Hilpert, Martin & Saavedra, David Correia & Rains, Jennifer. 2021. A multivariate approach to English clippings. *Glossa: a journal of general linguistics* 6(1): 104, pp. 1–30. DOI: https://doi.org/10.16995/glossa.5771

A multivariate approach to English clippings

Martin Hilpert, Université de Neuchâtel, CH, martin.hilpert@unine.ch
David Correia Saavedra, Université de Neuchâtel, CH
Jennifer Rains, Université de Neuchâtel, CH

This paper addresses the morphological word formation process that is known as clipping. In English, that process yields shortened word forms such as *lab* (< *laboratory*), *exam* (< *examination*), or *gator* (< alligator). It is frequently argued (Davy 2000; Durkin 2009; Haspelmath & Sims 2010; Don 2014) that clipping is highly variable and that it is difficult to predict how a given source word will be shortened. We draw on recent work (Lappe 2007; Jamet 2009; Berg 2011; Alber & Arndt-Lappe 2012; Arndt-Lappe 2018) in order to challenge that view. Our main hypothesis is that English clipping follows predictable tendencies, that these tendencies can be captured by a probabilistic, multifactorial model, and that the features of that model can be explained functionally in terms of cognitive, discourse-pragmatic, and phonological factors. Cognitive factors include the principle of least effort (Zipf 1949), an important discourse-pragmatic factor is the recoverability of the source word (Tournier 1985), and phonological factors include issues of stress and syllable structure (Lappe 2007). While the individual influence of these factors on clipping has been recognized, their interaction and their relative importance remains to be fully understood. The empirical analysis in this paper will use Hierarchical Configural Frequency Analysis (Krauth & Lienert 1973; Gries 2008) on the basis of a large, newly compiled database of more than 2000 English clippings. Our analysis allows us to detect regularities in the way speakers of English create clippings. We argue that there are several English clipping schemas that are optimized for processability.

1 Introduction

English words such as *lab*, *exam*, or *gator* are the product of a truncation process that is called clipping (Plag 2003: 116). In the literature on linguistic morphology, clipping is commonly seen as a minor word formation process that is relatively unpredictable (Durkin 2009: 116) or even outside the confines of morphology (Haspelmath & Sims 2010: 40). These positions are grounded in the observation that clipping is highly variable. Bauer (1994: 40) states that "there is no way to predict how much of a word will be clipped off in clipping, nor even which end of the word will be clipped off". This view is shared by Don (2014: 27), who argues that "clippings can result from deletion of just any part of the word, and there does not seem to be a clear pattern". The following examples illustrate that variability.

(1) a. sis (< sister)
b. gator (< alligator)
c. fridge (< refrigerator)
d. shroom (< mushroom)
e. bro (< brother)
f. bruv (< brother)
g. legit (< legitimate)

Clipping may involve the truncation of the end of a word (*sis*), the beginning (*gator*), or both end and beginning (*fridge*). The phonological material that is clipped off and deleted may be part of a syllable (*shroom*), an entire syllable (*sis*), or several syllables (*gator*, *fridge*). The clipped form may consist of one syllable (*sis*) or more syllables (*gator*). It may end in a consonant (*sis*) or in a vowel (*bro*). The same word may be clipped in different ways (*bro*, *bruv*). Clippings mostly instantiate nouns (*sis*), but also other word classes, including adjectives (*legit*). There is by now a growing literature on clipping that tries to account for this variability (Lappe 2007; Jamet 2009; Berg 2011; Alber & Arndt-Lappe 2012; Arndt-Lappe 2018) and that challenges the view of clipping as an unsystematic, unpredictable phenomenon. In this paper, we extend that work on the basis of new data and the use of multivariate statistics, with the goal of constructing a probabilistic, multifactorial model of the clipping process. That model should generate predictions about the form of a clipped word on the basis of the non-truncated source word, and it should allow us to explain why speakers produce that specific form. As has recently been pointed out by Arndt-Lappe (2018: 146), this aim has up to this point not been pursued with enough effort: "A general problem in assessing the predictability of output structures of truncatory processes is that, in much of the pertinent literature, the degree of predictability is not in the focus of investigation."

Previous research into the clipping process (Tournier 1985; Antoine 2000; López Rúa 2002; Jamet 2009; Balnat 2012, amongst others) has produced sizable collections of clipped forms, as well as several empirical generalizations. The position that clipping is largely unsystematic (Bauer 1994; Durkin 2009) is still common in current textbooks on word formation (Don 2014), but it has not gone unchallenged. In particular, studies that approach the phenomenon on the basis of quantitative evidence report probabilistic tendencies that suggest that clipping is at least to some extent predictable. In an early contribution, Tournier (1985: 298) notes the predominance of end-clipped forms in English words that are listed in the *Concise Oxford Dictionary*. Corroborating this finding, Jamet (2009: 17) analyzes a dataset of 290 clipped forms taken from Wikipedia, finding that more than 90% of them are nominal and that 75% involve end-clipping, rather than front-clipping or clipping of both the beginning and the end of the source word. Tournier (1985: 303) also finds that a majority of English clipped forms end in a consonant. Tournier (1985: 305) further argues that competing functional pressures give rise to the shape of clipped forms. While it would be economical for speakers to truncate as much material as possible from a source word, clipped forms need to retain enough substance to be recognizable. Put simply, clippings need to achieve a balance between the principle of least effort (Zipf 1949) and the principle of maximized recognizability.

Other early research has addressed the meaning of clipped forms. Two conflicting positions can be found. On one side, Adams (1973: 135) states that clipping is "the process by which a word of two or more syllables (usually a noun) is shortened without a change in the function taking place". Contradicting this view, there are accounts that view clippings either as informal or as belonging to specialized discourse. The informal nature of many clippings, which is illustrated by *cig* from *cigarette*, has been observed already in Kreidler (1979). Empirically, Adams' assessment is contradicted by clippings such as *bro* or *merch*, which have undergone lexicalization processes (Brinton & Traugott 2005; Hilpert 2019) that have rendered them less than fully synonymous with their respective source forms *brother* and *merchandise*. The clipped form *bro* serves as an intimate form of address, while the clipped form *merch* tends to refer to a very specific type of merchandise, namely clothing items that are purchased by the fans of pop artists and other celebrities. Plag (2003: 121) points out that clippings can differ from their non-truncated sources by conveying the speaker's familiarity with the denotation of the clipped form. Many clipped forms refer to highly specific referents that are only known to a limited community. As such, clipped forms function sociolinguistically as a marker of in-group identity. Rock musicians know that the clipped form *Strat* refers to a specific guitar model (*Stratocaster*), which can be plugged into an *amp* (*amplifier*) that is connected to a *cab* (*loudspeaker cabinet*). However, the community specificity of clippings is not a general feature. Many clipped forms have gained wider currency, so that a *mic* can be identified as a *microphone* not only by musicians.

A proposal that is relevant for our account of English clippings has been proposed by López Rúa (2002), who addresses clippings and other shortening phenomena in English from the perspective of prototype theory and radial networks (Lakoff 1987). López Rúa argues that clippings, initialisms, abbreviations, and blends can be analyzed as related phenomena. Her account, which is based on a database that includes 195 clippings (López Rúa 2002: 34), identifies prototypical schemas for each individual process, including clipping. The idea that clipping can be understood in terms of several different schemata will be of central concern for the analysis in this paper.

As was mentioned above, Lappe (2007) offers a pioneering analysis of clippings and other shortening phenomena in English that provides strong evidence against the view of clipping as an unpredictable phenomenon. The analysis is based on 2259 truncated names and 1158 suffixed and unsuffixed clippings (2007: 59). It is shown that a number of phonological factors, including stress, syllable structure, and consonant clusters can be used to account for systematic correspondences between shortened forms and their source words (2007: 164). Couching her analysis in an optimality-theoretic framework, Lappe (2007: 168) views the clipping process as the product of two interacting types of constraints, namely faithfulness constraints, which bias speakers against outcomes that differ strongly from their sources, and phonological markedness constraints, which bias speakers against phonologically unusual outcomes. Faithfulness constraints (cf. McCarthy 2011: 13) include the tendency to preserve the stressed syllable of the source word in the outcome (ˈ*doctor* > *doc*). Markedness constraints include the tendency for shortening to result in monosyllabic words, since these represent the unmarked case in English. Lappe (2007: 182–183) proposes different constraint rankings for monosyllabic and disyllabic clippings and shows that these rankings successfully address the variability that can be observed in the data. The present paper builds on these results, but it also differs from Lappe's approach in several respects. First, the present analysis restricts itself to the analysis of clipped common words to which no further material has been added. This means that hypocoristic names such as *Andy* (< *Andrew*) or *Betty* (< *Elizabeth*), diminutive forms such as *ciggy* (< *cigarette*) or *comfy* (< *comfortable*), *o*-suffixed words such as *pervo* (< *pervert*), or blends such as *brunch* (< *breakfast* + *lunch*) are not taken into account. Second, the present analysis draws on a new large database of more than 2000 English clippings. Third, it differs in its methodological approach, specifically its use of Hierarchical Configural Frequency Analysis (HCFA).

Another study that is close in spirit to the present study is offered by Berg (2011), who examines a database of 935 English clippings in order to determine the factors that govern speakers' choices between end-clipping and front-clipping. While end-clipping is generally the more frequent choice, Berg (2011: 9) uses logistic regression modeling to study the impact of different factors. The analysis shows several main effects and interaction effects. Source words with non-initial stress show a relatively greater likelihood of front-clipping (e.g. *gator* < *alligator*). Also, shorter words have relatively greater odds of front-clipping. First names are

front-clipped more often than common nouns, especially when they have non-initial stress (e.g. *Beth* < *Elizabeth*, *Tricia* < *Patricia*). Berg (2011: 12) proposes a discourse-pragmatic explanation for the observed differences between first names and common nouns, suggesting that front-clipping poses a greater risk to the recoverability of a source word (cf. Nooteboom 1981), and that first names, due to their contextual grounding, are more easily recovered than common nouns. This explains speakers' general preference for end-clipping and why first names deviate from this tendency more often than common nouns.

Alber & Arndt-Lappe (2012) propose a conceptual distinction between templatic truncation and subtractive truncation, which is relevant for the discussion in this paper. In their words, "[i]n templatic truncation, the size of the derived form is predictable; in subtractive truncation, predictability holds for the size of the truncated material. Prediction of form in both cases normally involves reference to higher-level phonological categories such as the syllable or metrical units such as the foot" (Alber & Arndt-Lappe 2012: 290). Templatic truncation thus allows the analyst to make generalizations about the shape of clipped forms. In phenomena that instantiate subtractive truncation, generalizations can be stated about the material that is taken away from a source word. As will be described in more detail in section 2, we have annotated a large database of attested English clippings for variables that capture characteristics of both the source words and the clipped forms, so that we are able to analyze different patterns of clipping in terms of either templatic or subtractive truncation.

In other work that assumes a cross-linguistic perspective on clipping, Alber (2010) and Arndt-Lappe (2018) study name truncation and clipping across languages such as Italian, German, and English in order to come to terms with the variability that shortening processes exhibit both within individual languages and across different languages. Instead of analyzing clipping as a unified phenomenon, different systematic truncation patterns are identified, whose formal characteristics are shown to correspond to different functional purposes. With regard to hypocoristics, strikingly systematic correspondences can be observed between non-truncated sources and their truncated targets (Arndt-Lappe 2018: 149). In Italian, truncated names tend to preserve either the first syllable (*Albi* < *Alberto*) or the main stress (*Betta* < *Elisabetta*). Truncated names with an *i*-suffix such as *Albi* overwhelmingly preserve the first syllable, while truncated names with a reduplicated consonant (*Pippo* < *Filippo*) show a strong preference for the preservation of the main stress. On the basis of these findings, Arndt-Lappe (2018: 145) argues, contra Haspelmath & Sims (2010), that it would be premature to exclude clipping from the study of grammatical morphology.

Our approach in this paper further aligns with Gries (2006), who focuses on English blends (*brunch* < *breakfast, lunch*) and complex clippings (*sysadmin* < *system*, *administrator*) with the aim of uncovering both the underlying systematic patterns that govern these processes and the cognitive factors that motivate them in the first place. Amongst other things, Gries (2006: 555)

finds that the recognition points of source words play a more important role for blends than for complex clippings. Blends also tend to recruit source words that are relatively more similar to one another in terms of their edit distance. The quantitative evidence leads Gries to conclude that the two processes are much more systematic than has been previously acknowledged.

To sum up the preceding paragraphs, a substantial body of research weakens the traditional view of clipping as unsystematic and unpredictable. A number of regularities have been detected, and the literature has identified several cognitive, discourse-pragmatic, and phonological factors that are implicated in clipping. Existing work has also laid important groundwork by drawing conceptual distinctions such as the contrast between templatic and subtractive truncation. The present paper will go beyond the current state by drawing on a larger database, by applying a method that up to now has not been used for the analysis of clipping, and by adopting Construction Grammar as a theoretical framework that is particularly suited for the study of clipped forms.

The paper is structured as follows. Section 2 discusses our data and methods. With regard to the data, a database of more than 2000 English clippings has been compiled from various sources. Each entry has been annotated for a range of variables that capture different structural criteria of the clipping and its respective source word. We then discuss the basic principles of Hierarchical Configural Frequency Analysis (Krauth & Lienert 1973; Gries 2008), which we adopt as the analytical method in order to model the clipping process. Section 3 presents the results of our analysis, which identifies fourteen structural patterns that occur more frequently than expected in the database. We further discuss another outcome of the analysis, namely two patterns that are significantly underrepresented. In section 4, we discuss our results in the light of earlier research on clipping, and specifically the role of functional pressures that shape the clipping process. Section 5 concludes the paper and outlines how we plan to follow up our results in future work.

2 Data and methodology

2.1 The clipping database

The analysis in this paper draws on a database of 2272 English clippings and their respective source words that has been annotated for several variables that pertain to morphological and phonological distinctions.[1] Section 2.1.1 identifies the sources from which the clipped words were compiled. Section 2.1.2 discusses the criteria for inclusion and exclusion that were used for the construction of the database. While the definition of clipped words may appear relatively straightforward, reductive word formation processes do in fact exhibit variation across several dimensions (López Rúa 2002), which makes it necessary to spell out the characteristics that were applied in our decision process. Section 2.1.3 presents the variables that are included in the database and discusses the process of annotation.

[1] The database is available upon request from the authors.

2.1.1 Sources

The clipped words in the database were compiled from various published sources, websites, a web-based survey, and from personal observations. A substantial number of clippings are listed in the online edition of the Oxford English Dictionary (OED). We retrieved 307 words from the OED that were characterized as shortened forms in their respective entries, which also identify the corresponding full forms. Another rich source of clippings is *An English-French Dictionary of Clipped Words* (Antoine 2000), which accounts for 1135 words in our database. Among the academic studies of clipping as a word formation process that we consulted, Marchand (1969), López Rúa (2002), and Jamet (2009) proved particularly useful. On Wiktionary (https://en.wiktionary.org), several lists of shortened words are available, which were cross-checked with the entries from other published sources. We included 526 words from those lists in the database. Further entries were gathered from the Urban Dictionary (www.urbandictionary.com), which is a crowdsourced online database for neologisms. The platform Qualtrics was used to conduct an online survey with 200 participants, who were asked to provide ten clipped words from their own language use. That procedure yielded 65 entries for the database. To the clippings that were collected from these sources, we added forms that we came across through personal observation. It follows from the above comments that we cast a very wide net, including clippings from different varieties of English, different registers, different historical periods during which the clippings were coined, and different modes of communication.

2.1.2 Criteria of inclusion and exclusion

This section discusses what elements were included as clipped words in our database. The decisions that were taken can be represented by six principles that are discussed below.

Principle #1: Only linguistic elements that have undergone truncation are included.

The first principle distinguishes clippings such as *prof* and *exam* from their non-truncated counterparts *professor* and *examination*. Plag (2003: 116) defines clippings in terms of truncation, which "is a process in which the relationship between a derived word and its base is expressed by the lack of phonetic material in the derived word." The principle also motivates the exclusion of cases that only involve mere sound changes or re-ordering of phonological substance (as in *pisketty* < *spaghetti*).

Principle #2: Acronyms, initialisms and orthographic abbreviations are not included.

Many linguistic elements show a relative lack of phonological substance that we did not want to include in the database, as for example acronyms (*SETI* < *search for extra-terrestrial intelligence*), initialisms (*IV* < *intravenous*), or abbreviations that are used only orthographically, but never in spoken language use (*Apr* < *April*). The second principle systematically excludes these forms.

Principle #3: Forms with simultaneous truncation and augmentation are not included.

According to the third principle, the present analysis does not include forms that simultaneously exhibit truncation and augmentation, that is, the addition of phonological substance. This is the case for example in blends (*staycation* < *stay*, *vacation*), which combine material from two independent words into a newly formed word in which either one or two of the input words are truncated (Gries 2006). It is also the case in hypocoristics (*Bobby* < *Robert*) or terms of endearment (*homie* < *homeboy*), which are the result of both truncation and augmentation with a diminutive suffix. While this principle most often draws a clear distinction, there are borderline cases that merit a brief discussion. An element such as *lesbie* (< *lesbian*) might be considered as an instance of suffixation. However, *lesbie* was still included in the database, because the phoneme /i/ is present in the source word. By contrast, the form *lezzie* (< *lesbian*) is the result of simultaneous truncation (*les* < *lesbian*) and augmentation (*lezzie* < *les*), so that *lezzie* was not in the database. Included in the database were further truncated plural forms with a retained suffix, as for example *abs* (< *abdominals*). Here, we argue against a scenario in which the rare singular form *abdominal* was first truncated to *ab* and then pluralized. Further included were truncated forms of multi-word compounds, such as *sitcom* (< *situational comedy*). Plag (2003: 122) points out that these forms are conceptually close to blends, since they can be segmented into different input words that are truncated. The inclusion of *sitcom* in our database is motivated by the fact that *situational comedy* exists as a compound. By contrast, the blend *staycation* does not have a corresponding non-truncated compound *stay vacation*.

Principle #4: Personal proper names are not included.

The fourth principle states that the database does not include personal proper names, which have been studied in detail by Lappe (2007) and Berg (2011), who have found that proper names and common nouns differ in the way they give rise to clippings. By contrast, names of geographical places (*Nash* < *Nashville*), companies (*Benz* < *Mercedes Benz*), and products (*Strat* < *Stratocaster*) have been included in the database.

Principle #5: Forms that exhibit simultaneous truncation and sound change are included if the sound change instantiates de-neutralization.

The fifth principle addresses clippings that exhibit simultaneous truncation and sound change. A prominent example is the clipping *bro* (< *brother*), which shows a change from the STRUT vowel to the GOAT vowel. Lappe (2007: 287) notes cases such as *ac* (< *accumulator*) or *comie* (< *comedian*), which undergo a sound change that is labeled schwa-deneutralization. Cases with sound changes were included in the database if a plausible link could be motivated between the corresponding sounds in the source word and in the clipping. For example, many clippings exhibit a tensed variant of the vowel that is found in the source word, as in *rubby* (< *rubbing*

alcohol) or *gynie* (< *gynecologist*). By contrast, no such motivation is possible in the word *aggro* (< *aggressive*).

Principle #6: Truncated phrases are not included in the database.

Finally, while truncated forms of established compounds were included in the database, truncated phrases were not included, even when those phrases were highly frequent. This is relevant for forms such as *sonova* (< *son of a bitch*), *catamount* (< *cat of the mountain*), or *abfab* (< *absolutely fabulous*). By contrast, truncated forms of multi-word units that are fully lexicalized are included in the database. For example, the form *all in* is a clipped variant of the lexicalized form *all inclusive*. While it is by no means trivial to distinguish between compounds and phrases in English, criteria such as the compound stress rule (Chomsky & Halle 1968: 17) were taken into account in order to guide the decision process. Phrasal multi-word units were included on the condition that they exhibit non-compositional meaning, which reflects advanced lexicalization (Hilpert 2019: 15). To take an example, the meaning of *private parts* is more specific than 'parts that are private', which thus warrants inclusion.

2.1.3 Variables in the clipping database

Each clipping in the database is represented alongside its source word. For example, the clipping *merch* corresponds to the source word *merchandise*. In addition to a unique identifier and information on the source from which the clipping was recovered, the full database contains a range of variables on language structure and use that is not described in full in this paper. For the present purposes, eight variables are crucial. The first of these, which captures the length of the clippings in the database, measured in terms of the number of syllables, is shown in **Figure 1**. Four variables that pertain to the nature of the source word, and to the preservation of stress are represented in **Figure 2**. Three further variables, which relate to segmental and morphological criteria of the clippings in the database, are visualized in **Figure 3**.

Variable #1: Number of syllables of the clipped word

This variable captures the number of syllables of each clipped word. The clipping *merch* (< *merchandise*) has one syllable, *chrono* (< *chronograph*) has two. The first panel of **Figure 1** shows that monosyllabic clippings are the most frequent category. With increasing length, fewer and fewer clippings are attested. For the present analysis, clippings with four and more syllables were subsumed under one single category.

Variable #2: Source word compound

The first panel of **Figure 2** distinguishes clipped words for which the source word is a non-compound word from clippings that derive from compounds or lexicalized multi-word units.

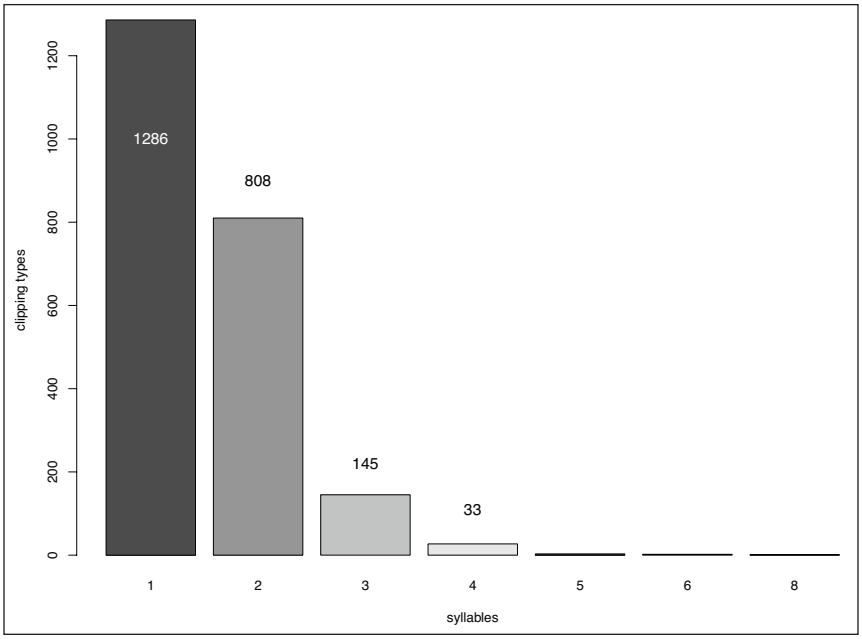

Figure 1: Word length in a database of 2272 English clippings.

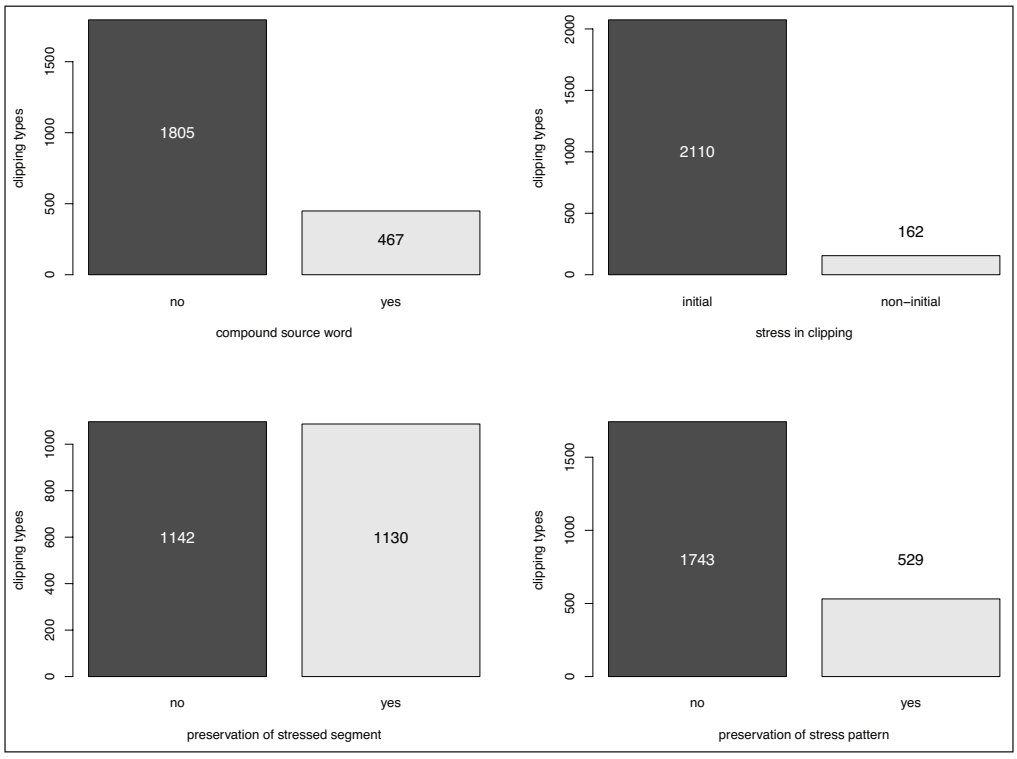

Figure 2: Compound source words, stress pattern, and stress preservation in a database of 2272 English clippings.

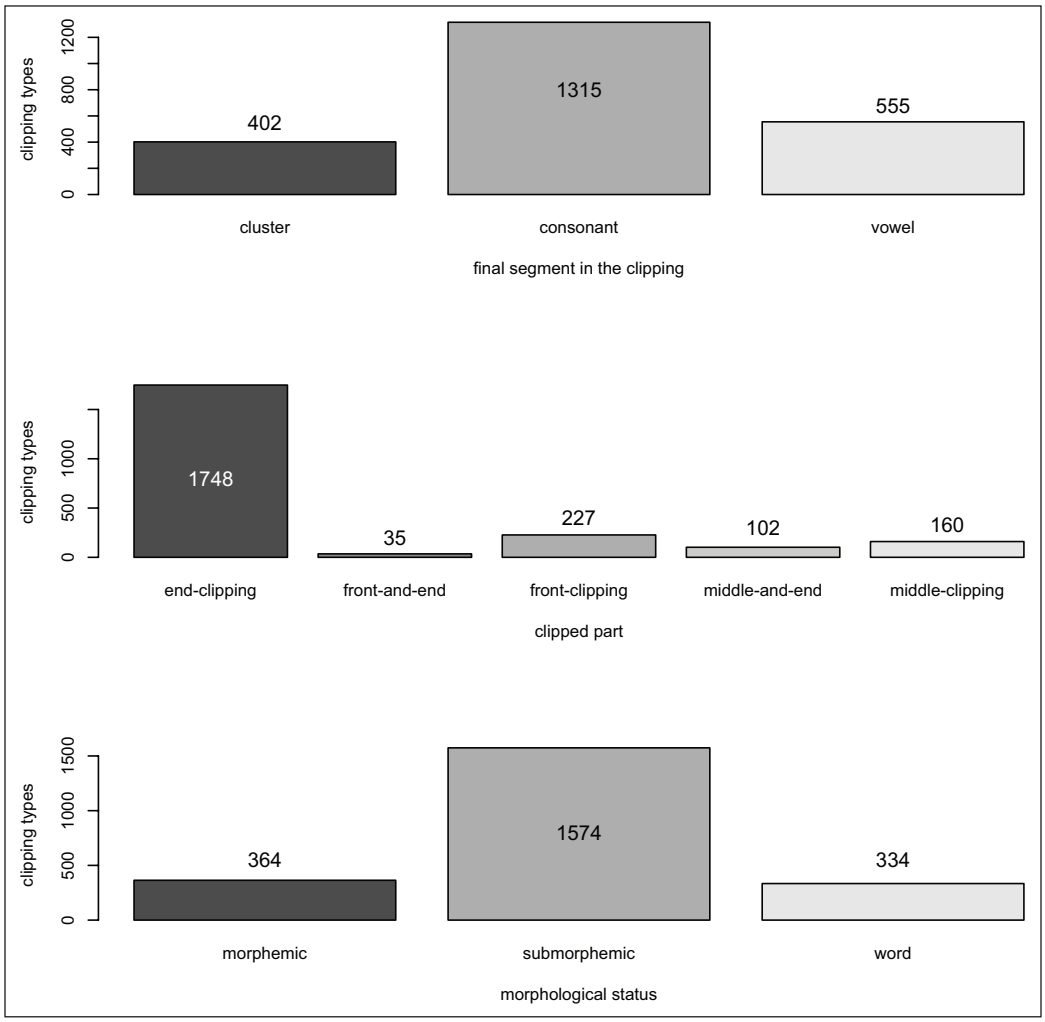

Figure 3: Segmental and morphological characteristics in a database of 2272 English clippings.

Clippings that have a compound as their source word include *pub* (< *public house*), *abs* (< *abdominal muscles*), *club* (< *nightclub*), or *coin-op* (< *coin-operated*). We followed Sánchez-Gutiérrez et al. (2018) in regarding classical morphemes such as *pre-*, *ultra-*, or *tele-* as affixes, so that we regard neoclassical compounds (*telephone*) as complex but non-compound words. The same principle was adopted for words with cranberry morphemes (*cauliflower*).

Variable #3: Stress pattern of the clipped word

This variable captures the stress pattern of the clipped word. We make a binary distinction between initial stress and non-initial stress. The second panel of **Figure 2** shows that the overwhelming majority of clippings are stressed on the initial syllable. Of course, a large proportion of the clippings in the database are monosyllabic, which partly accounts for this result.

Variable #4: Preservation of the stressed segment

As can be seen in the third panel of **Figure 2**, in about 50% of all entries in the database, the stressed segment of the clipped word preserves the stressed segment of the corresponding source word. We make a binary distinction between segmental stress preservation (‘*dino* < ‘*dinosaur*) and the absence of segmental stress preservation (‘*admin* < *admini’stration*).

Variable #5: Preservation of the stress pattern

The fourth panel of **Figure 2** shows that in 23% of the entries in our database, the clipping and its source word correspond to the same stress pattern. For example, both *admin* and *administration* show penultimate stress, whereas there is a difference between *dino* and *dinosaur*. While dino is stressed on the penultimate, dinosaur is stressed on the antepenultimate. The preservation of a stressed segment can thus be in conflict with the preservation of the stress pattern of a word. For our analysis, all clippings and their corresponding source words have been annotated in terms of antepenultimate, penultimate, and final stress. Cases that could not be assigned to any of those three types were labeled as “other”. If the clipping and its source word conform to the same stress pattern, we noted the presence of stress pattern preservation (ˈ*avo* < *avo*ˈ*cado*), for pairs with different patterns we noted the absence of stress pattern preservation (*de*ˈ*lish* < *de*ˈ*licious*).

Variable #6: Final phonological segment of the clipped word

This variable captures the phonological category of the last segment of the clipped word. Three types are distinguished: final vowels (*ammo* < *ammunition*), final consonants (*prof* < *professor*), and final consonant clusters (*fridge* < *refrigerator*). Large proportions of consonant-final clippings have also been observed by Kreidler (1979: 31) and Lappe (2007: 140).

Variable #7: The word-internal position of the truncated part

This variable distinguishes between end-clipping (*prof* < *professor*), front-and-end-clipping (*fridge* < *refrigerator*), front-clipping (*phone* < *telephone*), middle-and-end-clipping (*pram* < *perambulator*), and middle-clipping (*obstets* < *obtetrics*). It distinguishes thus between five different categories, out of which end-clipping is by far the most frequent one.

Variable #8: Morphological status of the clipped word

This final variable captures whether or not the form of a clipped word corresponds to an existing word (*robe* < *wardrobe*) or a morpheme (*homo* < *homosexual*), or if the clipping process yields a form in which the integrity of at least one of its morphemes is no longer maintained. For the latter case, we adopt the label “submorphemic”, since the resulting forms often represent a part of a morpheme (*merch* < *merchandise*). When a clipping corresponds to a morpheme that is present in

the source word, it is labelled as morphemic (e.g. *pyro* < *pyromaniac*, *psycho* < *psychopath*). This also applies to morphological roots (e.g. *teach* < *teacher*, *preach* < *preacher*). If this morpheme also happens to be an independent word, then it is labelled as such (e.g. *after* < *afternoon*, *ape* < *apeshit*). This also applies to clippings that result from a multi-word unit (e.g. *parts* < *private parts*, *sailor* < *sailor hat*). In some cases, while the clipping itself became a combining form in Modern English (e.g. *fem*, *tech*), what matters for the labelling is its status vis-à-vis its source word at the time of coinage (e.g. *fem* < *female/feminine*: submorphemic, *tech* < *technology*: submorphemic). Morphemes and words of foreign origin are also labelled as such, so that the entry *kinder* from *kindergarten* is categorized as a word. Cranberry morphemes are counted as morphemic (e.g. *cauli* < *cauliflower*, *cray* < *crayfish*). Words with obscure etymologies are labelled as submorphemic (e.g. *scouse* < *lobscouse*).

2.2 Hierarchical Configural Frequency Analysis

Hierarchical Configural Frequency Analysis, or HCFA for short (Krauth & Lienert 1973; Gries 2008), is usually applied with the aim of detecting so-called types in a population of subjects. In medical research, finding groups of patients with similar configurations of symptoms can help to define syndromes; in psychology, finding groups of people that display similar character traits can guide research on personality types. In the context of clipping, the goal is to find systematic patterns in the way source words are shortened. The method examines the features of clippings and their corresponding source words and determines whether certain configurations of features occur more often than would be expected by chance. To illustrate, if the annotated features include the difference between monosyllabic clippings (*sis*) and disyllabic clippings (*celeb*), as well as the difference between source words with initial stress (*sister*) and non-initial stress (*celebrity*), the HCFA can determine whether or not disyllabic clippings derive systematically from source words with non-initial stress. The HCFA performs multiple exact binomial tests to compare the observed frequencies of cases such as *celeb* against the frequencies that are expected by chance. If these cases are attested significantly more often than expected, that particular configuration is considered a type. Importantly, the HCFA may reveal that more than two variables interact in significant ways. For example, it could be that source words with non-initial stress produce disproportionately many disyllabic clippings if the resulting clipped form ends in a consonant (*celeb* < *celebrity*, *exec* < *executive*, *legit* < *legitimate*).

In the present study, HCFA is applied in order to arrive inductively at a typology of different clipping patterns. In order to illustrate the kinds of insights that can be provided by this method, it is useful to start by considering pairwise cross-tabulations of the variables that enter the analysis. **Figure 4** uses such a cross-tabulation to visualize how the retention of the stressed segment of the source word differs across clippings with different lengths. Each bar represents

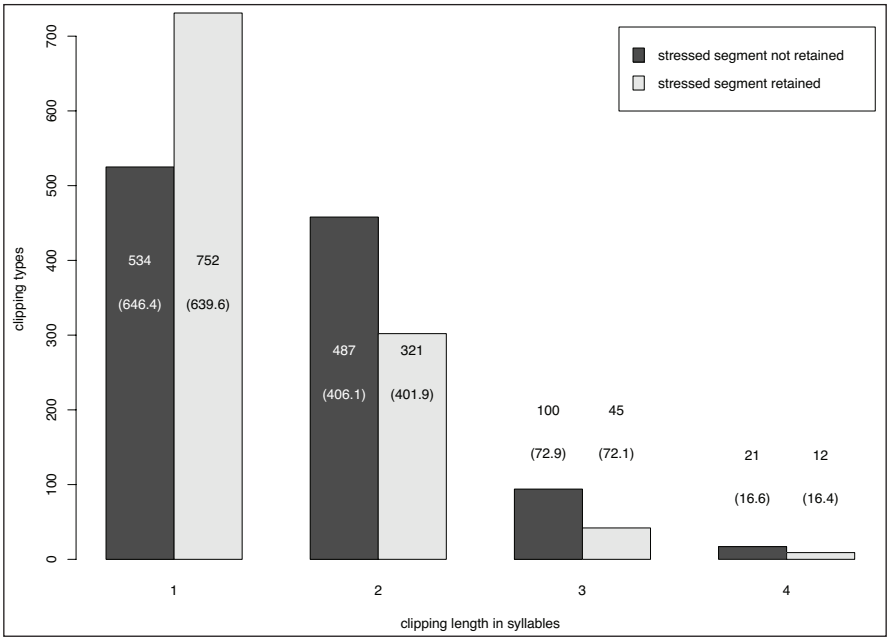

Figure 4: Retention of the stressed segment across clippings of different lengths.

observed frequencies of clippings in the database, expected frequencies are shown in brackets. What can be seen is an asymmetry between monosyllabic clippings on the one hand, which tend to retain the stressed segment of the source word, and polysyllabic clippings, which show the opposite tendency.

This asymmetry reflects the tendency of clippings to facilitate recoverability that has been pointed out by Tournier (1985: 305) and Lappe (2007: 303). Short clippings have a greater need to retain the stress of the source word because they have relatively little phonological substance that ensures recoverability. By contrast, long clippings are easy to recognize even with changed stress, since the segmental phonological cues are stronger.

Another pairwise cross-tabulation of variables is shown in **Figure 5**, which allows us to examine if front-clippings, back-clippings, and other truncation patterns differ with regard to their final segments. Again, the graph shows a number of asymmetries between the observed and expected frequencies. With regard to the most frequent type, end-clipping, it can be seen that end-clippings with final clusters (such as *tux*) are relatively under-represented. End-clippings with final consonants (*prof*, *sis*) are observed about as often as expected. End-clippings with final vowels (*dino* < *dinosaur*, *limo* < *limousine*) are overrepresented. Two more tendencies merit a brief discussion. The first concerns front-clipping, which is the only pattern for which final consonants are overrepresented. Words such as *quake* (< *earthquake*) or *shroom* (< *mushroom*) illustrate this. Why should front-clipping and end-clipping differ with regard to their relative preference

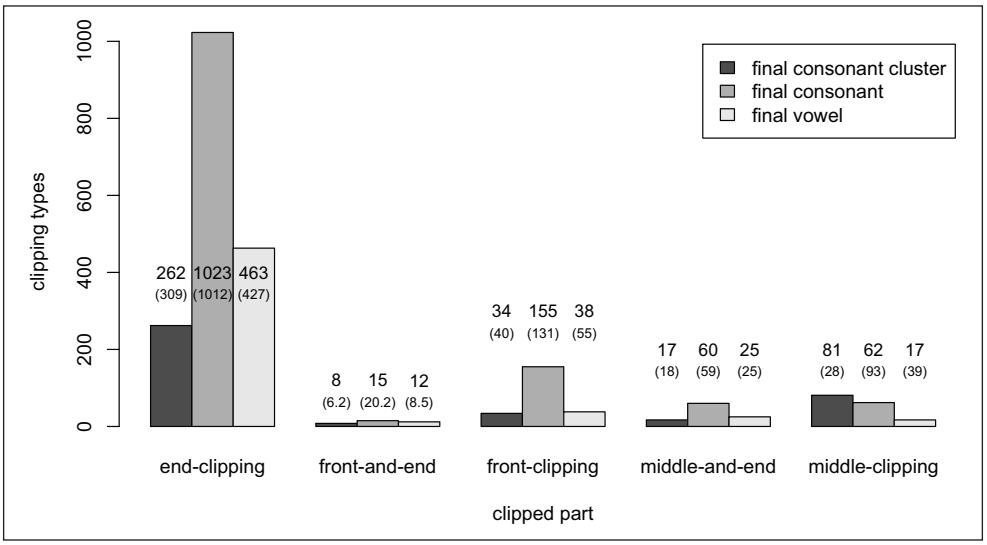

Figure 5: Final phonological segment(s) across clippings with different truncation positions.

for final consonants? The difference can be motivated functionally. All other things being equal, end-clippings are much less difficult to process, since the retained beginning of the source word serves as a strong perceptual cue (Berg 2011: 10). This means that front-clippings need to offer other cues that compensate for the missing beginning. The tendency for front-clippings to end in a consonant serves that purpose. It represents an economic strategy to enhance recoverability, since it adds a cue without requiring the speaker to pronounce an additional syllable.

The most pronounced asymmetry between observed and expected frequencies in **Figure 4** concerns the overrepresentation of final consonant clusters in middle-clippings. The database contains 81 clippings such as *circs* (< *circumstances*), *stats* (< *statistics*), *nugs* (< *nuggets*), or *deets* (< *details*). These four examples, and the overwhelming majority of clippings in the database, are end-clipped words that carry a plural *-s*. In accordance with principle #3 of the database construction (section 2.1.2), these forms were not excluded from the analysis. We argue that the attractivity of this particular clipping pattern can be explained in terms of its quasi-ideal recoverability. The beginning of the source word acts as a strong first cue, a closed syllable at the end offers further information to the hearer, and the final *-s* conveys the meaning of a plural, still within the scope of the final syllable. In other words, this pattern is extremely economic and very reliable at the same time.

An important question to ask at this point is whether the tendencies that can be seen in **Figures 4** and **5** are genuinely due to the variables that are cross-tabulated, or whether they are caused by intervening factors. It is the purpose of a multivariate analysis to simultaneously control for the effects of several interacting variables. Before the next section presents the results

of a HCFA, **Table 1** illustrates the general logic of that approach. The table lists all eight variables that were discussed in section 2.1.3, and it presents in bold the levels that characterize the specific clipping *croc* (< *crocodile*). The clipping is monosyllabic, its source word is not a compound, it preserves the stress of the source word, it ends in a single consonant, the final part of the source word is truncated, and the remaining part is submorphemic. The question that the HCFA allows us to answer is how often that particular configuration of structural characteristics is observed in the database, and whether the observed frequency differs from the expected frequency.[2] The HCFA will further not only examine configurations in which the levels of all variables are fully specified, but it will also calculate observed and expected frequencies for configurations in which individual variables are left unspecified. Results of this kind will be discussed in more detail in the following section. One finding that will play a role is that clippings such as *croc* (< *crocodile*) are indeed strongly overrepresented.

3 Results

The calculations for the present analysis were performed with HCFA 3.2 (Gries 2004), a script written for the statistical software package R (R Development Core Team 2020).[3] The analysis establishes that there are several significant interactions between the variables that were described in section 2.1.3. Section 3.1 discusses fourteen types that appear in the database more often than would be expected. Section 3.2 inverts the perspective and discusses two configurations that are

	Variable	Levels
1	Number of syllables:	**1**, 2, 3, 4+
2	Source word is a compound:	yes, **no**
3	Stress pattern:	**initial**, non-initial
4	Preservation of the stressed segment:	**yes**, no
5	Preservation of the stress pattern	yes, **no**
6	Final segment(s):	consonant cluster, **consonant**, vowel
7	Clipped part:	**end**, front-and-end, front, middle-and-end, middle
8	Morphological status:	**submorphemic**, morphemic, word

Table 1: Variables and their levels for the clipping *croc* (< *crocodile*).

[2] Expected frequencies are calculated on the basis of the marginal frequencies in a cross-tabulation of categorical data (cf. Levshina 2013: 211).

[3] Data and R code with which the analysis can be replicated are available upon request from the authors.

conspicuously absent from the database. In the terminology of the HCFA, these configurations are called anti-types.

3.1 Fourteen clipping types

Tables 2 and **3** summarize fourteen types, which represent a non-exhaustive selection of the types that are identified as significant by the HCFA. Our selection of types is based on descending type frequency and the implication of all eight variables. Each type is given a label, such as *croc* for the first type, which represents an instance of the general pattern. Below, we discuss the fourteen most frequent significant configurations that involve specific levels of all eight variables that entered the analysis. **Table 2** shows the quantitative results. Besides the observed and expected frequencies, the table presents for each type its contribution to the global significance of the HCFA, a p-value, and, as a measure of effect size that is independent of sample size, the Q-coefficient of pronouncedness (cf. Gries 2008: 252). As can be seen in **Table 2**, the values of the Q-coefficient are largely parallel to the observed frequency values.

Table 3 presents the types in terms of the variables and their respective levels. The information in the columns summarizes the structural profile of each type. The types are ordered

	Type	Observed	Expected	Cont. Chisq	P-value	Q-coefficient
1	*croc*	355	106,7	577,79	2,35E-82	0,122
2	*merch*	90	30,9	113,29	3,18E-14	0,028
3	*hydro*	57	2,1	1471,77	1,62E-56	0,026
4	*uni*	50	6,8	274,37	1,89E-22	0,020
5	*memo*	47	8,6	171,79	7,88E-16	0,018
6	*para*	45	6,8	216,43	2,87E-18	0,018
7	*delish*	31	4,9	137,44	4,57E-11	0,012
8	*circs*	22	2,8	132,43	5,76E-09	0,009
9	*stash*	21	4,3	65,96	8,29E-05	0,008
10	*romcom*	11	0,9	114,64	4,94E-05	0,005
11	*prior*	11	1,1	89,86	0,00038555	0,005
12	*physical*	10	0,6	134,66	3,22E-05	0,004
13	*milspec*	10	0,3	352,29	6,73E-09	0,005
14	*nilla*	10	1,1	69,83	0,0052	0,004

Table 2: Fourteen clipping types – quantitative results.

	Type	Syll	Source compound	Clipped part	Final segment	Status	Stress	Preserved stressed segment	Preserved stress pattern
1	*croc*	1	no	end	consonant	submorph	initial	yes	no
2	*merch*	1	no	end	cluster	submorph	initial	yes	no
3	*hydro*	2	no	end	vowel	morphemic	initial	no	yes
4	*uni*	2	no	end	vowel	morphemic	initial	no	no
5	*memo*	2	no	end	vowel	submorph	initial	no	yes
6	*para*	2	no	end	vowel	morphemic	initial	yes	no
7	*delish*	2	no	end	consonant	submorph	non-initial	yes	no
8	*circs*	1	no	middle	cluster	submorph	initial	yes	no
9	*stash*	1	no	front	consonant	submorph	initial	yes	yes
14	*nilla*	2	no	front	vowel	submorph	initial	yes	yes
10	*romcom*	2	yes	middle + end	consonant	submorph	initial	no	no
13	*milspec*	2	yes	middle + end	consonant	submorph	initial	no	yes
11	*prior*	2	yes	end	consonant	word	initial	no	yes
12	*physical*	3	yes	end	consonant	word	initial	no	no

Table 3: Fourteen clipping types – variables and values.

in the same way as in **Table 2**, with the exception of *nilla* (< *vanilla*), which appears next to *stash* (< *moustache*), and *milspec* (< *military specifications*), which appears next to *romcom* (< *romantic comedy*). This positioning reflects their mutual similarity: The table presents sets of types with similar characteristics in adjacent lines, such as for example type *croc* and type *merch*. Cells with grey shading indicate the features that differ. With regard to *croc* and *merch*, the only difference concerns the final segment. Similarly, *romcom* and *milspec* only differ with regard to the preservation of the stress pattern. Generally, what can be seen in the table is that there is considerable structural variety. The types that are identified have properties that draw from the full spectrum of the variables that were included in the analysis. The table further shows certain tendencies, such as the relative dominance of end-clipping, and the association of disyllabic clippings with final vowels. Only one type exhibits non-initial stress. The two variables that capture the preservation of the stressed segment or the stress pattern are in a near-complementary distribution. We will further comment on these observations in the descriptions of the individual types below.

3.1.1 Type *croc*

The structural characteristics of this clipping pattern have been discussed in the previous section, where they have been summarized in **Table 1**. The type is instantiated by 355 clippings in the database, whereas we would expect to find just over 100. Besides *croc*, clippings that instantiate this pattern include *sis* (< *sister*), *fam* (< *family*), *pup* (< *puppy*), and *dem* (< *democrat*). The HCFA identifies this type as the most strongly pronounced configuration, which corroborates existing claims from the literature on clippings (Kreidler 1979: 31; Lappe 2007: 140). The dominance of this structure is not a coincidence. Clippings that consist of a closed syllable and that maintain the initial stressed segment of the source word can be said to maximize both economy and recoverability.

3.1.2 Type *merch*

This type differs from the type *croc* only in the final segment, which is a consonant cluster instead of a single consonant. The type *merch* (< *merchandise*) subsumes all clippings that are monosyllabic, derive from a non-compound, preserve the initial stressed element from the source word but not its overall stress pattern, end in a consonant cluster, instantiate end-clipping, and are categorized as submorphemic. The database contains 90 clippings with these features, whereas only 31 would be expected. Since this type is structurally very close to the type *croc*, it benefits from the same advantages with regard to processing. Both types are observed about three times as often as expected.

3.1.3 Type *hydro*

Another frequent configuration in the database is instantiated by clippings such as *hydro* (< *hydrothermal*), *para* (< *paraplegic*), and *poly* (< *polyester*). These clippings are disyllabic, derive from a non-compound, have initial stress, end in a vowel, and illustrate end-clipping. They do not preserve the stressed element from the source word, but instead they preserve the overall stress pattern: *poly* and *polyester* converge on a penultimate stress pattern.

The HCFA identifies three further types that are highly similar to the type *hydro* and that directly follow it in **Table 2**, where they are labelled *uni* (< *university*), *memo* (< *memorandum*), and *para* (< *paragraph*). The types *hydro* and *uni* differ only with regard to the preservation of the stress pattern. Whereas both *hydro* and *hydrothermal* converge on a penultimate pattern, *uni* and *university* show a difference. While it would be possible to subsume the two types under a generalization that abstracts away from stress pattern preservation, the HCFA suggests that the type *hydro* is especially noteworthy. The clipping type that is instantiated by forms such as *memo* (from *memorandum*) is identical to the type *hydro* except that *memo* represents units with a submorphemic status. Finally, the type *para* (from *paragraph*) is highly similar to the type *uni*, except that *para* preserves the stressed element from the source word, which is not the case for *uni*. The four types thus show family resemblances that could be captured by an overarching generalization.

3.1.4 Type *delish*

An altogether different disyllabic clipping type is instantiated by clippings such as *delish* (< *delicious*), which bear non-initial stress, preserve the stressed segment but not the stress pattern of their source words, end in a consonant, represent end-clipping, and have submorphemic status. Further instances include *exec* (< *executive*), *celeb* (< *celebrity*), or *legit* (< *legitimate*). The database contains 31 clippings of this type, which significantly exceeds the expected frequency of 4.9. The type *delish* breaks away from the bias that disyllabic clippings exhibit towards initial stress and final vowels. Lappe (2007: 13) argues that speakers adopt clippings of this type despite their marked prosodic characteristics because forms such as *celeb* or *legit* are attractive in terms of faithfulness. A speaker would have the option to shorten *legitimate* to the stress-initial clipping *legi*, which would however make recoverability much more difficult.

3.1.5 Type *circs*

In section 2.2, **Figure 4** showed that middle-clippings with final consonant clusters, illustrated by forms such as *specs* (< *spectacles*), *nugs* (< *nuggets*), or *deets* (< *details*), are overrepresented in the database. The HCFA detects a more specific configuration of features that occurs more often than expected. The clippings of this type are monosyllabic, submorphemic, they derive from

a non-compound, and they necessarily have initial stress. The clippings preserve the stressed element from the source word, but not the overall stress pattern. The database contains 22 clippings that conform to this pattern, whereas only 2.8 would be expected. The HCFA thus adds further weight to our argument that clippings of this type should be a preferred option for speakers and hearers, since they are highly recognizable and very economical to produce.

3.1.6 Type *stash*

Table 2 shows only two types that instantiate front-clipping. One such type is represented by monosyllabic clippings such as *stash* (< *moustache*), *nam* (< *Vietnam*), or *vette* (< *Corvette*). In the database, 21 clippings of that type are attested, only 4.3 would be expected. These clippings are submorphemic, they derive from a non-compound word, have initial stress, and end in a consonant. Clippings of this type preserve the stressed element from the source word as well as the stress pattern, which is unusual. In section 2.2, we showed that final consonants are overrepresented in front-clipping. This type conforms to that generalization, and it further makes up for the processing disadvantage of front-clipping by also maintaining the stressed segment and the stress pattern of the source word.

It is useful to compare the type *stash* to the other front-clipping type that is listed in **Tables 2** and **3**, which is instantiated by forms such as *nilla* (< *vanilla*), *tato* (< *potato*), or *bacco* (< *tobacco*). The two types differ in length and in their final segment, but what is remarkable is that both converge on the preservation of both the stressed element from the source word and its stress pattern. The HCFA thus indicates that front-clippings can only gain traction if they are sufficiently faithful to their source words.

3.1.7 Type *romcom*

The next type exhibits a structural profile that has been noted by Jamet (2009: 17). It is disyllabic, derives from a compound, preserves neither the stressed segment of the source word nor its stress pattern, ends in a simple consonant, and instantiates the relatively rare pattern of middle-and-end-clipping. The forms that represent this type have at least one part that is submorphemic. In the database, the type is instantiated by 11 clippings, including *romcom* (< *romantic comedy*), *sitrep* (< *situation report*), or *mintech* (< *ministry of technology*). While 11 instances are not much in absolute terms, the HCFA identifies this pattern as a type because the configuration would only be expected to surface less than once in the database. Consonant-final disyllabic clippings are rare, middle-and-end clippings are rare, and so the combination of these features should be rarer still. With regard to recoverability, clippings like *romcom* strike a compromise. On the one hand, the source compound is shortened considerably and both the stressed element and the stress pattern are given up in the process. At the same time, the respective initial material of the two

component words is retained in the form of two closed syllables, which facilitates identification of the source word. It is worth pointing out that clippings such as *admin* (< *administration*) or *decaf* (< *decaffeinated*), which have the same prosodic characteristics, but which differ in terms of their respective source words, which are not compounds, are actually underrepresented in the database.

The type *romcom* is largely identical to the type *milspec*, which is listed as type 13 in **Tables 2** and **3**, and which is attested 10 times in the database. The only difference between the two is the preservation of the stress pattern. Both *milspec* and *military specifications* can be characterized as having penultimate stress, pairs such as *romcom* and *romantic comedy* do not converge on a single pattern. Given the substantial structural overlap between the two types and the similarity of the two in terms of their statistical significance, it is warranted to subsume the two under a single pattern that disregards the preservation of the stress pattern.

3.1.8 Type *prior*

A further type that involves compound source words is instantiated by disyllabic clippings such as *prior* (< *prior conviction*), *navel* (< *navel orange*), or *laptop* (< *laptop computer*). These end-clippings bear initial stress, preserve the stress pattern but not the stressed element of the source word, and crucially have the shape of existing English words. The database contains 11 examples, which exceeds the expected frequency of 1.1.

The type *prior* can be compared to the type *physical* (< *physical examination*), which differs only in terms of length and the preservation of the stress pattern of the source word, which is maintained in *prior* but not in *physical*. The statistical results are very similar for the two types. The type *physical* is attested 10 times, the expected frequency is under 1. In both types, the modifier of a compound source word is maintained in the clipping. It is worth pointing out that the pattern of *prior* and *physical* does not extend to monosyllabic compound modifiers. The database does contain such elements, as for example *step* (< *step sister*) or *soap* (< *soap opera*), but in contrast to *prior* and *physical*, those forms are not significantly overrepresented.

3.2 Two clipping anti-types

The HCFA does not only allow the identification of configurations that are overrepresented in a body of data, it can also reveal which configurations should be observed more often than they actually are. In other words, it can tell us which kinds of clippings would be expected to appear in the database but are conspicuously absent. Two such anti-types are presented in **Tables 4** and **5** and in the sections that follow. Our selection is motivated by statistical significance and the Q-coefficient. Amongst the anti-types that the HCFA judges to be significant, the two stand out as the ones with the highest Q-coefficients.

	Anti-type	Observed	Expected	Cont. chisq	P-value	Q-coefficient
1	*simul*	29	65.80	20.58	0.032	0.018
2	*dee*	9	46.47	30.21	2.61E-07	0.018

Table 4: Two clipping anti-types – quantitative results.

	Type	Syll	Source com-pound	Clipped part	Final segment	Status	Stress	Preserved stressed segment	Preserved stress pattern
1	*simul*	2	no	end	consonant	submorph	initial	no	no
2	*dee*	1	no	end	vowel	submorph	initial	no	no

Table 5: Two clipping anti-types – variables and values.

3.2.1 Anti-type *simul*

If monosyllabic clippings tend to end in a consonant, the reverse is true for disyllabic clippings, unless their source word is a compound, as has been discussed in connection with the type *romcom*. The HCFA yields the result that disyllabic, stress-initial end-clippings that derive from a non-compound are significantly underrepresented. Clippings such as *simul* (< *simultaneous*), *vocab* (< *vocabulary*), or *biog* (< *biography*) neither preserve the stressed segment nor the overall stress pattern of their source word, which negatively affects recoverability. The anti-type *simul* can be compared to the type *delish*, as the two differ only in stress position and in the preservation of the stressed segment. Due to the latter, stress-final clippings such as *exec* (< *executive*), *celeb* (< celebrity), or *delish* (< *delicious*) have better overall recoverability, which motivates the fact that they are overrepresented in the database.

3.2.2 Anti-type *dee*

This anti-type, which is illustrated by clippings such as *dee* (< *detective*), is the natural counterpart to the types *croc* and *merch* that were discussed in section 3.1. Whereas monosyllabic end-clippings that end in a consonant or a consonant cluster are overrepresented, the opposite is the case for monosyllabic end-clippings with a final vowel. The HCFA indicates that this tendency is particularly pronounced for clippings that preserve neither the stressed segment nor the stress pattern of the source word, as is the case for clippings such as *pneu* (< *pneumonia*) or *mo* (< *momentum*). The database lists 9 clippings that conform to the pattern, but more than 46 would be expected by chance. The broad conclusion that can be drawn is that speakers of English tend to avoid monosyllabic end-clippings that end in a vowel. This is not always possible. The clipping *noy* (< *noise*) would not be any shorter than its source word if the final consonant were

maintained. If the addition of consonants still yields a shortening of the source word, that option will often be taken. The database contains doublets such as *ma/mam* (< *mama*), *cue/cuke* (< cucumber), and *loo/lieut* (< lieutenant) that illustrate this tendency.

Taken together, the types and anti-types that have been identified by the HCFA yield a set of findings that can serve as the basis for a discussion of earlier generalizations on clipping that have been put forward in the literature. The following section will present four relevant points.

4 Discussion

The results that were obtained allow us to reaffirm the claim that the clipping process is not completely unpredictable (Lappe 2007; Jamet 2009; Berg 2011). To put this assessment into perspective, the 14 types that we presented in section 3.1 account for 770 of the 2272 clippings in the database (33.9%). There certainly is a lot of variability and even unpredictability, but a quantitative analysis of authentic data brings out several regular tendencies that reflect cognitive, discourse-pragmatic, and prosodic factors (cf. Arndt-Lappe 2018). The types that were identified by the HCFA are sensitive to segmental, suprasegmental and morphological structure, and they reflect patterns that balance recoverability and economy. This point has been articulated in several ways in the literature, and our results serve as another piece of empirical evidence that corroborates it. The following paragraphs outline three points that are raised by our results and that go beyond the issue of non-predictability.

First, our analysis suggests that it is useful to understand clipping in English as a process that involves several smaller-scale generalizations. Put simply, there are several ways to clip words in English. Each of these ways can be thought of as a schema that draws on the same general strategy of truncation, but does so in different ways. The types that we have discussed under labels such as *merch*, *romcom*, or *hydro* have the status of separate generalizations that speakers can use to generate new, similar clippings. This state of affairs can be usefully captured by adopting the perspective of Construction Grammar (Goldberg 2019), especially as it pertains to phenomena in morphology (Booij 2013; 2019). If clippings exhibit a large variety of forms, that is because there are several different constructional schemas that yield different outcomes, not because the clipping process as such would be unconstrained. In other words, clipping may appear to be unpredictable if it is viewed at the most general level, but at more specific levels, there is order. The importance of lower-level generalizations has been pointed out by Langacker (1999: 106), who argues that grammatical patterns that are specific have several processing advantages over patterns that are more abstract. The more specific a pattern is, the more features an individual instance will share with the schema that licenses it. This is especially important in the case of clippings, which incur a risk of miscommunication unless the clipping reliably allows the hearer to recover the source word. It is thus motivated that clippings should be organized around lower-level generalizations, which is a conclusion that resonates with Booij's (2019: 393) assessment

that " grammar and lexicon are not separated, and that the architecture of the grammar should be conceived as a multidimensional network of relations between syntactic and morphological constructions".

Second, based on our results we argue that clipping patterns are usefully understood as generalizations that map a source word to a shortened structure. This point takes us back to Alber & Arndt-Lappe (2012) and their distinction between templatic and subtractive truncation. Templatic truncation is oriented towards the clipped form, specifically the prosodic form that is the output of the clipping process. Subtractive truncation, by contrast, focuses on the source word and specifies which parts are truncated in the clipping process. A clipping type such as the type *romcom* can be usefully described in terms of templatic truncation. Clippings of this type conform to a template that has two closed syllables and carries initial stress. The speaker's knowledge of the template further specifies that the source word is a compound rather than a non-compound word, and that only initial material of the two component words is maintained. A parallel argument can be made for the clipping type that yields forms such as *stats*, *nugs*, and *deets*. The prosodic template of this clipping type is monosyllabic and it ends in a consonant cluster that contains the plural suffix *-s*. The source word needs to be a plural noun, which is affected by truncation in such a way that the stem is end-clipped and the plural suffix is retained. Crucially, *analysis* is not clipped to *ans*, and neither is *progress* clipped to *progs*. These potential clippings bear a superficial resemblance to the previously mentioned examples, but since they do not derive from plural forms, they do not arise. The relation between a clipping template and its source word can be seen as a paradigmatic link between linguistic forms that are mutually associated. Recent work that combines Construction Grammar with Relational Morphology has elaborated the concept of 'sister links' (Audring 2019) as an organizational principle of linguistic knowledge. Instead of thinking of clipping patterns as cognitive schemas that subsume aspects of the clipping template and the characteristics of the source word, an analysis in terms of sister links would rely solely on the connections between the two, without positing an overarching generalization. While it would be beyond the scope of this paper to work out such an account in detail, we do believe that this approach is highly promising for the analysis of clipping.

As a third point, we submit that the types and antitypes that were identified by the HCFA give rise to hypotheses that can be tested empirically. The issues that were raised in the preceding paragraphs allow us to formulate precise claims that are open to experimental investigation. Generally speaking, we hypothesize that in a forced-choice task in which one option corresponds to a type that has been identified by the HCFA, while the other option represents an anti-type, the type should be preferred. For example, if participants are presented with the word *probation* and the two hypothetical clippings *pro* and *prob*, we hypothesize to see a preference for the latter, which instantiates the type *croc* that was discussed above. Given the word *renovation* and the

options *reno* and *renov*, we expect to see a preference for the former, which corresponds to the type *hydro* that the HCFA has identified. There are more intricate predictions that our results allow us to make. Recall that the *romcom* derives from a compound source word, while this is not the case for the type *hydro*. We therefore predict that given the source word *high-potential*, the choice between the two alternative clippings *high-po* and *high-pot* should be biased towards the latter. However, given the source word *hypotension* and the phonologically identical alternatives *hypo* and *hypot*, we expect a preference for the former. In other words, we hypothesize that the same outcome structures are judged differently depending on the purported source word of the clipping.

Summing up this section, we believe that the points we have raised illustrate that clipping is a far cry from the trivial phenomenon as which it is sometimes portrayed. Developing a theoretically satisfying account of the kind of variation that is exhibited by clippings is a veritable challenge for different linguistic frameworks.

5 Concluding remarks

This paper has presented a newly-compiled database of more than 2000 English clippings, and it has discussed the results of a multivariate analysis that has identified several patterns that reflect regularities in the way speakers of English create shortened words. We have made the case that English clipping is best understood as a set of lower-level generalizations, which allow us to better perceive the reliable tendencies in what at first glance appears to be a bewildering, unconstrained process. While this point has been argued in similar form by López Rúa (2002), we believe that our analysis adds new empirical support and greater depth of description. We have further argued that clipping patterns contain information that relates to both the input and the output of the clipping process, and we have pointed to Relational Morphology as a useful theoretical framework for further analysis. We sketched possibilities to use experimental techniques in order to engage with the predictions that can be derived from our empirical results.

To conclude this paper, we acknowledge that we have had to sidestep several issues with regard to clipping that would have merited closer attention. For one thing, our analysis has not been able to take into account the frequencies with which the clippings in our database appear in authentic language use. The clipping types that we have discussed have been identified on the basis of their type frequency alone. A more thorough investigation of the productivity of clipping patterns would have to draw on relevant corpus-based measures (Baayen 2009), which is something we hope to do in future work. It is also clear that analogy is a phenomenon that cannot be ignored in the context of clipping, especially with regard to highly frequent forms that are not part of a more general pattern, as for example the clipping *bro* (< *brother*), which, according to our results, actually instantiates a clipping anti-type. Monosyllabic clippings with final vowels are underrepresented in our database, so if speakers nonetheless create new

clippings of this kind, this might receive a more compelling explanation in terms of analogy. Clippings further have fascinating semantic and socio-linguistic aspects, which we have not been able to work into the present analysis. Very often, the use of a clipping conveys the idea of shared common ground and a tight social bond between speaker and hearer. In future work, we further hope to contribute to the growing literature that investigates clipping from a cross-linguistic perspective (Nübling 2001; Alber and Arndt-Lappe 2012; Balnat 2012), specifically as regards a contrast between clippings in English and in French. There is thus no shortage of questions that can be asked about clippings and that deserve our close attention.

Supplementary File

The data on which our analysis is based is available as a supplementary file.

Supplementary Material. DOI: https://doi.org/10.16995/glossa.5771.s1

Funding Information

The research reported on in this paper was funded by the Swiss National Science Foundation under grant 100015_188788 ("Clipping in a cross-linguistic perspective", PI Martin Hilpert).

Acknowledgements

The authors would like to express their thanks to three anonymous reviewers who provided critical and constructive comments that we tried to implement to the best of our ability. All remaining errors and inconsistencies are of course our own responsibility.

Competing Interests

The authors have no competing interests to declare.

References

Adams, Valerie. 1973. *An introduction to modern English word-formation*. London: Longman. DOI: https://doi.org/10.4324/9781315504254

Alber, Birgit. 2010. An exploration of truncation in Italian. In Staroverov, Peter & Altshuler, Daniel & Braver, Aaron & Fasola, Carlos A. & Murray, Sarah (eds.), *Rutgers working papers in linguistics* 3. 1–30. New Brunswick: LGSA.

Alber, Birgit & Arndt-Lappe, Sabine. 2012. Templatic and subtractive truncation. In Trommer, Jochen (ed.), *The phonology and morphology of exponence – the state of the art*, 289–325. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199573721.003.0009

Antoine, Fabrice. 2000. *An English-French dictionary of clipped words*. Louvain-la-Neuve: Peeters.

Arndt-Lappe, Sabine. 2018. Expanding the lexicon by truncation: Variability, recoverability, and productivity. In Arndt-Lappe, Sabine & Braun, Angelika & Moulin, Claudine & Winter-Froemel, Esme (eds.), *Expanding the lexicon: Linguistic innovation, morphological productivity, and ludicity*, 141–170. Berlin: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110501933-143

Audring, Jenny. 2019. Mothers or sisters? The encoding of morphological knowledge. *Word Structure* 12(3). 274–296. DOI: https://doi.org/10.3366/word.2019.0150

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook*, 899–919. Berlin: De Gruyter. DOI: https://doi.org/10.1515/9783110213881.2.899

Balnat, Vincent. 2012. L'emprunt de procédés de formation lexicale est-il possible ? Le cas de l'abrègement des mots en allemand, en français et en anglais. In Aïno, Niklas-Salminen &

Steuckardt, Agnès (eds.), *Les langues germaniques*. Travaux du Cercle linguistique d'Aix-en-Provence 23. 181–191.

Bauer, Laurie. 1994. *Introducing Linguistic Morphology*. Edinburgh: Edinburgh University Press.

Berg, Thomas. 2011. The clipping of common and proper nouns. *Word Structure* 4(1). 1–19. DOI: https://doi.org/10.3366/word.2011.0002

Booij, Geert. 2013. Morphology in Construction Grammar. In Hoffmann, Thomas & Trousdale, Graeme (eds.), *The Oxford Handbook of Construction Grammar*, 255–273. New York: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780195396683.013.0014

Booij, Geert. 2019. The role of schemas in Construction Morphology. *Word* 12(3). 385–395. DOI: https://doi.org/10.3366/word.2019.0154

Brinton, Laurel J. & Traugott, Elizabeth C. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/cbo9780511615962

Chomsky, Noam A. & Halle, Morris. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Davy, Dennis. 2000. Shortening phenomena in Modern English word formation: An analysis of clipping and blending. *Franco-British Studies* 29. 59–76.

Don, Jan. 2014. *Morphological theory and the morphology of English*. Edinburgh: Edinburgh University Press.

Durkin, Philip. 2009. *The Oxford guide to etymology*. Oxford: Oxford University Press.

Goldberg, Adele E. 2019. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton: Princeton University Press. DOI: https://doi.org/10.2307/j.ctvc772nn

Gries, Stefan Th. 2004. HCFA 3.2 – A Program for Hierarchical Configural Frequency Analysis for R for Windows.

Gries, Stefan Th. 2006. Cognitive determinants of subtractive word-formation processes: a corpus based perspective. *Cognitive Linguistics* 17(4). 535–558. DOI: https://doi.org/10.1515/cog.2006.017

Gries, Stefan Th. 2008. *Statistik für Sprachwissenschaftler*. Berlin: De Gruyter.

Haspelmath, Martin & Sims, Andrea D. 2010. *Understanding morphology*. Second edition. London: Arnold. DOI: https://doi.org/10.4324/9780203776506

Hilpert, Martin. 2019. Lexicalization in morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. DOI: https://doi.org/10.1093/acrefore/9780199384655.013.622

Jamet, Denis. 2009. A morpho-phonological approach of clipping in English: Can the study of clipping be formalized? *Lexis – E-Journal in English Lexicology* HS1. 15–31. DOI: https://doi.org/10.4000/lexis.884

Krauth, Joachim & Lienert, Gustav A. 1973. *Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin: ein multivariates nichtparametrisches Verfahren zur Aufdeckung von Typen und Syndromen*. Freiburg: Alber.

Kreidler, Charles W. 1979. Creating New Words by Shortening. *Journal of English Linguistics* 13(1). 24–36. DOI: https://doi.org/10.1177/007542427901300102

Langacker, Ronald W. 1999. *Grammar and conceptualization*. Berlin: De Gruyter. DOI: https://doi.org/10.1515/9783110800524

Lappe, Sabine. 2007. *English prosodic morphology*. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-1-4020-6006-9

Levshina, Natalia. 2013. *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins.

López Rúa, Paula. 2002. On the structure of acronyms and neighbouring categories: a prototype-based account. *English Language and Linguistics* 6(1). 31–60. DOI: https://doi.org/10.1017/s136067430200103x

McCarthy, John. 2011. *Doing Optimality Theory: Applying Theory to Data*. Malden, MA: Blackwell.

Nooteboom, Sieb G. 1981. Lexical retrieval from fragments of spoken words: beginnings vs. endings. *Journal of Phonetics* 9. 407–424.

Nübling, Damaris. 2001. Auto–bil, Reha–rehab, Mikro–mick, Alki–alkis: Kurzwörter im Deutschen und Schwedischen. *Skandinavistik* 31(2). 167–199.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/cbo9780511841323

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Sánchez-Gutiérrez, Claudia H. & Mailhot, Hugo & Deacon, S. Hélène & Wilson, Maximiliano A. 2018. MorphoLex: A derivational morphological database for 70,000 English words. *Behavioral Research* 50. 1568–1580. DOI: https://doi.org/10.3758/s13428-017-0981-8

Tournier, Jean. 1985. *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Paris-Genève: Champion-Slatkine.

Zipf, George K. 1949. *Human behavior and the principle of least effort*. Cambridge: Addison Wesley.