word class, shows a flower-like architecture), and some others are different (e.g. the graph network for compounding in adjectives/adverbs that has demonstrated the lack of morphological integrity). Further, the established differences concern typological features of the formal paradigms, namely agglutination, fusion and isolation. Agglutination implies adding elements to the root mechanically, "i.e., without either of the elements being modified" (Greenberg 1960: 183), whereas isolation envisages expressing all relational concepts "by means of the one simple device of juxtaposing words in a definite order, the words themselves remaining unchangeable" (Sapir 1929: 64). Lastly, the fusional property of language is evident from "harmonious fusion of root and affix in a true unity" (Greenberg 1960: 181). In Sapir-Greenberg's classification, these features describe overall morphological structures of different languages. If we look for these typological characteristics in English word formation, they can be further refined with a new parameter: type valency. Overall, if the type polyvalency of a suffix is seen as a feature of agglutination, its monovalency as a feature of isolation and the presence of bound morphemes in word formation as a feature of fusion, it can be inferred from the performed graph analysis that different word-formation processes in different word formation classes show different typological characteristics.

The suffixation in nouns, adjectives, nouns/adjectives and adjectives/adverbs is prominent in that it has several important nodes which attach to a larger number of monovalent suffixes. For noun and adjective suffixation, the most significant node is that of verb, whereas in noun/adjective and adjective/adverb suffixation it is a noun. The role of the bound morpheme is less distinct in noun, adjective and adjective/adverb suffixation and more distinct in noun/adjective suffixation. Moreover, adjective suffixation displays a unique feature which is a more consistent distribution of the type frequency and type valency across the whole paradigm, as well as a fewer number of the monovalent suffixes. With these characteristics, it is possible to generalize that noun, noun/adjective and adjective/adverb suffixation tends to show isolation features, whereas adjective suffixation is more agglutinative in nature. Lastly, the importance of a bound morpheme in noun/adjective suffixation suggests that the fusional property in this conversive class is more strongly represented.

On the other hand, due to a smaller number of morphemes in the paradigms of verb and adverb suffixation, their structure is different. Adverb suffixation is heavily dominated by the combination of adjective and the suffix *-ly*, whereas in verb suffixation the central node is that of a noun. The role of a bound morpheme is more prominent in verbs and is insignificant in adverbs. Further, there is a greater number of polyvalent suffixes in verb suffixation. Thus, the agglutinative and fusional features are more prominent in verb suffixation and the isolation feature in adverb suffixation.

The properties of prefixation are more consistent across the word classes of nouns, verbs and adjectives: by preferring one word base, which is of the same word class as the paradigm, it displays a higher degree of isolation than suffixation. Further, prefixation in the conversive classes reveals the information about the nature of these classes. First, the noun/adjective class has two almost equally dominant nodes, which alludes to the fact that this conversive class has a higher degree of morphological integrity by equally combining the features of nouns and adjectives. In contrast, with three disentangled elements, adjective/adverb prefixation does not display a network structure and entirely lacks the agglutination feature, which points to the fact that without the polyvalency of affixes a network does not emerge (this is the direct evidence for the expression of agglutination through polyvalency of affixes). The possible explanation for this feature can be that the number of words in this class is relatively small to form a fully-fledged network. Another reason might lie in the fact that there is a greater difference between adjectives and adverbs, and

when they combine in a conversive class, their morphological integrity is lost, and in its word formation the adjective/adverb class relies more on semantic or syntactic resources. Lastly, the number of morphological elements in adverb prefixation is small, which is an indication of its low degree of productivity in English word formation.

The architecture of compounding networks is similar in nouns, adjectives and nouns/adjectives. Different word classes occupy the initiale and finale slots in a compounding construction, which can be seen as a property of agglutination. However, as evident from the combinations of word classes, as well as from the significance of nodes representing the finale slot in these paradigms, compounding is influenced by syntactic restrictions. Further, compounding in adverbs and verbs has a very low productivity. Finally, adjective/adverb compounding exhibits the same morphological disintegration as prefixation.

The degree of the expression of agglutination, fusion and isolation across the studied paradigms is illustrated in Table 6.42. The percentages of agglutination and isolation (the rows in the fourth and sixth columns) add up to 100% in this table, because they are two contrasting measures of the same morphemes in a class. However, fusion is a different measure, as it is based on the number of bound morphemes in a class.[46] As inferred from this table, compounding appears as the most agglutinative process for all the studied meta-constructions, except for adverbs and adjectives/adverbs. However, these results should be interpreted with caution, since the slots of the *determinantum* and the *determinant* have not been distinguished in these calculations, and the percentages of the discussed typological features reflect the overall type-valent behavior of word bases. Further, suffixation has the second largest agglutinative power of morphemes, with the exception of adverbs, which, in contrast, show the highest degree of isolation for the meta-construction {{C-a}}. Moreover, isolation is the most pronounced feature in prefixation, excluding adverbs, whereas fusion shows up as a distinct feature of compounding across all classes—again, with the exception of adverbs. The degree of fusion is also high in verb and noun/adjective suffixation. With these measures, we can conclude that the *typological features of different constructions display similarities with some variations, apart from adverbs. Their derivational typological profile is entirely different.*

[46] The proportions of word classes may also contribute to the typological derivational features of languages, but it seems that there are some universal grammatical tendencies in the proportions of word bases in derivational processes, which are common in many languages and which reflect some cognitive inclinations of the human mind.

Another significant observation from the graphs is that the most type-frequent affixes (in particular suffixes) also tend to be the most type-valent, which is obvious from the central position of the type-valent suffixes in the graphs, as well as from their larger number of connections, larger size of diameter and darker colours of nodes. From this observation, it can be hypothesized that there is a close relation between the type frequency of a suffix and its type valency. The next chapter looks at this and other possible effects of type frequency on English word formation. It also explores the distribution of different morphological patterns across years in a chronological order.

Table 6.42. The expression of agglutination, fusion and isolation across the meta-constructions in six major word classes

C	**Meta-construction**	**Agglutination**		**Isolation**		**Fusion**	
		No of polyvalent affixes/bases*	%	No of monovalent affixes/bases*	%	Type frequency of BM	%**
N	{{C-a}}	37	50.7	36	49.3	150	3.9
	{{a-C}}	8	22.9	27	77.1	9	3.6
	{{C-C}}	11	78.6	3	21.4	271	16.9
Aj	{{C-a}}	21	72.4	8	27.6	167	4.9
	{{a-C}}	13	36.2	23	63.8	3	0.73
	{{C-C}}	9	81.8	3	18.2	54	15.9
Verb	{{C-a}}	7	58.3	5	41.7	47	14.9
	{{a-C}}	12	48	13	52	14	3.6
	{{C-C}}	7	70	3	30	11	14.9
Ad	{{C-a}}	2	33.3	4	66.7	7	0.8
	{{a-C}}	3	75	1	25	0	0
	{{C-C}}	5	55.6	4	44.4	2	6.7
N/Aj	{{C-a}}	22	47.8	24	52.2	110	19.1
	{{a-C}}	8	33.3	16	66.7	3	4.2
	{{C-C}}	11	78.6	3	21.4	43	16.9
Aj/Ad	{{C-a}}	6	35.3	11	64.7	4	6
	{{a-C}}	0	0	6	100	0	0
	{{C-C}}	3	33.3	6	66.7	4	20

* The polyvalency of bases is calculated only for the compounding construction. The slots in the compounding constructions have not been distinguished.
** The percentage of bound morpheme bases in the constructions

7 Statistical analysis of different aspects of word formation

The previous chapters have identified different trends, patterns and regularities in English word formation with the help of structural analyses and descriptive tools, namely formal morphological analysis, matrix optimization, charts, bar graphs and graph networks. This chapter focuses on the statistical analysis of different aspects of the morphological metacorpus in an attempt to provide answers to the research questions of what the effects of type frequency in English word formation are (RQ3), how English word formation is represented in time (RQ4) and what clusters of affixes with similar characteristics can be identified in the data (RQ5).

Answers to these research questions complement the English word-formation picture, which has been constructed in the previous chapters. Namely, this chapter brings to light forces that shape English word-formation grammar and looks at the interaction between them, in a quest to understand why we observe this 'snapshot' of word formation. Obviously, it is not possible to account for all aspects of the observed picture only with the morphological lens. However, some of its aspects, as suggested by the relevant literature, are explicable with morphological parameters (see Section 3.6 for their description). For example, this chapter establishes that a high type frequency of affixes has an impact on their type valency and that the frequency of word bases are highly correlated on different levels of derivation—which implies that the observed combination of suffixes are base-driven. Moreover, it looks at how word-formation processes evolve in time and what general word-formation tendencies have led to the word-formation picture we witness today. In addition to identifying some relations between the forces that shape the word-formation grammar of English, the chapter outlines new directions for future studies.

In what follows, Section 7.1 explores type frequency effects in English word formation with correlations and a Poisson regression model (these statistical techniques are explained in Sections 4.1 and 4.2), and Section 7.2 analyzes morphological patterns from a diachronic perspective, presenting a picture of how English word formation in nouns, adjectives and verbs has evolved across years (the conversive classes have been excluded from the analysis, since their word-formation processes combine affixes from single word classes). In Section 7.3, the clusters of affixes are discussed that display similar features and the interaction between different parameters of word-formation processes is considered. The performed cluster analyses are particularly important, as they reveal the interaction between the parameters of word-formation

grammar. Finally, Section 7.4 gives an overview of the established key statistical trends in English word formation.

7.1 Type-frequency effects

The structural analysis in the previous chapters has hinted at a few effects of the type frequency of morphological patterns. It has been shown that, first, high type-frequent patterns display a larger number of orthographic and phonological changes on the boundaries of morphemes; and, second, they have a larger number of connections to other morphemes. Moreover, it has been suggested that the observed combinations of suffixes are driven by word bases and are determined by the frequency of word bases in word-formation processes. Hence, Section 7.1.1 looks into the relation between word bases and the combinations of suffixes, and Section 7.1.2 established the strength of a type-frequency effect on suffixes.

7.1.1 The effect of type frequency on suffix combinations

A considerable body of literature on derivation is devoted to the problem of suffix ordering, whose exploration has started with Level-Ordering Morphology in generative grammar (see Section 2.7.1). As discussed in Section 2.11, one aspect of this problem is whether the combinations of suffixes are driven by word bases or suffixes. Fabb (1988) suggests that selectional restrictions of affix combinations are determined by affixes which are involved in word-formation processes. In contrast, Plag (1996) maintains that English suffixation is a result of base-driven selectional restrictions, paradigmatic morphological processes, and independent principles and constraints of English derivation. In what follows, this base-driven hypothesis of selectional restrictions of suffixes is put to the test with the data from the morphological metacorpus.

The logic behind the test that establishes whether the observed suffix combinations are base-driven is simple. On the first level of derivation, suffixes attach to word bases, whereas combinations of suffixes emerge on the higher levels of word formation which allow for three or more morphemes in a word. If the combinations of suffixes are driven by word bases, then we would expect to observe a correlation between the word bases of suffixes on the first level of derivation and those on the higher levels, due to the similarity of the attachment patterns of word bases on different levels of derivation. By way of illustration, for *-ness* which has been registered as a polyvalent suffix in the metacorpus, adjectives are dominant bases on the first level of noun formation (e.g. *activeness*). This dominance is preserved on the higher derivational levels (e.g. *addictedness, unkindliness*) and determines specific combinations of affixes. Therefore, the

assumption is that a statistically significant correlation between word bases of different suffixes on the first level of derivation and those on higher levels provides evidence in support of the base-driven hypothesis argued by Plag (1996).

In order to test this hypothesis, first, matrices for the morphological constructions {C-a-a}, {C-a-a-a} and {C-C-a-a} (Figure 6.8 and 6.26) have been used to inform our view of suffix combinations. Then, the type frequency of the final suffixes in their combinations with different word bases has been recorded on the first and higher levels of derivation (Tables 7.1–7.2). For example, in the suffix combination *-ize + -er*, the preceding suffix *-ize* forms verbs. Hence, we know that a word base with this suffix (which is productive on the higher levels of noun derivation) belongs to the word class of verbs. Further, we register the type frequency of verb bases that attach to the suffix *-er* on the first level of derivation and the number of all types of this combination on the second and third levels. By this token, two variables have been created: one for the type frequency of a final suffix in combination with a word base of interest (marked as '**C+a**, **1st L**' in Tables 7.1 and 7.2) and another for the type frequency of the suffix combinations whose preceding suffix forms the word base of interest ('**C(')+a**, **2nd & 3rd L**'). Next, the Spearman correlation test (see Section 4.1) has been performed on these variables to establish the strength of the association between the studied morphological patterns on the first and higher levels of derivation.

Table 7.1. The type frequencies of word bases and suffixes on different levels of noun formation

Word base	Final Suffix	C+a 1st L	C(')+a 2nd & 3rd L	Observed combinations of suffixes on the second and third levels
Verb-	**Verb+ion**	64	36	ate+ion, ize+(at)ion
	Verb+or	38	3	ate+or
	Verb+er	724	38	en+er, ize+er, le+er
	Verb+ing	795	27	en+ing, ize+ing, ate+ing, eer+ing, age+ing
Aj-	**Aj+ism**	45	17	an+ism, al+ism, ent+ism, ic+ism
	Aj+ity	67	52	ic+ity, al+ity, ive+ity, able+ity
	Aj+ist	11	1	ic+ist
	Aj+hood	2	3	ly+hood, y+hood
	Aj+ness	204	106	ive+ness, able+ness, ed+ness, ful+ness, ous+ness, ing+ness, some+ness, less+ness, al+ness, y+ness, ly+ness
	Aj+s	10	6	ic+s
N-	**N+ist**	63	8	ion+ist, al+ist
	N+ism	51	5	er+ism, ee+ism
	N+ing	55	4	eer+ing, er+ing
	N+ship	42	8	er+ship
	N+ess	21	4	or+ess, er+ess
	N+hood	12	1	er+hood
	N+ry	11	1	ist+ry
	N+y	40	6	er+y
	N+dom	14	2	er+dom

Table 7.2. The type frequencies of word bases and suffixes on different levels of adjectival formation

Word base	Final Suffix	C+a 1st L	C(')+a 2nd & 3rd L	Observed combinations of suffixes on the second and third levels
Verb-	**Verb+ed**	715	57	ize+ed, ify+ed, en+ed, ish+ed
	Verb+ing	631	25	ize+ing, en+ing, le+ing
	Verb+able	299	4	ize+able, ify+able
	Verb+ive	42	3	ate+ive
Aj-	**Aj+ed**	11	1	ish+ed
	Aj+al	39	9	ic+al
	Aj+y	8	1	ed+y
	Aj+s	10	6	ic+s
N-	**N+ed**	183	8	ock+ed, er+ed, let+ed, ist+ed
	N+ing	11	1	eer+ing
	N+able	13	1	er+able
	N+al	146	10	ion+al, ment+al, er+al
	N+less	158	4	er+less, age+less, ing+less
	N+ish	33	2	y+ish
	N+ous	18	1	ion+ous
	N+ist	43	21	ist+ic
	N+y	225	2	er+y, th+y, ist+y
	N+ery	9	1	ion+ery
	N+ful	92	3	ing+ful

The data for this correlation analysis have been elicited only for nouns and adjectives. This is because the derivation of other word classes is limited and does not provide a sufficient sample of suffix combinations for running the correlation test. Further, the sufficiency of the sample size of nouns and adjectives (given in Tables 7.1 and 7.2) has been verified with the sample size calculator,[47] with a α-threshold set for 0.05 and β-threshold for 0.20.

The results are presented in Table 7.3. There is a very strong, statistically significant association between the type frequency of word bases and final suffixes in suffix combinations in the first and higher levels of derivation, which allows us to conclude that the combinations of suffixes are base-driven to the degree of the association between variables. Adjectives display a lower rho coefficient, as compared to nouns, which can possibly be explained by the fact that the role of adjectival suffixes as final suffixes is slightly weaker. The remaining 20% of the association can be accounted for by other factors, which are specific to each word-formation process.

Table 7.3. Statistics of the Spearman correlation for compared morphological patterns on the first and higher levels of derivation in nouns and adjectives

Word class	Rho	p-value
Nouns (n=19)	r = 0.824024	p < 0.05 (p = 7.96e-06)
Adjectives (n=17)	r = 0.7928628	p < 0.05 (p = 0.0001469)

[47]https://sample-size.net/correlation-sample-size/

7.1.2 The effect of the type frequency of suffixes on their type valency

The network graphs in Section 6.2 revealed a consistent trend in the organization of morphemes: the morphemes with larger circles tend to attach to a larger number of morphemes. That is, the type valency of morphemes with a larger type frequency is higher, which suggests that that these two parameters are related. Further, Krykoniuk (2020) has established the effect of the type frequency of suffixes on their type valency in Persian. Hence, this section aims to identify whether this effect is also present in English word formation.

For this purpose, I have identified the type frequency of the morphological construction {{C-a}} in nouns, adjectives and verbs, as well as the type valency of the suffixes which fill the *a* slot in this construction. The conversive classes have been excluded from the analysis, since they are hybrids with the combined derivation of suffixes from single word classes. In total, 114 noun, adjective and verb suffixes have been registered in the metacorpus. Their values of the type frequency and type valency constitute the data for the following correlation and regression analysis. This data is given in Appendix F.

The red line in Figure 7.1 visualizes an upward tendency in the relation between the type frequency and type valency of suffixes, which implies a positive correlation between them: an increase in one variable leads to an increase in another. This trend is fairly consistent, with the largest portion of suffixes showing a low type frequency and type valency (the suffixes clustered in the left bottom corner of the scatterplot). There are no obvious outliers in the plot, but there are a few suffixes that show a slightly atypical behavior. For example, the noun suffixes *-dom*, *-ship* and the adjectival suffix *-ing* have a lower type valency, and the noun suffix *-age* and the verb suffix *-en* have a lower type frequency, as compared to what the trend predicts. Another interesting observation that can be inferred from the scatterplot is that the suffix *-ing* displays a different picture for nouns and adjectives: with almost the same type frequency value, the type valency of the adjectival suffixes *-ing* is lower (3) than that of the nominal suffixes (5). This is an indication that syntactic constraints are tighter for the adjectival suffix *-ing*[48], preventing it from developing a higher type valency. Further, among nouns, adjectives and verbs, the verb suffix *-en*, the noun

[48] It can be argued that words with the adjectival suffix *-ing* are formed not by derivation, but conversion. However, the current study adopts the view that the suffix *-ing* is both adjectival and nominal. This is because, first, the time of their derivation largely overlaps and shows a similar pattern, and, secondly, both processes are identified as derivational by the OED.

suffixes *-age*, *-ness* and *-ing*, and the adjectival suffix *-y*, display the highest values of type frequency and type valency.

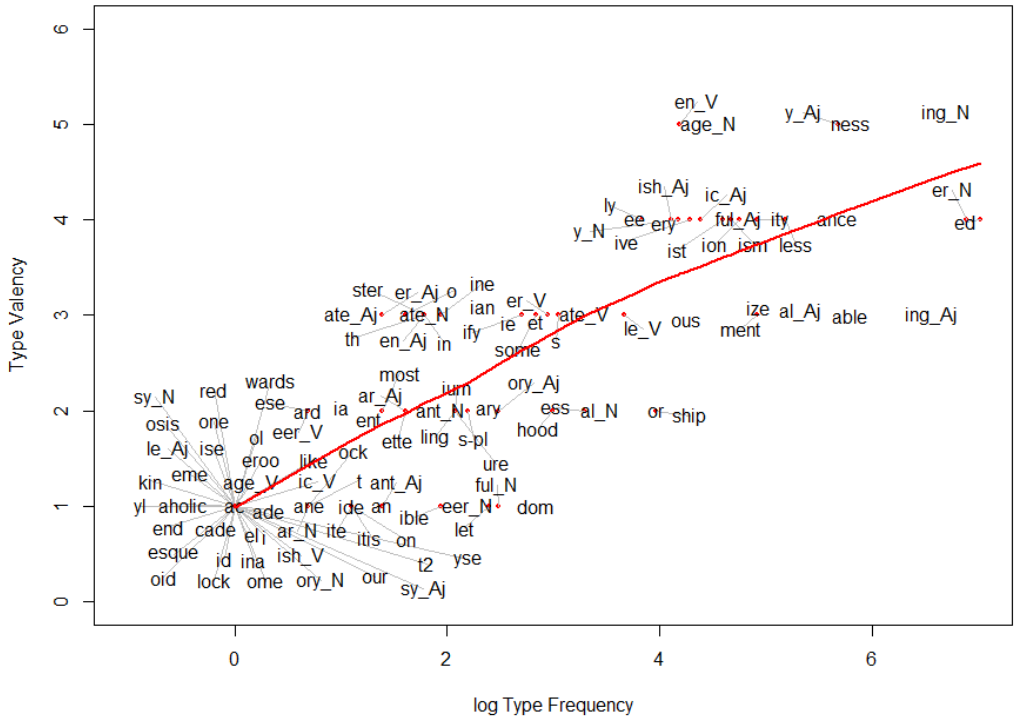


Figure 7.1. Scatterplot of the type valency and the log type frequency of nouns, adjective and verb suffixes.

The Spearman correlation test confirms the trend observed in Figure 7.1. There is a very strong correlation between the type frequency of suffixes and their type valency ($r = 0.86$; p-value < 0.001 which breaks down in a permutation test. Hence, the type-frequent suffixes also tend to be polyvalent.

To establish whether it is possible to predict the type valency of a suffix with its type frequency, Poisson regression models were fitted to the data. Table 7.4 presents statistics of the Poisson regression model jointly for nouns, adjectives and verbs, whereas Table 7.5 gives coefficients of the models fitted separately for nouns, adjectives and verbs. The α-level for the models has been set prior to running the regression analysis to $p < 0.05$.

Table 7.4. The coefficients of the Poisson regression model (GLM) for nouns, adjectives and verbs, fitted jointly

Predictor	**Estimate**	**SE**	**p-value (for the model)**	**R^2 Nagelkerke**	**AIC**
log type frequency (n=114)	0.21923	0.02929	p < 0.001	0.756	320.61

Table 7.5. The coefficients of the Poisson regression model (GLM) for nouns, adjectives and verbs, fitted separately

Word Class	Predictor	Estimate	SE	p-value (for the model)	R^2 Nagelkerke	AIC
Nouns (n = 73)	log type frequency	0.23729	0.03848	$p < 0.001$	0.776	199.93
Adjectives (n = 29)	log type frequency	0.17332	0.05365	$p = 0.001$	0.662	90.639
Verbs (n=12)	log type frequency	0.2681	0.1083	$p = 0.013$	0.856	36.597

The fitted Poisson model in Table 7.5 is statistically significant and explains up to 76% of the variation of the data, which is quite a considerable result. The model's standard error (SE) is low, and, as informed by its estimate, the expected number of the type valency changes by a factor of 1.25 for each additional morphological type of a suffix. Moreover, it is interesting that the statistics of the models for English derivation is very similar to the models fitted to Persian data (Krykoniuk 2020), which suggests that the type frequency effect is a derivation feature in languages showing derivational agglutination.

A similar picture emerges for the studied word classes fitted separately (Table 7.6). It is difficult to compare the three models presented in the table, since the number of their observations is not the same. There is a word-class derivational constraint for each word class, which is the strictest for verbs (operating with 12 derivational suffixes). Nevertheless, it is still worth noting that the type-frequency effect on the type valency is the most pronounced in the model for verbs, with the expected number of the type valency changing by a factor of 1.3 for each additional morphological type (containing a verbal suffix) and with a Nagelkerke R-squared of 0.86. The Poisson regression model for verbs also shows the least Akaike Information Criterion (AIC). The statistics for the verb regression model hint at the best model fit and suggest that the type-frequency effect of suffixes on their type valency is the most pronounced in verbs. Further, the regression model for nouns accounts for nearly 78% of the variability in the data, predicting the change in the type valency by a factor of 1.24. Finally, the regression model for adjectives explains up to 67% of the variability in the data and has the lowest regression coefficient, predicting that the type valency of adjectival suffixes increases by a factor of 1.2 for each additional morphological type.

For a Generalized Linear Model (GLM), it is not expected that the residuals should be normally distributed (Faraway 2016: 127). However, residual plots are informative since they provide a glimpse into the nature of the models. Figures 7.2–7.5 present residual vs fitted and residual QQ plots for each GLM discussed above. They share similar features: although tending

towards normality, the residuals of the models are not normally distributed and show the signs of a bimodal distribution, which is different for the Persian GLMs (Krykoniuk 2020). This observation suggests that there is a greater degree of non-linearity in the sample of English suffixes, where the number of monovalent suffixes is larger, and there is a large discrepancy of type-frequency values between highly type-frequent suffixes and the rest. Finally, in the residual vs fitted plots, the patterns of aslant, parallel lines are more distinct in the English GLMs, which emerge because the values on the *y*-axis are repeated in the data (Searle 2021[1988]).

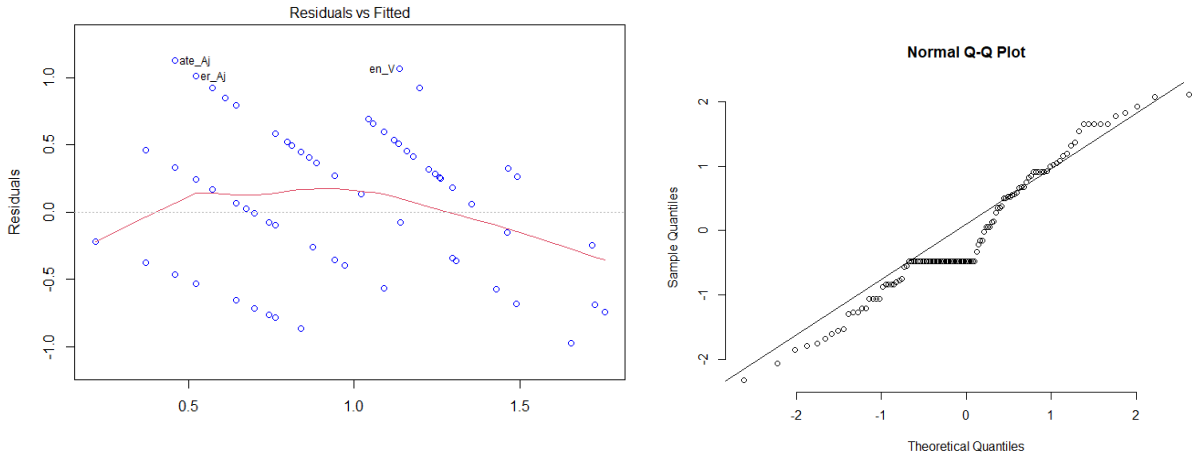


Figure 7.2. Residual vs fitted and residuals QQ plots for the GLM,
fitted to the suffixes of nouns, adjectives and verbs jointly

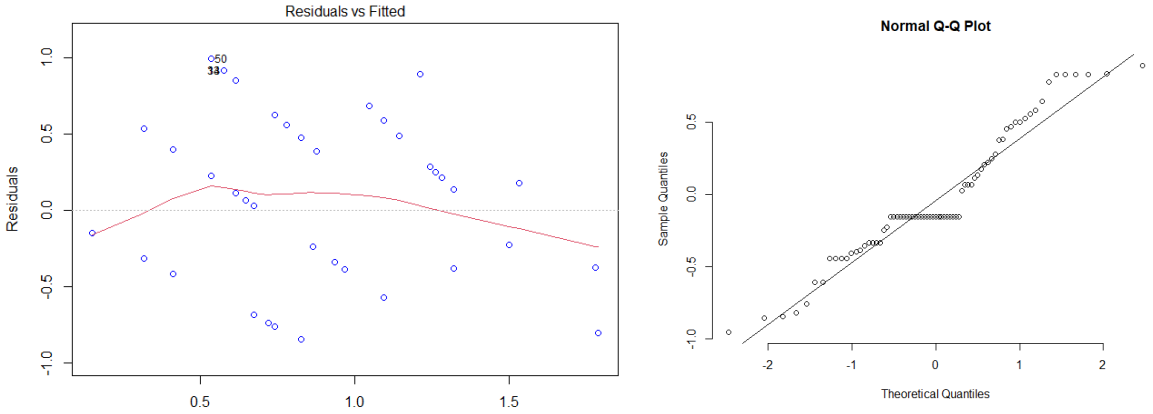


Figure 7.3. Residual vs fitted (left) and residuals QQ (right) plots for the GLM, fitted to nominal suffixes

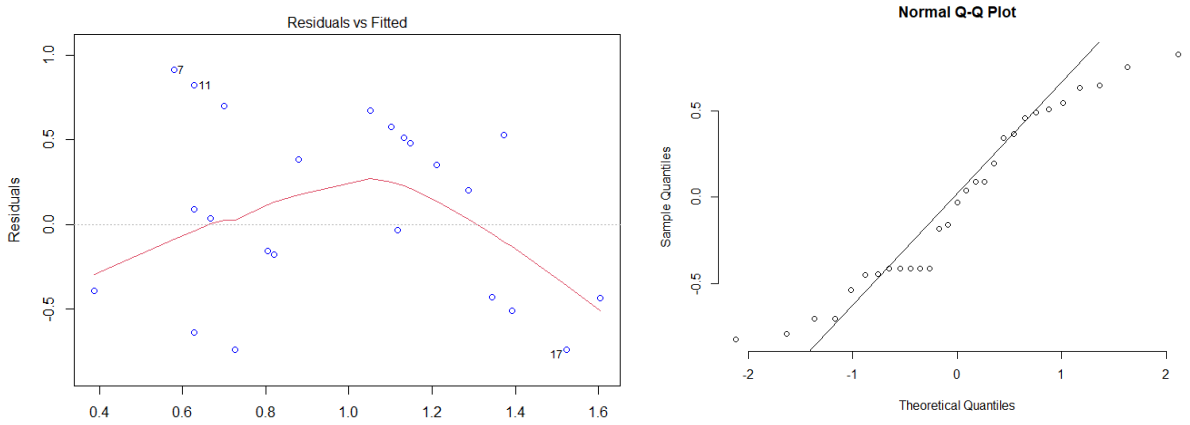


Figure 7.4. Residual vs fitted (left) and residuals QQ (right) plots for the GLM, fitted to adjectival suffixes

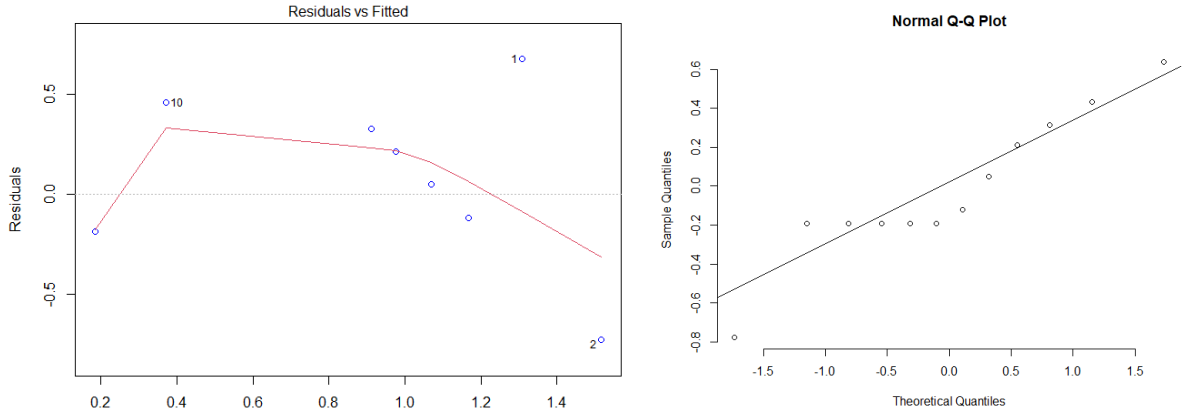


Figure 7.5. Residual vs fitted (left) and residuals QQ (right) plots for the GLM, fitted to verb suffixes

In order to validate the fitted GLMs, the bootstrap simulation[49] with prediction intervals[50] (n = 20,000) has been performed with the help of the *R* packages 'ciTools', 'trending', 'patchwork' and 'MASS' (without adding uncertainty). Considerations of space preclude a detailed description of all fitted regression models. Figure 7.6 illustrates the bootstrap simulation for only the GLM fitted jointly to all studied suffixes (described in Table 7.4). The two plots show both the prediction intervals for the Poisson regression model fitted jointly to all suffixes and those for the averaged model created through 20,000 simulations, with resampling the data of the real GLM model of interest. The bootstrap prediction intervals repeat the trend of the actual model and cover its true

[49] Bootstrapping is a statistical method for the validation of fitted regression models through random sampling with replacement.
[50] "A prediction interval for a single future observation is an interval that will, with a specified degree of confidence, contain a future randomly selected observation from a distribution" (Meeker et al. 2017: 27).

values, which is an indication that the model is adequate and that it can be generalized to population.

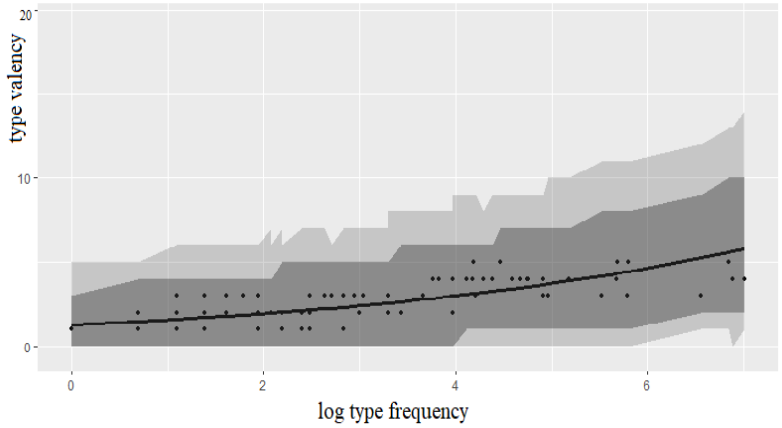


Figure 7.6. Prediction intervals for the GLM, fitted to the suffixes of nouns, adjectives and verbs (model fit – black line, bootstrap intervals – grey, parametric intervals – dark grey)

According to the fitted models, the type frequency of the English suffixes accounts for 66–86% of the variability in the data. The remaining variability is a consequence of other factors, such as the semantics of suffixes and the extent of their use, the prosodic or phonological features of suffixes and their historical and grammatical development (Krykoniuk 2020). These are the factors that constitute individual differences between suffixes.

The discussed regression models allow us to contemplate the nature of a morphological rule and the Unitary-Base Hypothesis (Aronoff 1976) in English, which assumes that suffixes tend to attach to one type of a word base. As shown above, the type-frequent suffixes develop a property to attach to more than one word base. Hence, it might be the case that this word-base constraint on a morphological rule is still true and occurs in the early stages of the development of a suffix, and as a suffix spawns a high number of types, the constraint loosens. This assumption is evident from the fact that, in English, there is a large number of monovalent suffixes—as shown above, their number is greater than that in the Persian language (Krykoniuk 2020). In fact, a larger number of monovalent suffixes in English make the Persian and English regression models different. From a typological point of view, it can be conjectured from this difference that Persian is more agglutinative in nature, whereas the isolation derivational feature of English is more distinct.

A piece of evidence in favour of the Unitary-Base Hypothesis for initial conditions is that some word bases are more dominant in specific word-formation processes and repeat themselves on the higher levels of derivations, resulting in specific combinations of suffixes (as substantiated

in Section 7.1.1). Hence, the Unitary-Base Hypothesis (Aronoff 1976) and the base-driven hypothesis (Plag 1996) are intimately linked and support each other: if the combinations of suffixes are driven by the word class of a base, then some word bases are more dominant than others, and this dominance can be understood as a word-base constraint on a morphological rule—which is a given initial condition of a 'suffix'-'word-base' relation. However, as a suffix develops a high type frequency, this condition is perturbed, and, as a result, some new combinations of suffix emerge.

7.2 The diachronic perspective on word formation

This section focuses on the diachronic perspective of the word-formation processes which have been identified as the most type-frequent in the sample. Their account is presented in the following order of word classes: nouns, adjectives and verbs. Each subsection corresponds to one word class.

7.2.1 The diachronic picture of the most type-frequent noun morphological patterns

In this subsection, 19 word-formation processes with a type frequency above 20 are explored: {C-ing}, {C-C}, {C-er}, {C-ness}, {C-ment}, {C-age}, {C-al}, {C-ee}, {C-|ence|}[51], {C-|ery|}[52], {C-ion}, {C-ism}, {C-ist}, {C-ity}, {C-or}, {C-ship}, {C-y}, {re-C} and {dis-C}. The development of these word-formation processes across years is visualized as overlapping histograms in Figures 7.7–7.15 that show the frequency of the first citation of words at three levels: the zero-level (marked as 0 in the histograms) that includes borrowed words, as well as words formed in Old English or inherited from Germanic, the first level (1) containing duomorphemic words, and the second and third levels (2&3) with three or more morphemes in a word. On the histograms, the zero level is shown in light sea green, the first level in dark green, and the second and the third levels in olive green. In order to better identify the trends in the histograms, the three levels have been plotted against each other in the same space as overlapping histograms, and a degree of transparency has been added to each colour such that when the bars from different levels coincide, they are still visible. Further, in the description of these histograms, I have relied on the concept of the 'realized productivity' which is a usage-based measure reflecting the size of a morphological category (Baayen 2009: 901) and which is estimated by the number of its types. Therefore, each bar in the histograms below is representative of the realized productivity in a specific point in time.

[51] Two vertical lines around the name of a suffix signify that all allomorphs of this suffix have been considered in creating its overall word-formation picture (e.g. the form |ence| encompasses the allomorphs *-ance*, *-ancy*, *-ence*, *-ency*, *-acy* and *-cy*).
[52] This suffix form includes allomorphs *-ry*, *-ery* and *-ary*.

The development of the realized productivity of word-formation processes across years is understood as their diachronic productivity.

The word-formation processes {C-ing} and {C-C}

With the largest number of types, these processes have the highest realized productivity. They were quite productive in Old English; in particular, many of these words were recorded around the year 900, which is evident from its high bars on the histograms in Figure 7.7. The process {C-ing} was at its peak between 1300 to 1700, and slowly its productivity had decreased by the year 2000 with another small productivity peak in the 1900s. However, a different trend is observed for compounding {C-C}, whose productivity has been constantly rising across years. This tendency suggests that compounding has become more productive in English relatively recently, with the peaks in 1800-2000. Another interesting observation from the histograms is that polymorphemic nouns in these two word-formation processes mirror the overall tendency at the first word-formation level, and they are later formations than nouns of the first level.

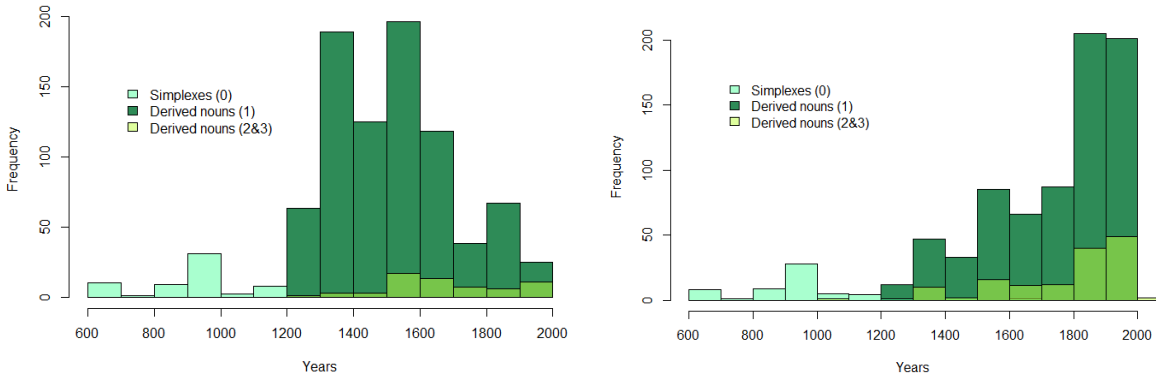


Figure 7.7. The distribution of the word-formation processes of {C-ing} (on the left) and {C-C} (on the right) across years

The word-formation processes {C-er}, {C-ness} and {C-ship}

{C-er},{C-ness} and {C-ship} are other derivation processes whose developmental trends show some similarities (Figure 7.8). The suffix *-er* was productive in Old English, with a relatively higher number of words formed in between 900 to 1000. At around the same period, a higher amount of words containing the suffix *-ness* was produced. In contrast, the suffix *-ship* was productive around 700 and, then, by 1000, its realized productivity dropped. The highest peaks of the first records of words with the suffix *-er* occurred around 1400, 1600 and 1900, which is almost the same for the suffixes *-ness* and *-ship*, with the only difference that, for *-ness*, the 1600's first

citation peak lasted longer and that its third peak was specifically productive on the higher levels of formation, and, for *-ship*, the first derivational peak started around 1200. In general, the suffix *-ness* displays a higher word-formation power on the second and third levels of formation, as compared to other word-formation processes, which is evident from higher bars of this category. This observation supports the idea of the suffix *-ness* as a closing suffix. On the other hand, the peculiarity of the suffix *-er* is that, along with the rich history of native formations, there are many borrowed words containing this suffix throughout the whole history of its development, the most recent of them in the sample being the word *waiver* (1628); and the suffix *-ship* is specific in that it spawns fewer types. Finally, as with other word-formation processes, the tendencies on the first and higher levels of formation are mirrored in these three suffixes.

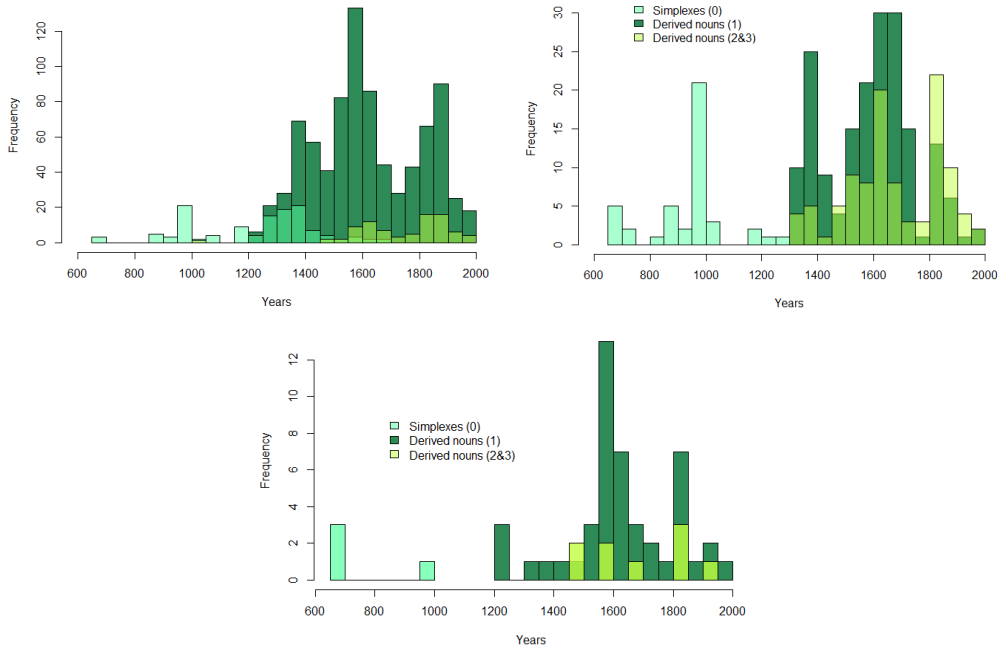


Figure 7.8. The distribution of the word-formation processes of {C-er} (on the upper left corner), {C-ness} (on the upper right corner) and {C-ship} (at the bottom panel) across years

The word-formation processes {C-|ence|} and {|C-ery|}

As evident from Figure 7.9, the suffixes *-ence* and *-ery* show a similar pattern, although there are a larger number of words containing the suffix *-ence*. These suffixes entered English with borrowed words in the 1200s, peaking in the 1400s. The derivational trend with these suffixes started before 1400 and culminated around the 1600s. Interestingly, the highest derivational peak for these suffixes coincides with the second peak of loan words, which defines the main feature of these processes, whose borrowing and derivation highly overlap, with loan words entering the language up to the 1900s. The derivation curve for the suffix *-ence* is smoother and approximates

a normal distribution more, whereas the curve for the suffix *-ery* displays a bimodal nature with another derivational peak in around the 1800s. The suffix *-ery* also has slightly greater productivity on the higher levels of derivation.

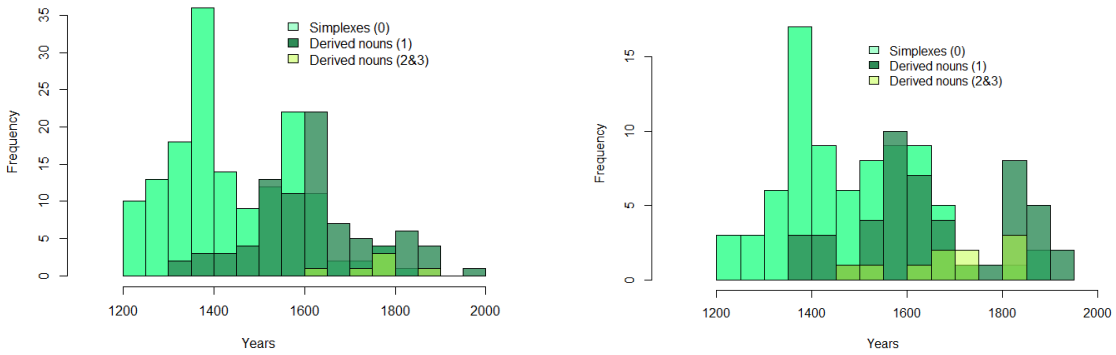


Figure 7.9. The distribution of the word-formation processes of {C-ence} (on the left) and {C-ery} (on the right) across years

The word-formation processes {C-ion} and {C-ity}

The first citation pictures for these processes resemble each other, with the difference being that the number of loan words containing the suffix *-ion* is almost five times greater than those containing the suffix *-ity* (Figure 7.10). The first record of words with these suffixes is in the early 1000s, and the borrowing trend abruptly increases, peaking in between 1300 and 1400, and in between 1500 and 1600. There are two waves of the first-level derivation in both suffixes, which is more clearly visible on the rightward histogram for the suffix *-ity*, due to its smaller discrepancy between borrowed and derived words. The first wave emerges around 1400 and peaks in around 1600. The second wave culminates in the 1800s. The second-level derivation appears later than that of the first level and almost overlaps with the second wave of the first-level derivation, peaking in between 1800 and 1900 and with the trend decreasing towards 2000.

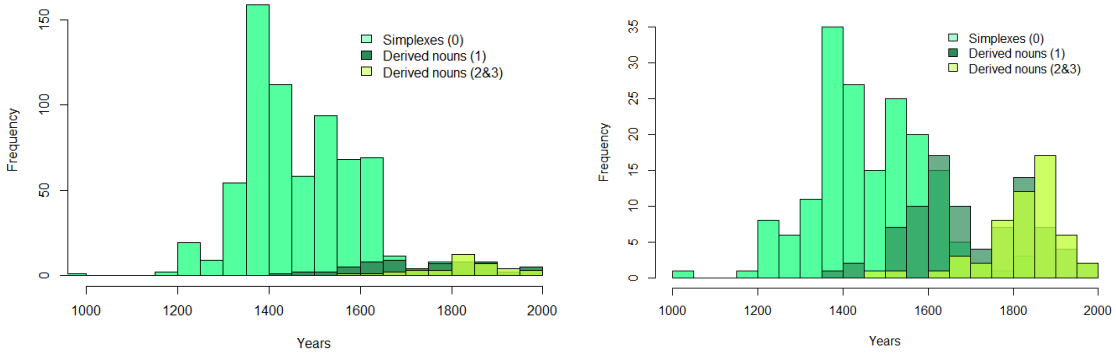


Figure 7.10. The distribution of the word-formation processes of {C-ion} (on the left) and {C-ity} (on the right) across years

The word-formation processes {C-ist} and {C-ism}

The prominent feature of these suffixes is that their derivation levels greatly overlap, specifically for the suffix *-ist* (the left panel of Figure 7.11), which means that they were realized as word-formation morphemes from the beginning of their appearance in the language. The first peak of their first-level derivation is between 1400 and 1500 for the suffix *-ist*, and the first half of the 1600s for the suffix *-ism*. The largest number of the first-level words with these suffixes were recorded in between 1800–1900, which allows us to hypothesize that this trend reflects the major historical changes in the world of that time triggered by the Industrial Revolution (e.g. leading to the formation of such words as *activism*, *personalism*, *corruptionist* and *scientist*). This is a vivid example of how a society's development is echoed in language. By 2000, the derivational tendency of these suffixes decreased. The polymorphemic words with these suffixes are later formations starting around 1700 and peaking in between 1800 and 1900. The trends in the polymorphemic derivations mirror the first-level derivation.

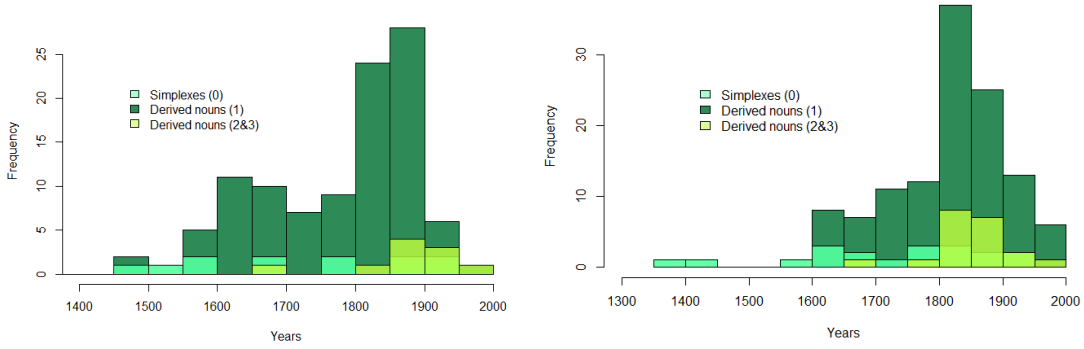


Figure 7.11. The distribution of the word-formation processes of {C-ist} (on the left) and {C-ism} (on the right) across years

The word-formation processes {C-ment} and {C-al}

As informed by the histograms in Figure 7.12, the derivational tendencies for the suffixes *-ment* and *-al* are similar. Borrowings with these suffixes started appearing in the language in the early 1100s, gradually increased by 1400 and regularly continued entering the language until the 1800s. The first-level derivation began around 1300 and culminated in the 1600s and then abruptly reduced to peak again in the 1800s. The second-level derivation emerged later in the development of these suffixes and was more pronounced for the suffix *-ment*. Further, the realized productivity of the suffix *-al* declined sharply by the 1900, whereas the suffix *-ment* continued to be productive in the first part of the 20th century.

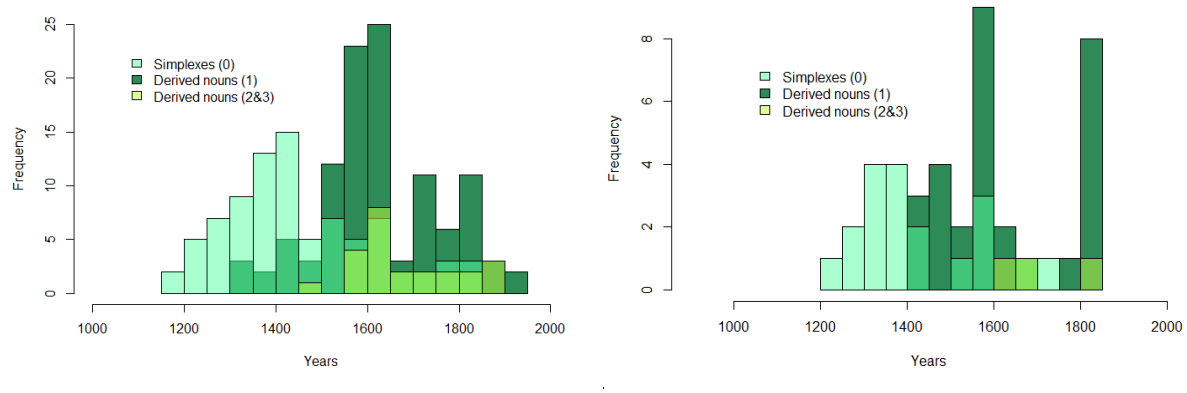


Figure 7.12. The distribution of the word-formation processes of {C-ment} (on the left) and {C-al} (on the right) across years

The word-formation processes {C-or} and {C-y}

Words with the later identifiable morphemes *-or* and *-y* existed in language since 600. The borrowing tendency began in 1200, reached its peak in 1400 for the suffix *-y* and in 1600 for the suffix *-or*, and then slowly decreased by 2000. The first-level derivation occurred after 1400, and peaked around the 1600s and 1800s. There are only a few polymorphic formations of these suffixes, which introduces them as mostly the first-level derivational units. Furthermore, the realized productivity of the suffix *-y* has continued up to the year 2000.

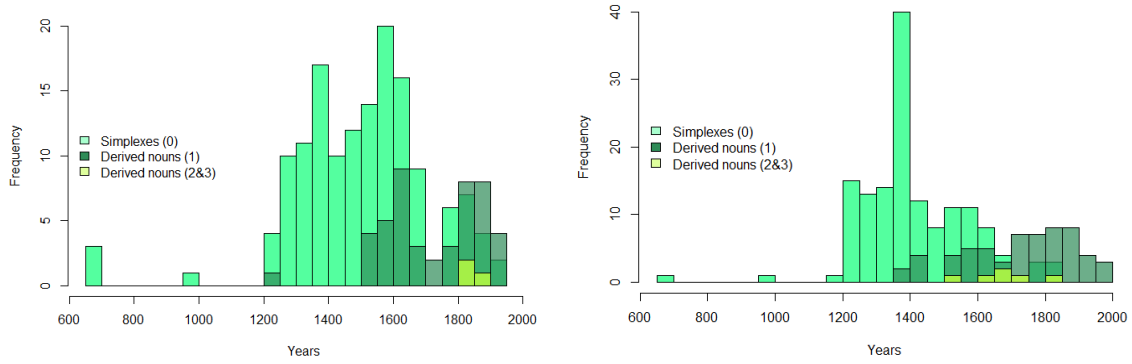


Figure 7.13. The distribution of the word-formation processes of {C-or} (on the left) and {C-y} (on the right) across years

The word-formation processes {dis-C} and {re-C}

These derivational processes are similar in that they were productive on the second level of word formation in between 1800-1900 (Figure 7.14). Their other trends are different. A larger number of loan words contained the prefix *dis-*, and these borrowings started in 1200 and continued up to 1700. The first peak of the first-level derivation with *dis-* happened in between 1300-1400, which

was a hundred years later for the prefix *re-*. Another difference is that the first-level derivation for the prefix *dis-* reached its culmination around 1600, whereas for the prefix *re-* around 1800. These observations allow us to conclude that the word-formation process with the prefix *re-* is more recent and that is has been more diachronically productive, specifically on the second and third levels of derivation.

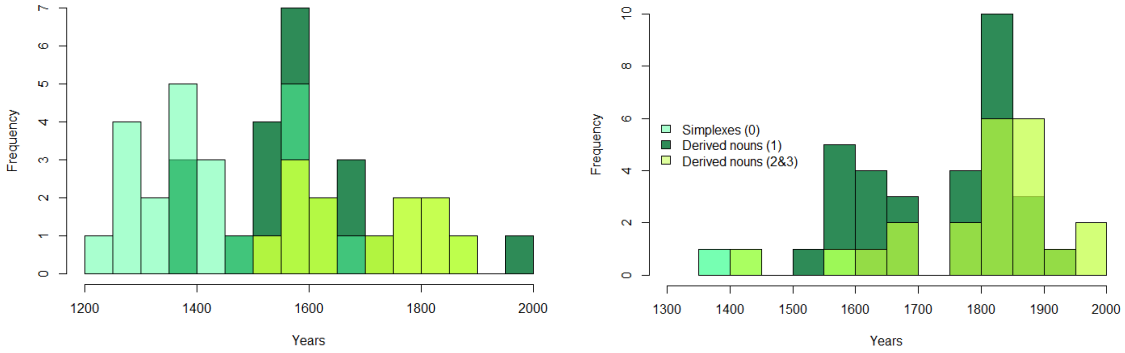


Figure 7.14. The distribution of the word-formation processes of {dis-C} (on the left) and {re-C} (on the right) across years

The word-formation process {C-age} and {C-ee}

The main feature of these processes is that their derivation happens on the first level of formation. The words with the suffix *-age* started appearing in the written language from 1100, with the highest amount borrowed between 1200 and 1300. As shown in Figure 7.15, the borrowing trend has continued up to the 20th century. The first-level derivation started in around 1300 and displayed an increasing trend up to 1900.

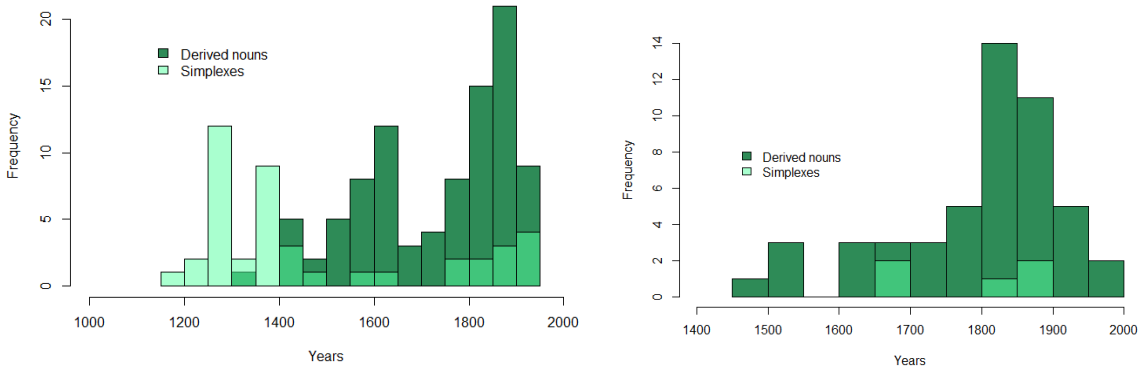


Figure 7.15. The distribution of the word-formation processes of {C-age} and {C-ee} across years

7.2.1 The overall picture of the diachronic development in nouns

The comparison of histograms has established that there are some common trends in the diachronic productivity of noun-formation processes. First of all, the years around 1600 and 1800 were the

most important in English noun formation constituting peaking points of two waves of derivation. Whereas all the noun derivational processes have an overall decreasing tendency towards the year 2000, compounding shows a significant rise in realized productivity over the last two centuries. Among the studied prefixes, *re-* is relatively more recent than the prefix *dis-*. Further, the second and third levels of noun formation are later than the first level (except for the suffix *-ness*) but mirror the tendencies observed on the first level. Moreover, the suffix *-ness* is the most productive of all suffixes on the higher levels of derivation, which is evident from the higher histogram bars in this category. This observation introduces the suffix *-ness* as the most productive closing suffix, followed by the suffixes *-ion, -ity* and *-ment*. In contrast, the suffixes *-age* and *-ee* are productive only on the first level of derivation. Finally, the suffixes *-ence, -ery, -ion, -ity, -or* and *-y* have a larger number of types in the category of loan words, as opposed to the rest of the studied suffixes.

7.2.2 The statistical comparison of the diachronic productivity in noun formation

The above-discussed histograms illustrate that some word-formation processes show similarities of the development of their diachronic productivity. The compared word-formation processes have been chosen solely on the basis of the visual similarities of their histograms, which does not imply that they are similar to the degree of a statistical significance. Hence, this subsection identifies how similar the diachronic productivity of the studied noun-formation processes is. The statistical comparison is performed on different pairs of the noun-formation processes with the help of three entropy estimators (see Section 4.3): KLD, symmetrized KLD and Turing's perspective estimator. The results are summarized in Table 40, which is organized in ascending order of the values for the KLD estimator (given in the second column).

For each compared pair of a word-formation process, the identical categories of one-hundred year spans were created in a chronological order, which included their type frequency. The oldest word-formation processes also contained an additional category that encompassed a six-hundred-year span (from 600 to 1200). The reason for expanding the boundaries of this category was that the number of words formed during this period was much smaller than that of other categories, and the year gaps between the first record of words were large (from 200 to 500 years). Further, the number of categories differs from one word-formation process to another and ranges from 6 to 10, depending on the specificity of their type frequency distribution across centuries.

The diachronic productivity of the studied word-formation processes is considered similar if the statistics of their obtained confidence intervals allow us to accept the null hypothesis of no difference between their distributions. As shown in Table 40 in Appendix G, the suffixes *-er* and *-or*, *dis-* and *-ment*, *re-* and *-ist*, *-ee* and *-ism*, *-ist* and *-ism*, *-ery* and *-ence*, and *-ment* and *-al* have been identified as similar with all three KLD estimators, with *-er* and *-or* showing the least difference in information. Further, the diachronic productivity of the affixes *re-* and *-ism*, and *-ity* and *-ment*, shows as similar with KLD and Turing's perspective estimators, and *-ship* and *-er* with symmetrized KLD and Turing's perspective estimator. Finally, the similarity of the diachronic productivity of the suffix *-ion* and *-ence*, *-age* and *-ity*, and *-ness* and *-ship* appears as statistically significant only with Turing's perspective estimator.

In addition, the pairs of noun-formation processes {C-ing} and {C-C}, {C-C} and {C-ness}, and {dis-C} and {re-C} have the largest information discrepancy in their diachronic productivity. It is also interesting to note that compounding—i.e.{C-C}—is different from all other noun-formation processes in that its diachronic productivity has been increasing towards modern times. Finally, the comparison of values of KLD and symmetrized KLD is informative about how symmetric the distributions are, and it has revealed that the pairs *-ion* and *-ness*, *-ity* and *-ness*, *-ment* and *-ness*, C-C and *-ness*, *-ee* and *-age*, and *re-* and *-ism* emerge as the most asymmetric.

7.2.3 The diachronic picture of the most type-frequent adjectival morphological patterns

This section focuses on the following 15 morphological patterns which have been identified as the most type-frequent in the metacorpus: {C-ed}, {C-ing}, {C-able}, {C-al}, {C-y}, {C-ful}, {C-less}, {C-C}, {un-C}, {C-ive}, {C-ous}, {C-ic}, {in-C}, {C-ish} and {C-ly}. Their diachronic productivity is visualized in Figures 7.16–7.22 and compared in Table 41 (Appendix G).

The word-formation processes {C-ed} and {C-ing}

It is interesting to observe that the diachronic productivity of the adjectival suffixes *-ed* and *-ing* is similar (Figure 7.16): the similarity of histograms has motivated me to consider them together. Their type frequency started developing in the 1200s, peaked in the 1600s and decreased by 2000. The first level of derivation in these word-formation processes is the most productive, with a higher level of derivation developing later in time.

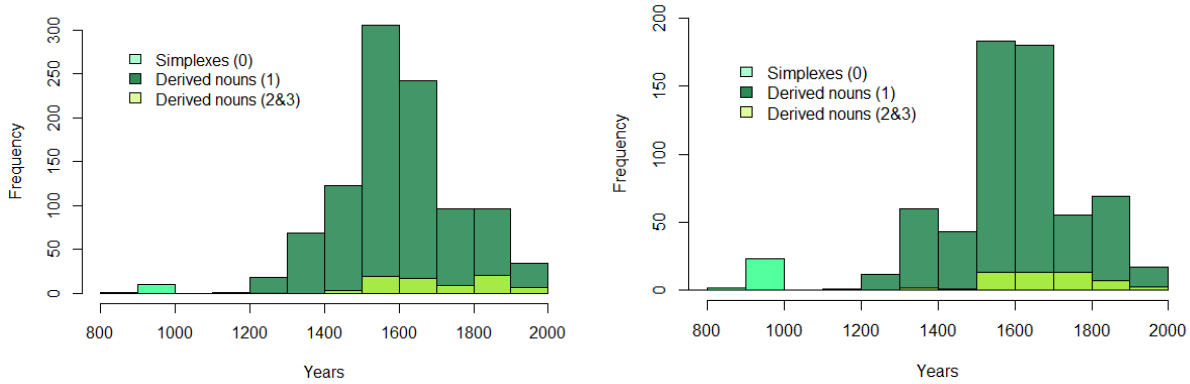


Figure 7.16. The diachronic productivity of the adjective-formation processes {C-ed} (left panel) and {C-ing} (right panel)

The word-formation processes {C-able} and {C-al}

The suffixes *-able* and *-al* are similar in that they entered the language in around 1300 with a number of loan words and were identified as native morphemes, with the highest productivity occurring on the first level of derivation (Figure 7.17). In the 1600s and 1800s, both processes reached the peak of their realized productivity, which then dropped considerably. Multimorphemic derivation in these processes is rare and appeared later in time.

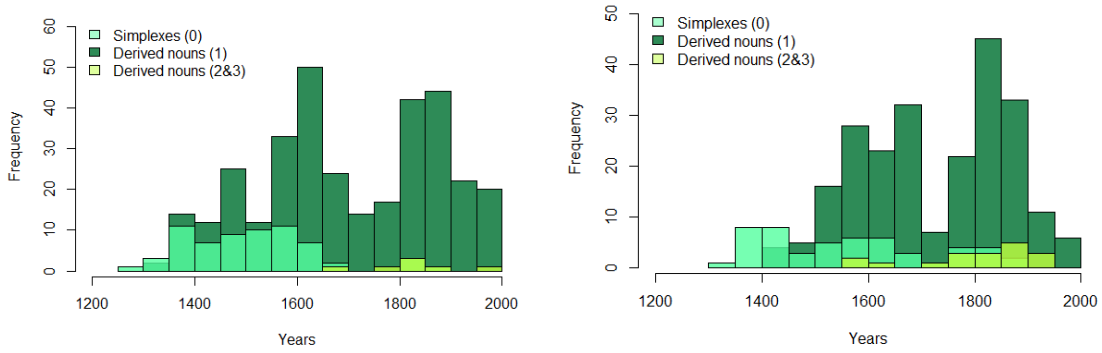


Figure 7.17. The diachronic productivity of the adjective-formation processes {C-able} and {C-al}

The word-formation processes {C-y}, {C-ful} and {C-less}

A distinct feature of these word-formation processes is that their realized productivity reached an optimum in three waves: in the 1400s, 1600s and in 1900s (Figure 7.18). As with other adjectival processes described above, the first derivational level of these suffixes has been the most productive in the course of their history.

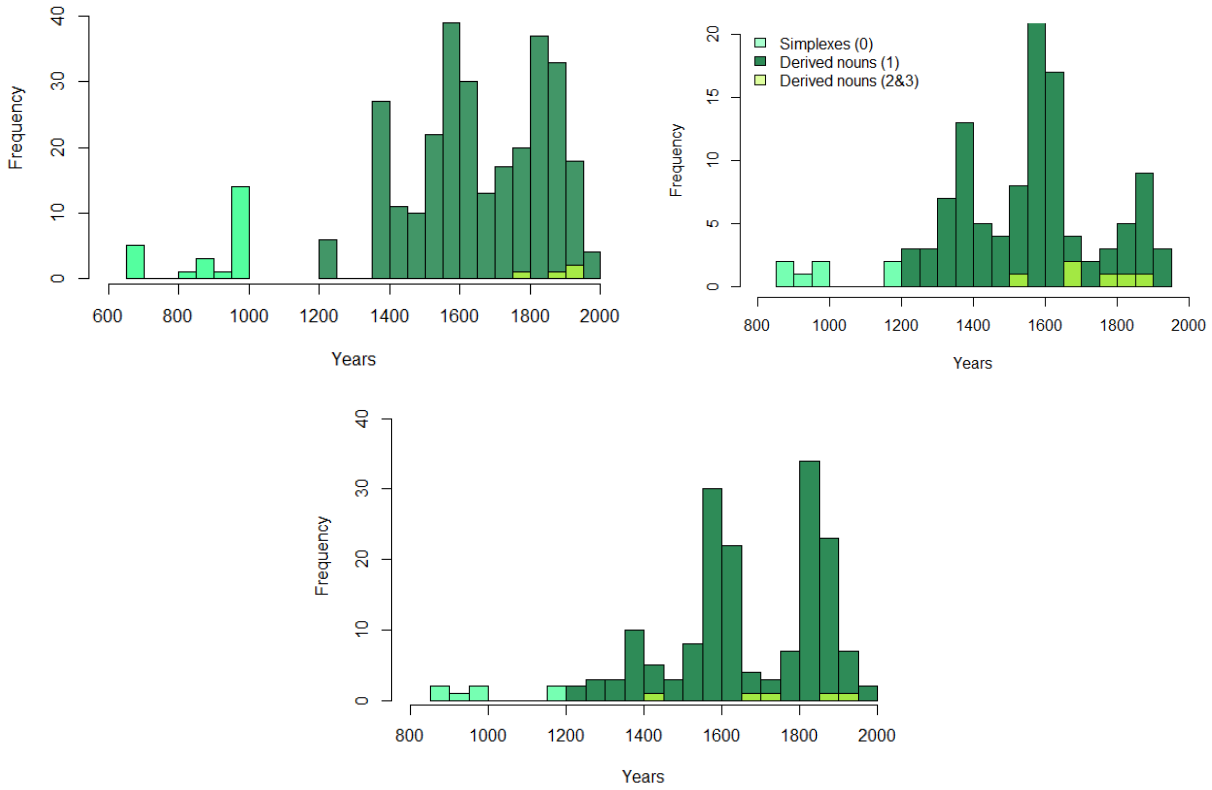


Figure 7.18. The diachronic productivity of the adjective-formation processes {C-y} (upper left panel), {C-ful} (upper right panel) and{C-less} (bottom middle)

The word-formation processes {C-C} and {un-C}

As informed by Figure 7.19, a common feature of compounding and prefixation with the involvement of *un-* is that, unlike other adjectival word-formation processes, they produce a large number of words on the second and third levels of derivation. However, these processes display a different dynamics of diachronic productivity: whereas the prefix *un-* had reached the peak of its productivity around the 1600s, the number of types for compounding continued to grow constantly until the 1900s.

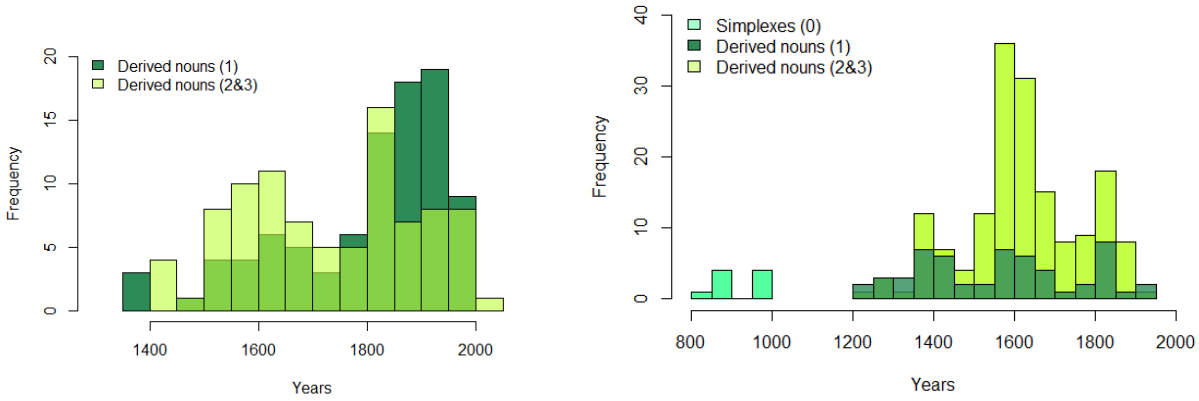


Figure 7.19. The diachronic productivity of the adjective-formation processes{C-C} (left panel) and {un-C} (right panel)

The word-formation processes {C-ive} and {C-ous}

The suffixes *-ive* and *-ous* developed in English from a high number of borrowed words (Figure 7.20). They were realized as native morphemes between 1300-1400, with the suffix *-ive* producing the largest number of types in the 1600s and 1800s, and the suffix *-ous* in the 1600s.

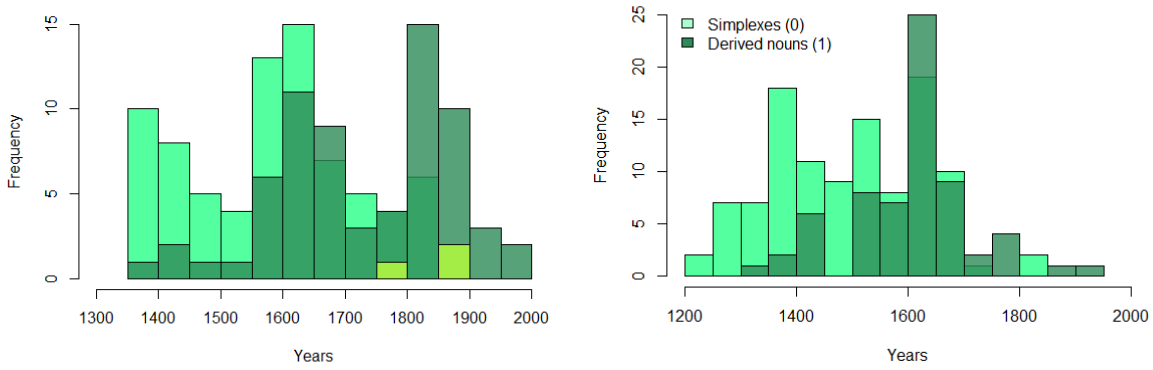


Figure 7.20. The diachronic productivity of the adjective-formation processes{C-ive} (left panel) and {C-ous} (right panel)

The word-formation processes {C-ic} and {in-C}

The development of these affixes show a similar trend (Figure 7.21). They appeared in the language with borrowings and showed two distinct waves of the first-level derivation around 1600s and 1800s. These processes are also productive on the higher levels of derivation.

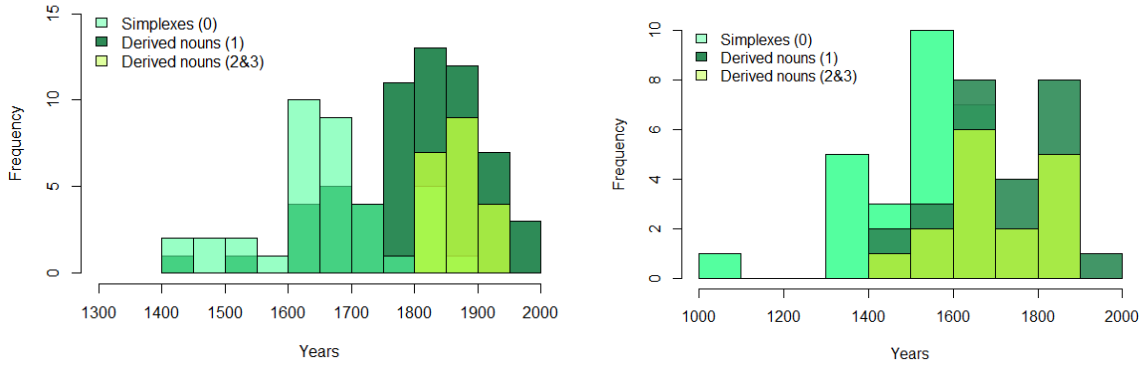


Figure 7.21. The diachronic productivity of the adjective-formation processes{C-ic} (left panel) and {in-C} (right panel)

The word-formation processes {C-ish} and {C-ly}

These suffixes have a long history in English. The first-level derivation of the suffix *-ish* started around 1300 and was consistently productive through the whole history of its development (Figure 7.22). The suffix *-ly* has been largely productive as adverb-forming but, as attested by the OED, it was also used in the formation of adjectives, peaking around the 1600s.

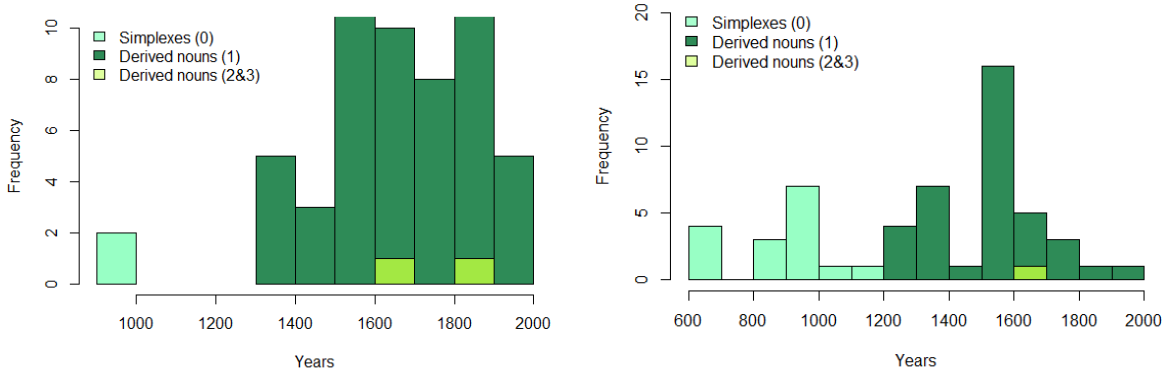


Figure 7.22. The diachronic productivity of the adjective-formation processes{C-ish} (left panel) and {C-ly} (right panel)

7.2.4 The overall picture of the diachronic development in adjectives

The histogram analysis of the diachronic productivity of adjectives has shown that their development is similar to that of nouns. The 1600s and 1800s were the most important time for English derivation, with the realized productivity decreasing towards the present day. As far as compounding in adjectives is concerned, however, its productivity had been rising from 1400 to

1900. Hence, the declining formation tendency of compounding in adjectives is more recent. The novel contribution of these comparisons is that they are performed with the consideration of different levels of word formation.

7.2.5 The statistical comparison of diachronic productivity in adjective formation

The goal of this subsection is to compare the diachronic productivity of the studied adjectival processes with the help of the KLD estimators (for more detail, see Section 4.3). The statistics of this comparison are given in Table 41 (Appendix G). The range of information discrepancies (KLD values) is larger for adjectives (0.03096907–1.433104 nats) than for nouns (0.03179659–0.7816313 nats).[53]

The pairs of the adjective-formation processes whose diachronic productivity has been identified as similar with all three KLD estimators are as follows: *-ish* vs *-less*, *-y* vs *-ish*, *in-* vs *un-*, *-y* vs *-less*, *-able* vs *-ive* and *-ish* vs *-ful*. Moreover, the pairs of *-ed* vs *-ity*, *-ed* vs *un-*, *-able* vs *-al*, *-able* vs *-less*, *-y* vs *-ful*, *-ful* vs *-less*, *-ic* vs *-ish*, {{C-C}} vs *-ish* and *-ish* vs *-ly* display similarities with the Turing's perspective estimator. These results allow us to conclude that, although each word-formation process is idiosyncratic, some of them have analogous patterns of development over a long period of time. In addition, the pairs of the suffixes *-ed* vs *-ing*, *-ish* vs *-less*, *-y* vs *-ish* show the least difference in information,[54] and the pair of *-ic* and *-ous* the greatest difference. The latter is also the most asymmetric, which is evident from a large discrepancy between the values of KLD and symmetrized KLD for this pair, whereas the most symmetric pairs are those with the least information discrepancy, which probably also adds to the similarities of diachronic productivity in these pairs.

7.2.6 The diachronic picture of the most type-frequent verbal morphological patterns

This section focuses on diachronic productivity of the verb-formation processes: {C-ize}, {re-C}, {C-ate}, {C-en}, {un-C}, {C-le}, {mis-C}, {C-C}, {de-C}, {dis-C} and {en-C}. As compared to nouns and adjectives, the histograms of diachronic productivity in verbs look more heterogeneous, which suggests a greater developmental diversity of realized productivity.

[53] It is difficult to explain why adjectives show a larger range of discrepancy in relative entropy. A possible explanation is that the diachronic productivity of adjectives is more diverse.
[54] Since the estimators of relative entropy are the information measures, the difference between distributions implies a difference in information between them.

The word-formation processes {C-ize}, {re-C} and {C-ate}

The suffix *-ize* started entering the language along with borrowed words (Figure 7.23). There were two peaks in the formation of verb types with this suffix: in the 1600s and 1800s. It also displays realized productivity on the higher levels of derivation, which are predominantly later formations. By contrast, the suffix *-ate* is productive only on the first level of derivation, which has probably emerged as a result of a high number of borrowed verbs. The prefix *re-* was also present in a large number of loan words and then, from the 1300s, participated in the derivation of new types in two distinct waves of productivity, peaking in the 1600s and the 1800s.

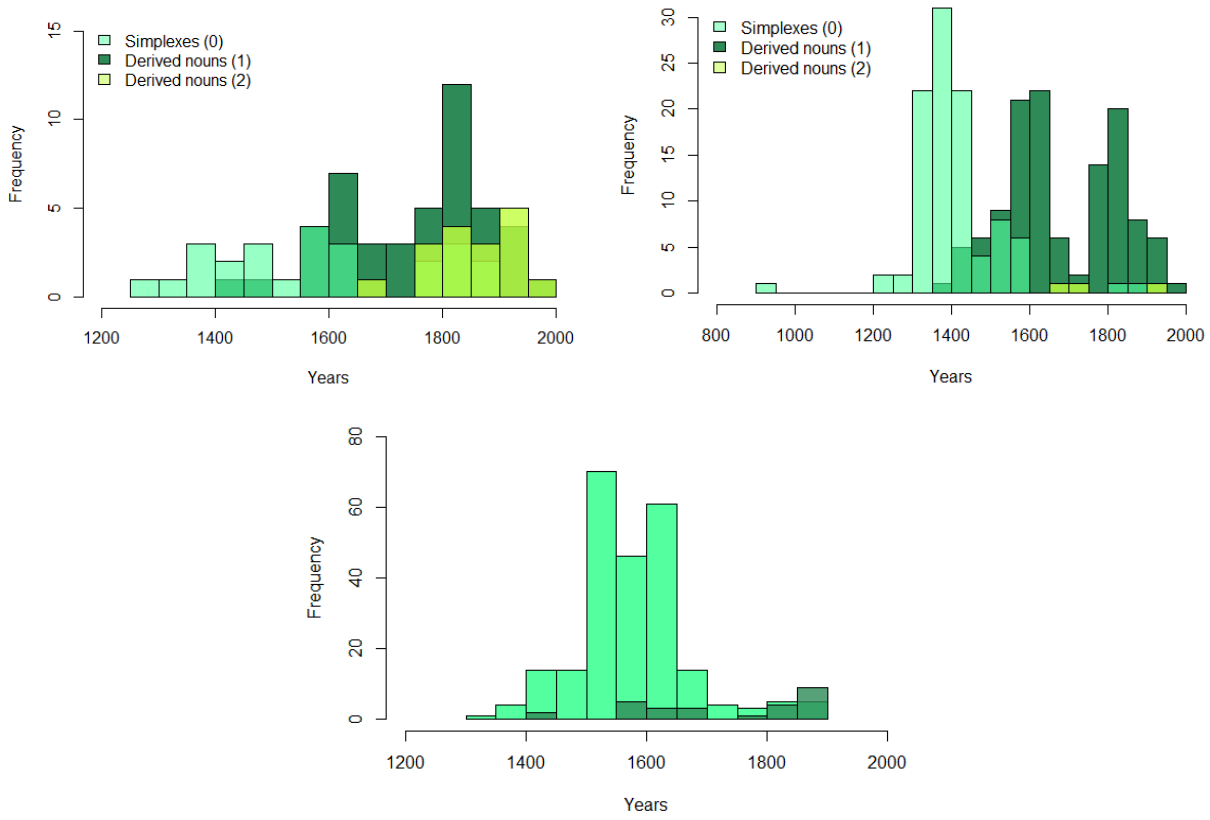


Figure 7.23. The distribution of the realized productivity in {C-ize} (upper left panel), {re-C} (upper right panel and {C-ate} (bottom middle panel) across years

The word-formation processes {C-en} and {un-C}

The developmental dynamic of these affixes is similar in that they were both productive in Old English and their productivity peaks in 1400, 1600 and 1800 almost coincide. However, the suffix *-en* had a higher realized productivity over the years, which dropped by 1900, whereas the prefix *un-* has maintained some degree of productivity to the present day.

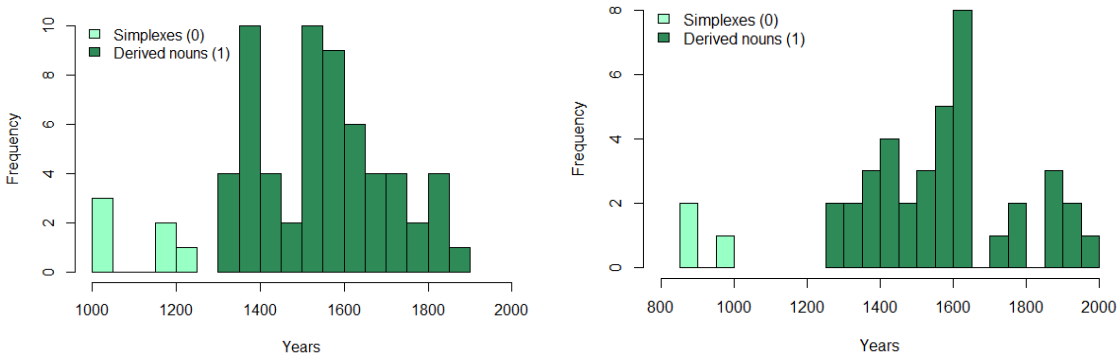


Figure 7.24. The distribution of the realized productivity in {C-en} (left panel) and {un-C} (right panel)

The word-formation processes {C-le}, {mis-C} and {C-C}

These processes are the most productive on the first level of verb formation (Figure 7.25). The diachronic productivity of the prefix *mis-* reached its peak in the 1400s and that of the suffix *-le* in the 1500s, and then it declined consistently by 2000. Conversely, similar to nouns, the productivity trend for compounding has increased over time, peaking between 1900 and 2000.

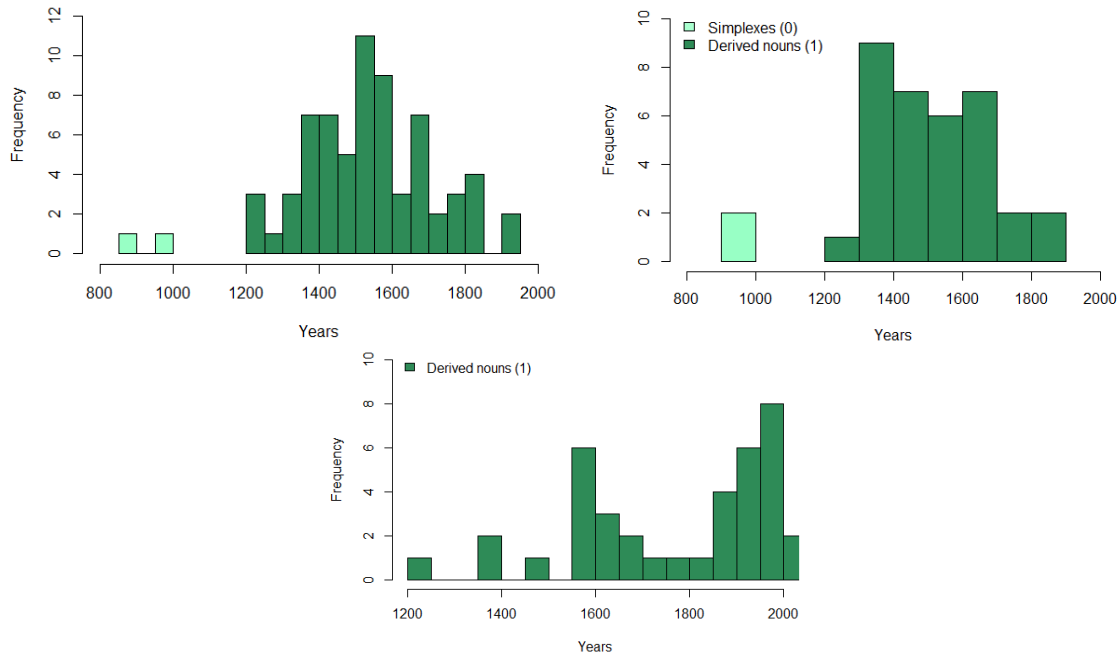


Figure 7.25. The distribution of the realized productivity in {C-le} (upper left panel), {mis-C} (upper right panel) and {C-C} (bottom middle)

The word-formation processes {de-C}, {dis-C} and {en-C}

These Latin prefixes appeared in English from 1200 onwards together with borrowed verbs (Figure 7.26). The productivity of borrowings peaked in the 1400s. Although the tendency of loan productivity in these processes is similar, this is not the case for their derivational dynamics. The

first-level derivation of the prefix *de-* developed between 1500 and 1600 and reached its peak (together with the higher levels of derivation) between 1800 and 2000. On the other hand, the derivation of *dis-* and *en-* began earlier, between 1300-1400, and reached its peak around the 1600s. Further, whereas the prefix *dis-* emerges as unproductive after 1800, the prefix *en-* had an increase in realized productivity between 1900 and 2000.

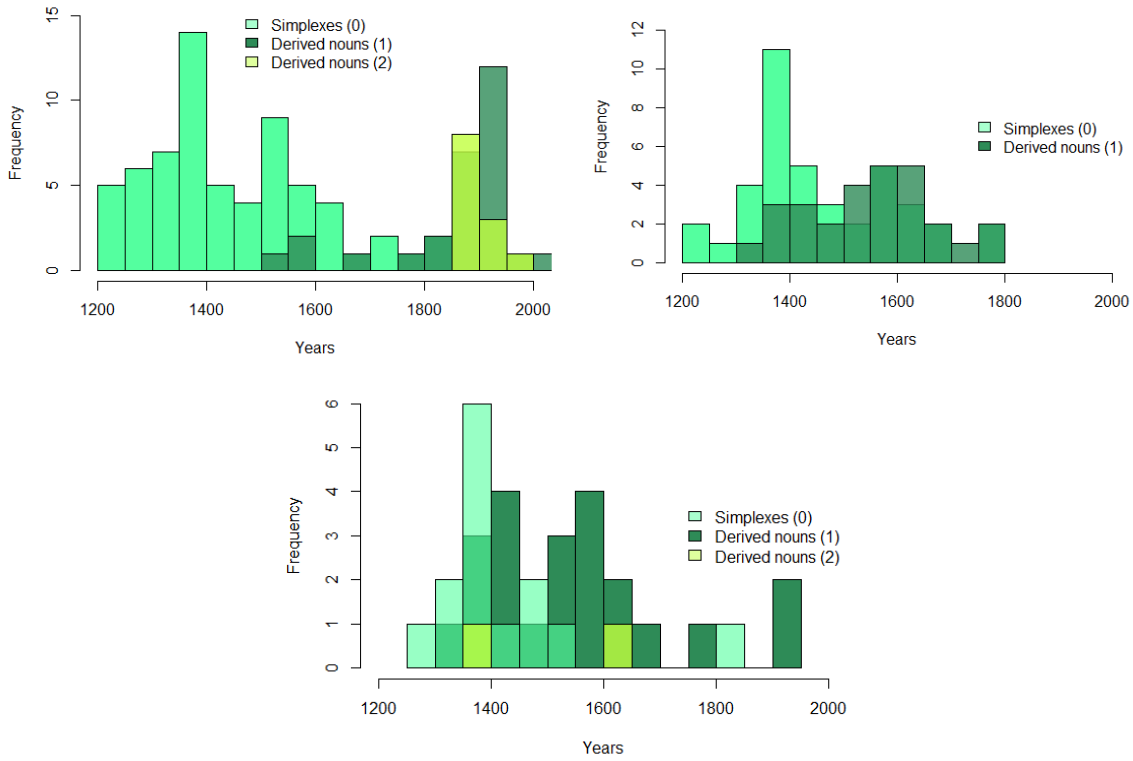


Figure 7.26. The distribution of the realized productivity in {de-C} (upper left panel), {dis-C} (upper right panel) and {en-C} (bottom middle)

7.2.7 The overall picture of the diachronic development in verbs

The diachronic productivity in verbs forms a diverse picture. It has been established that 1300 was an important year for lexical borrowings, and 1600 and 1800 were important years for derivation. From a derivational perspective, most of the verb affixes are only productive on the first level of morphological complexity, except for *-ize* and *de-* which also show productivity on the higher levels of derivation. As with nouns and adjectives, most of the morphological verb-formation processes have been on a declining trend, apart from compounding, whose realized productivity has increased to the present day.

7.2.8 The statistical comparison of the diachronic productivity in verb formation

The KLD values (presented in Table 42, Appendix G) for the pairs of verb processes show the largest variation (0.01905298–2.020724 nats), as compared to nouns and adjectives, which echoes with the heterogeneous patterns observed in the above-described histograms. However, it is interesting to observe that a larger number of pairs show similarity in their diachronic productivity, as compared to nouns and adjectives: *-le* vs *-en*, *re-* vs *dis-*, *-en* vs *un-*, *de-* vs *dis-*, *mis-* vs *en-*, *mis-* vs *re-*, *un-* vs *mis-*, *re-* vs *un-*, *re-* vs *en-*, *re-* vs *de-* and *en-* vs *-en*. These are the processes, whose similarity has been confirmed with all three KLD estimators. The pair *dis-* vs *mis-* has been identified as similar with the help of the KLD and symmetrized KLD estimators, and the pair *-le* vs *-ate* with the Turing's perspective estimator. Further, the smallest difference in information is registered for the pair *dis-* vs *mis-*, and the largest for the pair {C-C} vs *-ate*. As observed for nouns and adjectives, the largest information discrepancy between affixes implies the largest asymmetry between the distribution of their realized productivity, and the smallest discrepancy the symmetry between them. With the lowest difference between the values of KLD and symmetrized KLD, the pair *dis-* vs *mis-* is the most symmetric, whereas the pair {C-C} vs *-ate*, with the largest difference between these values, shows the largest asymmetry observed for all word classes.

7.2.9 The difference between the KLD estimators

The mathematical theory behind the KLD estimators is new (Zhang 2017), and no studies (to the best of my knowledge) have been performed to establish their properties in real life. Therefore, it is difficult to interpret these estimators conclusively. However, the KLD analyses carried out in this section have revealed thought-provoking observations which are worth outlining, in order to develop a better understanding of the KLD estimators as applied to language. Whereas most of the time KLD and symmetrized KLD behave similarly, the Turing's perspective estimator yields different statistical results. One possible explanation of the observed discrepancies in estimation is that the Turing's perspective estimator has a larger statistical power and might be more sensitive to distributional distortions of the data. Nevertheless, further research is needed to substantiate the nature of the introduced estimators.

7.3 The cluster analyses of affixes

The aim of this section is to identify clusters of affixes that show similar characteristics, based on their values of the type frequency, token frequency, productivity, type valency and type-token ratio (RQ5). Specifically, cluster analyses are helpful in understanding the interactions between

different forces that shape derivational processes. Moreover, they allow for identifying overall trends in the data and potential directions for future research.

The assumption behind using the above-mentioned variables is that on a highly general level of language, more universal currents can be identified. That is, although these variables are expected to be different in various texts/corpora/modes, when we consider them in their entirety, more general patterns would emerge. The ideal scenario for testing this assumption would be to collect the values for these parameters from all instances of use of affixes over certain periods of time—an unrealistic pledge since, at the current stage of science development, no such register seems possible. Moreover, even with the registers available nowadays (e.g. BNC, Brown Corpus, COCA, CELEX, OEC), this task is a Herculean task, because the precision of the automatic morphological parsers in corpora is low and requires thorough manual checking (e.g. Bauer et al. 2013: 42). Nevertheless, the field of morphology has made considerable progress in recent decades, and there are a few studies that offer morphological datasets, which can be used to identify the behavior of affixes on a more universal level. Although not sufficient for definitive conclusions, the combined picture of these datasets allows us to catch a glimpse of the profiles of affixes and of the interaction between their different parameters.

Thus, in what follows, subsection 7.3.1 describes the results of the hierarchical cluster and the k-medoids analyses, and subsection 7.3.2 the results of the PCA analysis. These analyses are used for the purpose of identifying similarities/differences between the most type-frequent affixes in the sample. The data set for these analyses has been compiled from the following sources[55]: MorphoQuantics (Laws & Ryder 2014), MorphoLex (Sánchez-Gutiérrez et al. 2018), CELEX (2001) and the morphological corpus of this study (abbreviated as MQ, ML, C and OED.s, respectively). From these sources, the values of type frequency (TF_MQ, TF_C, TF_ML and TF_OED.s), token frequency (TokF_MQ and TokF_ML), potential productivity (P_C and P_ML), expanding productivity (P._ML), type valency (TV) and type-token ratio (TTR_MQ and TTR_ML) have been collected. In addition, the dataset has been complemented with a supplementary variable of the origin of affixes, as informed by MorphoQuantics (Laws & Ryder 2014), containing the following categories: Old English (OE), Old English from Germanic (OEG),

[55] Different datasets have been used in this study for two reasons. First, in general, the method of cluster analysis requires a larger number of variables (specifically, PCA). Second, as described at the beginning of this section, the aim of the current study is to identify more general trends in how different morphological forces interact with each other.

Gothic (GT), Anglo-Norman (AN), Old French (OF), Old French from Latin (OFL), French from Greek (FG), French from Latin (FL) and Latin (L). The word-formation processes of nouns, adjectives and verbs, described above in the section on diachronic productivity (41 in total), have been understood as the observations of this eclectic dataset for the cluster analyses of this study. This dataset is introduced in Appendix H.

Since not all values for the chosen observations were present in the sources, an imputation has been performed with the *R* package 'mice' (van Buuren et al. 2021) to predict the missing values. For this purpose, the built-in univariate imputation method of 'classification and regression trees' ('cart') has been chosen, because the data consisted of mixed variables (ordered and continuous) displaying non-normal distribution. In Table 43 of Appendix H, the imputed values are coloured in blue. The use of imputation adds to the speculative nature of the created cluster models: in fact, these cluster analyses can be viewed as linguistic experiments. Therefore, the identified clusters are not absolute entities without a possibility of modification. Rather, they are hypothetical groups of affixes that show some similarities and hint at some trends in the data. A further verification of the clusters is feasible with a larger data set and may constitute content for future research.[56]

7.3.1 Hierarchical cluster and k-medoids analyses

The package 'shipunov' was used to perform the hierarchical cluster analysis. First, the most suitable method for its application was identified. As shown in Figure 7.27, the methods of 'centroid', 'median' and 'average' suit the data most.

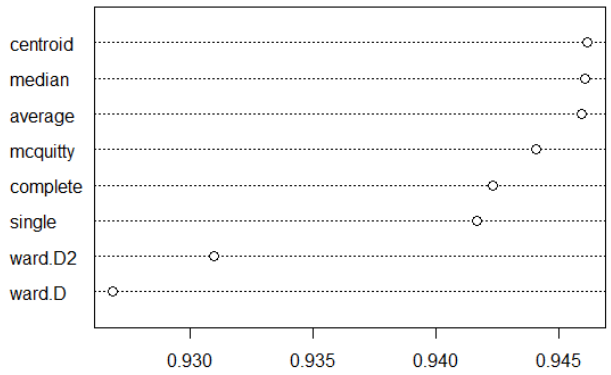


Figure 7.27. The importance plot of the methods for the hierarchical clustering

[56] As yet, morphological processes were largely studied with the data from one corpus. Collecting morphological data from different corpora/registers and processing morphological variables with different cluster techniques will help linguists identify more universal currents in language, otherwise hidden from our perception.

Hence, the method 'median' was applied in the creation of the averaged dendogram on 1000 replicates in Figure 7.28, which finds a proximity between two clusters on the basis of the proximity between their geometric centroid, i.e. a squared Euclidean distance between them. At the height of almost 1500, this dendogram depicts three clusters, which are further split into smaller clusters, as the height of the dendogram decreases. The first cluster is the largest and contains 32 suffixes. The second cluster consists of one suffix (*-ion*) and the third of three suffixes. The trend observed in the organization of suffixes in these clusters is largely related to their proportion of type and token frequency. The first cluster encompasses a mini-subcluster of four suffixes that have high type and token frequency, and a larger subcluster (from the adjectival suffix *-ing* to the adjectival prefix *in-*) with a lower type and token frequency. Suffixes in the second subcluster tend to have slightly higher values of type-token ratio and productivity, and their token frequency increases towards the end of the subcluster. The second cluster is composed of the suffix *-ion* which is characterized by the highest number of tokens and a relatively lower number of types. The third cluster, similar to the second, contains suffixes with a high token frequency (although slightly lower than in the previous cluster) and a medium type frequency. There is a rising tendency for the token frequency starting from the second subcluster on the left up the second cluster formed by the suffix *-ion* on the right. This observation highlights the impact of token frequency on the formation of clusters.

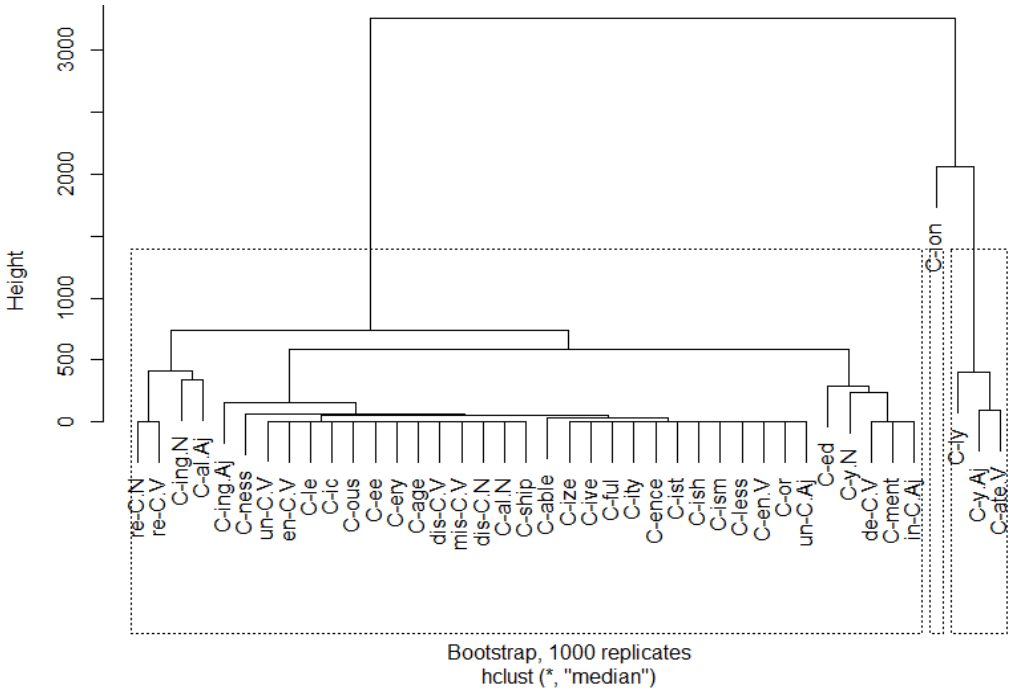


Figure 7.28. The consensus dendogram, averaged on 1000 replicates

A slightly different picture arises with a k-medoids cluster technique with the number of clusters specified as three (k=3), performed with the package 'cluster' (Maechler et al. 2021). Figure 7.29 visualizes these clusters: they overlap—which is an advantageous feature of the algorithm of Partitioning Around Medoids—meaning that some observations can be assigned to several clusters at the same time. The average silhouette width of the whole dataset is 0.68, which suggests a robust clustering. The shadow values for clusters (used as a diagnostic tool with the help of the package 'flexclust' (Leisch 2018)) are also low and away from 1, confirming the robustness of the identified clusters (the 1st: 0.4296299; the 2nd: 0.2573683: the 3rd: 0.5098882).

The general trend observed in these three clusters, however, is the same as in the above-discussed dendogram, but more vividly expressed and is largely driven by the proportions of type and token frequency. The first cluster (in pink) unifies suffixes with a higher type and token frequency and with a lower type-token ratio and productivity. The affixes in this cluster are influenced by a high token frequency. The second cluster, coloured in green, which is also the largest, contains suffixes with lower values of type and token frequency, and higher values of type-token ratio and productivity. Finally, the third cluster has a relatively lower value of type frequency and high values of token frequency, as well as medium values of productivity and type-token ratio. Further, the adjectival affixes *-able*, *-ing* and *in-*, and the nominal suffix *-y* lie in the overlapping area of three clusters, indicating that their profile can fit into all of the clusters. The identified clusters overlap in this way, because some of the quantitative characteristics of affixes coincide with the adjacent clusters, bringing to light individual differences between affixes and the variability within each category.

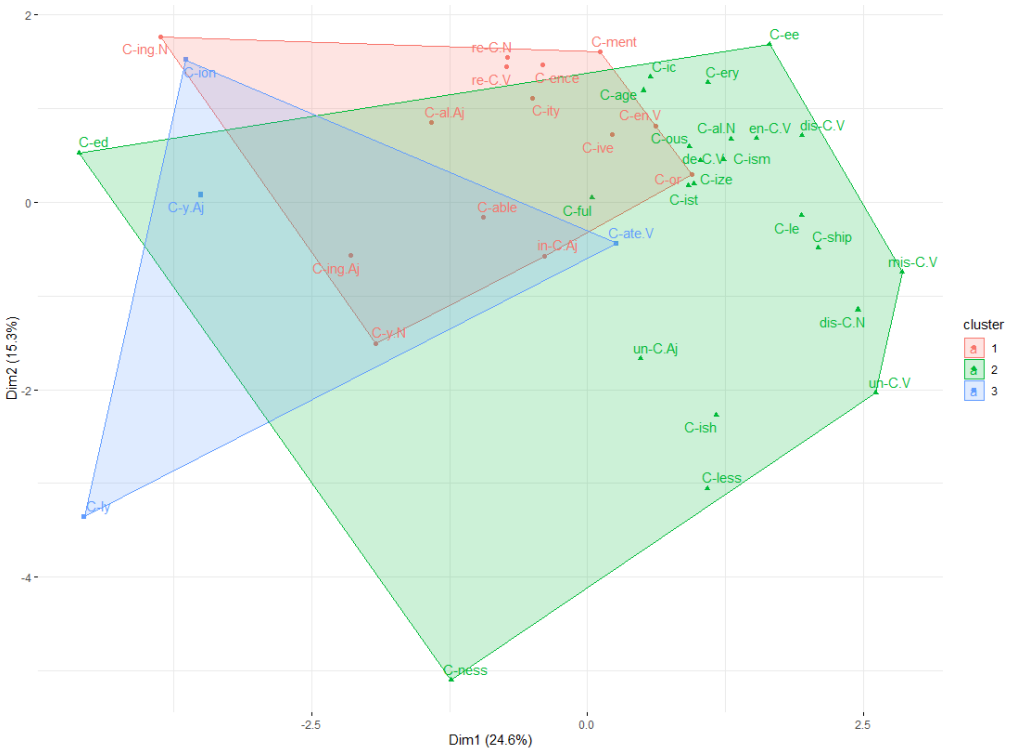


Figure 7.29. The k-medoids plot of affixes

7.3.2 The PCA analysis

In the previous section, the clusters of suffixes were identified, which give some clues about the interaction between the variables. It is possible to conclude from the performed cluster analyses that, in general, suffixes with a higher type and token frequency tend to have a lower productivity and type-token ratio, and vice versa. With the PCA analysis, in particular with its biplots of variables, we can get a closer look at the interaction between the variables and can establish more precisely what their impact on derivational processes is. The PCA analysis has been performed with the help of the package 'FactoMineR' (Husson et al. 2008). Before running the analysis, the data were log transformed with the formula *log(x – (min(x) – 1))*, because the PCA algorithm is based on Pearson correlation, and the data set for the analysis contains zeroes and discrete variables. Furthermore, during the PCA analysis, all variables were scaled.

Figure 7.30 features the biplot of the variables of the dataset for the first two components. In this biplot, the variables are represented by arrows coloured by a heat map according to their contribution to the model. To put it simply, we can think of each arrow as a force shaping a word-formation grammar of English. The biplot, thus, illustrates the dynamics between those forces.

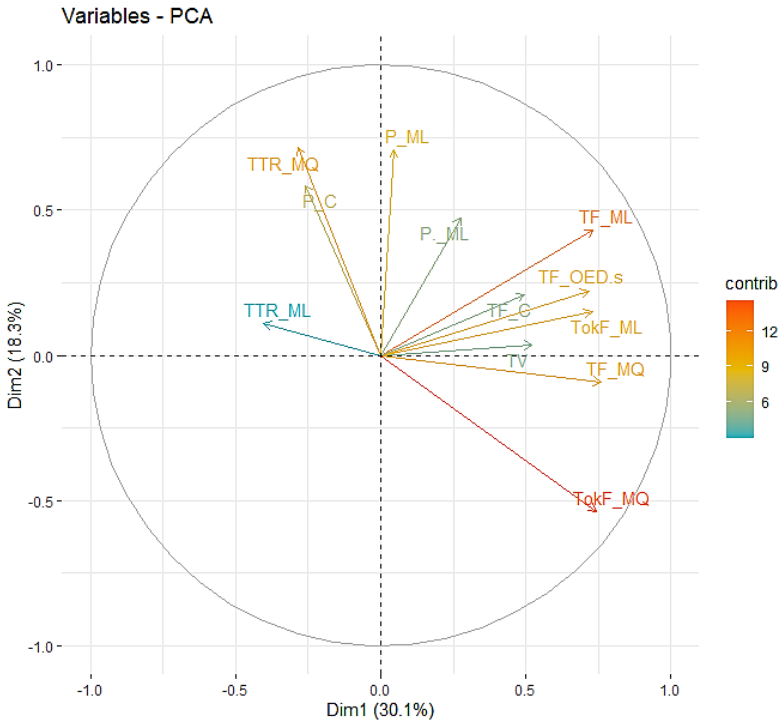


Figure 7.30. The biplot of variables in the PCA analysis
(the arrows are coloured with a heat map in accordance with their contribution to the model)

The first observation that catches the eye is that the token frequency (taken from MorphoQuantics) and the type frequency (MorphoLex)[57] have the greatest contribution to the model, which echoes with the clustering models of k-medoids and dendogram, driven by the values of the token and type frequency. Secondly, although spread across two quadrants, type and token frequencies are clustered together in a bunch of arrows with a small distance between them (except for TF_MQ). This observation conveys the idea that these measures are positively correlated and that the difference between them (within a category) can be considered as a variational pattern. Another general trend is that the tokens of suffixes are negatively correlated with the measures of productivity and type-token ratio (i.e. TokF_MQ vs TTR_MQ, TokF_ML vs TTR_ML and TokF_MQ vs P_C), which is evident from an obtuse angle between arrows. This trend asserts the view found in the literature that suffixes with high token frequency are not usually the productive ones (Bybee 2007a: 14). Moreover, the biplots show that the hapax-based measure of productivity, or expanding productivity (P._ML) is located between the arrows of potential productivity (P_ML) and of type frequency, which suggests its positive correlation with these adjacent variables.

[57] The used sources of data and the motivation for their use are explained in Section 3.6 and in the beginning of the current chapter. The variables used in all discussed cluster analyses were introduced on p. 253.

Finally, the type valency of affixes (TV) has little contribution to the model (the size of the arrow is smaller, and its colour is greenish), but it is worth noting that the arrow for the type valency is located very close to the arrow for type frequency of the sample of this study (OED.s), which asserts a strong positive correlation between these variables.

The biplot in Figure 7.31, then, shows how the data points are distributed across two principal components (the bar plot of components is given in Figure 11 of Appendix H). The greatest portion of the variance is explained by the first principal component on the *x* axis (30.1%), which has the largest spread of the data points, mostly because of the massive discrepancies between the values of the token frequency and those of type-token ratio and productivity, as well as the high significance of type and token frequencies for the model. The second portion of the variation is accounted for by the second principal component on the *y* axis (18.3%), featuring the number of types and productivity as the major course for this variation. Many of the data points are located across the vectors of type and token frequencies which introduces these variables as the most influential in the model.

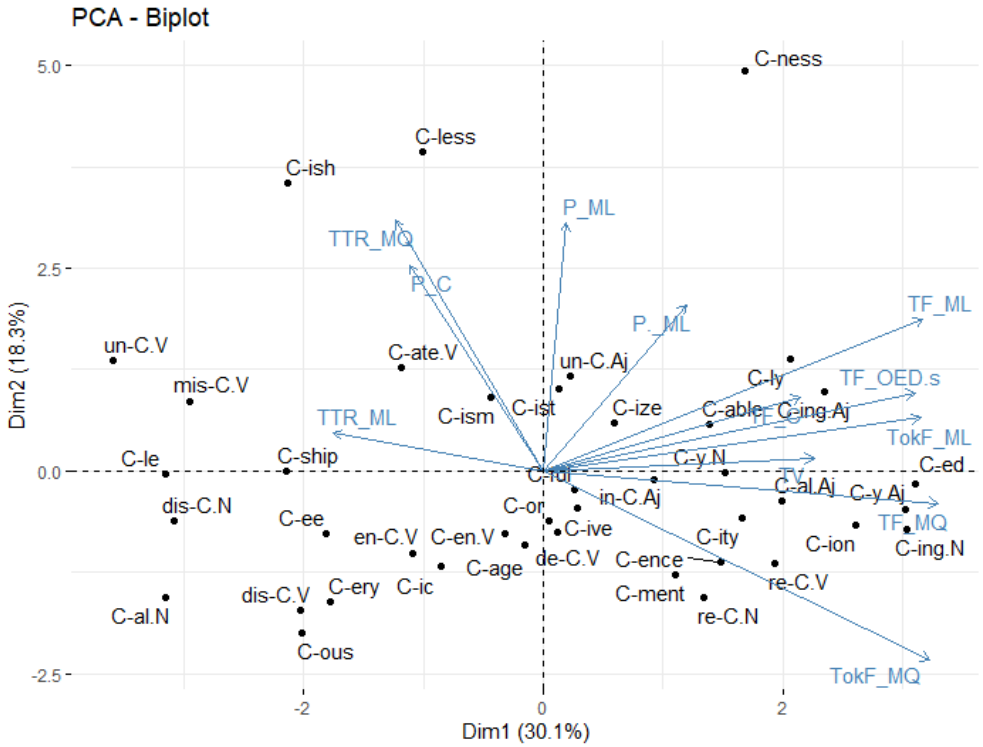


Figure 7.31. The PCA biplot for individuals and variables

More specifically, the derivational processes in the bottom right quadrant are greatly influenced by the token and type frequency in MorphQuantics (i.e. *re-*, *-ment*, *-ence*, *-ity*, *-ful* and *-or* by the

token frequency; *-al.Aj*, *-y.Aj*, *-ing.N*, *-ion*, *-ity*, *-ed* and *in-* by the type frequency), and the derivational processes in the top right quadrant by the type valency (*-y.N*), the type frequency in the morphological metacorpus (*-ing.Aj*), the type frequency of CELEX (*-able*), the type frequency of MorphoLex (*-ly* and *-ize*) and the potential productivity of MorphoLex (*un-* and *-ist*). Further, the data points in the bottom left quadrant are characterized by a considerably lower type frequency, as well as a higher potential productivity (as identified in CELEX) and a high values of type-token ratio. This quadrant encompasses the affixes *-less*, *-ish*, *-ate*, *-ism*, *-ship*, *mis-*, *un-* and *-le*. Finally, the derivational processes in the top left quadrant of the biplot are prominent for a lower type and token frequency, productivity, type valency and type-token ratio: *dis-.N*, *-al.N*, *dis-.V*, *-ee*, *-ous*, *-ery*, *en-*, *-en*, *-ic*, *de-* and *-age*. Figure 157 illustrates the discussed tendencies in the data with the help of a heat map based on the values of the type-token ratio (left panel) and the potential productivity (right panel): the affixes with the high type and token frequency tend to show a lower potential productivity and type-token ratio, as evident from the darker colour of affixes located below the *x*-axis line. Hence, we can observe an overlap between the impact of the potential productivity of affixes and their type-token ratio.

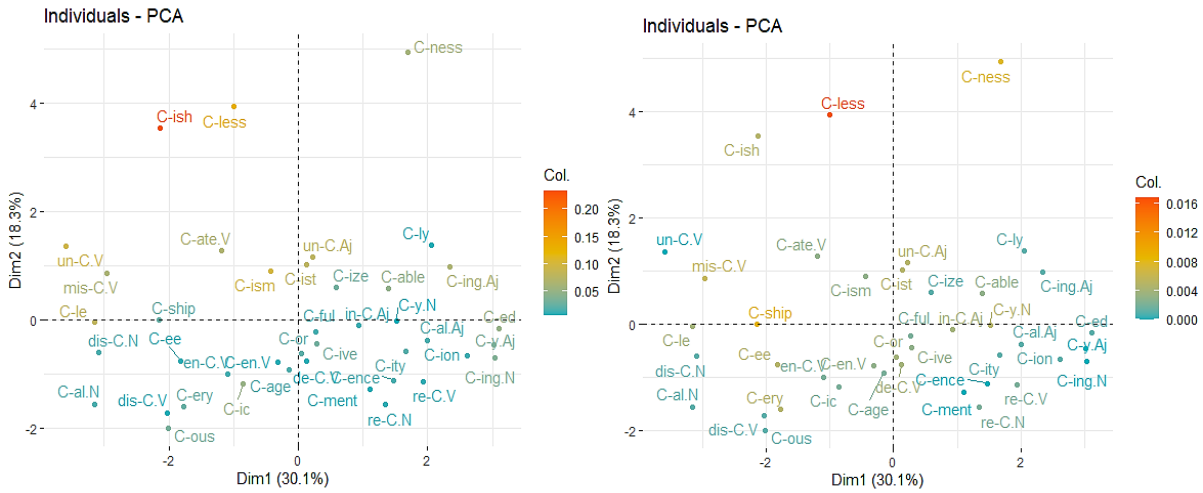


Figure 7.32. Biplots of individuals coloured by a heat map based on the type-token ratio in MorphoQuantics (left panel) and the potential productivity in CELEX (right panel)

Furthermore, there are three outliers in the model which add to the variation of the model in the second component: *-ness*, *-ish* and *-less*. They are located away from the rest of affixes for different reasons: the suffix *-ness* has a lower number of tokens, a higher number of types and a high value of potential productivity; the suffix *-ish* has the highest value of type-token ratio in MorphoQuantics and relatively higher potential productivity in CELEX; and the suffix *-less*

demonstrates the highest value of potential productivity in CELEX and a high value of type-token ratio in MorphoQuantics.

Furthermore, if we look at the biplot of individuals through the lens of word classes, the largest variation on the *x*-axis (the first component) and on the *y*-axis (the second component) is attributed to noun-forming affixes (*-al.N* vs *-ing* and *-ness* and *re-.N*). The second largest variation is observed for adjective-forming suffixes (*-ous* vs *-ed* and *-ish vs -ous*). Finally, the verbal affixes *un-* vs *re-* add to the variability in the first component and the affixes *-ate* vs *re-* in the second.

Lastly, the biplot of affixes in Figure 7.33 is marked for the origin of affixes. The observed trend is that affixes of Old English and those of Germanic origin (visualized in green) are driven more by their type frequency, potential productivity and type-token ratio, whereas the suffixes of Latin and French origin are characterized by a higher token frequency and they tend to be less productive.

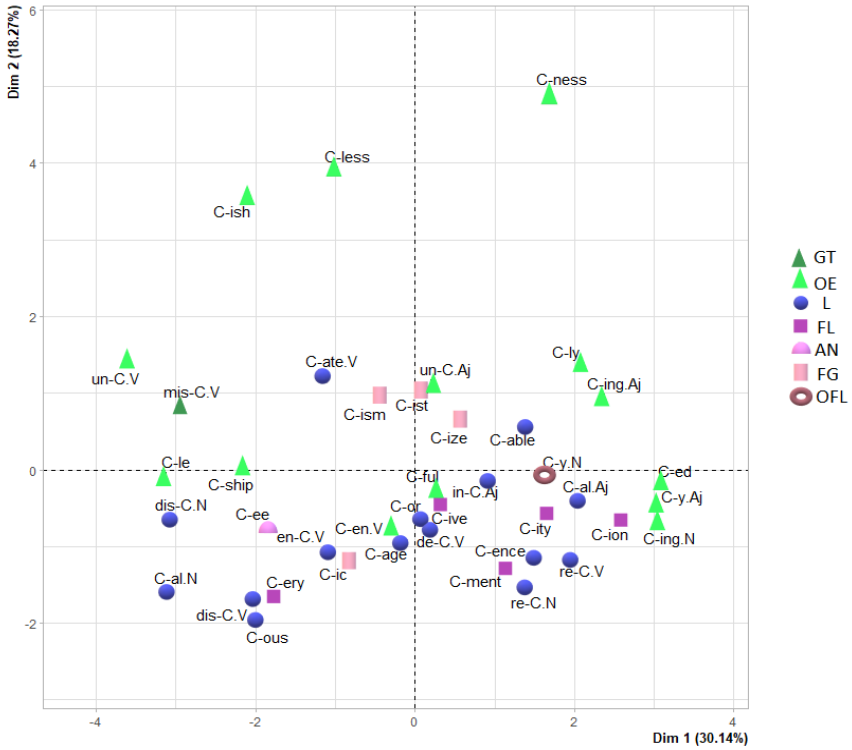


Figure 7.33. The biplot of affixes based on their origin

7.4 Conclusions

This section summarizes the main findings of Chapter 7. First, the performed correlation and regression analyses have established two statistically strong effects of the type frequency on the

word-formation system of English: (i) specific combinations of affixes are driven by the high type frequency of particular word bases; and (ii) the high type frequency of suffixes leads to their polyvalency. In the first case, the high type frequency seems to be a mediator of another grammatical/cognitive force that determines the dominance of specific word-bases in word-formation processes; in the second, it is a cause for a change of the initial condition of a word-formation rule that specifies a preferred word base for a suffix. The link between these two type-frequency effects has been theorized as a relation between the Unitary-Base Hypothesis (Aronoff 1976) and the base-driven hypothesis of selectional restrictions (Plag 1996).

Second, the comparison of the diachronic productivity of the most type-frequent morphological patterns in nouns, adjectives and verbs has revealed the similarities in the development of realized productivity of the following pairs of word-formation processes: *-er* and *-or*, *dis-* and *-ment*, *re-* and *-ist*, *-ee* and *-ism*, *-ist* and *-ism*, *-ery* and *-ence*, and *-ment* and *-al*; *-ish* vs *-less*, *-y* vs *-ish*, *in-* vs *un-*, *-y* vs *-less*, *-able* vs *-ive* and *-ish* vs *-ful*; *-le* vs *-en*, *re-* vs *dis-*, *-en* vs *un-*, *de-* vs *dis-*, *mis-* vs *en-*, *mis-* vs *re-*, *un-* vs *mis-*, *re-* vs *un-*, *re-* vs *en-*, *re-* vs *de-* and *en-* vs *-en*. Although each word-formation process is idiosyncratic, some of their developmental patterns of morphological productivity are alike, which suggests a more general derivational tendency—occurring on a wider scale and involving the whole language in the form of a typological shift. In this view, the analysis of the diachronic productivity of the studied word-formation processes has revealed a lexical significance of the years of 1300 and 1400, when a large number of loan words entered English, and a morphological significance of the years 1600 and 1800, when a large number of words were derived. In the considered timescale of the word-formation processes, a two-hundred-year span between peaks of realized productivity catches the eye, which might be the time when a word-formation process undergoes a full productivity cycle: from being morphologically productive to becoming unproductive and then productive again. If this assumption is true, a linguist who will study English diachronic productivity of word-formation processes in 2200 will find the 2000s, or perhaps 2100, as peak points for the realized productivity in many of these processes. This is because, as shown by the above-described histograms, the realized productivity of most derivational processes in English has been declining. The only exception is compounding, whose trend has been constantly increasing, specifically in nouns and verbs.

Another important finding of the current chapter concerns the KLD estimators. These statistical methods are new and have not been applied in the study of language. For this reason, our understanding of the properties of these estimators is limited.[58] Nonetheless, the study has found a difference in the behavior of the KLD and symmetrized estimators on the one hand and the Turing's perspective estimator on the other. This difference can be attributed to the fact that the latter is more precise and sensitive to the oscillations of the distributions in the data.

Finally, the cluster analyses performed with different techniques have identified three clusters of affixes, which are distinguished mainly on the basis of type frequency, token frequency and type-token ratio. It has been shown that suffixes with high type and token frequency have a lower potential productivity and type-token ratio—these are predominantly the suffixes of Latin or French origin. On the other hand, suffixes with a lower token frequency are more potentially-productive (and, thus, more salient), which agrees with the claims made in the corresponding literature. As shown with the PCA analysis, the affixes with a lower token frequency are mainly of Old English or Germanic origin. Furthermore, the PCA analysis has confirmed the nature of the associations between different parameters of word-formation processes: the token frequency of affixes has been demonstrated to have a negative correlation with their potential productivity, and the type valency of affixes a positive correlation with their type frequency. The current chapter has suggested that a study of word formation with data collected from different sources and registers would broaden our understanding of various morphological processes and substantiate a number of relevant hypotheses.

[58] The history of science, I believe, provides a large amount of evidence that new discoveries have become possible with new methods—they have always been driving forces behind the development of science. Although new methods might not be fully understood either due to their complexity or due to the lack of their empirical probing, this fact should not preclude us from taking courage to further explore them. The most evident modern example is the history of machine learning.

8 The overall conclusions

The first goal of this thesis has been empirical: to explore the English word-formation system with a sample of 32,000 individual lexemes. The methodology of this research marries formal and usage-based approaches to morphology. On the one hand, with formalization techniques, morphemes and words have been assigned to classes, which has led to the formation of the morphological metacorpus—a compilation of morphological patterns and constructions organized on the basis of word classes. On the other, the premise of the usage-based theory concerning the type frequency as a shaping force of grammar has opened up new avenues for the explanation of some morphological phenomena observed in the sample.

The second goal has been to create an overall picture of English word formation. It is helpful to think about this picture as a city image taken from a satellite. Aerial pictures of a city are different from those taken when exploring it on foot: they show overall boarders, connections and trends in the formation of the city. In some places, we can see buildings are cluttered tightly together with a larger number of pathways; in others, there are large spaces of green parks with only one or two walkways. Similarly, when we take an overall look into word-formation processes, we can see that some derivational morphemes, clustered together, display a larger number of connections to other morphemes, while others stand on their own. It turns out that these properties of morphemes are informative about the typological nature of word-formation processes in a language.

Although there is a rich body of literature on English word formation, their primary objectives were predominantly defined by theoretical premises and discussions. This study, however, is not driven by particular theoretical conjectures. Rather, it adopts a different strategy, which has not been implemented to date: instead of isolating specific categories of words—the approach taken by many studies in the field—the target of this research is a large sample of words, which is taken to represent the English lexicon. Then, this sample, a ‘snapshot’ of English words, has been analyzed as a whole with structural, descriptive and statistical tools in a quest to establish a general picture of English word formation using quantitative and qualitative characteristics. For this reason, the format of the current thesis may have seemed slightly unconventional. Theoretical generalizations about word-formation processes have been made purely from observations and analyses: some of them are what has been already established in the field (thus, they provide pieces

of evidence for hypotheses and discussions); others are novel and are not supported by existing literature. This distinctive feature of the current study comes with a number of advantages. First, the procedures of all performed analyses have been described step by step—from the individual instances of morphological patterns to general abstractions about English word formation—which makes this study consistent and transparent throughout. In fact, some of the morphological procedures are interesting in themselves and offer a new perspective on how we can treat various morphological units: for example, the analysis of constructions with the matrix optimization analysis leading to the compilation of matrices which are condensed representations of constructions. As a result, the whole English word-formation system is captured in a number of matrices that occupy only several pages. Such a perspective is particularly important for the field, because, in the present-day linguistic literature, the notions of 'construction', 'pattern' and 'slot' are widely used, but in a very abstract sense and, at times, not consistently. In this study, these abstract notions have been materialized in a concrete form and with uniform analyses. Thus, the concreteness of the description of the current study is valuable for projects whose aim is to create morphological models of languages. The second advantage of the empirical orientation of this thesis is that, with this approach, establishing new facts about English word formation has become possible, which have not been discussed in the corresponding morphological literature.

8.1 The main findings of this study

If we construct the English word-formation system on the basis of the compiled morphological metacorpus and with the established facts, a diverse picture emerges. In general, simplexes are more frequent in nouns, verbs, as well as grammatical and conversive classes, and multimorphemic words in adjectives and adverbs. Specifically, multimorphemic adjectives constitute 75% of all adjectives in the sample, and multimorphemic adverbs around 87% of all adverbs. The high proportion of derived words in these classes does not mean, however, that their derivation is more diverse, rather it suggests that, for the most part, adjectival and adverbial meanings in English are constructed morphologically. In contrast, the largest portion of verbs (nearly 90%) are simplexes, half of which are formed by conversion, largely from nouns—this established fact echoes with Plank's (2018) observation that, in languages with Germanic roots, "there is something verbal in many nouns, but not vice versa". Although there are other constraints (for example, etymological) that define which meanings are expressed by simplexes and which meanings by multimorphemic

words, the above-mentioned proportions suggest that there are more general tendencies—grammatical or cognitive—on the level of word classes, determining whether a word is morphologically simple or complex.

In typological theorizing of word formation, the question of 'basicness' and 'derivativeness' of words, as well the question of the predictability of these features, are important (Plank 2018). The comparison of the correlations of simplexes vs derived words across classes in English and those in Persian, for which similar quantitative characteristics have been collected (Krykoniuk 2014), reveals that they are different in the two languages: in Persian, there is a considerably larger number of simplexes in nouns (around 79%), a lower number of derivatives in adjectives (35%), and there are no monomorphemic verbs. However, similar to English, in adverbs, multimorphemic words are more frequent than simple (70%)—although Persian adverbial derivation is much richer than in English. Therefore, from these observations, it can be hypothesized that the ratios of basicness and derivativeness of nouns, adjectives and verbs are idiosyncratic in each language, but the derivativeness of adverbs tends to be greater.

A large number of simple nouns and adjectives are words borrowed from other languages (50% and 55% respectively), whereas borrowings are much less frequent in verbs and adverbs (24% and 7%). A similar picture emerges for Persian word classes (Krykoniuk 2014), with the difference that the number of loan words is higher in Persian nouns and adjectives (borrowed mainly from Arabic) and that there are no monomorphemic verbs in Persian. The fact that there are fewer borrowings in verbs and adverbs is evidence that these classes are more conservative and rely more on internal lexical and grammatical resources of language. Further, the structural analysis has shown that shortening (with the involvement of phonology) is more common in nouns, whereas back-formations are more common in verbs. Onomatopoeic formations are significant for nouns and verbs, and semantic word-formation processes are the least productive in simplexes across all word classes.

Another typological feature that has emerged from the performed structural analysis is the complexity of derivation: i.e. how many morphemes are involved in the formation of words. In English, nouns show the highest level of complexity (4 levels), followed by adjectives, nouns/adjectives, adverbs and adjectives/adverbs (3 levels), and verbs (2 levels). Hence, different word classes have different levels of complexity: noun formation is the most morphologically complex in English, and verb formation is the least complex (with the prefixation being the most

productive process on the second level of derivation). It is possible that the lower derivation complexity of verbs and the fact that half of the simple verbs (in the metacorpus) are formed by conversion are related. We can hypothesize that, in every morphologically-concatenative language, a word class which is largely formed by conversion on a zero-level of word formation would show less structural complexity: when the function of conversion is highly active in a word class, it tends to be less structurally complex.

Furthermore, verbs display another idiosyncratic feature: their median type valency in suffixation is 2 (whereas in other word classes it is 1). Possible explanations for this behaviour are that grammatically, a word-base slot in the construction of verb suffixation is almost equally open to two word bases (nouns and adjectives), or that verb suffixes have the property that allows them to more readily attach to these bases.

Conversive classes have been shown to combine the properties of the two single word classes that they merge. In particular, this is obvious from the 'division of labour' of their word bases: for example, noun bases are highly dominant in noun prefixation, and adjective bases in adjective prefixation, whereas in conversive noun/adjective prefixation, both word bases are powerful. For this reason, conversive classes are more morphologically diverse, which is also evident from a considerably smaller discrepancy between the number of word types and the number of constructions in these classes.

In addition, the current thesis has given a detailed account of all morphological constructions for major word classes (i.e. nouns, verbs, adjectives, nouns/adjectives and adjectives/adverbs). The internal structure of each construction can be perceived as a formal paradigm—a paradigmatic network of morphemes that occupy the slot of *a* (affix) or *C* (root). In this study, it has been demonstrated that the typological characteristics of morphological constructions differ across word classes. For example, the architecture of formal paradigms for suffixation in nouns, nouns/adjectives and adjectives/adverbs emerges as a fully-fledged network with several central nodes, surrounded by the rays of monovalent suffixes, but that of adjectives and verbs has a smaller number of rays around the nodes, which suggests that there are fewer monovalent suffixes in these word classes and that the connections between morphemes in their paradigmatic networks are tighter. In view of typology, this difference can be explained by various degrees of the expression of isolation and agglutination features: a high number of monovalent affixes in a word-formation process is an indication of derivational isolation, whereas a greater

number of connections within a network points to a higher degree of derivational agglutination. Further, the significance of a bound morpheme node in a word-formation process suggests a more pronounced expression of fusion. The graph-network analysis has revealed that the suffixation of verbs and the compounding of all studied word classes (except for adverbs) display a higher degree of fusion. Finally, as opposed to agglutination, morphological disintegration signifies a lack of cohesiveness between morphemes: in a graph, this feature surfaces as unconnected pieces of a network. The prefixation and compounding of the conversive class of adjectives/adverbs manifest morphological disintegration, which might have appeared due to the small number of items in this class or due the fact that the word-formation processes of this class are more semantically or syntactically driven.

A high frequency of certain word bases, recorded in the different analyses of the current thesis, suggests their grammatical significance for derivation. In nouns, for example, verb bases have the highest frequency. The derivational importance of verbs in noun formation is evidence for syntactic derivation, central to argument structure theories, which argue that verbs are "generally associated with a very rich conceptual meaning, including what is called semantic roles (Afarli 2007: 32). In this instance, high type frequency is a consequence of a grammatical cause. This study has also established that the type frequency of bases determines specific combinations of suffixes. A very strong and statistically-significant correlation between word bases in the first and higher level of derivation in nouns and adjectives provides evidence for the base-driven hypothesis of suffix combinations (Plag 1996). Finally, regression analysis of the type frequency of suffixes and their type valency has substantiated that the higher the type frequency of a suffix, the more likely it is that it will develop polyvalency. It has been hypothesized that the Unitary-Base Hypothesis (Aronoff 1976) holds true for the 'initial' conditions of word-formation processes.

Moreover, in this thesis, the new methods of relative entropy estimation (Zhang 2017)—KLD, symmetrized KLD and Turing's perspective—have been applied for the first time (in linguistics) to the study of the diachronic productivity of the most type-frequent word-formation processes of nouns, adjectives and verbs. For this purpose, an *R* package entitled 'kldtools' has been developed (it is now available as an open source in CRAN), which allows for the statistical comparison of frequencies in two distributions. The performed KLD analyses have demonstrated that the diachronic productivity of some word-formation processes is similar, suggesting more

universal derivational currents in the language. The years around 1300 and 1400 were marked by a lexical significance for English, when a large number of words were borrowed, and the years around 1600 and 1800 by a morphological significance, when a large number of words were derived. Whereas most derivational processes have been showing a declining productivity trend to the present day, the diachronic productivity of compounding has been increasing.

Finally, different cluster techniques have been applied to the combined morphological data collected from different sources, in order to establish groups of affixes with similar profiles. Whereas regression and KLD analyses allow for hypothesis testing—thus bringing more confidence to a linguistic description—cluster techniques hint at the main trends in the data and can be used for determining possible research directions. These analyses have substantiated three clusters of affixes on the basis of their proportions of type and token frequency. They also confirm claims made in the relevant literature that suffixes with high type and token frequency are less productive and that token frequency and potential productivity are negatively correlated. The biplots of the PCA analysis have also verified that the behaviour of the variable for the type-token ratio is similar to that of the potential productivity—this observation suggests a possibility of the prediction of the potential productivity of affixes with their type-token ratio. Another important observation concerns the origin of affixes: affixes of Old English and Germanic origin tend to be characterized by a higher type frequency, type-token ratio and potential productivity, whereas the suffixes of Latin and French origin are characterized by a higher token frequency and a lower potential productivity and type-token ratio.

8.2 The limitations of the current study and the potential for further research

There are some limitations to the methodology and interpretations of this study which we need to consider. First of all, the methodology focuses on the expression plane of language, excluding the content plane. Integrating semantics into the formal study of word formation will allow us to explore how morphological forms of language correlate with their meaning and what aspects of morphological regularities in word formation are driven by the content plane. A potential strand of research would involve, first, formal semantic analysis of the data and then the juxtaposition of the results of the formal morphological and formal semantic analyses. Moreover, the prosodic features of English word formation have been left out of the area of this research, which are known

to have an impact on the organization of morphemes in words. Future studies of word formation by means of formal morphology would also benefit from incorporating prosody.

In addition, the current thesis applies the methods of relative entropy to the study of diachronic productivity. Entropy measures are relatively new developments in the field of mathematics, and they have not been widely used in linguistics such that their properties and behaviours are fully understood. Since there are no benchmarks, against which the results of the KLD analyses performed in this research could be mapped, it is difficult to explain what the discrepancies of statistics between the KLD estimators mean in linguistic terms. The application of these estimators to the comparison of frequencies in different linguistic areas would help linguists clarify which estimator is more suitable for which types of distributions of linguistic units.

Lastly, morphological frequency and productivity data is limited to date. There are only a few sources that provide information about various quantitative characteristics of English morphemes, and their criteria for eliciting morphological data differ. For this reason, the cluster analyses of the current study have been performed with limited data, for which some of the values were generated with the help of imputation. Constructing the overall profiles of affixes with cluster analyses is one of the heuristic ways to understand their behavior and to formulate new hypotheses about morphological properties of word formation. Thus, another direction for future research in line with the framework of the current study is collecting more morphological data from different registers/corpora, in order to identify clusters of affixes with similar behavior and to establish how different parameters of the word-formation system interact with each other.

Notwithstanding, the current study has contributed to the field of morphology by constructing a general picture of English word formation, identifying its typological features and substantiating a few hypotheses. With its large amount of empirical findings, this research can be used as a reference source which can be consulted on different aspects of English word formation.

REFERENCES

Afarli, T.A. (2007). *Do verbs have argument structure?* In: Eric Reuland et al. (eds.), Argument Structure. John Benjamins.

Alok, M., Bradford, T. (2019). Applied Unsupervised Learning with R. Packt Publishing 1.

Anderson, S.R. (2015). *The morpheme: its nature and use.* In: Matthew Baerman (ed.), The Oxford Handbook of Inflection. Oxford University Press, pp. 11–33.

Arapov, M., Herz, M. (1973). Frequency and Age as Characteristics of a Word. *COLING*, Vol 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics, pp. 3–7.

Aronoff, M. (1976). Word formation in Generative Grammar. MIT Press.

Baayen, H. (1991). Quantitative aspects of morphological productivity. In Booij G., van Marle J. (eds.), *Yearbook of Morphology*. Springer, pp. 109−149.

Baayen, H. (1993). On frequency, transparency and productivity. In G. Boooij and J. van Marle J. (eds.), *Yearbook of morphology*. Kluwer Academic Publishers, pp. 181−208.

Baayen, H. (2001). Word frequency distributions. Kluwer Academic Publishers.

Baayen, H. (2003). *Probabilistic approaches to morphology*. In: Probabilistic linguistics. Cambridge, Massachusetts: The MIT Press, pp. 229−288.

Baayen, R. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge: Cambridge University Press.

Baayen, H. (2009). *Corpus Linguistics in Morphology: Morphological Productivity*. In Anke Lüdeling and Merja Kytö (eds.), Corpus Linguistics. An International Handbook. Mouton de Gruyter, pp. 899–919.

Baayen, H., Lieber, R. (1991). Productivity and English derivations: a corpus-based study. *Linguistics* 29, pp. 801−843.

Baayen, H., Lieber, R. (1997). Word frequency distribution and lexical semantics. *Computers and the Humanities*, 30, pp. 281–291.

Baghramian, M., Carter, J.A. (2017). Relativism. In: E.N. Zalta (eds.), *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/sum2017/entries/relativism/ [Accessed 27/04/2018].

Balota, D.A., Yap, M.J., Hutchison, K.A. et al. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, pp. 445–459.

Bauer, L. (1983). Word formation. Cambridge University Press.

Bauer, L. (2001). Morphological Productivity (Cambridge Studies in Linguistics). Cambridge University Press.

Bauer, L., Lieber, R., & Plag, I. (2013). Basic principles: methods. In: The Oxford Reference Guide to English Morphology. Oxford: Oxford University Press.

Bauer, L. (2019). Notions of paradigm and their value in word formation. *Word Structure* 12(2), pp. 153−175.

Berg, T. (2014). On the relationship between type and token frequency. *Journal of Quantitative Linguistics*, 2(3), pp. 199–222.

Bloomfield, L. (1935). Language. London: Allen & Unwin.

Bochner, H. (1993). Simplicity in Generative Morphology. De Gruyter Mouton.

Bolinger, D. (1948). On defining the morpheme. WORD, 4(1), pp.18–23.

Bondy, J.A., Murty, U.S.R. (1976). Graph Theory with applications. The Macmillian Press LTD.

Booij, G. (2010). Construction morphology. Oxford University Press.

Booij, G. (2015). *Word-formation in construction grammar*. In P. Müller, I. Ohnheiser, S. Olsen & F. Rainer (eds.), Word-Formation: An International Handbook of the Languages of Europe, Vol. 1, pp. 188−202.

Booij, G., Audring, J. (2017). Construction Morphology and the Parallel Architecture of Grammar. *Cognitive Science* 4, pp. 277–302.

Borer, H. (1994). The Projection of Arguments. In E. Benedicto and J. Runner (eds.) *University of Massachussetts Occasional Papers in Linguistics* 17. University of Massachusetts, Amherst.

Borer, H. (2003). Exo-Skeletal vs. Endo-Skeletal Explanations: Syntactic Projections and the Lexicon. In Moore, J. and Polinsky, M. (eds.). *The Nature of Explanation in Linguistic Theory*. Stanford: CSLI Publications, pp. 31−67.

Butts, C T. (2015) .network*: Classes for Relational Data*. R Package Version 1.13.0.

van Buuren S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), pp. 1-67.

Bybee, J. (1985). Morphology: A study of the relation between meaning and form. Amsterdam: Benjamins.

Bybee, J. (2001). Phonology and language use. Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language* 82, pp. 711–733.

Bybee, J. (2007a). Frequency of use and the organization of language. New York: Oxford University Press.

Bybee, J. (2007b). *Diachronic linguistics*. In: Dirk Geeraerts and Hubert Cuyckens (eds.), The Oxford handbook of cognitive linguistics. Oxford University Press, pp. 945−987.

Bybee, J. (2013). *Usage-based Theory and Exemplar Representations of Constructions*. In Thomas Hoffmann and Graeme Trousdale (eds.), The Oxford Handbook of Construction Grammar. Oxford University Press, pp. 49−69.

Carrier, J. (1979). Review of "Word Formation in Generative Grammar". In M. Aronoff (ed.), *Linguistic Inquiry Monograph*. Linguistic society of America, Vol. 55, No. 2, pp. 415−423.

Cetnarowska, B. (2001). Ingo Plag, Morphological productivity: Structural constraints in English derivation. *Journal of Linguistics*, 37(2). Mouton de Gruyter, pp. 451-462.

Chomsky, N. (1981). Lectures on government and binding. Foris.

Chomsky, N. (1982). Some concepts and consequences of the theory of government and binding. MIT Press.

Chomsky, N. (2002[1957]). Syntactic structures (2nd ed.). Mouton de Gruyter.

Chomsky, N. (2017). A question on morphology. [email].

Chomsky, N., Halle, M. (1968). The Sound Pattern of English. Harper and Row.

Chomsky, N. (1961). On the notion "Rule of grammar". Cambridge, Mass.: American Mathematical Society.

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs; P. Rosenbaum (eds.). *Readings in English transformational grammar*. Waltham (MA): Ginn, pp. 181–221; перепечатано в кн.: Chomsky, N. (1972). Studies on semantics in generative grammar. The Hague: Mouton, pp.11–61.

de Courtenay, Stankiewicz, E. (1972). A Baudouin anthology: The beginnings of structural linguistics. Indiana University Press.

Cover, T.M., Thomas, J.A. (2006). Elements of Information theory (2nd ed.). Wiley.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2), pp. 121–136.

Croft, W. (2007). *Construction Grammar*. In: Dirk Geeraerts and Hubert Cuyckens (eds.), The Oxford handbook of cognitive linguistics. Oxford University Press, pp. 463–508.

Csárdi, G., Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

Dąbrowska E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in psychology*, 6, 852.

Damasio, A. (2000). Descartes' error: Emotion, Reason and the Human Brain. Quill.

Demir-Lira, E., Applebaum, Ö., Goldin-Meadow, L. R. & Levine, S. C. (2019). Parents' early book reading to children: Relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental science, 22*(3), e12764.

Desagulier, G. (2017). Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics. Springer.

di Sciullo, A.,Williams, A. (1987). On the Definition of Word. Cambridge, Mass: MIT Press.

Dokulil, M. (1962). *Tvoření slov v češtině I. Teorie odvozování slov.* Prague: Nakladatelství ČAV.

Dokulil, M. (1994) The Prague School's theoretical and methodological contribution to 'word-formation' (derivology*). In Luelsdorff Ph. (ed.), *The Prague School of structural and functional linguistics: a short introduction*. J. Benjamins.

Don, J. (2014). Morphological theory and the morphology of English. Edinburgh University Press.

Fabb, N. (1988). English suffixation is constrained only by selectional restrictions. *Natural Language & Linguistic Theory*, 6(4), pp. 527–539.

Faraway, J. (2016). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition (2nd ed.). Chapman and Hall/CRC.

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. SAGE Publications.

Frawley, W.J. (2003). Generative morphology. In: *International encyclopedia of linguistics* (2nd ed.). Oxford University Press. Available at: http://www.oxfordreference.com/view/10.1093/acref/9780195139778.001.0001/acref-9780195139778-e-0402 [Accessed at 17/04/2018].

Gazdar, G., Klein, E., Pullum, G. K., & Sag, I. A. (1985). Generalized phrase structure grammar. Harvard University Press.

Gleick, J. (1987). Chaos: making a new science. London: Abacus.

Goldberg, A., Suttle, L. (2010). Construction Grammar. In *WIREs Cognitive Science*, vol. 1. John Wiley & Sons Ltd., pp. 468−477.

Greenberg, J. (1960[1954]). A quantitative approach to the morphological typology of language. *International Journal of Applied Linguistics* 26, pp. 178−194.

Grimshaw, J. (1990). Argument Structure. Cambridge, MA: MIT Press.

Hudson, R. (2007). *Word Grammar*. In: Dirk Geeraerts and Hubert Cuyckens (eds.), The Oxford handbook of cognitive linguistics. Oxford University Press, pp. 509–539.

Halle, M. (1973). Prolegomena to a theory of word-formation. *Linguistic Inquiry*, 4.1, pp. 3–16.

Halle, M., Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale, & S. J. Keyser (eds.), *The view from building* 20. The MIT Press, pp. 111−176.

Harley, H. (1995). Subject, Events and Licensing. Doctoral dissertation, MIT.

Harley, H. (2005). How do verbs get their names? Denominal verbs, manner incorporation and the ontology of verb roots in English. In Nomi Erteschik-Shir and Tova Rapoport (eds.), *The Syntax of Aspect*. Oxford University Press, pp. 42−64.

Haselow, A. (2011). Typological changes in the lexicon: analytical tendencies in English noun formation, Bernd Kortmann, Elizabeth Closs Trugott (eds.). De Gruyter Mouton.

Haspelmath, M., Sims, A. D. (2010). Understanding morphology. London: Hodder Education.

Hastie, T., Tibshirani, R. Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hay, J. (2000). Causes and consequence of word structure. PhD thesis, Northwestern University.

Hay, J. (n.d.). Causes and consequences of word structure. Dissertation Summary, Glot International.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics 39 (6)*, 1041−1070.

Hay, J. (2002). From speech perception to morphology: Affix-ordering revisited. In *Language* 78(3), pp. 527−555.

Hay, J. (2003). Causes and consequences of word structure. New York & London: Routledge.

Hay, J., Baayen, H. (2002). Parsing and productivity. In Booij and van Marle (eds.), *Morphology*, pp. 203−235.

Hay, J., Plag, I. (2004). What constraints possible suffix combination? On the interaction of grammatical and processing restrictions in derivational morphology. In *Natural Language and Linguistic Theory* 22, pp. 565−596.

Herdan, G. (1964). *Quantitative Linguistics*. Buttersworths.

Hilpert, M. (2014). Construction Grammar and its Application in English. Edinburgh University Press.

Hilpert, M. What is Construction Grammar? (video) Available at: https://www.youtube.com/watch?v=9DllnszLuM0&list=PLKgdsSsfw-fZyiK6ahhdg4N3n4NrpdgWk&index=3. [Accessed on 12/03/20].

Hjelmslev, L. (1939). La Structure Morphologique (Types de Système). Rapports V Congrès Intern. des Linguistes, pp. 66−93.

Hjelmslev, L. (1961). Prolegomena to a theory of language. The University of Wisconsin Press.

Hockett, Ch. (1958). A Course in Modern Linguistics. Oxford and INH Publishing Co.

Hooper, P. (1987). Emergent grammar. *BLS*, 13, pp. 139–157.

Horecký, J. (1983). Vývin a teória jazyka. Bratislava: SPN.

Howell, D. C. (2007). Statistical methods for psychology (6th ed.). Belmont, California: Duxbury Press.

Jackendoff, R. (1975). Morphological and Semantic Regularities in the Lexicon. In *Language*, Vol. 51, No. 3, pp. 639−671.

Jackendoff, R. (1983). Semantics and Cognition. Cambridge, MA: MIT Press.

Jackendoff, R. (1996). The proper treatment of measuring out, telicity, and perhaps even quantification in English. *Natural Language and Linguistic Theory* 14, pp. 305−354.

Jespersen, O. (1992[1924]). The Philosophy of Grammar. The University of Chicago Press.

Joshi, A. (2002). Hierarchical structure and sentence description. In *The legacy of Zellig Harris: Language and Information into the 21st century*, Vol. 2, Mathematics and computability of language, Bruce E. Nevin and Stephen B. Johnson (eds.). John Benjamins, pp. 121−143.

Kastovsky, D. (2005). Hand Marchand and the Marchnadeas. In: P. Štekauer and R. Liebers (eds.), *Handbook of Word Formation*. Springer, pp. 99−124.

Kastovsky, D. (2006). *Vocabulary*. In R. Hogg & D. Denison (eds.), A History of the English Language. Cambridge University Press, pp. 199−270.

Kemmer, S. (2003). *Schemas and lexical blends*. In: Hubert Cuyckens, Thomas Berg. René Dirvern, and Klaus-Uwe Panther (eds.), Motivation in Language: Studies in honor of Günter Radden. John Banjamins, pp. 69−97.

Kenstowicz, M. (1994). Phonology in Generative Grammar. Blackwell Publishers.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), pp. 263–276.

Kiparsky, P. (1982). Lexical phonology and morphology. MS, MIT.

Kiparsky, P. (1985). Some consequences of lexical phonology. In *Phonology* 2 (1), pp. 85−138.

Kratzer, A. (1994). *The Event Argument*. Unpublished manuscript, University of Massachusetts, Amherst.

Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing and linguistic change. *Journal of English Linguistics*, 26(4), pp. 286–320.

Krykoniuk, K. (2020). Predictive modelling of type valency in word formation grammar. *Journal of Quantitative Linguistics*. DOI: 10.1080/09296174.2020.1782720

Krykoniuk, K., Shipunov, A., Sekhon, J. (2021). kldtools: Kullback-Leibler Divergence and other tools to analyze frequencies. https://cran.r-project.org/web/packages/kldtools/index.html.

Kullback, S., Leibler, R.A. (1951). On information and sufficiency. In *Ann. Math. Stat*., 22, pp. 79–86.

Lai, P., Fu, H. (2011). Variance enhanced K-medoid clustering. *Expert Systems with Applications*, Vol. 38 (1), pp. 764–775.

Langacker, R. (2007). Cognitive Grammar. In: Dirk Geeraerts and Hubert Cuyckens (eds.), The Oxford handbook of cognitive linguistics. Oxford University Press, pp. 421−462.

Laws, J.V., Ryder, C. (2014). MorphoQuantics: http://morphoquantics.co.uk.

Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*. 25(1). pp. 1–18.

Legendre, P., Legendre, P. (2012). Numerical ecology. Amsterdam: Elsevier.

Leisch, F. (2006). A toolbox for k-centroids cluster analysis. *Computational Statistics and Data Analysis*, 51 (2), pp. 526-544.

Lerdahl, F., Jackendoff; F. (1983). An Overview of Hierarchical Structure. In *Music. Music Perception* 1 (2), pp. 229–252.

Levin, B., Rappaport-Hovav, M. (1988). Non-event *-er* nominals: a probe into argument structure. *Linguistics* 26: pp. 1067−83.

Levin, B., Rappaport-Hovav, M. (1995). Unaccusativity. Cambridge, MA: MIT Press.

Levin, B., Rappaport Hovav, M. (2005). Argument Realization (Research Surveys in Linguistics). Cambridge University Press.

Levine, T., René, W., Hullett, C., Park, H., & Massi, L. (2008). A Critical Assessment of Null Hypothesis: Significance Testing in Quantitative Communication Research. *Human Communication Research* 34, pp. 171−187.

Lieber, R. (1992). Deconstructing Morphology. The University of Chicago Press.

Lieber, R. (2004). Morphology and lexical semantics. Cambridge University Press.

Lieber, R. (2009). Introducing Morphology. Cambridge University Press.

Lieber, R. and Baayen, H. (1999). *Nominalizations in a calculus of lexical semantic representations*. In Geert Booij and Jaap van Marle, eds., Yearbook of Morphology 1998 . Dordrecht: Kluwer Academic Publishers. pp. 175–98.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), pp. 319−344.

Luraghi, S. (2005). *Hjelmslev, Lois*. In: Encyclopedia of linguistics / Philipp Strazny, ed. New York: Fitzroy Dearborn, pp. 470−472.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2013). cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.

Malmkjær, K. (ed.) (2010). The Linguistics Encyclopedia (3rd ed.). Routledge.

Mandelbrot, B. (1962). On the theory of word frequencies and on related Markovian models of discourse. In R. Jacobson (ed.), *Structure of Language and its Mathematical Aspects*, American Mathematical Society, pp. 190−219.

Manova, S. (2015). *Closing suffixes*. In: Peter O. Müller, Ingenbourg Ohnheiser, Susan Olsen, Franz Rainer (eds.), Word-formation: An International Handbook of the Languages of Europe, Vol. 2. De Gruyter, pp. 956−971.

Manova, S., Aronoff, M. (2010). Modelling affix order. In *Morphology* 20 (1), pp. 109−131.

Marantz, A. (1996). Cat as a Phrasal Category. Unpublished manuscript, MIT.

Marantz, A. (1997). No Escape from Syntax. *University of Pennsylvania Working Papers in Linguistics* 4.2, pp. 201−225.

Marantz, A. (2013). No escape from morphemes in morphological processing. In *Language and Cognitive Processes Language and Cognitive Processes* 28(7), pp. 905−916.

Marchand, H. (1951). Phonology, morphonology, and word-formation. *Neuphilologische Mitteilungen* 52, pp. 87−95.

Marchand, H. (1960). The categories and types of present day English word formation: a synchronic-diachronic approach. Wiesbaden: Otto Harrassowitz.

van Marle J. (1992). The relationship between morphological productivity and frequency: A comment on Baayen's performance- oriented conception of morphological productivity. In: Booij G., van Marle J. (eds.), *Yearbook of Morphology 1991*. Springer.

Mateu, J. (2014). *Argument structure.* In The Routledge Handbook of Syntax. Routledge. [Accessed on: 22/04/2019].

Mathesius, V. (1942) O soustavném rozboru gramatickém. [About systematic grammatical analysis.] *Slovo a slovesnost*, pp. 88−92.

Mathesius, V. (1947). *Čeština a obecný jazykozpyt*. [Czech and general linguistics]. Melantrich.

Matthews, P. H. (2014). The Concise Oxford Dictionary of Linguistics (3rd ed.), online version. [Accessed 20/04/2019].

McArthur, T., Lam-Mcarthur, J., & Fontaine, L. (eds.). (2018). *Construction.* In the Oxford companion to the English language (2nd ed.). Oxford University Press.

Meeker, W. Q., Hahn, G. J., and Escobar, L. A. (2017). Statistical intervals: A guide for practitioners and researchers. John Wiley.

OED Online, Oxford University Press, www.oed.com.

Oliveira, M. R. G. de, Cruz , D. V. da, & Cunha Filho, M. (2020). Non-hierarchical grouping: 'K-mean' and 'K-medoid' of plaques cisterns in the Pajeu region - PE. *Acta Scientiarum. Technology*, 42(1), e44378.

Park, Ch. (2017). Rochelle Lieber: English Nouns: The Ecology of Nominalization. *Cognitive Linguistics*, 28(4), pp. 799−805.

Pesteyovsky, J. (1995). The generative lexicon. Cambridge, MA: MIT.

Plag, I. (1996). Selectional restrictions in English suffixation revisited: a reply to Fabb (1988). *Linguistics* 34(4), pp. 769−798.

Plag, I. (2003). Word Formation in English. Cambridge University Press.

Plag, I., Baayen, H. (2009). Suffix ordering and morphological processing. *Language* 85(1), pp. 106−140.

Plank, F. (1981). Morphologische (Ir-)Regularitäten. Tübingen: Narr.

Plank, F. (2018). Direction of derivation: How predictable? [PowerPoint slides]. Somerville College, University of Oxford.

Pocket Oxford German Dictionary: English German, 4 ed. (2009). Oxford University Press.

Pollard, C., Sag, I.A. (1994). Head-driven phrase structure grammar. University of Chicago Press.

Popescu, I. (2009). Word Frequency Studies: Word Frequency Studies. De Gruyter.

del Prado Martín, F. M., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, 94(1), pp. 1–18.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, G. (1985). A Comprehensive Grammar of the English Language. London: Pearson Longman.

Reinhart, T. (1996). Syntactic Effects on Lexical Operations: Reflexives and Unaccusatives. *UiL OTS Working Papers*. Utrecht University.

Russell, B. (1940). An inquiry into meaning and truth. London: Allen & Unwin.

Salkind, D., Neil. J. (2006). Encyclopedia of Measurement and Statistics. Thousand Oaks: SAGE Publications.

Sánchez-Gutiérrez, C.H., Mailhot, H., Deacon, S.H. et al. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behav Res* 50, pp. 1568–1580.

Sapir, E. (1921). Language. New York: Harcourt, Brace and Co.

Scalise, S., Guevara, E. (2005). *The lexicalist approach to word formation and the notion of the lexicon*. In: P. Štekauer and R. Liebers (eds.), Handbook of Word Formation. Springer, pp. 147−188.

Schuchardt, H. (1985). Über die Lautgesetze: Geden die Junggrammatiker. Berli: Robert Oppenheimer.

Schützenberger, M.P. (1961). A remark on finite transducers. *Information and Control* 4, pp.185–196.

Searle, S. (1988). Parallel Lines in Residual Plots. *The American Statistician*, 42:3, 211. DOI: 10.1080/00031305.1988.10475569

Selkirk, E. (1982). The Syntax of Words. Cambridge, Mass.: MIT Press.

Sherman, D. (1975). Noun-verb stress alternation: an example of the lexical diffusion of sound change in English. *Linguistics* 13(159), pp. 43−72.

Siegel, D. (1974). Topics in English morphology. MIT dissertation.

Siertsema, B. (1955). A Study of Glossematics: Critical Survey of its Fundamental Concepts. Netherland: Springer.

Silverman, S., Ratner, N.B. (2002). Measuring lexical diversity in children who stutter: Application of vocd. *Journal of Fluency Disorders*, 27(4), pp. 289−304.

Shipunov, A., Murrell, P., D'Orazio, M., Turner, S., Altshuler, E., Rau, R. Beck, M., Gibb, S.. Qiu, W., Paradis, E., Koenker, R. and R Core Team (2021). shipunov: Miscellaneous Functions from Alexey Shipunov. R package version 1.15.

Schneider, T. (2013). Information theory primer with an appendix on logarithms (PDF version). 10.13140/2.1.2607.2000.

Sokal, R.R., Rohlf, F.J. (1969). *Biometry.* Freeman and Co.

Spencer, A. (2005). *Word formation and syntax*. In: P. Štekauer and R. Liebers (eds.), Handbook of word formation. Springer, pp.73−97.

Spencer, A. and Zwicky, A. (eds.) (1998). The Handbook of Morphology. Blackwell Publishers Ltd.

Štekauer, P. (1998). An onomasiological theory of English word formation. John Benjamins.

Štekauer, P. (2000). English word-formation: a history of research (1960-1995). Tübingen: Narr.

Štekauer, P. (2005). *Onomasiological approach to word formation*. In: P. Štekauer and R. Liebers (eds.), Handbook of Word Formation. Springer, pp. 207−232.

Švarný O. (1997). Úvod do studia hovorové čínštiny [An Introduction to the Chinese Language].− Olomouc:Vydavatelství Univerzity Palackého.

Szymanek, B. (1988). Categories and categorization in morphology. Lublin: Catholic University Press.

Taeger, D. and Kuhnt, S. (eds). (2014). Statistical hypothesis testing. In *Statistical Hypothesis Testing with SAS and R*. Wiley Publishing.

Ter Braak, C. (1994). Canonical community ordination. Part I: Basic theory and linear methods. *Écoscience*, 1(2), pp. 127-140.

Tesnière, L. (1959). Eléments de syntaxe structurale. Paris: Klincksieck.

Taylor, J.R. (2002). Cognitive grammar. Oxford University Press.

Tottie, G. (1991). Lexical diffusion in syntactic change: Freaquency as a determinant of linguistic conservatism in the development of negation in English. *Historical English Syntax*, ed. Dieter Kastovsky. Mouton de Gruyter, pp. 439−167.

Trask, R.L., Stockwell, P. (2007). Key Concepts in Language and Linguistics (2nd ed). Routledge.

Trubetzkoy, N.S. (2010). Grundzüge der Phonologie: Sprachwissenschaft: Ein Reader L. Hoffmann (ed.). De Gruyter. pp. 388−404.

Tyshchenko K. (2003). La morfologia formale del verbo italiano // Lingua e letteratura italiana dentro e fuori la penisola. Atti del III Congresso degli Italianisti Europei, Cracovia, pp. 521−527.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Waugh, L., R. (2005). *Jacobson, Roman*. In: Philipp Strazny (ed.), Encyclopedia of linguistics. Fitzroy Dearborn, pp. 549−551.

Wierzbicka, A. (1985). Lexicography and Conceptual Analysis. Ann Arbor: Karoma.

Williams, E. (1981). Argument Structure and morphology. *The Linguistic Review* 1, pp. 81−114.

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary-management profile. In: *Language* 67, pp. 763−789.

Yule, G.U. (1944). The statistical study of literary vocabulary. Cambridge University Press.

Zhang, Zh. & Grabchak, M. (2014). Nonparametric estimation of Kullback-Leibler divergence. Neural computation 26 (11), pp. 2570–2593.

Zhang, Zh. (2017). Statistical Implications of Turing's Formula. John Wiley.

Zipf, G. (1935). The Psychobiology of Language: An Introduction to Dynamic Philology. Cambridge, Mass.: M.I.T. Press.

Апресян Ю.Д. Идеи и методы современной структурной лингвистики. М.: Наука, 1966.

Братчиков И.Л., Фитиалов С.Я., Цейтин Г.С.О структуре словаря и кодировке информации для машинного перевода.− Л.: Изд-во ЛГУ, 1958. С.61−87.

Булыгина Т.В. Проблемы теории морфологических моделей. М.:Наука, 1977.

Гухман М.М., Ярцева В.Н. (отв.ред.). Исследования по общей теории грамматики. − М., 1968.

Ельмслев Л. Пролегомены к теории языка. / Пер. с англ. / Сост. В.Д. Мазо. − М.: КомКнига, 2006.

Иванова, Л.П. Курс лекций по общему языкознанию. – Киев: ООО «Освита Украины», 2006.

Кацнельсон С.Д. О грамматической категории // Вестник ЛГУ. 1948. N 2.

Криконюк К.М. Формальна мофологія питомих засобів сучасної перської мови: дис. ... канд. філ. наук. – Київ, 2014.

Кубрякова Е.С. Основы морфологического анализа (на материале германских языков). – М.: Изд-во «Наука», 1974.

Курилович, Е. (1962). Понятие изоморфизма. Очерки по лингвистике /под общ. ред. В.А. Звегинцева. М.: Изд-во иностранной литературы, с. 21-36 (translated from: Kyryłowicz, J. (1949). La notion de i'somorphisme. Recherches Structurales dediées à L. Hjelmslev, pp. 48-60, TCLC, 5.).

Мельчук И.А. Курс общей морфологии. – Москва-Вена: Издательская група «Прогресс», 1997.

Мурат В.П. Глоссематическая теория // Пролегомены к теории языка / Л. Ельмслев / Пер. с англ. / Сост. В. Д. Мазо. − М.: КомКнига, 2006.

Плунгян В.А. Общая морфология: Введение в проблематику: Учебное пособие. Изд. 2-е, исправленное. — М.: Едиториал УРСС, 2003.

Ревзин И.И. Современная структурная лингвистика: Проблемы и методы. – М.: Изд-во «Наука», 1977.

Сергиевский М.В. Современные грамматические теории в Западной Европе и античная грамматика // Учен. зап. I МГПИИЯ. Т.2.1940. С.5−40.

Тищенко К.М. Основи мовознавства: Системний підручник. – Київ: ВПЦ «Київський університет», 2007.

Тищенко К.Н. Глагольная парадигма романских языков: автореф. дис. на здобуття вченого ступеня канд. філ. наук. − К.: КГУ, 1969.

Успенский Л. Слово о словах. Почему не иначе? – Л.: «Детская литература», 1971.