QUANTITATIVE ASPECTS OF THE WORD CLASS CONTINUUM IN ENGLISH

Dissertation zur Erlangung eines

Doktors der Philosophie

am Fachbereich Philosophie und Geisteswissenschaften

der Freien Universität Berlin

vorgelegt von

Alexander Rauhut

Berlin 2023

Gutachten:

Erstgutachter:	Prof. Dr. Anatol Stefanowitsch
Zweitgutachter:	Prof. Dr. Matthias Hüning

Tag der Disputation: 25.05.2023

Contents

List of Figures

List of Tables

List of Abbreviations

ALD Average Logarithmic Distance. 38, 40, 46, 78

BNC British National Corpus. 43, 47–49, 53, 66, 80, 81, 99

BNC2014 British National Corpus 2014 (spoken). 47

CA Correspondence Analysis. 5, 66, 68, 86, 87, 91, 103, 106

CL Cognitive Linguistics. 1, 2, 4, 7, 10, 13, 14, 23–25, 27, 115, 116

COCA Corpus of Contemporary American English. 47, 75, 99

CWB Corpus Workbench. 36, 134, 135

CxG Construction Grammar. 5, 12, 22, 23, 114, 116

DP Deviation of Proportion. 40

$\mathbf{DP}_{\mathbf{norm}}$ Normalized Deviation of Proportion. 40, 72

DWG Word Growth Dispersion. 40, 41, 53, 63

GAM Generalized Additive Model. 63

GAMLSS Generalized Additive Models for Location, Scale and Shape. 46

GLMM Generalized Linear Mixed Effects Model. 46

KDE Kernel Density Estimation. 5, 46, 68

KLD Kullback-Leibler Divergence. 40, 41, 53, 54

LNRE Large Number of Rare Events. 44

NER Named Entity Recognition. 45, 100

NLP Natural Language Processing. 116

PoS Part-of-Speech. 2, 5, 43, 53, 82, 100, 107, 114

RCG Radical Construction Grammar. 22, 23, 96, 114

U_r Kromer's U_r. v, 38, 46, 56, 57, 60–63, 66, 78, 84, 85, 87, 89, 102, 104

UD Universal Dependencies. 47, 53, 100, 138

1. Introduction

1.1. Lexemes and lexical classes

This thesis is concerned with the structure of English word classes, and the quantitative distribution of lexemes over the feature space that defines them. The categorization of words is one of the most fundamental tasks in linguistics. Every naïve description of language starts with words, and the realization that there are at least two fundamentally different types—one for describing actions or events and one describing people or objects. The English language has a very flexible word class system. It is not uncommon to find words switching between nouns, verbs and adjectives. Furthermore, not every concept encoded in a lexeme is equally typical for a discrete object or a dynamic event. This variation was discovered to have a significant impact on grammar (Ross 1972; Ross 1973a). That grammatical structure alone does not account for this, which shows that word classes have a fundamentally semantic nature. Yet the properties are mostly defined in terms of morphological and syntactic properties. The fundamental intuition that there are nouns to refer to discrete objects and verbs to describe actions and events seems suspended in favor of form. Cognitive Linguistics (CL) (e.g., Givón 1979; Langacker 1987a; Langacker 1987b; Lakoff 1987a; Hopper & Thompson 1985; Bybee 2010) attempts to solve this discrepancy by assuming a symbolic relationship between form and function on every level of abstraction, including grammar. It also provides a framework for uncovering structure in the lexicon that is motivated by meaning and not immediately obvious from purely formal properties. The search for semantic properties within linguistic structure is central to the discussion of words and word classes.

A verb-noun distinction is one of the few typological universals that has endured linguistic debate so consistently (cf. Givón 1979; Hopper & Thompson 1984; Langacker 1998; Croft 2001; Schachter & Shopen 2007). Also within English linguistics, this distinction usualy seems clear-cut. Under scrutiny, the definition of word classes, however, becomes difficult. Early on, researchers have realized that not all nouns are distributed equally (Ross 1972; Ross 1973a), or that gerunds occupy a problematic place among nouns and verbs for sharing aspects of both nouns and verbs (Quirk 1965). It is also easy to find so-called exceptions in the definition of word classes. Stative verbs do not normally occur in the progressive construction, mass nouns are not inflected, neither are non-gradable adjectives. These are usually considered subcategories of the respective word class that they are otherwise most similar to. Subcategorization,

however, does not explain all idiosyncrasies. Some scholars have questioned the usefulness of the traditional concept of word classes altogether, especially from the field of linguistic typology (most notably Croft 2022; also see Croft 2001), especially on the grounds of their lack of homogeneity.

Grammar and the lexicon are indivisible (cf. Langacker 1998: 32; Janda 2006: 7). Not only the morphological and syntactic properties of a word define its function and its class. Lexical patterns play a crucial role, as well (cf. Firth 1957; Stefanowitsch & Gries 2003; Goldberg 2006). Conversely, the members of a mostly grammatically defined category—e.g., mass nouns, stative verbs, non-gradable adjectives—provide the lexical substance of the (sub)category. It is sometimes assumed that the organization of grammatical categories, and their lexical structure are evidence of iconic relationships between grammar and language use (Hopper & Thompson 1985). This iconicity together with the common cognitive facilities of language users may explain why nouns and verbs are universals. The noun-verb distinctions and other patterns across languages and within languages have to be understood as tendencies rather than clear-cut rules to match the evidence. Even the strongest grammatical rule is better described as strong statistical tendency (cf. Stefanowitsch 2006: 70). A probabilistic approach to word classes is a direct consequence of the discovery of word class gradience.

Word classes in CL are assumed to be prototype categories. They are formed around a prototype, which is a schematic representation of the core of the category (Hopper & Thompson 1985; Langacker 1987c). Known members of the category and new experiences are categorized in relation to this prototype, based on similarity. This entails that there are more or less similar members of a category. Word classes are therefore continuous categories that are defined by a multidimensional feature space, rather than a fixed set of discrete features. Gradience is also extensively described in historical linguistics, especially the in conjunction with the study of grammaticalization (cf. Hopper & Traugott 2003). Historical change is inherently gradual and grammatical markers regularly develop from lexical items. The change from free morpheme to a clitic to affix is gradual, and the associated lexemes display less and less of their original category distinctions. This type of gradience is one of the main sources for word class gradience. The difference between prototype-based gradience (clusters) and continuous variation (clines) is derivative of theoretical work in CL, but rarely given explicit focus, much less investigated directly in Corpus Linguistics. The main goal so far has been to show that continua exist, and to accumulate evidence from all disciplines. Going beyond that, some theoretical groundwork has been laid to formulate different types of gradience (cf. Aarts 2007).

In Corpus Linguistics, word classes are, more often than not, the starting point of investigation, and Part-of-Speech (PoS) tagging is a common first step in the annotation of data. Ironically, this imposes a discrete conception of word classes on the data that is then used to investigate continuous phenomena. For more fine-grained approaches to word classes, data is often hand-coded for abstract properties, usually also a discrete endeavor. The more abstract a posited schema, the harder it becomes to test

empirically. With increasing abstractness of categories and schemata, the explanatory power of linguistic models may diminish. A corpus approach to word classes, therefore, has to be balanced between theory-driven and data-driven techniques.

1.2. How much homonymy?

The concept of homonymy is central to the discussion of English word classes. Polysemy is related to homonymy and likely a different side of the same coin. With the inflectional system reduced to only a few grammatical morphemes, word classes are mostly distinguished by syntactic and arguably semantic properties. English lexemes occur in their bare form most of the time. This makes conversion a common phenomenon in English. There are many pairs of lexicalized nouns and verbs that share the same form (*sleep*, *run*, *drink*). Additionally, the remaining morphology is often ambiguous, e.g., the *-ing* suffix. An *-s* suffix serves as the regular plural suffix for English nouns (*two cats*), but it is identical in form to the possessive marker for both singular and plural forms (e.g., *the cat's toy*, *children's toys*), which exhibits properties of a clitic. Given the typical structure of English noun phrases, however, it is, more often than not, formally indistinguishable from the plural suffix. Additionally, the *-s* suffix can serve for the third-person singular present tense of regular verbs (e.g., *he walks*). The case of the *-s* suffix in these different functions are rather clear examples of morphological homonymy.

Similarly, the *-ing* form is used to create present participles, which are used to form the progressive aspect in English verbs (e.g., *I am eating*) and gerunds, which function as nouns (e.g., *Eating is my favorite hobby*). However, the *-ing* form also serves to form deverbal nouns and deverbal adjectives, such as *building* and *boring*. This particular example of syncretism is of special interest since it creates ambiguity all across the traditional word class categories. In particular, the distinction between gerunds and participles has been a matter of intense debate. It is not as easy to argue for homonymy in this case since the functional differences between the categories of participle and gerund are much more subtle. What makes matters worse is that the terminology itself stems from the description of Latin where it is referring to different phenomena.

Turning back to the *-s* suffix, there can be more subtle cases of homonymy hidden. Typologically, the singular-plural distinction is not the only functional distinction that is made in terms of grammatical number. In fact, English itself used to have a dual. Describing paired objects is a function for a grammatical system to cover; however, an opposition to other number markers is not always present in a language. In part due to the special status of paired objects, English nouns that refer to *scissors* and *trousers* display grammatical irregularities. Plurale-tantum nouns, such as *clothes*, *trousers* and *scissors*, do not occur in singular form. Still, it would be unusual to assume a homonymous plurale-tantum suffix. The suffix on *scissors*, *proceedings*, is likely the same as the one on *doors* and *arms*, contributing the same meaning. There is little

systematicity in the encoding of paired objects, and the status of pluralia tantum is unclear.

In the description of English lexemes, the question of how many separate forms there are is a central issue. If there is just one *-s* suffix covering all types of plurality, is there also just one *-ing* form covering gerunds, participles, and perhaps derivation? With the introduction of multi-level generalizations (Langacker 1990; Goldberg 2006), the question of morphological homonymy, and by extension lexical classes, becomes more complex. Are all the different uses of a word—constructional uses, uses in idioms, uses in syntactic structures, uses in specific discourse situation—polysemous or even homonymous? There are two major challenges resulting from this: first, the inventory of classes may get arbitrarily large, and second, not all structure that can be found in empirical data is equally meaningful. Especially given enough corpus data, spurious correlations with coded variables will become significant (cf. Schmid & Küchenhoff 2013: 540; also see Sönning & Werner 2021 on problems concerning statistical significance). After a period of data-driven research in Usage-based linguistics, the theoretical and methodological landscape may require a more rigorous approach to the basic idea of word classes based on empirical data. Likewise, empirical evidence on homonymy and polysemy is required for CL concepts to evolve (Stefanowitsch 2011a: 302ff).

1.3. Research questions

This thesis is concerned with the corpus linguistic description of word class gradience and implications for categories of problematic status in English, such as the gerund-participle and plurale tantum. The main research question is whether prototype clusters are detectable in corpus data and how they align with traditional word class categories. Furthermore, it is investigated whether prototype clusters can be distinguished from continuous clines that merely show a functional correlation within a category, but do not exhibit multi-modal distributions in the statistical sense. It is assumed that an even spread of lexemes across a feature continuum indicates the lack of an underlying prototype category.

Central to the corpus studies is the empirical search for an 'uncanny valley' that is necessary for prototype theory, i.e., a statistical gap between lexical categories that shows (a) that categories are maximally distinct with regard to the feature under investigation, and (b) that lexemes with in-between distributions are indirectly 'avoided'. A case study on adjectives and their more uncontroversial subclasses will serve as a benchmark for the investigation of more complicated cases, such as pluralia tantum and the gerund-participle. The techniques and analyses are mostly exploratory, since the issue of type distributions is not well-researched in corpus linguistic. The object of investigation is the type distributions themselves. What makes plurale-tantum nouns particularly interesting is the fact that absence is not the same as presence. The question is whether regular absence can be explained with the same mechanisms as

regular presence. Since absence and a morphologically bare form is the default, it is likely that negative preemption of missing inflection, derivation and co-occurrence is inherently more common than the negative preemption of simpler forms compared to complex forms.

A related question with regard to the lexicon (or 'constructicon' in Construction Grammar (CxG)) is whether the few English affixes cover many functions or whether there is a merger of function leading to one very abstract function and a network of closely related senses. If this is the case, a mono-modal distribution is expected, i.e., a single cluster of lexemes without a discernible subdivision. In the case of conceptually motivated categorical distinctions, multiple distributions of lexemes are expected.

The topics of word class and gradience are also of practical importance in Corpus Linguistics. Word class membership with PoS tags as indicator is commonly used as a sampling criterion. Annotation that is aimed at word class distinctions are inherently a means of disambiguation. The discrete decisions taken in the process of automatic annotation crucially depend on the validity of the classes that are annotated. Therefore, a secondary question pursued in this thesis is whether statistical techniques, such as Kernel Density Estimation (KDE) and Correspondence Analysis (CA) can be used to detect overlapping distributions in the data.

1.4. Outline

The theoretical discussion in Chapter 2 begins with a brief introduction to the cognitive basics of experience-based linguistic categories in Section 2.2. Special focus is given to the question of what constitutes the basic unit of analysis in the discussion of word classes. For this purpose, it will be established that grammatical categories and lexical items form a continuum that is defined by degrees of complexity and schematicity. Usage-based approaches to morphology are then discussed to establish a link to the phenomena explored in the corpus studies that are mostly morphological in nature. Section 2.3 is concerned with the concept of gradience and its implications for the description of lexical categories. The theoretical state of the art is reviewed, and approaches to the category space along which word classes vary are contrasted. The theoretical implications of Prototype Theory for the analytical category space are discussed and first constraints derived. Finally, Section 2.4 tries to solve the issue of how formal equivalence, i.e., polysemy and homonymy, can be consolidated with the idea of lexeme-based word classes.

Chapter 3 advances the discussion of the previous chapter with special focus on word class categories. Both traditional and more recent cross-linguistic conceptions are discussed in Section 3.2 with the aim of getting a full view of the category space in which the English word classes vary. A concept of different types of gradience is developed to differentiate prototype categories from other descriptive categories. Section 3.3

turns to English word classes, and the language-specific issues in the classification of sub categories. Different models of the English noun-adjective-verb continuum are contrasted, and further variables of variation are derived from the theoretical discussion.

The concepts discussed in the previous chapters are operationalized in Chapter 4. A variety of statistical measures are reviewed and discussed, including association, dispersion and productivity. The statistical techniques that make up the methodology are introduced in Section 4.3. It is established how type densities in combination with the aforementioned measures can be used to identify and explore prototype clusters. Finally, Section 4.4 describes the data sets and annotations used and assesses their challenges and limitations. The methodology of this thesis is designed in accordance with the principles summarized in Sönning (2021: table 1).

The studies in Chapters 5-6 are strongly theory-driven and heavily informed by the experimental and cross-linguistic literature. The objects of investigation are classics in English linguistics. The aim of the analyses is to replicate classic findings (e.g., Ross 1972; Ross 1973a), bring them into a corpus linguistic context, and refine the ideas based on converging evidence from other areas of linguistics. The first case study, in Chapter 5 investigates descriptive subcategories of adjectives and their categorical status. The second case study in Chapter 6 turns to the English pluralia tantum and compares them to more clearly established subcategories, such as mass nouns. Finally, Chapter 7 is devoted to the question of how many *-ing* forms can be described by applying the established methodology.

Chapter 8, provides a summary of the results and discusses methodological and theoretical implications from the case study. An outlook for further research is provided.

I am following the open science principle. All annotated data sets will be made available[1] as part of a software package alongside the code [2] used to create the results and this manuscript. To ensure the reproducibility of the results, *Docker* images will also be provided.

[1] with the exception of copyrighted full text corpus data
[2] only free and open source software was used in the creation of this thesis, see A.2

2. Cognitive Foundations

2.1. Overview

In the following chapter, I will outline the cognitive foundations of linguistic categories. This thesis is following a Usage-based approach, i.e., word class categories have to fulfill certain criteria in order to be considered psychologically plausible in the sense of CL (Lakoff 1987a: 7; Langacker 1998: 32; Janda 2006: 1). Additionally, actual usage data is the core object of analysis, rather than idealized representations. It is generally assumed that the cognitive mechanisms underlying linguistic categories are visible in the distributional patterns of linguistic data. Empirical data plays an essential role in the endeavor of identifying and describing linguistic categories (cf. Lakoff 1987a). A corpus linguistic approach to word classes is facing the challenge of finding the right operationalization of linguistic concepts to be able to quantify large amounts of data. As later chapters will show, many linguistic phenomena connected to word classes and their patterns are extremely rare and data collections of sufficient size can no longer be annotated by hand. I will consider a variety of theoretical concepts with the aim of facilitating the analysis of word classes at the low resolution of noisy corpus data. The goal is to be able to identify word classes, their subclasses, and boundaries.

The theoretical discussion will be supplemented by a review of related experimental studies, and is also informed by insights from linguistic typology. Cross-linguistic universals that are established on the basis of evidence from a wide variety of languages will serve vital for the analysis of more subtle English word class distinctions. This is especially true since Present-Day English is mostly a non-inflecting language and its word class system is very flexible. Many distinctions present in Old English, such as case, are not available anymore to distinguish the main word classes. The inflectional paradigm is also very limited, which makes it considerably harder to evaluate typicality of members of a word class based on how many oppositions they have (cf. Hopper & Thompson 1984).

Experiential concepts lie at the heart of Usage-based approaches to linguistics, and the next section will turn to this fundamental notion. Special focus will be given to concepts that are central for a Corpus Linguistics approach. The notions of LEXEME and LEXICON are central in defining an appropriate unit of analysis and understanding its challenges and limitations. Formal identity in the form of homonymy and polysemy require special treatment. The debate between nativist and non-nativist approaches still lingers (e.g., review of the 'poverty of stimulus' argument in Pearl 2022). However,

decades of Usage-based Linguistics have successfully demonstrated the value of quantitative data, and a scientific methodology in general (cf. Stefanowitsch 2011a). Therefore, this discussion will not be reiterated here.

2.2. Mental representation of categories

2.2.1. Experience-based categories

Linguistic categories can be divided into descriptive categories and conceptually motivated categories. Descriptive categories can range from 'words that are spelled with the letter *u*' to 'nouns without a singular form (plurale tantum)'. The former is arbitrary and of little use in linguistic analysis, while the latter is useful because it describes a non-obvious property of a language. A different dimension is whether the category reflects a cognitive or a social pattern. This could relate to actual neurological processes, or surface phenomena that are the result of such processes or social interactions. With pluralia tantum, it is not clear whether the category has any cognitive motivation, or whether it is a statistical anomaly that became conventional. In order to be able to assess this, we need to delimit what is conceptually motivated and what is not.

From a Usage-based perspective, the mental representation of linguistic categories is crucially dependent on experience. Langacker (1998: 33) postulates various types of mental experience that play a key role in semantic conceptualization:

1. Established vs novel
2. Abstract (e.g. intellectual concepts) vs concrete (sensory, motor, and emotive experience)
3. Processing time
4. Physical, social, cultural and linguistic context

These can serve as starting point to establish the necessary dimensions that play a role in the formation of grammatical categories. 'Established versus novel' and the connected notion of information structure is cleary manifested in some English morphosyntactic phenomena, such as DEFINITENESS. Information structure, topic and focus, and related concepts, however, are not normally considered defining features of word classes. The distinction between ABSTRACT and CONCRETE, on the other hand, can lead to categorical distinctions. For example, there is a connection to the use of plural marking in English (*love*, *hate*, *information*). Processing time is closely connected to the notion of prototypicality. Prototypical members of a category are processed faster than atypical members. This, in turn, is connected to frequency and other statistical properties of the structures in question. Finally, the linguistic context is easily the most important factor in determining word classes in English. Co-occurrence, collocation,

collostruction, and syntactic structure all ultimately depend on the spatio-temporal proximity of linguistic units.

From these experiential dimensions, linguistic symbols derive their meaning and function. In cognitive grammar (Langacker 1987a; Langacker 1990; Langacker 1987b), symbols are defined as conventional pairings of form and function. These symbols vary along two important dimensions: conventionality and schematicity. Symbols with a high degree of specificity are concrete phonological labels (Langacker 1998). More abstract conceptual structures are also encoded as form-function pairings. Schematicity defines a hierarchical system of symbols. More schematic generally means more grammatical (Langacker 1999; Goldberg 2006). It also means that schematic structures can be used in more contexts since the meaning and function are more abstract and therefore more flexible. In addition, structures vary in terms of their conventionality. Both schematicity and conventionality are fluid concepts and vary across speech communities, individuals, text, and time. It is to be expected that conventionality is strongly correlated with commonness of a linguistic structure, which is grounded in the requirements language has to fulfill in order to describe the experiential world.

Linguistic categories that reflect form-function pairings are also available for analogical extension. Analogy is one of the key mechanisms for categories to accumulate members, and for properties to converge. Analogical extension is based on the generalizations of form-function pairings. The availability of analogical processes and the fact that linguistic categorization is inherently continuous, means that the boundaries between categories are not well-defined. The implication of this is that semantic and distributional properties need to be maximally distinctive on both a conceptual and formal level. This is analogous to distinctiveness in phonology, which is rather well understood. A phonetic feature such as palatalization is not always involved in the formation of categories within a particular language even though it is a descriptive feature of speech sounds in every language. English does not distinguish between palatalized and plain consonants while many Slavic languages do. This does not mean that there is no palatalization. It simply means that the group of palatalized sounds in English is not maximally distinct from the group of plain (and other) sounds, and it is not conventionally used to indicate any conceptual opposition. Even without distinctiveness, there are discernible emergent patterns (cf. Bush 2001), since conventionality is a matter of degree. Therefore, one of the major tasks in linguistic categorization is to determine which features show such a high degree of distinctiveness that they can be used as evidence for the encoding of underlying cognitive concepts.

2.2.2. Grammar and the lexicon

Grammar and the lexicon form a continuum. Diachronically, this is a result of grammaticalization processes, and constant change (Hilpert 2013; Diessel 2019). Synchronically, the continuum is a result of the hierarchical conceptualization of linguistic units that

ranges from the most concrete structures, such as monomorphemic words, to the most abstract schemas, e.g. argument structure constructions, and syntactic phrases (Langacker 1999: 25ff.; Goldberg 1995a). Processes like automation lead to varying degrees of chunking of multi-word constructions (Bybee 2010: 38). Some chunks may lose their internal structure, and there are varying degrees of transparency. Words may be contained in multi-word units that may have more or less fixed slots. These constructions are contained within yet more abstract syntactical structures. In CL, the lexicon-grammar continuum is usually modelled along two orthogonal dimensions: complexity and schematicity (Croft 2001: p. 17; see also Langacker 1987a). From simple to complex, these models identify the following contrasting structures:

(1) simple words: monomorphemic words < syntactic categories
(2) complex words: multi-morphemic words < affixes < affix schemas
(3) multi-word expressions: fixed expressions < semi-fixed constructions < syntactic structures

Complexity of form and schematicity are orthogonal dimensions according to this model. The most specific, and therefore most lexical structures, are monomorphemic words. Bare forms make up the core vocabulary of natural languages. In highly inflecting languages, there may be some degree of schematicity even in bare forms since they can be generalizations from otherwise more common, complex inflected forms. Words (as in units associated to a phonological form) like *spite* in *in spite of* or *spoils* in *spoils of war* or *to the victor go the spoils* always coincide with a limited set of multi-word expressions. Additionally, *spoils* is fixed to a certain inflected form, which is relatively uncommon for a noun. Therefore, they are not typical instantiations of a specific lexical class. However, the word class is still derivable as 'noun' based on generalizations derived from similar constructions. For multi-word units, such as complex prepositions, this assignment, however, may be less informative than for typical monomorphemic items.

There are multiple possible extensions and constraints to this model that apply on a language-specific level. For example, suffixes in English are more abundant than prefixes, both in type and token frequency. English also does not have inflectional prefixes. This has to do, among other factors, with the general word order preference of Indo-European languages (Greenberg 1990) and stress patterns (Molineaux 2012). Berg (2015) argues that the temporal order of prefixes versus suffixes causes the former to be more grammatical and the latter to be more lexical. This can also be seen across derivational suffixes. The major nominalization, verbalization and adjectivization affixes are all suffixes. Those functions are rather abstract, and therefore more grammatical. That does not mean that there are no highly schematic prefixes, negative prefixes being an example. However, being more grammatical, inflectional negation on auxiliary verbs comes in the form of a suffix.

According to this model of the lexico-grammatical continuum, word classes in the traditional sense are hierarchically related to simple words on the lowest level of

complexity. Where the concept of traditional word classes differs is complex inflected forms. With regard to inflected words it is non-reductionist. If the model is taken literally, singular [N] and plural [Ns] would be considered different classes due to their differences in complexity. This makes sense in this case since many distributional properties of plural nouns are complementary with those of singular nouns, e.g., the set of correlated determiners and quantifiers. However, paradigmatically associated forms share many other if not most of their distributional properties with their non-inflected counterparts. The overlap is large enough in some cases to justify the conceptual union of complex forms with their base form. In other case, this is not so clear-cut. The *-ing* form, and other non-finite verb forms in general, cannot be used in the same contexts as their simple counterpart. For *-ing*, this has contributed to theoretical issues of categorizing it as either inflectional or derivational suffix.

I will return to the question of how this continuum is related to word classes in Chapter 3. First however, I will turn to theoretical approaches more specific to morphology since the objects of interest in later chapters are mostly morphologically defined.

2.2.3. Usage-based approaches to morphology

Constructionist approaches to morphology consider multi-morphemic words themselves constructions. This is in line with the notion of constructions as the most abstract linguistic units (Booij 2010; Hoffmann & Trousdale 2013; also Croft 2001). Constructions are stored at different levels of abstractness (Diessel 2016) with the most specific level being lexemes. Bound morphemes are also typically considered lexemes that are stored in the lexicon with their own distributional and functional properties. There is empirical support for the idea of such multiple storage (Bybee 2006; Diessel 2016). Furthermore, Bybee (2006) concludes that high-frequency constructions with bound morphemes are stored in the lexicon, and those with low frequency are analogically derived. Affixes in this sense are an important generalization of complex words. The evidence also suggests that morphology is intrinsically graded (cf. Hay & Baayen 2005 for overview).

The meaning and function of bound morphemes is therefore both highly schematic in isolated bound morphemes, and specific in derived forms. Such forms are more or less lexicalized, partly depending on how much meaning they derive from the schematic meaning of the bound morpheme. This makes rare and derived words carry more of their affix's meaning. Marslen-Wilson et al. (1994) argue that phonological transparency does not matter as much as might be suspected in analogical formation. The abstract morpheme, however, undergoes decomposition when semantically transparent. Semantic transparency of complex words is also required for morphological priming (Marslen-Wilson et al. 1994; Feldman 2000; Longtin, Segui & Hallé 2003). If purely orthographic similarity has an effect, it is less strong. The strength of priming effects seems to be on a cline from orthographically similar to morphologically related to

semantically transparent. Speakers are subconsciously sensitive to the formal similarity between morphemes in different derivational and inflectional constructions. However, they are also sensitive to common allomorphy patterns, for example, *kid(s)* versus *child(ren)*. In addition to this, Construction Morphology (Booij 2010) extends the idea of a hierarchical schema-based lexicon to the level of multi-morphemic words. In such a lexicon, there are intermediate schemas, "which express generalizations about subsets of complex words of a certain type" (Booij 2007: 1; Booij 2005). For example, this allows for the fact that not all nouns are pluralized in the same way and that not all plural forms refer to the same type of plurality, without the concept of plurality losing its categoriality.

Concerning the gerund-participle, it is remarkable that it makes few appearances in the Construction Morphology literature, considering the traditionally strong interest in the topic. This is likely due to the fact that the gerund-participle and the controversy surrounding it are tied very closely to syntactic phenomena, while Construction Morphology is more commonly concerned with word formation. However, in a CxG approach, different uses of the gerund-participle are themselves constructions, and the flexibility of *-ing* compared to other morphemes requires explanation within Construction Morphology.

There is an intricate interaction between generalizations of bound morphemes and generalizations of more specific words within their respective semantic fields. Pluralia-tantum nouns show very specific generalizations leading to what may appear as a coherent class. In some of these cases, the generalization can be attributed very clearly to a dominant exemplar from which the behavior is analogically derived. In analogy to the prominent exemplars *trousers* and *glasses*, there are related words, like *slacks* and *goggles*, that share the same grammatical peculiarites. Goldberg (2006) calls *trousers*, *slacks*, *knickers*, etc., the 'lower-trunk-wear construction' (2006: 218), and considers the plural form motivated due to the bipartite nature of the objects. However, this stops short of explaining whether the lack of a plural is motivated. Strictly speaking, such a motivation is not necessary. *trousers* has no singular, therefore, words describing trouser-like object also have no singular. The question becomes how much generalization is plausible beyond this simple case of analogy. Through a process of template unification (Booij 2007: 38ff), for words like *knickers* to exist, there does not need to be an intermediate noun *knicker* first that refers to a singular version. Analogical derivation can, in a sense, skip levels of abstractions and levels of complexity. Therefore, both the highly specific schema LOWER-TRUNK-WEAR and a more general schema of PLURALE TANTUM are possible sources for the common idiosyncrasies of the words in this class.

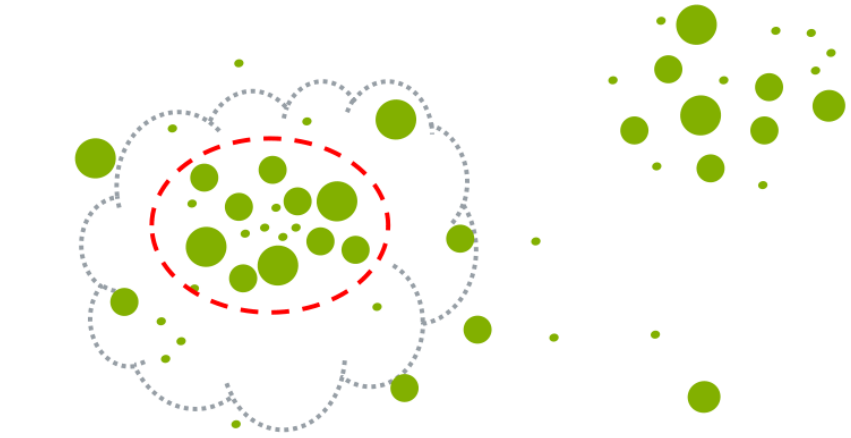


Figure 2.1.: Prototype categories with fuzzy category boundaries

2.3. Gradience

2.3.1. Prototypes and exemplars

Gradience of linguistic categories is typically attributed to a series of experiments investigating natural categories, but has been shown to apply to grammar as well (Rosch 1973a; Rosch 1973b; Labov 1973; Rosch 1978). Prototype Theory (cf. Lakoff 1987b) has been one of the most influential ideas in the study of natural categories. Converging evidence suggests that linguistic categories also exhibit a prototype structure (Langacker 1987a; Lakoff 1987b; Goldberg 1995a; Bybee 2010), and the idea has also been successful in linguistic typology (Croft 2000; Van Der Auwera & Gast 2010; Haspelmath 2010). Grammatical categories have been found to exhibit prototype effects. Initially, such effects were described as 'fuzzy' categorization (Ross 1972; Ross 1973a; also cf Janda 2006: 66ff.). Prototype theory has become one of the most prominent models of categorization in CL. It has also been applied to the study of word classes (Lakoff 1987a; Langacker 1987c; Hopper & Thompson 1984). Semantic and grammatical categories alike are clustered around a prototype center. Figure 2.1 visualizes this idea. Such centers are either schematic representations, concrete exemplars, or a group thereof. Other members of the category are less typical, and there is a gradual decline in similarity until the borders of the category are reached. Prototype clusters of similar members are expected to be more dense towards the center than the periphery. The importance of gradience is undeniable for its theoretical impact and for the empirical turn in linguistics.

In-between categories became an attractive solution where a lack of clear-cut distinctions caused problems in discrete categorization. In some areas, this led to a lumping approach, wherein all non-prototypical data are lumped together into a single in-between category (e.g., the 'nomiverb,' see Haspelmath 2021). The gerund-participle

is a very prominent attempt to consolidate two classical categories whose distinction has caused theoretical issues in English linguistics (Huddleston & Pullum 2002; Duffley 2006). In other areas, the acceptance of gradience has led to a splitting approach, where non-prototypical data is split off into separate categories that vary across semantic-pragmatic clines. In construction-centered approaches, this may seem especially useful, due to the complexity of functional variables and the sheer size of the collective schematic inventory that has been proposed within the paradigm of CL.

An exemplar refers to a specific instance or example of a concept that a speaker encounters. These exemplars are stored in memory. A prototype is defined by the convergence of the most salient members of a category. Salience is correlated strongly with frequency. Many of the most common nouns and verbs show irregularities. Exemplar approaches to linguistic categories are based on the observation that meaning correlates strongly with distribution. Structures that are similar to each other, are expected to have similar distributions (Firth 1957). However, frequency of occurrence alone is not sufficient to determine important exemplars. Bybee (2010) criticizes the idea that the overall frequency of a lexeme outside a given construction has a direct influence on the way it is associated to the construction. She argues that there is a lack of a cognitive mechanism. At least indirectly, such mechanisms have been proposed and tested in the form of cue validity, entrenchment, associative learning (e.g., Rosch 1973b; Goldberg 2006; Ellis 2007a; Stefanowitsch & Flach 2017). Exemplar mechanisms and schematization co-exist (Diessel 2016).

In the later chapters, I am closely following the view outlined in (Hopper & Thompson 1984; Hopper & Thompson 1985; Thompson 1989). Nouns, verbs, adjectives, etc., are prototype categories that exhibit varying degrees of overlap. Nevertheless, they are also maximally distinct within the category space. Distinctiveness plays a major role in the perception of prototype categories (cf. Ellis & Fernando Gonçalves Ferreira-Junior 2009). This ultimately derives from the underlying conceptual motivations. Typical exemplars of the category nouns describe a discrete discourse entity, and typical verbs describe a discrete discourse event (Hopper & Thompson 1985: 151). Even though there is gradience between those two functions, they are diametrically opposed. Hopper & Thompson (1985) dubbed this the 'Iconicity of Lexical Categories' principle. Following an approach that embraces Prototype Theory and the notion of categorical gradience, it is easy to forget that most classical descriptive categories in linguistics describe tendencies so strong that they can be conceptualized as rules without sacrificing too much explanatory power of the model. The importance of gradience in language description has been somewhat exaggerated in some areas. Nevertheless, the same methodological rigor has to be applied to nearly binary grammatical categories as to more fuzzy semantic concepts.

2.3.2. Category space

The variation that prototype categories exhibit is modeled as a category space. The most common approach to the category space is a binary feature analysis (e.g., Quirk 1965; Ross 1972; Ross 1973a; more recently Rosenbach 2003; cf. also Aarts 2007). Gradience arises through the overlap of features. The following table shows a sketch of this approach.

Table 2.1.: A binary feature space

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
word 1	+	-	-	-	-
word 2	+	+	+	-	-
word 3	+	+	+	+	-
word 4	+	+	+	+	+
word 5	-	-	-	-	+

Features may include any type of formal marking, co-occurrence, (con)textual information or semantic property. This type of gradience is fundamentally discrete, or 'quantized' and, there is techincally a finite number of possible combinations. All of the approaches to gradience have a discrete starting point, namely an utterance (idealized or observed). An occurrence either does or does not exhibit a category feature. For example, there is either a plural suffix on *cats* or there is not. There is no in-between state of the suffix being a bit present. At latest in comprehension, a language user typically perceives either plural or bare form. When it comes to lexical classes, however, the accumulation of experiences leads to the actual phenomenon of gradience. A binary feature space as presented above, therefore, works for the description of individual instances, but not categories thereof. Lakoff (1987b) calls this approach 'feature bundles', and rejects it for the lexical level. Sometimes, word class categories like adjectives and verbs are presented in the same way (Quirk 1965; cf. also Aarts 2004). Instead of the occurrence of 'word 1', it is a lexical entry, and instead of the features that describe it in a particular instance, it is the properties of a lexeme that are deemed possible or grammatical. In reality, non-gradable adjectives have a non-zero chance to be graded, mass nouns have a non-zero chance to be counted and pluralia tantum to be singularized. If the object of investigation is the more abstract 'lexeme', the starting point must be probabilistic. The only discreteness that is plausible is the one that comes from the phonological form, and even this is a simplification of the real world (cf. Berg 2000).

Aarts (2004) distinguishes subsective gradience from intersective gradience. The former is a continuum between two categories, e.g. nouns and adjectives, while the latter is gradience within a category. The categories in question are also considered prototype categories, but are presupposed in this view, and there seems to be no straightforward

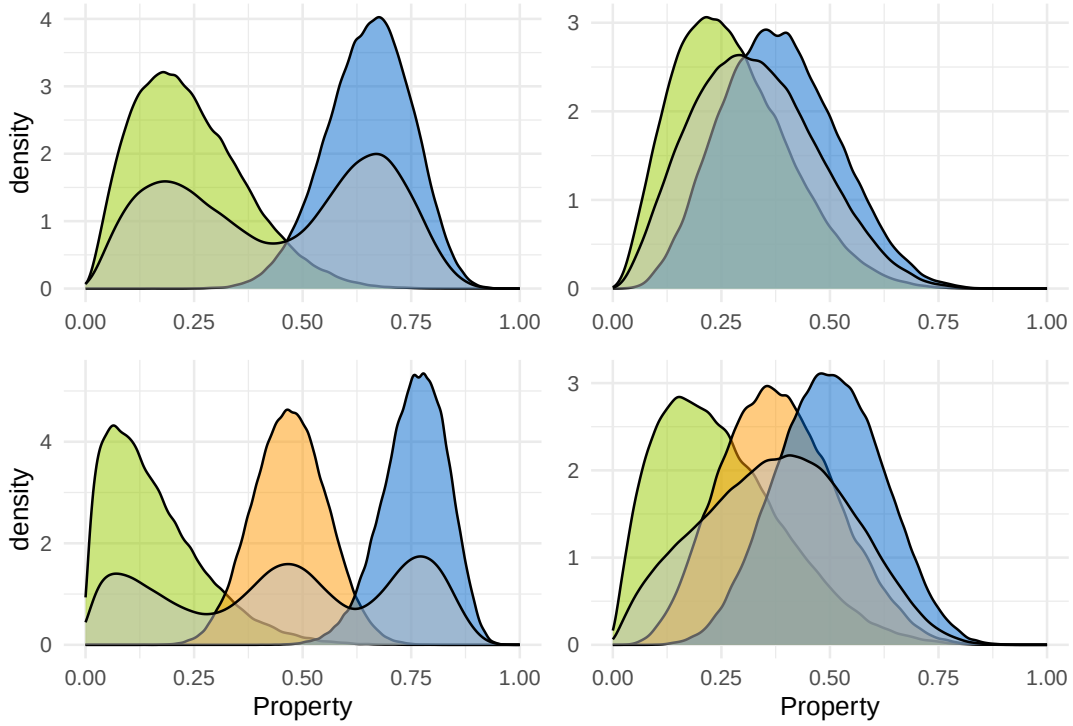


Figure 2.2.: Schematic layouts of overlapping categories.

way of testing whether something actually is a prototype category. This makes the distinction between subsective and intersective gradience rather difficult to apply when there is no clear a priori conception, as is the case with in-between categories. Aarts attempts to solve this simply by positing that intersective gradience is subsective gradience with strong convergence (2004: 32f.), but does not elaborate what would be considered strong. However, distinguishing two types of gradience is promising, since it can be interpreted to imply different types of distributions. Inter-categorical gradience should be much less continuous and should have much sparser transition areas than intra-categorical gradience. This is due to the requirement of maximal distinction. Dense regions in category space are expected to be associated with a single category, while categories adjacent to sparse regions are expected to be associated with multiple categories.

Figure 2.2 shows schematic representations of categories with different degrees of overlap (based on and extended from Aarts 2007: 31). The first panel shows two distributions that have a clear overlap, but their probability masses stay distinct and cause two clear modes, i.e., peaks that represent the region where the most typical elements exist. The analogy to probability distributions also offers an additional aspect. When distributions, overlap (i.e., categories overlap), the distinctiveness of the individual distributions decreases. This can model the historical syncretism of many

English categories that became less and less distinct morphologically. It would also predict the flexibility that English lexemes exhibit with respect to word class conversion. Panel 2 at the top shows what happens if two probability distributions become too similar to each other. The combined distribution appears as mono-modal. The overall density of the cluster, however, is also higher compared to panel 1. Panel 3 shows three sufficiently distinct distributions and panel 4 shows three converging distributions. Often, such mixed distributions show a skew or sometimes what is called 'a shoulder' as can be seen on the left of the distribution in panel 4. Such cases may represent transitional states where a more dominant distribution fuses with a less dominant one, (e.g., competing Old and Middle English plural inflections). In general, major categories are, however, maximally distinct from each other with little overlap. Phenomena like nominal and verbal gerunds that regularly share features of both verbs and nouns are the exception. The conceptual region between categories is expected to be thinly populated.

2.3.3. The uncanny valley

The uncanny valley is a phenomenon originally described in robotics (Mori, MacDorman & Kageki 2012), but it has since been applied to various other fields. The concept describes the feeling of eeriness or discomfort experienced when robots or other human-like entities appear almost, but not quite, human. As a robot's appearance becomes more human-like, there is a corresponding increase in its likability and perceived familiarity. However, at some point, the robot's appearance becomes too similar to a human, and the likeability suddenly drops. This point of sudden drop in likeability is known as the uncanny valley. The effect happens in a transitional region of a category continuum where the stimulus is not distinct enough from either category. Such cases can lead to cognitive dissonance and avoidance. Since it is related to cognition, especially categorization, it can serve as a metaphor for non-prototypical feature combinations in language or places in feature space that have a low probability density.

The uncanny valley effect is related to the concept of prototypicality. In language, some structures are more prototypical than others, meaning that they are more frequent, more regular, and more easily processed. Prototype theory predicts that lexemes avoid a very similar uncanny valley. While English lexemes can undergo conversion rather flexibly, it should be expected that most lexemes tend to behave very close to their category prototype. If this was not the case, there would not be the sense of different classes in the first place. Uncanny valley lexemes that exist at a statistical category boundary should be in the minority. The smoother the transition between categorical prototypes, the weaker is the categorical distinction. For example, there should be a core vocabulary of mass nouns that is sufficiently different from count nouns if the count-mass distinction is a conceptually motivated category grounded in cognitive reality. Nouns that are balanced between count and mass noun uses are

expected to be underrepresented since they exist in an uncanny valley. In other words, if speakers were to use *water* equally often in count noun contexts as in mass noun context, it would not be prototypical for a class of mass nouns and with too may of such 'exceptions' there would not be a basis for distinct categories. The idea of purely semantic properties without any effect on the distributional structure of the class they are encoded in are untenable.

The concept of the 'uncanny valley' can be applied to a variety of linguistic concepts, e.g., to language acquisition and language processing. The idea is that there is a range of linguistic forms or constructions that are more difficult for language learners to acquire or process because they fall in a 'valley' of familiarity, where they are similar enough to known forms, but different enough to cause confusion or difficulty. One example of an 'uncanny valley' in linguistics can be found in the study of the processing of non-native speech sounds. Researchers have found that listeners are better able to discriminate between sounds that are clearly distinct from their native language sounds or that are very similar to their native language sounds, but have difficulty with sounds that are in between these two categories (Werker & Tees 1984; Davidson 2017). Phonetic concepts often have an analog in grammar. It is conceivable that the phenomenon of avoidance of in-between forms applies to lexical classes, as well.

2.4. Multi-level generalizations

2.4.1. Indeterminacy of word meaning

Despite emergent structure through historical change and gradience through varying degrees of schematicity and conventionality, there are yet other reasons for the gradience of word classes. Lexemes exhibit a high degree of overlap with each other and also have significant internal variation. Most lexical items show high degrees of polysemy. The different senses have varying degrees of homogeneity. Polysemy has been presented as a network of senses (Rosch 1973b; Lakoff 1987b; Langacker 1987b) that "[...] form a complex category [...] usually centered on a prototype" (Langacker 1998: 33). It is not always clear whether the nodes of such a network themselves are also embedded prototypes. The idea of a hierarchical schematic system seems to strongly suggest that. A network without this property would be another example of an inherently discrete conception of a category. In a fully probabilistic model, an instance of an utterance should have varying similarity to all senses at the same time that are available for a given form. One of the probabilities within the network is likely dominant in a fully specified communicative context (Wasow 2015).

Homonymy and polysemy are two sides of the same coin. Langacker puts it this way:

> Homonymy is better analyzed as the endpoint along the cline of relatedness—it is the limiting or degenerate case of polysemy, where

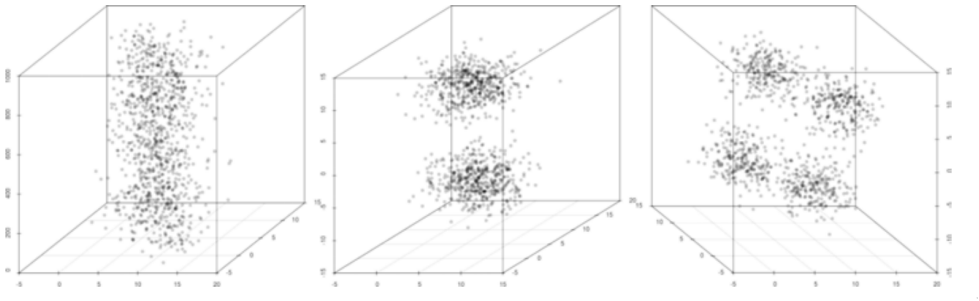


Figure 2.3.: Polysemy in three-dimensional category space, from three different angles, taken from (Gries 2019a: 483)

> the only relationship between two senses consists in their common phonological realization. (Langacker 1990: 268)

Some researchers suggest that homonymy is rather the norm than the exception, especially in earlier stages of language acquisition (Rice 2003: 275). The fact that most words have a discernible meaning is itself based on the schematization of individual experiences with a word form. In that sense, not even the simplest mono-morphemic lexemes are perfectly specific, but only an individual, fully contextualized occurrence in discourse.

The fact that forms are redundantly stored at different levels of complexity and schematicity (Bybee 2006; Diessel 2019) not only leads to extremely vast networks of senses, but also to different levels of generalization. Figure 2.3 shows a three-dimensional representation of a polysemy network (Gries 2019a: 483). The point of view on the network determines distinct entities are perceived. In analogy, categorical distinctions might disappear or neutralize from certain perspectives, i.e., when certain features are focused on and others demoted.

Homonymy is one of the biggest methodological problems in Corpus Linguistics. Homonyms are among the most common reasons for noise in corpus data. Approaches for automated disambiguation exist, but are not in common practice, and researchers often default to manual coding. I argue that this might not be conceptually desirable because it is necessarily a biased disambiguation of lexemes of the same form. With related senses, i.e., polysemy, the problem is even more severe. Manually selecting senses of a polysemous word or different homonyms (with the exception of pure homographs) strongly resembles non-empirical approaches based on speaker intuition. If any type of disambiguation is done, it should be on the basis of distributional criteria. At that point, however, the phenomenon in focus changes from simple lexemes to more complex and more schematic constructions (see Section 2.2.2). It is not possible to investigate, for instance, the categorical affinity and related semantic-functional porperties of the *-ing* suffix, and at the same time exclude adjectival and nominal uses from the data that appear distinct enough for the researcher. At that point, any

generalization is made on a somewhat arbitrary subset, and not a word class or suffix. It is likely that not only one sense is activated when a lexical item is encountered. Moreover, homonyms in actual discourse are disambiguated not only on the basis of co-occurring forms and syntactic context, but also on the basis of complex extra-linguistic contextual information. Once the discourse participant's world knowledge becomes important, disambiguation becomes nearly impossible with traditional corpus linguistic methods.

Recently, advances in deep learning and language models have opened up new possibilities since they are able to encode vast amounts of contextual information in language models with rather little loss and high dimensionality (Mikolov et al. 2013; Wiedemann et al. 2019; Bevilacqua et al. 2021; Beekhuizen, Armstrong & Stevenson 2021). Whether there is some sort of inherently schematic representation that emerges inside the model is an open question. However, in light of the discussion above, disambiguation may be the wrong goal in the first place.

From a theoretical perspective, multi-level generalizations allow for a useful operationalization of lexical units in corpus linguistics. In fact, their existence explains the apparent difficulty in homonym disambiguation. A form should be expected to have different prominent homonymy and polysemy structures given different domains of analysis. The simplest example of this is in lexical ambiguities that do not exist once variety or text type is taken into account. The sensible description, and therefore detection of homonymy, is only possible with respect to certain lexical, grammatical, and/or contextual properties. A form can represent a single lexical item and at the same time multiple polysemes of homonyms. This idea is also in line with findings that forms are stored multiple times in the lexicon.

2.4.2. Competition

Generalizations of linguistic structures are made on multiple levels of abstraction. That is not only true on an analytical level, but there is evidence that suggests that the same is true on a cognitive level (Bybee 2006; Goldberg 2006). Goldberg (2006) argues that both exemplar knowledge and generalized schema play a crucial role. Barsalou (1990) concludes that exemplar memory is indistinguishable from abstraction. Another aspect is the competition of cognitive-functional features and other aspects of lanuguage, such as processing, information structure, and discourse (Diessel 2005; Goldberg 1995a).

Different senses of a lexeme are connected to different concepts. Different concepts are associated with different grammatical structures and collocation patterns. Therefore, due to polysemy and homonymy alone, there is substantial competition between different forms. Additionally, the reduction of English morphology has led to widespread syncretism, where distinct forms of a morpheme have merged into a single form. This can create ambiguity and lead to homonymy if the original functional properties

survive the process. Arguably the most contentious example of syncretism in the English language is the *-ing* suffix for which there has been a lot of debate about how many different functional units it represents, and wether they are in competition (De Smet 2014).

Other times, multiple forms may compete for the same functional space. The *-ing* suffix as a nominalizing suffix competes with the bare infinitive and also sometimes the *-ed* suffix. This type of competition is an integral part of language and leads to a constantly evolving and changing system, where idiosyncrasies and transitional forms arise regularly due to competing motivations (Tomasello 2003; Wulff 2008; Diessel 2019). The key to creativity in language may lie in the competition of different co-existent generalizations. Goldberg (2019) proposes that constructions are partially productive. Booij (2007) calls a similar idea 'embedded productivity'. Langacker (1987a) distinguishes full and partial schematicity.

3. The word class continuum

3.1. Overview

The 'word' is the most intuitive unit of language, and it makes sense that the classification of words is one of the oldest linguistic topics. The terms *word class*, *lexical class*, and *part of speech* are often used interchangeably to refer to the categories of words in a language based on their syntactic, morphological, semantic and functional properties. They are sometimes used to make subtle distinctions between different types of word categories. Nevertheless, they all refer to the same general concept. Additionally, Croft (1991) distinguishes *lexical classes* from *syntactic categories* to refer to a related idea. He uses the term syntactic category to refer to the smallest, non-compositional units of schematic constructions (see (1)-(3)). In his sense, slots commonly labeled as [N] or [V] in CxG depend on the construction in question, and do not refer to a lexical class of words. The consequence is that [N] in [DET ADJ N] and [N] in [DET N of-phrase] refer to separate lexical classes (Croft 2001: 50f; Croft 2022). This non-reductionist approach is the essence of Radical Construction Grammar (RCG), but also characteristic of some approaches to CxG in general. In this thesis, the theoretical idea of word class and its operational definition generally does not go as far as that. I will investigate word classes in the more traditional sense as functionally motivated and distributional distinct categories of English lexemes. The assumption is that the [N]s above do form a class whose members are related and analogically derivable from each other. However, I will draw from insights of RCG and empirical evidence from related fields of Linguistic Typology in order to select distributional categories to focus on.

The lexical idiosyncrasies of word classes have led to a myriad of different theoretical approaches to word classes ranging from Aristotelian models of categorization to fully data-driven, probabilistic approaches found in modern Machine Learning. These models of categorization range from simple, reductionist frameworks with complex mechanisms to account for exceptions (e.g., transformations rules, Chomsky 1957; Chomsky 1970) to complex, non-reductionist frameworks that are typically focused on smaller units of observation (e.g., Construction Grammar, Goldberg 1995a). What they all have in common is that they are simplifications, and each approach has its own strengths and limitations. On the one hand, in more traditional, rationalist theories, the status and discreteness of nouns, verbs and adjectives is rarely challenged. They are treated as primitives for analysis (Jackendoff 2002). Despite the discrepancies mentioned, it is far from the worst model for English (and Latin, and Greek, etc.). On

the other hand, in Usage-based linguistics, and related theories, the very concept of word classes is not as straightforward (cf. Croft 2022; Croft 2001; Diessel 2019). From a typological perspective, word classes distinctions may not be clear-cut, and sometimes absent (e.g., adjectives may be hard to define, cf. Dixon 1977).

The basic unit of analysis determines the types of classes that can be identified. Since (proto-)typicality plays such an important role in the formation of linguistic classes, it makes sense to start from the simplest and least schematic units of language. Syntactically, the units in question can be identified as the terminal node or ultimate constituent (Haspelmath 2012). As such, their form and functional profile embody lexemes that make up a lexicon of form-function pairings. The traditional notion of the lexicon is considered to be made up of mostly single-word and fewer multi-word units. With advances in CL, it became evident that a lexicon must accommodate larger structures, such as idioms and fixed expressions, and other prefabricated units (cf. Bybee 2002; Arnon & Snider 2010), but also constructions of varying schematicity. More recently, the term 'constructicon' was coined to move on from the classical lexicon (Goldberg 1995b; see also Herbst 2019); however, this notion has mostly been adopted in lexicography, and 'lexicon' is still mostly used as hyperonym in most of the CL and CxG literature. In either case, monomorphemic words are stored alongside constructions, and they are the simplest members of the lexicon/constructicon. They are also the locus of word class distinctions, which is not true for constructions. This conceptual mismatch of words as primary unit versus construction as primary unit has been approached in different ways. One notable approach is to consider even the simplest meaningful structure as a construction, which is the approach taken in Construction Morphology (Booij 2007; Booij 2010), and RCG.

The categories noun and verb can be viewed as universal lexicalizations of the prototypical discourse functions (Hopper & Thompson 1984; Croft 1991; Langacker 1999; Croft 2001; Croft 2000).

> The more a form refers to a discrete discourse entity or reports a discrete discourse event, the more distinct will be its linguistic form from neighboring forms, (...) (Hopper & Thompson 1985: 151)

The more typical a lexical item for the category noun, the more likely it is to occur in contexts typical for nouns. However, nouns do not strictly require such a context to be identified as noun. Their class association, to some degree, is built into the lexicon. The word class is commonly considered part of the statistical information on a lexeme that is available to a speaker. Forms and accompanying forms of lexemes can be understood as cues for identifying its category membership, which in turn determines its functional interpretation.

Word class categories, like other categories in linguistics, are centered around a conceptual prototype that is characterized by a gradient from less to more typical. Nouns and verbs are commonly considered as maximally distinct (Langacker 1987c). The functional motivation for this is the salient experiential differences between discourse

events and discourse participants (Hopper & Thompson 1984). If we deal with a highly distinctive grammatical category, we see extreme lexical clustering. In fact, so extreme, that it took some time for the linguistic community to embrace the idea that there is a noun-verb continuum at all (cf. McClelland et al. 2010). However, the noun-verb distinction is not so clear-cut in all languages. In fact, even English nouns and verbs have fewer morphosyntactic cues available than those in strongly inflecting languages.

3.2. Cross-linguistic perspectives

3.2.1. Essentialist word classes

Where English does not have formal distinctions, a cross-linguistic perspective can provide insight into other structuring factors that might be distinctive on a higher level. Linguistic categories are language-specific, so are word class categories, especially subcategories (Croft 2000; Haspelmath 2010). Nevertheless, it has been shown that some tendencies concerning word class categories seem to be universal, e.g. the verb-noun distinction (Croft 1991: ch. 2; also cf. Baker 2003) and adjectives (Dixon 2004). Furthermore, some categorical distinctions made in one language can be found as tendencies in another language representing 'soft generalizations' (Aarts 2007: 74). Langacker (1998) argues very strongly for the universality of the noun-verb distinction.

> I personally find it hard to imagine that fundamental and universal categories like noun and verb would not have a conceptual basis [...], I believe that such categories reflect inborn cognitive abilities that are initially manifested in the category prototype (Langacker 1998: 46)

This reflects the general consensus within CL. This is immensely important for the study of language-specific categories because it implies that there is a definitive set of categories that are conceptually motivated, in contrast to purely descriptive categories that have arisen through other mechanisms. Stefanowitsch (2008: 527) distinguishes between arbitrary and motivated restrictions. Semantic motivations may not be necessary for the emergence and learning of patterns.

Lexical fluidity between the categories noun and verb is especially evident in non-inflecting languages (cf. Arcodia 2014 on Mandarin Chinese; Hendrikse & Poulos 1994 on Southern Bantu). Haspelmath discusses the idea of a class of 'nomiverbs' (2021: 20f.), which are lexical roots that are not inherently associated with either word class and can manifest as either nouns or verbs. As examples, he lists the English forms *hammer*, *dance*, *walk*. The idea is tempting since the formal and semantic similarity between the nominal and verbal uses of those lexemes is clear. Assuming different lexical entries based on the word class is not satisfactory under CL assumptions. Croft goes as far as to characterize word classes as epiphenomenon of the elements within the constructions

they appear in (Croft 2022), shifting the focus to individual constructions of varying degrees of complexity and schematicity. In that sense, lexical roots are inherently classless. Croft makes a clear distinction between 'essentialist' and 'language-specific' word classes, with the latter being constructions. This non-reductionist approach is necessary for typological comparison, but it significantly complicates the description of word classes within a given language. However, the idea of a more abstract conceptual system underlying word classes that is not dependent on the object language is very true to CL, and offers interesting perspectives. The conceptual layout in Croft's model of essentialist word classes can be seen in Table 3.1.

Table 3.1.: Word classes according to Croft (1991: 53, 67)

	Reference	Modification	Predication
Objects	**unmarked nouns**	genitive, PPs on nouns, adjectivalizations	predicate nominals, copulas
Properties	de-adjectival nouns	**unmarked adjectives**	participles, relative clauses action
Actions	action nominals, complements infinitives, gerunds	predicate adjectives, copulas	**unmarked verbs**

Unmarked in this case refers to a variety of properties. Croft distinguishes three types: formal, behavioral and textual markedness. A form that is formally unmarked has fewer morphemes than its marked counterpart. A deverbal noun derived with an affix is more marked than a monomorphemic nominal root. Predicative adjectives are more marked in English since they require a copula. Behavioral markedness refers to the potential of a form to be used with category defining morphology and syntax. For example, bare de-adjectival nouns in English do not inflect and are mostly restricted to a few nominal constructions, such as uses with the definite article: *the poor*, **a poor*. Textual markedness refers to the frequency of a form in a given text. A

The lexico-grammatical continuum (cf. Section 2.2.2) and the related grammaticalization cline is mostly orthogonal to these dimensions. Languages vary with respect to their set of grammatical structures and how they are encoded. English predicative adjectives, e.g., are linked to the copula construction, but are otherwise morphosyntactically identical to attributive adjectives in that they can be modified by adverbs, be a head of an adjective phrase, take degree modifiers, etc.

Croft proposes two distinct types of in-between category: intermediate categories (Croft 1991: 23, 133), and transitory categories (Croft 1991: 142ff.). Intermediate categories share grammatical properties of their neighboring categories. Numerals and quantifiers, for example, vary between nominal and adjectival behavior, i.e., a modifying function, mostly in attributive position. Adjectival syntax is more likely for smaller

numerals, and nominal syntax for larger numerals. Consider the following examples 4-7 that demonstrate uses from predicative over attributive to nominal.

(4) You can see erm **one** example of this , a striking example (BNC:KRL)
(5) I was **twelve** . I did n't have the language to explain (...) (COCA: 4000283)
(6) This guy is the **one** , Malcolm said (BNC:A6E)
(7) Another Astra change is the upgrading of specification on the **1** . (BNC:K2P)

Quantifiers are another example, and they share aspects of adjectival and nominal class features (cf. Chapters 5 and 6). Most intermediate categories that belong to closed classes are unlikely to exhibit a prototype structure due to their low type frequency. However, there is usually a clear functional motivation for their in-between status. Transitory categories, on the other hand, are emergent categories that are the result of historical change. They also have irregular grammatical behavior and show less of the typical properties of the category they are historically derived from. Croft proposes that transitory categories do not have a prototypical core, but rather display "a cline of grammatical behavior" (Croft 1991: 143). Examples of this are English auxiliaries that are less verb-like than prototypical verbs, but not yet affixes. Other minor word classes in English include determiners, prepositions, conjunctions, determiners, and demonstratives, all of which can be mapped onto a noun-verb or noun-adjective-verb continuum, as well. The only word class in this list that does not seem to be derived from members of one of the major word classes are demonstratives (Diessel 1999). This is considered a typological universal. However, they can be the source of other minor word classes. The English definite article originates from demonstratives, and determiners are considered transitory word classes since they are likely to develop into clitics and affixes (e.g., in languages of the Balkan Sprachbund).

With gradience across all descriptive and conceptual concepts, one of the questions becomes on which level of schematicity to operate. English word classes tend to be associated to very abstract schemas, e.g., slots in argument structure constructions (cf. Goldberg 2006), possessive constructions (Rosenbach 2003; Rosenbach 2014), or near-atomic extremely schematic syntactic structures like [DET N]. Rosenbach (2003)'s empirical findings on the genitive alternation line up well with Croft's framework if the inherent difference between the *of*-genitive and the *s*-genitive is considered (cf. Section 2.2.2). The *of*-genitive is more syntactic than the *s*-genitive, which is more morphological, being marked by a clitic. Likewise, the more morphological *s*-genitive is more likely to be populated with more 'nouny' nouns, i.e., those that are animate, objects, and concrete. The *of*-genitive is 'more marked' and populated by less typical nouns, i.e., those that are inanimate, stative, and abstract, etc. (Rosenbach 2014: 232 for full list of identified properties). This variation is rather fluid, and many nouns can be found in both constructions. This can be explained in terms of different frames. An inanimate noun is most common with the *of*-genitive, but an *s*-genitive can be used to emphasize its potential animacy metaphorically. Consider (8), where *pride* is personified, which is further supported by *dictates*.

(8) Sighing , she ignored **pride 's dictates** and sank back again (BNC: H9L)

The general conclusion from the typological literature is that conceptual distinctions, such as 'modification', 'predication', 'gradability', 'animacy', etc., may be more important dimensions in the discussion of the word class system, even if the concepts are not fully grammaticalized in English. This is congruent with the general ideas of CL laid out in Chapter 2. Furthermore, at multiple points in the discussion so far, the notions of clines and prototype categories have been contrasted with each other. Therefore, the following section will provide a more detailed account of why a difference between the two is important.

3.2.2. Clines vs prototype categories

Clines and continua are two important concepts in linguistic analysis that are used to describe patterns of variation and change in language. While they share some similarities, they are also distinct in important ways. A cline describes a gradual and continuous pattern of variation in a linguistic feature usually across a geographic or social space or diachronically. Clines are typically visualized as a line or gradient that represents the gradual change in the linguistic feature being studied. In the context of grammaticalization, clines are typically directional (Haspelmath 1999; Hopper & Traugott 2003).

Hopper and Traugott provide a detailed account on grammaticalization clines and recognize that not all points on a cline are equal:

> The metaphors ["cline" and "continuum"] are to be understood as having certain focal points where phenomena may cluster. [...] The precise cluster points are to a certain extent arbitrary. Linguists may not agree on what points to put on a cline [...] (Hopper & Traugott 2003: 6)

However, they do not elaborate on how to identify such focal points and only pick up this notion again once in the context of polysemy. This demonstrates that the question of what even constitutes a linguistic class is a problem both synchronically and diachronically. Croft (1991) distinguishes between clines and prototype categories in the context of transitory and intermediate categories. The most complete notion of this distinction can be found in Aarts (2004; see also Aarts 2007). Aarts argues that there are different types of gradience, which can be distinguished by whether they are between categories or within categories (see Section 3.3.4). There is some overlap between these different notions.

There are few quantitative studies directly concerned with prototypicality and their associated distribution(s) in corpus data. Experimental evidence suggests that word-frequency distributions generally correlate with prototypicality (e.g., Ellis & Fernando Gonçalves Ferreira-Junior 2009; Wulff et al. 2009). Zipf's law may be connected to

prototype categories and may be less pronounced for pure descriptive categories (see Section 4.3.1). Moving forward, the working hypothesis is that functional properties create prototype effects when they are contingent on formal patterns, while purely transitional gradience from historical change (phonetic processes, syncretism, mergers, automation, optimality) create clines, i.e., evenly spread continua.

3.2.3. Markedness hierarchies

The concept of markedness has been criticized for being too vague (e.g., Haspelmath 2006). Nevertheless, it is a staple in linguistic terminology. Among the types of *markedness* discussed in Haspelmath (2006: 26), I will mostly refer to a structure as *marked* if it has overt coding, i.e., if it is marked by morphological or syntactical means. A complex word is more marked than a monomorphemic one; a noun phrase with an of-phrase is more marked than one without. This is in line with the notion of structural coding:

> Structural coding: the marked value of a grammatical category will be expressed by at least as many morphemes as is the unmarked value of that category (Croft 2003: 92)

The second main type of markedness according to Croft (2003), "inflectional potential", is less obvious in English:

> Inflectional potential: if the marked value has a certain number of formal distinctions in an inflectional paradigm, then the unmarked value will have at least as many formal distinctions in the same paradigm (Croft 2003: 97)

Inflectional paradigms in English are rather restricted, but it can be hypothesized that this tendency is true for derivational morphology as well. Complex words themselves often show rather restricted derivational morphology. The selection of derivational suffixes is one of the main defining features of word classes. It is not obvious, however, whether derivability is a feature of typical members of a word class. For adjectives, *-ish* is extensively used on non-gradable adjectives that normally do not occur in the context of quantification: *blueish*, *greenish*; in spoken language even on numerals, like *nine-ish*. Neo-classical suffixes and combining forms often occur on nouns that are not typical members of their word class either due to them being borrowed as a whole or created through backformation. Different derivational suffixes likely correlate with bases of varying degree of typicality.

The third type according to Croft is textual markedness, which is concerned with the frequency and commonness of a structure. A marked form is rarer. For example, there are fewer mass nouns than count nouns, and fewer ungrabable adjectives than gradable adjectives. This property is rather straightforward in a quantitative corpus study and will be discussed in detail in Chapter 4.

Various hierarchical tendencies have been derived from markedness patterns and observed cross-linguistically. For example, Givón introduced the idea of a time-stability scale (Givón 1979: 14; Givón 1985). Prototypical verbs, i.e., verbs that describe dynamic events are also more time-stable than stative verbs. At the same time, stative verbs can be used in fewer contexts than dynamic verbs. Words describing time-stable concepts tend to be less marked relative to verbal features. In English, the progressive construction is less common with stative verbs. Hopper & Thompson (1985) propose a number of hierarchies based on cross-linguistic evidence that are all related varying degrees of nouns and verbs fulfilling their prototypical discourse functions.

3.3. The English noun-adjective-verb continuum

3.3.1. Between nouns and verbs

The noun-verb continuum as a concept has been around for many decades and numerous studies found evidence for gradience in English word classes on all classic levels of linguistic analysis, i.e., (morpho)syntax (e.g., Ross 1972; Ross 1973b; Ross 1973a; Comrie 1975; Aarts 2007), semantics (e.g., Givón 1984; Winter & Lievers 2017), pragmatics (e.g., Thompson 1989), phonology (e.g., Berg 2000), pragmatics (Lehmann 2018). The issue was initially explored mostly with respect to adjectives. Adjectives serve functions that cover a wide range, including verbal to nominal uses. Furthermore, it has been debated whether they are universal, since some languages are considered to have small, closed adjective classes, or none at all (cf. Rijkhoff 2000; Haspelmath 2012). I will return to the topic of adjectives in Section 3.3.2 and Chapter 5.

Numerous studies have successfully leveraged the model of a verb-noun continuum, finding semantic heterogeneity of nominalizations (Bekaert & Enghels 2019; Hartmann 2019; Fonteyn 2019a). This heterogeneity is considered the result of different stages in the historical development of nominals with respect to a 'cycle of categorical shift' (Fonteyn 2019b). Deverbal nominalizations are initially less 'nouny' than other nouns (Bekaert & Enghels 2019). Certain nominalizations are semantically similar to their base verbs, whereas others approximate the prototypical semantics of a noun. This divergence in semantics is reflected in their distributional properties. Hartmann (2019) argues that those differences in 'nouniness' can be explained in terms of a substantivization cline.

Another pattern that has been observed is that subjects are more prototypical than objects that are more typical than obliques in terms of 'nouniness'. This means that nouns in subject position are more likely to show other noun-specific marking.

(9) SBJ < OBJ < OBL (Croft 1991: 186)

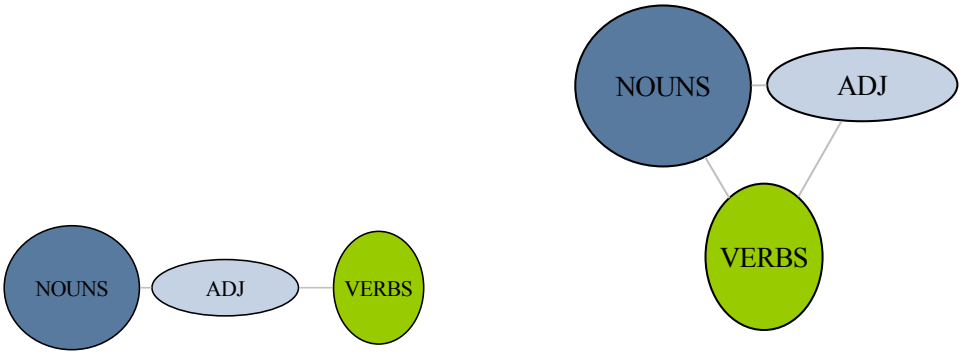
English does not have morphological case marking. Subject and object function are generally marked by means of word order. The syntactic analog to oblique case marking is the use of prepositions. Even though formal marking of these concepts is limited in English, the expectation is that prototypicality of nouns is correlated with this hierarchy. For example, *home* is less concrete than *house* and is very common in bare form and/or in prepositional phrases, such as in *at home*.

Many other patterns can be derived from Hopper & Thompson (1985; also Hopper & Thompson 1984). For example, predicate nominals are less typical for nouns, and definiteness is more 'nouny' than indefiniteness. German predicate nominals typically occur without article (*Sie ist Lehrerin*). In English, these 'role predicates' (Doron 1988) are not as common, but there are cases, especially with *of*-phrases: *he was president of the United States*. Such bare uses are heavily restricted in English, and generally a sign of a less prototypical use. It is most common with proper names and mass nouns. Mass nouns are grammatically less typical than count nouns. They are also perceived as less concrete than count nouns (Lievers, Bolognesi & Winter 2021). It can be suspected that other predicative uses might also be less 'nouny'. Most variation between nouns and verbs is related to Givon's (1980) time-stability scale. Nouns are more time-stable than verbs, and verbs are more time-stable than adjectives. Physical objects are usually perceived as time-stable entities compared to events. This is reflected in the fact that less time-stable concepts, such as *fire* are also correlated with non-typical constructions like the predicate position as in, *there was a fire*. Inherently more dynamic concepts, such as sound, have been shown to map more strongly onto verbal structures, while non-changing, static concepts, such as color, onto nominal ones (Winter & Lievers 2017).

It is important to emphasize that universal markedness patterns are tendencies. In English, there are some notable exceptions. In verbal morphology, it is usually the form of the 3rd person singular present tense which is the unmarked form (Greenberg 1990: 259f.). English shows exactly the opposite pattern with only the 3rd person being marked. It is likely that this is a transitory state of the English inflectional system. Basing categorization on typological tendencies poses the risk of introducing the same fallacy based on which Latin case categories used to be assigned to English. A universal tendency or concept can only serve as a starting point in the search for English prototype categories. Due to the probabilistic nature of such concepts, English may or may not exhibit expected distinctions.

3.3.2. The position of adjectives

Adjectives are typically considered one of the primary open word classes in English, alongside nouns and verbs. However, this perspective is heavily influenced by a Eurocentric viewpoint. In fact, some languages have a small, closed set of adjectives, while others do not differentiate between adjectives and nouns, or between adjectives



(a) linear (b) triangular

Figure 3.1.: Noun-Adjective-Verb continuum models

and verbs. For example, some Niger-Congo languages have a small, closed class of adjectives (Welmers 1973; also see Dixon 1977; Dixon 2004), and Chinese Mandarin is commonly analyzed to lack a distinct class of adjectives, with some scholars describing them as a subset of verbs (cf. McCawley 1992). Whether adjectives are universal is more controversial than the noun-verb distinction. Croft (2001) assumes that modification as an essentialist category is universal; but it does not need to be a lexicalized in a given language. In terms of markedness and syntactic independence, English adjectives are more restricted than nouns and verbs. Croft (1991) considers adjective prototypes a subtype of noun and verb prototypes (Croft 1991: 130). However, it is likely that every language has some type of adjective class (Thompson 1989; Dixon 2004).

A significant part of the prior theoretical and empirical work on adjectives in English is concerned with the position of adjectives on the noun-verb continuum (Comrie 1975; Ross 1972; Hopper & Thompson 1984; Givón 1984; Berg 2000). There are two main types of models that are used to explain the distribution of adjectives in English: the linear model, and the triangular model. Both types were already considered in Ross (1972). Moreover, in linear models, there have been variations placing adjectives outside the noun-verb continuum or in the middle of it.

Applying Booij (2010)'s notion of template unification, the triangular model captures the fact that there need to be no intermediate forms for movement across the continuum. Categorical variability as a result of constant diachronic change is also captured better by a model where all major categories are interconnected. Nouns shifting to a more verbal use do not have to undergo a stage of being an adjective first. The linear model, on the other hand, captures conceptual notions of functional continua better where modification is considered as intermediate. Croft (1991)'s model of essentialist word

classes (cf. Section 3.2.1) combines aspects of a linear model since unmarked adjectives lie between unmarked nouns and verbs, but it also allows different routes.

The shape of the continuum is not the only property that can be assessed. There is also a degree of similarity to either verbs and nouns. In Germanic languages, at least those that still inflect, adjectives share similar inflectional paradigms with nouns, not verbs. Generally, English adjectives have been shown to show more similarity with nouns.

3.3.3. Obligatory presence versus obligatory absence

The distributional features that define language-specific word classes and their subcategories are mostly based on the presence of morphological or syntactic markers. Inflectional morphology is the clearest example of mutually exclusive features. Subclasses, such as gradable adjectives and mass nouns, however, are additionally described by the absence of some of these features. Mass nouns lack plural marking; non-gradable adjectives lack comparative and superlative marking; stative verbs cannot be used in the progressive construction. However, such subcategories also have other associated morphosyntactic properties not defined by absence. The absence of a feature is inherently different from the presence of a feature.

The statistically significant absence of a structure can provide evidence for the avoidance of a specific feature and also discriminate from accidental absence (Stefanowitsch 2006; Stefanowitsch 2008; Ambridge et al. 2015). Langacker proposes that any encounter with a linguistic structure has a positive impact on its entrenchment in memory, and that its absence over extended periods of time has a negative effect (Langacker 1987a: 59). The relationship between negative and positive entrenchment in this model is inherently asymmetric when other aspects of experience-based categorization are taken into account. A structure can be salient for other reasons than its frequency or regular use, such as surprise or high contrast, and explicit emphasis (cf. Langacker 1990). Arguably, these types of salience do not apply to absent features.

Research in language acquisition suggests that children neither have systematic access to nor do they require explicit negative evidence to learn grammatical patterns (Marcus 1993; Rohde & Plaut 1999). Recasts with positive evidence have been described as triggers for self-correction (cf. Goldberg 2019: 84). Whether this is necessary or even beneficial, however, is controversial (cf. Morgan, Bonamo & Travis 1995). Error-based learning does not seem to be a requirement in acquisition models (McClelland et al. 2010; Perfors, Tenenbaum & Wonnacott 2010). The statistical evidence for non-existence of a lexeme within a construction may not be enough (Stefanowitsch 2011b; Pinker 1988). However, acceptability of lexico-grammatical structures has been found to correlate with statistical evidence from corpora (Stefanowitsch 2008).

In any given data set, there are a large number of words that never occur in their inflected form, and it is likely that many words that a speaker experiences are never

encountered in their inflected form (or a certain construction, syntax, collocation, etc). If probabilistic knowledge is required for the preemption of obligatory absence, then the structure in question would require a certain minimum commonness in order to be used as negative evidence. In the very least, there would need to be enough evidence from other frequent exemplars from which to abstract.

Another asymmetry lies in the fact that marked forms are less frequent than their unmarked forms (cf. Section 4.3.3). That makes forms that always occur with a certain feature less likely. In this case, their unmarked form would have to be absent enough to be negatively entrenched, which takes more input for inflected forms due to their generally lower commonness. From these considerations, a ranking of the expected likeliness of a categorization can be derived:

1. marked and unmarked forms available
2. only unmarked forms available
3. only marked forms available

Note that the fourth logical option—absence of both marked and unmarked forms—is mostly trivial, except in perhaps cross-linguistic comparison or taboo language.

3.3.4. Stucture of the multidimensional category space

In the following section, I will argue that well-known processes from historical linguistics and linguistic typology can provide useful constraints on a multivariate analysis. While multidimensional concepts and multivariate data have become more and more prevalent, there is often little regard for the inherent structure of the category space itself. Not every formal cue or contextual feature contributes equally to category formation. The idea is that an utterance provides many cues that can be contingent with many form-function pairings due to multi-level generalizations. This idea represents associative categories better than the vague notion of multidimensionality. It provides an opportunity to obtain simpler, more intuitive models that are not statistical black boxes. It is otherwise easy to attribute shortcomings in a model simply to missing variables. The complexity of linguistic structures has even been used as an argument against statistical methodology proper.

(10) I hope you do n't mind **my asking** so many questions (BNC: G3E)

There is layered information available from an utterance like in (10). The word *asking* has it own object, which is a cue for being a verb. It is also preceded by a determiner, which is a cue for being a noun. In line with Aarts (2007)'s idea of intersective gradience, and Figure 2.3, there can be two different perspectives taken on this utterance. It can be approached from the nominal side or from the verbal side. In both cases, the whole construction is likely to be non-prototypical due to the specific construction not being available for other members of the classes nouns and verbs. Consequently, the

lexeme *asking* is not necessarily member of an in between category, rather the instance contributes to *asking* incrementally moving away from either prototype center of noun and verb. This shift is not too drastic here due to the presence of a class-defining property. Only if enough lexemes are balanced between both nouns and verbs does an in-between category emerge.

Categorical distinctions can be connected to a variety of formal markers. However, not all formal marking is equally informative. Consider the following groups of words:

(11) string, strap, strike, strip, striation
(12) clothes, trousers, scissors, spectacles
(13) food, water, blood, dust
(14) earnings, tidings, proceedings

In every of these groups, there is a common formal marker that is shared by all of its members. In (11), it is a phonological pattern, in (12), the obligatory presence of the plural marking, in (13), the lack of plural marking, and in (14), the *-ing* suffix plus the obligatory presence of the plural marking. All of these categories may be conceptually motivated. However, phonological markers alone are not too commonly found to be distinctive since phonological form is largely arbitrary. There is some experimental evidence that supports certain cases of sound symbolism (Kwon 2017; Nuckolls 1999; Winter & Lievers 2017; Mompean, Fregier & Valenzuela 2020). The motivation of these patterns is controversial with one side of the argument proposing (universal) iconicity and the other (language-specific) analogy (Bergen 2004). Nevertheless, it is safe to assume that sound symbolism does not play a major role in word class formation.

Absence of certain class-defining features like the lack of plural marking in mass nouns is much more common. Most classes have subclasses based on such an absence. Less prototypical members usually show fewer category distinctions (Hopper & Thompson 1984; Greenberg 1990), which is common enough for class formation. Therefore, absence of morphological marking is a strong indicator for a categorical distinction. On the other hand, obligatory absence is not as common. Plurale tantum nouns are cross-linguistically unsystematic (Acquaviva 2008; Corbett 2019; Mackenzie 2019). Mass nouns also have their own morphosyntactic profile that is different from count nouns. Plurale tantum nouns on the other hand are more similar to regular plurals, and only show few significant lexical patterns, such as the *a pair of* construction (Huddleston & Pullum 2002: 341). In general, morphological markedness provides stronger cues for category membership. For example, only the most lexicalized nominalizations occur frequently with plural marking (*building*, *meaning*). Form differences through affixation are simpler and require less cognitive effort than syntactic and lexical patterns, such as collocation. Likewise, associations and frequency patterns are typically more extreme toward the grammatical side of the continuum. So extreme in fact that they sometimes resemble fixed rules.

In summary, categorization needs to be conceptually motivated, and semantic concepts are established by “drawing on all available resources” [langacker10: p. 33]. However,

it is equally important to recognize that not all resources contribute equally. Langacker acknowledges that some types of experience have 'inherent cognitive salience' (Langacker 1987c: p. 37). Associated conceptual dimensions include whole-part relationships and by extension metonymy, animate-inanimate, physical-abstract, spatial and temporal primacy, and contiguity. English lacks a nominal class or derivation mechanism whose primary function is to distinguish between animate and inanimate, but animacy can still be found to play a significant role in the choice of syntactic structure (e.g., in the choice of possessive construction Rosenbach 2003). English has few morphosyntactic constraints on the choice between *s*-genitive and *of*-genitive, but those trigger very strong tendencies. So strong, in fact, that they have been called 'categorical contexts' Rosenbach (2019) in variationist studies. It is important to note that such categorical contexts are mostly morphosyntactic.

4. Methodology

4.1. Overview

The following chapter will outline the methodological concepts that are crucial to the corpus studies. For the investigation of type distributions, every step of the corpus linguistic pipeline from tokenization, to tagging, to the choice of statistical measures require careful consideration. One of the main assumptions is that conceptually motivated structures and correlations are distinguishable from spurious ones. The resolution of corpus data is generally low and there is a lot of noise. A variety of statistical measures can be used in order to filter out meaningful signal. Especially association measures and dispersion measures will play a key role in this endeavor. A common criticism of corpus approaches is that "there are way too many variables, and [that] all the data is contaminated" (Janda 2006: 8). However, the same is true for real-world use of language. It is not just a feature of empirical data but intrinsic to linguistic interaction. Empirical data should be considered as multivariate, and not all noise is a result of methodological shortcomings. A language learner is also subjected to noisy input.

The approach followed here is multivariate. This does not necessarily mean that all the modeling is multivariate, but that a variety of different measures are taken into account. This is in line with the idea of a tupelized approach (Gries 2019b).

I wrote a full implementation in *R* (R Core Team 2021) for all lexical statistical measures used in this thesis. It is available in the `occurR` package (Rauhut 2022a). The package solves common performance problems with large data sets, especially concerning dispersion measures. An additional auxiliary package `linguio` (Rauhut 2023) for data import, export and communication with the Corpus Workbench (CWB) is also available. Selected implementation details and examples of the use of the package can be found in Appendix A.1.

Considering that word-hood is also continuous due to processes like grammaticalization and lexicalization (cf. Hopper & Traugott 2003; Bybee & Scheibman 1999), using lemma as the basic target of analysis is a somewhat arbitrary, but necessary heuristic. To prevent fixed multi-word expressions and other idiomatic expressions from confounding the data, a measure of lexical fixedness is going to be used as a control.

A comprehensive analysis of the distributional patterns of linguistic structures requires the integration of multiple perspectives. Collostructional and collocational analysis

(Stefanowitsch & Gries 2003; Evert 2005), for instance, tends to focus on the most frequent structures in a corpus. In contrast, productivity analysis (Baayen 1992; Plag, Dalton-Puffer & Baayen 1999; Evert 2004) examines the distribution of the least frequent structures, and is usually based on hapax legomena and other low-frequency types. Finally, dispersion is most informative on structures that occur at medium frequencies, where it displays the most variance (Gries 2008; Gries 2010). These might be overlooked by approaches that prioritize either high or low frequency structures.

Grammatical categories, while existing in a highly multidimensional category space, are still determined to a large degree by few very salient dimensions. This is one of the reasons why discrete models, feature models, rule-based models, and even prescriptivism have been so popular and still are to this day. In the usage-based literature, a lot of emphasis has been placed on non-obvious dimensions and very specific types of variation. This was a natural progression away from strict rule-based grammar. However, methodologically speaking a mixture of univariate, bivariate and multivariate approaches is necessary to capture the full complexity of language use.

4.2. Measures

4.2.1. Frequency

Token frequency is one of the most used, most discussed, and most basic measures in linguistics. Its importance, especially for the field of corpus linguistics, is well established. In short, frequency of occurrence is a crucial factor in the formation of linguistic categories, and frequency effects are ubiquitous across all linguistic fields (see Bybee 2006; Schmid 2010; Baayen 2010; Diessel 2016; Baayen, Milin & Ramscar 2016). After great initial success of using raw frequencies in corpus analyses and psycholinguistic studies, the measure has been criticized, mostly for its overuse and misuse (Stefanowitsch 2006; Baayen 2010; also see Gries 2022a). Frequency is often used as a measure for commonness. As such, its use is not always a reliable indicator (see also Brysbaert, Mandera & Keuleers 2018). In the following, I use token frequency mostly for two distinct purposes: (1) for exploration and representation, (2) as weighting factor in density calculations. In the latter case, it is always supplemented by other statistics.

Due to the Zipfian distribution of word frequencies, frequency is usually reported as log-frequency, both as a normalization technique in regression analysis or simply as a cosmetic transformation for the presentation of data. Due to the distributional anomalies of word frequency distributions, this may not be advisable in the context of regressions (O'Hara & Kotze 2010; Winter & Bürkner 2021). There is some evidence; however, that in principle, log-frequencies have a theoretical basis in the human perception of frequency. Absolute differences in frequency-based stimuli become exponentially less informative (cf. Kromer 2003). This is sometimes referred to as the

Weber-Fechner law (cf. Portugal & Svaiter 2011). This is true for word frequency effects as well (Tryk 1968; Szmrecsanyi 2006; Brysbaert, Mandera & Keuleers 2018). Kromer derived an adjusted frequency measure from this notion, referred to as U_r. Other log-based adjusted frequency measures exist, such as Savickỳ & Hlavácová (2002)'s adjusted frequencies based on Average Logarithmic Distance (ALD). These measures highlight slightly different aspects of frequency (see Figure 5.10).

4.2.2. Association

In experimental studies, statistical association has been linked to contingency and associative memory. It plays a crucial role in models such as the Competition Model (e.g., Ellis 2007a; Ellis 2007b; Fu & Li 2019). And it has also had a significant role in the acquisition of collocations in studies on second language learning studies. In Corpus Linguistics, the use of association is wide-spread thanks to the family of Collostruction Analyses (Stefanowitsch & Gries 2003; Stefanowitsch 2012; Gries 2019b). Strongly associated items within a context (most commonly words in a construction) show a high degree of semantic homogeneity, which makes association especially useful for the exploration of constructional or syntactic meaning. Furthermore, it is related to preemption and entrentchment which plays a crucial role in grammaticality judgments (cf. Stefanowitsch 2008; Stefanowitsch 2011b).

Finding the strongest collocates or collexemes is a common objective. Naturally, there has been intense discussion on appropriate indeces in the methodological literature (Evert 2005; Wulff 2008; Bartsch & Evert 2014; Evert et al. 2017; Uhrig, Evert & Proisl 2018; Gries 2019b; Gries 2022a). The index to be used for association strength is a point of contention (cf. Schmid & Küchenhoff 2013; Gries 2015a; Küchenhoff & Schmid 2015). In the original influential papers on 'Collostruction Analysis', the index used to measure association (Stefanowitsch & Gries 2003) was the p-values of the Fisher-Yates exact test. Recently, it has fallen out of favor. The reason for this is that the Fisher-Yates exact test is computationally expensive; with increasing corpus size, some values involved in the calculation become big enough for overflow[1] to become a practical issue[2]. The Fisher p-values are also strongly correlated with other indices. Today, the quasi standard for collocation and collostruction analysis is the log-likelihood test and the G^2 value (Dunning 1993). It has been shown to perform well across many applications and types of co-occurrence, especially collocation (Bartsch & Evert 2014). Schmid & Küchenhoff (2013) makes a case for odds ratios which is also taken up in Gries (2022b) because it does not conflate frequency and association. Both G^2 and Odds Ratio take into account all four cells of a contingency table, meaning that it takes into account frequencies of all constructions/contexts, frequencies of all elements, and overall corpus size. This

[1] *overflow* refers to the fact that default number formats in programming languages usually operate with a finite number of digits and replace larger values with 0 or placeholders

[2] 'arbitrary precision' libraries can operate with smaller numbers, but at the cost of even more computational resources and a much less straight forward implementation

makes the values robust against varying sample sizes, and is also interesting from a theoretical perspective, as it potentially captures negative preemption. Another advantage of odds ratio is that it allows better comparison between different sample sizes (Gries 2006).

The wealth of association measures can be explained by the interaction of at least 3 different dimensions: first, different types of structures that are under investigations (collocates, collexeme—construction, keyword—text, ...). It is evident that there is no one answer for a perfect index. In fact, the choice of index alone is actually of secondary importance. Ultimately, association measures always represent a conflation of properties (cf. Gries 2022b). The most important correlate is frequency, but indices are also sensitive to textual dispersion to varying degrees. This explains why so many measures are applied with such varying success. Different objects of investigation have differing sensitivity to frequency or dispersion. More specialized measures might have a stronger validity and exhibit less co-linearity with related indices. The choice of corpus and processing may also influence the appropriateness of a given index.

The association measures selected for the following studies are log-likelihood G^2 for univariate analysis since it includes information on frequency, and the log odds ratio, in particular the discounted variant. Discounted log odds ratios are computed by adding a constant of 0.5 to every observed frequency in a contingency table in order to prevent the log odds ratio from becoming infinite when the observed frequency is 0 (see Evert 2005: 86). Evert states that this has no theoretical background from statistical perspective. From a Corpus Linguistic perspective, it is not a totally implausible heuristic since no construction, even one that is deemed completely ungrammatical, has a non-zero probability of occurrence.

4.2.3. Dispersion

Lexical items vary in terms of how evenly they are distributed over the sample. Bursts of occurrences can skew perception of commonness since they inflate the frequency of an item. There is evidence that repetition in rapid succession loses its impact in terms of memory and entrenchment (Kromer 2003; Bybee 2006). Numerous measures have been proposed to capture the dispersion of items across corpora (Gries 2010; Gries 2020; Gries 2021; Gries 2022b; also cf. Kromer 2003; Savickỳ & Hlavácová 2002; Baayen 2001). While these studies heavily emphasize the importance of lexical dispersion, the measures proposed are still not in common use among corpus linguists. Uses are usually restricted to methodological discussion or demonstration (e.g., Egbert, Burch & Biber 2020). Part of the reason for this is the lack of accessible implementations, most of which are either closed source or hard to integrate into a research project. An implementation of the most commonly discussed measures is available in an R package that was developed alongside this dissertation (Rauhut 2022a).

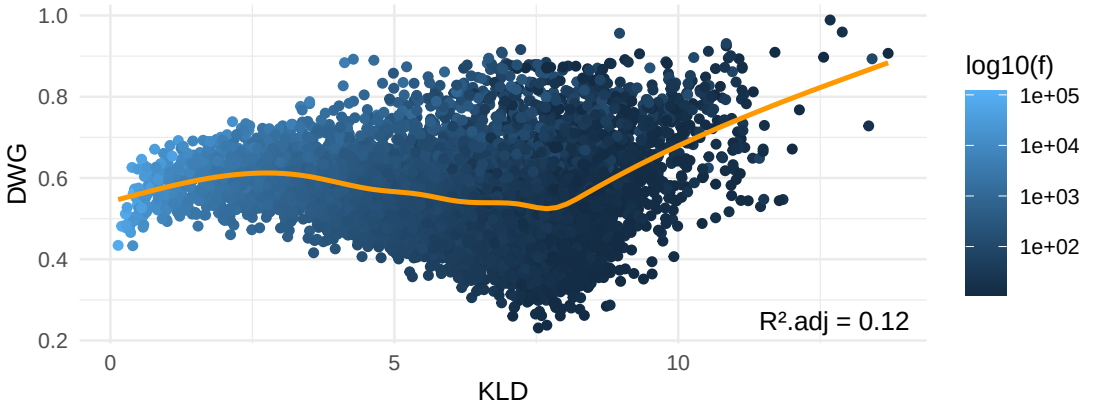


Figure 4.1.: Kullback-Leibler Divergence (KLD) versus Word Growth Dispersion (DWG)

As it is the case with association measures, most dispersion measures are strongly correlated with frequency, and therefore conflate those two dimensions. Deviation of Proportion (DP) and its normalized variant Normalized Deviation of Proportion (DP_{norm}) (Gries 2008: 414ff.; see also Gries 2010; Lijffijt & Gries 2012) can be used to find the most evenly dispersed items across different corpus parts. It, too, is correlated with frequency and is mostly designed to be used on its own. Lately, Gries (2022b) proposed an extension of this measure that attempts to eliminate the contribution of frequency. Kullback-Leibler Divergence (KLD), also known as the relative entropy, is an alternative to DP that is less correlated with frequency (Gries 2020: 104). In Gries (2022b), it is part of the highest ranking dispersion measures when tested on external data (2022b: 29).

There are two types of dispersion measure: distance-based and part-based (Gries 2008). Distance-based measures have been given little attention. Examples include ALD (Savický & Hlavácová 2002) and Word Growth Dispersion (DWG) (Rauhut 2021: 273; Zimmermann 2020). The aforementioned lack of implementation is an even stronger factor for the uncommonness of such measures in linguistic discourse. Gries claims that a significant downside of distance-based measures is that they are relatively intensive computationally (2020: 9). However, given the right implementation, this turns out as mostly unproblematic. [3]

Figure 4.1 shows the relationship between the KLD and DWG for adjectives (frequency > 10). The correlation is low with an adjusted R^2 of 0.12. It is also non-linear. Low-frequency items have increasingly bad dispersion scores in both measures. The blue shading indicates the log-frequency, and it visualizes that it correlates more strongly with KLD along the x-axis than with DWG. Both measures capture different aspects of

[3] implementations of dispersion measures available in the R package *occurR* (Rauhut (2022a)) show sufficient performance for corpora up to 1 Billion words; and there is a lot of room for optimization, especially concerning memory requirements

the dispersion of the data. Especially mid-frequency items can be both well-dispersed over corpus parts and clumped up within them (high KLD, low DWG), or occur in few corpus parts but in regular intervals (low KLD, high DWG).

4.2.4. Fixedness

To estimate the fixedness of a given item, I obtained prediction scores with BERT (*bert-large-uncased*, Devlin et al. 2019). The prediction score is the probability that the model assigns to the item in the given context as a result of masked language modeling. It ranges from 0 to 1 with 1 being the highest possible probability, i.e., a fully fixed structure. The window of prediction was restricted to the sentence. This decision was made to prevent the model from being able to use contextual information from the surrounding text, which would lead to vast improvements in prediction, and therefore an overestimation of the syntactic and lexical fixedness of the item. The average prediction score also captures surprise, i.e., the degree to which the item is unexpected in the given context.

4.3. Modeling prototype clusters

4.3.1. Distributional properties

A fundamental property of linguistic categories is that members of the category are distributed following a Zipf distribution (Zipf 1935). The distribution follows a power-law relationship in which the frequency of a given word is inversely proportional to its rank in the frequency table. The most frequent word in a corpus appears approximately twice as often as the second most frequent word, three times as often as the third, and so on. The same can be observed for more specific linguistic categories. Zipf distributions are ubiquitous even outside linguistics.

Many linguistic categories exhibit a Zipf distribution (Zipf 1935). Numerous mechanisms have been suggested to explain the emergence of Zipf distributions. More recently, cognitive mechanisms have been connected to the emergence of Zipf distributions in language. Language learning may be facilitated by Zipf distributions, especially the fact that one or few high frequency items in the head of the distribution can serve as a prototype for the category to be learned (Goldberg 2006). The lexical prototype(s) that fill construction slots tend to be learned first by language learners (Ellis & Fernando Ferreira-Junior 2009).

However, not all frequency distributions are Zipfian (cf. Piantadosi 2014). When observing the distribution of letters instead of words the frequencies decrease logarithmically (Kanter & Kessler 1995). Wulff et al. (2009) test whether tense-aspect categories follow a Zipf distribution as well, and note that perfect and progressive tense distributions

are less Zipfian than the present and past tense distributions. Unfortunately, they do not formally test this claim or investigate it any further. Piantadosi (2014) shows that verbs including modals create less Zipfian distributions. The reason for this is that high-frequency modals are uncharacteristically frequent, and low-frequency modals forming a cluster in the mid-range of the distribution (2014: Fig. 7f). This may be interpreted as evidence for modals belonging in a different category, which is in line with findings that modals are both less typical for the category of verb and have different distributional properties (see section 3.3).

These studies suggest that adherence to Zipf's law and deviations from it can serve as evidence for the existence of subcategories. In a simulation experiment, Lestrade (2017) demonstrates that word frequency distributions emerge from a combination of syntactic properties and semantic properties that are selected over multiple meaning dimensions. If this is the case, and the categorization of constructions is also subject to such selection, then it can be expected that spurious categories are less "Zipfy" than actual categories, and that this can be quantified and tested.

More specific mathematical models exist to describe word frequency distributions (Baayen 2001; Evert 2004). These distributions crucially depend on the lower frequency bands. In many corpus studies, those lower frequency bands are discarded, which restricts them to analyzing the most frequent items. Since type frequencies are of interest here, this approach is not feasible. If prototype structures are to be identified, it is important to include all items in the data.

4.3.2. Productivity

Productivity is related to word frequency distributions, and an important property of lexical items. Productivity in the linguistic sense is the possibility to use a structure to create novel expressions. This notion has been thoroughly investigated in the context of derivational morphology (Baayen 1992; Baayen 2001; Plag, Dalton-Puffer & Baayen 1999), but is also applicable to constructions of any kind. According to Goldberg (2006), productivity of constructions is constrained by semantic factors of potential lexemes to bind to a construction. Productivity in general is a gradual property, ranging from not productive at all (*-en* plural suffix, the combining form *-ceive*) to highly productive (*-ing* suffix, *-ize* suffix). Productivity of constructions also depends on the level of complexity. Seemingly unproductive classes can be locally productive if (more complex) constructions are treated equally as members (e.g., Stefanowitsch, Smirnova & Hüning 2020 on complex adpositions).

In the narrow sense, the term 'productivity' is used to refer to specific measures to quantify this property. The most common measures are the type-token and hapax-token ratios (Baayen 1992; Baayen 2001). Even though the following case studies are not concerned with productivity itself and productivity measures are not used explicitly, the concept is still inherently present in the methodology. As will be shown in the

following sections, estimating the number of types and accounting for their frequency is an integral part of the process.

4.3.3. Methodological challenges

Automatic processing of corpus data is prone to errors. Large data sets are full of noise caused by irregular spelling, typos, use of unusual characters, missing detectable word or sentence separation, text that doesn't represent actual language use (e.g., HTML code), and much more. Annotations are full of false positives if not carefully cleaned. Derived type frequencies are sensitive to low-frequency noise. The lowest frequency bands, especially hapax legomena, are often caused by tokenization errors. Another very common, and potentially more harmful type of noise is biases within lemmatization and PoS-tagging. Due to so-called out-of-dictionary errors, lemmatizers and taggers will often fail to recognize formal identity between complex and simple words if they are not frequent enough to appear in the language model's training data. If a PoS-tagger is reported to be 99% accurate, it is inevitable that the remaining 1% of mistagged items tend to exist in those lowest frequency bands due to out-of-dictionary items, and also in ambiguous cases that may be of particular interest.

Statistically speaking this causes 0- and 1-inflated distributions of co-occurrences proportions. For example, lemmatization may fail to correctly assign *wackiest* as form of *wacky* causing the word to be counted as a separate type. Since *wacky* is a mid-to-low frequency item, it will appear as a lemma that never occurs in its superlative form among a large amount of lemma for which this is genuinely the case. This both inflates the 0s of non-inflected types and the 1s of always inflected types. In collocation and collexeme studies, the focus is usually on the most strongly associated and most frequent types, which can be cleaned opportunistically. The rest of the distribution is usually ignored or even cut off at an arbitrary frequency threshold. The most commonly used measures Fisher p-values and G^2 correlate with frequency, which makes cases like the hypothetical lemmatization of *whacky*/*whackiest* hard to detect. This is where the use of odds ratios (simple ratios if it is merely for data cleanup) is beneficial since it ranks ratios of 0 and 1 at extreme ends of the scale. If a co-occurrence ratio is not perfectly 0 or 1, the chance for errors decreases substantially since lemmatizers and taggers are usually either categorically incorrect or accurate within the reported limits. This makes manual cleanup of lemmas a viable option even for moderately large corpora, such as the British National Corpus (BNC).

Figure 4.2 shows the effect that low-frequency items have on the ratio of plural to singular forms in the Brown corpus (Kučera & Francis 1967). The left panel shows the raw distribution with the arithmetic mean and median. Hapax legomena make up the largest group in a corpus and can only occur in either the singular or plural form. Therefore, their ratio is either 0 or 1. Dis legomena make up the second-largest group, and they can also have a ratio of 0.5. This additional outcome produces another

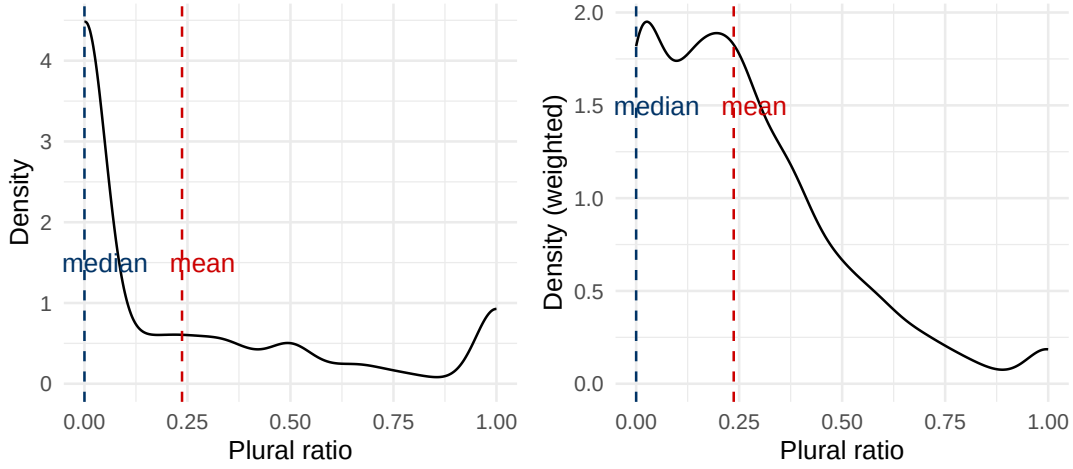


Figure 4.2.: Raw singular-plural ratios in the Brown corpus

peak at 0.5. They also introduce bias to the mean. The right panel shows the same distribution weighted by the overall frequency. There is now a peak at around 0.2, just below the mean. The picture still suggests that there are two groups of singular-only and plural-only items. These items, however, are both extremely strongly effected by categorical tagging decisions and errors.

This is especially problematic for any calculations involving type frequencies, such as type-token ratios, hapax-token ratios, and Large Number of Rare Events (LNRE) models (Evert 2004; Evert & Baroni 2007). In Rauhut (2022b), I attempted to mitigate this problem by estimating type ratios by resampling, which may make measures of type frequency more robust to noise since false positives, especially spelling variations, are not distributed evenly across corpus parts. Resampling can also be used effectively to improve other measures, like association and dispersion. It can also be used to provide measures of uncertainty.

Figure 4.3 shows the mean of the ratios across the frequency registers, starting with hapax legomena on the left. There is a slight tendency for low-frequency items to show a lower ratio of plural to singular forms. It is unclear whether this is due to bias or due to a real frequency effect. In either case, the distribution made up of at least three distributions, and a single measure of central tendency is not sufficient to describe the data. The solution to this is to use the modes. The mode is the most frequent value within a distribution. In this case, the most frequent value refers to a dense region of types that all have the same probability to be inflected. The mode here is slightly lower than the mean of the sample.

Annotation errors mostly cause a lower resolution in the search for distinctive clusters in the data since they distort the position of lemmas on a given scale. Lexemes might be accidentally pushed into the uncanny valley between category clusters. The variation

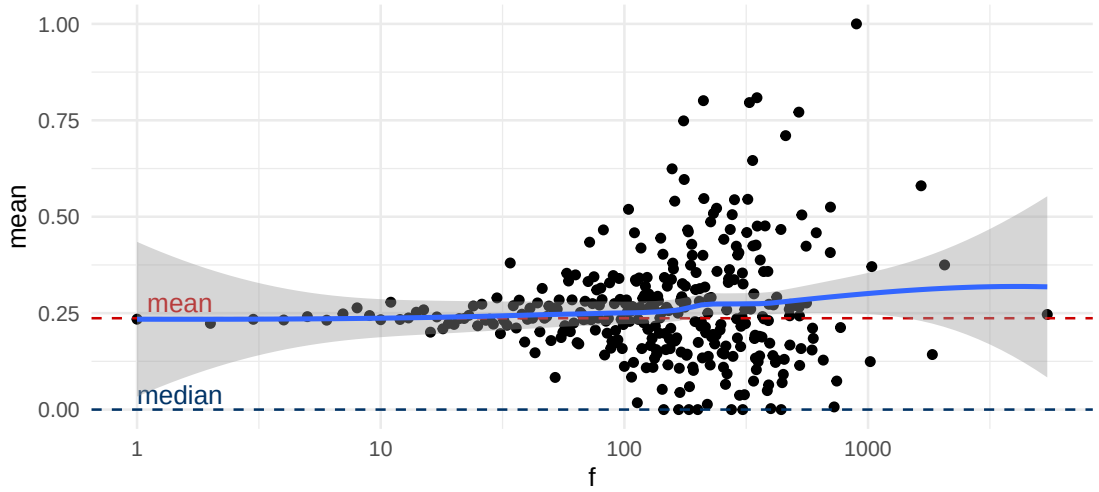


Figure 4.3.: Mean of singular-plural ratios in the Brown corpus across frequency bands

in noisy corpus data is lower than it appears. Nevertheless, the general tendency can still be uncovered given a large enough data set. Most noisy mistagging appears to happen rather regularly and across the entire lexicon and is mostly conditioned by frequency (see Figure 4.4).

Several decisions were made in the following case studies to reduce the noise by means of automatic filtering. However, only the final type lists were filtered. All calculations were carried out on the full corpus. All lemmas containing special characters were filtered, except for hyphenated tokens in the case of the *-ing* analysis. Proper nouns were also filtered due to their extremely high type frequency, irregularity, and low token frequency. These turned out to be the worst conditions for the dependency tagger, however, Named Entity Recognition (NER) tagging (Qi et al. 2020; based on Akbik et al. 2019) proved to be able to detect proper nouns and parts of proper nouns with a high degree of accuracy. In all following analyses, any type of numerals were filtered. Type frequencies of numerals are unusually high due to their inherent nature and due to the fact that spelling is also extremely variable (e.g., 4, *four*, 4*th*, *fourth*, etc.). In all steps of automatic text processing (most notably tokenization and lemmatization), numerals cause irregularities. Yet, they are part of the word class continuum, and vary in function and use along the noun-verb continuum (e.g., Corbett 1978; also see Mengden 2010: 262f.). The exclusion here is mostly due to technical reasons since type-frequency based quantification of numerals requires an additional theoretical framework and further processing steps. Nevertheless, numerals offer an interesting route for further research into the topic of word-class continuum.

4.3.4. Multivariate density

More researchers have recently emphasized the importance of multivariate modelling techniques, such as Generalized Linear Mixed Effects Model (GLMM) (Gries 2015b; Gries 2015c; Sönning & Werner 2021). In Rauhut (2021), I used a Generalized Additive Models for Location, Scale and Shape (GAMLSS) to explore the distribution of nouns and verbs and their conversion behavior by accounting for their frequency, dispersion, and association in the model. The approach taken in the following case studies is similar in that it is based on non-parametric smoothing to explore the distributional properties of the measures in question and to mitigate the non-normal, zero-inflated and highly skewed nature of corpus data. Since the object of investigation is the distribution of lexemes itself, I will mainly use KDE in univariate and bivariate form.

Kernel density estimation (cf. Silverman 1986; Sheather & Jones 1991) is a widely used non-parametric technique for estimating the probability density function of a random variable. This method involves creating a smooth estimate that represents the data, instead of using a mathematical equation that describes the shape of the distribution. This is useful when the data is not normally distributed or when there are outliers. It is also commonly used to detect multi-modality, i.e., the existence of multiple modes/peaks in the data set. One important extension of kernel density estimation is the use of weights, which allows for the incorporation of prior information or 'expert knowledge' about the data. Weighted kernel density estimation can be particularly useful when dealing with data that is sparse, skewed or has outliers. The lexicon is full of outliers and word frequency distributions often have long extremely sparse tails even without annotation artifacts. Another important extension of kernel density estimation is its multivariate variant (cf. Scott 1992; Chacon & Duong 2018), which is used to estimate the probability density function of a pair (bivariate) or a set (multivariate) of variables. Overall, kernel density estimation and its extensions offer tools for exploring the underlying structure of complex data sets.

The first case study on adjectives in Section 5 will serve to explore the use of different frequency weights, including the adjusted frequencies ALD (distance-based) and U_r (part-based) mentioned in Section 4.2.3. One of the biggest challenges in the interpretation of the data is the reduction in dimensionality. Assessing multidimensional distributions with Zipf-distributed frequency data makes visual and numerical interpretation difficult. Since frequency in a corpus does not measure commonness directly and is dependent on a variety of variables, I essentially use it as an indicator for signal strength. That means low-frequency items contribute less to the final model than high-frequency items. In a sense, the weighted densities represent a multivariate type-token relationship. Using adjusted frequencies includes textual dispersion into the calculation. The fact that many dispersion measures are correlated with frequency—normally considered a disadvantage—is helpful in this case and also has some theoretical plausibility. For example, U_r essentially applies a penalty that is proportional to the number of repetitions within the same text. Due to their statistical

properties, some other dispersion and association measures are not as useful as weights.

4.4. Corpora and materials

4.4.1. Data set

For the following corpus studies, I mainly used data from the BNC (2007). Both Bartsch & Evert (2014) and Uhrig, Evert & Proisl (2018) suggest that corpora that are carefully balanced perform better in tasks like collocation candidate extraction than more noisy data sets such as web corpora. In Rauhut (2022b), I found the BNC's composition and text ordering to be beneficial for the identification of lexical clusters. Distance-based dispersion measures are expected to perform better with meaningful text ordering.

The original word and sentence tokenization was preserved. However, additional tokenization of hyphenated tokens was performed. For rarer structures, additional qualitative data was drawn from the Corpus of Contemporary American English (COCA) (Davies 2008). The results were compared across other corpora, including the British National Corpus 2014 (spoken) (BNC2014), and Brown Corpus (Kučera & Francis 1967; Love et al. 2017).

The data sets were annotated with Universal Dependencies (UD) tags (Marneffe et al. 2021) using the *stanza* library (Qi et al. 2020). The UD project offers a promising framework to annotate corpora with cross-linguistic comparison in mind. The available annotations closely match the functional dimensions that are of interest here. Its theoretical aspiration is mostly congruent with the background of this thesis. Furthermore, a universal annotation scheme holds the door open for follow-up studies in other languages. The annotated corpora were indexed and encoded with the *Corpus Workbench* (Evert & Hardie 2011).

4.4.2. Note on phonetic and orthographic form

The phonological level is not directly observable in the data that is used in the following studies, so it is not an actual annotation level. That means, the default annotation level, by which formal identity is determined, is derived from orthographic tokens. This is, of course, far from ideal, especially when discussing potential homonymy of affixes. Homonymy requires an identity of form for which graphemes are not the best indicator. While phonological annotation is possible, it is questionable whether it even makes sense to assign some sort of idealized phonological form to an orthographic token. Such an idealized form would only slightly shift the problem of formal identity and complicate it. It would also not reflect phonetic variation or variation through variety or idiolect. Phonological form, however, has a special status in linguistic categorization since it

Table 4.1.: Top 10 orthographic trigrams in the British National Corpus

Rank	TRIGRAM	FREQUENCY
1	the	275621
2	ing	117934
3	and	105159
4	hat	70294
5	her	69160
6	ion	60928
7	ent	56194
8	ere	54799
9	tha	54086
10	you	49971

is usually considered to have an arbitrary relationship to meaning and function. As long as it can be ascertained that orthographic idiosyncrasies do not cause systematic biases, it is therefore reasonable to use orthographic tokens as the default annotation level. Homography causes systematic skew for the affected types, but there are no clear solutions to this problem to date.

In the case of the *-ing* suffix, there are very few homographs or homophones that have accidental similarity. Most words with the phoneme sequence [ɪŋ] that are clearly unrelated to the *-ing* suffix have monosyllabic stems. Within English, the phoneme [ŋ] has a very low type frequency and typically ranks among the rarest consonants (Hayden 1950; Mines, Hanson & Shoup 1978). The sound [ŋ] is not contrastive in most European languages and if so it is phonotactically restricted (Anderson 2013). It can be considered typologically marked compared to other nasal consonants. However, in English, the *-ing* morphemes cause a very high token frequency. <ing> is the second most frequent orthographic trigram in the BNC after <the> and before <and>.

Morphological relatedness may not strictly be required for association between sound sequences (cf. Bergen 2004; also Longtin, Segui & Hallé 2003). This is crucial in the discussion of morphological homonymy. Some research suggests that even two full homonyms can over time be affected by their formal identity. At least it could act as a mediating effect. Interestingly, this can be a further justification for the use of orthographic tokens as the default annotation level, and may explain why corpus linguistic studies have been successful despite this inherent flaw of the method.

4.4.3. Note on morphological form

Since the phenomena in question are themselves morphemes, the case studies are concerned with complex words. There is only limited morphological variation when it comes to those forms. Inflection plays only a minor role. Most variation comes from compounding or derivation. Word-class changing derivation is naturally of most interest. Nominalizations are often less 'nouny' than typical nouns (Mackenzie 1985; Bekaert & Enghels 2019; Maekelberghe, Fonteyn & Heyvaert 2021). In the same context, verbal and nominal gerunds are often described on a functional cline from nominal to verbal.

Neither the original BNC lemmatization, nor *stanza*'s lemmatization, are designed to include derivational forms into the lemma annotation. In an attempt to find and assign derivational forms to their base lemma, I used both *Wordpiece* tokenization through *BERT* (Devlin et al. 2019), and morphological segmentation through *Morfessor* (Virpioja et al. 2013). Unfortunately, neither method yielded satisfactory results. Many complex nouns were not recognized as containing multiple pieces, and in many cases even common derivational suffixes were not separated correctly. In the case of *Wordpiece* annotation, very common complex words are not split at all, mid-frequency words perform best, and low-frequency words often get split off in the wrong place, especially with low-frequency suffixes. However, the performance of the *-ing* suffix appears to be the worst despite its high frequency. The split into morphemic or non-morphemic parts is illustrated in figure 4.4.

Interestingly, the fact that high-frequency words are not further tokenized is congruent with the observation that complex high-frequency items are stored in the lexicon as a single item alongside its parts (Bybee 2006; Bybee 2010). This may be one of the reasons the language models perform well in practical application. In a way, the training methods mimic real language learning in the sense that it is exemplar-based (based on occurrences in the training set) and handles new data based on something akin to analogical derivation, only that the training data is not as rich in experiential cues as real language. Orthographic and allomorph variation pose another problem. More research is needed to fine-tune the models to the specific needs of linguistic research. Even if unfortunate for the core research question of this study, this finding is interesting in its own right.

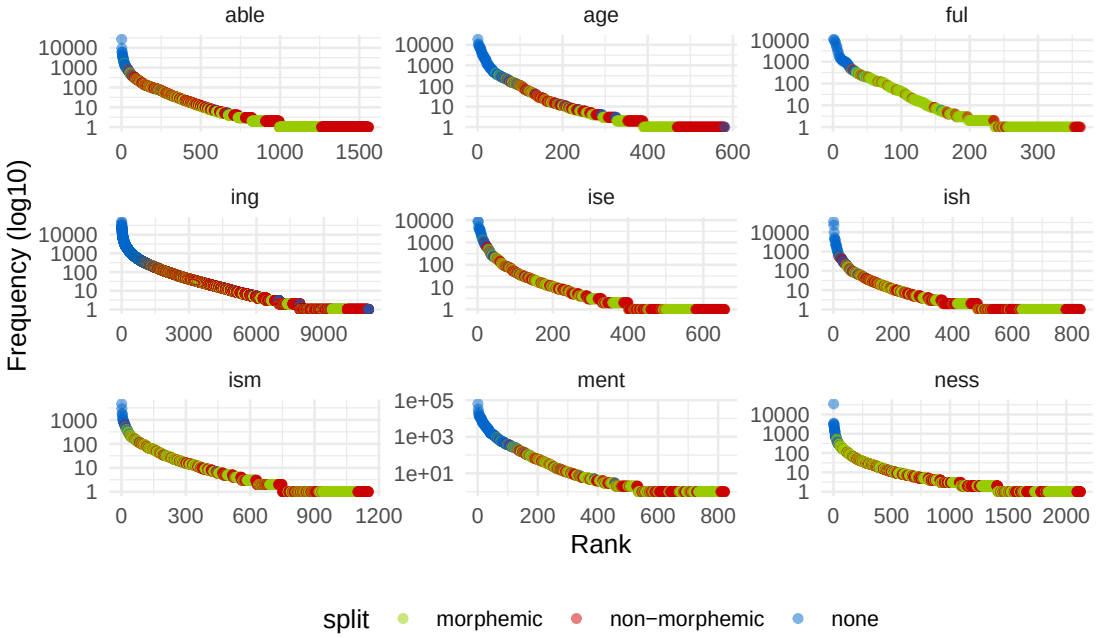


Figure 4.4.: Wordpiece tokenization accuracy detecting derivational suffixes

5. Adjectives and their subclasses

5.1. Overview

The following chapter explores the statistical distribution of adjectives in the search for subtypes that manifest as type clusters of lexemes. At the same time it lays the methodological foundation for the following case studies in Chapters 6 and 7 and serve as a benchmark.

English adjectives are mostly classified into inflecting versus non-inflecting, attributive versus predicative, and gradable versus non-gradable (cf. Huddleston & Pullum 2002). Different adjectives show different preferences for the related constructions. Croft (2022) puts strong emphasis on the idea that, "different constructions define different distributions" (2022: 2). He claims that the distribution of adjectives is ignored when a word class of adjectives is defined as only one category (2022: 11). For illustration, he provides the following examples for four main constructions usually associated with adjectives. He claims that these four constructions define 4 different overlapping word classes.

Croft (2022: 11):

(15) Modification of a referent:

a. *This insect is alive.*
b. **an alive insect*

(16) Predication with a copula:

a. *An entire chapter is devoted to this problem.*
b. **This chapter is entire.*

(17) Degree inflections:

a. *tall-er*, *tall-est*
b. **intelligent-er*, **intelligent-est*

(18) Degree modifiers:

a. *a very tall tree*
b. **a very even number*

However, the data presented only superficially supports such a claim. The grammaticality judgment implied in these examples has to be relativized as probabilistic once empirical data is taken into account. Three of the four structures can be found easily in a sufficiently large data set:

(19) – and making it **an alive** part in the alive Ireland (COCA: ACAD)
(20) The financial sector 's move **is almost entire** (BNC: EAK)
(21) Cold-blooded creatures would need **very even** temperatures if their intelligence was not to switch on and off (BNC: C9A)

Only *intelligenter* and similar cases seem absent. It can be speculated, however, that spontaneous analogical formation occurs on occasion, be it for humorous reasons or due to transfer. It is important to note that it is the only example in this list that is based on morphological form.

The idea that English has separate classes of 'gradable' and 'absolute' adjectives has its origin in English prescriptivism (Huddleston & Pullum 2002: 530), even though for very different reasons. Whether there actually is a difference between gradable and non-gradable adjectives has to be determined empirically. Huddleston & Pullum (2002) note that gradability applies to senses and not lexemes (2002: 531). *very unique* and *very ideal* are valid gradable uses of words that are traditionally assumed non-gradable. Moving the idea of language-specific word class onto construction level does not answer the question of whether there are systematic differences in the lexicon with respect to the selection of adjectives. If words like *unique* and *ideal* can regularly derive senses that are gradable, and actually do so in language use, then there is no real reason to distinguish gradable from absolute adjectives.

If any member of these four classes has a non-zero possibility of being used in any of the other classes, then the overlap happens for every lexeme and not across the lexicon. If similar enough, four interacting univariate probability distributions can still be described by one multivariate distribution and perceived as one population. A multidimensional conceptual space (Croft 2001: 93), in theory, accomodates distributional word class categories just fine, despite the individual idiosyncrasies.

Moreover, these classes do not necessarily interact. There is no obvious correlation between the ability to be inflected and the ability to be used in predicative position. The reason that some adjectives in English inflect and not others is not conceptually motivated, but a result of historical processes and dependent on their frequency and length (which is indirectly correlated with frequency as well since frequent words tend to be shorter). It should be rather uncontroversial that there is no plausible semantic property warranting a distinction between short adjectives and long adjectives (other than perhaps vague tendencies connected to iconicity, cf. Section 2.3.2).

5.2. Preparation

The population of lemmas that makes up the data set for this chapter was prepared by annotating the BNC with the values and measures listed in Chapter 4. With the help of prediction scores from BERT and by relying on the heuristic that tagging errors are mostly categorical (cf. Section 4.3.3), the lemmatization was semi-manually cleaned. Word lists with manually corrected lemmas will be made available alongside the rest of code. The final sample contains 25873 lemmas representing 4708594 tokens. No frequency thresholds were used.

Any adjective lemmas containing non-alphabetical characters were removed. For inflected forms, PoS tags were used for initial identification. Suppletive forms of *good* and *bad* were annotated according to their suppletion pattern, i.e., *best* was annotated as superlative form of *good*. Differences within association patterns between lemmas and their superlative and comparative forms were negligible, and thus reduced to one binary category of INFLECTED versus UNINFLECTED.

For the second annotation level, a combination of UD dependency relations and PoS tags was used. Attributive uses were determined by selecting tokens tagged as adjectives whose dependency relation was tagged as AMOD and that are headed by a token tagged as noun, including proper nouns.[1] Predicative uses were determined by selecting occurrences tagged with the dependency relation of COP (copula) and that are headed by an adjective. In both cases, the distance between the head and the modifier was measured in tokens. However, the distance of the dependent was mostly directly adjacent and larger distances seemed to have a negligible effect, so the variable was dropped from the analysis.

Dispersion and association measures were calculated for the entire lemma and separately for each of the distributional categories.

Table 5.1 summarizes the 'tuple' of values that was used.

Table 5.1.: Variables and measures used

Type	Measure
Observation	Lemma (manually corrected)
Variable(s)	based on PoS and/or UD dependency tags
Unit	Text (text ID)
Association	Odds ratio$_{\text{discounted}}$
Part-based dispersion	KLD
Distance-based dispersion	DWG
Frequency	f (raw)
	$f_{\log}$

[1] cf. B.1 for reference and tag set

Type	Measure
Frequency dispersion-adjusted (per part)	Kromer's U_r
Frequency dispersion-adjusted (distance)	f_{ALD}

5.3. Morphosyntactic subclasses

5.3.1. Inflection

Adjectives are commonly considered to have two inflectional suffixes, *-er* and *-est*. There is little irregularity to speak of, except for adjective-like quantifiers *much*/*more*/*most* and *little*/*less*/*least* which could be regarded as suppletive forms. *more* and *most* are themselves used as grammatical markers to grade the vast majority of regular adjectives. These six special lexemes do occur in some contexts where adjectives also do, even inflection for *much* and *little*, if suppletion is included definitionally via allomorphy. Additionally, *little* exists with the regular inflectional suffixes when used in a more specific sense of size rather than quantity. *much* arguably lacks this sense, and also the form **mucher*. As we will see in the following sections, they are among the least typical, but most frequent forms in the data set. Due to the low number of 6 types, they do not change the overall picture and were kept as a reference guide.

Figure 5.1 shows the association of adjectives to the inflectional suffixes *-er* and *-est* against distance-based dispersion. The odds ratios (x-axis) range from -2 (repelled) to 2 (attracted), with the center point at slightly above 0 (dotted line), since the discarded version is used. DWG (y-axis) ranges from 0 to 1, but the bulk of the distribution is between 0.4 (evenly dispersed) and 0.8 ('clumpy'). Only tokens such as punctuation marks and the definite article appear at such regular intervals that they can score DWGs of much smaller than 0.3. The KLD and absolute frequencies are symbolized as size and color, respectively. Without those two variables, most of the data points would be a diffuse mass given the Zipfian distribution of the data. There is simply not enough information about the association and 'clumpiness' of most of the lemmas in the data set. At least visually, two clearly distinct groups emerge once those variables are added. In later sections, we will see that this is also quantifiable.

The y-axis shows the distance-based dispersion of the adjectives. Some adjectives like *social* are extremely common in the BNC (34910 tokens). Compared to *intelligent* (only 1737), this is a huge difference, and unexpected if it is 'commonness' that is intended to be measured. If corrected for short bursts of occurrences, *social* moves to the margins of the distribution. This now multidimensional picture allows the quantification of

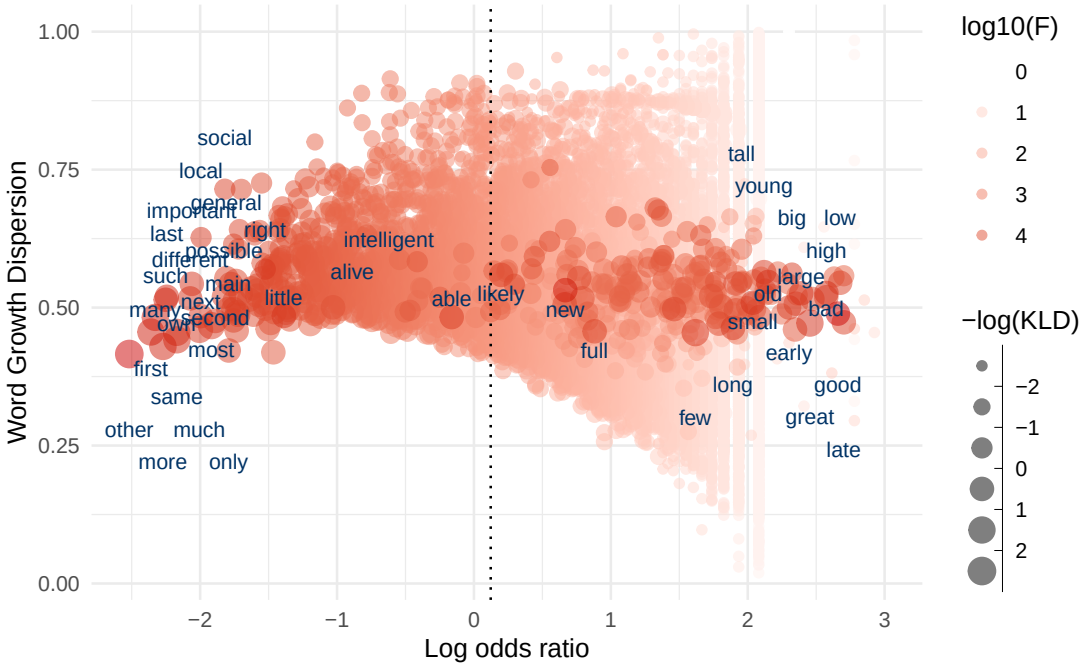


Figure 5.1.: Association of adjective lemmas to the inflectional suffixes *-er* and *-est* (x-axis) against distance-based dispersion (y-axis), absolute log frequency (color gradient), and part-based dispersion (size). The text labels are fixed along the x-axis; their vertical position only reflects relative order. The dotted line represents the center point of the discarded log odds ratios. A negative log transform of *KLD* is done to improve scaling.

social as less typical than *possible* with regard to their membership of the UNINFLECTED group, even though it is more frequent, and more strongly repelled to the inflectional suffixes. *intelligent*, due to its relatively low frequency, does not appear as central to the cluster of non-inflecting adjectives. Later, however, we will see that the left cluster, in fact, forms more of a ridge rather than a circle, within which *intelligent* is located (cf. Figure 5.9).

As expected, the two groups of inflecting and non-inflecting adjectives are split from this perspective. Typical gradable adjectives dominate the center of the right cluster (*old*, *small*, *long*). Inflecting but less gradable adjectives (*new*, *full*) are still attracted, but in the periphery of the cluster. Something worth noting is that among the most common and most repelled items, there are some monosyllabic adjectives, such as *main*, *right*, *real*, that should be able to inflect. In fact, with the exception of *main*, their inflectional forms are attested in the corpus. A similar case is *little*. These lemmas share common features with adjective-like quantifying expressions such as *same*, *only*, *whole*. All of these examples, be it quantifiers or more typical adjectives, are extensively used in determiner constructions such as *the main thing*, *the right thing*, *the real deal*,

a little. Another group in the mix is ordinals such as *first*, *second* and *next*. These two groups are more grammatical in nature, which allows them to be used more flexibly. It causes them to be more common and evenly distributed, which makes them break out vertically from the rest of the point cloud at around [-2, 0.4] (see also regression curve dipping at these coordinates in Figure 5.9). This can be interpreted as evidence for a cline from adjective to quantifier, i.e., from lexical to grammatical. Adjectives may lose their ability to inflect when they become too grammatical in general as a result of grammaticalization.

Many adjectives that are statistically repelled from the inflectional suffixes have non-accidental occurrences with them:

(22) This is the world 's **arcanest** grove (BNC: J0R)
(23) I 'm not going to exploit the **basest** aspects of sexuality (BNC: CGC)
(24) You are a thousand times a **properer** man (BNC: A06)
(25) They are all very hardy and may be planted in the **openest** places (BNC: ALU)
(26) I went **oftener** to Uncle Geordie 's by that time (BNC: BN1)

An interesting type of these exceptions appears to be in coordination and otherwise symmetrical constructions where the unexpected inflectional form is triggered by a more common one. Unfortunately these forms are too rare for quantitative assessment.

(27) The longer we live the **oftener** we find how wrong we can be (BNC: FTX)

In a next step, I added 1-dimensional densities in both the raw version, and two weighted versions, using raw frequency and U_r. As can be seen in Figure 5.2, the unweighted density contains little information about the underlying distribution since most of the probability mass is in the lowest frequency registers. It might be counter-intuitive that most low-frequency items are positively associated with the inflectional suffixes. One way of thinking about this is that the entire sample is biased towards uses with inflectional suffixes. This is due to the fact that the inflectional suffixes tend to be used in conjunction with adjectives of very high-frequency. Therefore, rare words, for which there is less information available, are by default expected to occur with inflection. This is also plausible from a usage-based perspective of learning. In analogy to rare words in the data set, new words that a learner encounters are likely to be categorized based on existing information about the most typical members of the associated category and used analogously. This is ultimately connected to phenomena like morphological levelling. In the case of inflection, there are (morpho-)phonological factors that play a role in the choice of comparative construction; therefore, this simplified picture is incomplete.

The weighted density curves, on the other hand, convey a different picture. Assigning weights to the density calculations to specify the relative importance of each lemma accounts for the fact that more frequent lemmas in the corpus are more informative than others. Accounting for raw frequencies (green line) produces 3 clear peaks,

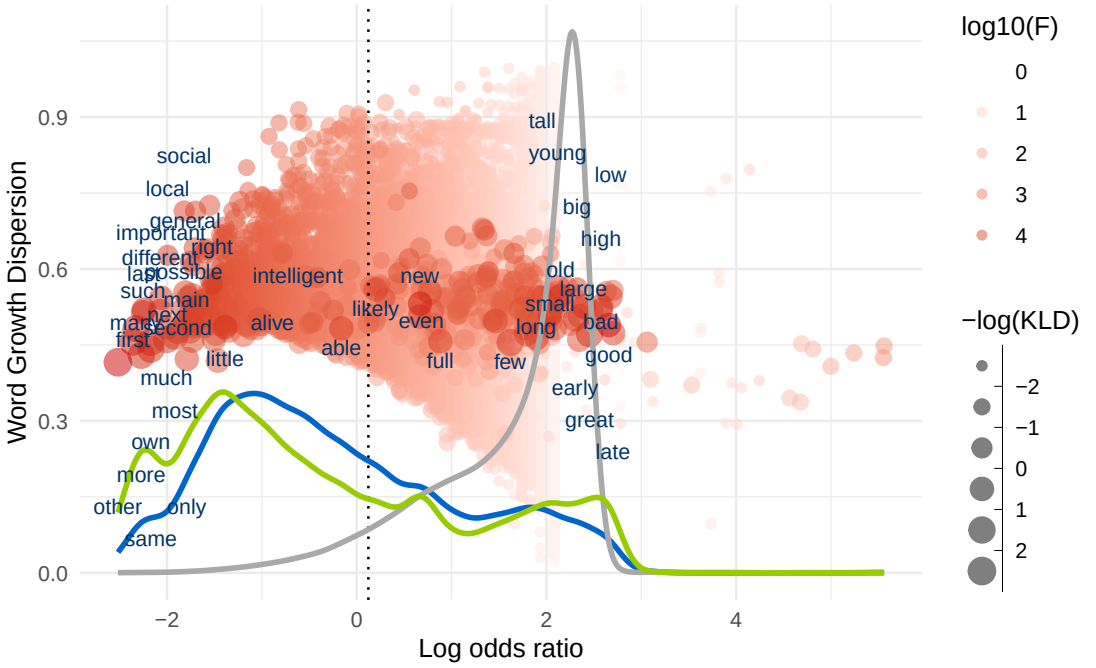


Figure 5.2.: Association ADJ to *-er*/*-est* (x-axis) distance-based dispersion (y-axis), absolute log frequency (color), part-based dispersion (size). Same as in 5.1, but with added 1-dimensional density: simple type density (gray), weighted by raw frequency (green), weighted by dispersion-adjusted frequency (U_r, blue).

suggesting 4 different overlapping distributions. The main peak is located at around x=-1.75, and contains most of the high-to-mid-frequency, and moderately evenly dispersed adjectives that are statistically repelled to the inflectional suffixes. A second dense area is located at around x=2.5 containing those adjectives that most commonly inflect. Two more smaller peaks are formed around high frequency items. One of them (x=0.5) is caused by items such as *new*, *full*, and *likely*, that are almost neutral in association toward the inflectional suffixes. Some very frequent quantifiers (*many*, *such*) and ordinal-like modifiers (*first*, *last*) potentially cluster separately on the leftmost side of the distribution at approximately x=-2.5. After correcting for repetitions within the same corpus parts (blue density curve), those two clusters appear much more faint. Those modifiers belong to classes distinct from adjectives even though similar in function. Therefore, the answer to the question of how many different distributions are present in the data strongly depends on what properties of the lemmas are considered.

Log odds ratios have an asymptotic normal distribution (Agresti 2002: 581). However, all three density curves are non-normal, i.e., even the dispersion-adjusted density curve hints at the multi-modal nature of the distribution, and possibly a mix of prototype categories. This makes sense if there is no firm conceptual motivation for the

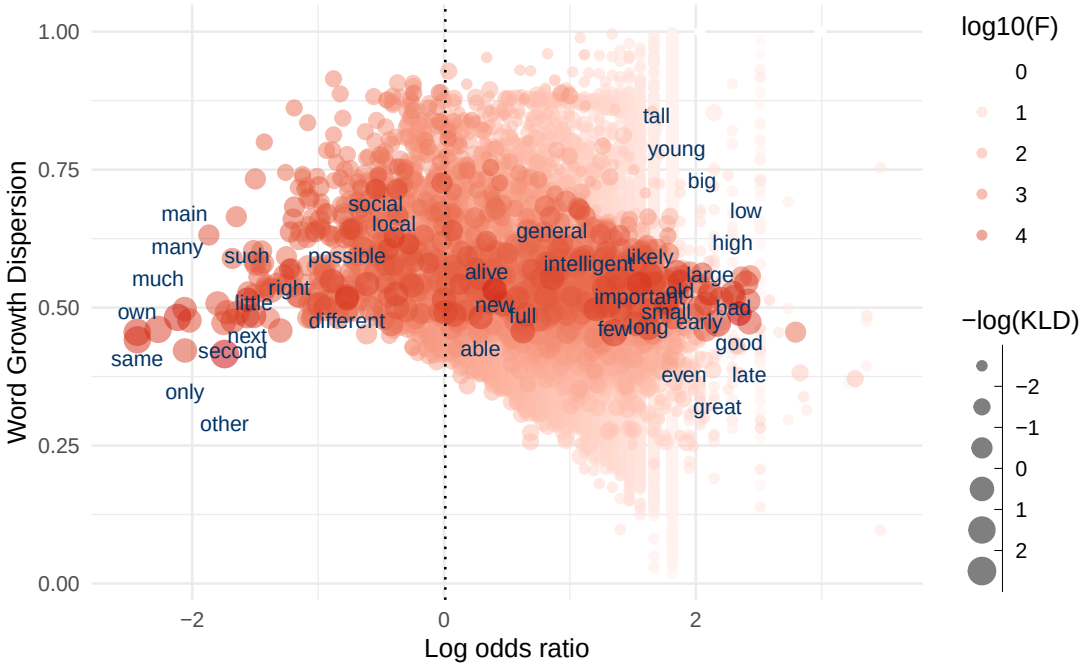


Figure 5.3.: Association ADJ to *-er*/*-est* (x-axis) distance-based dispersion (y-axis), absolute log frequency (color), part-based dispersion (size).

inflectional behavior. If phonologically conditioned, the relationship is likely arbitrary and the different modes are indicative of a third variable, such as general gradability, to which I will turn in the next section.

5.3.2. Gradability

Once the analytic ways to form comparatives and superlatives are taken into account, the picture shifts. After this conflation, the opposition is more conceptual in nature by contrasting gradable versus non-gradable properties. Figure 5.2 can be imagined as a projection of Figure 5.1 from the side at a 90° angle in a 3rd dimension.

Quantifying expressions like *much*, *little*, and *only* are once again the most repelled items. Most of the top 50 most frequent adjectives are split between repelled and associated with few items near the neutral line. Adjectives like *big*, *old* and *small*, which are strongly associated to the inflectional form are also strongly associated to being graded in general. The distribution is much more uniform from this perspective. The group of non-gradable adjectives overlap significantly with the group of gradable adjectives. Items like *full*, *likely*, and *new* are in the middle of the two extremes. Those are in the uncanny valley of gradability because they have an almost equal share of gradable and non-gradable uses. *alive* in its biological sense is non-gradable, but is

also frequently used metaphorically, referring to certain aspects of being biologically alive that are gradable.

(28) It was then six thirty and I was beginning to feel a little **more alive** (BNC: FAP)
(29) The tongue was impossibly extended , pointed and wet and **more alive** than the rest of the thing (BNC: ALJ)

It is tempting to disambiguate these uses, after which a strong split between gradable and non-gradable adjectives would likely appear. Despite it being methodologically unfeasible, it is also an inherent part of the variation of the form *alive*. Furthermore, if all adjectives were to show such a 50-50 split between gradable and non-gradable uses, the distinction would disappear. The pattern in the data is interesting since it shows that such cases might blur the line between gradable and non-gradable adjectives, making the class more coherent. Additionally, the example of *alive* is not restricted to non-gradable adjectives. People can also *feel younger*, referring not to age, but energy levels, and *seem taller*, which may refer to perceived attitude or confidence rather than height. Conversely, gradable adjectives can be framed as non-gradable, i.e., discretized. *old people* and *young people* are often understood as belonging to certain discrete age groups.

Occasionally, adjectives are used in both inflected form and analytic form. *able* for example occurs as both *abler* and *more able*. The former, however, is rather rare (3 in the BNC; 8 in the COCA).

(30) a number of people a great deal **abler** than ourselves have worked on the problem (BNC: CEG)
(31) Mr Kinnock is surrounded by men who are **more able** and winning than he is (BNC: A8K)

This overlap potentially boosts the association to comparative and superlative forms for individual lemmas, but at these low frequencies it is likely negligible.

Figure 5.4 again shows the three different densities as described above. Accounting for raw frequencies produces a very diffuse picture with numerous modes. The reason for this is that the highest frequency adjectives that contribute most are scattered along the x-axis. After accounting for repetition within texts, only one clear main peak remains. The distribution has a strong skew and a prominent 'shoulder' at around x=-0.4, which is where the non-gradable adjectives would be expected to cluster. From left to right, we are presented with the following groups

1. x=-1.7: quantifiers, determiner-like items, ordinals—*same*, *own*, *much*, *only*, *first*
2. x=-0.4: non-gradable adjectives—*possible*, *particular*, *local*, *different*
3. x=1.1: mixed gradable adjectives—*few*, *intelligent*, *long*, *small*, *old*
4. x=2.2: only inflecting adjectives—*great*, *high*, *low*, *large*

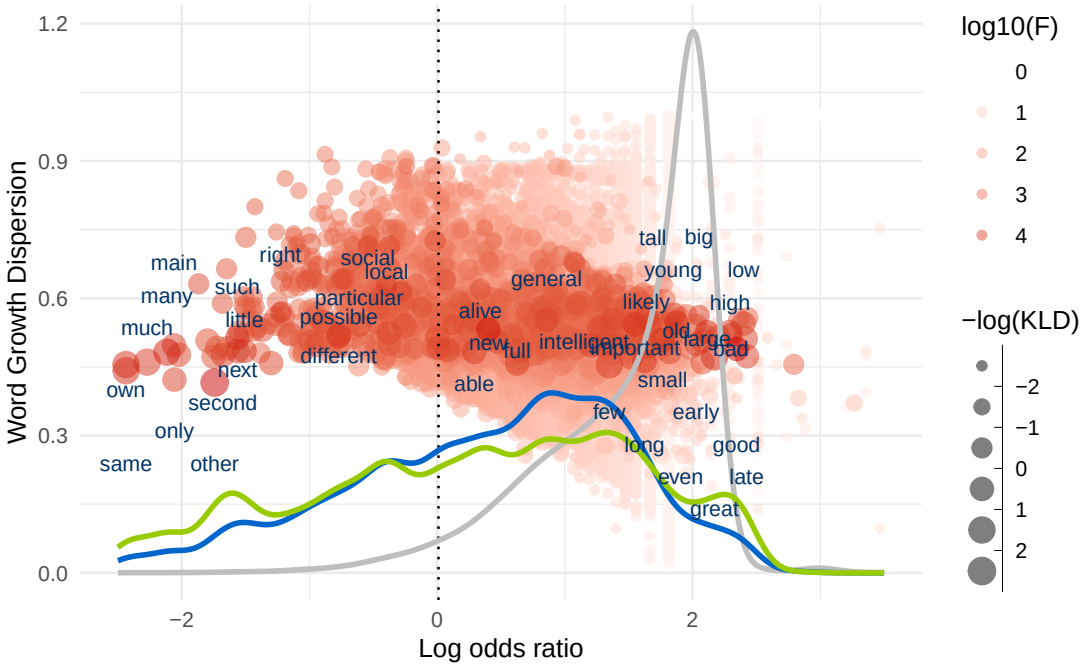


Figure 5.4.: Association ADJ to comparative and superlative uses (x-axis) distance-based dispersion (y-axis), absolute log frequency (color), part-based dispersion (size). Same as 5.3, but with added 1-dimensional density: simple type density (gray), weighted with raw frequency (green), weighted with dispersion-adjusted frequency (U_r, blue).

Strictly speaking, there is no valley between the gradable and non-gradable adjectives. It can be considered a feature of non-gradable adjectives to be used in a gradable context, e.g. via conventional metaphor. The group of short, frequent and inflecting adjectives causes a small shoulder on the right of the distribution. Inflecting adjectives tend to be a bit more strongly associated to grading than non-inflecting adjectives. In summary, the gradable-non-gradable distinction does not show a strong separation and is rather continuously shaped along these focal groups: quantifiers/ordinals < non-gradable < mixed < gradable non-inflecting < gradable inflecting.

5.3.3. Copula construction

The next construction to be investigated is the copula construction in order to determine whether predicative adjectives form a group distinct from attributive adjectives.

The results can be seen in Figure 5.5. Clustering behavior is not immediately apparent from this perspective. The overall distribution is much denser and most adjectives are attracted to copula uses. This alone is interesting due to the expected dispreference

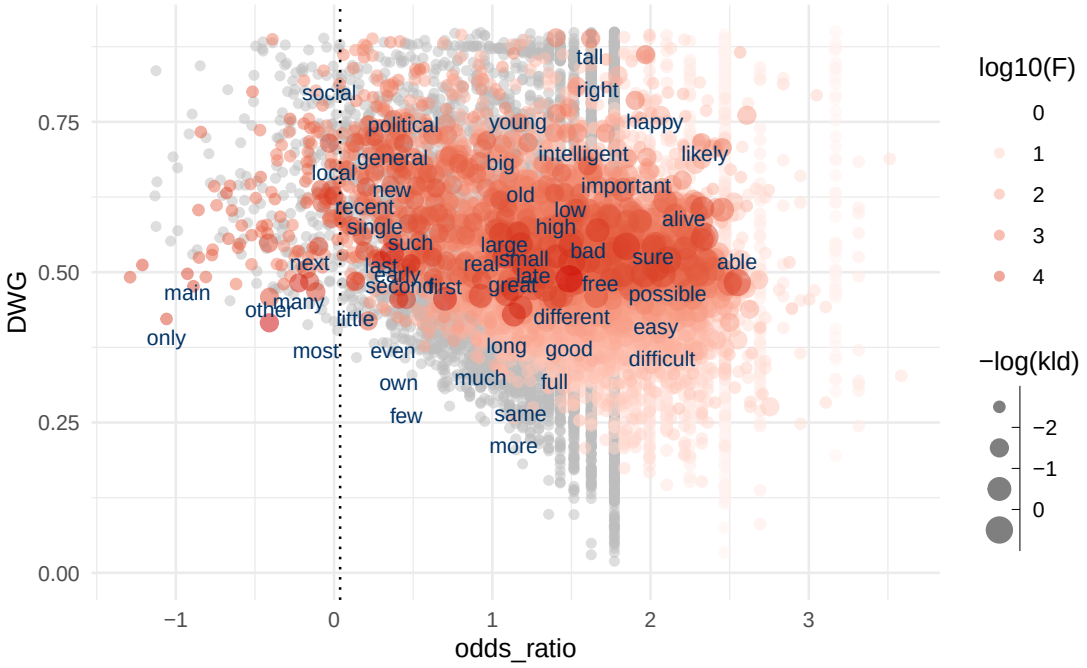


Figure 5.5.: Association ADJ to COP (x-axis) distance-based dispersion (y-axis), absolute log frequency (color), part-based dispersion (size).

of some adjectives. The only repelled lemmas are the group of quantifers and ordinals that were already distinct from the other perspectives. Adjectives that are almost at the center line include *social*, *political*, and *recent*.

(32) will your local brewery **be next** ? (BNC: A13)
(33) the underlying crisis **was social** and economic (BNC: A7Y).

The majority of copula uses of *recent* were actually correlated with the comparative *more recent*. Therefore, there is some interaction between the features for some lemmas.

Figures 5.6 and 5.6 are a bit different from previous figures in that grey points symbolize words that are not attested in the copula construction. In the case of the copula few highly frequent and well dispersed adjectives never occurred in the copula construction.

The three density curves in figure 5.6 all represent U_r-weighted densities. The darkgrey line represents the distribution of adjectives that do not occur in the construction. There is a clear peak at x=2 which is a dense region of very low-frequency lemmas. Low-frequency items that did not have a chance to show significant attraction or repulsion can still show relative frequency of copula constructions as a whole. The density curve indirectly represents this split. The density of the peak at 2 is almost as high as the

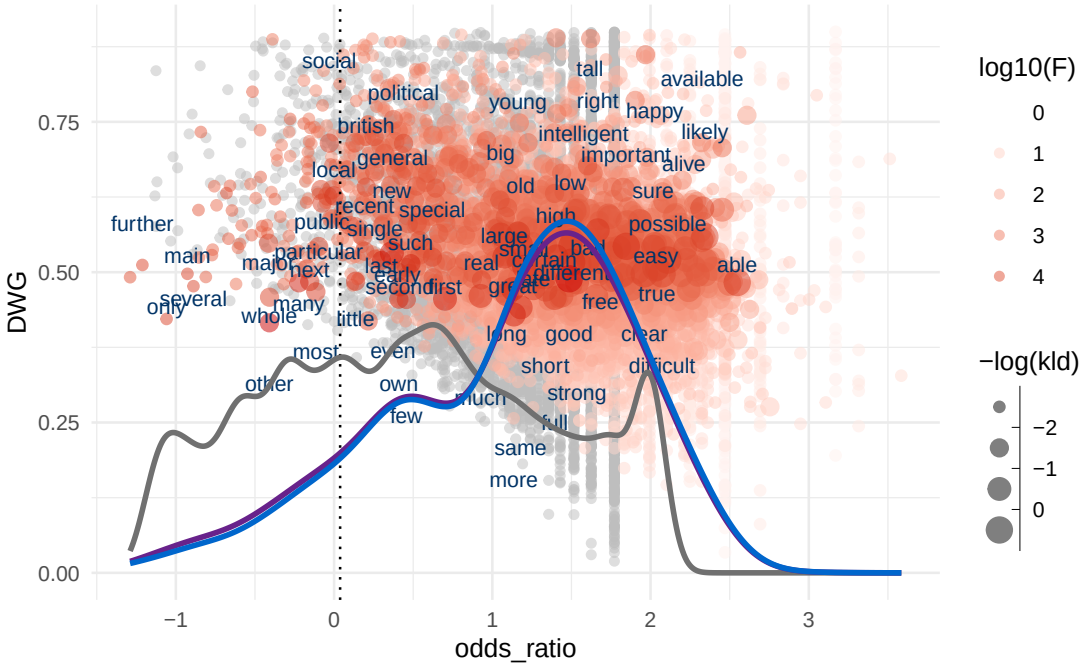


Figure 5.6.: Association ADJ to COP (x-axis) distance-based dispersion (y-axis), absolute log frequency (color), part-based dispersion (size). Same as 5.3, but with added 1-dimensional density: simple type density (gray), weighted with raw frequency (green), weighted with dispersion-adjusted frequency (U_r, blue).

other peaks. Based on the distribution of hapax legommena, dis legommena etc., and their relationship to the odds ratio of the sample, it is likely that nearly as many of the low-frequency items would be attracted to the copula construction with an odds ratio of about 2 (x=2) as are likely to be neutral or repelled with an odds ratio of -0.5. In general, there are a lot of lemmas that are attracted, statistically speaking. The picture for adjectives that are not attested in the construction is much noisier due to reasons discussed in Section 4.3.3.

The blue line represents the weighted density of forms accounted in the construction. Two main peaks can be observed at x=.75, and x=.4, and a heavy left tail that contains special cases of quantifiers that are most strongly repelled. The purple line shows the additive densities of both distributions, which is almost identical. Adjectives that are deemed ungrammatical in the predicative position are not separated well, and there is a lot of overlap. However, a bimodal distribution is clearly visible.

Finally, Figure 5.7 shows the results of classifying the data into two clusters as the result of fitting a mixture model with two Gaussian components. This so-called *rootogram* shows the density of the posterior probabilites. Peaks at 1.0 and little mass in the center are indicative of good cluster separation. Especially component 1 does not

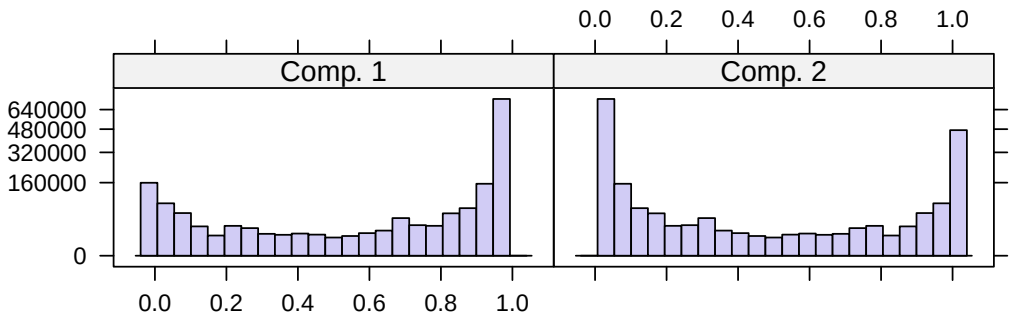


Figure 5.7.: Rootogram of mixture model of two Gaussian components

separate well, as visual inspection can confirm in figure 5.8. In this case, component 1 corresponds to non-predicative adjectives. Figure 5.8 shows predicted clusters of the data points. The expected attributive adjectives are very diffuse and masked by the predicative adjectives. For non-predicative adjectives in this model, strong association to the copula construction would be within the limits of class variation. This does not make a lot of sense if a word class based on the copula construction alone is assumed. It is more likely that other confounding variables cause the multi-modal nature of the data set, such as other constructions. If this is the case, the traditional conception of word class would be more suitable than a purely construction-based one as in Croft (2001). It could also be interpreted as emergent pattern stemming from the underlying semantic concept of gradability that is simply not fully lexicalized in English. Alternatively, the construction in question, the copula construction, might not be captured at the necessary resolution. In conclusion, from the perspective taken here, it is hard to imagine that adjectives that do not appear in predicative position form a separate class that is maximally distinct from other adjectives.

5.3.4. Interim conclusion

The three different perspectives on the data set are summarized in figure 5.9. The red line shows the smoothed estimate of a Generalized Additive Model (GAM) of DWG scores dependent on the odds ratios. The regression was also weighted by the U_r scores. This is an improved view on the data since the density is now multivariate and affected by dispersion. The only clear separation can be observed for inflectional forms. Interestingly, for gradability (center panel), the DWG at around x=-0.7. This means that adjectives that are moderately repelled from comparative use are also less well dispersed. This relativizes the peak that was visible from the one-dimensional perspective, and explains the lack of a mode in the 2D density since the variation in

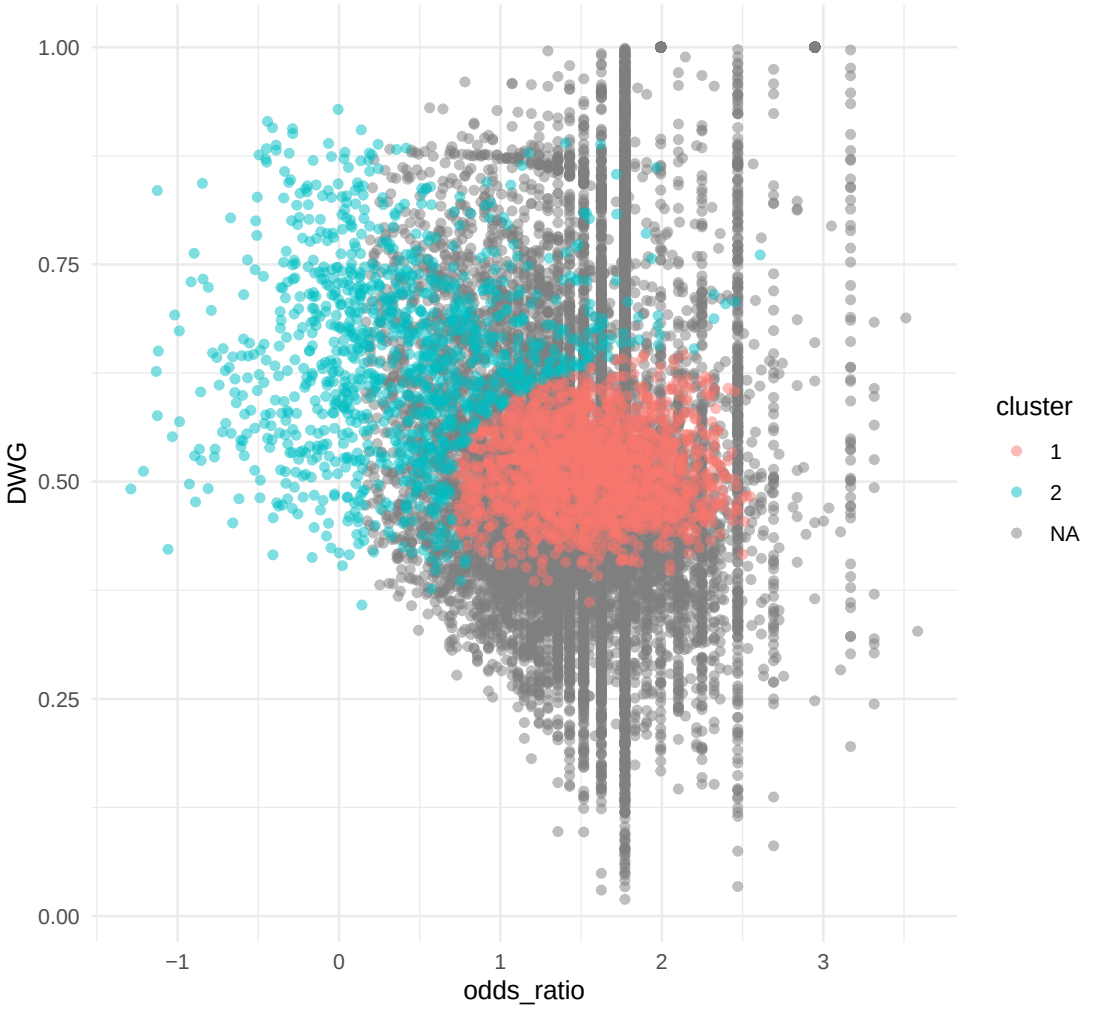


Figure 5.8.: Predicted clusters of mixture model

DWG is higher in that region. Either non-gradable adjectives are less common, or short bursts of occurrences of words like *political*, *social*, *local* in specific contexts cause an overestimation of the association strength, which is negative in this case.

5.4. Attributive vs. predicative

5.4.1. Multivariate perspective

To go one step further from specific to abstract, I will take a multidimensional approach by taking into account the association between adjective lemmas and their nominal,

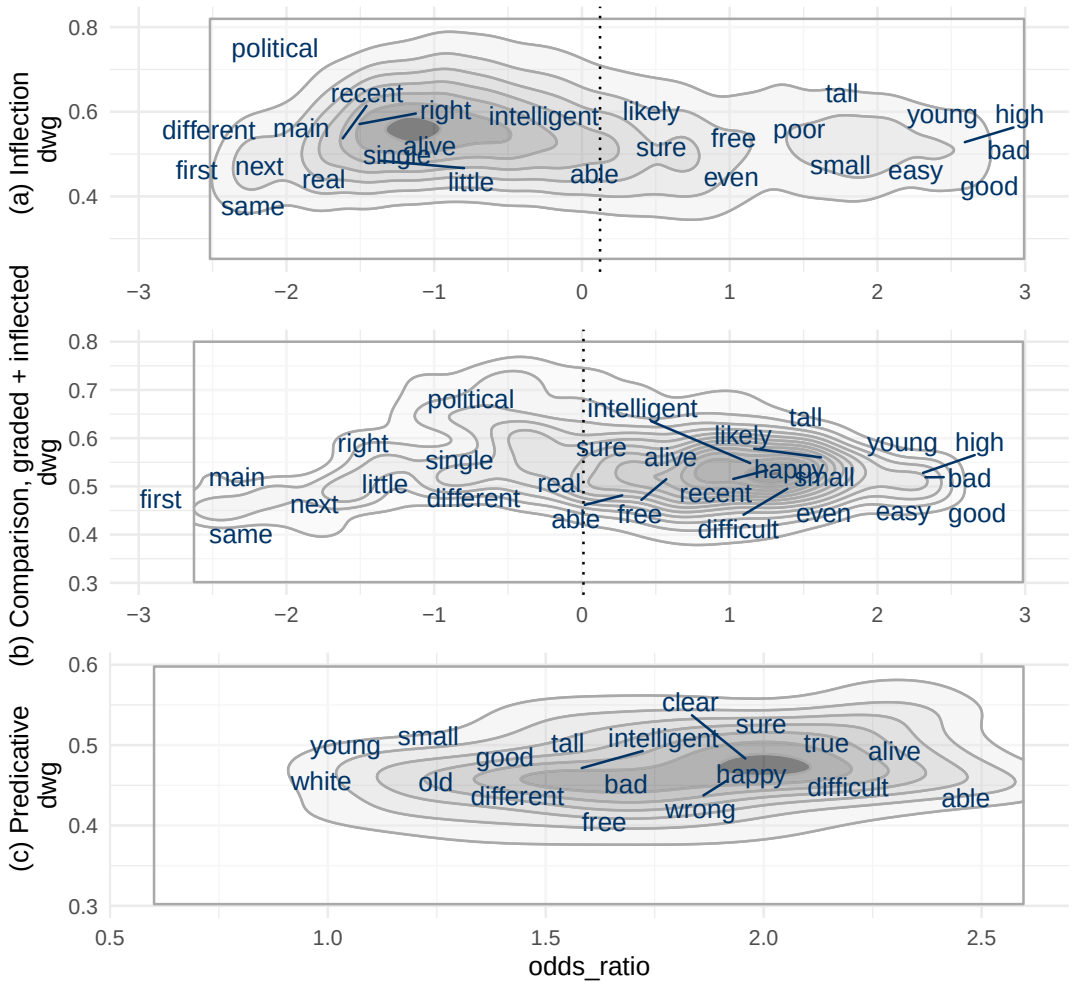


Figure 5.9.: Adjective features with weighted 2D densities

modifying, and predicative uses. Firstly, I took the different annotated dependency relation of the adjective and categorized them into the four groups: the three listed above plus separate categories for temporal modification, adverbial modification, and bare uses. The coding scheme can be seen in Table 5.2.

Table 5.2.: Coding scheme for adjective dependency relations

Group	UD tags	Example
adjectival	amod	*a great idea*
adverbial	advmod (dependent)	*very painful*
bare	discourse, appos	*right?*
predication	cop (dependent),	*it is real*

Group	UD tags	Example
	clausal tags	
nominal	all nominal tags	*the unemployed*
temporal	nmod.*	*do it first*

With those coded uses, a CA was calculated. Correspondence analysis can be used to analyze and visualize the associations between more than two categorical variables (cf. Baayen 2008: 139ff.). The technique involves converting the contingency table of the observed categorical data into a low-dimensional space that can be easily visualized, allowing for the identification of underlying patterns and relationships. This is similar to the approach taken in Collostruction Analysis and collocation analysis, but with multivariate data. Another difference to the previous approach is that the calculations are based on the Chi-squared statistic. CA is particularly useful for investigating the distribution of linguistic features across different populations, in this case potentially different word classes.

In order to reduce the dimensionality of the analysis further and since there is no straightforward way to include dispersion into the CA, I used adjusted frequencies instead of raw frequencies in the calculations. How the adjusted frequencies correlate with raw frequencies is shown in Figure 5.10. The distance-based dispersion-adjusted frequency f_{ALD} penalizes bursts of repeated occurrences and also clustering of occurrences within neighboring corpus parts. The BNC has a meaningful text order, therefore, this behavior makes some conceptual sense (cf. Rauhut 2022b). The penalty is a bit larger than the one for the second dispersion measure, U_r. This is due to the fact that U_r penalizes repetition per text. This is systematically more likely for the most frequent types in a corpus. The distance-based dispersion measure was picked for the CA. And the U_r was later used as additional weight variable in the density calculation.

Figure 5.11 shows a 3-dimensional view on the results of the CA for the six coded contexts. The vast majority of variation is explained by the opposition of predicative and attributive uses, which is to be expected. The only other variable with any influence to speak of were the nominal uses. The most common adjectives have notable uses as discourse markers, such as *good* and *right*. These uses and other bare forms fall close to the middle of the distributionn. Lexemes with dominant adverbial uses are extremely distinct. The same is the case for temporal uses. Therefore, only the categories corresponding to attributive, predicative, and nominal uses were kept and everything else lumped into a single category OTHER. The results of the reduced CA are shown in Tables 5.3 and 5.4.

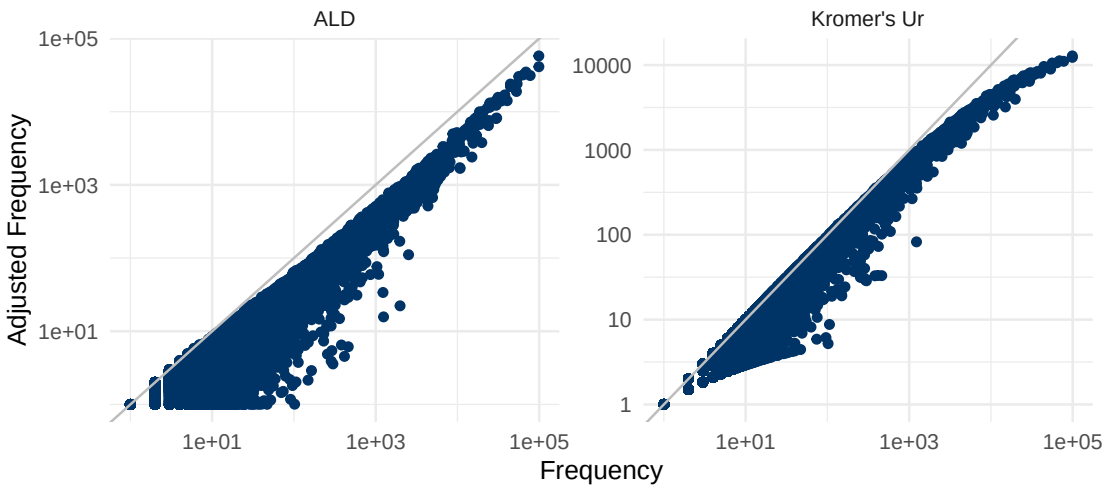


Figure 5.10.: Frequency adjusted by the Average Log Distance (f_{ALD}) versus raw frequency on a log-log scale (left); Kromer's *Ur* (right)

Table 5.3.: Correspondence Analysis of adjective features (reduced set): Principal inertias (eigenvalues)

dim	value	%	cum%	scree plot
1	0.347	85.6	85.6	**********
2	0.042	10.3	95.9	***
3	0.017	4.1	100.0	*
Total:	0.405497	100.0		

Table 5.4.: Correspondence Analysis of adjective features (reduced set): Columns

name	mass	qual.	iner-tia	k=1	COR	contr.	k=2	COR	contr.
modi-fica-tion	753	1000	191	-318	985	220	39	15	28
predi-cation	222	1000	665	1103	1000	777	12	0	1
nomi-nal	17	866	86	-261	32	3	-1329	834	697
other	9	494	58	-79	2	0	-1122	492	275

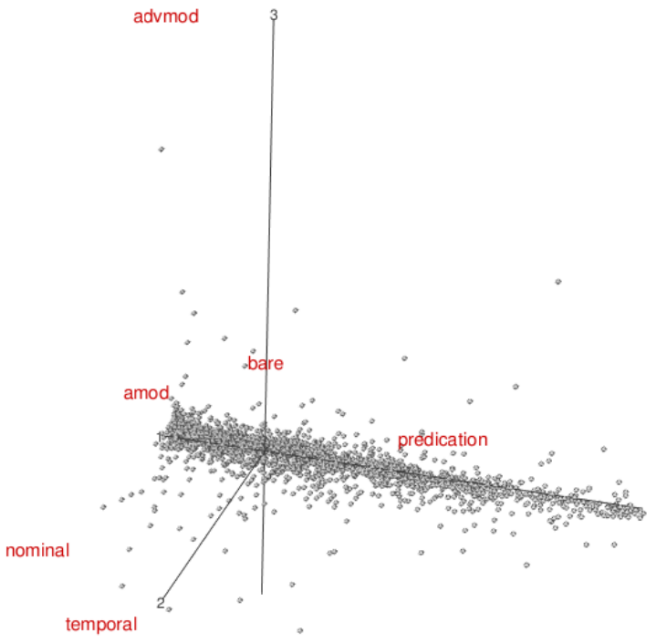


Figure 5.11.: Correspondence Analysis of Adjective dependency relations (all variables)

Even though adjectives can occur in a wide variety of syntactic contexts their variation between attributive and predicative contexts accounts for an overwhelmingly large amount of the distribution. 85.6% of the variation is explained by the first dimension, which has the modification group on one end and the predication group on the other. The nominal group is orthogonal in the third dimension but more similar to modification, as would be expected.

The strongest concentration of adjectives is found in lemmas that are strongly associated to attributive uses. Most adjectives across the frequency spectrum show a tendency for this function. This is not surprising since modification is assumed to be the prototypical function of adjectives. Fewer adjectives prefer predicative uses, however, among these are some of the most frequent adjectives in the data set, such as *wrong*, *happy* and *difficult*.

In neither the full CA in Figure 5.11, nor in the reduced version with weighted KDE in Figure 5.12 is there a clear separation between the predicative side and the nominal side. The bandwidth of the KDE was adjusted to be more sensitive on the x-axis since that is where most variation happens. We can see a small number of high frequency lemmas (at around [-0.5, 0.25]) that concentrate on the more nominal side of the distribution, these are the same group of quantifiers, ordinals and pronominals that were exceptional from all other perspectives. Some adjectives can convert to nouns, e.g., *the poor* and *the rich*. Those are usually strongly lexicalized uses. The lack of

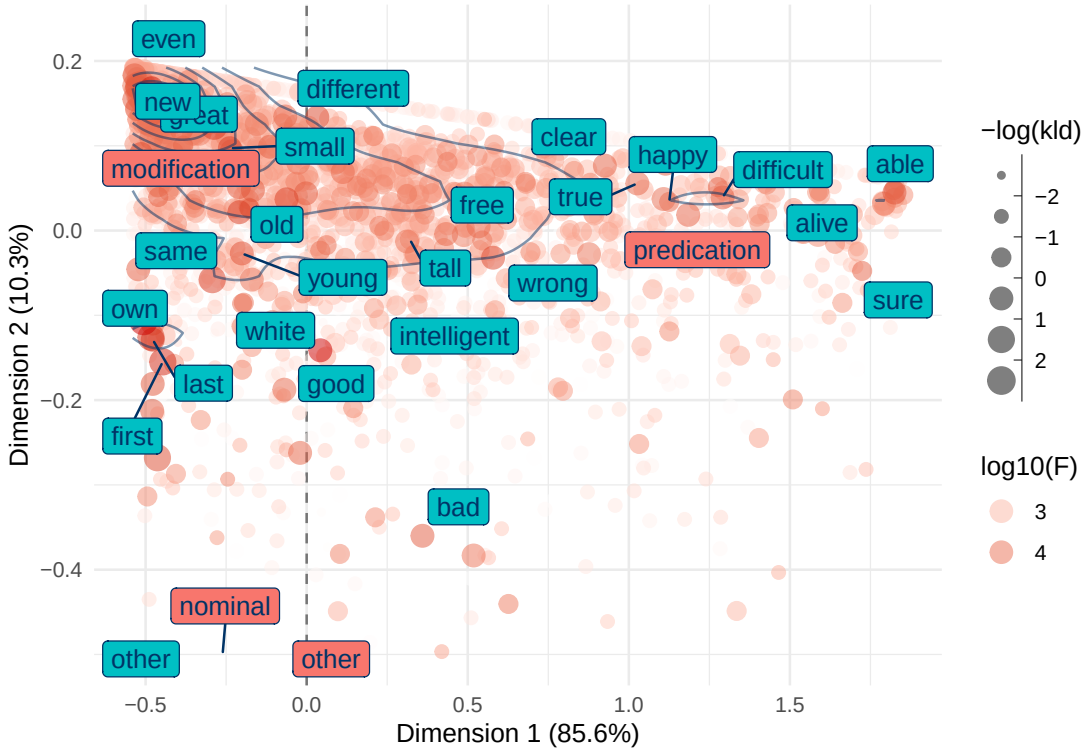


Figure 5.12.: Correspondence Analysis of Adjective dependency relations (outliers removed and zoomed in)

separation could have several reasons. The distinctiveness could get drowned out by random fluctuation in the data, or the tagged dependency relations have too much overlap due to the conceptual basis and/or tagging error. Only a very faint shallow mode can be seen on the predicative side where the lemmas *happy* and *difficult* are located. This mode readily disappears at different bandwidths.

Alternatively, the homogeneity of the adjective classes across this dimension is just that high. In the latter case, the interpretation would be that there is a cline from modification to predication along which adjectives vary, but no clear evidence for separate prototype categories. In this more complex picture, adjectives like *tall* and *free* appear as equally attracted to both functions of modification and predication. There is no uncanny valley of unusual adjectives between those two poles. In fact, some high-frequency exemplars, such as *intelligent*, *good*, and *bad*, are rather balanced between modification and predication (between x=0, and x= 0.4) once the their non-adjectival uses are accounted for. Intuitively, *good* and *bad* are some of the most typical adjectives, and being balanced between all the different adjectival functions might be what makes them typical.

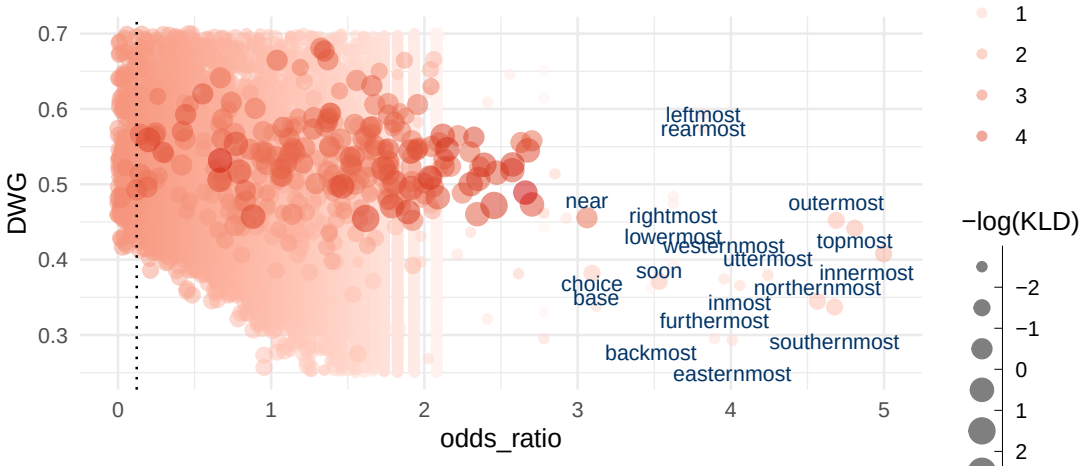


Figure 5.13.: Same as Figure 5.1, but zoomed on the area of strongest association.

5.4.2. Derivations of *most*

An interesting subclass that can be found in the data is words derived from *most*, which are a productive class of the schematic form $[x_{LOC}[most]]_{ADJ}$. Their derivation with *most*, in a sense, makes them superlative tantum if they are analyzed as superlatives.

(34) The first bit of every line is located at the **leftmost** bit of a byte (COCA: ACAD)
(35) where only the **anteriormost** BX-C gene turns on (COCA: MAG)
(36) Clamp a crimping iron on the first two inches from the scalp of hair 's **undermost** layers (COCA: ACAD)
(37) *leftmore / *leftmuch ...

Their distribution regarding adjectival features is almost identical to their base *most*, and they form a small local cluster on their own for all features. Figure 5.13 shows that they barely share this extreme region with other adjectives. *near* is rather strongly associated to comparative and superlative forms, however also exists as a preposition. Those forms are not included in this data set, and would strongly reduce this association. *base* and *choice* are tagging artifacts since most of their positive forms are formally ambiguous with compound nouns. Therefore, *base instinct* and *choice words* are tagged as nouns, while *basest instincts* and *choicest words* are tagged as superlatives. *soon* in predicative position is ambiguous with a temporal adverb. In attributive position, it is truly comparative/superlative tantum (*a soon date).

5.4.3. Deverbal adjectives in *-ed*

Deverbal adjectives in *-ed* show an interesting split between abstract properties on the modifying side and more concrete properties on the predicative side. Table 5.5

Table 5.5.: Top 30 best dispersed V-ed sorted along the modification-predication axis (Dimension 1, Correspondence Analysis)

lemma	Dim1	Dim2	f	dp_norm
impressed	1.8328438	0.0318102	1259	0.7362448
pleased	1.8250454	0.0387484	4126	0.5915239
delighted	1.7864335	0.0345803	1390	0.7555605
satisfied	1.6952614	0.0191157	2627	0.5775956
surprised	1.6699704	-0.0103831	1038	0.7774991
concerned	1.6400740	-0.0157161	5117	0.4561011
interested	1.5907420	0.0466418	7097	0.4532794
disappointed	1.5347149	-0.0036429	998	0.7815128
worried	1.4532096	0.0449049	1667	0.7136513
shocked	1.4155621	0.0358740	931	0.7972371
excited	1.4029712	0.0837356	963	0.7907227
tired	1.3376683	0.0158432	2572	0.6955474
aged	1.3339855	-0.0515768	1575	0.7131437
determined	1.2211944	0.0505160	1364	0.7013531
relaxed	1.0456204	0.0483693	1043	0.7475609
related	0.8799398	0.0563629	5221	0.5744236
married	0.7032334	0.0126168	1585	0.7520576
complicated	0.5863438	0.0514813	2307	0.5880974
unemployed	0.5370928	-1.0804321	1249	0.7971363
qualified	0.1393643	0.0761174	1237	0.7546058
armed	0.1360336	0.0730814	2848	0.7442831
sophisticated	0.0096501	0.0594412	2021	0.6154410
skilled	-0.0364608	0.0253039	1196	0.7391579
unexpected	-0.0662145	-0.0028003	1500	0.6492617
unprecedented	-0.1541361	0.1474556	641	0.7946708
advanced	-0.1687624	0.0765051	2217	0.6662051
experienced	-0.2792397	0.1358664	1224	0.7229760
detailed	-0.2998947	0.1465342	4868	0.5329008
distinguished	-0.3080703	0.0423850	845	0.7868632
limited	-0.4425142	0.1398125	3753	0.5021670

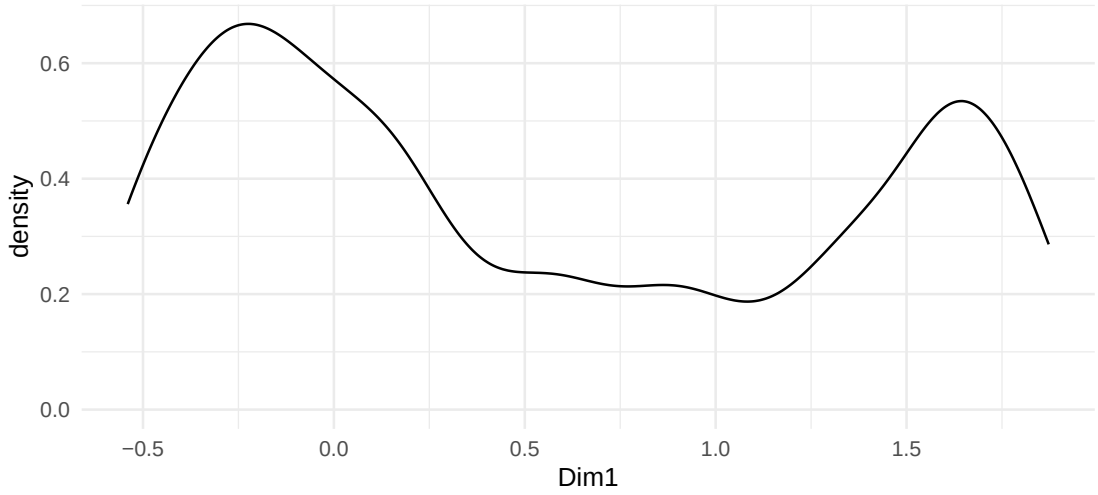


Figure 5.14.: Weighted 1-dimensional KDE of V-*ed* along the modification-predication axis (Dimension 1)

shows the top 30 adjectives in *-ed* by dispersion (DP_{norm}). There is a quantifiable difference between psychological states on the top of the list, and more abstract, permanent and social properties on the bottom. The weighted density in Figure 5.14 shows a clear separation. The top of the list is dominated by short-lived dynamic states, like *impressed*, *ashamed*, *pleased*, etc. The bottom features states that are socially rather concrete, but only indirectly related to human psychology. Those adjectives are time-stable and often imply a telic event structure. Among those, there are numerous adjectives of the form *un*-V-*ed*, like *unpublished*, *unauthorised*, *unspecified*. Another semantic field contains social constructs, such as *skilled*, *qualified*, *talented*, *sophisticated*.

This pattern is predicted by Givón (1980)'s time stability scale and in line with Dixon (1977)'s cross-linguistic observation that the most time-stable properties tend to be the ones lexicalized as adjectives. The 2nd most adjective-like item in the top 100 (cf. Table B.2) is literally *sustained*, which is as close to a prototypical exemplar for a time-stable property as one could wish for. However, its per-part dispersion is very low, as is the one for *uniformed* which is restricted to contexts containing police, soldiers, and other civil servants. As we move along the x-axis, we get to longer lasting psychological states, such as *relaxed* and *depressed*, after which this semantic field ceases to contribute.

The valley in between is populated by *married*, *related*, *complicated*, *unemployed*, among others. The latter two vary wildly in their time-stability, and even *married* is not always happily ever after. These are the 'uncanny valley' states where the variation may reflect the varying time-stability of the properties they describe. There might be substantial differences in framing between cases like *woman (who) is/was married*

and *married woman*. Compare 38-41.

(38) " Little Children " is about a young **woman** named Sarah , **who is married** , raising her child in an upper middle-class suburb (COCA: SPOK)
(39) the situation of reconciling the issue of being a Jewish **woman** [...] **who is married** , and fights to be a feminist (BNC: ARW)
(40) It 's not right for a **married woman** to have a man in the house and he not her husband (sic., BNC: A7J)
(41) they did n't believe in **married women** working and er they thought a **married woman** should be at home you see (BNC: K62)

A more detailed sociolinguistic analysis, however, is beyond the scope of this thesis.

related is the odd one out in this line-up since it should be expected to be very time stable. Its relatively strong association with predicative uses is caused by the construction [COP related to]—roughly half of its occurrences. The other half is mostly adjectival modifiers. Incidently, *married to* is also frequent, but only makes up less than 10% of the occurrences, so does not explain the intermediate position.

This shows quite clearly that semantic properties create prototype effects that are correlated with syntactic categories. It also shows that there is an uncanny valley. The pattern holds for more than just the top 30, see B.2 for the top 100. Interestingly, the correlation with time-stability is not as pronounced in the overall picture of all adjectives. Neither is the split between attributive and predicative uses. It is likely that the group of all adjectives contains many different overlapping populations that are distinguished only by more abstract variables. It can be hypothesized that homogeneous groups, such as *-ed* forms and *-ing* forms (cf. Section 7.3.4) exhibit variation along the time-stability scale, which otherwise gets blurred by the mixture of different populations.

5.5. Discussion

Combining the insights from the discussion on inflection and gradability, we can conclude that the four permutations (inflected-gradable, inflected-non-gradable, non-inflected-gradable, non-inflected-non-gradable) all form clusters to some extent. Many of the non-inflecting adjectives are also non-gradable; however, this group is also mixed with other marginal members of the category, such as quantifiers. The distribution is, therefore, best described as a cline. The overlap is large after accounting for all the lexical statistics presented in the data. Especially figures 5.9b-c essentially each show one domninant cluster.

Spatio-temporal contiguity as a cognitive factor in experience-based class formation might explain why the morphological difference is the strongest. The suffix is more likely to be perceived as part of the lexeme. Analytic comparison and its paradigmatic

relationship to the inflectional counterpart is more schematic. If 'gradability' is a conceptually motivated functional property, the separation between gradable and non-gradable adjectives should increase with the inclusion of both patterns. Based on the data, it can only be concluded that inflection causes a stronger clustering behavior.

Binary feature analysis, and analyses based on grammaticality judgment, which are a common approaches to gradience (e.g., Ross 1972; Croft 2022), cannot account for the variation found in the data. Unusual adjectives were found to be unusual across the category space even if they are attested in all typical constructions. Likewise, typical adjectives were also typical across the categories. For example, *happy* both shows a typical association to the comparative and superlative forms and to the copula form. Purely construction-based clustering could not be observed, and there is no indication for the 4 different categories listed by Croft. The lexical overlap is too high, which makes the entire class appear rather homogeneous, despite its internal structure. The patterns observed may be best described in terms of Aart's (2007: 6ff.) concept of 'subsective gradience'. The members of the class adjective vary in their degree of resemblance relative to the category prototype, which, in this case, can be characterized as abstract idea of an adjective that is gradable, but not graded too often, and is balanced between uses as modifier and predicate, with a stronger preference for attributive uses.

6. Pluralia tantum

6.1. Overview

The plurale tantum (pl. pluralia tantum), also known as uncountable plural, or categorical plural, is a type of noun that only exists in the plural form and lacks a singular counterpart. These nouns are found in various languages, including English. There are relatively few lexemes that can be identified as plurale tantum (cf. Huddleston & Pullum 2002: 341ff.), and even fewer lexemes that only have senses that are plurale tantum. There has been a lot of discussion about the categorical status of pluralia tantum due to the fact that it is difficult to define the class semantically, and due to cross-linguistic inconsistencies (Acquaviva 2004; Acquaviva 2008; Alexiadou 2011; Corbett 2019; Mackenzie 2019). This type of nouns is often compared and contrasted with mass nouns and other nominal subcategories that exhibit non-canonical behavior concerning grammatical number (Corbett 2019).

Not all languages have plurale-tantum nouns (Alexiadou 2019). For those languages that do, there is generally little overlap between the exact types of lexical items. For example, English *clothes* may correspond to German *Kleidung* which is a singular noun. There is the plurale tantum *Klamotten*, which is more restricted in use. *scissors* and *trousers* are regular countable nouns in standard German with singular and plural uses. The strongest candidate for an underlying semantic concept underlying some pluralia tantum is 'bipartite structure' (Huddleston & Pullum 2002; Goldberg 2006). Goldberg admits that bipartite structure does not predict the existence of items like *trousers*, and is satisfied with the prediction that a language with plural nouns for non-bipartite clothing, but singular nouns for bipartite clothing, is impossible (2006: 220). In terms of the larger group of pluralia tantum, the general consensus is that there is little common ground for generalization. Even within English, the same semantic field can contain both pluralia tantum and regular count nouns (Alexiadou 2019). For example, a *onesie* is a type of clothing that combines a part for the upper body and a bipartite portion for the lower body. A *suit* is comprised of multiple pieces of clothing, but singular. Both of these items are, in a sense, semantically more plural than *trousers*.

There are pluralia tantum that lack a singular counterpart with a similar meaning (*glasses*, *spectacles*), and others that have no matching singular form altogether (*clothes*). Even if rare, some singular forms of plurale-tantum nouns are possible, while others are not. Examples can be found in large enough corpora. The following examples include data taken from the COCA:

(42) Start with **a skinny pant** (COCA: SPOK)
(43) it tends to be the shorter the woman , the longer the **pant** (COCA: SPOK)
(44) Then he fished in his **trouser and** brought out a key (COCA: FIC)
(45) Stuck into a clever slot in the end , a toothpick and **a tweezer** (COCA: FIC)
(46) * the/a good
(47) * the/an odd

Pluralia tantum are used in their bare form in compounds.

(48) So out went the playful **jean and mechanic 's dungarees** (BNC: ADR)
(49) The salmon 's lipless **plier mouth** snapped open (COCA: FIC)

Most plural-only nouns listed in Huddleston & Pullum (2002) that have no semantically related bare form have frequent homonymous verbal or adjectival forms (2002: 341ff.). There are *trouser legs*, but no *good vendors* meaning vendors of goods.

On a specific level, constructions containing plural forms are different from constructions containing a singular form (e.g., Goldberg 2006: 5). The same could be said about comparatives. *A is healthier than B* is a construction specific to the comparative forms. Yet, *healthy* would not be considered to have a sense that is 'comparative tantum'. The difference between pluralia tantum and other grammatical patterns like this is that the conditioning is on a much more abstract lexical level, sometimes up to discourse level. *pyramids* has a sense that is plurale tantum when referring to the famous Egyptian pyramids. That is due to the high conventionalization of the phrase *the Pyramids*, which is basically a proper noun. *respects* is plurale tantum when occurring in the constructional idiom *pay (one's) respects*. In the same vein, *cats* is also construction-specific, and there are cases where there is little overlap. *I like cats* is different from *I like cat*. The main difference is in specificity, and the range of possible existing constructions that contain the same form. *cats* occurs in a large range of constructions that overlap with *cat* (e.g., many determiner constructions). The form *respects* is mainly restricted to the construction mentioned above plus some adverbials. In [QUANTIFIER (ADJ) *respects*], as in *in many respects*, the only overlap with *respect* exists in similar adverbials like *in some respect*. The singular is, therefore, likely negatively entrenched for all but these adverbial use. The same would be true for singular uses of respect, as in *have respect* or *respect for*.

Pluralia tantum vary with respect to their range of syntactic constraints. For example, some are more likely to be modified by a numeral than others.

(50) revealing **two scissors** she turned back to Patrick (BNC: EVG)
(51) *two clothes

The lack of quantifying modifiers for some pluralia tantum leads some researchers to argue that cases such as *clothes* are in fact to be characterized as plural mass nouns (Acquaviva 2004: 391; also see Alexiadou 2011). Based on similar observations, Acquaviva (2008) argues that there is no lexical basis for a class of plurale tantum.

However, his conclusion is equally applied to the mass-count distinction. It is still interesting since it is one of few studies that make an overt distinction between the descriptive concept and a lexical class. Adding gradience to the picture, I will argue in the following that the mass-count distinction resembles that of a lexical class more so than a potential third distinction of plurale-tantum nouns.

There are some notable subtypes of plurale-tantum nouns, such as pluralia tantum nominal gerunds: *proceedings*, *earnings*. Mackenzie (2019) suggests that even this small, restricted group displays too much variability to qualify as a constructionally defined category. It is important to note that the categorical status of nominal gerunds is presupposed. Other pluralia tantum are restricted to fixed expressions. Examples include *make amends* and *get the jitters/creeps* (cf. Acquaviva 2008). In these cases, there is no singular form, but any other nominal uses are also pretty much non-existent.

Quirk (1989) discuss examples of pluralia tantum allowing indefinite articles with modification. In the BNC, there is exactly one occurrence of *scissors* used like this, but it is in the sense of a sports maneuver. In the larger COCA, more examples can be found.

(52) Hastings responded with a brilliant diagonal run off **a dummy scissors** (BNC: K5A)
(53) cut herbs into a bowl or glass with **a sharp scissors** (COCA: ACAD)

This structure is interesting since it represents a construction that is not available for regular plurals. Unfortunately, its low frequency does not allow for a more detailed analysis in a corpus study, and likely has little effect on the quantitative data. In general, plurale-tantum nouns are an interesting case since it is a class mostly based on the absence of a related pattern. This is an oddity across the word class system and may point to a more general pattern that obligatory presence and obligatory absence are fundamentally different (cf. Section 3.3.3).

6.2. Morphosyntactic properties

6.2.1. General singular preference

To start, consider this simple sounding question: how often does a typical English noun occur with a plural marker? If a noun is almost never inflected and common enough, it is likely a mass noun. If it is always inflected, it is a plurale tantum. So the question is relevant for both of these types of noun. There must be a range of proportions of inflected to non-inflected forms that prevents negative preemption of the complementary form. Alternatively, there could be a diffuse, continuous increase in the proportion of inflected forms with most nouns having a low proportion of forms and the mass-noun distinction being completely determined by semantic properties or

other aspects of their distribution that they do not share with count nouns. To quantify this is surprisingly complex due to the nature of corpus data and word frequency distributions. Section 4.2 explored this issue in more detail. Rough estimates of relative frequencies of plural forms exist. Corbett (2019) claims that plural nouns tend to make up about 30% of the occurrences of a noun in text. The general preference for singular forms is well-attested cross-linguistically (cf. Greenberg 1990). However, the number seems much lower in the corpora used here. It is likely that this is due to the fact that the numbers cited in Corbett (2019) relied on some kind of manual disambiguation of homonymous mass noun uses. In reality, forms of both mass and count nouns have considerable overlap with their non-mass and non-count counterparts. Additionally, a simple percentage is an over-simplification since the distribution is heavily skewed. There is much more variation in relative frequency with nouns that are more common with it. Generally, relative frequencies only have their intended meaning when applied to normally distributed data from one population. Proportions are non-normal, and on top of this, the underlying data is Zipf-distributed count data.

Based on the distributions of relative frequencies in figure 4.2, most count nouns occur with plural markers about 20-25 per cent of the time. The dispersion measures presented in 4.2.3 can be used to mitigate some anomalous occurrences of plural nouns. Spurious plural-only and singular-only forms often occur within a single document and/or in rapid succession. Researchers have suggested various adjusted frequencies that penalize such word occurrences. U_r and ALD are two of examples (Kromer 2003; Savickỳ & Hlavácová 2002). Gries (2008) criticizes them for being too similar to raw frequencies, and points out that dedicated dispersion measures capture the idea of dispersion better. However, adjusted frequencies have some advantages precisely for their similarity to raw frequencies. They have a range and distribution that is familiar and can be interpreted intuitively. In the following studies, I will make use of these adjusted frequencies as frequency weights in density calculations when separating frequency from dispersion would otherwise lead to too much complexity. Figure 5.10 shows the relationship between these adjusted frequencies and raw frequencies. ALD tends to have stronger penalties for lower frequencies, while the penalty of U_r becomes increasingly stronger with frequency, in line with the Weber-Fechner law (cf. Section 4.2.1).

The only inflectional property that English nouns have is plural marking. The main marker is the regular suffix *-s*. There are some semi-regular forms, mainly with neo-classical borrowings, such as *alumnus*/*alumni* and *cactus*/*cacti*. Borrowings are often irregular, sometimes due to borrowing of the plural forms themselves, sometimes because of analogical word formation, often through meta-linguistic knowledge. There is a number of native Germanic nouns that have irregular plural forms, including the *-en* suffix; zero plural, such as *sheep* and *fish*; plurals involving ablaut, such as *foot—feet*, and *goose—geese*. Most irregular plurals are part of the base vocabulary and tend to be very high in frequency compared to their singular form, or at least can be assumed to have been high in frequency in older, spoken varieties of English. The most frequent

irregular plural forms were identified automatically, and confirmed manually. In the following, I will focus on the regular plural marker *-s*.

6.2.2. Homonymy of the *-s* suffix

There has been a wealth of research on the phonetic and phonological properties of the *-s* suffix investigating whether it is just one form at all. Yung Song et al. (2013), in a lexically restricted study, did not discover any significant differences between morphological forms in child language. However, a variety of follow-up studies suggest that there may be perceivable variation in duration (Plag, Homann & Kunter 2017; Zimmermann 2016a; Zimmermann 2016b; Seyfarth et al. 2018; Tomaschek et al. 2021; Schmitz, Baer-Henney & Plag 2021; Schlechtweg & Corbett 2021). Morphologically distinct homonyms with word-final [s] were found to be most strongly associated to length differences. Plag, Homann & Kunter (2017) argue that this heterophony is likely to influence the lexical representation in memory. Despite being subtle, it is possible that these differences are learned and assist in disambiguation. Although the effects were much smaller, significant differences were also observed between morphemic types of *-s*, particularly those that occur with nouns (plural, genitive, clitics). The duration was discovered to decrease with increasing contextual ambiguity, which points to a more general underlying process that is not related to categoriality. If there is a word class category of plurale-tantum nouns, this makes plurale suffixes, mostly the *-s* suffix, potentially homonymous. In a study specifically focused on pluralia tantum, Schlechtweg & Corbett (2021) found no evidence for homophony.

Pluralia tantum are similar to regular plurals in many ways. Even though some pluralia tantum refer to individual items, they are still used in the same way a plural noun would be used. They cannot be used with singular determiners such as *a* and *an* in English. Instead, plurale-tantum nouns are typically used with determiners that signal quantity, such as *some, a lot of,* or *many*. However, they share a subsets of these quantifiers with both count and mass nouns. There is little evidence that the suffix on plurale-tantum nouns is different from the suffix on regular plurals (see also Goldberg 2006; Langacker 1987c; Huddleston & Pullum 2002). In an attempt to describe the semantics of the plural suffix, Acquaviva (2008) comes to the conclusion that the meaning of the plural is not defined by the meaning of the singular.

6.2.3. Plural versus bare form

Other asymmetries can be found in the relationship between the unmarked (bare) forms and their marked counterparts. Adding a marker is possible on an individual occurrence, while omitting a marker often enough to lead to negative preemption requires long lasting processes. A speaker can creatively draw from patterns in the lexicon/constructicon and create new expressions. Any mass noun can be made a

count noun by adding a plural marker. If this does not match any lexicalized use, it may be perceived as unusual and possibly ungrammatical, but the schematic meaning derived from other count nouns is inherited by the new expression. Conversely, count nouns can also be used as a mass noun even if it is rare (Drożdż 2020). In that regard, the categories mass and count noun are both productive. However, the mass noun sense only comes across in a clear mass noun context, such as the use in partitive constructions. Another example is the bare singular construction (Croft 2001: 40).

(54) in search of the perfect 2,000 metres of flat water . It was a blessing that (BNC: CLP)
(55) consumption is about 100 units (kg of coal equivalent per person) in a developing (BNC: A1H)

In (55), we can see that nouns that are not necessarily mass nouns can occur in the bare singular construction as head of a compound with a mass noun. In a sense, the modifier infects its head with its mass noun properties. The mass noun meaning is 'inherited' in the sense of Krieger & Nerbonne (1993; see also Booij 2005). The noun *equivalent* otherwise shows a rather normal plural-singular distribution with about 16% of its occurrences BNC being plural. Outside the compound, *equivalent* does not occur in similar constructions.

(56) ?? a piece of equivalent
(57) ?? 3 pounds of equivalent

With compounds like *coal equivalent* or *meat equivalent* becoming more and more common, it is possible that those examples are not completely unacceptable.

Pluralia tantum, on the other hand, cannot be created ad hoc unless they are extremely close in meaning to a specific exemplar. Thus, we can find very specific lexical fields of plurale-tantum nouns, such as EYEWEAR with members like *glasses*, *sunglasses*, and *goggles*; LEGWEAR with *trousers*, *jeans*, and *shorts*; TWO-PRONGED TOOLS like *scissors* and *tweezers*. In the latter case, two-pronged might even be too general yet because they also have to be movable and tend to be hand-operated (cf. Huddleston & Pullum 2002: 341). Members that are added to this group not only inherit the aspects of meaning from the dominant exemplar, but likely also the grammatical patterns and restrictions.

There are other very small classes of nouns that have very specific lexical niches. For example, i.e., animals, in particular, those that are hunted, form such a minor class: *bore*, *deer*, *elephant*. In some specific senses, they show morphosyntactic irregularities, mostly plural agreement without an overt plural inflection. From a perspective of word class, they do not show any other striking morphosyntactic patterns that makes them distinct from other mass nouns or collective nouns. Its niche productivity can be sufficiently explained by analogical processes within its specialized semantic field, and its associated lexical structures, such as collocates and collostruction (e.g., *hunt some X*). This class distinction lives on the lexical end of the lexical-grammatical continuum.

The same can be said about the specific subclasses of pluralia tantum mentioned above.

Interestingly, other tantum-like categories do not exist in the English inflectional system. Also cross-linguistically, they are not described as such very commonly. Slavic languages may have an imperfective tantum (cf. Eckhoff, Janda & Lyashevskaya 2017). In English, there are arguably no past-tense-tantum verbs. For some reason, there does not seem to be a need for a category of lexemes that exclusively describe past events. The closest in resemblance to such a phenomenon is isolated examples of frozen items like *born*, where the distinction is rather one of aspect than tense. There are also no comparative/superlative-tantum adjectives or 3rd-person-tantum verbs.

Plural-only nouns have much fewer morphosyntactic idiosyncrasies than mass nouns. They share many distributional similarities with count nouns and some with mass nouns. They are restricted to certain specialized quantifying constructions (*a pair of N*, *some N*). These, however, are also available for mass and count nous.

(58) oil/oils
(59) forest/forests
(60) metal/metals

For other mass nouns, there are plural forms available that decribe so-called 'abundance plurals'. Examples include *water/waters*.

This does not work with all uncountable nouns, however, as there are also a class of singulare tantum, such as *luck/??lucks*. There are exactly two occurrences of *lucks* across the BNC, BNC2014 and COCA, one of which is from a non-native speaker, for which is unclear how intentional it was (consider Arnold Schwarzenegger's best *advices*). The other one can be seen in 61.

(61) It was one of the **lucks** of my life that I was able to go in and out (COCA: FIC)

Even though the phenomenon is not as famous as the debate on the *-ing* and gerunds, it is the same in essence. If there is a linguistic category of plurale tantum nouns, there needs to be conceptual basis for it, and provide evidence that it behaves like other linguistic categories, such as prototype effects.

6.3. Preparation

Similarly to Chapter 5, the population of lemmas was drawn from the dependency-annotated BNC with the values and measures listed in Chapter 4. Manual correction was carried out for lemmas with 0 singular or 0 plural uses up to an overall frequency of 5. Otherwise, no frequency thresholds were used. Any adjective lemmas containing non-alphabetical characters were removed. Irregular forms were treated as allomorph

variants, and retained in the sample. The coding scheme for the final groups is shown in Table 6.1. The distance between the head and the modifier in tokens once again showed little influence on the bigger picture.

Table 6.1.: Coding scheme for noun dependency relations

Group	UD tags
inflected	NNS, NNPS (PoS)
indefinite	det
	word forms: *a*, *an*
definite	det
	word forms: *the*, *this*, *these*, *that those*
modifier	amod, nmod:poss, compound
case	case (left)
clitic	case (right)
cop	cop
pp	nmod

The following examples illustrate some of the non-obvious structures captured by this.

(62) amod: **useful contacts**, **practical homecare** (BNC: A00)
(63) case: Alright , she gets a lot **of money** . (BNC: KSW)
(64) case: I could get us chucked **outside the cinema** (BNC: KSW)
(65) case: it 's like a ring **around the room** , you know hmm (BNC: KSV)
(66) nmod:poss a hut in **Roche 's garden** (BNC: A05)
(67) pp: The **care of people** in the community , (both overlapping cases included, BNC: A00)
(68) pp: After a short **interview with the BBC** (BNC: A00)

Determiner genitives and noun modifiers are not distinguished here (cf. Rosenbach 2019 for discussion on their similarity). Both cases can be broadly summarized under a modifying function.

The annotations are rather reliable in finding dependencies a few tokens away, therefore, examples like the following are also included. The distance between the noun and its dependency did not seem to have a significant effect on the results.

(69) The **interactive menu-driven access** (BNC: ALW)
(70) case: what she gone **to er photography lessons** or something ? (BNC: KSV)
(71) case: **With an even more resounding name** (BNC: KS8)
(72) nmod:poss **his grafting emergent politicians** (BNC: A05)

False positives do occur, however. The following examples show some of those artifacts:

(73) Bash **in my windows** ! (as ‘case’; BNC: KSW)
(74) provide **help in many different ways** (as ‘pp’; BNC: A00)

Unfortunately, it is hard to judge how pervasive such false positives are. On manual inspection, they seemed to be rare and unsystematic enough for the purposes of this study. Having a more reliable benchmark for linguistic purposes would be desirable Qi et al. (2020).

6.4. Data

6.4.1. Plurality

Like in the previous chapter, the first view is going to be on occurrences with the inflectional form, and is visualized in Figure 6.1. In contrast to adjectives, the bare form is not necessarily the grammatical opposite of the plural form. In addition to the reasons discussed above, there is also wide-spread ambiguity between compounds and adjective noun combinations. In general, the population of lemmas of potential nouns is much larger by an order of magnitude. It includes about 4.7 million tokens and 227,000 types.

Interestingly, many mass nouns describing substances, such as *water*, *oil* and *gas* have enough plural occurrences that, from a quantitative perspective, they look very similar to regular count nouns in distribution.

Some place names are also plurale tantum. For example, *Balkans* as a plural refers to the countries in the *Balkan*. *A Balkan* is not used to refer to one of the countries in *the balkans*. There is a clear metonymical relationship. Additionally, both words mostly occur with the definite article. On a morphosyntactic level, the plurale tantum form is not distinct from its related singular form. Analogous with plurale tantum nouns, such as scissors, the bare form is also common as modifier in compounds or attributive adjective, e.g., *Balkan countries*, *Balkan Mountains*.

In the data, there are many nouns that refer to scientific concepts and always occur in their plural form. There is always a singular form that is entailed by the terminology, but reference to the individual is irrelevant in language use. A range of combined forms ending in *-cysts* have virtually no singular forms since it is always a colony of cells that is relevant. The bare form of *antibody* is mostly used as attributive modifier—like most pluralia tantum—, more rarely to refer to the generic concept, and more rarely yet to an individual protein. The singular form also often lacks other nominal markers, and often resembles a mass noun. Very rarely, prototypical uses can be found, however (76).

(75) HIV-1 preparation or adjuvant control , showed significant increases in **antibody** (COCA: ACAD)

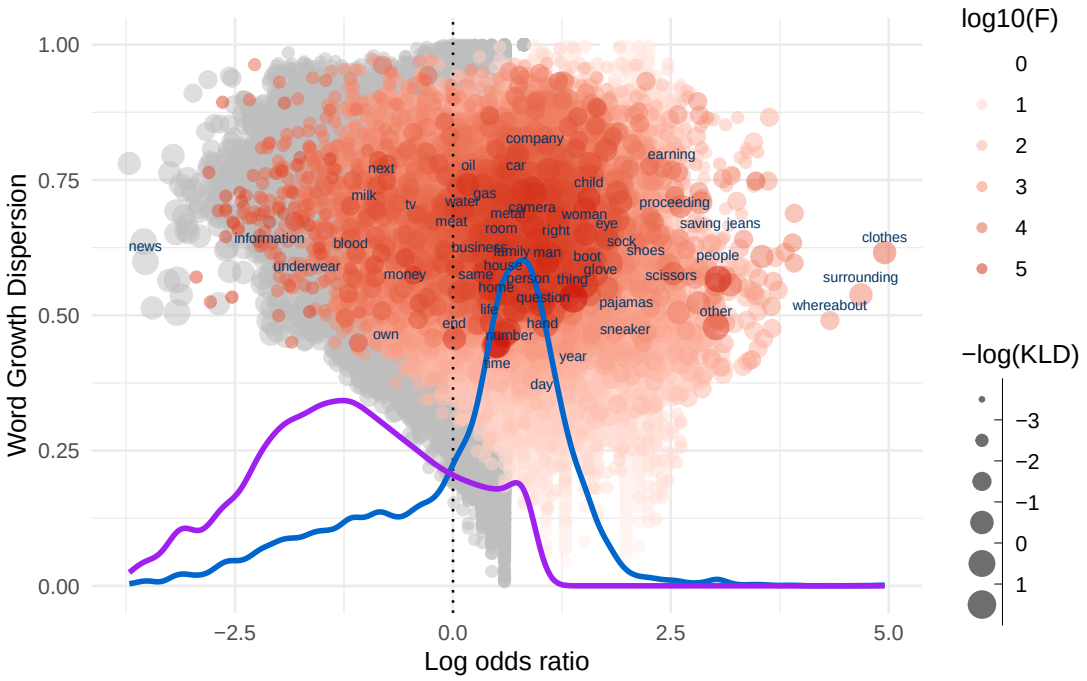


Figure 6.1.: Association of noun lemmas to the inflectional suffix -s (x-axis) against distance-based dispersion (y-axis), absolute log frequency (color gradient), and part-based dispersion (size). Lines refer to U_r-weighted density estimates of the odds ratios: all lemmas (blue), only with attested plural (purple). Lemmas for which the form is not attested appear grey. The text labels are fixed along the x-axis; their vertical position only reflects relative order.

(76) The method takes advantage of **an antibody** 's ability to bind to a unique antigen in pathogen cells (COCA: ACAD)

(77) Her therapy uses **an antibody** that targets the CD-45 protein on white blood cells and most leukemias

Such lexemes present one of the classes that bridge the gap between plurale tantum and count nouns as they exhibit a very similar distributional profile by having very rare singular forms. Meta-linguistic reasoning can suspend the negative preemption of singular forms even in *trousers*, and *scissors*, but it is more jarring due to their use as core vocabulary. It may require extremely high frequencies of use to become truly plurale tantum, something most scientific concepts lack outside of their domain.

Some forms that are strongly associated to the plurale derive from adjectives, such as *singles*. This noun contrasts with a bare form mostly used as an adjective which is common in predicative position. Such bare forms are not prototypical nouns because they are focused on the state of an individual.

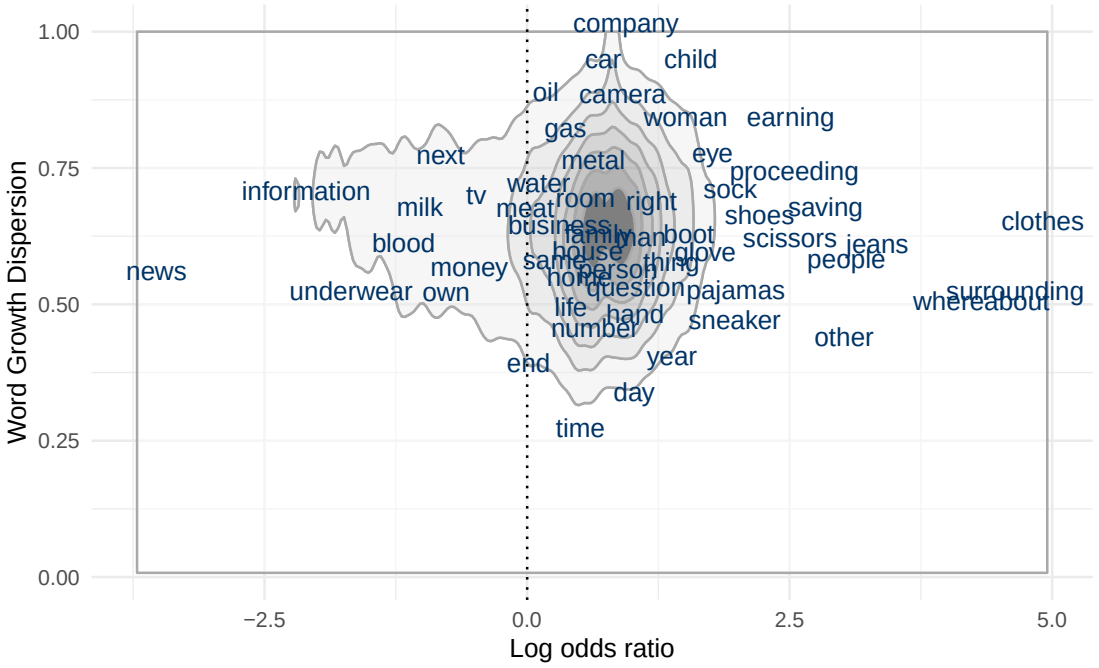


Figure 6.2.: Association of noun lemmas to the inflectional suffix *-s* with U_r-weighted 2-dimensional density. The text labels are fixed along the x-axis; their vertical position only reflects relative order.

(78) They were **single**
(79) One of the items available from the display was a simple measure to ensure that all spirits drunk at home **are singles** (BNC: K4V)

On the basis of inflection alone, it is not possible to find any clustering of pluralia tantum.

parents is not plurale tantum in English, but it is in German. Even in English, the distribution is very strongly skewed towards plural uses. The most frequent plurale tantum is *clothes*, which also lacks an attributive use.

The two-dimensional density can be observed in Figure 6.2. There is a clear shoulder with odds ratios of below 0 that represent items that are repelled from the plural suffix. Based on the plural alone, there is no separation, which in itself is not surprising as it hints at the fact that there is are other variables defining mass nouns. In other words, focusing on the property of plural marking, the category noun appears as one continuous class.

Plurale tantum nouns are relatively rare in both type and token frequency. They also vary substantially in their dispersion. This means that some appear rather sporadically and in short bursts, and few are more evenly distributed. Among those evenly

distributed, yet fewer never occur in their bare form. That means, while the chances were almost equally high to come across a gradable or non-gradable adjective, there are just not enough types that appear regularly. There is no dense pattern vertically or horizontally. Gradable and non-gradable adjectives had similar dispersion, frequency and statistical association with class-defining properties. These properties can be found with mass nouns. Mass nouns overlap very strongly with regular nouns in terms of association to the -s suffix, but form a distinct shoulder in the distribution.

6.4.2. Combined morpho-syntactic features

As a next step, I combined the features above to see whether the nominal subclasses become distinct from a multi-dimensional perspective. The result of a CA can be seen in Tables 6.2 and 6.3. The class of nouns is significantly more complex, and the selection of nominal features includes more dimensions. Consequently, the explanatory power of a 2-dimensional projection is much lower than was the case with adjectives. The first two dimensions account for only 56.52% of the variation. A visualization of the first 2 dimension can be seen in Figure 6.3.

Nevertheless, three modes are visible. A main cluster at around [0, 0] contains the majority of the data. Another cluster containing mostly mass nouns is visible with a center at around [0, 0.4]. The largest region in x direction is covered by a rather thinly populated strip of proper nouns. The reason for this is that none of the features are particularly distinctive for proper nouns except for the possessive clitic. However, the group of proper nouns does not show a uniform preference for the clitic, and for most members of this class it is a simple matter of marking versus no marking. At that point, the association strength is contingent on frequency alone, preventing any clustering. A similar pattern can be observed with non-inflecting adjectives in Chapter 5, Figure 5.1.

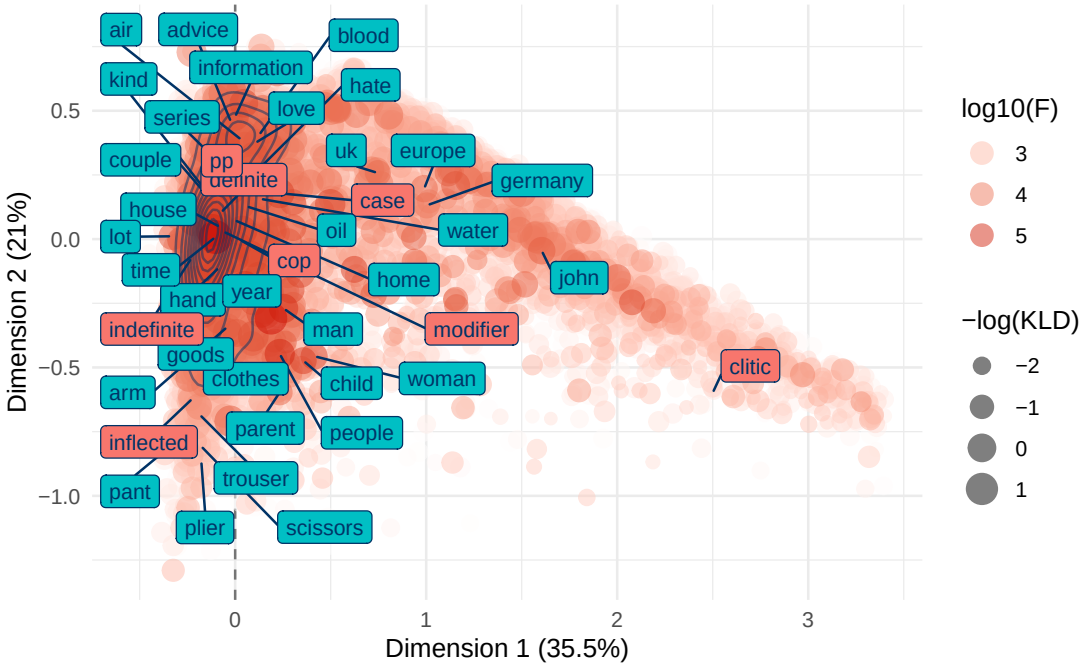


Figure 6.3.: Nominal morpho-syntactic features: 2-dimensional view on Correspondence Analysis with 2-dimensional KDE. The CA is based on the U_r adjusted frequencies.

Table 6.2.: Correspondence Analysis of noun features: Principal inertias (eigenvalues)

dim	value	%	cum%	scree plot
1	0.152632	37.3	37.3	*********
2	0.079592	19.5	56.8	*****
3	0.054688	13.4	70.2	***
4	0.038436	9.4	79.6	**
5	0.032865	8.0	87.6	**
6	0.028575	7.0	94.6	**
7	0.021976	5.4	100.0	*

Table 6.3.: Correspondence Analysis, of noun features: Columns

name	mass	qual-ity	iner-tia	k=1	COR	contr.	k=2	COR	contr.
mod.	217	62	60	-80	57	9	23	5	1
case	234	522	115	200	199	61	255	322	191
cop	35	1	60	12	0	0	-22	1	0

name	mass	quality	inertia	k=1	COR	contr.	k=2	COR	contr.
clitic	16	974	337	2827	938	847	-550	35	61
pp	132	201	91	-147	77	19	188	125	58
infl.	129	878	160	-197	77	33	-637	801	657
def.	156	70	80	-88	37	8	83	33	14
indef.	81	125	97	-209	90	23	-131	35	18

Figure 6.4 shows a 3-dimensional view on the same data. The three contour levels represent the 25th, 50th and 75th percentiles. This means that the red area comprises about 25% of the lexemes, the beige area 50%, and the gray area 75%. 25 per cent of the data are well separated into the three densest clusters, but as a whole the data set forms one coherent class. The patterns become clearer and the clusters are more distinct. The three dimensions now account for 70.22%. The biggest cluster of lexemes is well-balanced between the main morphosyntactic markers, i.e., INDEFINITE, DEFINITE, adjective modifiers (MODIFIER), and prepositional modification PP. The clusters of mass and count nouns are mostly distinguished by DEFINITENESS and CASE rather than number. Count nouns are more strongly associated to uses with the indefinite article, while mass nouns prefer the definite article and demonstratives. They are also more associated to CASE, which mostly represents occurrences in prepositional phrases that can be roughly described as 'oblique'. These are most often *of*-phrases, such as *a glass of water*. In fact, DEFINITENESS alone does not account for the difference in the clusters. Together with adjectival modifiers, these are features that actually connect the two groups as they are almost perfectly in between.

The continuum between INDEFINITE and DEFINITE is arranged orthogonally to plurality. The absence of the indefinite article with mass nouns is often seen as a sign for countability, but does not correlate with plurality since plural-dominant nouns cannot co-occur frequently with the indefinite article due to their singular form being less frequent. Countability is, therefore, at least a two-dimensional phenomenon.

Proper nouns are distinguished by the use of the possessive clitic. The majority of rarer proper nouns do not show any specific preference for the markers investigated here. Therefore, the group as a whole is rather diffuse.

Figure 6.5 shows the same data, but with significant modes (Chaudhuri & Marron 1999; Duong et al. 2008; Godtliebsen, Marron & Chaudhuri 2002) added as yellow contours. Only two of the cluster modes are significant (at a 0.05 significance level). The count noun cluster is too variable and more like a moderately dense continuum. At least with the features tested, it does not significantly differ from the other clusters. If the members of this long strip of nouns are inspected, it becomes apparent that there are different conceptual types of nouns arranged along the continuum. Many nouns refer to humans that are often referred to as group or collective, including *children*, *refugees*, *immigrants*, but also *workers*. They show similar distributional patterns

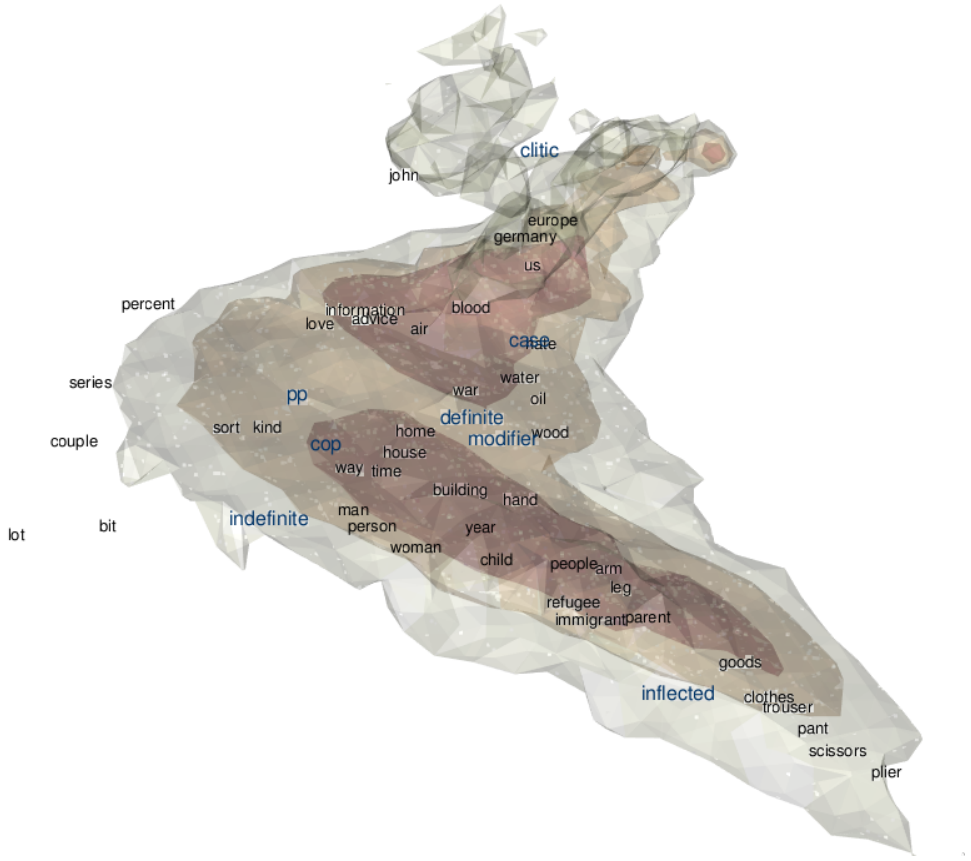


Figure 6.4.: Nominal morpho-syntactic features: 3-dimensional view on Correspondence Analysis with 3-dimensional KDE. The correspondence analysis is based on the U_r adjusted frequencies. The colored contours represent the 25th, 50th and 75th percentiles.

to the more general *people*. Inside of the mass noun cluster, the association to the plural ranges from repelled abstract nouns like *information* and *love* to mass nouns describing materials that are commonly classified into different types, therefore often pluralized. An intermediate type that is connecting these two groups can be found as mass nouns that are not as commonly used as resources, such as *air* and *blood*. The plural uses of blood is restricted to certain metaphors and strong collocations such as *young bloods*. *air* has some rare homonymous plural uses (*airs and graces*) that are historically unrelated. Again, disambiguation would create a stronger negative association for these two examples, but there is no reason to assume that cases like these are systematic for this whole region in feature space. In fact, removing these frequent forms from the subset does not change the picture dramatically. It is also important to note that uses like *airs and graces* are not well-dispersed over the corpus and already penalized due to the use of the adjusted frequencies.

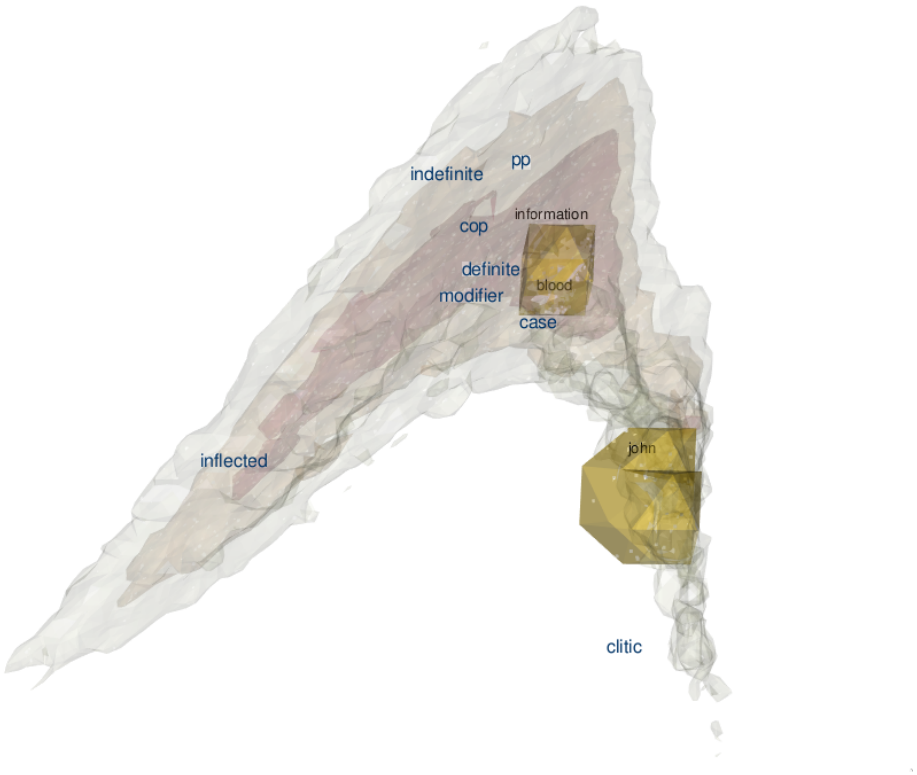


Figure 6.5.: Nominal morpho-syntactic features: 3-dimensional view on Correspondence Analysis with 3-dimensional KDE and significant modes (yellow contours).

The following examples outline a cline of countable nouns from least to most associated to the plural:

(80) singular pronominal: *none*, *nobody*, *something*
(81) singular dominant:

a. complex quantifiers: *percent of*, *(a) couple (of)*
b. partitives, hedging expressions: *bit(s) of*, *kind(s) (of)*, *sort(s) of*

(82) core nouns: *time*, *way*, *house*, *home*
(83) balanced:

a. animate individuals: *woman*, *person*, *man*
b. time increments: *day*, *week*, *year*

(84) plural dominant:

a. animate groups: *people*, *workers*, *immigrants*, *parents*
b. paired concepts: *arms*, *legs*, *eyes*
c. pluralia tantum: *goods*, *clothes*, *scissors*

(85) plural pronominal: *others*

These groups are based on the most frequent and best dispersed exemplars. The overlap is significant and becomes larger with less frequent items.

A similar cline of countability can be described for the mass noun cluster, again from least countable to most countable:

(86) pronominal quantifiers: *everything*, *everyone*, *rest*
(87) locatives and directives: *front*, *back*, *top*, *south*, *north*
(88) abstract nouns: *information*, *love*, *production*, *attention*
(89) mass nouns: *food*, *water*, *oil*, *energy*
(90) collective mass nouns: *industries*, *sciences*, *minds*, *arts*

There is no clear group of plural-dominant nouns in the mass noun cluster since this is where its distribution merges with the count noun cluster. In a sense this gives the idea of pluralia tantum as special mass noun some support since they would be the continuation of the mass noun distribution. In other words, count noun and mass nouns converge at the plural-dominant end of the continuum; therefore, pluralia tantum are both mass nouns and count nouns at the same time since there is no opposition.

6.5. Discussion

The results demonstrate that mass nouns form a distinct cluster already based on a relatively small set of features. The count noun cluster is more diffuse, but definitely distinct from both mass nouns and proper nouns. The general impression is that there are many distinct populations included in count nouns from the perspective of the selected morphosyntactic properties. This is not too surprising since the class of nouns is by far the largest. It also shows that perhaps the count-mass distinction is not an opposition on morpho-syntactic level, since there are many types of countability. Maybe the question is ultimately whether the category of 'count nouns' makes sense, and not whether it is mass nouns, proper nouns or pluralia tantum. Overall, the CA conveys a good impression of the network of nominal subclasses. All distributions have significant overlap.

Based on the data, the best explanation for the plurale tantum is that it is the end point of a contiguous cline of various types of count nouns that at some point lose their ability to be used in the singular. Whether it is possible to determine this point reliably, and whether there are prototype effects past that point remain open questions due to the uncertainty in the data. Considering the cross-linguistic literature, it seems likely that there are no prototype effects at least on a grammatical level. On a more specific constructional level, it is likely that smaller groups of pluralia tantum (e.g., *trousers*, *pants*, *knickers*, or *scissors*, *pliers*, *tongs*), create local, specific prototype

clusters. From a perspective of more abstract word classes, they are definitely not on par with proper nouns or mass nouns.

In Rauhut (2022b), I presented data that suggests that there is a strong textual preference to either the plural use or singular use of an individual noun. Texts tend to either refer to *cats* in general or individual countable *cat(s)*. In specific communicative situations, even some regular plurals are unlikely to co-occur with their singular counterpart. Consequently, pluralia tantum may be considered defective cases where there is simply no available communicative context that would trigger singular use. This is reminiscent of Langacker's (1990) treatment of homonyms (cf. Section 2.4.1). These communicative contexts that could trigger singular use are rare or negatively preempted due to the relatively high frequency of words like *scissors*, *trousers*, etc.

7. How many *-ing*?

7.1. Overview

The *-ing* suffix is a versatile morpheme in English, serving a wide range of functions in the English language. Its most common use can be found in the progressive construction as part of the English tense-aspect system. As such, it is traditionally considered a participle. The present participle is also used in non-finite clauses and verbal gerunds. Furthermore, functions of *-ing* include the derivation of adjectives, nouns and prepostions.

(91) Present participle:

a. progressive construction: *I am eating*
b. adverbial clause: *Going down the hall, she saw ...*

(92) Gerund:

a. nominal gerund: *the writing of a story*
b. verbal gerund: *be good at writing a story*

(93) Deverbal adjective: *interesting*, *challenging*
(94) Deverbal noun: *building*, *meaning*
(95) Deverbal preposition: *during*, *concerning*, *regarding*

These uses span almost the entire English word class system. This is likely a consequence of formal syncretism and the competing functions of originally distinct forms that become encoded by one and the same form (De Smet et al. 2018). The difference between nominal and verbal gerunds is most commonly described in terms of its 'internal syntax'. Nominal gerunds have the internal syntax of a noun phrase, while verbal gerunds have the internal syntax of a verb phrase (cf. De Smet 2008; Langacker 1987b; Pullum 1991; Ross 1973a).

There is a high degree of terminological confusion and overlap when it comes to the concepts of nominal gerund, verbal noun, gerundive nominal, and gerund participle. Distinctions between neighboring categories in this classification are often difficult. There is significant overlap between the nominal gerund and verbal nouns, as well as verbal gerunds and participles. There can also be mixed gerunds that share verbal and nominal features. According to Givón (1980), *-ing* forms are higher on the binding

scale than verbs. That means they are inherently more noun-like than regular verbs, even as participles.

7.2. The problem

7.2.1. *-ing* as in-between category

The *-ing* form can be used extremely flexibly, with varying degrees of nominal and verbal marking:

(96) **Walking** was nearly impossible (BNC: AT3)
(97) He thought **walking** was old-fashioned (BNC: C86)
(98) we do have a large number of visitors to Scorton who enjoy **walking the paths** (BNC: HPK)
(99) Normally , **walking along tarmac is** a piece of cake (BNC: AS3)
(100) so people can enjoy **walking around the farm** (BNC: K1T)
(101) You obviously enjoy **walking** , ' he countered (BNC: JYF)
(102) Charity **kept walking** (BNC: JY6)
(103) I **am walking** over hot coals (BNC: CEK)

These uses exhibit different degrees of nominality. Functionally, it is also not always clear which cases represent an event nominalization, or simply an event. In (96) the event encoded by *walking* likely has a concrete discourse referent, while in (97) the event is more abstract. This makes (96) more verbal, and (97) more nominal. The syntactic structure, however, is nearly identical. The only available cue comes from the pragmatic interpretation of *nearly impossible* via implicature. This implicature itself requires the right discourse context since the utterance can be framed to refer a generic concept. Considering the subject-object-oblique hierarchy, these examples range from highly noun-like to not noun-like at all. The last two examples are typically considered participial uses, but they can contrast to some degree with nouns. If there is an in-between category of gerund-participle, it should be expected that this cline is represented in the data and that prototypical *-ing* forms should not cluster at either end but somewhere in the middle.

Aarts (2004) rejects the idea of a gerund category on the basis of his idea of subsective gradience (see Section 2.3). Gerunds in his view can be both members of the classes noun and verb, and assignment depends on balancing the morphological features. This approach does not take into account whether gerunds typically occupy the intermediate position between nouns and verbs. In his model of subsective gradience, this would have to be assumed. A model without gerunds or gerund-participles would predict that lexemes vary between mostly nominal and mostly verbal uses with more mixed uses being possible, but the exception.

Some deverbal nouns, such as *building*, and *meaning* are very strongly entrenched as noun. However, the verbal sense is still fully available in most cases, and even rather frequent. Some exceptions include words such as *ceiling* that have no transparent root. Senses across nominal and gerundial uses are often also fully transparent and clearly related; however, some denominal verbs have a non-obvious semantic relationship to the nominal sense.

(104) He was involved in **the building of** nearby Lyndon Hall (BNC: AB4)
(105) She thought , I 'll have him to **do the building** but not the rest of it (BNC: KST)
(106) prevention as such has always had lower status than the task of catching those who do **the taking** (BNC: A0K)
(107) they saw **the taking of** law into their own hands as temporary (BNC: A07)

The semantic relationship between (104) and (105) is not as close as in the case of *taking*. The nominal sense of *building* is focused on the resul of the action. In (106) and (107) both instances are action nominalizations. It is not too difficult; however, to find action nominalizations of the verb *build* when they have clear constructional uses that are typical for gerunds, such as the [do the V-*ing*] construction seen in (106). It is unclear whether the difference in senses between the gerundial uses of *building* are intransparent and warrant a homonymy analysis. More likely, the deverbal use referring to an object is an extension in the network of senses that is not available for all *-ing* forms, but connected to the verbal senses via action nominalizations, such as in (105).

Both English participles can be used as predicative adjectives (Huddleston & Pullum 2002: 541). This is likely due to a clash with other dominant copula uses of the same form. The past participle is used to form the passive and the present participle is used to form the progressive. The *-ing* form's use as deverbal noun may also clash with the copula use. Exceptions to this are compounds, such as *time-consuming*, *awe-inspiring*, *self-assuring*, *Russian-speaking*. These are usually unambiguous and lack nominal and verbal uses. Even in this group, there are forms that cannot be used predicatively, such as *Oscar-winning*. It is difficult to say, however, whether this is connected to the *-ing* suffix or due to the general variation of adjectives. In many regards, these compound adjectives are less behaviorally marked than participles.

7.2.2. The Gerund-Participle

The gerund-participle analysis, which treats gerunds and participles as different forms of the same underlying category, is a prominent view that is held by some linguists, such as Huddleston & Pullum (2002) and Quirk (1989). However, it is not universally accepted, and there is ongoing debate and discussion about the nature of these forms (also see Aarts 2007). Quirk (1989) also favors the gerund-participle. He describes gerunds and participles as different surface forms of a single underlying form. They argue that the gerund-participle has both nominal and verbal properties, and that its

use is determined by the syntactic context in which it appears. They also note that it has properties that are intermediate between those of full-fledged nouns and verbs. The account of Huddleston & Pullum (2002) is generally similar.

Mixed gerunds are part of the reason why the categories are often lumped. Mixed cases sometimes balance properties of nominal and verbal forms. In some cases, nominal morphosyntax is present combined with verbal syntax. The example in (108) has a possessive determiner, which is a cue for nominality, but the *-ing* form also has its own direct object *you*. This structure shows significant variation with pronominal forms (in this case *me*).

(108) But if you really do n't like **my kissing you** in public , why do n't we go upstairs (BNC: JY0)

(109) The idea of **their having** a say in the running of a club appears to make officials recoil with horror (BNC: A8C)

Duffley (2006) represents the most drastic case for a unified category by also including deverbal adjectives and nouns. Using image schemas (Lakoff 1987a; Langacker 1999), he offers a unified conceptual description of *-ing* forms. With the possible uses and meanings of *-ing* being so varied, this leads to a highly schematic abstraction. He posits that all *-ing* uses can be explained by the idea of 'interiority' (Duffley 2006: 16). This includes its derivational uses.

One major criticism of the gerund-participle analysis is that it fails to capture the distinct syntactic and semantic properties of gerunds and participles. Some linguists argue that gerunds and participles are actually separate categories with different underlying structures, rather than just different surface forms of the same underlying form (e.g., Bresnan 2001: 287f.). Bresnan (2001) only presents very limited data in the form of constructed examples including the same verbs, which does not exclude the possibility of lexical confounds.

To my knowledge, a specific stance on the gerund-participle debate has not been made explicit, within the context of RCG. However, RCG's emphasis on constructions suggests that it may be more sympathetic to the view that gerunds and participles are distinct constructions rather than members of a single category. Croft (2001), emphasizes the importance of constructional idioms, which are fixed expressions that cannot be analyzed in terms of their component parts. This strong role of (constructional) idioms supports the view that gerunds and participles may be distinct constructions, as they often exhibit different patterns of meaning and use.

7.2.3. Phonological form

Like with the *-s* suffix, there may be an important phonological confound. The phonological form of *-ing* is itself variable, in some varieties more so than others. The so-called reduction is probabilistic in nature and has been found to be associated with

the more verbal uses, especially the progressive, and such verbal gerunds that are similar to the progressive. The change of [ɪŋ] to [ɪn] is strongly associated with verbal instances as opposed to nominal ones (Pullum & Zwicky 1988; Houston 1991). If verbal uses are not in fact homonymous with nominal uses, the problem with *-ing* forms would have to be framed very differently. It is possible, however, that the phonological variation is itself gradual and does not imply homonymous uses.

More data on this is required. However, a phonological analysis is beyond the methodological scope of this study. Since the possible heterophony of verbal *-ing* and nominal *-ing* is continuous and correlated with categorical status, it is not expected to be a major confound. In a variety with a conventionalized formal difference, the potential homonymy may be replaced by an alternation between strongly associated forms.

Corpus data includes some evidence for the so-called reduced form; however, this is mostly restricted to fictional text and the spoken part of the corpus and not nearly enough data for the methodology applied here.

7.2.4. Historical perspective

The gerund-participle has been a very active field of research in historical linguistics (De Smet 2014; De Smet et al. 2018; Fonteyn 2016; Fonteyn & Hartmann 2017; Fonteyn 2019a). Especially the nominal gerund has seen a wealth of corpus-based analyses. It is of such high interest because it does not have a counter-part in other Germanic languages, where cognate forms like German *-ung* are clearly nominal, without any obvious verbal features. In Modern English, derived words with the *ing* suffix are less 'nouny' than other nominalizers, such as *-age*, *-(at)ion*, There is no consensus about just how nominal or verbal *-ing* nominals are. *-ing* nominals are extremely versatile with respect to their discourse function (Fonteyn 2016; Fonteyn 2019b). The variation spreads from very verb-like to (proto)-typically noun-like. Iordăchioaia & Werner (2019) argue that *-ing*-nominals have not fully completed what they refer to as a 'nominalization cycle', and propose that this is due to competition with Romance nominals in *-age*, *-al*, *-ance*, *-ion*, and *-ment*. Additionally, they compete with zero-derived nominals and verbal gerunds (Fonteyn 2019b). Those nominals share some of the peculiarities with nominal gerunds, such as being able to head a phrase with objects.

(110) the least I can do **is help you** with it (COCA: FIC)
(111) ?? All I wanna do **is helping you**
(112) ?? **Help me** would be appreciated

Considering that *-ing* forms have gone from a clear two-way distinction to a cluster of formerly three different categories that came to share the same category space, it is not surprising that nominal gerunds became more verbal after the merger (Fonteyn, De Smet & Heyvaert 2015). Consider 2.2.

7.2.5. Inflection

-ing forms are already complex words, and as such do not have as much morphology available as bare forms. This already makes them less typical members of their respective word class than bare forms. Nominal uses of *-ing* can be inflected. The more strongly lexicalized the nominalization, the more likely it occurs with plural marking, e.g., *buildings*, *meanings*. The plural marker is a strong cue for nounhood, so complex forms are normally unproblematic in classification.

The enclitic *-s* can attach to a non-nominal base if it happens to be the last constituent of the noun phrase. Quantitatively speaking, this is extremely rare. The most likely non-nominal bases are pronouns, pronoun-like phrasal elements (*someone else's*, *each other's*), adjectives and adjective-like elements (numbers), and occasionally adverbials. The clitic appearing on verbs is very unlikely and not attested in the data set.

(113) ?? [People who are fighting]'s weapons

However, it has to be noted that taggers are likely to be biased towards identifying non-nominal bases as nouns, or the clitic of non-nominal bases as copula, which makes the search considerably more difficult.

Therefore, even though the possessive *-s* can attach to a non-nominal base, it is mainly used on nouns, and a strong indicator for 'noun-hood'. On a cline from morphological to syntactic, the possessive *-s* is closer to affixes than, for example, prepositional modifiers.

The possessive clitic is virtually non-existent with most *-ing* forms. The only examples that can be found are deverbal nouns that refer to communities or organizations. Another notable type is part of the constructional idiom [for V-ing's sake].

(114) Ken Wilson [...] is **climbing 's** most famous publisher (BNC: CG1)
(115) they were **racing 's** cannon fodder (BNC: BP7)
(116) he was just asking for **asking 's** sake (BNC: K8V).

The choice of the possessive clitic over the *of*-genitive is contingent on the animacy of the possessor (Rosenbach 2003); therefore, it makes sense that those verbal clitics are a bit more common. However, the reliability of the orthographic representation in the spoken parts of the corpus is unclear, and for the written parts, the distinction between clitic and long form may often be lost or does not apply for higher registers. Therefore, verbal clitics have to be treated the same as their 'long' form for the purpose of this study.

7.2.6. Derivations and other minor forms

-ing forms can themselves be roots of derivations; however, few derivational suffixes occur in actual use. For example, it is unlikely to find action nominals derived with the *-er* suffix. More complex derivations exist, such as *meaninglessness*. In the BNC, such examples are mostly restricted to derivations of *meaning*, a strongly lexicalized deverbal noun. Even in the larger COCA, very few examples exist. Apart from words derived from *meaning*, only two other examples exist in either corpus. Both of the following examples are hapax legomena:

(117) It is a focused devotional **feelingfulness** , a self-aware , non-naming amplification of faith (COCA-S: FIC)
(118) it is also a comment about the **caringlessness** about modern society (BNC: HPG)

All examples found are cases of nominal derivation. Verbal derivational suffixes are absent, which is not surprising since the *-ing* suffix de-verbalizes its root and the most verby uses, such as in the progressive constructions, are invariant. Deverbal nouns can be found with suffixes that are productive in spoken English (cf. Plag, Dalton-Puffer & Baayen 1999).

(119) I think we 're looking at the **beginning-ish** of November (COCA-S: SPOK)

-ing forms, therefore, are rather unproductive as bases, or at least the frequency of their derivations is too low to judge their productivity independently due to sample size restrictions. This alone makes them untypical for both nominal and verbal bases. Recursive derivation, in general, can be assumed to be rather unproductive.

There is a range of deverbal prepositions that are grammaticalized *-ing* forms, (Huddleston & Pullum 2002: 611). Some of these have no corresponding bare form, such as *according*, while others have no transparent semantic connection to their homonymous base, such as *during*. Individually, those forms are frequent, but they have a very low type frequency and are not expected to cause any clustering effects. Similar to quantifiers and pronouns in the last chapters, they can serve as a benchmark for very untypical forms.

Some plural forms are frozen and lack a singular form, which are referred to as 'Pluralia tantum nominal gerunds'. Examples include *belongings*, *goings-on* and *surroundings*. Researchers have not been able to find any semantic class that is correlated with this type (Mackenzie 2019; Acquaviva 2008: 16). It is likely a pattern that lacks a conceptual basis, at least in modern English (cf. Chapter 6).

Another very small class of *-ing* forms is those of the form V+*ing*+*s*-*PARTICLE*. As in *goings-on*, *sendings-off*, *castings-on*, *crossings-out*. Those are minor in frequency, even in the COCA, and cannot be meaningfully fitted by a Zipf-Mandelbrot Model (LNRE: X2=2.92, df=3, p=0.40). Therefore, it is unclear whether they form a productive class on their own.

7.3. Corpus analysis

7.3.1. Preparation

Due to the nature of the phenomenon in question, PoS tagging was not a viable annotation strategy. Likewise, lemmatization proved unreliable (cf. see Figure 4.4). Instead, a more basic strategy was employed based on regular expressions, with additional filtering by UD dependency relations and NER tagging, which proved very accurate in the detection of names ending in *-ing*. In contrast to the previous chapters, the population of types consists of the complex word form, and not lemmas or stemmed forms. To obtain the population of *-ing* types, the following heuristics were applied to clean the data set:

1. Split hyphenated types
2. Remove double and triple consonant grapheme onsets for 2- and 3-character stems
3. Remove any tokens containing special characters
4. Filter NER tags indicating the name of a person or organization

Cleaning short matches proved to be most efficient since false positives are extremely rare with longer matches due to multi-syllabic monomorphemic forms ending in *-ing* being extremely rare and mostly part of proper names. Tokens with special characters other than hyphens are most often tokenization errors or rare spelling variants, so they were discarded.

The distributional variables under investigation are summed up in 7.1.

Table 7.1.: Coding scheme for *-ing* dependency relations

Group	UD tags
PROGRESSIVE	aux (dependent)
HAS-OBJ	obj (dependent)
ADVERBIAL	advcl, acl
SUBJECT	nsubj, nsubj:pass
OBJECT	obj
OBLIQUE	obl, iobj, appos, vocative, expl, dislocated
MODIFIER	amod (dependent)
INDEFINITE	det (dependent) word forms: *a*, *an*
DEFINITE	det (dependent) word forms: *the*, *this*, *these*, *that those*
DETERMINER	det (excluding INDEFINITE and DEFINITE)

The variable HAS-OBJ was independent of whether the *-ing* form itself appeared as an object, allowing for mixed forms to be correctly classified as mixed. In general, there is a lot of overlap between the uses, unlike in the previous data sets. The potential for most of the class defining features to co-occur is itself a feature that makes *-ing* forms untypical and seem like an intermediate category. However, this potential does not entail that lexical items are actually balanced between the class-defining uses.

The class OBLIQUE contains many different types of nominal uses other than uses as object argument. Some of these are themselves ambiguous. Most notably, expletives like *fucking*, *flipping*, etc., could also be described as adjectives in many cases. However, only a small proportion of types had dominant uses like this, and this is not expected to be a major confound for the entire distribution. The ambiguity between COP and AUX tags, and the tagging error was too high. Due to this and its generally low frequency, COP was not used for category selection. Uses of *-ing* as indirect object were very rare and restricted to strongly lexicalized deverbal nouns. The only possible exceptions for this can be found in (120):

(120) a degree of blandness **gives the playing** a slightly ‘ automatic ’ quality (BNC: BMC)

7.3.2. *-ing* on the noun-adjective-verb continuum

-ing forms are extremely abundant and extremely productive. About 2% of the tokens in the BNC are *-ing* forms and about 10% of the distinct lemmas that do not include special characters have associated *-ing* forms. The present progressive is the most common use in the BNC. Over 15% of all tokens in the data set are instances of the present progressive.

Figure 7.2 shows a 3-dimensional view on the same data. As in Chapter 6, the three contour levels represent the 25th, 50th and 75th percentiles. The three dimensions account for 88.76%. Table 7.2 shows the variance explained by each dimension. The dimension whose variance is dominated by the distinction of progressive and adverbial uses (y-axis) contributes least. Therefore, most variation can be found across three poles that roughly represent nominal, adjectival and verbal uses. Most lemmas can be found in either the nominal cluster or the verbal cluster. Furthermore, lemmas that are associated to adjectival uses are most distinct. Since the gerund-participle debate revolves around the difference between verbal and nominal uses, this is an expected result. The reason between nominal and verbal uses is diffuse and there is no real separation. Many lexemes are rather well-balanced between verbal and nominal uses, however few lexemes exist that are balanced between nominal and adjectival uses or verbal and adjectival uses.

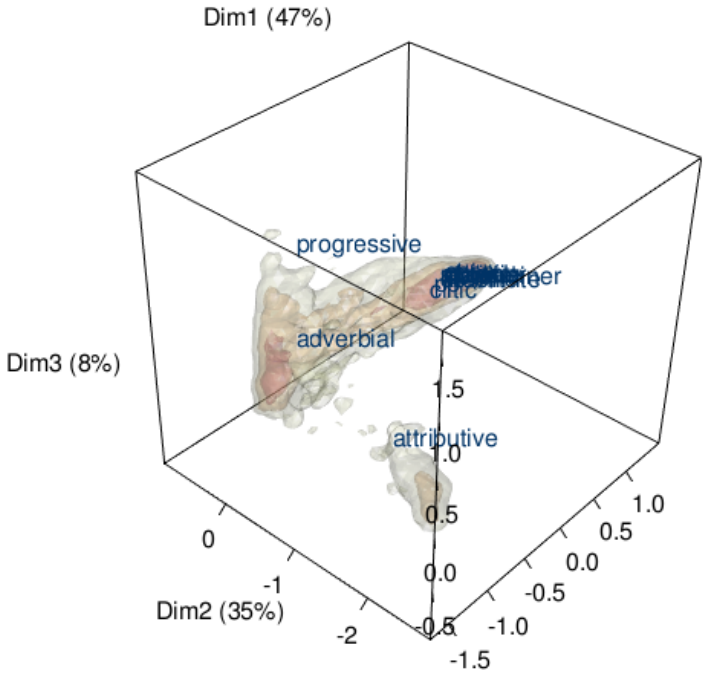


Figure 7.1.: morphological and syntactic features of *-ing*: 3-dimensional view on Correspondence Analysis with 3-dimensional KDE. The correspondence analysis is based on the U_r adjusted frequencies. The colored contours represent the 25th, 50th and 75th percentiles.

Table 7.2.: Correspondence Analysis, of noun features: Columns

name	mass	quality	inertia	k=1	COR	contr.	k=2	COR	contr.
modifier	46	844	96	1098	556	126	-790	288	96
definite	31	808	72	1039	443	76	-945	366	92
indefinite	13	757	36	1066	409	35	-984	348	44
determiner	5	592	15	1068	330	12	-954	263	14
attrib.	121	995	288	1011	411	280	1205	584	588
subject	19	773	41	1069	504	49	-780	268	39
object	22	825	54	1137	500	64	-917	325	62
oblique	27	833	57	1096	551	74	-786	283	57
has_obj	268	491	120	-470	476	134	-83	15	6

name	mass	quality	inertia	k=1	COR	contr.	k=2	COR	contr.
adverbial	303	632	67	-381	630	100	22	2	0
prog	145	144	153	-392	140	51	67	4	2

In a next step, the adjectival uses were unfocused by means of Subset Correspondence Analysis (Greenacre & Pardo 2006). The idea is to keep the overall structure of the data, i.e., maintain the relative frequencies and associations of adjectival uses. Thus, the numeric results in Table 7.2 are not affected. Removing the data would simply create a large 'rest' category and/or simply result in information loss. With Subset Correspondence Analysis the relevant variables can be focused on by re-scaling the data. Even though the first perspective showed that there is a clear tendency for lemmas to be concentrate at the ends of the verb-adjective-noun continuum, the variation within *-ing* forms is immense. Forms with strongly lexicalized nominal, adjectival, or even prepositional uses also have corresponding verbal and gerundial uses.

(121) a player/coach position at Exeter City and that **'s obviously interesting him**
(122) Arune **keeps amazing me** with her quickness and eagerness to learn

Subset CA allows modeling how having a strong adjectival, prepositional, or nominal homonym affects the distribution of the verbal and gerundial uses of the same lexeme. Figure 7.2 shows the result of excluding attributive uses from the subset of variables in the described fashion. The result shows two clearly distinct clusters. The top cluster contains *-ing* types with adjectival uses and types describing properties. An example for this is *rising*, which has frequent adjectival uses, and when used as a verb often describes long-term trends or processes that are time stable and more property- than event-like. These types are moderately repelled from nominal morphology, however. The second cluster contains the largest mass of types, and interestingly is not strongly associated to the progressive. This is due to a few types of extremely high frequency that are significant collexemes of the progressive constructions. Among these are *going*, *doing*, and *talking*. For *going*, this is partly due to the fact that it occurs as an auxiliary in the going-to future. Like in the data presented in the previous case studies, more grammatical and otherwise unusual lexemes (e.g., *during*) reliably show as outliers in the category space. The nominal side from this perspective is not so clear-cut. The nominal features still cluster reliably, but the types do not concentrate in one particular area.

To focus the perspective even more, additional features were collapsed. The different types of determiners are highly correlated and did not show any continuous effect in the expected way (definite < indefinite < other). With verbal uses on one side and nominal uses on the other, the different determiners were actually arranged orthogonally to the noun-verb axis. Another distinction that did not yield any interesting results is that

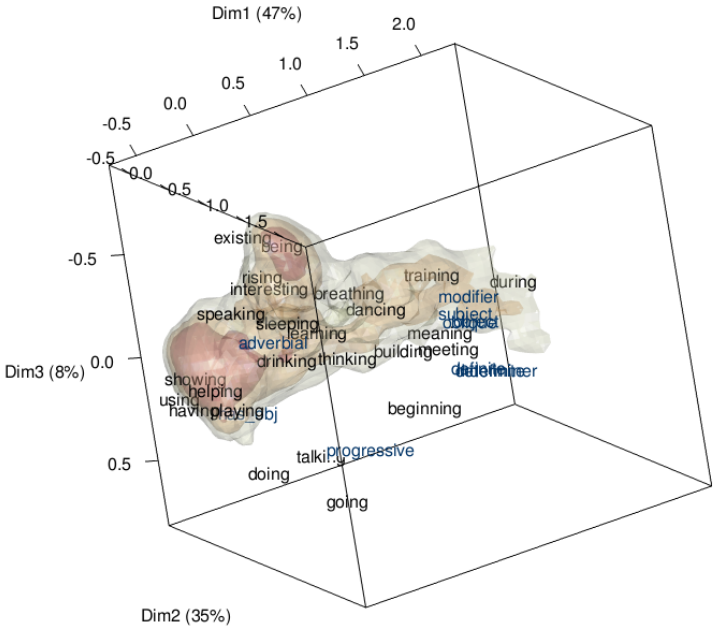


Figure 7.2.: subset of morphological and syntactic features of *-ing* defocussing adjective uses: 3-dimensional view on Subset Correspondence Analysis with 3-dimensional KDE. The correspondence analysis is based on the U_r adjusted frequencies. The colored contours represent the 25th, 50th and 75th percentiles.

of different types of syntactic functions on the nominal side. Uses as subject, object, and oblique were also highly correlated. Consequently, these were dropped from view with only the most distinct class of SUBJECT remaining in the subset.

After this folding of dimensions, the result is less diffuse. The results can be seen in Figure 7.3. The lemmas are now clearly distributed over 3 distinct clusters. One cluster is strongly associated with verbal uses, especially those where the *-ing* form has an object. Interestingly, having an object and nominal features are most distinctive. The means that *-ing* types, despite the fact that mixes of nominal and verbal syntax are possible, generally tend to be used with either. Nevertheless, there is a continuum of types that are balanced between those two poles. Dominantly progressive types are still distinct from other uses. These types can be interpreted as the most verb-like types.

There are two smaller modal regions between verbal and nominal uses. It is tempting to take these as evidence for verbal and nominal gerunds, however, they are not distinctive enough to be significant. The only significant modes (yellow contours), correlate with the three main regions described above. Interestingly, the significant modal region on the nominal side is shifted concentrated at the more verbal end of the

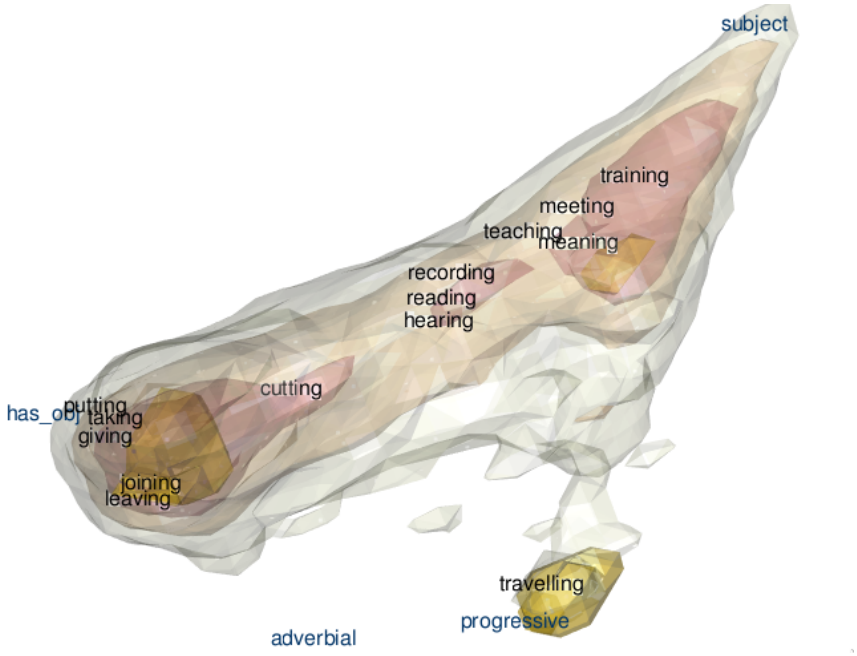


Figure 7.3.: Nominal morpho-syntactic features: 3-dimensional view on Correspondence Analysis with 3-dimensional KDE and significant modes (yellow contours).

nominal cluster. The outermost distinct nominal lemmas include *meaning*, *training*, *building*, which are mostly highly lexicalized deverbal nouns.

The overall picture suggests that there are four main clusters of *-ing* types, depending on the features that are focused. The most distinctive cluster is that of deverbal adjectives. Lemmas associated to adjectival uses generally have the least overlap with other lemmas. There is a clearly nominal cluster; however, this cluster includes both deverbal nouns and types associated to uses as nominal gerund. The verbal side is between a small group of *-ing* types that is highly associated to uses in the progressive construction. Finally, the main group contains other verbal uses, mostly uses with an object.

7.3.3. Revisiting nouns

Taking a closer look at the subclass of *-ing* nominalizations shows a clear split between typical nouns where the nominalization is fully lexicalized. Among such types are the frozen forms *ceiling*, *morning*, and *evening*. They show almost exactly average attraction to the plural form and average dispersion. The same behavior is shown by words like *building*, *setting* and *wedding* that are as strongly lexicalized, but have transparent bases.

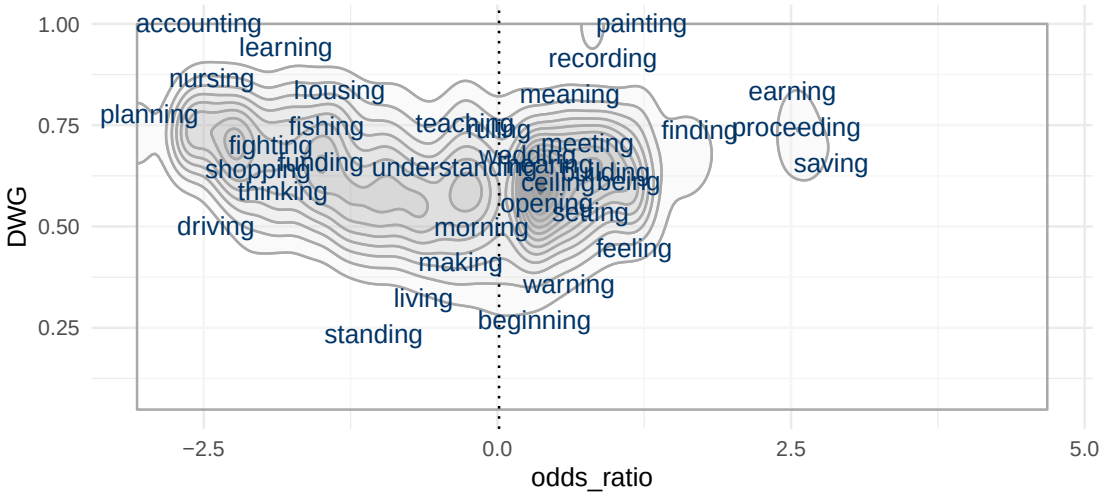


Figure 7.4.: Odds ratios of nominal *-ing* versus *-s* against DWG with weighted 2-dimensional KDE

The repelled side of this continuum contains multiple modes and a much fuzzier picture, which is congruent with the results from the CA. Most *-ing* forms that otherwise occur with nominal morphosyntax are repelled to the plural form in varying degrees. A small group of plurale-tantum nominal gerunds can be found at around x=2.5. On a constructional level—contrasting singular *-ing* with plural—, there seems to be a separation.

7.3.4. Revisiting adjectives

Finally, to focus on adjectival uses of *-ing* forms, I will revisit the adjective data from section 5. *-ing* competes with the other English participial form *-ed* which I have already investigated in the context of adjectives. In Section 5, *-ed* adjectives showed a clear split between lemmas that were associated with attributive uses and lemmas associated with predicative uses (cf. Figure 5.14). The same data can be used to investigate how *-ing* contrasts.

Figure 7.5 shows the distribution of *-ing* adjectives in the same way as Figure 5.14. Interestingly, the distribution of *-ing* adjectives shows three clear peaks, hinting at least at three separate clusters. The first and last peaks at roughly x=-0.2 and x=1.4 match the pattern observed on *-ed*. The attributive lemmas represent time-stable properties, such as *corresponding*, *ongoing*, *missing*, *promising*. The predicative side features less stable psychological properties, such as *willing*, *surprising*, *tempting*, *confusing* (cf. Tables 7.3 and B.2). However, there is a third peak that is much more prominent at around x=0.6. At its center lies the most frequent and one of the most regularly dispersed *-ing* adjectives *interesting*. This form is very strongly lexicalized.

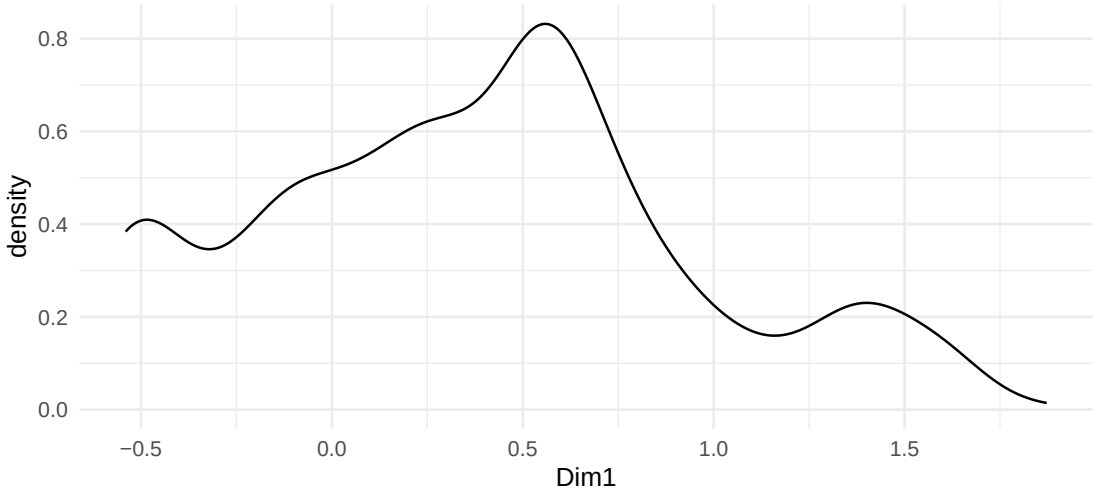


Figure 7.5.: Odds ratios of nominal *-ing* versus predicative uses against DWG with weighted 2-dimensional KDE

Yet it is more associated to predicative uses than typical adjectives. In fact, it occupies almost exactly the center of the distribution. Considering that linguistic categories tend to be as distinctive as possible, the intermediate clustering of *-ing* adjectives cannot be a coincidence. It seems that *-ing* adjectives tend to switch very flexibly between the two major functions of modification and predication, bridging the gap between the two.

Another pattern that can be observed in the data lies in the relative distribution of competing *-ing* and *-ed* forms. When there are forms with the same bases, the *-ed* form tends to be more strongly associated to the predicative function. Among those pairs that have frequencies >50, there is only one case where the opposite can be observed (*englighten*/*englightening*, cf. Figure 7.6). Like *interesting*, *interested* is central to its own cluster, which features the most distinctly predicative adjectives. The general impression is that *-ing* adjectives are repelled from the predicative region that *-ed* forms occupy. Exceptions include *complicated*/*complicating* and *willed*/*willing* where either form is very strongly bound to certain constructions, like *complicating factor*.

While the predicative cluster and the main *-ing* cluster are separated very well, the left side of the distribution is rather diffuse and does not show a clear valley. There can be several explanations for this. Firstly, it could be a caused by general noise in the data and a bias for *-ing* forms to be over-represented on the attributive side of the spectrum. Secondly, it could be a sign of a range of constructions that are not sufficiently covered by the annotation scheme that cause a stronger bias for some *-ing* forms and not others. Since the data set was drawn on the basis of PoS tagging. There may be a systematic lack of predicative uses of *-ing* forms for some lemmas that are more likely to be tagged as verbs than others. The distinction between predicative

Table 7.3.: Top 30 best dispersed V-ing sorted along the modification-predication axis (Dimension 1, Correspondence Analysis)

lemma	Dim1	Dim2	f	dp_norm
willing	1.5942136	0.0419464	2649	0.5549819
surprising	1.4060261	0.0077907	3532	0.4923871
unwilling	1.3513761	0.0527819	609	0.7963051
tempting	1.2247806	0.0647569	449	0.8321303
misleading	1.0869198	0.0068762	835	0.7565023
disappointing	0.9735615	-0.0361254	831	0.7863114
confusing	0.9331732	0.0664576	555	0.8286523
boring	0.7569824	0.0595457	1105	0.7829273
damaging	0.7508408	-0.0832552	724	0.8047006
frightening	0.7204205	0.0147852	688	0.8140138
embarrassing	0.7069016	0.0489833	766	0.7817310
forthcoming	0.5786421	0.0858111	1057	0.7255175
disturbing	0.5651924	-0.0242673	700	0.7958845
interesting	0.5451981	-0.0117249	7969	0.4441303
convincing	0.5312242	-0.0413720	862	0.7586876
amazing	0.3675554	0.1126879	1419	0.7196095
exciting	0.3390843	0.0305410	2424	0.6346581
alarming	0.3011929	0.0072783	434	0.8395282
fascinating	0.2808483	0.0780195	1205	0.7256330
charming	0.2427980	0.0753554	953	0.8069943
devastating	0.1821226	0.1056511	551	0.8031754
appalling	0.1306613	0.0572632	784	0.7843618
striking	0.1276750	-0.0731849	975	0.7583707
promising	0.0819103	-0.0758642	730	0.7783095
overwhelming	-0.0234788	0.1633538	1003	0.7109097
missing	-0.0608428	0.0987153	667	0.8114348
outstanding	-0.1078816	0.0173736	2056	0.6549164
ongoing	-0.1847976	0.1723579	479	0.8413746
underlying	-0.5294295	0.1712191	1435	0.6886146
corresponding	-0.5314654	0.1917074	741	0.8041068

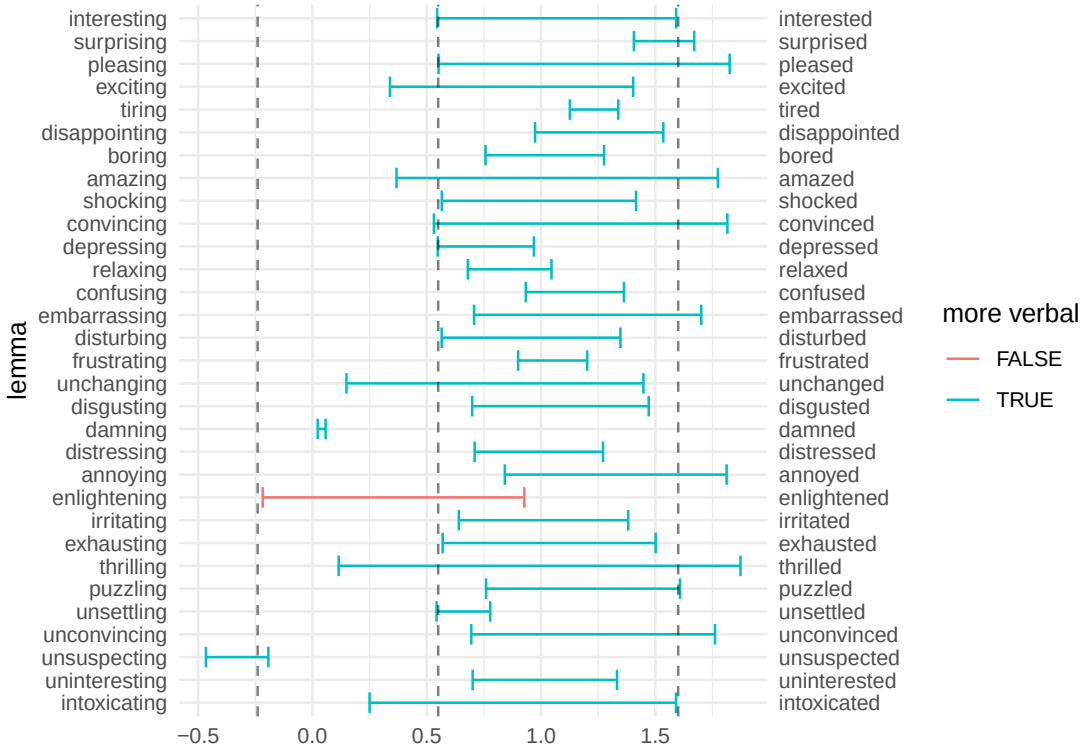


Figure 7.6.: Differences in attributive and predicative uses of bases that have both *-ed* *-ing* occurrences with f>50. Blue lines indicate that the *-ing* form is more verbal, red lines that it is more adjectival.

adjectives in *-ing* and the progressive construction is difficult and inherently fuzzy. Thirdly, it could be a sign of multiple distributions that are distinguished by factors external to the ones measured in the data. Some *-ing* adjectives may be more strongly lexicalized as adjectives than others, and in transition. It is interesting to note that in this region of constructional overlap, there is little lexical overlap between the bases. Figure 7.7 shows a conditional density plot that visualizes how the proportion of competing forms varies.

In regions where a form has a corresponding form with the other suffix, the green density region is larger. The picture confirms the pattern observed in 7.6. The predicative cluster is mostly made up of *-ed* forms that have a corresponding *-ing* form. As little as 5% of the lemmas are *-ing* forms without a corresponding *-ed* form, even though there is about a 50-50 split overall. On the other side, this is not the case. At around x=0.0, only about 10% of the forms have a corresponding form attested in the data set.

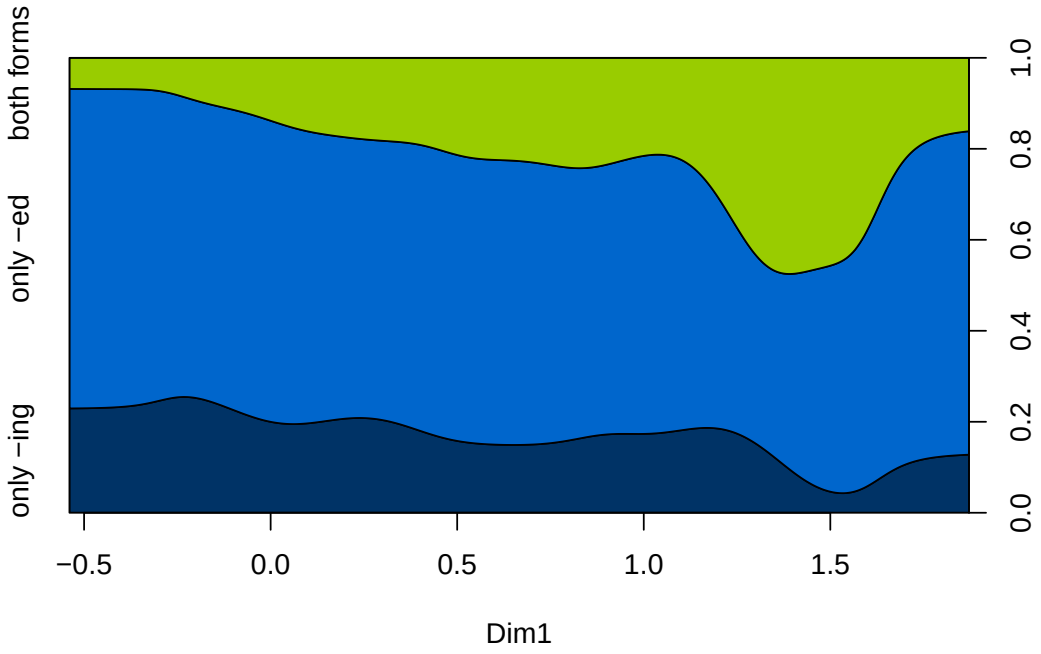


Figure 7.7.: Conditional density of bases with -ing and/or -ed forms

7.4. Summary

The presented data attempted to determine the number of distinct clusters that *-ing* forms produce in the BNC. Based on the features selected for this endeavor, the method was able to identify the expected clusters of nominal, verbal and adjectival types. In terms of in-between categories, such as participles, verbal and nominal gerunds, no significantly distinct clusters could be found between the most nominal and most verbal types. However, the uncertainty in the data is rather high due to the lower frequencies of *-ing* forms compared to both nouns and adjectives in previous chapters. Larger data sets may be required to identify more coherent clusters, and especially ones based on much smaller scale variation. In general, the overlap between the clusters was large. However, the general opposition of verbal and nominal types, provides some evidence against a gerund-participle.

A number of conclusions can be drawn from the results. Surprisingly, the progressive construction compared to related adverbial constructions and other non-finite clauses is not the pole of the strongest concentration of associated types. This is due to highly frequency types, such as *having*, *doing*, and *going* that have more grammatical uses in auxiliary constructions. There is also a smaller group of progressive types

that seems distinct from other participial uses. Despite potential differences between participles in the progressive and in non-finite clauses, there is no indication for separate groups. By taking into account more features specific to this distinction, preferences between types may appear on a more specific constructional level. From a more global perspective of the word class continuum, there is only one distinct cluster (cf. Figure 2.3).

The clear separation between deverbal adjectives and other *-ing* forms may not be too exciting on its own. However, adjective types were on the opposite of the verbal types in feature space with other non-finite uses in the middle. This demonstrates once again that a triangular model of the noun-adjective-verb continuum is more suitable. Furthermore, radically reductionist approaches such as Duffley (2006) lack a quantitative basis since adjective types seem to exhibit some degree of prototype effect. Contrasting adjectival uses of *-ing* yields strong evidence for Givón (1980)'s time-stability scale.

Any clustering between the verbal and nominal pools was inconclusive. The two distributions of more verbal *-ing* types and more nominal *-ing* types rather show a smooth continuum. Neither the determiner hierarchy nor the syntactic subject-object-oblique hierarchy showed the expected gradient from nominal to verbal in the case of *-ing* forms. Deverbal nouns are distinct from more verbal forms when focusing on nominal features. However, from a more global perspective, the transparent homonymy/polysemy of words like *meaning*, *building* and *training* makes the distributions collapse into one more generally nominal *-ing*. A difference between types that prefer nominal gerund constructions and deverbal nouns is not apparent from the data. Deverbal adjectives and deverbal nouns, on the other hand, are clearly distinct, solely based on their morphosyntactic distribution, and show prototype effects.

8. Conclusion

8.1. Empirical results

The data presented in this thesis confirms several empirical findings from previous studies on the word class continuum. It also offers additional insight. In Chapter 5, I showed that a 2-dimensional model of a noun-verb-adjective continuum is preferable to a 1-dimensional one, at least in English. Adjectives in the data set varied mostly between modification and predication, but the noun-verb opposition appeared to be orthogonal and not opposite. Adjective lexemes were found to be more similar to nouns than verbs, with the exception of *-ing* forms that are more 'verby'. The studies also confirmed once more that there is a clear correlation between conceptual and morphosyntactic properties of lexemes. Deverbal *-ed* showed a very distinct pattern. The vast majority of predicative uses of *-ed* described are lexemes describing psychological state, while those used primarily for modification showed a mixture of abstract properties, including many socio-culturally evaluative properties pertaining to humans, or their lifestyle (see 5.5). The continuum from verbal to adjectival uses can be explained very well with Givón (1980)'s time-stability scale.

The investigation into nominal subcategories in Chapter 6 revealed well-defined clusters of count nouns, mass nouns and proper nouns, based on selected morphosyntactic properties. Pluralia tantum did not show any clear tendency to form clusters. Plurality in general caused a lot of variation in both count and mass nouns. In both cases, a cline of associatedness to plural forms was suggested. These clines represent an overlapping network of noun populations that range from constructionally restricted singular-dominant lexemes over a diverse group of core members to plural-dominant lexemes, most notably lexemes describing groups of humans. Pluralia tantum are likely not distinctive enough from other plural-dominant forms to form a separate class, which is in line with previous studies.

Finally, the question of *-ing* forms mostly confirmed established classes. A group of deverbal adjectives showed to be clearly distinct. The rest of the lexemes was split between nominal and verbal uses, with separate small group of strong collexemes of the progressive construction. The other nominal and verbal uses showed a very strong overlap, pointing to a continuum of forms, rather than in-between categories associated to either verbal gerunds or verbal nouns. The distinction between deverbal nouns and other strongly associated nominal types was also non-existent, except when

focusing on the plural inflection. The clear verbal-nominal opposition of types can be interpreted as evidence against a gerund-participle.

8.2. Not all structure is meaningful

The results of the case studies show that purely formal morphosyntactic features do not show sufficiently distinctive clustering behavior in many cases. At the same time, the data emphasizes why gradience is such an important property for the description of grammatical features. From a more global perspective of word classes and their strongest morphosyntactic markers, there is little evidence for prototype formation of concepts like plurale-tantum nouns, and verbal/nominal gerunds. The confusion about gerunds and participle is symptomatic of the diffuse distribution of *-ing* types. Between the three populations that were investigated, the *-ing* types show by far the most overlap.

Plurale tantum nouns show very specific generalizations. In some of these cases the generalization can be attributed very clearly to a dominant lexical exemplar (trousers, glasses) from which the behavior is analogically derived. *trousers* has no singular so words describing trouser-like object also have no singular. There is a realistic possibility that pluralia tantum form a grammatical structure without functional motivation.

Given the sheer size of the lexicon in a language, there are bound to be plural-only words for reasons of probability alone. Those cases need to be frequent enough to be negatively entrenched as lacking their bare form (cf. Stefanowitsch 2008). In that regard, it is worth noting that nouns have the highest token and type frequencies among word classes, which would increase the chance of a statistical plurale tantum. It could also explain why there is no counterpart in other word classes. There may also be functional reasons for why verbs do not pattern like this. Event descriptions are often relative to the speaker's perspective and may be differently framed. The present tense is regularly used to refer to past events. Plurality, on the other hand, is no matter of perspective. Countability can be subject to varying types of framing since most mass nouns that are usually conceptualized as unindividuatable can be pluralized productively to refer to multiple types or instances of the same concept, e.g., consider (58)-(60).

8.3. Methodological contributions

This thesis has presented an approach to lexical categories that capitalizes on a wealth of well-established descriptive and analytic techniques rather than being restricted to just association or just productivity. It followed a tupelized approach (Gries 2021), by combining measures of association, productivity and dispersion. Some measures, such

as adjusted frequencies, proved useful to mitigate the influence of high-frequency items, and also to correct for bursts of repetitions. Distance-based dispersion, in addition, uncovered interesting patterns, such as the tendency of more grammatical items to be more well-dispersed. The value of odds ratios, both as tool for visualization and as a statistical measure, was demonstrated. Due to its straightforward distribution, it is well suited for the task of identifying type clusters. Furthermore, heuristic cut-offs of low-frequency items that are often used are far from ideal from a conceptual point of view. Low-frequency words and constructions are a reality and contribute to the overall structure of linguistic categories, as the German saying goes: "*Kleinvieh macht auch Mist*". This thesis presented various ideas on how to deal with low-frequency items in a principled way. Some of the methods were experimental and show a promising direction for future application.

Another key takeaway from the case studies is that elicited corpus samples are heavily multi-modal in the statistical sense of having multiple modes; i.e., the data presented has multiple peaks in the distribution of its values. Chapters 5-7 demonstrate how many different distributions are hidden behind as inconspicuous a label as 'adjective'. Neither more accurate PoS tagging nor manual annotation can solve this problem as it is built into the nature of language. The annotation techniques that are common in Corpus Linguistics rarely capture homogeneous groups. Especially, RCG provides a theoretical explanation for this, which is present to a somewhat lesser degree in other CxG approaches. Traditional word classes are not as homogeneous as they are often assumed to be, even if subcategorization is taken into account. This is especially problematic with PoS-tagging as one of the most common classifications in corpus linguistics. PoS-tagging and related annotation methods are often based on manually annotated training sets, and dictionaries that are external to the data. The word-class bias is, therefore, built into Corpus Linguistics at its current state. Multimodality is a violation of the underlying assumptions of most statistical models. However, the results from the density analyses are encouraging. The central tendency of lexeme groups relative to distributional properties can be unearthed, and the major categories and their subcategories converge on distinct lexical statistical properties once frequency and different kinds of dispersion are accounted for. Multivariate clustering approaches may be a method flexible enough to extract more homogeneous groups of lexemes, e.g., mainly mass nouns. There must be more focus on how the right population is drawn from a corpus. Multivariate clustering techniques, e.g., via model mixtures seem promising in that regard. Assigned clusters could also be used for selection or as variable (random or fixed effect) in regressions.

The presented methodology is well-suited for both reductionist and non-reductionist approaches. The methodology can be further refined and extended in several ways. Residuals of the density modes could potentially be used to identify prototypical exemplars. Validation against external data is required. The data emphasizes that multivariate factors such as association, part-based dispersion, distance-based dispersion and productivity, all contribute to the overall structure of the data in non-obvious ways. In the case studies in this thesis, I presented a combination of measures that

were chosen to be intercorrelated as little as possible. The weaknesses of such a methodology are not unique within Corpus Linguistics. The data is very noisy, and especially automatically processed type frequencies become unreliable. This is especially problematic if identifying prototype clusters is the goal (or identify statistical multi-modality in general). Manual annotation is often still required (also see Lüdeling, Evert & Heid 2000; Evert & Lüdeling 2001). Automatic processing has come a long way, and morphological processing like WordPiece shows promising results in practical application, but needs fine-tuning for linguistic tasks, and become more accessible and more commonplace. Another weakness of the methodology is that it is very data-hungry. Low-frequency phenomena do not carry enough signal in a corpus of the size of the BNC. Larger corpora, on the other hand, are noisier and less well-balanced. Chapters 5-7 provided a qualitative overview of many structures that are simply to rare to make equal impact on all lexemes of a class. The uncertainty of a structure's dispersion is also rather high, leading to a very blurry picture. Phenomena such as the plurale tantum are difficult to capture by the approach taken.

A significant part of the work for this thesis went into the creation of code for the corpus analyses and the endeavor to turn it from one-off scripts into robust, reusable, and accessible packages. All statistical measures used in this thesis are implemented and available in an *R* package (*occurR*, Rauhut 2022a) alongside with a selection of data processing tools tailored for corpus linguistic tasks (*linguio*, Rauhut 2023). The packages are written in an idiomatic functional style, fully unit tested, and have virtually no dependencies. All of this hopefully makes the code easier to use, easier to maintain, and therefore more future-proof. Code examples are available in Appendix A.1.

8.4. Theoretical Implications

Another conclusion from the case studies is that classes based on absence of a feature pattern differently from classes based on presence. Both non-inflecting adjectives and non-predicative adjectives showed rather diffuse distributions in terms of their negative association compared to inflecting and predicative adjectives. Mass nouns have a distinct distribution not merely due to the absence of the plural form, they also correlate with a number of other constructions, as opposed to plurale tantum nouns.

There is some evidence that even solely distributionally defined word classes (cf. Croft 2022) do not line up with observable prototype clusters. The data showed that multi-modality, i.e., multiple overlapping distributions of lexical items commonly appear in quantitative data. There needs to be a better theoretical understanding of continuous versus discontinuous prototype categories, and how their overlap forms larger categories. Most traditional CL literature does not describe this issue directly, even though the means to do so are available through prototype, exemplar, and grammaticalization theories. I argued in this thesis that the concept of 'cline' as it is in common usage in the context of historical development is to be distinguished different from the concept

of 'continuum' between linguistic categories. The former does not entail the latter and not all emergent grammatical distinctions produce prototype clusters in the same way.

12 years ago, Stefanowitsch (2011a: 303) remarked that there was a considerable mismatch between CL and cognitive sciences. The same seems to be true right now with Computational Linguistics and Natural Language Processing (NLP). Recent advances in language models have been greatly successful in application with little input from CL or CxG even if concerned with linguistic topics (e.g., Schneider et al. 2021). For example, in a recent review by Church & Liberman (2021), there is no mention of advances in usage-based linguistics even though the overlap between the fields is undeniable (cf. Linzen & Baroni 2021). In the same vain, they acknowledge the lack awareness of other disciplines as a downside of the recent methodology in AI and Computational Linguistics. Linguistics, especially Usage-based Linguistics, can be the key to understanding when and why those methods work, and perhaps more importantly, when they fail. This is especially relevant for non-obvious socially relevant biases concerning gender, race, and social status. Church & Liberman (2021) call it "unfortunate" that both synonyms and antonyms score high similarity scores (2021: 6). Reasons for this are well-studied in linguistics (e.g., Justeson & Katz 1991), and results presented here show that antonyms are likely part of the same prototype distributions that are mixed with others. Awareness of multi-level generalizations and the associated linguistic patterns open up the potential for much more focused fine-tuning of models for specific analytic or practical purposes.

Moving forward, it is of utmost importance to be able to tell apart motivated linguistic patterns from spurious ones. This ability facilitates the falsifiability of research hypotheses related to more abstract functions and schemas (cf. Stefanowitsch 2011a on falsifiability). The underlying distribution of linguistic categories has to be a central focus. Theoretical concepts, such as prototype clusters, clines, and networks, need to be connected to the data. There are direct analogs in data science and statistics, and techniques to model such phenomena (regression, cluster analysis, graphs etc.). The data presented in this thesis shows aspects of all of these types of gradience and how they may be modeled. Prototype clusters can be found in the convergence of distributional and statistical properties of lexical items. Such prototype clusters overlap with other distributions, forming a multidimensional network. Absence of clustering can also be found as the continuous correlation of lexemes along a number of dimensions. Empirical data, including corpus data and multivariate techniques cannot only help in the delineation of linguistic categories, but also consolidate seemingly contradictory theoretical models by localizing competing motivations in cluster formation. Multiple classifications can accurately describe the same data set depending on the functional dimension they highlight. The data set showed that data-driven identification of prototype clusters could also potentially aid in the comparison of languages and enhance typological methodology provided that enough corpus data can be compiled and annotated.

Even with many questions left open, it is safe to say that the word-class continuum is real and can be observed in quantitative data. Both essentialist categories and more traditional word classes offer interesting perspectives on the structure of the lexicon. Careful operationalization and selection of distributional properties can uncover the underlying functional structure of lexical categories and their subcategories in corpus data. Central tendencies of categories, i.e., prototypes, are an empirical reality, even in pure Corpus data.

9. Bibliography

Aarts, Bas. 2004. Modelling linguistic gradience. *Studies in Language* 28. 1–49. doi:10.1075/sl.28.1.02aar.

Aarts, Bas. 2007. *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford: Oxford University Press.

Acquaviva, Paolo. 2004. Plural mass nouns and the compositionality of number. *Verbum*. Presses Universitaires de Nancy 26(4). 387–401. http://hdl.handle.net/10197/3898 (19 January, 2023).

Acquaviva, Paolo. 2008. *Lexical plurals: A morphosemantic approach*. Oxford: Oxford University Press.

Agresti, Alan. 2002. *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/0471249688.

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter & Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics (demonstrations)*, 54–59. Minneapolis, Minnesota, USA: Association for Computational Linguistics. https://aclanthology.org/N19-4010.pdf.

Alexiadou, Artemis. 2011. Plural mass nouns and the morpho-syntax of number. In Nina Topintzy, Nikolas Lavidas & Maria Moumtzi (eds.), *Proceedings of the 28th west coast conference on formal linguistics*, 33–41. doi:10.26262/istal.v23i0.7317.

Alexiadou, Artemis. 2019. On plurals and plurality. *Selected Papers on Theoretical and Applied Linguistics* 23. 3–18. doi:10.26262/ISTAL.V23I0.7317.

Ambridge, Benjamin, Adam Bidgood, Kathleen E. Twomey, John M. Pine, Christopher F. Rowland & David Freudenthal. 2015. Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE*. Public Library of Science 10(12). e0123723. doi:10.1371/journal.pone.0123723.

Anderson, Gregory D. S. 2013. The velar nasal (v2020.3). In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Zenodo. doi:10.5281/zenodo.7385533.

Arcodia, Giorgio F. 2014. The Chinese adjective as a word class. *Word classes: Nature, typology and representations*, 95–118. Amsterdam/Philadelphia: John Benjamins. doi:10.1075/cilt.332.06arc.

Arnon, I. & N. Snider. 2010. More than words. Frequency effects for multi-word phrases. *Journal of Memory and Language* 62. 67–87. doi:10.1016/j.jml.2009.09.005.

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In G. E. Booij & J. van Marle (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Kluwer. doi:10.1007/978-94-011-2516-1_8.

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*. John Benjamins 5(3). 436–461.

Baayen, R. Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology*. Routledge 30(11). 1174–1220. doi:10.1080/02687038.2016.1147767.

Baker, Mark C. 2003. *Lexical categories*. Cambridge: Cambridge University Press.

Barsalou, Lawrence W. 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (eds.), *Advances in social cognition, volume III: Content and process specificity in the effects of prior experiences*, 61–88. Hillsdale, NJ: Erlbaum.

Bartsch, Sabine & Stefan Evert. 2014. Towards a firthian notion of collocation. *Vernetzungsstrategien Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern* 2(1). 48–61. https://stephanie-evert.de/PUB/BartschEvert2014.pdf (14 March, 2023).

Beekhuizen, Barend, Blair C. Armstrong & Suzanne Stevenson. 2021. Probing lexical ambiguity: Word vectors encode number and relatedness of senses. *Cognitive Science* 45(5). e12943. doi:10.1111/cogs.12943. https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12943.

Bekaert, Elisa & Renata Enghels. 2019. On the edge between nouns and verbs. The heterogeneous behavior of spanish deverbal nominalizations empirically verified. *Language Sciences* 73. 119–136. doi:10.1016/j.langsci.2018.08.012.

Berg, Thomas. 2000. The position of adjectives on the noun-verb continuum. *English Language & Linguistics*. Cambridge University Press 4(2). 269–293. doi:10.1017/S1360674300000253.

Berg, Thomas. 2015. Locating affixes on the lexicon-grammar continuum. *Cognitive Linguistic Studies* 2(1). 150–180. doi:10.1075/cogls.2.1.08ber.

Bergen, Benjamin. 2004. The psychological reality of phonaesthemes. *Language* 80(2). 290–311. doi:10.1353/lan.2004.0056.

Bevilacqua, Michele, Tommaso Pasini, Alessandro Raganato & Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In Zhi-Hua Zhou (ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence survey track*, 4330–4338. Vienna: International Joint Conference on Artificial Intelligence, Inc. doi:10.24963/ijcai.2021/593.

Booij, Geert. 2005. Compounding and derivation: Evidence for construction morphology. *Morphology and its demarcations*, 109–132. Amsterdam: John Benjamins. doi:10.1093/oxfordhb/9780199695720.013.0010.

Booij, Geert. 2007. Construction morphology and the lexicon. *Selected proceedings of the 5th décembrettes: Morphology in Toulouse*, 34–44. Somerville, MA: Cascadilla Press. https://www.academia.edu/12336371/Construction_morphology_and_the_lexicon (14 March, 2023).

Booij, Geert. 2010. Construction morphology. *Language and linguistics compass*. Wiley Online Library 4(7). 543–555.
Bresnan, Joan. 2001. *Lexical-functional syntax*. Oxford: Wiley-Blackwell.
Brysbaert, Marc, Paweł Mandera & Emmanuel Keuleers. 2018. The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science* 27(1). 45–50. doi:10.1177/0963721417727521.
Bush, Nathan. 2001. Frequency effects and word-boundary palatization in English. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 255–280. Amsterdam/Philadelphia: John Benjamins. doi:10.1075/tsl.45.15ber.
Bybee, Joan. 2002. Phonological evidence for exemplar storage of multiword sequences. *SSLA* 24. 215–221. doi:10.1017/S0272263102002061.
Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733. http://www.jstor.org/stable/4490266 (9 March, 2023).
Bybee, Joan L. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511750526.
Bybee, Joan L. & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4). 575–596. doi:10.1515/ling.37.4.575.
Chacon, J. E. & T. Duong. 2018. *Multivariate kernel smoothing and its applications*. Boca Raton: Chapman & Hall/CRC. doi:10.1080/01621459.2020.1721247.
Chaudhuri, P. & J. S. Marron. 1999. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* 94. 807–823. https://www.jstor.org/stable/2669996 (10 March, 2023).
Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton. doi:10.1515/9783110218329.
Chomsky, Noam. 1970. Remarks on nominalization. In Noam Chomsky (ed.), *Studies on semantics in generative grammar*, 1–52. Berlin, Boston: De Gruyter. doi:10.1515/9783110814231.11.
Church, Kenneth & Mark Liberman. 2021. The future of computational linguistics: Beyond alchemy. *Frontiers in Artificial Intelligence* 4. 625341. doi:10.3389/frai.2021.625341.
Comrie, Bernard. 1975. Polite plurals and predicate agreement. *Language* 51. 406–418. https://www.jstor.org/stable/412863 (14 March, 2023).
Corbett, Greville G. 1978. Universals in the syntax of cardinal numerals. *Lingua* 46(4). 355–368. doi:10.1016/0024-3841(78)90042-6.
Corbett, Greville G. 2019. Pluralia tantum nouns and the theory of features: A typology of nouns with non-canonical number properties. *Morphology*. Springer 29(1). 51–108. doi:10.1007/s11525-018-9336-0.
Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago/London: University of Chicago Press.
Croft, William. 2000. Parts of speech as language universals and as language-particular categories. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes*, 65–102. Berlin/New York: De Gruyter. doi:10.1515/9783110806120.65.
Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press. doi:10.1093/acprof:oso/9780198299554.001.0001.

Croft, William. 2003. *Typology and universals*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511840579.
Croft, William. 2022. Word classes in radical construction grammar. In Eva van Lier (ed.), *Oxford handbook of word classes*. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199281251.013.0015.
Davidson, Lisa. 2017. Cross-language speech perception and production. *Oxford Bibliographies in Linguistics*. doi:10.1093/obo/9780199772810-0152.
Davies, Mark. 2008. *The corpus of contemporary American English: 450 million words, 1990-2012*. http://corpus.byu.edu/coca.
De Smet, Hans. 2008. Functional motivations in the development of nominal and verbal gerunds in middle and early modern english. *English Language and Linguistics* 12(1). 55–102. doi:10.1017/S136067430800001X.
De Smet, Hendrik. 2014. Constrained confusion: The gerund/participle distinction in late modern English. In Marianne Hundt (ed.), *Late modern English syntax* (Studies in English Language), 224–238. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139507226.017.
De Smet, Hendrik, Frauke D'hoedt, Lauren Fonteyn & Kristel Van Goethem. 2018. The changing functions of competing forms: Attraction and differentiation. *Cognitive Linguistics*. De Gruyter 29(2). 197–234. doi:10.1515/cog-2016-0025.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Thamar Solorio Jill Burstein Christy Doran (ed.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423.
Diessel, Holger. 1999. *Demonstratives : Form, function, and grammaticalization*. Amsterdam/Philadelphia: J. Benjamins.
Diessel, Holger. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics* 43(3). 449–470. doi:10.1515/ling.2005.43.3.449.
Diessel, Holger. 2016. Frequency and lexical specificity in grammar: A critical review. In Heike Behrens & Stefan Pfänder (eds.), *Experience counts: Frequency effects in language*, 209–238. De Gruyter. doi:10.1515/9783110346916-009.
Diessel, Holger. 2019. Usage-based construction grammar. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Cognitive linguistics - a survey of linguistic subfields*, 50–80. Berlin, Boston: De Gruyter Mouton. doi:10.1515/9783110626452-003.
Dixon, Robert M. W. 1977. *Where have all the adjectives gone?* Berlin: De Gruyter.
Dixon, Robert M. W. 2004. Adjective classes in typological perspective. In Robert M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Adjective classes: A cross-linguistic typology*, 1–49. Oxford: Oxford University Press.
Doron, Edit. 1988. The semantics of predicate nominals. *Linguistics*. De Gruyter 26(2). 281–302. doi:10.1515/ling.1988.26.2.281.
Dowle, Matt & Arun Srinivasan. 2021. *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table (1 March, 2023).
Drożdż, Grzegorz. 2020. New insights into english count and mass nouns – the

cognitive grammar perspective. *English Language & Linguistics* 24(4). 833–854. doi:10.1017/S1360674319000480.
Duffley, Patrick J. 2006. *The English gerund-participle: A comparison with the infinitive*. New York: Peter Lang.
Dunning, Ted E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61–74. doi:10.5555/972450.972454.
Duong, Tarn. 2022. *Ks: Kernel smoothing*. https://CRAN.R-project.org/package=ks (3 April, 2023).
Duong, T., A. Cowling, I. Koch & M. P. Wand. 2008. Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis* 52. 4225–4242. doi:10.1016/j.csda.2008.02.035.
Eckhoff, Hanne M., Laura A. Janda & Olga Lyashevskaya. 2017. Predicting Russian aspect by frequency across genres. *The Slavic and East European Journal* 61(4). 844–875. doi:10.18710/BIIGT6.
Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115. doi:10.1075/ijcl.18010.egb.
Ellis, Nick C. 2007a. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24. doi:10.1093/applin/ami038.
Ellis, Nick C. 2007b. The associative-cognitive CREED. In Bill VanPatten & Jessica Williams (eds.), *Theories of second language acquisition: An introduction*, 77–95. Erlbaum.
Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3). 370–385. doi:10.1111/j.1540-4781.2009.00896.x.
Ellis, Nick C. & Fernando Gonçalves Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220. doi:10.1075/arcl.7.08ell.
Evert, Stefan. 2004. A simple LNRE model for random character sequences. In Gérald Purnelle, Cédrick Fairon & Anne Dister (eds.), *Proceedings of the 7èmes journées internationales d'analyse statistique des données textuelles*, 411–422. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.
Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. Friedrich-Alexander-Universität Erlangen-Nürnberg PhD thesis. http://www.collocations.de/phd.html (14 March, 2023).
Evert, Stefan & Marco Baroni. 2007. zipfR: Word frequency distributions in r. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 29–32. Stroudsburt, PA: Association for Computational Linguistics. doi:10.3115/1557769.1557780.
Evert, Stefan & Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. *Proceedings of the corpus linguistics 2011 conference*. University of Birmingham, UK.
Evert, Stefan & Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In Paul Rayson, Andrew Wilson, Tong McEnery, Andrew Hardie & Shereen Khoja (eds.), *Proceedings of the corpus linguistics 2001 conference*, 167–175. Lancaster: UCREL.

Evert, Stefan, Peter Uhrig, Sabine Bartsch & Thomas Proisl. 2017. E-VIEW-affilation–a large-scale evaluation study of association measures for collocation identification. In M. Jakubíçek I. Kosem C. Tiberius (ed.), *Proceedings of eLex 2017 – electronic lexicography in the 21st century: Lexicography from scratch*, 531–549. https://elex.link/elex2017/wp-content/uploads/2017/09/paper32.pdf (28 February, 2023).

Feldman, L. B. 2000. Are morphological effects distinguishable from the effects of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(6). 1431–1444. doi:10.1037/0278-7393.26.6.1431.

Firth, John Rupert. 1957. *A synopsis of linguistic theory 1930-1955*. Oxford: Oxford University Press.

Fonteyn, Lauren. 2016. From nominal to verbal gerunds: A referential typology. *Functions of Language*. John Benjamins 23(1). 60–83. doi:10.1075/fol.23.1.04fon.

Fonteyn, Lauren. 2019a. *Categoriality in language change: The case of the English gerund*. Oxford: Oxford University Press. doi:10.1093/oso/9780190917579.001.0001.

Fonteyn, Lauren. 2019b. A corpus-based view on the (aspectual-)semantics of modern English nominalizations. *Language Sciences* 73. 77–90. doi:10.1016/j.langsci.2018.08.008.

Fonteyn, Lauren, Hendrik De Smet & Liesbet Heyvaert. 2015. What it means to verbalize: The changing discourse functions of the English gerund. *Journal of English Linguistics*. John Benjamins 43(1). 1–25. doi:10.1177/0075424214564365.

Fonteyn, Lauren & Stefan Hartmann. 2017. Usage-based approaches on diachronic morphology: A mixed-methods approach towards English ing-nominals. *Linguistics Vanguard* 2(1). doi:lingvan-2016-0057.

Fu, M. & Shaofeng Li. 2019. The associations between individual differences in working memory and the effectiveness of immediate and delayed corrective feedback. *Journal of Second Language Studies* 2(2). 233–257. doi:10.1075/jsls.19002.fu.

Givón, Talmy. 1979. *Understanding grammar*. New York: Academic Press.

Givón, Talmy. 1980. The binding hierarchy and the typology of complement. *Studies in language* 4. 333–377. doi:10.1075/sl.4.3.03giv.

Givón, Talmy. 1984. *Syntax. A functional-typological introduction, vol 1*. Amsterdam: John Benjamins.

Givón, Talmy. 1985. *Iconicity, isomorphism, and non-arbitrary coding in syntax*. (Ed.) John Haiman. Amsterdam: John Benjamins.

Godtliebsen, F., J. S. Marron & P. Chaudhuri. 2002. Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics* 11. 1–22. doi:10.1198/106186002317375596.

Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, Adele E. 1995a. *Constructions: A construction grammar approach to argument structure*. Chicago: University Of Chicago Press.

Goldberg, Adele E. 1995b. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton, NJ: Princeton University Press.

doi:10.2307/j.ctvc772nn.
Greenacre, Michael & Rafael Pardo. 2006. Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Journal of Official Statistics* 32(2). 195–212. doi:10.2478/v10078-006-0003-1.
Greenberg, Joseph H. 1990. Some universals of grammar with particular reference to the order of meaningful elements. In Keith Denning & Suzanne Kemmer (eds.), *On language: Selected writings of joseph h. greenberg*, 40–70. Redwood City: Stanford University Press. doi:10.1515/9781503623217-005.
Gries, Stefan Th. 2006. Exploring the variability within and between corpora: Some methodologicalconsiderations. *Corpora* 1. 109–151. doi:10.1075/corp.1.09gri.
Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*. John Benjamins 13(4). 403–437. doi:10.1075/ijcl.13.4.02gri.
Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: Further explorations. In Stefan Th. Gries, Stefanie Wulff & Mark Davies (eds.), *Corpus-linguistic applications: Current studies and new directions*, 197–212. Amsterdam: Rodopi. https://stgries.info/research/2010_STG_DispersionAdjFreq_CorpLingAppl.pdf (14 March, 2023).
Gries, Stefan Th. 2015a. More (old and new) misunderstandings of collostructional analysis: On schmid and küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536. doi:10.1515/cog-2014-0092.
Gries, Stefan Th. 2015b. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1). 93–117. doi:10.1177/1606822X14556606.
Gries, Stefan Th. 2015c. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10. 95–125. doi:10.3366/COR.2015.0068.
Gries, Stefan Th. 2019b. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3). 385–412. doi:10.1075/ijcl.00011.gri.
Gries, Stefan Th. 2019a. Polysemy. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Cognitive linguistics - key topics*, 15–36. Berlin/Boston: De Gruyter Mouton. doi:10.1515/9783110626438-002.
Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of Corpus Linguistics*. Cham: Springer International Publishing. doi:10.1007/978-3-030-46216-1_5.
Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33. doi:10.32714/ricl.09.02.02.
Gries, Stefan Th. 2022a. On, or against?, (just) frequency. In Hans C. Boas (ed.), *Directions for pedagogical construction grammar* (Learning and Teaching (with) Constructions), 47–72. Berlin/Boston: De Gruyter. doi:10.1515/9783110746723-002.
Gries, Stefan Th. 2022b. What do (most of) our dispersion measures measure (most)? dispersion? *Journal of Second Language Studies* 5(2). 171–205.

doi:10.1075/jsls.21029.gri.
Hartmann, Stefan. 2019. Up and down the substantivization cline: Response to Bekaert & Enghels. *Language Sciences* 73. 137–145. doi:10.1016/j.langsci.2018.07.007.
Haspelmath, Martin. 1999. Why is grammaticalization irreversible? *Linguistics*. De Gruyter 37(6). 1043–1068. doi:10.1515/ling.37.6.1043.
Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1). 25–70. doi:10.1017/S0022226705003683.
Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687. doi:10.1353/lan.2010.0021.
Haspelmath, Martin. 2012. How to compare major word-classes across the world's languages. In Anna Szabolcsi Thomas Graf Denis Paperno (ed.), *Theories of everything: In honor of Edward Keenan*, 109–130. Los Angeles: University of California at Los Angeles.
Haspelmath, Martin. 2021. Word class universals and language-particular analysis. In Fulano (ed.), *Oxford handbook of word classes*. Oxford University Press.
Hay, Jennifer B. & R. Harald Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348. doi:10.1016/j.tics.2005.04.002.
Hayden, Rebecca E. 1950. The relative frequency of phonemes in General American English. *WORD*. Taylor & Francis 6(3). 217–223. doi:10.1080/00437956.1950.11659381.
Hendrikse, Andries P. & George Poulos. 1994. Word categories—prototypes and continua in Southern Bantu. *South African Journal of Linguistics*. Routledge 12(20). 215–245. doi:10.1080/10118063.1994.9723955.
Herbst, Thomas. 2019. Constructicons—a new type of reference work? *Lexicographica* 3(1). 3–14. doi:10.1515/lex-2019-0001.
Hilpert, Martin. 2013. *Constructional change in english: Developments in allomorphy, word-formation and syntax*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139004206.
Hoffmann, Thomas & Graeme Trousdale. 2013. *The oxford handbook of construction grammar*. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199609205.001.0001.
Hopper, Paul J. & Sandra A. Thompson. 1985. The iconicity of the universal categories "noun" and "verb". In John Haiman (ed.), *Iconicity in syntax* (Typological Studies in Language 6), 151–183. Amsterdam: John Benjamins. doi:10.1075/tsl.6.08hop.
Hopper, Paul J. & Elizabeth C. Traugott. 2003. *Grammaticalization*. 2nd edition. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139165525.
Hopper, Paul & Sandra A. Thompson. 1984. The discourse basis for lexical categories in universal grammar. *Language* 60(4). 703–752. doi:10.2307/413797.
Houston, Ann. 1991. A grammatical continuum for (ING). In Peter Trudgill & J. K. Chambers (eds.), *Dialects of English*, 241–257. London: Routledge. doi:10.4324/9781315505459-18.
Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Iordăchioaia, Gianina & Martina Werner. 2019. Categorial shift via aspect and gender change in deverbal nouns. *Language Sciences* 73. 62–76.

doi:10.1016/j.langsci.2018.08.011.
Jackendoff, Ray. 2002. *Foundations of language. Brain, meaning, grammar, evolution*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198270126.001.0001.
Janda, Laura A. 2006. Cognitive linguistics. *Glossos* 8(3). 1–60.
Justeson, John & Slava M Katz. 1991. Co-occurrences of antonymous adjectives and their contexts. *Computational linguistics* 17(1). 1–20. https://dl.acm.org/doi/abs/10.5555/971738.971739 (20 April, 2023).
Kanter, Ido & David A. Kessler. 1995. Markov processes: Linguistics and zipf's law. *Physical Review Letters* 74(19). 4559. doi:10.1103/PhysRevLett.74.4559.
Krieger, Hans Ulrich & John Nerbonne. 1993. Feature-based inheritance networks for computational lexicons. In Ted Briscoe, Ann Copestake & Vera de Paiva (eds.), *Inheritance, defaults and the lexicon*, 90–136. Cambridge: Cambridge University Press. doi:10.22028/D291-24827.
Kromer, Victor. 2003. A usage measure based on psychophysical relations. *Journal of Quantitative Linguistics*. Taylor & Francis 10(2). 177–186. doi:10.1076/jqul.10.2.177.16718.
Kučera, Henry. & W. Nelson Francis. 1967. *Computational analysis of present-day American English*. *Brown University Press*. Providence, R.I. https://clu.uni.no/icame/manuals/ (20 November, 2022).
Küchenhoff, Helmut & Hans-Jörg Schmid. 2015. Reply to "more (old and new) misunderstandings of collostructional analysis: On schmid & küchenhoff" by stefan th. gries. *Cognitive Linguistics* 26(3). 537–547. doi:10.1515/cog-2015-0053.
Kwon, Narae. 2017. Empirically observed iconicity levels of English phonaesthemes. *Public Journal of Semiotics* 7(2). 73–93. doi:10.37693/pjos.2016.7.16470.
Labov, William. 1973. The boundaries of words and their meanings. In Joshua A. Fishman (ed.), *New ways of analyzing variation in english*, 340–373. Washington, D.C.: Georgetown University Press.
Lakoff, George. 1987a. *Women, fire, and dangerous things*. Chicago: University Of Chicago Press.
Lakoff, George. 1987b. Cognitive models and prototype theory. In Ulric Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, 63–100. New York: Cambridge University Press.
Langacker, Ronald W. 1987a. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*. Stanford, CA: Stanford university press.
Langacker, Ronald W. 1987b. *Foundations of cognitive grammar: Volume II: Descriptive application*. Stanford, CA: Stanford university press.
Langacker, Ronald W. 1987c. Nouns and verbs. *Language* 63(1). 53–94. doi:10.2307/415384.
Langacker, Ronald W. 1990. *Concept, image, and symbol: The cognitive basis of grammar* (Cognitive Linguistics Research 1). Berlin; New York: De Gruyter.
Langacker, Ronald W. 1998. Conceptualization, symbolization, and grammar. In Michael Thomasello (ed.), *The new psychology of language: Cognitive and functional approaches*, 1–39. Hillsdale, NJ: Erlbaum. doi:10.4324/9781315085678-1.
Langacker, Ronald W. 1999. *Grammar and conceptualization*. Berlin: De Gruyter.
Lehmann, Christian. 2018. Adjective and attribution. Category and operation. In Carolin Baumann, Viktória Dabóczi & Sarah Hartlmaier (eds.), *Adjektive*, 13–76.

doi:10.1515/9783110584042-002.
Lestrade, Sander. 2017. Unzipping zipf's law. *PLOS ONE*. Public Library of Science 12(8). 1–13. doi:10.1371/journal.pone.0181987.
Lievers, Francesca Strik, Marianna Bolognesi & Bodo Winter. 2021. The linguistic dimensions of concrete and abstract concepts: Lexical category, morphological structure, countability, and etymology. *Cognitive Linguistics*. De Gruyter 32(4). 641–670. doi:10.1515/cog-2021-0007.
Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to stefan th. Gries' "dispersions and adjusted frequencies in corpora" international journal of corpus linguistics 13: 4 (2008), 403-437. International Journal of Corpus Linguistics 17. 147–149. doi:10.1075/IJCL.17.1.08LIJ.
Linzen, Tal & Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics* 7(1). 195–212. doi:10.1146/annurev-linguistics-032020-051035.
Longtin, Catherine-Marie, Juan Segui & Pierre A Hallé. 2003. Morphological priming without morphological relationship. *Language and Cognitive Processes*. Routledge 18(3). 313–334. doi:10.1080/01690960244000036.
Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3). 319–344. doi:10.1075/ijcl.22.3.03lov.
Lüdeling, Anke, Stefan Evert & Udo Heid. 2000. On measuring morphological productivity. *Proceedings of the KONVENS 2000*, 57–61. Ilmenau: UCREL.
Mackenzie, J. Lachlan. 2019. Is there a pluralia tantum subcategory of nominal gerunds? Developing Gaeta's notion of morphological differentiation. *Language Sciences* 73. 179–189. doi:j.langsci.2018.08.015.
Mackenzie, Lachlan. 1985. Nominalization and valency reduction. In A. M. Bolkestein, C. de Groot & J. L. Mackenzie (eds.), *Predicates and terms in functional grammar*, 28–47. Dordrecht: Foris.
Maekelberghe, Charlotte, Lauren Fonteyn & Liesbet Heyvaert. 2021. Categoriality in the English gerund system: Lessons learned from Cognitive Linguistics. In Guro Kristiansen, Katerina Franco, Stefania de Pascale, Liesbet Rosseel & Wen Zhang (eds.), *Applications of cognitive linguistics; cognitive sociolinguistics revisited*. De Gruyter. doi:10.1515/9783110733945-033.
Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition* 46(1). 53–85. doi:10.1016/0010-0277(93)90022-n.
Marneffe, Marie-Catherine de, Christopher Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. doi:10.1162/coli_a_00402.
Marslen-Wilson, William, Lorraine K. Tyler, Rachelle Waksler & Lianne Older. 1994. Morphology and meaning in the English mental lexicon. *Psychological review* 101(1). 3–33. doi:10.1037/0033-295X.101.1.3.
McCawley, James D. 1992. Justifying part-of-speech assignments in mandarin chinese. *Journal of Chinese Linguistics*. Cambridge University Press 20(2). 211–246.
McClelland, James L., Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy

T. Rogers, Mark S. Seidenberg & Linda B. Smith. 2010. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14(8). 348–356. doi:10.1016/j.tics.2010.06.002 (5 March, 2023).

Mengden, Ferdinand von. 2010. *Cardinal numerals: Old english from a cross-linguistic perspective*. Berlin, New York: De Gruyter. doi:10.1515/9783110220353.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Mines, M. Ardussi, Barbara F. Hanson & June E. Shoup. 1978. The frequency of occurrence of phonemes in conversational English. *Language and Speech* 21(3). 221–241. doi:10.1177/002383097802100302.

Molineaux, Benjamin J. 2012. The construction of aspect. *English Language and Linguistics*. Cambridge University Press 16(3). 427–458. doi:10.1017/S1360674312000184.

Mompean, Jose A., Amandine Fregier & Javier Valenzuela. 2020. Iconicity and systematicity in phonaesthemes: A cross-linguistic study. *Cognitive Linguistics* 31(3). 515–548. doi:10.1515/cog-2018-0079.

Morgan, James L., Katherine M. Bonamo & Lisa L. Travis. 1995. Negative evidence on negative evidence. *Developmental Psychology* 31(2). 180–197. doi:10.1037/0012-1649.31.2.180.

Mori, Masahiro, Karl F. MacDorman & Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19(2). 98–100. doi:10.1109/MRA.2012.2192811.

Nenadic, O. & M. Greenacre. 2007. Correspondence analysis in r, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3). 1–13. http://www.jstatsoft.org (3 April, 2023).

Nuckolls, Janis B. 1999. The case for sound symbolism. *Annual Review of Anthropology* 28. 225–252. http://www.jstor.org/stable/223394 (9 March, 2023).

O'Hara, Robert B. & Johann Kotze. 2010. Do not log-transform count data. *Nat Prec*. doi:10.1038/npre.2010.4136.1.

Pearl, Lisa. 2022. Poverty of the stimulus without tears. *Language Learning and Development* 18(4). 415–454. doi:10.1080/15475441.2021.1981908.

Perfors, Amy, Joshua B. Tenenbaum & Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of child language* 37(3). 607–642. doi:10.1017/S0305000910000012.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5). 1112–1130. doi:10.3758/s13423-014-0585-6.

Pinker, Steven. 1988. Learnability theory and the acquisition of a first language. In Frank S. Kessel (ed.), *The development of language and language researchers: Essays in honor of roger brown*, 97–119. Hillsdale, NJ: Erlbaum. doi:10.4324/9781315801919.

Plag, Ingo, Christiane Dalton-Puffer & R. Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics* 3(2). 209–228. doi:10.1017/S1360674399000222.

Plag, Ingo, Julia Homann & Gero Kunter. 2017. Homophony and morphology:

The acoustics of word-final S in English. *Journal of Linguistics* 53(1). 181–216. doi:10.1017/S0022226715000183.
Portugal, R. D. & B. F. Svaiter. 2011. Weber-fechner law and the optimality of the logarithmic scale. *Minds & Machines* 21(1). 73–81. doi:10.1007/s11023-010-9221-z.
Pullum, Geoffrey K. 1991. English nominal gerund phrases as noun phrases with verb-phrase heads. *Linguistics*. De Gruyter 29(3). 763–799. doi:10.1515/ling.1991.29.5.763.
Pullum, Geoffrey K. & Arnold M. Zwicky. 1988. The syntax-phonology interface. In Fred J. Newmeyer (ed.), *Linguistics: The cambridge survey, vol 1*, 255–280. Cambridge: Cambridge University Press.
Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-demos.14. https://aclanthology.org/2020.acl-main.0 (14 March, 2023).
Quirk, Randolph. 1965. Descriptive statement and serial relationship. *Language* 41(2). 205–217. doi:https://doi.org/10.2307/411874.
Quirk, Randolph. 1989. *A comprehensive grammar of the English language*. 1. publ., 7. corr. impr. London [u.a.]: Longman.
R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (19 January, 2023).
Rauhut, Alexander. 2021. Exploring the effect of conversion on the distribution of inflectional suffixes: A multivariate corpus study. *Zeitschrift für Anglistik und Amerikanistik* 69(3). 267–290. doi:10.1515/zaa-2021-2024.
Rauhut, Alexander. 2022b. The status of nominal sub-categories: Exploring frequency densities of plural -s. *Yearbook of the German Cognitive Linguistics Association* 10(1). 59–76. doi:10.1515/gcla-2022-0004.
Rauhut, Alexander. 2022a. occurR: Dispersion and association measures for linguistic corpus data. https://github.com/alex-raw/occurR (19 March, 2023).
Rauhut, Alexander. 2023. linguio: R tools and interfaces for handling data formats commonly used in linguistics. https://github.com/alex-raw/linguio (14 March, 2023).
Rice, Sally A. 2003. Growth of a lexical network: Nine English prepositions in acquisition. In Hilde Cuyckens, Renate Dirven & John Taylor (eds.), *Cognitive approaches to lexical semantics*, 243–260. Berlin/New York: Mouton de Gruyter. doi:10.1515/9783110219074.243.
Rijkhoff, Jan. 2000. When can a language have adjectives? An implication an universal. In Petra M. Vogel & Bernard Comrie (eds.), *Approaches to the typology of word classes*, 217–257. Berlin/New York: De Gruyter. doi:10.1515/9783110806120.217.
Rohde, Douglas L. T. & David C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* 72(1). 67–109. doi:10.1016/S0010-0277(99)00031-1 (5 March, 2023).

Rosch, Eleanor H. 1973a. Natural categories. *Cognitive psychology* 4(3). 328–350. doi:10.1016/0010-0285(73)90017-0.
Rosch, Eleanor H. 1973b. On the internal structure of perceptual and semantic categories. In Timothy E. Moore (ed.), *Cognitive development and acquisition of language*, 111–144. San Diego: Academic Press. doi:10.1016/B978-0-12-505850-6.50010-4.
Rosch, Eleanor H. 1978. Principles of categorization. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale, NJ: Erlbaum. doi:10.1016/B978-1-4832-1446-7.50028-5.
Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive. In Günther Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 379–411. De Gruyter. doi:10.1515/9783110900019.379.
Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18(2). 215–262. doi:10.1017/S1360674314000021.
Rosenbach, Anette. 2019. On the (non-)equivalence of constructions with determiner and noun modifiers in English. *English Language and Linguistics* 23(4). 759–796. doi:10.1017/S1360674319000273.
Ross, John R. 1972. The category squish: Endstation Hauptwort. In Paul M. Peranteau & Gloria C. Phares Judith N. Levi (eds.), *Papers from the eighth regional meeting of the chicago linguistic society*, vol. 8, 316–328. Chicago Linguistic Society.
Ross, John R. 1973b. A fake NP squish. In Charles-James N. Bailey & Roger W. Shuy (eds.), *New ways of analyzing variation in English*, 96–140. Georgetown University Press.
Ross, John R. 1973a. Nouniness. In Osamu Fujimura (ed.), *Three dimensions of linguistic theory*, 137–257. Tokyo: TEC Corporation.
Savickỳ, Petr & Jaroslava Hlavácová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics* 9(3). 215–231. doi:10.1076/jqul.9.3.215.14124.
Schachter, Paul & Timothy Shopen. 2007. Parts-of-speech systems. In Timothy Shopen (ed.), *Language typology and syntactic description*, 1–60. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511619427.001.
Schlechtweg, Marcel & Greville G. Corbett. 2021. The duration of word-final s in English: A comparison of regular-plural and pluralia-tantum nouns. *Morphology* 31(4). 383–407. doi:10.1007/s11525-021-09381-x.
Schmid, Hans-Jörg. 2010. Does frequency in the text instantiate entrenchment in the cognitive system? In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 101–133. Berlin/Boston: De Gruyter. doi:10.1515/9783110226423.101.
Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577. doi:10.1515/cog-2013-0018.
Schmitz, Dominic, Dinah Baer-Henney & Ingo Plag. 2021. The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords.

Phonetica 78(5-6). 571–616. doi:10.1515/phon-2021-2013.
Schneider, Florian, Björn Barz, Philipp Brandes, Sophie Marshall & Joachim Denzler. 2021. Data-driven detection of general chiasmi using lexical and semantic features. *Proceedings of the 5th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 96–100. Punta Cana, Dominican Republic (online): Association for Computational Linguistics. doi:10.18653/v1/2021.latechclfl-1.11.
Scott, David W. 1992. *Multivariate density estimation. Theory, practice and visualization*. New York: John Wiley & Sons.
Seyfarth, Scott, Marc Garellek, Gwendolyn Gillingham, Farrell Ackerman & Robert Malouf. 2018. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33(1). 32–49. doi:10.1080/23273798.2017.1359634.
Sheather, S. J. & M. C. Jones. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society Series B* 53. 683–690. doi:10.1111/j.2517-6161.1991.tb01857.x.
Silverman, B. W. 1986. *Density estimation*. London: Chapman; Hall.
Sönning, Lukas & Valentin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59(5). 1179–1206. doi:10.1515/ling-2019-0045.
Stefanowitsch, Anatol. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2(1). 61–77. doi:10.1515/CLLT.2006.003.
Stefanowitsch, Anatol. 2008. Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics* 19(3). 513–531. doi:10.1515/COGL.2008.020.
Stefanowitsch, Anatol. 2011b. Constructional preemption by contextual mismatch: A corpus-linguistic investigation. *Cognitive Linguistics* 20(1). 107–129. doi:10.1515/COGL.2011.005.
Stefanowitsch, Anatol. 2011a. Cognitive linguistics as a cognitive science. In Michael Callies, Wolfgang R. Keller & Andreas Lohöfer (eds.), *Bi-directionality in the cognitive sciences: Avenues, challenges, and limitations* (Human Cognitive Processing), 295–310. Amsterdam & Philadelphia: John Benjamins. doi:10.1075/hcp.30.18ste.
Stefanowitsch, Anatol. 2012. Collostructional analysis. In G. Trousdale & T. Hoffmann (eds.), *The oxford handbook of construction grammar*. Oxford/New York: Oxford University Press. doi:10.1093/oxfordhb/9780195396683.013.0016.
Stefanowitsch, Anatol & Susanne Flach. 2017. The corpus-based perspective on entrenchment. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, 101–127. Boston, MA: De Gruyter. doi:10.1037/15969-006.
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics* 8(2). 209–243. doi:10.1075/ijcl.8.2.03ste.
Stefanowitsch, Anatol, Elena Smirnova & Matthias Hüning. 2020. Complex adpositions in three west germanic languages: German, Dutch, and English. In Bernard Fagard, João Pinto de Lima, Danica Stosic & Elena Smirnova (eds.), *Complex adpositions in european languages: A micro-typological approach to complex nominal relators* (Empirical Approaches to Language Typology), vol. 65, 65–138. Berlin: De Gruyter.

doi:10.1515/9783110686647-003.
Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken english: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin/New York: De Gruyter. doi:10.1515/9783110197808.
The BNC Consortium. 2007. The british national corpus, version 3 (BNC XML edition). Oxford: Bodleian Libraries, University of Oxford. http://www.natcorp.ox.ac.uk/ (19 January, 2023).
Thompson, Sandra A. 1989. A discourse approach to the cross-linguistic category "adjective." In Roberta Corrigan, Fred Eckman & Michael Noonan (eds.), *Linguistic categorization: Proceedings of an international symposium in milwaukee, wisconsin, april 10–11* (Current Issues in Linguistic Theory 61), 245–265. Amsterdam: John Benjamins. doi:10.1075/CILT.61.16THO.
Tomaschek, Fabian, Ingo Plag, Mirjam Ernestus & R. Harald Baayen. 2021. Phonetic effects of morphology and context: Modeling the duration of word-final s in English with naïve discriminative learning. *Journal of Linguistics*. Cambridge University Press 57(1). 123–161. doi:10.1017/S0022226719000203.
Tomasello, Michael. 2003. *Constructing a language*. Cambridge, MA: Harvard University Press. doi:j.ctv26070v8.
Tryk, H. Edward. 1968. Subjective scaling and word frequency. *American Journal of Psychology* 81. 170. doi:10.2307/1421261.
Uhrig, Peter, Stefan Evert & Thomas Proisl. 2018. Collocation candidate extraction from dependency-annotated corpora: Exploring differences across parsers and dependency annotation schemes. In Pascual Cantos-Gómez & Moisés Almela-Sánchez (eds.), *Lexical collocation analysis: Advances and applications*, 111–140. Cham: Springer International Publishing. doi:10.1007/978-3-319-92582-0_6.
Van Der Auwera, Johan & Volker Gast. 2010. Categories and prototypes. In Jae Jung Song (ed.), *The oxford handbook of linguistic typology*. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199281251.013.0010.
Virpioja, Sami, Peter Smit, Stig-Arne Grönroos & Mikko Kurimo. 2013. *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*. Aalto University. http://urn.fi/URN:ISBN:978-952-60-5501-5 (19 March, 2023).
Wasow, Thomas. 2015. Ambiguity avoidance is overrated. In Susanne Winkler (ed.), *Ambiguity*, 29–48. De Gruyter. doi:10.1515/9783110403589-003.
Welmers, W. E. 1973. *African language structures*. Berkeley: University of California Press.
Werker, Janet F. & Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1). 49–63. doi:10.1016/S0163-6383(84)80022-3.
Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org (19 January, 2023).
Wiedemann, Gregor, Steffen Remus, Avi Chawla & Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*. doi:10.48550/arXiv.1909.10430.
Winter, Bodo & Paul-Christian Bürkner. 2021. Poisson regression for linguists: A tutorial

introduction to modelling count data with brms. *Language and Linguistics Compass* 15(11). e12439. doi:10.1111/lnc3.12439.
Winter, Bodo & Francesca Strik Lievers. 2017. Sensory language across lexical categories. *Lingua* 204. 45–61. doi:10.1016/j.lingua.2017.11.002.
Wulff, Stefanie. 2008. *Rethinking idiomaticity. A usage-based approach*. London; New York: Continuum. doi:10.4000/lexis.1749.
Wulff, Stefanie, Nick C. Ellis, Ute Römer, Kathleen Bardovi-Harlig & Chelsea J. LeBlanc. 2009. The acquisition of tense–aspect: Converging evidence from corpora and telicity ratings. *Language Acquisition* 16(3). 193–230. doi:10.1111/j.1540-4781.2009.00895.x.
Xie, Yihui. 2014. Knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch & Roger D. Peng (eds.), *Implementing reproducible computational research*. Boca Raton, Florida: Chapman; Hall/CRC. doi:10.1201/9781315373461-1.
Xie, Yihui. 2015. *Dynamic documents with R and knitr*. Boca Raton, Florida: Chapman; Hall/CRC. doi:10.1201/b15166.
Xie, Yihui. 2016. *Bookdown: Authoring books and technical documents with R markdown*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/bookdown (19 January, 2023).
Xie, Yihui, J. J. Allaire & Garrett Grolemund. 2018. *R markdown: The definitive guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown (19 January, 2023).
Xie, Yihui, Christophe Dervieux & Emily Riederer. 2020. *R markdown cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook (19 January, 2023).
Yung Song, Jae, Katherine Demuth, Karen Evans & Stefanie Shattuck-Hufnagel. 2013. Durational cues to fricative codas in 2-year-olds' american English: Voicing and morphemic factors. *The Journal of the Acoustical Society of America*. Acoustical Society of America 133(5). 2931–2946.
Zimmermann, Julia. 2016a. Morphological status and acoustic realization: Findings from New Zealand English. In Christopher Carignan & Michael D. Tyler (eds.), *Proceedings of the 16th australasian international conference on speech science and technology*, 201–204. Sydney, Australia: University of Western Sydney.
Zimmermann, Julia. 2016b. Morphological status and acoustic realization: Is there a difference between Bra[d] Pitt and a grille[d] cheese omelet, or between Kate Mo[s] and killer robot[s]? In Johannes Wahle, Marisa Köllner, R. Harald Baayen, Gerhard Jäger & Tineke Baayen-Oudshoorn (eds.), *Proceedings of 6th conference on quantitative investigations in theoretical linguistics (QITL-6)*. Tübingen: Universität Tübingen. http://hdl.handle.net/10900/68949 (14 March, 2023).
Zimmermann, Richard. 2020. Word growth dispersion—a single corpus part measure of lexical dispersion. Paper presented at ICAME41 Heidelberg. https://www.youtube.com/watch?v=k8etOvRcF4c (19 January, 2023).
Zipf, George K. 1935. The psychobiology of language: An introduction to dynamic philology. 1935(1). 1–1. doi:10.2307/9855.

A. Appendix A

A.1. Packages

This section demonstrates some of the functionality of the packages I have developed as part of this thesis. Ironically, the corpus of code from studies in corpus linguistics is still lackluster. Therefore, all code that was written to create this thesis will be made available, some packaged. I hope that this can prove useful as reference for fellow researchers.

I developed two software packages developed as a side-product of this thesis. The R package `occurR` is an efficient implementation of the part-based dispersion, distance-based dispersion and association measures. The R package `linguio` provides convenience tools for input and output of common linguistic data format (.vrt, frequency lists, frequency signatures, etc.). It also provides wrappers for the communication with a local CWB installation.

The following notebook-style script demonstrates the functionality and interlocking mechanism of the two packages. More information can be found in the package documentation.

Assume the aim is to investigate adjectives and their inflectional forms in predicative position in a *stanza*-annotated corpus. `linguio` offers wrappers for communication with CWB and data import. Queries can be defined within the same R script or session.

```
library(linguio)
library(occurR)

query <- r'(
  # find all adjectives that that have a dependent tagged as "cop"
  [deprel = "cop"] []* [xpos = "JJ.*"]
  :: target.head = keyword.id within s;

  # export corpus position, id of corpus part, lemma, and PoS tag
  tabulate Last match, text_id, match lemma, match xpos;
)'
```

```
adjectives <- cwb_query(query, corpus = "BNC-STANZA") |>
  linguio::read_text(stringsAsFactors = TRUE)
```

The selected corpus is parametrized for easy reproduction on multiple data sets. However, any part of the query can be parametrized by string interpolation. The `read_text()` function disables all features of *R*'s `read.table` that may cause issues with tab-separated raw text data.

Next, lemmas can be counted and annotated with dispersion measures. The corpus position or the complete corpus are required for distance-based dispersion.

```
names(adjectives) <- c("cpos", "text_id", "lemma", "pos")
dispersions <- dispersion(
  adjectives,
  tokens = lemma,
  parts = text_id,
  cpos = cpos,
  measures = c("f", "dp_norm", "kld_norm", "dwg", "f_ald", "Ur")
)
```

All measures from Gries (2008) and Gries (2010) are available, plus DWG. The result is a `data.frame` with the types and the specified measures. Association measures can also easily be computed. The function `coll_analysis()` has methods for `data.frame` and `table` objects. The package is also optimized for `data.table` and `tidyverse` workflows.

```
adjectives_annotated <- table(adjectives[, c("lemma", "pos")]) |>
  coll_analysis(o11 = JJR + JJS, o12 = JJ, n = corpus_size("BROWN")) |>
  merge(dispersions, by = "lemma")
```

The `corpus_size` function is a helper function in `linguio` that can determine the size of an installed CWB corpus. All the most common input formats to collocation or collostruction analysis can be used alternatively, such as frequency signatures. The resulting `data.frame` (or `data.table` or `tibble`, depending on the input) contains a table with all specified measures for further data manipulation and analysis. Future features will include bootstrapping (cf. Rauhut 2022b), confidence intervals, and more recent measures from (Gries 2022b).

An example implementation of DWG that mirrors the one found in `occurR` can be seen in the following:

```r
#' Word Growth Dispersion
#'
#' A distance-based dispersion measure
#'
#' @param tokens character raw corpus
#' @param corr logical whether or not to apply geometric correction
#'
#' @return numeric
#' @examples
#' n <- 50
#' tokens <- sample(letters, n, replace = TRUE)
#' dwg(tokens)
#'
#' @export
dwg <- function(tokens, corr = TRUE) {
  vocab <- unique(tokens)
  itokens <- match(tokens, vocab)
  f <- tabulate(itokens)

  s <- sort.int(itokens, index.return = TRUE)
  sort_ids <- s$ix
  i <- s$x
  l <- length(tokens)

  d <- c(sort_ids[-1], l) - sort_ids
  last <- cumsum(f)
  first <- c(1, last[-length(last)] + 1)
  d[last] <- sort_ids[first] + l - sort_ids[last]

  mad <- rowsum(abs(d - l / f[i]), i)[, 1] / f
  worst_mad <- (l - f + 1 - l / f) / (f / 2)
  ans <- mad / worst_mad
  if (corr) {
    ans <- ans / (2 * atan(worst_mad) / atan(mad))
  }
  names(ans) <- vocab
  ans
}
```

The full source code is free and open source and available on GitHub:

- occurR
- linguio

A.2. Third-party software used

- Corpus Workbench (Evert & Hardie 2011)
- R Core Team (2021)
  - data.table (Dowle & Srinivasan 2021)
  - ca (Nenadic & Greenacre 2007)
  - ks (Duong 2022)
  - Rmarkdown (Xie, Allaire & Grolemund 2018; Xie, Dervieux & Riederer 2020)
  - knitR (Xie 2015; Xie 2014)
  - bookdown (Xie 2016)
  - ggplot2 (Wickham 2016)

B. Appendix B

B.1. UD dependency relations and other coding schemes

The following is a list of the UD dependency tags taken from the UD website (https://universaldependencies.org/u/dep/index.html):

- acl: clausal modifier of noun (adnominal clause)
- acl:relcl: relative clause modifier
- advcl: adverbial clause modifier
- advmod: adverbial modifier
- advmod:emph: emphasizing word, intensifier
- advmod:lmod: locative adverbial modifier
- amod: adjectival modifier
- appos: appositional modifier
- aux: auxiliary
- aux:pass: passive auxiliary
- case: case marking
- cc: coordinating conjunction
- cc:preconj: preconjunct
- ccomp: clausal complement
- clf: classifier
- compound: compound
- compound:lvc: light verb construction
- compound:prt: phrasal verb particle
- compound:redup: reduplicated compounds
- compound:svc: serial verb compounds
- conj: conjunct
- cop: copula
- csubj: clausal subject
- csubj:outer: outer clause clausal subject
- csubj:pass: clausal passive subject
- dep: unspecified dependency
- det: determiner
- det:numgov: pronominal quantifier governing the case of the noun
- det:nummod: pronominal quantifier agreeing in case with the noun
- det:poss: possessive determiner

- discourse: discourse element
- dislocated: dislocated elements
- expl: expletive
- expl:impers: impersonal expletive
- expl:pass: reflexive pronoun used in reflexive passive
- expl:pv: reflexive clitic with an inherently reflexive verb
- fixed: fixed multiword expression
- flat: flat multiword expression
- flat:foreign: foreign words
- flat:name: names
- goeswith: goes with
- iobj: indirect object
- list: list
- mark: marker
- nmod: nominal modifier
- nmod:poss: possessive nominal modifier
- nmod:tmod: temporal modifier
- nsubj: nominal subject
- nsubj:outer: outer clause nominal subject
- nsubj:pass: passive nominal subject
- nummod: numeric modifier
- nummod:gov: numeric modifier governing the case of the noun
- obj: object
- obl: oblique nominal
- obl:agent: agent modifier
- obl:arg: oblique argument
- obl:lmod: locative modifier
- obl:tmod: temporal modifier
- orphan: orphan
- parataxis: parataxis
- punct: punctuation
- reparandum: overridden disfluency
- root: root
- vocative: vocative
- xcomp: open clausal complement

B.2. Supplementary data

This Section includes more data from the relevant sections. Table B.2 extends table 5.5 in Section 5. Table B.2 extends table 7.3 in Section 7.

Table: Top 100 best dispersed V-*ed* sorted along the modification-predication axis

(Dimension 1 of a Correspondence Analysis)

lemma	Dim1	Dim2	f	dp_norm
impressed	1.8328438	0.0318102	1259	0.7362448
ashamed	1.8280067	0.0227876	602	0.8536825
pleased	1.8250454	0.0387484	4126	0.5915239
convinced	1.8149156	-0.0015130	378	0.8740603
annoyed	1.8116932	-0.1142829	194	0.9335266
delighted	1.7864335	0.0345803	1390	0.7555605
amazed	1.7735194	0.0630532	214	0.9242619
relieved	1.7465076	0.0645609	328	0.8978404
inclined	1.7340687	0.0652552	548	0.8262742
satisfied	1.6952614	0.0191157	2627	0.5775956
surprised	1.6699704	-0.0103831	1038	0.7774991
concerned	1.6400740	-0.0157161	5117	0.4561011
scared	1.6215803	-0.1022355	837	0.8298798
interested	1.5907420	0.0466418	7097	0.4532794
involved	1.5651217	-0.0592566	370	0.8689970
disappointed	1.5347149	-0.0036429	998	0.7815128
dissatisfied	1.5042454	0.0048313	288	0.8975392
exhausted	1.5013326	0.0087203	180	0.9344711
worried	1.4532096	0.0449049	1667	0.7136513
unchanged	1.4478471	0.0578484	551	0.8363327
shocked	1.4155621	0.0358740	931	0.7972371
excited	1.4029712	0.0837356	963	0.7907227
confused	1.3626481	-0.0824932	480	0.8460772
tired	1.3376683	0.0158432	2572	0.6955474
aged	1.3339855	-0.0515768	1575	0.7131437
bored	1.2757116	-0.0456398	697	0.8315565
distressed	1.2708431	0.0166739	258	0.9107607
determined	1.2211944	0.0505160	1364	0.7013531
frustrated	1.2016486	0.0949725	270	0.9021746
terrified	1.1314178	-0.1095711	464	0.8726963
unfounded	1.1074860	-0.1070655	145	0.9291782
flawed	1.0572003	-0.2254037	274	0.8961554
relaxed	1.0456204	0.0483693	1043	0.7475609
depressed	0.9687506	0.0860038	726	0.8097032
unjustified	0.9125163	0.1111106	139	0.9258069
related	0.8799398	0.0563629	5221	0.5744236
minded	0.8665790	-0.3963415	210	0.9222906

(continued)

lemma	Dim1	Dim2	f	dp_norm
married	0.7032334	0.0126168	1585	0.7520576
unrelated	0.7001633	0.0893001	365	0.8703400
complicated	0.5863438	0.0514813	2307	0.5880974
unemployed	0.5370928	-1.0804321	1249	0.7971363
misguided	0.5202768	-0.0409519	180	0.9203955
impaired	0.5201262	-0.1982173	270	0.9283100
muted	0.4984938	0.1342195	207	0.9230248
inexperienced	0.4113123	-0.0433938	267	0.9064721
outdated	0.4098764	0.0393276	177	0.9195407
unresolved	0.3476786	-0.1015592	167	0.9307169
unmarried	0.3180139	-0.0813455	370	0.9074480
isolated	0.2473747	0.0926514	502	0.8153064
unused	0.2168699	0.0728842	254	0.8947712
varied	0.2079454	0.0996697	721	0.8000857
privileged	0.1962718	-0.0146517	660	0.8130226
detached	0.1746170	-0.0138162	201	0.9245579
handicapped	0.1570996	-0.7923347	775	0.9057300
crowded	0.1450394	0.0804269	514	0.8416544
endangered	0.1404538	0.0134860	271	0.9260257
qualified	0.1393643	0.0761174	1237	0.7546058
disabled	0.1391132	-0.5346672	1601	0.8333133
dignified	0.1385870	0.0648955	246	0.9026963
armed	0.1360336	0.0730814	2848	0.7442831
impoverished	0.1021822	-0.0485266	180	0.9219268
civilised	0.0595450	-0.0537055	303	0.8902976
damned	0.0573286	-0.2593380	457	0.8979264
coloured	0.0472996	-0.0556021	193	0.9359214
polished	0.0184887	0.1075801	266	0.9088535
sophisticated	0.0096501	0.0594412	2021	0.6154410
unfinished	-0.0146301	0.0005431	224	0.9171634
talented	-0.0222346	0.0045109	619	0.8485415
skilled	-0.0364608	0.0253039	1196	0.7391579
unqualified	-0.0546751	0.1029913	187	0.9253284
sacred	-0.0620855	0.0318042	689	0.8582609
unexpected	-0.0662145	-0.0028003	1500	0.6492617
wicked	-0.0742551	-0.1458831	592	0.8646983
unified	-0.0854644	0.1365395	502	0.8445699
unskilled	-0.0918305	-0.2729534	233	0.9351314

(continued)

lemma	Dim1	Dim2	f	dp_norm
unlimited	-0.1165600	0.1403629	493	0.8567282
unprecedented	-0.1541361	0.1474556	641	0.7946708
uncontrolled	-0.1650563	0.0836983	161	0.9336154
advanced	-0.1687624	0.0765051	2217	0.6662051
rugged	-0.1756813	0.0669723	241	0.9201479
enlightened	-0.2177151	0.0616571	340	0.8977886
ragged	-0.2412070	0.1755064	310	0.9055499
unauthorised	-0.2789090	0.0869728	213	0.9330025
experienced	-0.2792397	0.1358664	1224	0.7229760
unidentified	-0.2949368	0.0874418	186	0.9226572
detailed	-0.2998947	0.1465342	4868	0.5329008
distinguished	-0.3080703	0.0423850	845	0.7868632
unpublished	-0.3270246	0.0606172	250	0.9260794
belated	-0.3297169	0.0993276	136	0.9326236
unwanted	-0.3425910	0.0830453	513	0.8312571
protracted	-0.3463248	0.1813736	186	0.9251738
unspecified	-0.3549469	0.1818549	166	0.9330840
prolonged	-0.3737259	0.1829030	576	0.8163494
armoured	-0.3776831	0.1097811	273	0.9226979
undoubted	-0.3934972	0.1288976	209	0.8998489
allied	-0.4155172	-0.3036392	596	0.8880518
beloved	-0.4336644	-0.2956180	330	0.8901011
limited	-0.4425142	0.1398125	3753	0.5021670
sustained	-0.4626715	0.1878676	116	0.9355164
uniformed	-0.5103635	0.1310325	237	0.9222105

Table: Top 100 best dispersed V-*ed* sorted along the modification-predication axis (Dimension 1 of a Correspondence Analysis)

lemma	Dim1	Dim2	f	dp_norm
willing	1.5942136	0.0419464	2649	0.5549819
surprising	1.4060261	0.0077907	3532	0.4923871
unwilling	1.3513761	0.0527819	609	0.7963051
encouraging	1.3032584	0.0352958	449	0.8641847
tempting	1.2247806	0.0647569	449	0.8321303
misleading	1.0869198	0.0068762	835	0.7565023
insulting	1.0649105	-0.3422213	154	0.9389295

(continued)

lemma	Dim1	Dim2	f	dp_norm
disappointing	0.9735615	-0.0361254	831	0.7863114
revealing	0.9698295	0.1079116	158	0.9365987
confusing	0.9331732	0.0664576	555	0.8286523
frustrating	0.8995988	0.0162375	422	0.8505605
appealing	0.8637949	-0.0585567	413	0.8528983
reassuring	0.8550018	0.0450711	374	0.8606370
annoying	0.8415369	-0.1002820	281	0.9247026
illuminating	0.7936596	0.1177447	149	0.9424019
disconcerting	0.7729652	-0.0200969	157	0.9438966
puzzling	0.7595466	-0.0462900	195	0.9297552
boring	0.7569824	0.0595457	1105	0.7829273
rewarding	0.7546574	-0.0847499	406	0.8726958
damaging	0.7508408	-0.0832552	724	0.8047006
frightening	0.7204205	0.0147852	688	0.8140138
distressing	0.7094753	-0.0032004	263	0.9061163
embarrassing	0.7069016	0.0489833	766	0.7817310
disgusting	0.6985351	-0.1142824	517	0.8752442
unconvincing	0.6945736	-0.0812730	104	0.9438998
flattering	0.6569649	0.0594345	211	0.9172846
irritating	0.6402828	-0.0370603	301	0.8987166
patronising	0.6020757	-0.0044302	116	0.9479995
satisfying	0.5820658	-0.0119783	522	0.8458962
forthcoming	0.5786421	0.0858111	1057	0.7255175
amusing	0.5786412	0.0073911	517	0.8522843
daunting	0.5745370	0.0695814	374	0.8568521
exhausting	0.5694195	-0.0024746	168	0.9344033
shocking	0.5657147	0.0347035	389	0.8743229
disturbing	0.5651924	-0.0242673	700	0.7958845
entertaining	0.5599222	-0.1280475	379	0.8807281
pleasing	0.5513321	0.0401063	378	0.8747582
depressing	0.5480494	0.0756367	473	0.8459037
interesting	0.5451981	-0.0117249	7969	0.4441303
unsettling	0.5427432	0.1317497	136	0.9433092
convincing	0.5312242	-0.0413720	862	0.7586876
threatening	0.4871396	-0.1266244	228	0.9196344
fitting	0.4734854	0.1118420	508	0.8535179
comforting	0.4473655	0.0514503	278	0.9070899
exhilarating	0.4463507	0.1371299	173	0.9376737

(continued)

lemma	Dim1	Dim2	f	dp_norm
scathing	0.4226855	0.0484850	124	0.9438374
amazing	0.3675554	0.1126879	1419	0.7196095
exciting	0.3390843	0.0305410	2424	0.6346581
alarming	0.3011929	0.0072783	434	0.8395282
humiliating	0.2925165	0.0294728	238	0.8997354
compelling	0.2855869	-0.0276888	392	0.8650185
breathtaking	0.2839754	-0.0462086	251	0.9152142
uncompromising	0.2829437	0.0592513	157	0.9432497
fascinating	0.2808483	0.0780195	1205	0.7256330
charming	0.2427980	0.0753554	953	0.8069943
horrifying	0.2302968	-0.0990880	143	0.9362616
challenging	0.2196324	0.0032682	389	0.8658174
terrifying	0.2177482	-0.0136664	386	0.8782259
devastating	0.1821226	0.1056511	551	0.8031754
welcoming	0.1776547	0.0897928	212	0.9258381
cunning	0.1583828	-0.1019357	189	0.9380373
appalling	0.1306613	0.0572632	784	0.7843618
striking	0.1276750	-0.0731849	975	0.7583707
thrilling	0.1145358	0.0271239	251	0.9151967
soothing	0.0904345	0.0823398	189	0.9387112
promising	0.0819103	-0.0758642	730	0.7783095
bewildering	0.0507506	0.0880832	166	0.9245301
startling	0.0485353	0.0649339	452	0.8436380
stunning	0.0483327	-0.0320057	690	0.8524051
overwhelming	-0.0234788	0.1633538	1003	0.7109097
caring	-0.0405573	0.0096734	265	0.9038706
staggering	-0.0407677	0.1643188	132	0.9333189
missing	-0.0608428	0.0987153	667	0.8114348
outstanding	-0.1078816	0.0173736	2056	0.6549164
outgoing	-0.1325552	0.0546412	428	0.9012925
dazzling	-0.1535708	0.0604809	310	0.8895018
fleeting	-0.1607934	0.1246007	257	0.9164806
imposing	-0.1642245	0.0878646	156	0.9472520
sickening	-0.1745836	0.0161293	149	0.9395822
harrowing	-0.1817030	0.1721852	103	0.9492112
ongoing	-0.1847976	0.1723579	479	0.8413746
loving	-0.2257871	-0.0539145	387	0.8960441
enterprising	-0.2336346	-0.0362743	181	0.9342424

(continued)

lemma	Dim1	Dim2	f	dp_norm
sweeping	-0.3365027	0.1808254	184	0.9113274
painstaking	-0.3404948	0.1810482	123	0.9479475
lasting	-0.3768314	0.1830764	233	0.9142969
contrasting	-0.3776016	0.1831194	279	0.8965702
gruelling	-0.3889245	0.1837514	131	0.9439678
commanding	-0.4159819	0.1852616	172	0.9157566
resounding	-0.4626172	0.0386661	145	0.9252363
unsuspecting	-0.4661459	0.0877917	106	0.9440361
longstanding	-0.4665988	0.1880868	112	0.9484774
ageing	-0.4775119	0.1886959	245	0.9080103
impending	-0.5175735	0.1909320	209	0.9017441
incoming	-0.5239810	0.1912896	314	0.8988721
accompanying	-0.5255578	0.1913776	358	0.8654360
underlying	-0.5294295	0.1712191	1435	0.6886146
corresponding	-0.5314654	0.1917074	741	0.8041068
opposing	-0.5400288	0.1921853	210	0.9076614
ensuing	-0.5400288	0.1921853	140	0.9418890

Aus der Dissertation hervorgegangene Publikationen

Rauhut, Alexander. 2021. Exploring the effect of conversion on the distribution of inflectional suffixes: A multivariate corpus study. Zeitschrift für Anglistik und Amerikanistik 69(3). 267–290. doi:10.1515/zaa-2021-2024.

Rauhut, Alexander. 2022. The status of nominal subcategories: Exploring frequency densities of plural -s. Yearbook of the German Cognitive Linguistics Association 10(1). 59–76. doi:10.1515/gcla-2022-0004

Kurzzusammenfassung in deutscher Sprache

Die vorliegende Arbeit befasst sich mit der Frage, wie sich die Wortklassengradienz und Prototypikalität in der englischen Sprache quantitativ beschreiben lassen. Auf Grundlage von kognitionslinguistischen Modellen wird zunächst ein Überblick über die verschiedenen Aspekte von Wortarten gegeben. Besondere Aufmerksamkeit wird dabei den verschiedenen Arten von Gradienz gewidmet. Der aktuelle Stand der Forschung zum Wortartenkontinuum wird dargestellt und als Grundlage für die deskriptive Analyse der Wortarten des Englischen genommen. Erkenntnisse aus der Sprachtypologie werden genutzt, um relevante funktionale Dimensionen zu identifizieren und mit distributionellen Eigenschaften von Englischen Nomen und Verben in Beziehung zu setzen.

Diese konzeptionelle Grundlage führt zu den Kernhypothesen der Arbeit:

1. Prototypencluster basieren auf sprachspezifischen distributionellen Eigenschaften und derer Kontingenz mit lexikalischen Formen.
2. Semantisch-pragmatische Eigenschaften tendieren zur Formation von kontinuierlichen Gradienten, von denen nur manche Prototypencluster aufweisen.

Drei Korpusfallstudien werden vorgestellt, die die statistische Verteilung von verschiedenen ambivalenten Kategorien untersuchen. Die erste Fallstudie dient zur Exploration der Methode und befasst sich mit Unterkategorien von englischen Adjektiven.

Die zweite Fallstudie ist auf die Untersuchung der Pluralformen von englischen Nomen und die damit assoziierten Unterkategorien fokussiert. Es werden das Cluster-Verhalten von Massenomen, Eigennamen und dem Pluraletantum untersucht und illustriert. Die Ergebnisse zeigen klare Abgrenzungen zwischen Eigennamen, Massenomen und zählbaren Nomen, allerdings kann eine Verdichtung von Lexemen, die auf eine Prototypenkategorie eines Pluraletantum hinweisen, nicht demonstriert werden.

Die dritte Fallstudie befasst sich mit dem umstrittenen Gerund-Partizip. Die Daten zeigen klare Abgrenzungen zwischen verbalen und nominalen Gruppen. Ein enges Netzwerk lässt mehrere Intreprätationsmöglichkeiten zu.

Die Ergebnisse der Fallstudien werden in Bezug zu den Hypothesen gesetzt und diskutiert. Eine der Hauptschlussfolgerungen ist, dass verschiedene Arten von Gradienz sowohl konzeptuell als auch quantitativ unterschieden werden müssen, was bisher in der linguistischen Forschung selten explizit gemacht wird. Die Arbeit schließt mit einer Zusammenfassung der Ergebnisse und einer Diskussion zu den methodischen und theoretischen Implikationen.

Kurzzusammenfassung in englischer Sprache

This dissertation investigates the quantitative aspects of word class categories. Based on cognitive linguistic models, the dissertation provides an overview of the different aspects of word classes. Special attention is given to the different types of gradience. The current state of research on the word class continuum is presented and used as a basis for the descriptive analysis of the word classes of English. Insights from typological research are used to identify relevant functional dimensions and to relate them to distributional properties of English nouns and verbs.

This conceptual foundation leads to the core hypotheses of the work:

1. Prototypical clusters are based on language-specific distributional properties and their contingency with lexical forms.
2. Semantic-pragmatic properties tend to form continuous gradients, only some of which contain prototype clusters.

Three corpus case studies are presented that investigate the statistical distribution of various ambiguous categories. The first case study serves to explore the method and focuses on subcategories of English adjectives. It will used as a benchmark for the subsequent case studies.

The second case study is focused on plural forms of English nouns and associated subcategories. It investigates the cluster behavior of mass nouns, proper nouns and the pluralia tantum. The results show clear boundaries between proper nouns, mass nouns and count nouns, although a concentration of lexemes that point to a prototype category of the pluralia tantum cannot be demonstrated.

The third case study deals with the controversial gerund-participle. The data show clear boundaries between verbal and nominal groups. A tight, and diffuse network allows for several interpretations.

Eigenständigkeitserklärung

Rauhut, Alexander

Name, Vorname

Erklärung zur Dissertation

mit dem Titel: Quantitative aspects of word class categories: The noun-verb continuum in English

1. Hiermit versichere ich,

    - dass ich die von mir vorgelegte Arbeit **selbständig** abgefasst habe, und
    - dass ich **keine weiteren Hilfsmittel** verwendet habe als diejenigen, die im Vorfeld explizit zugelassen und von mir angegeben wurden, und
    - dass ich die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen) entnommen sind, unter Angabe der Quelle kenntlich gemacht wurden, und
    - dass die Arbeit nicht schon einmal in einem früheren Promotionsverfahren angenommen oder abgelehnt wurde.

2. Mir ist bewusst,

    - dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und entsprechend der Promotionsordnung geahndet werden.

Berlin, den 11.04.2023

Ort, Datum Unterschrift