ICAME Journal • 47(1) • 2023 • 19–37
DOI: 10.2478/icame-2023-0002

sciendo

ICAME Journal

Meaning differences between English clippings and their source words: A corpus-based study

Research Article

Martin Hilpert*, David Correia Saavedra, Jennifer Rains

Université de Neuchâtel

Received 28 September 2022; accepted 2 January 2023

Abstract: This paper uses corpus data and methods of distributional semantics in order to study English clippings such as dorm (< *dormitory*), *memo* (< *memorandum*), or quake (< *earthquake*). We investigate whether systematic meaning differences between clippings and their source words can be detected. The analysis is based on a sample of 50 English clippings. Each of the clippings is represented by a concordance of 100 examples in context that were gathered from the Corpus of Contemporary American English. We compare clippings and their source words both at the aggregate level and in terms of comparisons between individual clippings and their source words. The data show that clippings tend to be used in contexts that represent involved text production, which aligns with the idea that clipped words signal familiarity with their referents. It is further observed that individual clippings and their source words partly diverge in their distributional profiles, reflecting both overlap and differences with regard to their meanings. We interpret these findings against the theoretical background of Construction Grammar and specifically the Principle of No Synonymy.

Keywords: *clipping • COCA • distributional profile • Construction Grammar • Principle of No Synonymy*

1 Introduction

This paper uses corpus data in order to study English clippings such as *dorm* (< *dormitory*), *memo* (< *memorandum*), or *quake* (< *earthquake*), which are formed on the basis of existing lexical items from which phonological material is deleted. More specifically, we investigate meaning differences between clippings and their source words. Are there semantic or pragmatic differences between pairs such as *dorm* and *dormitory*, and if so, what are they? Is it possible to formulate generalizations that capture how clippings and their source words relate to one another with regard to meaning? How should these relations be described? While questions of this kind have been raised in previous work on clippings, research that addresses these issues on the basis of quantitative empirical data is scarce. Commenting on this gap in the research landscape, Alber and Arndt-Lappe (2012: 314) observe that semantic questions have received relatively little attention in the study of clippings, and they remark that systematic studies of meaning in truncatory processes are virtually absent. The present paper aims to take a step towards addressing these issues.

Our study focuses on a set of 50 English clippings and their corresponding source words. For each clipping and each source word, we retrieve corpus data from the Corpus of Contemporary American English (Davies 2008), which serves as the basis for our analyses. We undertake comparisons at the level of the entire set in order to determine if there are general tendencies that distinguish clippings and their source words, and we also present individual comparisons that identify how specific clippings differ semantically from their respective source words. The corpus analysis reveals general distributional asymmetries that suggest a difference relating to involved vs. informational text production (Biber 1988). Clippings have a relatively greater tendency to appear in texts with contextual elements that reflect text production in which the speaker shows a high degree of involvement, for example by making

* *Corresponding author: Martin Hilpert, E-mail: martin.hilpert@unine.ch*

reference to the listener and to themselves. Clippings appear to be preferred in contexts in which there is substantial common ground between speaker and hearer, which aligns with the notion that clippings signal familiarity with the ideas that are conveyed (Wierzbicka 1984; Quirk et al. 1985; Plag 2003; Katamba 2005). For the individual comparisons between clippings and their source words, we draw on the distributional semantic method of token-based semantic vector spaces (Hilpert and Correia Saavedra 2020). Pairs such as *cardio-cardiovascular*, *chemo-chemotherapy*, and *intro-introduction* are analyzed in terms of their collocational behavior. The method allows us to pinpoint aspects of meaning that are specifically associated with a clipping, rather than its source word, and vice versa, while also revealing how their respective meanings overlap. Our findings show that many clippings are semantically distinct from their source words, but we also document cases in which the collocational profile of the clipping is indistinguishable from that of the source word.

The remainder of this paper is structured as follows. Section 2 offers a discussion of how the issue of meaning has been treated in the existing literature on clippings. We outline several theoretical positions and define the assumptions that we bring to our analysis. Section 3 addresses the corpus data that forms the basis for the present study, and it explains how that data has been collected and analyzed. Section 4 presents the results. In Section 4.1 it is shown that clippings differ from their source words with regard to the frequency of specific contextual elements, such as common nouns, present tense verb forms, and first and second person pronouns. Section 4.2 offers semantic comparisons between individual clippings and their respective source words. We discuss cases of semantic divergence as well as cases in which the clipping and its source word are near-synonymous. Section 5 takes a step back and discusses what the empirical results imply for the study of clippings in general. We interpret our findings against the theoretical background of Construction Grammar (Goldberg 1995; Hilpert 2019) and specifically discussions of the Principle of No Synonymy (Uhrig 2015; Levshina and Lorenz 2022). Section 6 summarizes the main conclusions of the paper and points towards issues for further research.

2 The meaning of clippings

Is the clipping *fridge* distinct in meaning from its source word *refrigerator*? Speakers of English have a choice between the two options, but does that mean that there is a tangible difference? It is difficult to address that question on the basis of linguistic intuition alone, and also a qualitative inspection of individual examples from corpus data will not provide a satisfying answer. In the existing literature on clippings, various positions have been expressed with regard to their meanings. For the sake of exposition, these perspectives can be arranged on a continuum that assume greater or lesser degrees of meaning difference between clippings and their source words. At one extreme of the continuum, the semantic distance between the two is thought to be minimal, or even non-existent. For example, Dressler (2000: 4) argues that abbreviations and clippings such as *GOP* (< *Grand Old Party*) or *mic* (< *microphone*) do not, as a matter of principle, change the meaning of their source words. Clippings and their source words are thus viewed as mutual alternatives that only differ in form. Speakers' choices between those alternatives are thought to be contingent on contextual factors. For example, a speaker may want to verbalize the concept 'microphone' in a speech situation in which all participants are highly familiar with the concept and with one another, and a referent is clearly identifiable. The speaker would thus retrieve the word *microphone* from memory and, influenced by the context of the speech situation, realize it as *mic*.

Whereas many clippings indeed appear to be very close in meaning to their respective source words, some clippings conventionally express meanings that are clearly distinct from those of their respective source words. For example, the clipping *bro* (< *brother*) is not a kinship term, but it is a word that is used to address a close friend within specific communities of speakers. This meaning has become lexicalized over time, so that in current language use *bro* is to be seen as a lexical item of its own, rather than as a variant of *brother*. Accounts that view clippings and their source words as semantically indistinguishable in principle do allow for diachronic meaning change through lexicalization, so that forms such as *bro* can be fully accommodated in such accounts.

A position that is somewhat different, but still compatible with the semantic identity of clippings and their source words, is that clippings may distinguish themselves from their source words in terms of their connotations, even when their referential meanings are assumed to be exactly the same. The social dimension of these connotations is especially important. Plag (2003: 117) states that clippings such as *lab* (< *laboratory*) are used to express familiarity, and he argues that this is a feature that is shared across clippings of common nouns and truncated names. Similar statements can be found in Wierzbicka (1984) and Quirk et al. (1985: 1584). Through their usage in situations in which there is substantial common ground between speakers, clippings come to serve as markers of group membership and shared expertise. By the same token, clippings can mark the boundaries of social groups, for example when a speaker uses a clipping in a situation in which only a part of the audience is able to make sense of it. A doctor who is in a consultation with a patient and a nurse might address the nurse using specialized terminology, including clippings, that are not easily understood by the patient, who might thus feel excluded. Antoine (2000: xxxviii) comments on such effects by stating that clippings "have an undeniable social meaning beyond their lexical meaning: they denote a stance on the part of the speaker, who affirms and claims his familiarity with a closed milieu". Plag (2003: 121) points out that this does not apply to clippings in general, but only to a specific subset, as there are clippings that do not seem to be tied to in-groups of speakers. Clippings such as *fridge* (< *refrigerator*) or *ad* (< *advertisement*), which are found in general language use, do thus not appear to signal familiarity, expertise, nor membership in a specific community.

Allowing for the fact that different types of clippings may have their own specific semantic characteristics, Jamet (2009: 19) proposes an account in which three scenarios are highlighted. The first of these, which Jamet argues to be the most frequent case, is the semantic identity of clippings and their source words. A clipping such as *memo* (< *memorandum*) would illustrate this. In a second scenario, the clipped form is marked as colloquial. Whereas the referential meaning is thus identical to the meaning of its source word, the clipped form connotates informal language use. The clipping *fam* (< *family*) falls into that category. The third scenario is exemplified by clippings that have taken on specialized meanings in particular fields. To illustrate, the clipping *subs* (< *subscribers*) is used in the context of online media, where it for example describes the persons that have subscribed to a given YouTube channel. Crucially, the clipping *subs* is not used to the same extent in order to talk about people who subscribe to a print newspaper or a book club. Similarly, the clipping *cab* is used by rock musicians to refer to a loudspeaker cabinet, or by wine lovers to talk about a Cabernet Sauvignon, or by car experts who use it to denote a cabriolet. Kreidler (2000: 962) notes similar examples. For cases of this kind, Jamet (2009: 19) argues that the meaning of the clipping has narrowed and is perceived to convey a more technical meaning than the full form. For the purposes of the present paper, Jamet's account makes an interesting prediction, namely that meaning change in clippings will typically result in semantic narrowing. The empirical comparisons that we discuss in later sections of this paper allow us to examine that issue.

The present paper adopts a view of clippings and their meanings that owes its general orientation to what Goldberg (1995: 67) has termed the Principle of No Synonymy. That principle holds that a difference in linguistic form will always indicate a difference in meaning, either with regard to semantic aspects or concerning discourse-functional characteristics of the form in question. We thus work with a broad notion of meaning that includes information-structural, discourse-functional and interpersonal aspects, among other facets of linguistic meaning and function. A prediction that follows from the Principle of No Synonymy is that in authentic language use, clippings and their source words should have distinct distributional characteristics that reflect their functional differences. In line with usage-based theories of language (Bybee 2010), we adopt the position that language use shapes and reflects speakers' knowledge of language. The meaning of a linguistic form relates to the speech situations in which it has been witnessed, so that patterns in corpus data allow conclusions not only about how clippings are used, but also about how they are stored and processed by actual speakers. We further subscribe to what is known as the distributional hypothesis (Firth 1957; Turney and Pantel 2010), which holds that the meaning of words is reflected in their contextual elements in language use. Words that appear in similar contexts can be shown to share aspects of their meanings. For example, the words

cardiovascular and *hypertension*, which are semantically related, appear in contexts that share many common collocates such as *heart*, *disease*, *diabetes*, *stroke*, and others. In this paper, we use corpus data from the COCA (Davies 2008) to compare clippings and their source words, as for example *cardio* and *cardiovascular*, in terms of their distributional behavior. We present evidence in order to suggest that there are tangible differences between clippings and their source words, and that despite a fair amount of semantic overlap, it is reasonable to maintain that once a clipping has established itself in language use, speakers will treat it as a separate lexical element. To make an important caveat right away, the clippings that we analyze in this paper are all well-established in the variety that the corpus represents, namely American English. Clippings that are formed in an ad hoc fashion naturally have not had the opportunity to establish conventionalized usage patterns in a community of speakers. Yet, it can be hypothesized that even newly coined clippings will express familiarity in the way that is described by Antoine (2000: xxxviii) or Plag (2003: 117). We argue that the meaning of familiarity is reflected in measurable characteristics of corpus data, which thus suggests a general difference between clippings and their source words.

3 Data and methodology

For the present study, 50 clippings were selected from a database of English clippings that was created by Hilpert et al. (2021), who used the data in order to argue that English clippings instantiate a range of different structural types that are statistically overrepresented. For example, the database contains many clippings such as *croc* (< *crocodile*), which are monosyllabic, end in a consonant, maintain the stressed syllable of the source word, and cut into a morpheme boundary. Another clipping type is instantiated by forms such as *hydro* (< *hydrothermal*), which are disyllabic, end in a vowel, break up at a morpheme boundary, and bear initial stress, whereas their source words are stressed on a different syllable. Since the analysis presented by Hilpert et al. (2021) does not focus on meaning per se, it has no direct implications for the analysis in the present paper beyond the fact that it provides data as input. The 50 clippings that were selected for the present study are listed in Table 1, along with their source words. The selection of clippings is a convenience sample. Inclusion was motivated by the following considerations. First, the text frequency of the clipping needed to be high enough to yield at least 100 true examples for both the clipping and its corresponding long form. Many clippings in the database are infrequent even in contemporary mega-corpora and thus do not yield enough data for a thorough analysis. For example, clippings such as *depresh* (< *depression*) or *nilla* (< *vanilla*) only yield low counts of true hits in the COCA (Davies 2008). Second, the meanings of the clippings needed to be clearly identifiable on the basis of their linguistic context. Clippings such as *comp*, *prog*, or *sub* are used with meanings that correspond to various different source words. The clipping *comp* may thus refer to *compensation*, *compression*, *composite*, *complementary*, or *comprehensive school*, among other meanings, and in many cases the context of a single concordance line is not sufficient to fully disambiguate the clipped form. By contrast, with clippings such as *ammo* (< *ammunition*) or *expat* (< *expatriate*), the rate of true positives is much higher, and even a small context window allows a reliable identification of the intended meaning. As can be seen in Table 1, a large proportion of the selected clippings are nouns, which reflects the general noun bias that has been observed in the literature on clippings (Tournier 1985: 299; Jamet 2009: 17).

For each clipping type in Table 1, concordance lines with a context window of ten words to the left and right were retrieved from the COCA (Davies 2008). We used the 2014 release of the off-line version of the corpus, which contains 440 million word tokens. The concordance lines were manually checked in order to exclude false positives. For example, the form *alum* is not only a clipping with the meaning ‘alumnus’, but it also appears in texts where it means ‘aluminum’. The form *lab* instantiates clipped forms of *laboratory* and *Labrador*. Non-target concordance lines were identified and removed from the data. For the analysis, 100 concordance lines instantiating the correct meaning of each clipping were randomly selected. Following the same procedure, matching concordances of 100 examples were retrieved from the COCA for each of the source words listed in Table 1.

Table 1. Selected clippings and their long forms.

	clipping	source word		clipping	source word		clipping	source word
1	admin	administration	18	expat	expatriate	35	mic	microphone
2	alum	alumnus	19	fab	fabulous	36	pic	picture
3	ammo	ammunition	20	fave	favorite	37	porn	pornography
4	amp	amplifier	21	frat	fraternity	38	prefab	prefabricated
5	app	application	22	fridge	refrigerator	39	prenup	prenuptial agreement
6	boomer	baby-boomer	23	gent	gentleman	40	prof	professor
7	burbs	suburbs	24	glam	glamorous	41	psych	psychology
8	carbs	carbohydrates	25	goth	gothic	42	pup	puppy
9	cardio	cardiovascular	26	hippo	hippopotamus	43	quake	earthquake
10	celeb	celebrity	27	improv	improvisation	44	rehab	rehabilitation
11	chemo	chemotherapy	28	indie	independent	45	roach	cockroach
12	chimp	chimpanzee	29	intro	introduction	46	stats	statistics
13	croc	crocodile	30	lab	laboratory	47	teen	teenager
14	decaf	decaffeinated	31	legit	legitimate	48	tux	tuxedo
15	dorm	dormitory	32	limo	limousine	49	undergrad	undergraduate
16	exam	examination	33	lube	lubricant	50	vid	video
17	exec	executive	34	memo	memorandum			

For the subsequent analytical steps, the search terms, that is, the clippings and their source words, were removed from the concordance lines, so that only the contextual elements remained in the concordance files. Each concordance file thus comprises 100 lines with 20 context elements, so that each clipping and each source word is represented by exactly 2000 context elements. These context elements contain both lexical and grammatical words, as well as punctuation, as no stop words were removed from the concordance. We would like to point out that it was a deliberate choice for us to work with a modest amount of data. It was our goal to see if reliable differences could be observed on the basis of relatively small samples of corpus data. Our analysis can thus be seen as a pilot study that aims to deliver a proof of concept, which can subsequently be scaled up to larger amounts of data and applied to other sets of clippings and their source words.

The set of 50 clipping concordances and 50 source word concordances that was selected for this study allows us to make comparisons of how clippings and their source words differ in terms of their contexts. At the level of the entire set, we can determine whether certain types of contextual elements, such as for example pronouns or prepositions, are asymmetrically distributed across the data, so that they occur more frequently in the clipping concordances, and less frequently in the source word concordances. In order to focus on an individual clipping and its source word, we can check if and how the respective concordances differ, for example with regard to the presence or absence of specific collocates. The technique that we use for that purpose is token-based semantic vector space modeling (Hilpert and Correia Saavedra 2020), which is described below.

Semantic vector space modeling (Turney and Pantel 2010) is a technique that operationalizes the meaning of linguistic elements in terms of their distributional characteristics. If two linguistic elements share overlapping sets of collocates that occur with similar frequencies, this is taken to reflect a close semantic relationship between these two elements. In Section 2, we used the example of the words *cardiovascular* and *hypertension* to illustrate this point. The two words are semantically related, as both

of them relate to aspects of the blood flow inside the human body. In actual language use, both appear in contexts in which collocates such as *heart*, *disease*, *diabetes*, and *stroke* are common. Semantic vector space models make use of such regularities and represent the meaning of linguistic elements in terms of vectors that specify for a large set of contextual elements how strongly each of those elements is attracted to a target word, as for example the target word *cardiovascular*. Such vectors are created on the basis of corpus data. Concordance lines of a target word are retrieved, co-occurrence frequencies of the target word and its contextual elements are determined, and the co-occurrence frequencies of the target word and each of its contextual elements are weighted via an association measure. A commonly used method for this purpose is Positive Pointwise Mutual Information (Levshina 2015: 327). The resulting vector reflects the target word as a type, as it is based on many instances of the target word and the contextual elements that occur before and after those instances. The semantic similarity between two target words can be captured through calculations of the cosine distance between their context vectors. For larger sets of target words, pairwise comparisons between all possible target word pairings yield similarity matrices that can be used to visualize semantic relations in the entire set via hierarchical clustering or through dimension-reduction techniques such as multi-dimensional scaling (Wheeler 2005) or t-SNE (van der Maaten and Hinton 2008).

The analyses in the present paper use an application of semantic vector space modeling that aims to differentiate between the meanings of individual tokens of language use. Token-based semantic vector space models (Schütze 1998; Heylen et al. 2012, 2015; Hilpert and Correia Saavedra 2020) try to account for meaning differences across instances of the same linguistic unit. For example, the English noun *syntax* has several related senses, one of which relates to the ordering of constituents in human language, and another one that refers to conventions in computer programming languages. The two senses can be distinguished in terms of the contextual elements that appear to the left and right of a given token of *syntax*. Contextual elements such as *ditransitive* or *morphology* indicate that the linguistic sense is intended, whereas elements such as *error* or *HTML* are cues for the computational sense. Token-based semantic vector space models distinguish between instances of language use that are represented by concordance lines, but importantly, the analysis is not solely based on the context elements that are found in those concordance lines. Rather, the technique enriches the information that is present in any single concordance line by taking into account second-order collocates. For example, a concordance line of the word *syntax* may contain the contextual elements *ditransitive* and *accusative*, among others. For each of these contextual elements, further information is collected by means of retrieving concordances from additional corpus data and determining the frequencies of elements in those concordances. Along the lines of what was described above, type-based context vectors for *ditransitive*, *accusative*, and all remaining elements of the concordance line are gathered and merged into a single vector that offers a rich semantic representation of the concordance line. By virtue of including information about second-order collocates, the representation of the concordance line now captures that the specific token of *syntax* is semantically related to words such as *dative* or *morphological*, even though these are not directly present in the concordance line itself.

Token-based semantic vector space models have been applied successfully to the study of both lexical semantics (Montes 2021) and meaning differences in grammatical items such as modal verbs (Hilpert and Flach 2020). In this paper, we follow the methodology that is described in Hilpert and Correia Saavedra (2020), whose implementation relies on Pointwise Mutual Information as an association measure, uses the cosine as a similarity measure, and draws on metric multidimensional scaling as a reduction technique. We use the method to create token-based semantic vector spaces that are based on concordance lines of clippings and their respective source words. The visualizations of these vector spaces allow us to determine if there are distributional differences between clippings and their source words, and more specifically, it makes it possible to identify both aspects of meaning that show overlap and other aspects that allow us to distinguish between clippings and their source words.

4 Results

This section presents the findings that were obtained on the basis of the dataset and the analytical methods that were described in Section 3. We discuss observations at the aggregate level of the entire dataset before moving on to individual comparisons between clippings and their source words.

4.1 Aggregate comparisons between clippings and their source words

In the introduction to this paper we asked if there were any systematic meaning differences between clippings and their source words. We specified in Section 2 that we take a broad perspective on meaning that includes information-structural and discourse-functional aspects. Taking as our point of departure the observation that clippings tend to be used to signal familiarity with a referent (Antoine 2000: xxxviii; Plag 2003: 117), we derive the prediction that our data should reveal frequency asymmetries between clippings and their source words that pertain to linguistic markers of enhanced common ground between the speaker and the hearer. As a way of operationalizing this, we turned to work by Biber (1988), specifically the distinction between involved text production and informational text production. As one central dimension of textual variation among others, Biber (1988: 104–107) notes striking distributional differences between two types of discourse. The first represents interpersonal interaction and the expression of affect and involvement, whereas the second is detached from interpersonal matters and high in informational density. The two types of discourse are characterized by linguistic features that show gradient degrees of association with either involved or informational text production. For the purposes of the present study, we have selected twelve markers from those identified by Biber (1988: 102). Table 2 lists the elements in question according to their respective textual preference.

A few comments are in order to explain how the features shown in Table 2 reflect the circumstances of language use. Private verbs such as *know*, *mean*, or *think* are used by speakers to express ideas to which they have privileged mental access and that they would like to share with hearers. Texts that are meant to convey information in a detached, objective way will feature fewer private verbs than spontaneous conversation between speakers who know each other well. Contractions such as *it's* or *you're* are typical for informal spoken conversation with a high degree of intersubjective alignment. In speech, contractions can lead to homophony, so that for example the phonological string [ɪts] may represent *it is*, *it has*, or possessive *its* (Biber 1988: 106). In such cases, the speaker relies on the hearer to identify the correct interpretation. Present tense verbs are used to verbalize ideas that are of immediate relevance and often related to the context of the speech situation. First and second person pronouns serve to make direct reference to the speaker and the hearer. Most verbalizations of their

Table 2. Twelve features of involved and informational text production (Biber 1988: 102).

Involved Text Production	Informational Text Production
private verbs*	complementizer THAT
contractions	common nouns
present tense verbs	prepositions
1st person pronouns	type/token ratio
2nd person pronouns	
indefinite pronouns**	
pronoun IT	
demonstrative pronouns	

*Following Biber (1988: 242), the verbs that were included in this category were *anticipate*, *assume*, *believe*, *conclude*, *decide*, *demonstrate*, *determine*, *discover*, *doubt*, *estimate*, *fear*, *feel*, *find*, *forget*, *guess*, *hear*, *hope*, *imagine*, *imply*, *indicate*, *infer*, *know*, *learn*, *mean*, *notice*, *prove*, *realize*, *recognize*, *remember*, *reveal*, *see*, *show*, *suppose*, *think*, and *understand*.
**Following Biber (1988: 226), the pronouns that were included in this category were *anybody*, *anyone*, *anything*, *everybody*, *everyone*, *everything*, *nobody*, *none*, *nothing*, *nowhere*, *somebody*, *someone*, and *something*.

thoughts and actions necessitate the use of first and second person pronouns. Indefinite pronouns such as *anything* or *everyone*, the third person pronoun *it*, and demonstrative pronouns such as *this* or *those* have in common that their deictic nature requires the hearer to identify the referent on the basis of contextual information. Pronouns of this kind are associated with spontaneous informal discourse (Biber 1988: 113). On the right-hand column of Table 2, the first feature is the presence of *that* in complement clause constructions. Whereas the complementizer is commonly left unexpressed in involved texts, informational texts show a greater tendency for *that* to be present. A high ratio of common nouns and prepositions is typical for texts with increased information density. In the same vein, a high type-token ratio indicates that a text is lexically diverse and that repetitions of the same items are relatively rare.

How do the twelve features relate to clippings? Our data indicates that the features shown in Table 2 differ in usage frequency in the respective contexts of clippings and their source words. The boxplots in Figure 1 show distributional differences between the 50 clippings and their source words with regard to the first six types of contextual elements that are listed in the first column of Table 2. All six panels show normalized frequencies of the contextual elements on the y-axis, as well as central tendencies and the distribution around those tendencies for clippings and their source words. From the existing literature, we derived the expectation that clippings should be associated with features of involved text production. In accordance with this idea, all six types of contextual elements are relatively more frequent in the clipping concordances. Figure 2 visualizes the remaining features that are listed in Table 2. The frequency of the pronoun *it* is higher in the clipping concordances, which is in line with our expectations. We also expected that the respective frequencies of nouns, prepositions, and overt complementizers would be higher in the source word concordances, which turns out to be the case.

The remaining two panels in Figure 2 show results that were unexpected. First, there is no measurable frequency difference for demonstrative pronouns, which have been identified as a feature of involved text production (Biber 1988: 102). In our data, demonstrative pronouns are not more prevalent in the clipping concordances. Furthermore, informative texts are generally observed to have a higher type-token ratio than involved texts. Our data indicates that there is no difference between the clipping concordances and the source word concordances. With regard to this result, it is important to keep two factors in mind. First, our comparisons are not based on running texts. Our data is assembled from individual concordance lines that are randomly sampled from larger corpus files. Also, at 2000 words, the concordances that we compare are relatively short.

Table 3 summarizes the results of paired t-tests that indicate significant differences for ten of the twelve features that we have investigated. The tests for two features (demonstrative pronouns, type-token ratio) yield non-significant results.

These findings suggest that clippings and their source words exhibit systematic differences with regard to the text types in which they are found. The observed differences map onto the distinction between involved and informational text production that has been described by Biber (1988: 107), with only two features that do not align with the general contrast. We argue that that clippings exhibit this distributional behavior because speakers use them to mark familiarity with the entities and ideas that are referred to. Importantly, an alternative interpretation is possible. Involved text production typically happens under real-time conditions that exert various pressures on the speaker, so that there is limited time to formulate an utterance, and there are limited cognitive resources available to support that process. It could be that clippings are systematically favored under such conditions, simply because they can be produced by the speaker with less effort. If the meaning of an utterance is relatively predictable, and alternative forms are available to verbalize that meaning, speakers will tend to choose a variant that is easier to produce (Levshina and Moran 2021). From the data that we have presented above, it is unfortunately not possible to decide whether the distributional differences are primarily due to an intended expression of familiarity or the tendency to use language efficiently. Can the question be settled at all? Levshina and Lorenz (2022) present empirical data that is highly relevant for this issue. Specifically, they investigate variation in speakers' choices between English *want to* and *wanna*. The reduced variant should occur in contexts in which it is highly predictable. However, data from the British National Corpus shows that the effect of predictability is limited and appears only in contexts in which the speech rate is high. Levshina

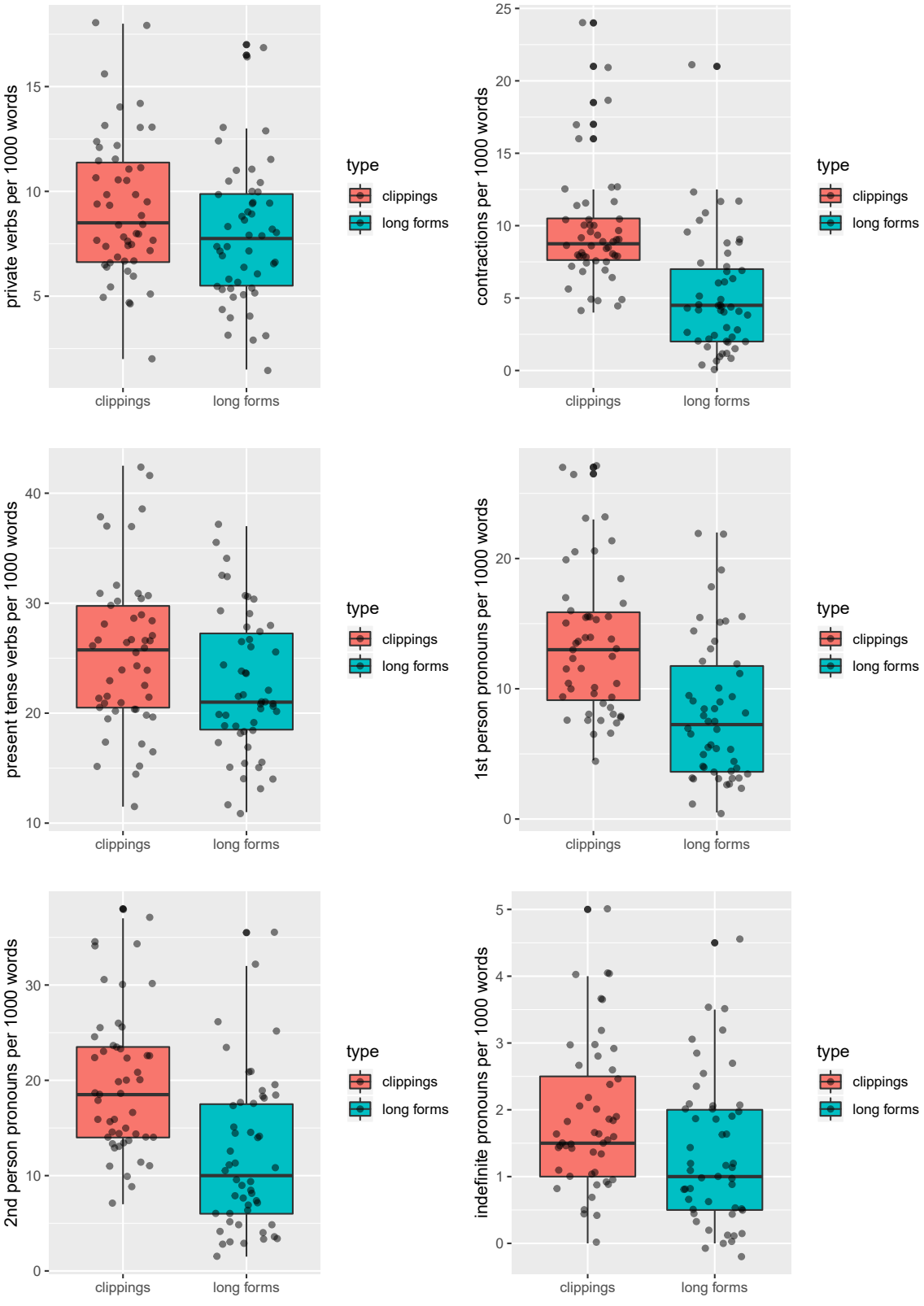

Figure 1. Contextual elements that distinguish between clippings and their source words.

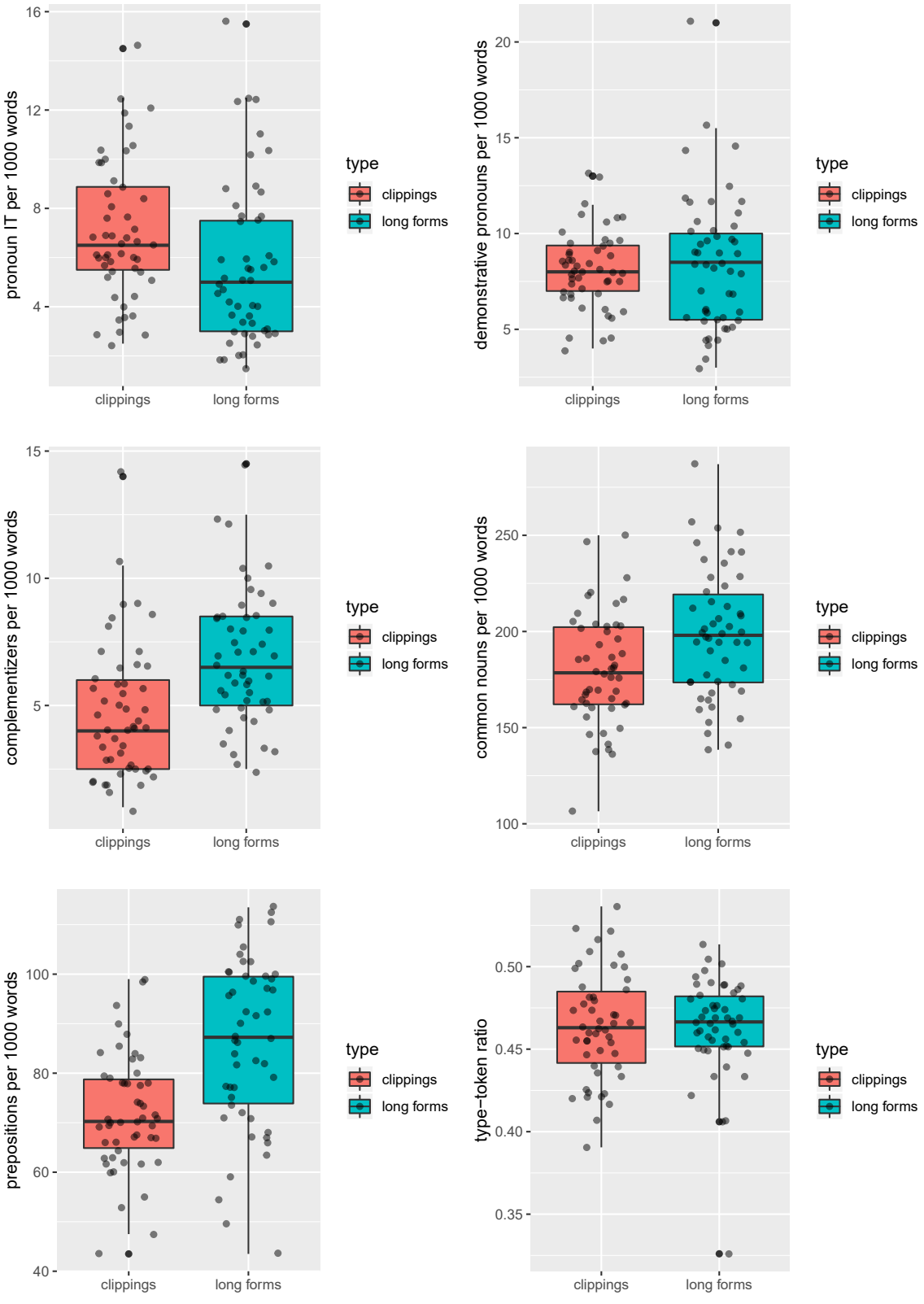

Figure 2. Contextual elements that distinguish between clippings and their source words.

Table 3. Results of paired t-tests for twelve features of involved and informational text production.

	Feature	t	df	95% CI	p	Sig	Attraction
1	private verbs	2.38	49	0.41 4.87	0.021	**	clippings
2	contractions	6.92	49	6.31 11.49	<0.001	***	clippings
3	present tense verbs	2.65	49	1.40 10.28	0.011	**	clippings
4	1st person pronouns	5.79	49	6.99 14.41	<0.001	***	clippings
5	2nd person pronouns	4.16	49	2.66 7.62	<0.001	***	clippings
6	indefinite pronouns	2.59	49	0.25 1.99	0.012	**	clippings
7	pronoun IT	2.73	49	0.71 4.69	0.008	***	clippings
8	demonstrative pronouns	-0.56	49	-2.82 1.59	0.57	n.s.	
9	complementizer THAT	-5.05	49	-5.9 -2.54	<0.001	***	long forms
10	common nouns	-4.16	49	-54.9 -19.2	<0.001	***	long forms
11	prepositions	-6.69	49	-37.1 -19.9	<0.001	***	long forms
12	type/token ratio	0.19	49	-0.01 0.01	0.84	n.s.	

and Lorenz (2022: 269) explain this result in terms of meaning differences between the two variants. Predictability will influence speakers' choices when two forms have largely identical meanings, but it will not do so if the two forms differ in terms of their semantics or pragmatics or their sociolinguistic characteristics. With regard to clippings, this means that efficiency could come into play if and only if a clipping and its source word have more or less identical meanings, as perhaps in the case of *fridge* and *refrigerator*. In order to arrive at a more conclusive interpretation of the results that were discussed in this section, we thus need to analyze the data at a finer level of granularity, contrasting individual pairs of clippings and their source words in terms of their distributional behavior. If it can be shown that clippings and their source words typically differ in aspects of their meanings, the conclusions from Levshina and Lorenz (2022) detract from the idea that efficiency can be held responsible for the asymmetries we observe in our data.

4.2 Individual comparisons between clippings and their source words

In this section, we use token-based semantic vector space modeling in order to determine if the meanings of word pairs such as *fridge* and *refrigerator* can be distinguished on the basis of their respective distributional behaviors. The analyses are based on the same data that was used for the aggregate comparisons of clippings and their source words that we presented in the previous section. Our approach thus starts with concordance lines such as the ones that are shown in (1), which illustrate the clipping *cardio* and its source word *cardiovascular*:

(1) a. fatigue your muscles by the final two reps. The fat-blasting **cardio** workout Use this weekly cardio calendar to gradually boost your
b. ticker to delicious dishes built around the best foods for **cardio** health # BY LORI POWELL # Bananas # These potassium-and
c. Welfare; 2008. # 9. Prevention of **Cardiovascular** Disease, Diabetes and Chronic Kidney Disease: Targeting Risk
d. small airways, setting the stage for respiratory and **cardiovascular** disease. these long effects appear to trace to Fresno's

All four concordance lines express ideas that relate to the heart and the blood vessels of the human body. Yet, it is also apparent that the examples focus on distinct semantic aspects. Concordance lines (1c) and (1d), which feature the long form *cardiovascular*, verbalize problems that can affect the human body. The word *disease* appears in both concordance lines, along with other words from the medical domain. Intuitively, it is easy to see that concordance lines (1a) and (1b), which feature the clipping *cardio*, express semantic facets that are generally more positive, relating to training routines and nutritional habits that support cardiovascular health. Token-based semantic vector space models

create representations of concordance lines such as the ones shown in (1) on the basis of second-order collocates. To illustrate how this is done, consider concordance line (1b), which contains the words *delicious*, *dishes*, *foods*, and *bananas*. In order to capture the meaning of that concordance line, the method creates a vector representation that takes into account the collocational preferences of those words. That is, the semantic vector that represents concordance line (1b) characterizes it in terms of words that are associated with the items that appear in it. To take an example, the word *delicious* is strongly associated with words such as *recipes* or *fruit*. The specific token of *cardio* that is used in concordance line (1b) is thus associated with the second-order collocates *recipes* and *fruit*. By contrast, the token of *cardio* that appears in concordance line (1a), which contains words such as *fatigue*, *muscles*, and *workout*, has different second-order collocates.

In order to determine the profiles of second-order collocates for the tokens of clippings and source words in our data, we utilize a type-based semantic vector space that was prepared by Hilpert and Correia Saavedra (2020: 396), and that contains semantic vectors for a set of 12,621 English words, each of which is represented by a vector of PMI values for 12,621 contextual elements. For all concordance lines in our data, we determine which words of the concordance line are represented in the type-based semantic vector space. If there are not at least four such words, the concordance line is discarded. This step considerably reduces the number of concordance lines in the data, so that on average 80 of the 100 concordance lines are discarded. Yet, this procedure guarantees that only concordance lines containing a sufficient amount of information enter the analysis. If a concordance line contains at least four words that are contained in the type-based semantic vector space, those words are looked up, and their semantic vectors are retrieved and averaged, so that each concordance line is represented by a single vector. We have implemented the threshold of at least four valid context elements with the goal of enhancing the reliability of our results. The data that remains in the analysis is semantically more representative of the key words than the data that was discarded. At the same time, reducing the data in this way comes with a risk. If the data contains only relatively few reliable examples, this means that the results might not show a clear difference because of data sparsity. While including more data points could have counteracted that risk, we have preferred to err on the side of caution. Table 4 offers an illustration of the results of that process. For the concordance lines shown above in (1), it lists the second-order collocates with the highest PMI values. What can be seen is that the method considerably enriches the semantic information that is present in any single concordance line. Whereas some of the words that are present in the concordance lines reappear as second-order collocates, the lists contain further items that are conceptually related and that flesh out the semantic representations of the concordance lines. For example, the top second-order collocates of concordance line (1c) include *inflammatory*, *arthritis*, and *liver*, which are not present in the concordance line itself, but which are semantically related to the words that appear in the concordance line.

Once these vector representations have been obtained, it can be assessed how the concordance lines of a given clipping, such as for example *cardio*, differ from one another, and from concordance lines of the source word *cardiovascular*. For the purpose of such comparisons, the vectors that represent concordance lines of pairs such as *cardio* and *cardiovascular* are combined into a data frame, so that pairwise similarities between all concordance lines can be calculated, using the cosine as a similarity measure. The resulting similarity matrix can then be used in order to analyze both overlap and differences in the meanings of clippings and their source words, and it can further reveal whether there is semantic variation within the concordance lines of a clipping, or within those of a source word. For the present analysis, we used metric multidimensional scaling in order to project the similarity matrices of clippings and their source words onto two-dimensional scatterplots.

Two sets of pairings of clippings and their source words are shown in Figures 3 and 4. The pairs have been selected from the available data so as to include a set with clear differences and a set with less clearly pronounced differences. The motivation for this selection is to showcase the kinds of result that the method produces for different word pairs. In Figure 3, we visualize data from the pairs *cardio-cardiovascular*, *chemo-chemotherapy*, *dorm-dormitory*, *indie-independent*, *intro-introduction*, and *pic-picture*. Each dot in the panels of Figure 3 represents a concordance line. The colors of the dots

Table 4. Second-order collocates of tokens of *cardio* and *cardiovascular*.

Concordance line	Key word	Second-order collocates
(1a)	cardio	muscles, slam, abdominal, eyelids, stamina, circulation, nerves, headache, calves, ...
(1b)	cardio	recipes, oranges, cooked, pies, dishes, salads, sweets, foods, fried, dried, fruit, ...
(1c)	cardiovascular	diabetes, inflammatory, disorders, chronic, disease, renal, arthritis, kidney, liver, ...
(1d)	cardiovascular	chronic, obstruction, lung, asthma, epithelial, disease, hypertension, prevalence, intermittent, ...

distinguish concordance lines of clippings (shown in red) and concordance lines of source words (shown in green). Dots that are placed in close mutual proximity represent concordance lines with highly similar collocational profiles. The panels of Figure 3 show that for all six pairs, the profiles of second-order collocates allow us to achieve a fairly accurate discrimination between the clippings and their source words. The following paragraphs discuss what specific insights can be gleaned from the graphs.

With regard to the pairing of *cardio* and *cardiovascular*, it was suggested above that the source word is used in contexts in which various medical conditions are described, whereas clippings tend to appear in contexts that discuss sports and fitness. In the first panel of Figure 3, there is only one token of *cardiovascular*, shown in the upper left of the graph, that goes against this generalization. That token corresponds to the concordance line shown in (2), which relates to the topic of fitness and is thus similar to uses of *cardio*:

(2) flexibility, endurance, posture, balance, coordination and **cardiovascular** health. A trainer can help you learn what to

The remaining tokens of *cardiovascular*, which form a relatively tight cluster, instantiate medical uses of the term. It can be seen that the dots that represent the clipping *cardio* are distributed relatively more widely across the graph. While all concordance lines of *cardio* relate to the topic of fitness in some way, they include a broad set of words that leads to more diffuse profiles of second-order collocates.

The pairing of *chemo* and *chemotherapy* yields a highly similar result. The clipping and its source word differ clearly in terms of their second-order collocates. The concordance lines of the source word contain precise medical vocabulary and typically come from scientific writing. By contrast, concordance lines with the clipping *chemo* tend to verbalize the experience of individual patients. While the concordance lines with *chemo* also occasionally contain medical terms, they are less prevalent and there is greater semantic variety in the elements that are present. The examples in (3) offer illustrations of this:

(3) a.left leg above the knee. She also underwent intensive **chemo**, six rounds of three treatments each. Chemo
b.Once my temperature was stabilized, I lay in a **chemo** fog while Nancy fought for my release and got me

These results from *cardio* and *chemo* stand in contradiction to the idea that clippings that differ in meaning from their source words have a tendency to be semantically narrower and more technical in meaning (cf. Jamet 2009: 19). Rather, it appears that clippings of technical medical terms have been co-opted into everyday language use, where they appear in collocational contexts that are wider than those in which the source words are commonly found. We will come back to this point in Section 5.

In the middle row of Figure 3, the second-order collocates of *dorm-dormitory* reveal significant semantic overlap on the right of the graph and differentiation towards the left. The examples that represent the semantic overlap commonly include words that relate to university life, as shown in the examples in (4). This contrasts with the uses in (5), which refer to more comfortable tourist lodgings, and which appear to be restricted to *dorm*:

(4) a.infrastructures. First, the campus is connected. All **dormitory** rooms, classrooms, offices, libraries, laboratories
b.out the admissions policy and a plan to finance a **dorm** at Southern Tech in Marietta before all regents were informed

(5) a.apart. A larger cottage sleeps five in an upstairs **dorm** with a spacious living room and fireplace, kitchen and
b.by William Morris. There was a fireplace in my **dorm** room. When I saw that I laughed out loud

The pairing of *indie* and *independent* is distinguished clearly on the basis of second-order collocates. The clipping has a specific meaning that is related to underground culture and that finds its expression in collocations such as *indie rock*, *indie film*, *indie label*, and others. It thus exemplifies a type of clipping that has narrowed in meaning, vis-à-vis its source word, along the lines envisioned by Jamet (2009). Similarly, the contrast of *intro* and *introduction* is partly due to specialized meanings of the clipping, which can for instance denote the initial part of a song or musical piece. Still, *intro* and *introduction* semantically overlap in examples such as the ones shown in (6), which denote academic subjects. By contrast, the examples of *introduction* in (7) would not be felicitous with the clipping instead of the long form:

(6) a.on horizontal notebook paper, was taped to it: **INTRODUCTION** TO GRAPHIC DESIGN Formerly mislabeled Introduction to Commercial Art
b.into the lexicon. Schools were overcrowded. In her **Intro** to Sociology text a map of the geometric increase in

(7) a.become a major preoccupation of theoretical physicists. Regarding his **introduction** of the cosmological constant in 1917, Einstein's real
b.The major features of this design included (a) independent **introduction** of the treatment to each parent group with continued baseline

In the last panel, Figure 3 visualizes the differences between *pic* and *picture*. The clipping is semantically more narrow than its source word. A *pic* usually refers to a photo, whereas a *picture* can refer to a much wider range of meanings. One exceptional use of *pic*, in which the clipping refers to a movie, is shown in the upper left corner of the graph.

Not all clippings in our data can be distinguished reliably from their source words. Figure 4 visualizes concordance lines for six clippings and their source words that exhibit considerable overlap in their semantic potential. Still, the graphs and the concordance lines that are represented by them can be used to identify different aspects of meaning that characterize the respective semantic potentials. For example, the contrast between *exam* and *examination* reveals that the upper right area of the graph is exclusively populated by concordance lines featuring the long form. The examples in (8) illustrate the meaning of *examination* that is at issue, which is that of a journalistic investigation, rather than a medical or academic test. The clipping *exam* would be inappropriate in these contexts:

(8) a.published Bureaucrats in Power: Ecological Collapse, a detailed **examination** of the inefficiencies of Soviet centralized decision-making. Zabelin
b.International Relations: Understanding the Behavior of Nations. In-depth **examination** of why nations compete, cooperate, and sometimes go

Figure 4 also allows us to address the intriguing question of whether *fridge* actually means the same thing as *refrigerator*. Our analysis points to an answer in the affirmative. The concordance lines in our data refer to different semantic facets of refrigerators. Some concordance lines mention the beverages or food items that are stocked, others discuss decorative items such as magnets that are found on the outside of a refrigerator, still others refer to technical specifications or other household appliances. Importantly, either of these facets can be verbalized with either the clipping or the long form. Very similarly, the concordance lines for *porn* and *pornography* relate to semantic aspects such as the industry behind pornography, psychological effects of pornography consumption, as well as legal concerns. Both the clipping and its source word can appear in these contexts.

To briefly summarize the findings that have been presented in this section, we can first of all point out that token-based semantic vector space modeling serves to detect meaning differences between clippings and their source words. In some cases, the distributional characteristics of a clipping and its source word diverge considerably, underscoring the fact that the two have semantically diversified, so that there are some senses that are exclusively expressed by either the clipping or the source

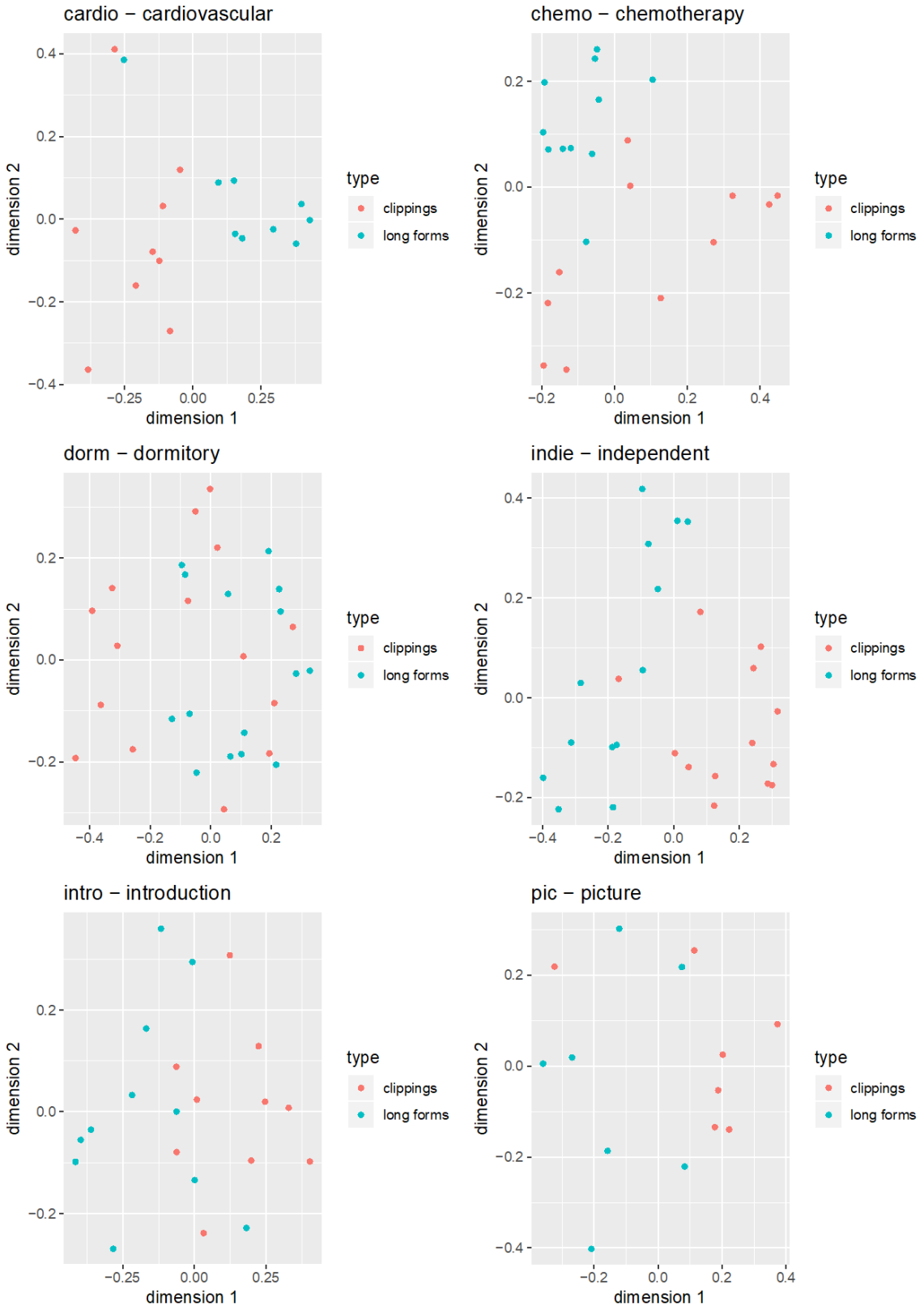

Figure 3. Clippings and source words with divergent distributional behaviors.

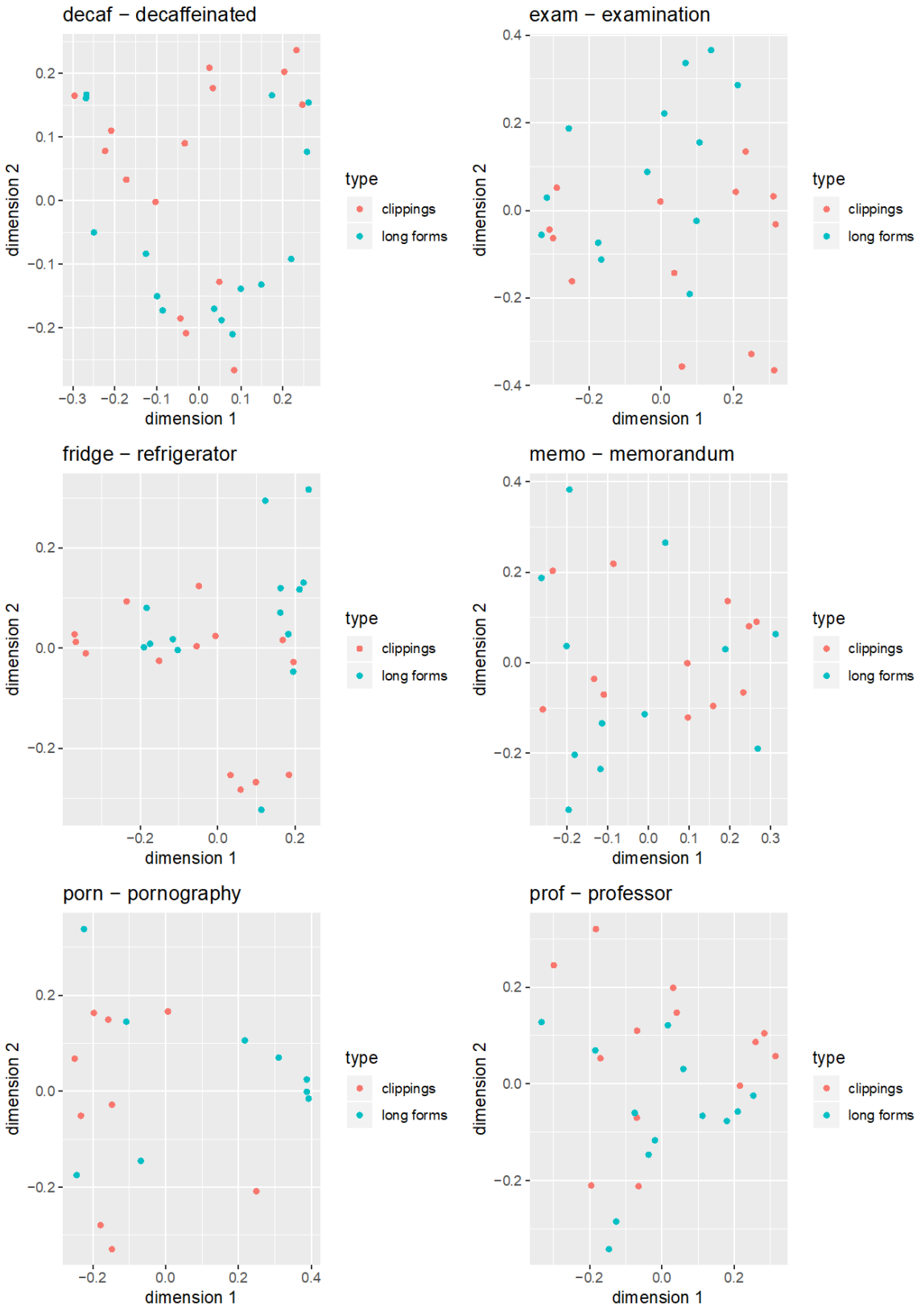

Figure 4. Clippings and source words with convergent distributional behaviors.

word, while other senses constitute areas of semantic overlap. We have also shown that there are pairs of clippings and source words that are indistinguishable in terms of their profiles of second-order collocates. For these cases, the analysis allows us to distinguish between facets of meaning that are characterized by different sets of second-order collocates. In the following section, we interpret these findings against a broader theoretical background.

5 Discussion

Do clippings differ semantically from their source words, and if so, how? Earlier in this paper, we stated that our theoretical position with regard to this question aligns with the Principle of No Synonymy (Goldberg 1995: 67), which makes the working assumption that a difference in form will reflect a difference in meaning, allowing for a broad conception of the latter. In other words, we hypothesized that pairs such as *intro* and *introduction* should differ with regard to their denotation, their contexts of use, their register, or any combination of those factors. In Section 4.1, we showed that clippings and their source words differ in terms of ten types of contextual elements, such as first and second person pronouns, present tense verbs, nouns, and prepositions. These asymmetries in language use are consistent with the idea that speakers use clippings in order to mark familiarity with their referents (Wierzbicka 1984; Quirk et al. 1985; Plag 2003; Katamba 2005). We also considered the efficiency of language use as a possible explanation. Given two linguistic forms that express the same meaning, speakers will prefer the form that is easier to produce in contexts where that form is easily predicted by the hearer. The findings that were presented in Section 4.2 document clippings that are semantically distinct from their source words. Taking our cues from findings presented by Levshina and Lorenz (2022), this leads us to argue that speakers' choices in these cases are motivated by meaning rather than efficiency. For cases in which clippings overlap semantically with their source words, we documented facets of meaning that are preferentially or even exclusively expressed by one of the two alternatives. The scenario in which efficiency could have an effect is the one in which there is near-complete semantic overlap between the clipping and its source word, and in which the semantic facet that the speaker wants to verbalize is associated in equal measures with the clipping and with the source word. Our data suggest that this kind of scenario is relatively rare. We would also like to submit that a heterogeneous meaning potential that is expressed both by a clipping and its source word, as was observed for example with the pair *porn-pornography*, may over time lead to a semantic division of labor between the two alternatives.

The association of clippings with the meaning of familiarity or personal involvement in the subject of conversation is further supported by several qualitative observations that we have made on the basis of our data. In pairs such as *chemo* and *chemotherapy*, the clipping is used in contexts in which a personal experience is at stake, whereas the source word appears in discussions that are technical and detached. This is also true for *cardio* and *cardiovascular*, and to some extent for *exam* and *examination*. Cases such as these serve as counterexamples to a generalization formulated by Jamet (2009: 19), according to which clippings should undergo semantic narrowing and convey specialized, technical meanings.

While we take these results as an indication that the Principle of No Synonymy holds up well with regard to English clippings, it is useful to consider critical issues that call into question the blanket application of the principle. Uhrig (2015: 331) identifies a problem that is relevant for our analysis. He points out that a strict interpretation of the principle would make it difficult to account for phenomena such as intra-speaker variation. The examples in (9) show that speakers occasionally use the words *memo* and *memorandum* in the same utterance, with the intent of denoting the same referent:

(9) a. Whit Johnson now joins us from the White House this morning. And Whit, we've heard about that Secret Service **memorandum**, the internal **memo** about new rules and regulations. What can you tell us about it?
b. My point was, in the context of this **memo**, in the context of this **memorandum**, the question was whether or not the court should in fact have a heightened scrutiny.

How could this variation be accounted for? If we grant the idea that one function of clippings is to signal familiarity, example (9a) can be interpreted in terms of the speaker introducing a referent and then treating it as mutually known. Example (9b) shows the inverse order, so that the clipping appears first. The repetition in the example suggests that the speaker backtracks in order to issue a repair and present a more appropriate word choice. Uhrig's (2015: 331) point is well-taken, and a systematic analysis of intra-speaker variation in the use of clippings is definitely needed. What the two examples illustrate is that our general conclusions do not preclude such variation, but rather provide hypotheses that could be tested in future work.

6 Concluding remarks

In the literature on English clippings, the topic of meaning, and specifically the question of meaning differences between clippings and their source words, has not received a lot of attention. It was the aim of the present paper to address this gap in the research landscape and to develop a few suggestions for the corpus-based study of clippings and their meanings. Our general conclusion is that corpus data can yield useful insights into the semantics of clippings. We found general tendencies such as the affinity of clippings with linguistic elements that signal involvedness, we observed that many clippings diverge semantically from their source words, and we documented and analyzed semantic contrasts between individual clippings and their source words. It goes without saying that a lot of work remains to be done. For example, in the preceding section we briefly touched on the subject of variation and the factors that would govern speakers' behavior in scenarios in which they are faced with the choice between a clipping and its source word. Since the focus of the present paper has been squarely on meaning, we have not taken into account the full scope of factors that would be at play in such a scenario, which is bound to involve effects of priming, frequency, efficiency, cognitive complexity, and others, alongside the semantic factors that we have discussed above. In future work, we hope to address these factors, along with other dimensions of variation that relate to text type, to different varieties of English, and to other sociolinguistic factors. Another question that would merit attention is how clippings are affected by meaning change. We have argued that pairs of clippings and their source words would be prone to undergo semantic diversification over time. It remains an open question whether any unifying tendencies can be identified that would account for clippings in general, or at least for a large number of cases. Pairs such as *fridge* and *refrigerator*, which are largely synonymous and have been semantically stable for quite some time, indicate that broad generalizations may be hard to come by. Lastly, it would be a worthwhile, if challenging, endeavor to investigate if speakers of English show any convergent tendencies when they process the meanings of newly formed, ad-hoc clippings. For this, we would need to turn to experimental, psycholinguistic methods. Our assessment that clippings signal the involvedness of the speaker could be tested for example in experimental designs in which one condition presents ad-hoc clippings in contexts that express familiarity, whereas another condition presents the clippings in a neutral context. We hope that the results that we have reported in this paper inspire further discussions of how the meanings of clippings can be usefully investigated, and we look forward to continuing the discussion. This was only the intro, the exam is not over.

References

Alber, Birgit and Sabine Arndt-Lappe. 2012. Templatic and subtractive truncation. In J. Trommer (ed.). *The phonology and morphology of exponence – the state of the art,* 289–325. Oxford: Oxford University Press.

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

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Dressler, Wolfgang U. 2000. Extragrammatical vs. marginal morphology. In U. Doleschal and A. M. Thornton (eds.). *Extragrammatical and marginal phonology*, 2–10. Munich: Lincom.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, 1–32. Oxford: Blackwell.

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

Heylen, Kris, Dirk Speelman and Dirk Geeraerts. 2012. Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH: Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, 16–24. https://dl.acm.org/doi/pdf/10.5555/2388655.2388658.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman and Dirk Geeraerts. 2015. Monitoring polysemy. Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157: 153–172.

Hilpert, Martin. 2019. *Construction Grammar and its application to English*. 2nd edition. Edinburgh: Edinburgh University Press.

Hilpert, Martin and David Correia Saavedra. 2020. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 16 (2): 393–424. https://doi.org/10.1515/cllt-2017-0009.

Hilpert, Martin and Susanne Flach. 2020. Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities* 36 (2): 307–321. https://doi.org/10.1093/llc/fqaa014.

Hilpert, Martin, David Correia Saavedra and Jennifer Rains. 2021. A multivariate approach to English clippings. *Glossa: A Journal of General Linguistics* 6 (1): 104. https://doi.org/10.16995/glossa.5771.

Jamet, Denis. 2009. A morphophonological approach to clipping in English: Can the study of clipping be formalized? *Lexis: Journal in English Lexicology* 1: 15–31.

Katamba, Francis. 2005. *English words*. 2nd edition. New York: Routledge.

Kreidler, Charles W. 2000. Clipping and acronymy. In G. Booij, C. Lehmann and J. Mugdan (eds.). *Morphologie/ morphology: Ein internationales Handbuch zur Flexion und Wortbildung/An international handbook on inflection and word-formation*. Vol. 1, 956–963. Berlin: Walter de Gruyter.

Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis.* Amsterdam: John Benjamins.

Levshina, Natalia and Steve Moran. 2021. Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard* 7 (Suppl. 3): 20200081.

Levshina, Natalia and David Lorenz. 2022. Communicative efficiency and the Principle of No Synonymy: Predictability effects and the variation of *want to* and *wanna*. *Language and Cognition* 14 (2): 249–274. https://doi:10.1017/langcog.2022.7.

Montes, Mariana. 2021. *Cloudspotting. Visual analytics for distributional semantics*. Doctoral dissertation. University of Leuven.

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

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. New York: Longman.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24 (1): 97–124.

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

Turney, Peter and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37: 141–188.

Uhrig, Peter. 2015. Why the Principle of No Synonymy is overrated. *Zeitschrift für Anglistik und Amerikanistik* 63 (3): 323–337. https://doi.org/10.1515/zaa-2015-0030.

Van der Maaten, Laurens J. P. and Geoffrey E. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.

Wheeler, Eric S. 2005. Multidimensional scaling for linguistics. In R. Koehler, G. Altmann and R.G. Piotrowski (eds.). *Quantitative linguistics. An international handbook*, 548–553. Berlin: De Gruyter.

Wierzbicka, Anna. 1984. Diminutives and depreciatives: Semantic representation for derivational categories. *Quaderni di semantica* 5 (1): 123–130.